Journée Francophones de la Toile - JFT'2003

30 juin, 1 et 2 juillet 2003

École Polytechnique de l'Université de Tours
Département Informatique
64 avenue Jean Portalis
37200 Tours - France

Table des matières

Avant propos

L'idée qui a motivé la création de ces journées est le rassemblement des chercheurs francophones travaillant sur toutes les théories, techniques et applications liées au Web, aussi bien dans le secteur académique et universitaire que dans l'industrie. Jusqu'à présent plusieurs conférences nationales traitaient de thèmes plus spécifiques ou reliés indirectement au Web mais il manquait certainement un événement regroupant toutes les tendances de la recherche scientifique dans ce domaine. Ce manque a été apparemment ressenti par d'autres chercheurs si l'on en juge par le grand nombre de propositions que nous avons reçues (60, dont 46 articles et 14 démonstrations), ce qui pour une conférence toute neuve, est fort encourageant. Le point central des JFT'2003 a donc été de faire partager les travaux innovants effectués dans le domaine du Web dans une atmosphère scientifique conviviale (et tourangelle !).

Il me faut d'abord adresser mes plus vifs remerciements à tout le comité scientifique, où chacun a bien voulu croire au succès de l'aventure et répondre présent alors qu'aucun historique ne laissait présager de la suite. La sélection des articles publiés (25 en communication orale, 6 en poster, 11 démonstrations) a été menée avec rigueur pour aboutir à un taux de sélection garantissant la qualité des travaux présentés pour ces premières journées.

Je dois remercier également les institutions qui ont soutenu ces journées et dont les aides financières et matérielles ont permis de réaliser ces journées dans les meilleures conditions pour tous : le Conseil Général d'Indre et Loire, le Conseil Régional du Centre, l'Université de Tours, le Département Informatique de l'Ecole Polytechnique de l'Université de Tours et le Laboratoire d'Informatique de l'Université de Tours. Sans l'aide de ces institutions, les journées n'auraient pas pu voir le jour sous leur forme et leur dimension actuelles.

Pour finir, tout le comité d'organisation ainsi que l'équipe Réseaux-TIC espérons que ces journées et les travaux qui y ont été présentés vous seront profitables. Nous espérons également pouvoir compter sur votre présence et votre intérêt pour les suites qui seront données dans le futur à cette conférence.

Gilles Venturini,

Président du Comité Scientifique
des Journées Francophones de la Toile 2003.

Les thèmes abordés dans la conférence sont principalement les suivants et restent très largement ouverts à toute la recherche scientifique et technique sur le Web :

- Interfaces et Utilisateurs : Interfaces non visuelles, Interfaces et sites adaptatifs, Esthétisme, Analyse d'audience et de profils utilisateurs, Analyse comportementale, Facteurs humains
- Web sémantique (WS) : Langages pour le WS, Ontologies pour le WS, Méta-données et annotation dans le WS, Systèmes de médiation, Adaptation/personnalisation dans le WS, Bases de données et de connaissances
- Calcul global et services Web : Calcul global et mobilité, Architectures pair à pair, Services et données, Description de services
- Web et 3D : Visualisation, Réalité virtuelle, Serveur de données 3D
- Documents et multimédia : Hypertextes pour le Web (génération, analyse, utilisation), Art numérique sur le Web, Documents électroniques et multimedia, Formats, Accessibilité des contenus aux personnes handicapées
- Web mining : Moteurs de recherche, Recherche d'informations, Veille stratégique, Fouille de textes, Extraction de connaissances
- Hébergement et génération automatique de sites : Générateurs de sites, Sites portails, Evaluation automatique, Utilisabilité

Cette manifestation a été organisée avec le soutien de :

- le Conseil Général d'Indre et loire
- le Conseil Régional de la région Centre
- l'Université de Tours
- le Département Informatique de l'École polytechnique de l'Université de Tours
- le Laboratoire d'Informatique de l'Université de Tours

Comité scientifique

- Dominique ARCHAMBAULT, INOVA, INSERM, Université P. et M. Curie
- Bernd AMANN, INRIA et CNAM
- Jean-Pierre BALPE, Paragraphe, Université de Paris 8
- Dominique BURGER, INOVA, INSERM, Université P. et M. Curie
- Frank CAPPELLO, LRI, Université de Paris 11
- Daniel DARDAILLER, W3C France et Europe
- Alexandre DELTEIL, France Télécom R&D
- Jérôme EUZENAT, INRIA
- Christian FLUHR, CEA
- Benoît HABERT, LIMSI, Université de Paris 11
- Christian JACQUEMIN, LIMSI, Université de Paris 11
- Yves KODRATOFF, LRI, Université de Paris 11
- Benedicte LE GRAND, LIP6, Université de Paris 6
- Alain LELU, Paragraphe, Université de Paris 8
- Christian LICOPPE, France Telecom R&D
- Denis MAUREL, LI, Université de Tours
- Nicolas MONMARCHÉ, LI, Université de Tours
- Emanuel NÉRON, LI, Université de Tours
- Antoine OLIVER, LI, Université de Tours
- Vincent QUINT, INRIA et W3C
- Chantal REYNAUD, LRI, Université de Paris 11
- Marie-Christine ROUSSET, LRI, Université de Paris 11
- Jean-David RUVINI, Bouygues' e-lab
- Imad SALEH, Paragraphe, Université de Paris 8
- Mohamed SLIMANE, LI, Université de Tours
- Michel SOTO, LIP6, Université de Paris 6
- Brigitte TROUSSE, INRIA
- Gilles VENTURINI, LI, Université de Tours (Président)
- Djamel A. ZIGHED, ERIC, Université de Lyon 2

Comité d'organisation locale

- Ali ALARABI
- Sébastien AUPETIT
- Hanène AZZAG
- Christophe GUÉRET
- Thierry HÉNOCQUE
- Nicolas LABROCHE
- Christophe LENTÉ
- Patrick MARTINEAU
- Denis MAUREL
- Nicolas MONMARCHÉ
- Antoine OLIVER
- Fabien PICAROUGNE
- Mohamed SLIMANE (Président)
- Ameur SOUKHAL
- Fabrice TERCINET
- Gilles VENTURINI

Index des Auteurs

Première partie

Communications

Journées Francophones de la Toile - JFT'2003

Web Sémantique

Représentation de métaconnaissances pour le développement de Webs Sémantiques d'Organisation

C. CORMIER, J.-Y. FORTIER, G. KASSEL, C. BARRY
Laboratoire de Recherche en Informatique d'Amiens
Université de Picardie Jules Verne
33, rue Saint-Leu
80039 Amiens Cedex 1, FRANCE
Mail : cormier@u-picardie.fr
Tél : +33 3 22 82 88 75

Résumé

Les travaux que nous menons visent à développer des Webs Sémantiques d'Organisation hybrides, réalisant un couplage entre une Base de Connaissances (BC) et une Base de Documents (BD). Les connaissances capitalisées se trouvent donc, pour partie, dans des modèles de connaissances et, pour partie, dans des textes. Concernant le contenu de la BC, nous préconisons de modéliser en priorité l'*organisation* pour laquelle la mémoire est développée et l'*information* que contient la mémoire, en faisant abstraction de son mode de spécification (modèle de connaissances ou texte). Dans cet article, nous présentons l'apport du langage réflexif d'ontologie DefOnto (défini et implanté dans notre équipe) pour représenter les différents modèles (de l'organisation et de l'information) et leur ontologie associée. Plus particulièrement, nous montrons l'intérêt de la représentation de métaconnaissances pour rendre compte du modèle de l'information. Nous montrons également l'utilisation des services inférentiels de DefOnto pour le développement d'un moteur de recherche exploitant les différents modèles de connaissances.

Abstract

Our research project aims at elaborating hybrid Organizational Semantic Webs, based on a coupling between a Knowledge Base (KB) and a Documents' Base (DB). The capitalized knowledge is therefore distributed among the KB and the DB. Concerning the contents of the KB, we advocate to model the organization for which the memory is developed in priority and the information that the memory contains, while making abstraction of its mode of specification. In this paper, we present the contribution of DefOnto, a reflective ontology language defined in our team, to support the different models and their associated ontology. In particular, we emphasize the necessity of having metaknowledge representation capabilities in order to deal with the model of the information. We also present the use of the inferential services of DefOnto for the development of a search engine exploiting the different knowledge models.

1 Introduction

Les travaux que nous menons concernent la conception et le développement de Webs Sémantiques d'Entreprise, ou plus généralement d'« Organisation » (WSO). Il s'agit de systèmes de « mémoires d'entreprise », au sens où les définissent [Dieng-Kuntz et al., 2001] : *une mémoire d'entreprise est la représentation persistante, explicite, désincarnée, des connaissances et des informations dans une organisation, afin de faciliter leur accès, leur partage et leur réutilisation par les membres adéquats de l'organisation, dans le cadre de leurs tâches*, dont l'implantation exploite les technologies du Web, notamment du Web sémantique. Plus particulièrement, notre projet de recherche vise à développer des WSOs hybrides, réalisant un couplage – fort – entre une Base de Connaissances (BC) et une Base de Documents (BD).

Les architectures courantes de WSOs reposent sur le couplage d'une BD (des ressources Web) avec des annotations et une ontologie, le rôle de ces deux dernières ressources « sémantiques » étant de constituer un index pour les documents [Gandon et al., 2000][Staab et Maedche, 2001]. Un tel couplage peut être qualifié de « faible » dans la mesure où le seul rôle de la BC est de faciliter l'exploitation de la BD, les connaissances capitalisées ne se trouvant que dans la BD. Au contraire, nous préconisons de réaliser un couplage « fort » en distribuant les connaissances à la fois dans la BC et la BD. Concernant le choix des connaissances à modéliser, nous proposons de modéliser prioritairement l'organisation pour laquelle le WSO est développé [Fortier et al., 2002][Kassel, 2002].

Un tel couplage fort pose toutefois le problème de la diffusion des connaissances modélisées. En effet, la BC étant spécifiée dans un langage formel, son contenu est difficilement appréhensible par un être humain. De plus, contrairement aux textes, la structure des modèles de connaissances n'obéit pas à un objectif de transmission d'informations ciblées, sur un sujet précis et pour un lectorat précis.

Pour pallier ces difficultés, nous avons proposé récemment d'ajouter à la BC un nouveau modèle, à savoir un modèle des informations que le WSO est capable de présenter à propos de l'organisation [Fortier and Kassel, 2003]. Un tel modèle offre deux principaux avantages pour le développement de fonctionnalités de recherche et de diffusion des informations :

- D'une part, il permet de construire un index du contenu informationnel du WSO en faisant abstraction de la façon dont l'information est spécifiée et, de ce fait, de sa localisation. Un tel index constitue pour l'utilisateur une aide à la définition de son besoin d'information.
- D'autre part, il offre un cadre bien adapté pour représenter des métaconnaissances, à savoir des connaissances portant sur la confidentialité des informations, un modèle de l'utilisateur et également une indication de la localisation de l'information (BC ou BD).

Dans cet article, nous présentons l'apport du langage de représentation d'ontologies DefOnto, défini et implanté dans notre équipe, pour rendre compte de ces différents modèles et de leur ontologie associée. Plus particulièrement, l'objet de l'article est de justifier l'apport des capacités de représentation de métaconnaissances de DefOnto pour représenter le modèle de l'information et son articulation avec le modèle de l'organisation. Nous suivons pour cela le plan suivant :

- La section 2 décrit le contenu d'un modèle de l'organisation et montre comment représenter en DefOnto un tel modèle. La représentation correspondante mobilise des propriétés d'objets. Il est ainsi fait usage de capacités de représentation analogues à celles de langages comme DAML-OIL [Horrocks et al., 2002a] ou OWL [Dean and Schreiber, 2003].
- La section 3 décrit le contenu d'un modèle de l'information et sa représentation. Cette fois-ci, la représentation mobilise des propriétés de concepts et de propositions. Elle fait ainsi appel à des capacités de représentation de métaconnaissances que ne possèdent pas les langages DAML-OIL et OWL, mais qui sont pourtant préconisées pour les langages d'ontologies pour le Web sémantique [Heflin, 2003]
- La section 4 décrit l'utilisation des services inférentiels de DefOnto pour implanter un moteur de recherche basé sur un index du contenu informationnel du WSO.
- La section 5 positionne DefOnto par rapport aux autres langages d'ontologie pour le Web actuellement définis et présente les développements en cours sur le langage.

2 Représentation du modèle de l'organisation au moyen d'une théorie d'objets

2.1 Contenu d'un modèle de l'organisation

Un modèle de l'organisation correspond à la description, selon différents points de vue, de cette organisation. Prenons l'exemple d'un projet, exemple que nous suivrons tout au long de l'article.

En tant qu'organisation, c'est-à-dire en tant qu'un groupe de personnes réalisant un projet, un projet comporte des participants, a généralement un chef de projet et peut être doté d'un comité de pilotage. En tant que processus complexe, ce projet peut être décomposé en tâches (ou « lots »), elles-mêmes donnant lieu à la réalisation d'activités de natures diverses (ex : réunion de travail, rédaction de document, développement de logiciel, *etc.*). Enfin, un projet produit des résultats, certains tangibles (ex : des logiciels, des documents) d'autres intangibles (ex : une méthode conceptuelle).

Un tel modèle fait donc intervenir des objets de natures très diverses. Le rôle de l'ontologie associée au modèle est justement d'expliciter le sens de ces différentes catégories d'entités. L'ontologie contient ainsi une spécification de notions comme les notions de « tâche », de « rapport intermédiaire » ou de « comité de pilotage ». Se doter d'une telle ontologie d'organisation est du reste un préalable à l'expression d'un modèle de l'organisation.

Tel que nous l'avons défini, le modèle de l'organisation inclut un modèle de sa documentation. On peut donc considérer que modéliser l'organisation selon l'ensemble des points de vue revient à étendre l'approche des méta-données, qui ne sont qu'un ensemble de descriptions restreintes aux seuls documents. Quant au contenu de ce modèle et aux points de vue à considérer, il faut considérer qu'ils dépendent des tâches de gestion des connaissances que le WSO doit assister, tâches qui dépendent à leur tour des objectifs ayant motivé le développement du WSO (ex : faciliter la diffusion des documents, aider à l'intégration d'un nouvel employé dans l'organisation)[1].

2.2 Représentation du modèle de l'organisation

La représentation en DefOnto du modèle de l'organisation et de son ontologie associée fait appel au « noyau » de DefOnto, autrement dit à ses capacités de représentation de base.

Ce noyau correspond à la représentation d'une *théorie d'objets*, c'est-à-dire en un ensemble d'axiomes assertant des propriétés vérifiées par des objets appartenant à un certain domaine de discours. En l'occurrence, le domaine de discours rassemble l'ensemble des objets évoqués en §2.1, autrement dit des personnes, des activités, des organisations, des documents, *etc*. Trois types de primitives de représentation sont pour cela distinguées et définies : des *objets individuels*, des *classes d'objets* et des *relations*.

Une définition d'objet individuel (*cf.* figure 1) – introduite par le constructeur *DefIndObject* – consiste ainsi en l'assertion d'un ensemble de propriétés vérifiées par un unique objet. Par convention, les propriétés exprimant l'appartenance de cet objet à une (ou plusieurs) classe(s) sont placées dans l'en-tête de la définition. Elles sont repérées par le mot-clef « isA ». Les autres propriétés sont placées après le mot-clef « ObjectProperties ». Ces dernières sont construites à partir de relations binaires (ex : *aPourMembre*, *estAuteurDe*) et expriment un lien avec un objet individuel (ex : *J. Dupont*, *Rapport-17*) ou une donnée, placée entre guillemets (ex : *"Dupont"*). Par définition, nous appelons « concept individuel » l'ensemble des propriétés portant sur l'objet individuel. Nous utilisons de fait le terme « concept » au sens d'une « notion », ou encore d'un « objet intensionnel ». Informellement, un concept individuel correspond à l'idée que se fait un agent d'un objet individuel.

[1] Pour une synthèse récente sur les ontologies et modèles d'entreprise, le lecteur pourra consulter la référence [Gandon, 2002].

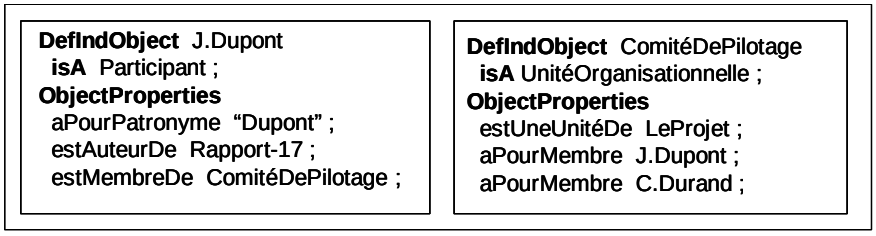


FiG. 1 – Représentation en DefOnto d'objets individuels

Le modèle de l'organisation se trouve ainsi représenté au moyen d'une collection de définitions d'objets individuels. L'ontologie associée est pour sa part représentée au moyen de classes d'objets et de relations.

Une définition de classe d'objets (*cf.* figure 2a) - introduite par le constructeur *DefClassObject* – rassemble un ensemble de propriétés vérifiées par tous les objets de la classe. Sa structure est analogue à celle de la définition d'un objet individuel. On note toutefois quelques différences :

- Les axiomes incluent des quantificateurs (ex : AE, AI, ME, MI) attachés syntaxiquement au nom de la relation. Ainsi en est-il de l'axiome *DocumentÉlectronique (AE)aPourFormat String* pouvant se paraphraser en français par : « tout document électronique a un format qui est une chaîne de caractères »[2]. La signification logique des axiomes avec quantificateurs (utilisés dans l'article) est précisée dans le tableau 1 ci-dessous. Nous avons distingué pour cela deux logiques : la Logique du 1er ordre et la Logique de Description (avec sa notation standard).
- Les liens de subsomption peuvent être complétés d'une différence, comme dans l'exemple de la définition des concepts *ResponsableDeTâche* et *Document-Électronique*. Les quantificateurs associés sont 'ME' ou 'MI'[3]. De tels liens reviennent à exprimer une condition nécessaire et suffisante d'appartenance d'un objet à la classe.
- Des propriétés portant sur la classe peuvent compléter la définition. Elles sont placées après le mot-clef « ClassProperties ». Les seules propriétés de classe autorisées reposent sur la relation prédéfinie *disjoint*.

Une définition de relation (*cf.* figure 2b) – introduite par le constructeur *DefRelation* – rassemble, d'une part, un unique lien de subsomption (sans différence) mentionnant une sur-relation et, d'autre part, des propriétés de la relation (analogues aux propriétés d'une classe d'objets). Ces dernières reposent sur un ensemble de relations prédéfinies : *domain, range, inverse*.

À l'instar des concepts individuels, nous nommons « concept générique » (resp. « relation conceptuelle ») l'ensemble des propriétés portant sur une classe d'objets (resp. une relation).

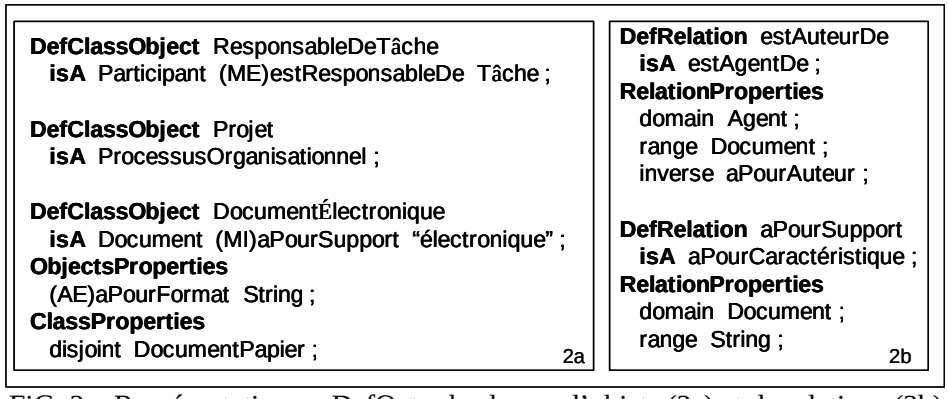


FiG. 2 – Représentation en DefOnto de classes d'objets (2a) et de relations (2b)

[2] D'après cette traduction, on observe que le caractère 'A' tient pour le quantificateur universel « All » tandis que le caractère 'E' tient pour le quantificateur existentiel « Exist ». Nous revenons en section 5 sur la puissance d'expression propositionnelle de DefOnto.

[3] Le caractère 'M' tient pour un « modificateur » et revient en français à introduire une proposition relative. Exemple : « un responsable de tâche est un participant QUI est responsable d'une tâche ».

DefOnto	Logique du 1er ordre	Logique de Description
C_1 **(AE)**R C_2	$\forall x\ C_1(x) \supset \exists y\,(C_2(y) \wedge R(x,y))$	$C_1 \sqsubseteq \exists R.C_2$
C_1 **(AI)**R C_2	$\forall x\ C_1(x) \supset R(x,C_2)$	$C_1 \sqsubseteq \exists R.\{C_2\}$
C_1 **isA** C_2 **(ME)**R C_3	$\forall x\ C_1(x) \equiv C_2(x) \wedge \exists y\,(C_3(y) \wedge R(x,y))$	$C_1 \equiv C_2 \sqcap \exists R.C_3$
C_1 **isA** C_2 **(MI)**R C_3	$\forall x\ C_1(x) \equiv C_2(x) \wedge R(x,C_3)$	$C_1 \equiv C_2 \sqcap \exists R.\{C_3\}$

TaB. 1 – Extrait de la sémantique logique de DefOnto

3 Représentation du modèle de l'information au moyen d'une théorie de propositions et d'une théorie de concepts

3.1 Contenu d'un modèle des informations

Une information est une entité à part entière, mais d'une nature ontologique différente des objets modélisant l'organisation. Dans cette section, nous commençons par proposer une ontologie de l'information. Ceci nous amène à décrire les liens qu'entretient l'information avec les autres objets modélisant l'organisation.

Nous posons comme point de départ que l'information tire son identité du rôle qu'elle joue dans un acte de communication, au cours duquel un agent émetteur communique une information à un agent récepteur. Ce dernier pouvant considérer l'information comme vraie ou fausse, nous l'assimilons à une *proposition*.

Une proposition rend compte d'une situation du monde, par exemple : « *le numéro de téléphone de C. Durand est le 03 22 82 54 08* » et fait ainsi référence à des objets du monde, par exemple : *le numéro de téléphone de C. Durand*. L'histoire de la logique nous enseigne toutefois que la référence peut être directe ou indirecte, suivant qu'elle porte sur l'*objet* ou bien sur le *concept* dénotant l'objet [McCarthy, 1979]. Dans le cas présent, nous considérons que l'information porte sur le concept individuel *numéro de téléphone de C. Durand* plutôt que sur l'objet[4]. Ainsi, nous considérons qu'une information « porte sur », ou « a pour sujet », un concept. Il peut bien sûr s'agir de n'importe quel concept organisationnel, et d'un concept individuel comme d'un concept générique.

Finalement, pour compléter la notion d'information, nous admettons les liens suivants : un document contient des informations ; une information complexe (ex : le contenu d'un livre) est composée d'informations plus élémentaires ; une information peut être connue d'un agent.

L'ontologie de l'information, que l'on vient de décrire, permet d'exprimer un modèle de l'information, ou plutôt des informations. Il s'agit d'informations qui, d'une part, sont capitalisées dans le WSO, sous une forme d'explication ou une autre, et qui, d'autre part, sont susceptibles d'intéresser un utilisateur du WSO. La détermination de ce modèle résulte donc d'une analyse des besoins des utilisateurs. Par exemple, à propos d'un projet, un utilisateur peut désirer connaître son financement, quelles sont les organisations partenaires du projet, quels sont les objectifs visés ou savoir s'il existe des projets concurrents. Nous verrons en section 4 qu'à chaque type d'information retenu correspond un document, élaboré dynamiquement par le WSO, présentant cette information.

[4] L'argument mis en avant est le même que celui développé par [McCarthy, 1979]. À supposer que J. Dupont et C. Durand soient collègues de bureau et partagent un même poste téléphonique, donc un même numéro. De la proposition : « *P. Martin compose le numéro de téléphone de J. Dupont* », nous pouvons inférer que : « *P. Martin compose le numéro de téléphone de C. Durand* ». Par contre, dans la mesure où l'information dépend de l'état de connaissance de l'agent récepteur, de la proposition : « *l'information porte sur le numéro de téléphone de J. Dupont* », nous ne pouvons pas inférer que : « *l'information porte sur le numéro de téléphone de C. Durand* ». Nous en concluons que la référence est directe pour la relation *compose* et indirecte pour la relation *porte sur*. Partant d'un même constat, [Horrocks et al., 2002b] ont souligné récemment les limites des Logiques de Description. En effet, faute de pouvoir considérer des concepts comme des objets à part entière, ces langages ne permettent pas de représenter le(s) sujet(s) de documents.

3.2 Représentation du modèle de l'information

La représentation du modèle des informations et de son ontologie associée fait appel aux capacités de représentation de métaconnaissances de DefOnto. Ces capacités reposent sur le fait de considérer les propositions et les concepts comme de nouveaux objets, auxquels il est possible d'attribuer des propriétés. On aboutit ainsi à considérer une *théorie de propositions* et une *théorie de concepts*.

DefOnto permet tout d'abord d'attribuer des propriétés à des propositions et des concepts, considérés en tant qu'individus.

Une définition de proposition (*cf.* figure 3a) – introduite par le constructeur *DefIndProposition* – consiste en un ensemble de propriétés vérifiées par la proposition. L'exemple de la définition de la proposition *Info-5* montre qu'une définition de proposition est structurellement analogue à une définition d'objet. Simplement, comme il est indiqué plus loin, les relations *estContenuDans* et *aPourSujet* sont définies en tant que relations propositionnelles et non en tant que relations d'objets.

La définition d'un concept suit pour sa part un mécanisme différent. Considérant qu'un ensemble de propriétés attachées à une entité quelconque (ex : objet individuel, classe d'objets, relation, proposition individuelle, *etc.*) constitue un concept de cette entité – le concept réifie l'ensemble des propriétés -, DefOnto permet de compléter chaque définition d'entité par une liste de propriétés vérifiées par le concept. Ces propriétés apparaissent après le mot-clef « *ConceptProperties* ». Cette possibilité est illustrée dans la définition de la proposition *Info-6*, dont le concept est déclaré être le sujet de *Info-7*[5]. Elle est également illustrée dans la définition de l'objet *C.Durand* (*cf.* figure 3b) dont le concept est le sujet de *Info-5*. On voit du reste dans cette définition que chaque proposition contenue dans une définition d'entité peut être nommée, ce qui permet de faire le lien avec sa définition et ainsi de lui attribuer des propriétés.

La figure 4 montre enfin que DefOnto permet de représenter des classes de propositions (ex : *InformationConfidentielle*), des relations propositionnelles ayant pour domaine une proposition (ex : *estContenuDans*), des classes de concepts (ex : *CentreIntérêt*) et des relations conceptuelles ayant pour domaine un concept (ex : *estSujetDe*). Les constructeurs *DefClassProposition* et *DefClassConcept* offrent la même puissance d'expression, pour représenter une théorie de propositions et de concepts, que le constructeur *DefClassObject*, pour représenter une théorie d'objets. L'exemple de la définition de la classe *InformationConfidentielle* montre à ce propos que le concept *InformationConfidentielle* est déclaré dans les propriétés du concept comme sujet du document *AccordDeCollaboration.*

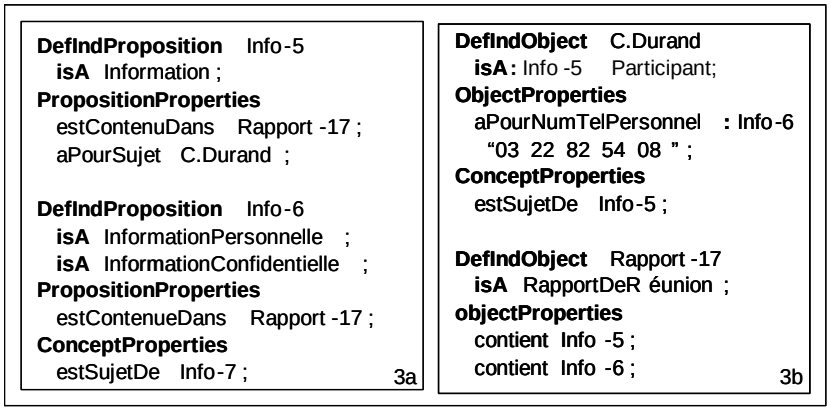


FiG. 3 – Représentation en DefOnto d'un modèle de l'information (3a) et de son lien avec le modèle de l'organisation (3b)

[5] Ceci suppose que l'information *Info-7* mentionne explicitement l'information *Info-6*, par exemple pour lui attribuer un auteur ou discuter de sa validité.

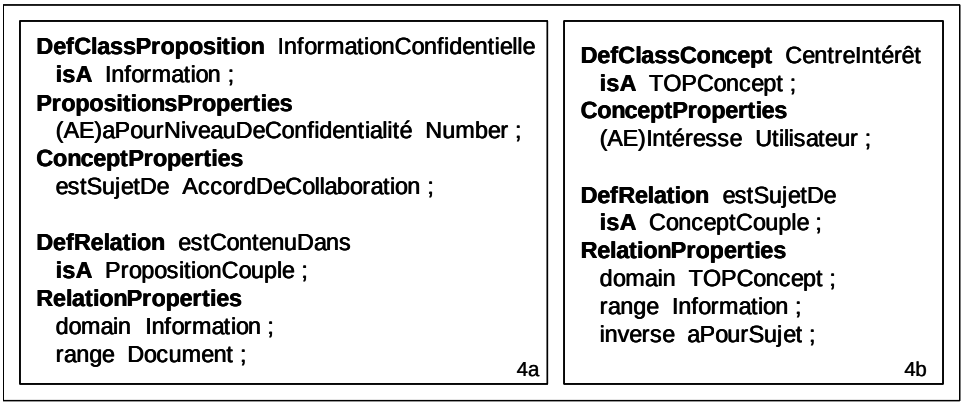


FiG. 4 – Représentation en DefOnto d'une ontologie de propositions (4a) et d'une ontologie de concepts (4b)

4 Construction d'un « index informationnel »

Dans cette section, nous présentons une exploitation des modèles de l'organisation et de l'information pour développer des fonctionnalités d'aide à la recherche d'information dans un WSO. Nous proposons en effet d'utiliser ces modèles pour construire un « index informationnel », c'est-à-dire un index du contenu informationnel d'un WSO. L'objectif de cette présentation n'est pas tant de justifier l'apport d'un tel index pour la recherche d'information (apport présenté dans [Fortier et Kassel, 2003]) que d'illustrer comment les services inférentiels associés à DefOnto permettent de raisonner sur les modèles en question et leur ontologie associée.

Un index informationnel (*cf.* figure 5) est un objet complexe, composé de deux index : un index de l'organisation (sur la gauche de la figure) et un index de l'information que le WSO est capable de présenter (au milieu). Chacun de ces index a la forme d'un arbre d'entrées, que l'utilisateur peut déplier en fonction du niveau de détail souhaité. Ils sont présentés dans deux fenêtres séparées dans le portail du WSO.

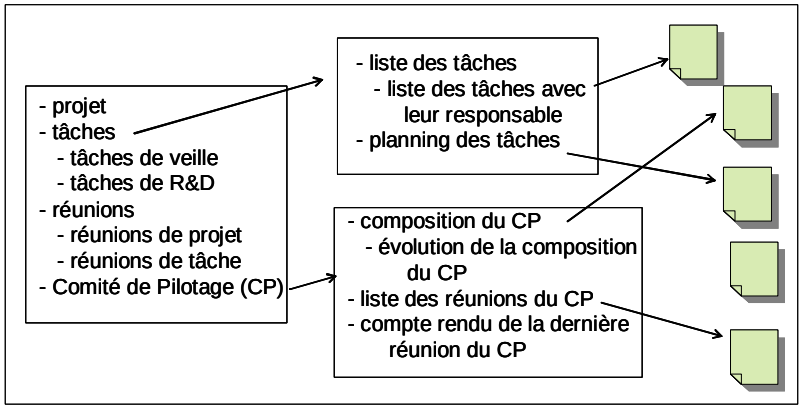


FiG. 5 – Exemple schématique d'index informationnel

4.1 Index de l'organisation

Les entrées de cet index sont matérialisées par l'étiquette d'un concept, qui peut être générique (ex : tâches de R&D, réunions de projet) ou individuel (ex : projet, comité de pilotage).

L'index est structuré par la relation « est un », ce qui signifie qu'il existe un lien de subsomption entre une sous-entrée et son père (ex : une tâche de veille est une tâche). Lorsque l'utilisateur demande que l'index soit déplié, un raisonnement est effectué sur l'ontologie de l'organisation pour connaître les concepts subsumés par un concept donné. Une requête est alors envoyée par l'Interface (ex : *[estInclusDans *x tâche]*) qui est évaluée par le Moteur d'Inférences de DefOnto. Il s'agit, sur cet exemple, de déterminer quelles sont les sous-classes directes de la classe des tâches.

Un clic de l'utilisateur sur une entrée de l'index provoque la présentation de la partie correspondante de l'index de l'information, c'est-à-dire de la partie concernée par le concept organisationnel qu'il/elle a sélectionné (ce saut est matérialisé par des flèches en figure 5). Un raisonnement sur le modèle de l'information est pour cela réalisé, qui conduit à l'évaluation de nouvelles requêtes (ex : *[(AI)aPourSujet *x ComitéDePilotage]*). Il s'agit ici de déterminer l'ensemble des classes d'information ayant pour sujet le concept *ComitéDePilotage*.

4.2 Index de l'information

Une entrée de cet index est matérialisée par l'étiquette du sujet d'une classe d'informations. Pour définir son besoin d'information, l'utilisateur n'a donc qu'à sélectionner une entrée de cet index. Cet index est également structuré par la relation « est un ». Un lien de subsomption entre deux (concepts d') information repose sur l'existence d'une proximité sémantique entre leur sujet (ex : une information sur l'évolution de la composition du comité de pilotage est considérée comme une information à propos de la composition de ce comité). Lorsque l'utilisateur demande à déplier cet index, un raisonnement est réalisé sur l'ontologie de l'information pour accéder à des classes d'information plus spécifiques. On retrouve le même type de raisonnement que sur l'ontologie de l'organisation.

Finalement, un clic sur une entrée de cet index provoque la génération par le WSO d'un document présentant l'information en question (le mécanisme de génération de ces documents est décrit dans [Fortier and Kassel, 2002]). L'élaboration du contenu de ces documents est réalisée en extrayant des parties de textes de documents présents dans la BD et en raisonnant sur le modèle de l'organisation. De nouvelles requêtes sont alors évaluées, par exemple : *[TâcheDeVeille *x]* (Quelles sont toutes les tâches de veille ?) ; *[estAuteurDe J.Dupont *x]* (Quels sont tous les documents ayant pour auteur J.Dupont ?).

En bilan, concernant les services inférentiels de DefOnto, on voit qu'on a affaire à de la classification d'instances. Le rôle de l'ontologie est de compléter les descriptions d'instances en inférant des informations implicites. Par exemple, à partir des faits : *C.Durand isA Participant, Rapport-18 aPourAuteur C.Durand, aPourAuteur inverse estAuteurDe, Auteur isA Personne (ME)estAuteurDe Document*, on va inférer que : *C.Durand isA Auteur*.

5 Positionnement de DefOnto

Le langage DefOnto est issu de travaux menés sur le langage Def-*, visant à définir un langage de haut niveau (plus déclaratif qu'un langage de règles) pour opérationnaliser des modèles d'expertise de résolveurs de problèmes réflexifs [Kassel et al., 2000]. Les notions de « réflexivité au niveau connaissance » et de « réflexivité au niveau symbolique », sur lesquelles repose Def-*, sont à l'origine des capacités de méta-représentation de DefOnto.

Ces capacités conduisent à décomposer l'univers de discours en quatre sous-ensembles disjoints : un ensemble d'objets, un ensemble de concepts, un ensemble de propositions et un ensemble d'entités de représentation[6]. Une Base de Connaissances en DefOnto est ainsi constituée d'une théorie d'objets (ce que nous avons appelé le « noyau » de DefOnto), complétée d'une théorie de concepts, de propositions et d'entités de représentation.

Pour chacune des théories, DefOnto offre une puissance d'expression analogue à celle du langage DAML-OIL [Horrocks et al., 2002a]. La comparaison entre ces puissances d'expression repose sur la définition pour DefOnto de deux sémantiques logiques : une sémantique de type Logique du 1er ordre avec égalité et une sémantique de type Logique de Description[7]. À partir de cette dernière, il est possible de vérifier que les axiomes de DefOnto reprennent l'ensemble des constructeurs de DAML-

[6] Ce dernier sous-ensemble permet d'assurer la réflexivité au niveau symbolique du langage. Il revient à considérer les entités de représentation elles-mêmes en tant que nouveaux individus, auxquels il est possible d'attribuer des propriétés. De telles propriétés peuvent être ainsi ajoutées à la définition de n'importe quel(le) individu, classe d'individus ou relation. Syntaxiquement, elles figurent après le mot-clef « EntityProperties ».

[7] Ces sémantiques sont construites selon le principe du tableau 1 en faisant correspondre à chaque axiome de DefOnto une expression équivalente dans ces deux logiques. Ces sémantiques peuvent être consultées sur le site de DefOnto : http://www.laria.u-picardie.fr/EQUIPES/ic/defOnto/.

OIL, excepté la possibilité de déclarer la transitivité d'une relation[8]. Cette sémantique sert du reste de base actuellement à la réalisation d'un traducteur de DefOnto vers la Logique de Description implantée dans le système FACT [Horrocks et al., 1999] de façon à pouvoir bénéficier des services inférentiels offerts par ce système, à savoir la vérification de la cohérence logique d'une ontologie et l'explicitation de liens de subsomption implicites. Ces services d'aide à la construction d'ontologies vont venir compléter les services de classification d'instances dont dispose déjà DefOnto.

DefOnto se rapproche donc du langage DAML-OIL, en apportant la possibilité de représenter des métaconnaissances, une capacité reconnue comme une caractéristique souhaitable pour les langages d'ontologies pour le Web sémantique [Heflin, 2003]. Par contre, contrairement au langage DAML-OIL, DefOnto n'a pas été défini comme une extension du langage RDF(S), la compatibilité avec ce standard étant considérée comme une caractéristique essentielle pour ces langages [Patel-Schneider and Fensel, 2002].

Un tel positionnement, tant sur le plan syntaxique que sémantique, figure parmi nos objectifs de recherche. Comme l'ont montré récemment [Pan and Horrocks, 2002], ce positionnement passe justement par un examen des capacités de méta-représentation du langage RDF(S).

6 Conclusion

Dans cet article, nous nous sommes situés dans le contexte du développement de mémoires organisationnelles hybrides, réalisant un couplage fort entre une BC et une BD et nous avons rappelé nos propositions quant au choix des connaissances à modéliser. Notamment, pour faciliter l'exploitation des documents, nous préconisons de modéliser en priorité l'organisation pour laquelle le WSO est développé et les informations dont dispose le WSO à propos de l'organisation.

Partant de ces propositions, nous avons surtout montré que la modélisation de l'information nécessite de pouvoir rendre compte de métaconnaissances, sous la forme de propriétés portant sur des propositions et des concepts. Nous retrouvons ainsi un besoin dont l'importance a déjà été soulignée pour la définition des langages d'ontologie pour le Web sémantique.

Nous avons également introduit le langage DefOnto qui dispose de telles capacités de représentation de métaconnaissances et qui autorise des raisonnements sur ces métaconnaissances. DefOnto est en cours de développement. Un de nos objectifs de recherche est de positionner DefOnto par rapport aux langages d'ontologie pour le Web sémantique, ce qui revient à étudier sa compatibilité sur les plans syntaxique et sémantique avec le standard RDF(S).

Références

[Dean and Schreiber, 2003] Dean, M. and Schreiber, G. (eds.) (2003). OWL Web Ontology Language Reference. W3C Working Draft 31 March 2003, http://www.w3.org/TR/owl-ref/.

[Dieng-Kuntz et al., 2001] Dieng-Kuntz, R., Corby, O., Gandon, F., Giboin, A., Golebiowska, J., Matta, N. and Ribière, M. (2001). *Méthodes et outils pour la gestion des connaissances ; une approche pluridisciplinaire du Knowledge Management*. 2e édition, Dunod, Paris.

[Fortier et al., 2002] Fortier, J.-Y., Cormier, C., Kassel, G., Barry, C., Irastorza, C. and Bruaux, S. (2002). To supply organization views, suited to users: an approach to the design of organizational memories. In *Proceedings of the ECAI'2002 Workshop on Knowledge Management and Organizational Memories*, Lyon (France), pages 61-69.

[Fortier and Kassel, 2002] Fortier, J.-Y. and Kassel, G. (2002). Génération de documents virtuels personnalisés à partir de modèles de connaissances. In *Actes de la Conférence Documents Virtuels Personnalisables : DVP 2002*, Brest (France), pages 115- 126.

[Fortier and Kassel, 2003] Fortier, J.-Y. and Kassel, G. (2003). Building adaptive Information Retrieval Systems for Organizational Memories: a case Study. In *Proceedings of the 12th*

[8] Ce choix est délibéré. Il a en effet été prouvé que la combinaison de cette propriété de relation avec le fait de définir la relation en tant que relation inverse d'une autre relation correspondait à la complexité dans le pire des cas pour les algorithmes des services inférentiels évoqués dans la phrase suivante.

International Conference on Intelligent and Adaptive Systems and Software Engineering: IASSE-2003, San Francisco (CA, USA).

[Gandon et al., 2000] Gandon, F., Dieng, R., Corby, O. and Giboin, A. (2000). A Multi-Agent System to Support Exploiting an XML-based Corporate Memory. In *Proceedings of the 3rd International Conference on Practical Aspect of Knowledge Management (PAKM'2000)*, pages 10-1, 10-12.

[Gandon, 2002] Gandon, F. (2002). Ontology Engineering: a Survey and a Return on Experience. Rapport de recherche N°4396 de l'INRIA.

[Heflin, 2003] Heflin, J. (ed) (2003). Web Ontology Language (OWL) Use Cases and Requirements. W3C Working Draft 31 March 2003, http://www.w3.org/TR/2003/WD-webont-req-2003031/

[Horrocks et al., 1999] Horrocks, I., Sattler, U. and Tobies, S. (1999). Practical reasoning for expressive description logics. In Ganzinger, H. et al. (eds), *Proceedings of LPAR'99*, Springer-Verlag, pages 161-180.

[Horrocks et al., 2002a] Horrocks, I., Patel-Schneider, P.F. and van Harmelen, F. (2002). Reviewing the Design of DAML+OIL: An Ontology Language for the Semantic Web. In *Proceedings of the 18th American National Conference on Artificial Intelligence: AAAI-2002*.

[Horrocks et al., 2002b] Horrocks, I., McGuinness, D.L. and Welty, C. (2002). Digital Libraries and Web-Based Information Systems. In Baader, F. *et al.* (eds), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press.

[Kassel, 2002] Kassel, G. (octobre 2002). Une approche du développement de webs sémantiques d'entreprise centrée sur un modèle de l'entreprise. In Laublet, P. et al. (eds), *Actes des journées de l'Action Spécifique : Web Sémantique*. Les actes sont accessibles à l'adresse : http://www.lalic.paris4.sorbonne.fr/stic/octobre/programme0209.html.

[Kassel et al., 2000] Kassel, G., Barry, C. and Abel, M.-H. (2000). Programmer au niveau connaissance en Def-*. In Charlet, J. et al. (eds.), *Ingénierie des Connaissances, évolutions récentes et nouveaux défis*, Eyrolles, pages 145-160.

[McCarthy, 1979] McCarthy, J. (1979). First Order Theories of Individual Concepts and Propositions. In Hayes *et al.* (eds), *Machine Intelligence 9*, England: Ellis Horwood, Ltd., pages 129-147.

[Pan and Horrocks, 2002] Pan, J.Z., and Horrocks, I. (2002). Metamodeling Architecture of Web Ontology Languages. In Cruz, I. et al. (eds), *The Emerging Semantic Web*, IOS Press, pages 21-45.

[Patel-Schneider and Fensel, 2002] Patel-Schneider, P.F. and Fensel, D. (2002). Layering the Semantic Web: Problems and Directions. In Horrocks, I. and Hendler, J. (eds), *Proceedings of the First International Semantic Web Conference: ISWC 2002*, Sardinia (Italy), LNCS 2342, pages 16-29.

[Staab and Maedche, 2001] Staab, S. and Maedche, A. (2001). Knowledge Portals, Ontologies at Work. *AI Magazine*, Summer 2001, pages 63-75.

Une application Web Sémantique basée sur les Topic Maps : projet MEMORAe

A. Benayache, M-H. Abel, D. Lenne
UMR CNRS 6599 Heudiasyc Université de Technologie de Compiègne
BP 20529, 60206 Compiègne Cedex, France
{Ahcene.Benayache, mhabel, Dominique.Lenne}@hds.utc.fr

Résumé

Le but du projet MEMORAe est de réaliser une mémoire organisationnelle de formation et de l'exploiter dans le cadre d'une formation e-learning. Cette mémoire capitalise les connaissances liées à la formation et facilite l'accès aux informations et aux documents pertinents à la fois pour les apprenants et pour les enseignants. Dans cet article, nous présentons l'organisation de la mémoire, ainsi que les ontologies et le formalisme des Topic Maps sur lesquelles elle s'appuie.

Mots-clés : *Web Sémantique, Topic Maps, Ontologie, e-learning.*

Abstract

The MEMORAE Projects aims at building an organisational memory for an e-learning application. This memory capitalises training knowledge and facilitate access to pertinent documents and information, by both students and and teachers. In this paper we describe first the memory organisation and then we present the ontologies and the Topic Maps formalism on which it is based.

Keywords : *Semantic Web, Topic Maps, Ontology, e-learning.*

1 Introduction

Avec l'émergence des Technologies de l'Information et de la Communication (TIC), une nouvelle forme de formation, ou plus exactement un nouveau mode d'apprentissage, est apparu. Souvent appelé "e-learning", ce mode est basé sur l'accès à des formations en ligne, interactives et parfois personnalisées, diffusées par l'intermédiaire d'un réseau (Internet ou intranet) ou d'un autre média électronique. Cet accès permet de développer les compétences des apprenants, tout en rendant le processus d'apprentissage indépendant du temps et du lieu.

Dans le cadre du projet MEMORAe[1] (MEMoire ORganisationnelle Appliquée au e-learning) nous nous intéressons à la problématique du contenu mis à disposition des apprenants. Nous avons choisi de découper et de représenter le contenu pédagogique d'une formation au moyen de grains de connaissance. La gestion de ce contenu pose des problèmes de structuration, d'organisation, d'acquisition, de maintenance, de diffusion des informations et des connaissances. Nous proposons de traiter cette problématique au moyen d'une mémoire organisationnelle qui sera mise à la disposition des utilisateurs (apprenants, enseignants, etc.). Le partage de cette mémoire repose, entres autres, sur des mécanismes de consultation de son contenu via l'utilisation d'Internet. Les grains de connaissance représentent les notions à appréhender et servent d'index pour accéder aux ressources (livres, sites Web, etc.) qui traitent de ces notions. De ce fait on se retrouve dans une problématique Web Sémantique.

Le projet MEMORAe est donc situé au carrefour de trois domaines de recherche : Ingénierie des Connaissances, Ingénierie Éducative et Web Sémantique. Il s'agit tout d'abord d'effectuer un travail pédagogique afin de découper le contenu de la formation en grains de connaissance et d'articuler ces derniers les uns par rapport aux autres (domaine de l'ingénierie éducative) ; ensuite, le choix d'organiser et de gérer un tel contenu au moyen d'une mémoire organisationnelle basée sur des ontologies concerne le domaine de l'ingénierie des connaissances. Enfin, le choix de la norme ISO Topic Maps [IEC, 99] pour structurer les informations afin de présenter l'accès aux ressources de la mémoire organisationnelle mise à disposition des apprenants concerne le domaine du Web Sémantique. Cette norme permet de distinguer un niveau ressources et un niveau de représentation sémantique décrivant, indexant, ces ressources.

Dans cet article, nous présentons tout d'abord notre conception du e-learning, puis nous montrons comment nous lions ontologies et norme ISO Topic Maps pour concevoir une mémoire organisationnelle de formation.

2 E-learning

Le e-learning est un mode d'apprentissage qui va au-delà d'une simple utilisation des TIC au service d'anciens modes de formation. Pour nous [Abel & al, 02], une application e-learning est une formation à distance :

- Exploitant une logique réseau (c'est l'apprentissage lui-même qui se déroule à distance) ;
- Mettant à disposition un contenu sous forme granulaire et non un cours linéaire, magistral en ligne ;
- Dans laquelle les apprenants sont actifs et travaillent de façon collaborative ;
- Dans laquelle le rôle des formateurs est plus celui d'accompagnateur, de tuteur (on passe d'une logique de transmission du savoir à une logique d'accompagnement).

[1] Projet soutenu dans le cadre du pôle NTE de la région Picardie.

Ainsi, il est plus juste de traduire le terme e-learning par "e-apprentissage" plutôt que "e-formation" (e-training). Dans ce qui suit, nous précisons le lien que nous faisons entre e-learning et organisation avant de définir le concept de mémoire organisationnelle.

2.1 e-learning et organisation

Une formation s'organise autour d'acteurs (intervenants, apprenants, secrétaires, etc.), de connaissances (celles des acteurs) ainsi que d'informations de différents types (définitions, exercices avec ou sans corrigé, études de cas, etc.) sous différentes formes (rapports, livres, sites web etc.). En ce sens, une formation est une organisation.

Une solution souvent adoptée pour gérer les connaissances d'une organisation est la réalisation d'une mémoire organisationnelle. Une telle mémoire peut être considérée comme la représentation explicite et persistante des connaissances et des informations dans une organisation, afin de faciliter leur accès et leur réutilisation par les membres adéquats de l'organisation pour leur tâche [Dieng & al, 00].

Une mémoire organisationnelle permet de capitaliser les grains de connaissance du contenu de formation au même titre que les informations concernant les acteurs (leurs spécificités, leurs parcours, etc.) et la gestion administrative (inscription, notes, etc.) de la formation pour partager des informations au sein d'une organisation, particulièrement lorsque les acteurs sont géographiquement distants, il est nécessaire d'utiliser un vocabulaire commun et que les termes signifient pour tous la même chose. C'est pourquoi la construction et la gestion d'une mémoire organisationnelle s'effectuent souvent à l'aide d'ontologies.

Dans le projet MEMORAe, pour matérialiser notre mémoire, nous avons fait le choix d'utiliser deux formalismes de représentation des connaissances : les ontologies et la norme ISO Topic Maps. Nous justifions ce choix dans la suite.

2.2 e-learning et Web Sémantique

Une application e-learning est mise en ligne (à disposition des apprenants) via l'utilisation du Web. De ce fait, elle partage le même problème de pertinence avec le Web lorsque les apprenants veulent accéder au savoir mis à leur disposition. Parmi les efforts visant à résoudre ce problème (de pertinence) nous trouvons la notion de "Web Sémantique". Cette notion, lancée en 1998 par Tim Berners-Lee, a pour objectif d'améliorer les rapports des utilisateurs avec le Web. Le Web Sémantique peut être défini comme un substrat supportant des fonctions avancées pour la collaboration (homme-homme, homme-machine, machine-machine), qui permet de partager des ressources et de raisonner sur le contenu de ces dernières [Berners-Lee & al, 01]. L'idée est de rendre explicite la sémantique des documents au travers de meta-données ou d'annotations [Gandon & al, 02].

On peut distinguer deux types d'approche pour le Web Sémantique [Caussanel & al, 02], l'une qualifiée de "Web computationnellement sémantique" et l'autre de "Web cognitivement sémantique". La première concerne l'automatisation de la recherche au moyen d'agents logiciels. La deuxième s'intéresse à la structuration des contenus et vise une semi-automatisation de certaines tâches [Caussanel & al, 02].

L'approche « Web cognitivement sémantique » privilégie la problématique de l'indexation à celle de l'inférence. Elle permet néanmoins des inférences "simples" à partir de représentations dont la "sémantique opérationnelle" est plus faible que celle de représentations basées sur des langages formels supportant des traitements puissants [Caussanel & al, 02]. C'est pourquoi nous avons choisi cette approche pour constituer notre mémoire organisationnelle de formation. Elle permet aux utilisateurs d'accéder aux diverses informations concernant la formation choisie de manière signifiante et selon des points de vues multiples.

3 Mémoire organisationnelle de formation appliquée à l'e-learning : projet MEMORAe

Dans le cadre du projet MEMORAe, nous proposons de réaliser une mémoire organisationnelle de formation et de l'exploiter comme contenu mis à disposition des apprenants d'une formation e-learning. Cette mémoire est constituée des ressources nécessaires à la formation ainsi que d'index permettant d'accéder, de décrire ces ressources. Ces index représentent les granules ou notions à appréhender traités dans les ressources. Notre mémoire permet finalement de capitaliser toute connaissance, document, ou information traitant de ces notions. Parmi les documents représentés, certains (documents électroniques) sont directement stockés dans la mémoire, alors que les autres ne figurent que sous la forme de références. Cette capitalisation touche aussi toute information relative à l'environnement de la formation qui peut être utile à un utilisateur, telle que les informations concernant : les acteurs (membres), les dossiers d'inscription, etc. Afin que les utilisateurs de la mémoire puissent accéder aux ressources capitalisées, et puissent communiquer et collaborer entre eux, il est nécessaire :

- qu'ils emploient un vocabulaire commun ;
- que l'accès aux ressources soit le plus signifiant possible.

Dans ce contexte, nous avons choisi une approche « web cognitivement sémantique » basée sur des ontologies et sur la norme Topic Maps.

Afin de développer un premier prototype de mémoire, nous avons choisi deux applications. La première traite de l'unité de valeur NF01, un enseignement d'initiation à l'algorithmique et à la programmation Pascal suivi par les étudiants de première année à l'Université de Technologie de Compiègne (UTC). La deuxième traite de l'unité de valeur B31, un enseignement de statistiques suivi par les étudiants de MIAGE et e-MIAGE de l'Université Jules Verne d'Amiens. Les exemples que nous présentons par la suite sont tirés de ces applications.

Nous précisons dans un premier temps ce que nous entendons par "grain de connaissance". Dans un second temps nous présentons notre mémoire organisationnelle qui est basée à la fois sur l'utilisation d'ontologies et sur l'utilisation de la norme ISO Topic Maps.

3.1 Grain de connaissance

Le but d'une application e-learning est de permettre aux apprenants d'acquérir de nouvelles connaissances en utilisant les TIC. Cela implique un recentrage pédagogique complet sur l'apprenant, notamment pour lui faire comprendre les ressources et les informations mises à sa disposition et lui apprendre à les chercher et à les utiliser. Articuler un cours autour de grains de connaissance offre davantage de possibilités d'individualisation de la formation. Selon Boullier (2001), il s'agit d'un découpage du « monde des savoirs », acte préalable à tout enseignement. Il propose de matérialiser ce découpage par les supports. Les grains de connaissance sont alors le résultat d'une délimitation du texte, un balisage sémantique.

Contrairement à ces travaux, nous n'utilisons pas le terme "grain de connaissance" pour désigner une portion de texte d'un cours linéaire, mais pour désigner une notion à appréhender. Ainsi, il n'est pas nécessaire de produire des supports matérialisant un tel découpage. Les auteurs restent libres quant à la réalisation de leurs supports, ils n'ont pas à respecter une quelconque charte de rédaction, qu'elle soit graphique ou relative au contenu.

Les grains de connaissance (notion à appréhender) servent d'index pour accéder aux documents qui traitent de ces derniers. Le grain de connaissance « boucle » indexe donc les ressources de la mémoire traitant de la notion de « boucle ». Choisir de représenter les grains de connaissance comme notions (concepts) qui permettent d'indexer des documents les concernant laisse une souplesse de rédaction de ces documents ou bien la possibilité de réutiliser des documents déjà rédigés. Le fait qu'un grain de connaissance fasse référence à différents documents permettra différentes manières de l'appréhender.

3.2 Représentation de la mémoire organisationnelle

Pour représenter la mémoire de notre projet et assurer un bon moyen d'accès aux informations, nous nous sommes intéressés d'une part aux ontologies (définition d'un vocabulaire commun) et d'autre part aux Topic Maps (navigation, accès aux ressources pédagogiques).

3.2.1 Ontologies

En Ingénierie des Connaissances, une ontologie est définie comme une spécification explicite d'une conceptualisation [Grüber, 93]. Une ontologie fournit un cadre unificateur pour réduire et éliminer les ambiguïtés et les confusions conceptuelles et terminologiques et assurer une compréhension partagée par la communauté visée [Uschold & Gruninger, 96].

Articuler notre mémoire sur des ontologies, est un choix stratégique et essentiel, stratégique pour définir un vocabulaire commun aux utilisateurs de la mémoire, et essentiel pour associer du sens aux concepts à appréhender. Pour cela, nous définissons deux ontologies :

- Une ontologie du domaine formation : elle définit des concepts du domaine formation, qui restent générique pour ce domaine. Y sont définis des concepts tels que : rapports, livre, support de cours pour les documents. Cette ontologie sera exploitée par chaque formation particulière.
- Une ontologie d'application : elle définit les concepts d'une application donnée. Y sont définis des concepts plus spécifiques que ceux définis dans une ontologie de domaine. Par exemple, pour une ontologie concernant une formation en algorithmique, pourront être définis des concepts tels que : 'algorithme', 'boucle', 'boucle à bornes définies'. Une ontologie d'application doit être définie pour chaque formation envisagée. C'est au moyen de cette ontologie que sont définis les grains de connaissance.

Pour construire ces ontologies, nous avons choisi de suivre la méthode développée par l'équipe IC de LaRIA [Kassel & al, 00]. Cette méthode est constituée de deux étapes : l'ontologisation et l'opérationnalisation. La première consiste à construire une spécification structurée en langue naturelle d'une ontologie conceptuelle. Cette étape est réalisée manuellement à partir de différentes sortes de données telles que des glossaires de termes, des livres, des cours, etc. La deuxième étape consiste à coder l'ontologie conceptuelle informelle obtenue à la première étape à l'aide d'un langage de représentation d'ontologies DefOnto [Barry & al, 01]. Ce choix de méthodologie se justifie par la diversité des ressources entrant en jeu pour constituer une formation : livres, support de cours électroniques ou non, interviews de formateurs, etc.

3.2.2 Choix de la norme ISO Topic Maps (TM)

Les Topic Maps (TM) ont été créées au début des années 1990 par le groupe de documentalistes Davenport pour répondre à une problématique d'échange de documents

électroniques et plus particulièrement celui de leurs index [XTM, 01]. Elles sont devenues un standard au début des années 2000.

La norme ISO Topic Maps, permet de représenter des connaissances contenues dans une base documentaire sous forme de sujets (Topics) et d'associer ces connaissances à des ressources d'information. Les topic maps contiennent des documents, les sujets traités par ces documents et les relations entre ces sujets [IEC, 99].

Les raisons qui nous amenés à choisir ce formalisme pour représenter notre mémoire de formation sont multiples :

- Les topic maps sont un nouveau standard pour décrire des structures de connaissances associées à des ressources d'information. A ce titre, elles constituent un bon candidat pour la problématique de gestion des connaissances. Les topic maps permettent aussi bien l'expression de méta-données que l'exploitation des relations entre éléments pour différentes tâches.
- Les méta-données sont associées aux ressources mais ne sont pas incluses dans la ressource décrite. Les ressources ne sont donc pas modifiées et peuvent être ré-exploitées par d'autres communautés ou différentes tâches. Les TM distinguent ainsi un niveau ressource et un niveau description sémantique.
- Les TM permettent d'offrir un accès navigationnel aux informations représentées, de décrire les routes, les liens entre ces différentes informations.
- Afin d'implémenter les TM, un langage opérable pour le Web sémantique a été créé : XTM (ou *XML Topic Maps, eXtensible Markup Langage Topic Maps*). De plus, il existe des langages de requête de TM, par exemple TMQL (*Topic Map Query Language*).

Les TM sont essentiellement basées sur les notions de Topics, d'Associations et d'Occurrences. Nous précisons dans la suite comment nous les exploitons pour notre problématique.

3.2.2.1 Les Topics

Un *Topic* est la représentation informatique d'un *Sujet* appliqué à un ensemble de localisations (*Contexte*) [Le Grand, 01]. Il peut désigner un apprenant, une entité, un concept, etc. Strictement parlant, le terme Topic réfère à un objet ou nœud d'une TM qui représente un sujet. En d'autres termes, à un topic correspond un sujet unique et inversement. Cependant, il peut arriver qu'un même sujet soit représenté par plusieurs topics, dans le cas de la création de plusieurs TM par exemple (e.g. les topics « algorithme » et « algorithm »). Dans une telle situation, il est nécessaire d'établir une seule et même identité pour les différents topics. Ceci est réalisé par le concept de subject indicator. Tout topic partageant un ou plusieurs subjects indicators sont considérés comme sémantiquement équivalents. Précisons qu'un sujet peut être adressable (*Adressable Subject*) s'il est lui-même une ressource, dans ce cas, son URL ou URI permet de l'identifier, il n'est donc pas besoin de recourir à un subject indicator. Si le sujet n'est pas une ressource (cas de nos grains de connaissance), il est nécessaire d'utiliser un subject indicator. Un subject indicator est ni plus ni moins qu'une ressource créée afin de définir de façon non ambiguë un sujet (elle possède une URI et est donc adressable).

En ce qui concerne notre mémoire, nous avons donc des sujets adressables tels que des sites web, des articles, des forums etc. et des sujets non adressables tels que les grains de connaissance tableau, boucle, variable etc. ou bien une ressource non électronique (livre classique). Notons qu'en permettant qu'un sujet soit adressable ou non, la norme TM nous permet de décrire de la même façon un grain de connaissance ou bien un document comme par exemple une annotation, un site web etc.

Dans une Topic Map, les topics peuvent être classés selon leur type. Ainsi, *Jean* pourrait être un topic de type *Formateur*, *Rose* un topic de type *Apprenant*, *NF01* un topic de type *UV*, etc. La relation entre le topic et son type de topic est une relation classe/instance. Le

type d'un topic doit lui-même être défini en tant que topic, par exemple *Formateur*, *Apprenant* et *UV* sont aussi définis comme des topics dans la Topic Map de l'application "formation d'algorithmique". La définition des types des topics dépend de l'utilisation de leur TM, des besoins de l'application et de la nature d'information [Pepper, 00]. Pour notre application e-learning, nous retrouvons les topics formateurs, apprenants, les grains de connaissance (structure de données, structure itérative, structure répétitive, etc.), etc.

Un topic peut avoir un ou plusieurs ***noms***. Ces noms doivent permettre d'identifier un topic sans ambiguïté au sein de la TM, c'est-à-dire que deux topics distincts ne peuvent pas partager un même nom à l'intérieur d'une même TM. Sinon, lors du traitement de la Topic Map en utilisant le standard XTM, ils seront fusionnés. Le choix de ces noms doit être conforme aux usages dans lequel la Topic Map est utilisée. Notons qu'un topic peut ne pas avoir de nom mais une simple référence telle que « voir page 22 », référence considérée comme un lien vers un topic sans nom explicit. Possibilité que nous utilisons dans notre mémoire afin de préciser qu'un grain de connaissance est traité à partir d'une certaine page d'un livre.

3.2.2.2 Les Occurrences

Un topic peut être lié à une ou plusieurs ressources d'information, ressources pertinentes pour décrire ce topic (sujet). Une ***occurrence*** (ressource d'information) peut être un article, une image, une vidéo, un commentaire, etc. Ces occurrences se trouvent généralement en dehors de la topic map ce qui représente une séparation en deux niveaux (couches) des TM : d'une part les topics et d'autre part leurs occurrences. L'accès aux occurrences d'un topic se fait en utilisant des mécanismes d'indexation (par exemple : Xlink '*XML Linking Language*' [W3C, 01]) ou de marquage des ressources sans modifier le document indexé. C'est ce dont nous avions besoin pour la constitution de notre mémoire. Ces occurrences peuvent être classées selon leurs types : document texte, image, son, statistique, etc.

3.2.2.3 Les Associations

Jusqu'à présent, les notions que nous avons abordées (topic, sujet, type de topic, occurrence, rôle d'occurrence) nous ont permis d'organiser nos ressources d'information selon un topic ou sujet et de créer une indexation « directe ». La notion d'Association va nous permettre de créer des liens entre les différents topics. Créer des liens entre topics nous permet de structurer les topics les uns par rapport aux autres et donc d'offrir aux utilisateurs de notre mémoire un bon moyen de navigation et d'accès à l'information.

Une association permet de lier deux ou plusieurs topics. Dans le standard des topic maps, les associations ne sont pas directionnelles, c'est-à-dire le nom attribué à l'association n'implique aucune direction particulière [Le Grand, 01]. C'est la notion de ***rôle*** joué par chacun des topics impliqués dans l'association qui donnera le sens de la lecture. Notons que puisque qu'un topic peut très bien représenté une ressource (sujet adressable) ou bien un grain de connaissance (sujet non adressable), il nous est facile dans notre mémoire de représenté l'exemple suivant de notre application : « Programmer en turbo Pascal 7.0 a été écrit par Delannoy », cet exemple est représenté par l'association « a été écrit par», le topic « Programmer en turbo Pascal 7.0 » dont le rôle est « livre » et le topic « Delannoy » dont le rôle est celui d' « auteur ». On peut par la même précisé que ce premier topic a fait l'objet d'une annotation dont l'auteur est Julien.

3.2.2.4 Notion de contexte

L'accès à l'information peut devenir plus clair si on utilise le concept de ***Scope*** (contexte), qui permet de préciser la validation des caractéristiques d'un topic (noms, occurrences et associations) à un contexte particulier. Ce concept permet ainsi d'éviter les ambiguïtés et de réduire les risques d'erreurs. Par exemple, supposons que *Algorithme* puisse avoir plusieurs significations selon l'utilisateur de la TM, alors l'association entre *Algorithme*

et *Boucle à bornes définies* n'est valable que dans le contexte de *structure de donnée*. Ce la est appréciable mais pas suffisant pour fournir du sens et assurer une exploitation intégrale de la structure des topics. Afin de définir sémantiquement les topics, nous utilisons dans le cadre de notre projet les Ontologies.

Une Topic Map est constituée d'un ensemble de topics et de leurs associations. Une convention d'écriture des concepts d'un topic map et de les rendre opérable par le Web est celle du syntaxe XTM. L'énoncé de notre exemple « Programmer en turbo Pascal 7.0 a été écrit par Delannoy », peut être représenté par l'extrait suivant :

```
<topicMap xmlns="http://www.topicmaps.org/xtm/1.0/"
          xmlns:xlink="http://www.w3.org/1999/xlink">
    <topic id= "t-Programmer-turbo-pascal"> < ! --  Identifiant du topic -- !>
        <instanceOf> < ! --  Type de topic (Super classe) -- !>
            <topicRef   xlink:href= "#t-Livre-turbo-pascal" />
        </instanceOf>
        < ! – Le Nom de base du topic pour le Français -- !>
        <baseName> < ! --  Nom du topic -- !>
            <scope><topicRef xlink:href="#FR"/></scope>
            <baseNameString> Programmer en turbo pascal </baseNameString>
        </baseName>
    </topic>
    <topic id= "t-Delannoy"> < ! --  Identifiant du topic -- !>
        <instanceOf> < ! --  Type de topic (Super classe) -- !>
            <topicRef   xlink:href= "#t-Auteur-Livre" />
        </instanceOf>
        <baseName> < ! --  Nom du topic -- !>
            <baseNameString> Claude Delannoy </baseNameString>
            <variant> < ! --  Variante de Nom du topic -- !>
                <parameters> <topicRef xlink:href="#t-Nom-CD" > </parameters>
                <variantName>
                    <resourceDate id="Nom-CD01"> C. Delannoy </resourceDate>
                </variantName>
            </variant>
        </baseName>
    </topic>
    <association id= "A-LA001" > < ! --  Identifiant de l'association -- !>
        <instanceOf>
            <topicRef   xlink:href= " #t-Ecrit-par" /> < ! --  Type de l'association  -- !>
        </instanceOf>
        <membre> < ! --  Membre n°1 avec son rôle  -- !>
            <roleSpec> <topicRef   xlink:href= "#t-Livre"/> </ roleSpec>
            <topicRef   xlink:href="#t-Programmer-turbo-pascal"/>
        </membre>
        <membre> < ! --  Membre n°2 avec son rôle  -- !>
            < roleSpec> <topicRef   xlink:href= "#t-Auteur"/> </ roleSpec>
            <topicRef   xlink:href="#t-Delannoy"/>
        </membre>
    </association>
</topicMap>
```

Figure N°1 : extrait d'une Topic Map

3.2.3 Lien Ontologie-Topic Maps

Notre mémoire est constituée principalement de grains de connaissance et de ressources informationnelles traitant de ces grains. La mise en place d'une telle mémoire nécessite d'assurer un bon environnement de navigation qui permet de mettre l'apprenant en situation d'agir, c'est-à-dire lui offrir un accès facile au contenu. A cette fin, notre mémoire est

articulée autour d'ontologies et de topic maps. Cela permet à la fois de représenter les grains de connaissance comme concepts ontologiques et de réifier les informations décrivant ces concepts au moyen des topic maps. En fait, ces deux formalismes sont complémentaires, le premier (Ontologie) permet de fournir du sens en donnant une signification claire des grains. Le deuxième (Topic Maps) permet de compléter cette définition au moyen de contexte de validation, d'associations et d'occurrences (ressources de description). Par exemple, représenter la notion "*évènement*" qui peut désigner à la fois "*sous ensemble de l'univers en statistique*", "*événement populaire*", etc. pose un problème d'identification. Ce problème peut être résolu en utilisant une ontologie. Cette dernière permet de spécifier qu'il s'agit dans notre application de "*sous ensemble de l'univers*". Présenter une notion à des apprenants peut se faire par le biais de ressources bibliographiques (utilisation des occurrences des topic maps). Enfin, le formalisme des topic maps permet de lier une notion avec d'autres concepts et de limiter son domaine de validation à un contexte bien défini (dans ce cas : statistiques).

La figure 2 présente l'interface d'interrogation de la mémoire en cours de développement. Cette interface est constituée principalement de trois parties :

- Le bandeau supérieur permet l'accès à d'autres interfaces : structurer, rechercher, etc.
- Le bandeau de gauche permet de choisir une formation donnée (ex : NF01, B31, etc.) et de parcourir la hiérarchie des notions correspondant à la formation choisie.
- Le bandeau de droite permet d'afficher le chemin parcouru, la liste des sous-notions liées à une notion choisie, la définition de la notion choisie, ainsi que la liste des différentes ressources indexées par cette notion.

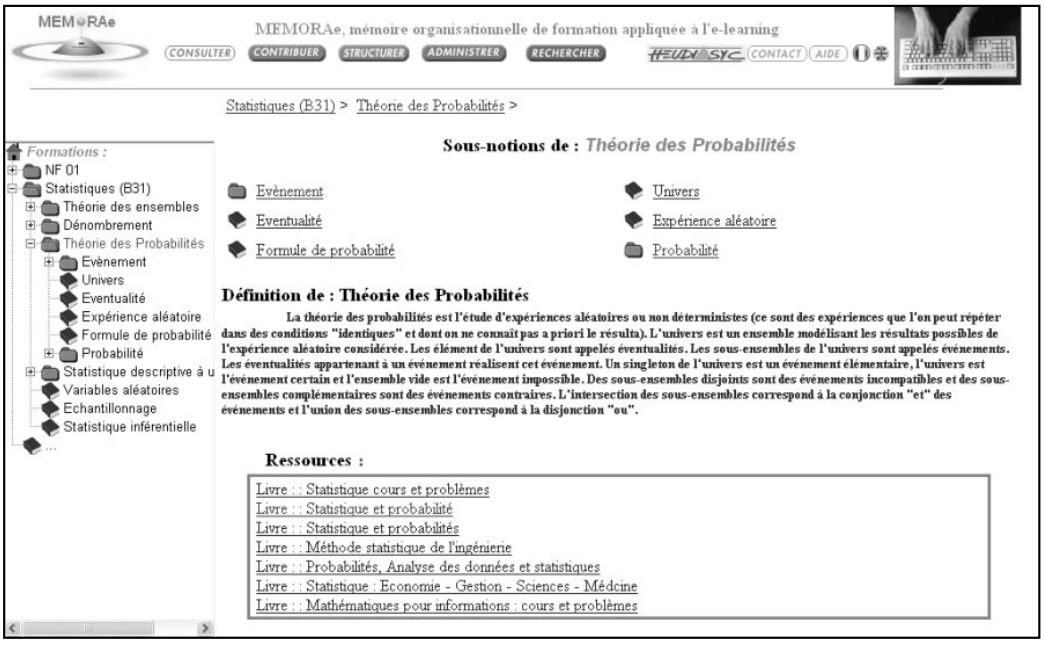


Figure N°2 : Interface d'interrogation de la mémoire.

4 Conclusion

Dans cet article, nous avons présenté l'approche que nous avons choisie dans le projet MEMORAe pour la mise en place d'une mémoire organisationnelle de formation. Cette approche consiste à utiliser des méta-données reposant sur des ontologies : une ontologie générique du domaine de la formation et une ontologie spécifique pour chaque application donnée. Afin d'assurer une bonne exploitation de la mémoire, MEMORAe intègre aussi le

formalisme des Topic Maps qui permet de décrire au mieux les ressources informationnelles de la mémoire.

Cette approche regroupe trois domaines de recherche : le découpage du contenu de la mémoire en grains de connaissance concerne le domaine de l'Ingénierie Éducative ; la gestion du contenu de formation au moyen d'une mémoire organisationnelle basée sur des ontologies concerne le domaine de l'Ingénierie des connaissances. Enfin, la structuration des informations en utilisant la norme ISO topic maps concerne le domaine du Web Sémantique. De ce fait, le projet MEMORAe se trouve au carrefour de ces trois domaines.

Nous nous intéressons actuellement aux travaux liés à la normalisation des objets pédagogiques [Bourda & al, 00]. Pour évaluer l'apport de notre mémoire dans le cadre d'une formation e-learning, nous avons commencé à développer deux applications : une unité de valeur concernant l'algorithmique et la programmation, et une autre concernant les statistiques.

Références

[Abel & al, 02] Abel, M.-H., Lenne, D., Cisse O. (juin 2002) "E-Learning and Organisational Memory", *Proceedings of IC-AI'02*, Las Vegas.

[Barry & al, 01] Barry, C., Cormier, C., Kassel, G., Nobécourt, J. (Juin 2001) "Évaluation de langages opérationnels de représentation d'ontologies", *Actes de IC'2001*, Grenoble.

[Berners-Lee & al, 01] Berners-Lee, T., Hendler, J., and Fensel, D. (2001) "The Semantic Web", *Scientific Americain 78(3)*, p 20-88.

[Boullier, 2001] Boullier, D. (2001) "Les choix techniques sont des choix pédagogiques : les dimensions multiples d'une expérience de formation à distance", *Sciences et Techniques Éducatives, vol. 8, n°3-4, pp. 275-299*.

[Caussanel & al, 02] Caussanel, J., Cahier, J.-P., Zacklad, M., Charliet J. (Mai 2002) "Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web Sémantique ?", *Actes de IC'2002*, Rouen.

[Dieng & al, 00] Dieng, R., Corby, O., Giboin, A., Golebiowska, J., Matta, N., Ribière, M. (2000). "Méthodes et outils pour la gestion des connaissances", Dunod.

[Gandon & al, 02] Gandon, F., Dieng, R., Corby, O., Giboin, A. (Mai 2002) "Web Sémantique et Approche Multi-Agents pour la Gestion d'une Mémoire Organisationnelle Distribuée", *Actes de IC'2002*, Rouen.

[GRU 93] Gruber, T. (1993)"A transaction approch to portable ontology specification", *Knowledge Acquisition*, 5(2) : 199-220.

[IEC, 99] International Organisation for Standardization (ISO), International Electronical Commission (IEC), Topic Map, International Standard ISO/IEC 13250, 19 april 1999.

[Kassel & al, 00] Kassel, G., Abel, M.-H., Barry, C., Boulitreau, P., Iraztorza, C., Perpette, S. (10-12 Mai 2000) "Construction et exploitation d'une ontologie pour la gestion des connaissances d'une équipe de recherche", *Actes de IC'2000*, Toulouse, p 251-259.

[Le Grand, 01] Le Grand, B. (décembre 2001) "Extraction d'information et visualisation de systèmes complexes sémantiquement structurés", *thèse de doctorat de l'université de pierre et marie curie*.

[Pepper, 99] Pepper, S. (June 2000) *The TAO of Topic Maps : Findinq the way in the age of infoglut*, Inproceedings of XML Europe 2000 Conférence, Paris.

[Uschold & Gruninger, 96] Uschold, M., Gruninger, M. (1996) "ONTOLOGIES : Principles, Methodes and Applications", *Knowledge Engineering Review*. Vol. 11; n°2.

[W3C, 01] World Wide Web Consortiun, *XML Linking Language (Xlink) Version 1.0*, W3C Recommendation, 27 June 2001.

[XTM, 01] TopicMaps.org XTM Authoring Group, *XML Topic Maps (XTM) 1.0 : TopicMaps.org Specification*, 3 March 2001.

Comment représenter les ontologies pour un Web Sémantique Médical ?

C. GOLBREICH[1], O. DAMERON[2], B. GIBAUD[2], A. BURGUN[1]

(1)Laboratoire d'Informatique Médicale
Faculté de Médecine, Av du Pr. Léon Bernard, 35043 Rennes France
Mail : christine.golbreich@uhb.fr, anita.burgun@univ-rennes1.fr

(2)Laboratoire IDM, UPRES-EA 3192
Faculté de Médecine, Av. du Pr. Léon Bernard, 35043 Rennes Cedex France
Mail :{odameron, bernard.gibaud}@univ-rennes1.fr

Résumé

La communauté biomédicale a des besoins réels et concrets d'utilisation d'un futur Web Sémantique. Une question importante est de savoir si les langages standards d'ontologies du Web permettront de répondre à ses principaux besoins. L'objectif de l'article est d'apporter une contribution à ce sujet en évaluant sur une ontologie médicale concrète deux des langages-outils disponibles pour développer des ontologies, Protégé et DAML+OIL. Cette expérience a permis de dégager certaines limitations d'expressivité de ces langages et des exigences qui semblent importantes à satisfaire pour représenter les ontologies médicales en vue d'un futur Web Sémantique.

Abstract

The biomedical community has concrete needs of a future Semantic Web. An important issue is to know whether the standard Web Ontology Languages will satisfy its main needs. This paper aims at contributing to this question in evaluating two languages, Protégé and DAML+OIL, on a medical ontology. This experiment has pointed out some expressiveness limitations and requirements of the biomedical domain that seem important for the future Semantic Web

1 Introduction

Avec l'explosion du Web et la prolifération des connaissances biomédicales, les utilisateurs du monde médical ont « potentiellement » accès à des informations de plus en plus nombreuses. Mais en réalité, obligés de naviguer dans un véritable labyrinthe de pages, il leur est encore difficile d'obtenir de façon « satisfaisante », c'est-à-dire rapidement, avec le minimum de bruit et de silence possible l'information médicale récente voulue. Si ce problème revêt une importance particulière pour les communautés biomédicales, il n'est pas propre à ce domaine et rejoint des questions d'ordre plus général. Le défi majeur actuel pour le Web est de tendre vers un « Web Sémantique » où l'information a un sens explicite, permettant aux machines de mieux exploiter et intégrer les données disponibles pour en faciliter l'accès. Le moyen d'y parvenir est le marquage sémantique des données.
Les ontologies ont un rôle central pour un Web Sémantique, puisqu'elles définissent les concepts à utiliser pour le marquage sémantique des ressources, sans en donner d'acception particulière, mais en

visant une signification partagée et réutilisable pour différentes applications et différents usagers. Toutefois, la représentation des ontologies pose des problèmes difficiles, comme en témoigne la création par le W3C en 2001 d'un groupe de travail Web Ontology chargé de concevoir un langage d'ontologies du Web « Web Ontology Language ». Le domaine biomédical présente des besoins réels et concrets d'ontologies pour le Web. L'objectif de l'article est de contribuer à évaluer deux des langages-outils actuels disponibles pour les ontologies du Web, sur une ontologie médicale concrète, et en dégager certaines caractéristiques qui semblent importantes pour représenter les ontologies du domaine biomédical. L'enjeu est de contribuer à la mise à disposition de langages apportant des solutions adéquates en vue d'un Web Sémantique Médical (WSM). On peut penser que ces solutions trouveront des applications dans d'autres domaines du Web, manisfestant des besoins similaires.
Après un rapide survol section 2 des grands besoins du domaine biomédical par rapport au Web, la section 3 donne un aperçu des principaux langages candidats pour modéliser et représenter les ontologies du Web. La section 4 vise à déterminer à partir de la représentation de l'ontologie de l'anatomie du cortex cérébral en PROTEGE et DAML+OIL, l'apport et les limitations de ces langages pour la représentation d'ontologies du domaine biomédical, dans la perspective d'un Web Sémantique.

2 Besoins du domaine biomédical

Les besoins majeurs attendus en médecine du Web Sémantique sont de pouvoir trouver facilement des informations médicales sur le Web, partager des informations grâce au Web et ainsi pouvoir les exploiter pour l'aide à la décision.
La *recherche d'informations* sur le Web concerne de multiples utilisateurs du monde médical. Trouver rapidement sur le Web, avec le minimum de bruit et de silence possible, une information scientifique récente a un intérêt non seulement pour le chercheur qui doit accéder à des bases hétérogènes et réitérer régulièrement les interrogations sur ces bases, pour les patients à la recherche d'informations, mais aussi dans la pratique médicale quotidienne où médecins et industriels pharmaceutiques sont amenés à rechercher de l'information en disposant de peu de temps. Les moteurs de recherche sur le Web existants basés sur des mots-clé, peuvent retourner des documents non pertinents, à cause par exemple d'homonymie ou de mauvais contexte, ou rater des documents à cause de l'utilisation de mots différents (synonymie), de mots plus spécifiques, ou plus généraux (hypo-hyperonymie). La solution des banques documentaires est d'indexer et rechercher les documents à l'aide de « descripteurs » d'un thesaurus plutôt que par mot-clé. C'est l'approche retenue pour des thesaurus comme Digital Anatomist [Rosse et al.], ou MEDLINE basée sur le thesaurus MeSH. Si les thesaurus et langages existants représentent un énorme acquis qu'on ne peut ignorer, ils ont toutefois certaines limitations. L'imprécision de la définition de certains concepts peut conduire à des significations non partagées liées à des interprétations différentes de leur sens, ce qui pose un problème pour la réutilisation et l'interopérabilité des thesaurus (ontologies). Certains concepts plus pointus peuvent être absents (cf.§4). L'évolution rapide des connaissances en médecine et le caractère fondamentalement dynamique du Web, nécessitent des mises à jour fréquentes, d'où l'importance de la gestion de versions, de la détection d'incohérences lors d'ajout ou de modifications.
De même, le *partage d'informations* revêt en médecine de multiples aspects, notamment du fait de l'association étroite entre activités cliniques, recherche, et formation. Longtemps empirique et exercée de façon individuelle, la médecine est devenue plus scientifique, de plus en plus spécialisée, et donc exercée plus collectivement. Le partage des données médicales, de plus en plus nombreuses et complexes compte tenu des progrès de la biologie et de l'imagerie - devient indispensable pour assurer la délivrance et la continuité des soins. La constitution d'entrepôts de données, ou la mise en place d'entrepôts Web de bases d'informations fédérées, « Webhouse », articulés autour d'ontologies communes est un besoin primordial. L'enjeu est particulièrement important dans les disciplines à caractère fortement pluri-disciplinaire – comme par exemple les neurosciences [Toga], [Brinkley et al.] – qui requièrent le partage à la fois de données et de connaissances, et suppose donc « l'alignement » d'une façon ou d'une autre de plusieurs ontologies de domaine (clinique neurologique ou psychiatrique, neuroimagerie, anatomie, génome, modèles neuronaux, neurochimie, etc.). Le cas de l'imagerie est aussi particulièrement important, compte tenu de la nécessité de décrire les signaux et les images de façon adéquate, pour expliciter le contexte et le protocole d'acquisition et de traitement, ainsi que les résultats. Le problème ne se limite pas au partage des données médicales.

L'enjeu est aussi - et peut-être surtout - *l'utilisation* de ces données dans le cadre de l'aide à la décision [Mendonca et al.]. Ceci recouvre encore une fois de multiples aspects. Citons par exemple l'accès à des protocoles et guides de bonnes pratiques (prise en charge, prescription) et la référence à des modes de prise en charge optimaux (evidence-based medicine), domaines dans lequel le web sémantique doit apporter, à relativement peu de frais, des contributions notables [Sakai]. Il peut également s'agir à proprement parler d'aide à la décision dans le cadre de systèmes intelligents ; ce domaine s'avère beaucoup plus exigeant, car il nécessite la formalisation des données médicales sous une forme assimilable par ces machines, et suppose des *connaissances*, elles aussi suffisamment formalisées pour interpréter et traiter correctement ces données.
Les nouveaux langages, méthodes, outils du Web Sémantique, devraient contribuer de façon notable à mieux répondre à ces besoins en facilitant l'indexation, la recherche et le partage d'informations médicales, en vue d'une meilleure utilisation de ces données. Un langage formel adéquat avec une sémantique bien définie pour représenter les ontologies (§3) devrait contribuer à apporter des réponses à ces questions en permettant de donner une définition formelle précise des concepts, en fournissant des mécanismes de raisonnement puissants permettant la vérification automatique de cohérence, la classification automatique des concepts, de faire des inférences assurant une recherche plus « intelligente », et d'interroger des sources hétérogènes et réparties de documents [Rousset et al.].

3 Langages pour les ontologies

Si on part de la définition d'une ontologie du domaine, comme un modèle formel décrivant les concepts du domaine et leurs relations, un langage pour les ontologies doit offrir les primitives épistémologiques nécessaires pour décrire des concepts de l'ontologie (classes), leurs propriétés et leurs relations (attributs ou rôles), des restrictions sur ces propriétés (facettes ou restriction de rôles).

3.1 Modélisation

Différents éditeurs d'ontologie qui supportent ces définitions sont actuellement proposés en particulier, Protégé 2000 [Noy et al.] http://protege.stanford.edu, OntoEdit et OILEd [Bechhofer et al.] qui sont deux éditeurs[1] pour le langage DAML+OIL.

- Protégé-2000 offre un environnement graphique interactif pour la conception d'ontologies (Figure 1). Un arbre permet une navigation rapide et simple dans la hiérarchie. Le modèle de Protégé-2000 est basé sur les frames. Il aide à définir les classes et les hiérarchies avec héritage multiple ; les attributs, les restrictions de valeurs de ces attributs, leurs facettes, comme les restrictions de cardinalité, valeurs par défaut, ainsi que des attributs inverses, des métaclasses et la hiérarchie de métaclasses. Il faut noter que des interfaces à l'UMLS™ [Lindberg et al] et WordNet permettent aux utilisateurs d'intégrer et importer des éléments de ces grandes sources de connaissances *on-line* dans leurs ontologies.
- OILEd est un éditeur graphique développé par l'Université de Manchester qui permet à l'utilisateur de construire des ontologies représentées en DAML+OIL. Le modèle de OILEd est basé sur DAML+OIL [DAML+OIL]. Tout en offrant une interface de modélisation de type « frame », paradigme plus familier aux utilisateurs que celui de la logique de description, il supporte toute l'expressivité de la logique de description OIL et DAML+OIL. Les classes sont définies en terme de leurs super-classes, de leurs propriétés avec restrictions de type, et en outre la possibilité de définir des axiomes, par exemple pour définir des classes disjointes (Figure 2). Le modèle permet de définir des descriptions complexes comme valeur des attributs, à l'opposé de la plupart des éditeurs de frames où les classes doivent être nommées pour pouvoir être utilisées.

3.2 Représentation

Il existe de nombreuses présentations des langages standards ou en cours de standardisation du W3C (http://www.w3.org): XML, RDF, RDFS, DAML+OIL, OWL, leur

[1] cf. A survey on ontology tools : OntoWeb Ontology-based information exchange for knowledge management and electronic commerce IST-2000-29243 Deliverable 1.3, 31 st May, 2002 pour une présentation plus exhaustive des outils existants

structuration en couche, et leur évolution. Cette section se limite à en rappeler les très grandes lignes, en se focalisant sur les langages d'ontologie du Web et les aspects en rapport avec un WSM.

- *"XML is like HTML, where you make up your own tags" :* si HTML est un langage d'annotation pour décrire la présentation du contenu d'un document, XML [XML] est un langage pour décrire sa structure. Il est possible d'utiliser ses propres balises. XML et XML Schema [XML Schema] sont des langages qui pourraient être suffisants pour publier ou échanger des données médicales.
- RDF [RDF] est un langage pour décrire les ressources du Web et leurs méta-données. Elles sont décrites par des triplets [Propriété Sujet Objet]. RDFS introduit en plus la possibilité de définir des classes et des hiérarchies, des propriétés, de contraindre leur domaine (range, domain). RDF [RDF] et RDFS peuvent être éventuellement suffisants pour une exploitation des méta-données (documentation) ou une navigation classiques sur le Web. Mais pour une recherche de documents ou une navigation plus intelligente sur le Web, un langage formel plus expressif est nécessaire.
- Le langage DAML+OIL [DAML+OIL] est un langage conçu par le groupe du W3C WebOnt dans le but de dépasser la simple « présentation » d'informations sur le Web pour aller vers l'interopérabilité, la compréhension et le raisonnement sur ces informations. DAML+OIL est le résultat de la fusion de OIL résultant du projet européen OntoKnowledge, et de DAML-ONT, issu du projet DARPA DAML (American Agent Markup Language) [Hendler et al.]. Il doit ses primitives de modélisation intuitives aux « frames », sa syntaxe aux standards XML et RDF, sa sémantique formelle et ses mécanismes de raisonnement aux logiques de description. D'un point de vue formel DAML+OIL est basé sur la logique de description expressive $\mathcal{SHIQ}$ étendue du constructeur oneOf et de types de données (cf. colonnes 1 et 3 Tableau 1). DAML+OIL permet de définir en outre un ensemble d'axiomes (cf. colonnes 1 et 3 Tableau 2). DAML+OIL est accompagné de différents outils : un éditeur mais surtout un classifieur FaCT (Fast Classification of Terminology) qui permet de détecter automatiquement les incohérences, et classer automatiquement les concepts d'une ontologie.
- Le futur standard OWL [OWL] est le successeur de DAML+OIL. OWL fournira trois sous-langages d'expressivité croissante : OWL Lite, OWL DL et OWL Full. La sémantique formelle des logiques de description est définitivement acquise, et ces différentes couches de langages, de niveau d'expressivité et de complexité différents, sont en phase avancée de standardisation au W3C (au 10 février 2003). OWL Lite aura moins de constructeurs de base, en particulier pas la disjonction ni la négation (mais qui pourraient être capturés). Les différentes fonctionnalités souhaitées pour chacun de ces langages détermineront les constructeurs finaux retenus pour chacun d'eux, le centre de la question étant l'opposition entre *expressivité* et *complexité - propriétés* attendues des algorithmes (décidabilité, correction, complétude). OWL-DL garantirait la complétude et la décidabilité, tandis que OWL Full offrant un maximum d'expressivité et la liberté de syntaxe de RDF, serait sans garantie computationelle. Il reviendra au développeur de l'ontologie de choisir le langage qui convient à ses besoins.

En bref, XML fournit la syntaxe de la couche transport, RDF/RDF(S) les primitives ontologiques de base (le modèle simple de données de RDF et les schémas de RDF), DAML+OIL la couche logique (la sémantique formelle de OIL et DAML+OIL); enfin d'autres couches (notamment règles) viendront étendre DAML+ OIL [Bechhofer et al.]. La section suivante vise à évaluer les langages Protégé et DAML+OIL sur une ontologie concrète du domaine médical en cours de développement.

4 Ontologie Web de l'anatomie du cortex cérébral

Il y a plusieurs raisons de construire une ontologie Web de l'anatomie du cortex cérébral. D'abord, les notions anatomiques sont très souvent référencées par les autres connaissances médicales, d'où l'intérêt d'une ontologie générale de l'anatomie qui soit partageable. Ainsi, divers travaux comme Galen [Galen] ou Digital Anatomist [Rosse et al.] portent sur l'anatomie. Pour autant, ces travaux sont axés sur la notion d'organes et n'offrent pas de description suffisamment fine de l'organisation du cortex cérébral. De plus, une ontologie Web de l'anatomie du cortex pourra être réutilisée par divers champs d'application : (1) l'enseignement, (2) l'aide à la décision lors de la pratique clinique, (3) ou en neuroimagerie pour partager des données entre plusieurs sites et ainsi améliorer la qualité statistique des résultats de recherche, par une plus grande taille d'échantillons. Cette ambition de généricité et d'indépendance du domaine applicatif trouve un écho dans le Web Sémantique.

4.1 Ontologie en Protégé

Le langage de frames Protégé avec son éditeur Protégé2000 ont été utilisés pour représenter une partie de l'ontologie. Les définitions de l'exemple s'appuient sur des descriptions données dans les atlas d'anatomie e.g. [Ono et al.] ou des terminologies comme NeuroNames [Bowden et al.]. Un « hémisphère cérébral » est défini comme une partie anatomique du cortex *latéralisée* (située soit du côté gauche soit droit), qui *a* normalement cinq subdivisions anatomiques appelées lobes (lobe frontal, temporal, pariétal, occipital et limbique), *occupe une région* spatiale bien spécifique, etc. Un «hémisphère» est représenté en Protégé par la classe Hémisphère, avec les attributs hasLocation, hasSide, hasDirectAnatomicalPart, et ses facettes de restriction at least, at most (valeur : 5). On note qu'on exprime ainsi qu'un hémisphère comprend cinq lobes, tous types confondus. Il serait compliqué (mais possible) d'exprimer qu'il a exactement un et un seul lobe de chaque type (frontal, temporal, pariétal, occipital, limbique) (cf. Ex14 §4.3). Une « partie anatomique du cortex latéralisée à gauche » est représentée par la classe LeftLateralizedAnatomicalConcept avec des contraintes sur les valeurs de ses attributs : restriction de hasSide à LeftSide et sur hasDirectAnatomicalPart (les partie directes). Un «hémisphère gauche» est représenté par la classe LeftHemisphere, il n'a pour parties que des lobes gauches (restriction : LeftLobe) (Figure 1). On voit sur la hiérarchie que LeftHemisphere est une sous-classe de Hemisphere et de LeftLateralizedAnatomicalConcept.

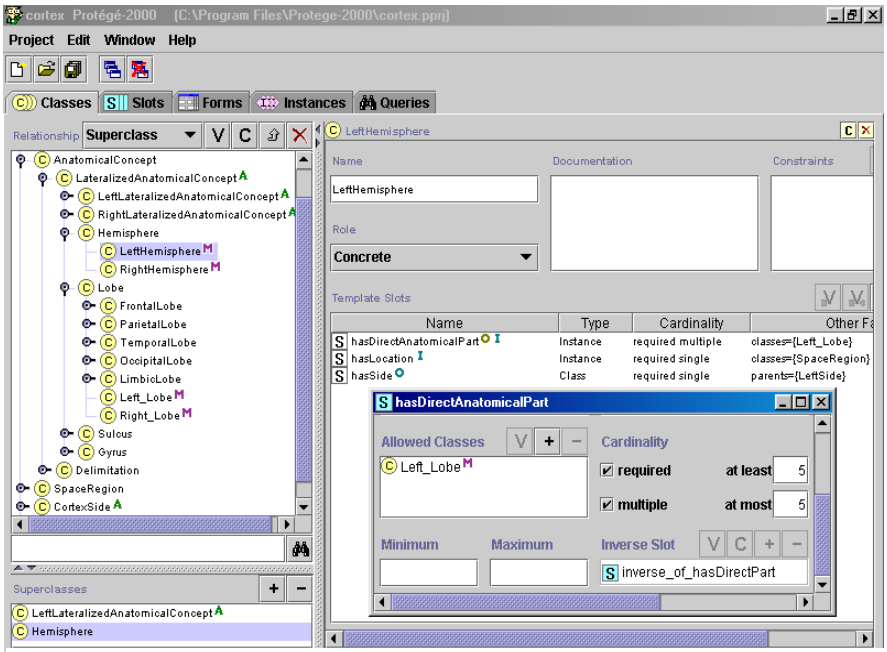


Figure 1 Hiérarchie de classes, concept LeftHemisphere, slot hasDirectPart

4.2 Ontologie en DAML+OIL

La logique de description $\mathcal{SHIQ}$ du langage DAML+OIL avec l'éditeur OILEd ont été utilisés. pour représenter l'ontologie du cortex cérébral, dont voici quelques exemples de définitions.

Ex1. /Un concept anatomique a pour parties directes des concepts anatomiques, a pour localisation exactement une région de l'espace/
AnatomicalConcept := (∀ hasDirectAnatomicalPart AnatomicalConcept) ∧ (= 1 hasLocation SpaceRegion)

Ex2. /Un concept latéralisé a pour côté soit le côté droit du cortex soit le côté gauche, on distingue ceux de type droit et ceux de type gauche/
LateralizedAnatomicalConcept := AnatomicalConcept ∧ (= 1 hasSide LeftSide ∨ RightSide)
LeftLateralizedAnatomicalConcept := LateralizedAnatomicalConcept ∧ (∀ hasSide LeftSide)
RightLateralizedAnatomicalConcept := LateralizedAnatomicalConcept ∧ (∀ hasSide RightSide)

Ex3. /Un hémisphère est un concept latéralisé dont toutes les parties directes sont des lobes, et qui a exactement un lobe de chaque type/
Hemisphere := LateralizedAnatomicalConcept ∧ (∀ hasDirectAnatomicalPart Lobe) ∧ (= 1 hasDirectAnatomicalPart FrontalLobe) ∧ (= 1 HasDirectAnatomicalPart ParietalLobe) ∧ (= 1 hasDirectAnatomicalPart OccipitalLobe) ∧ (= 1 hasDirectAnatomicalPart LimbicLobe) ∧ (= 1 hasDirectAnatomicalPart TemporalLobe)

Ex4. /Un hémisphère (resp. un lobe etc.) gauche (resp. droit) est un hémisphère de côté

```
gauche (resp.droit)/
LeftHemisphere := LeftLateralizedAnatomicalConcept ∧ Hemisphere
```

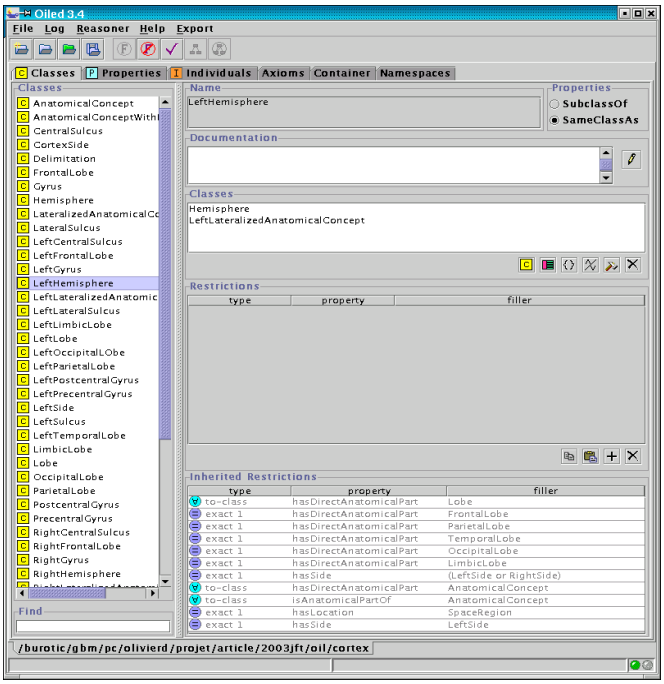


Figure 2 Définition avec OILEd du concept LeftHemisphere

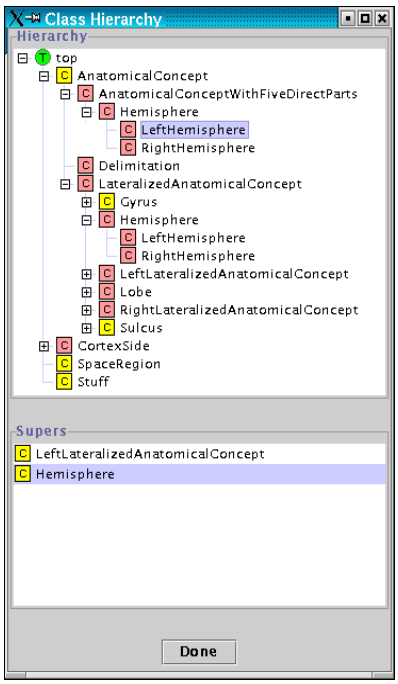


Figure 3 Hiérarchie postclassification

Ces exemples mettent en évidence qu'un tel formalisme logique permet d'une part des définitions rigoureuses et précises de concepts, donc partageables sans faux sens, et d'autre part l'automatisation de certaines tâches, indispensable pour les vastes ontologies du domaine médical. Par exemple, la définition du concept Hemisphere précise qu'un hémisphère n'a pour parties directes que des Lobe, dont un et un seul Lobe de chaque type : FrontalLobe, ParietalLobe etc. (Ex3). Il résulte des contraintes définies sur ses propriétés que Hemisphere est automatiquement placé par le classifieur FaCT dans la hiérarchie comme subsumé par FiveDirectPartAnatomicalConcept (Figure 3), et donc LeftHemisphere aussi, vu sa définition comme sous-classe de LeftLateralizedAnatomicalConcept et de Hemisphere, alors qu'il ne l'était pas dans la hiérarchie initiale pré-classification.

4.3 Expressivité requise et limites des langages utilisés

L'utilisation de Protégé et de DAML+OIL a permis de répertorier certaines limitations d'expressivité des formalismes de représentation utilisés et d'identifier certains besoins du domaine biomédical importants pour les futures ontologies du Web. D'abord l'utilisation de ces deux langages illustre le vieux débat en représentation des connaissances entre logique et frames, au cœur duquel est toujours la même question : d'une part l'opposition entre expressivité et propriétés attendues des algorithmes (complexité, décidabilité, correction, complétude), et d'autre part la facilité d'usage des frames. Ensuite, la représentation de l'ontologie de l'anatomie du cortex a soulevé des difficultés dans les deux formalismes, mais il n'est pas étonnant que les limitations d'expressivité observées avec Protégé ont été pour la plupart résolues avec la logique de description $\mathcal{SHIQ}$ de DAML-OIL, plus expressive. Les exemples suivants sont représentatifs de besoins récurrents pour les ontologies médicales. Ils montrent que pratiquement toutes les primitives de DAML+OIL (OWL-DL) ont été utilisées (§ 4.3.1), mais aussi que d'autres éléments absents de Protégé ou DAML+OIL (OWL-DL) seraient nécessaires pour répondre aux besoins rencontrés en médecine, en particulier les règles (§ 4.3.2 et § 4.3.3).

4.3.1 Besoin d'une expressivité importante

Les exemples 1 à 4 (cf. 4.2) et 5 à 14 montrent les primitives qui ont été utilisées pour l'anatomie du cortex cérébral. Pour chaque exemple le numéro renvoie au constructeur ou à l'axiome de Protégé et DAML+OIL qui a été utilisé (listés respectivement dans les colonnes 2 ou 3 des Tableaux 1 et 2).

Ex5. **disjointWith** nécessaire pour exprimer l'exclusion de concepts (pas d'instance en commun) (n°10)

```
/Les classes LeftLateralizedAnatomicalConcept RightLateralizedAnatomicalConcept sont disjointes, ou encore Hemisphere, Lobe, Gyrus et Sulcus sont des concepts anatomiques disjoints deux à deux/
disjointWith (Lobe | Gyrus)
```

Ex6. **unionOf** nécessaire pour exprimer une alternative (n°2)

```
/Une partie anatomique latéralisée est droite ou gauche/
LateralizedAnatomicalConcept := AnatomicalConcept ∧ (= 1 hasSide LeftSide ∨ RightSide)
```

Ex7. **complementOf (négation)** nécessaire pour définir un concept par opposition à un autre (n°6)

```
/Une partie anatomique non latéralisée e.g. corps calleux/
NonLateralizedAnatomicalConcept := AnatomicalConcept ∧ ¬ LateralizedAnatomicalConcept
```

Ex8. **disjointUnionOf**[2] nécessaire pour exprimer une partition de A en List (n°11)

```
/Un côté du cortex est soit le droit soit le gauche/
DisjointUnionOf(CortexSide [LeftSide RightSide])
/Un lobe est soit frontal, soit temporal, pariétal, occipital, soit limbique/
DisjointUnionOf(Lobe [FrontalLobe ParietalLobe OccipitalLobe LimbicLobe TemporalLobe])
/Un hemisphere est soit un hémisphère droit, soit un hémisphère gauche/
DisjointUnionOf(Hemisphere [LeftHemisphere RightHemisphere])
```

Ex9. **equivalentClass** (≡) pour exprimer l'équivalence de classes (n°9)

```
/le concept lobe gauche est équivalent à partie latéralisée gauche et lobe/
LeftLobe ≡ LeftLateralizedAnatomicalConcept ∧ Lobe
```

Ex10. **subpropertyOf** (⊆) pour hiérarchiser les relations (n°12)

```
/la relation hasAnatomicalPart est un type particulier de la relation hasDirectAnatomicalPart/
hasDirectAnatomicalPart ⊆ hasAnatomicalPart
```

Ex11. **transitiveProperty** pour exprimer des propriétés de relations comme la transitivité (n°15)

```
/has-part est transitive (hasDirectPart ne l'est pas/
Slot-def has-part Properties transitive
```

Il est important de pouvoir définir des propriétés comme la réflexivité, la symétrie ou la transitivité. Ainsi, la transitivité rend explicite la différence entre hasDirectAnatomicalPart et hasAnatomicalPart, la dernière correspond à la fermeture transitive de la seconde. Par exemple, les parties anatomiques directes d'un Hemisphere sont des Lobes, celles d'un Lobe sont des Gyri, tandis que les parties anatomique d'un Hemisphere sont des Lobes ou des Gyri. Protégé ne permet pas d'expliciter de telles propriétés, mais c'est possible avec DAML+OIL.

Ex12. **inverseOf** pour exprimer une relation inverse (n° 14)

```
isLocatedIn inverseOf hasLocation
```

Ex13. **equivalentRelation** (n°13)

```
/ un concept est partie anatomique d'un autre si et seulement si la région occupée par le premier est incluse dans celle du second /
isAnatomicalPartOf ≡ (isLocated o isSubAreaOf o hasLocation)
```

Cette équivalence est nécessaire pour déduire des contraintes sur les régions de concepts liés par isAnatomicalPartOf, ou à l'inverse déduire que deux concepts sont liés par isAnatomicalPartOf à partir de leurs régions respectives. Etablir explicitement des équivalences entre relations est important pour établir des ponts entre plusieurs ontologies.

Ex14. **cardinalité** et contraintes non exclusives sur la même relation (n°7)

```
/Un hémisphère est un concept latéralisé dont toutes les parties directes sont des lobes, et qui a exactement un lobe de chaque type/
Hemisphere := LateralizedAnatomicalConcept ∧ (∀ hasDirectAnatomicalPart Lobe) ∧ (= 1 hasDirectAnatomicalPart FrontalLobe) ∧ (= 1 HasDirectAnatomicalPart ParietalLobe) ∧ (= 1 hasDirectAnatomicalPart OccipitalLobe) ∧ (= 1 hasDirectAnatomicalPart LimbicLobe) ∧ (= 1 hasDirectAnatomicalPart TemporalLobe)
```

Un langage de frame comme Protégé ne permet pas de représenter ce type de contrainte, il serait au mieux possible de définir cinq relations spécifiques pour exprimer qu'un Hemisphere a exactement un Lobe de chaque type (sans toutefois pouvoir représenter que ces relations sont des spécialisations de hasDirectAnatomicalPart). Par contre, c'est possible en DAML+OIL avec le constructeur = n r.C.

Il ressort de ces exemples que l'expressivité de OWL Lite n'est pas suffisante pour l'ontologie de l'anatomie du cortex puisque l'union, la négation, entre autres, sont nécessaires et donc au moins l'expressivité de OWL DL requise.

[2] cette primitive existait dans Oil mais n'existe plus dans OWL, néanmoins elle peut être capturée grâce à unionOf et disjointWith

4.3.2 **Besoin d'une couche « règles »**

DAML+OIL n'a pas permis d'exprimer certaines connaissances. Les exemples suivants 15 à 17 montrent des propriétés complexes hors de portée de l'expressivité de langages comme DAML+OIL (OWL), et les avantages qu'offriraient des règles Web pour l'ontologie de l'anatomie du cortex. « The Rule Markup Initiative » (http://www.dfki.uni-kl.de/ruleml/#Papers-Publications ou le langage CARIN–$\mathcal{ALN}$[Rousset et al.] sont centrés sur cette question.

`Ex15.` **besoin de règles d'inférence** (n° 21)

Règle 1 : si A est une partie de B, alors A a même côté que B

`isAnatomicalPartOf (A B) ∧ hasSide (B,C) → hasSide (A,C)`

Règle 2 : si C partie-de D, non (A partie-de D), S sépare (A,C) alors S sépare (A, D)

`isAnatomicalPartOf (C D) ∧ ¬ isAnatomicalPartOf (A D) ∧ separation (S A C) → separation (S A D)`

Les règles sont nécessaires pour exprimer des dépendances entre des concepts ou des propriétés, ou des contraintes de cohérence, comme : *si un sillon sépare deux gyri (G1 et G2), et que ces gyri sont des parties de lobes différents (G1 est une partie anatomique de L1 et G2 de L2), alors le sillon sépare également G1 de L2, G2 de L1 et L1 de L2* (cf. Règle 2 Ex15). C'est le cas par exemple du sillon central, qui sépare le gyrus précentral, appartenant au lobe frontal, du gyrus postcentral, appartenant au lobe pariétal. De telles règles, permettent de générer 221 des 370 relations que compte actuellement le modèle de l'anatomie proposé par [Dameron et al.]. La première règle ci-dessus pourrait être utilisée à la place de certains concepts définis dans l'ontologie, mais sans grand intérêt pour l'utilisateur. De plus de telles règles sont importantes pour la construction et pour assurer la cohérence de l'ontologie lors de mises à jour, ou de sa liaison à d'autres ontologies.

`Ex16.` **besoin de composer des relations**

/ soit C un concept qui occupe une région qui est incluse dans la région occupée par un autre concept. /

$\texttt{hasLocation}^{-1} \circ \texttt{isSubAreaOf} \circ \texttt{hasLocation}$

`la relation g°f peut être représentée par un prédicat h défini par f(x,y)et g(y,z) ↔ h(x, z), avec f,g,h déclarés fonctionnelles (FunctionalProperty)`

`Ex17.` **besoin de représenter des relations d'arité supérieure à 2.**

/Un Sulcus sépare deux Lobes, deux Gyrus ou un Lobe et un Gyrus/

`Separation := AnatomicalConcept∧(= 1 separator Sulcus)∧(= 2 parts Lobe∨Gyrus)` (1)

`parts(S, V) ∧ 1stPart(V, A) ∧ 2ndPart(V, B) → separation(S, A, B)` (2)

Les frames et logiques de description ne permettent de représenter que des relations binaires. Il est nécessaire de représenter des relations d'arité supérieure, par exemple pour exprimer qu'un sillon (Sulcus) sépare deux Gyri, deux Lobes ou un Gyrus et un Lobe. Une première possibilité (1) est de représenter la relation par un concept `Separation`, avec un role `separator` de cardinalité 1 à valeur dans `Sulcus` et un role `parts` de cardinalité 2 à valeur dans `Gyrus` ou `Lobe`. Une autre solution (2) pour représenter une relation n-aire serait d'utiliser une règle avec des prédicats binaires comme dans [Rousset et al.] qui donne également d'autres cas d'usage intéressants de règles CARIN–$\mathcal{ALN}$.

Il ressort de ces exemples que l'expressivité de DAML+OIL et OWL DL, n'est pas suffisante pour l'ontologie de l'anatomie du cortex et qu'une couche de règles semble fortement nécessaire.

4.3.3 **Besoins de métaclasses**

Comme le montre l'exemple suivant, il est nécessaire de définir pour l'ontologie du cortex des métaclasses. Protégé2000 le permet mais pas DAML+OIL ni OWL-DL.

`Ex18.` **besoin de méta-classe** (n°22)

```
<!-- La classe FrontalLobe, instance de la métaclasse MetaAnatomicalConcept, est liée
par la propriété UMLS-ID à l'identifiant UMLS C0016733 -->
<MetaAnatomicalConcept rdf:ID="FrontalLobe">
<UMLS-ID rdf:resource="&rdfs;Literal">C0016733</UMLS-ID>
```

Les thesaurus et langages existants représentent un énorme acquis qu'on ne peut ignorer pour un futur Web Sémantique. Il parait indispensable de pouvoir relier les concepts de l'ontologie de l'anatomie du

cortex aux concepts de terminologies comme l'UMLS™. Chaque concept UMLS étant identifié par un « Concept Unique Identifier », e.g. C0016733 pour FrontalLobe, C0458332 pour CerebralLobe (Source : 2003AA release), l'idée est de définir une propriété liant les concepts de l'ontologie aux UMLS-ID correspondants. La solution pour définir des slots *propres* à chaque classe, non hérités par leurs sous-classes et leurs instances, est d'utiliser des métaclasses dont les instances sont des classes. Protégé-2000 permet de définir des métaclasses. C'est donc la solution retenue en Protégé par [Noy, Musen et al.], qui ont créé une métaclasse `Anatomical entity metaclass` avec des slots pour les IDs, UMLS-ID, et Terminologia Anatomica ID. Chaque nouvelle classe est une instance de `Anatomical entity metaclass`. Ce même besoin est généralisable aux autres ontologies médicales. Une option, serait de s'orienter pour certains usages vers OWL Full et ses facilités de méta-modélisation. Ex18 exprime schématiquement que la classe FrontalLobe est instance de la métaclasse MetaAnatomicalConcept où la propriété UMLS-ID liant un concept anatomique à un identifiant UMLS, a pour valeur C0016733.

Cette expérimentation a mis en lumière que

- la plupart des constructeurs de DAML+OIL, en particulier la négation, la disjonction (et partition), l'inverse, ont été tout à fait indispensables et semblent nécessaires pour les ontologies médicales du Web.
- L'équivalence de classes ou de relations, de sous-classe ou sous-propriétés sont des axiomes clé pour indiquer les relations entre classes et relations d'ontologies développées séparément, et donc particulièrement utiles pour fusionner, ou connecter plusieurs ontologies du Web
- Les limitations d'expressivité les plus importantes rencontrées avec DAML+OIL et sui se retrouveront avec le standard OWL-DL sont l'absence d'opérateur de composition, et surtout le besoin de règles et enfin de métaclasses (par contre possible en OWL FULL).

5 Discussion

Il est intéressant de comparer le standard OWL à d'autres langages formels existants qui pourraient être utilisés pour des ontologies du Web comme CARIN-$\mathcal{ALN}$ [Rousset et al.]. Le Tableau 1 donne les principaux constructeurs des langages PROTEGE 2000, DAML+OIL (sensiblement similaire à OWL DL), OWL Lite, CARIN-$\mathcal{ALN}$, et le Tableau 2 les axiomes, référençant le cas échéant un exemple du § 4.

Constructeur du langage	Protégé 2000	DAML+OIL syntaxe	Exemple	OWL Lite syntaxe	CARIN-$\mathcal{ALN}$ syntaxe
1. Conjonction	Non	C1 ∧ C2	Ex4	C1 ∧ C2 où C1 et C2 nommés ou restrictions	C1 ∧ C2
2. Alternative	Non	C1 ∨ C2	Ex2	Non	Non
3. Universalité	Oui	∀ r.C	Ex1	∀ r.C	∀ r C
4. Existentialité	Non	∃ r.C		∃ r.C	Non
5. OneOf	Non	{x1 … xn}		{x1 … xn}	Non
6. Négation	Non	¬ C	Ex7	Non	¬ C sur C de base
7. Cardinalité	simple ou multiple	≤ n r C ≥ n r C = n r C	Ex14	≤ n r C ≥ n r C = n r C pour n = 0 ou 1	≤ n r ≥ n r

Tableau 1 : Constructeurs de différents langages pour les ontologies

D'un point de vue formel DAML+OIL est équivalent à la logique de description $\mathcal{SHIQ}$ étendue du constructeur oneOf et de types de données, et par la possibilité de définir un ensemble d'axiomes tandis que CARIN-$\mathcal{ALN}$ combine le pouvoir d'expression de la logique de description $\mathcal{ALN}$ à celui d'un formalisme à base de règles. Le classifieur FaCT qui offre un raisonneur pour $\mathcal{SHIQ}$ avec des algorithmes de tableaux corrects et complets, peut être utilisé pour raisonner sur les ontologies en DAML+OIL ONTOCLASS offre les mêmes fonctionnalités pour CARIN–$\mathcal{ALN}$, mais la subsomption et la satisfiabilité y sont polynomiales, alors qu'elles sont exponentielles pour $\mathcal{DAML+OIL}$.

Axiome	Protégé 2000	DAML+OIL syntaxe	Exemple	OWL Lite syntaxe	CARIN-$\mathcal{ALN}$ syntaxe
8. Subsomption	Sous classe	subClassOf $C1 \subseteq C2$	Ex5	subClassOf $C1 \subseteq C2$	$C1 \subseteq$ Cexpression où A de base
9. Equivalence de classes		SameClassAs $C1 \equiv C2$	Ex9	equivalentClass $C1 \equiv C2$	Non
10. Exclusion	Non	DisjointWith $C1 \subseteq \neg C2$	Ex5	Non	$C1 \wedge C2 \subseteq \perp$ pour A et B de base
11. Partition	Non	DisjointUnionOf $C \equiv C1 \vee C2$ $C1 \wedge C2 \subseteq \perp$	Ex8	Non	Non
12. Sous relation	Non	subslot-of $r1 \subseteq r2$	Ex10	SubPropertyOf $r1 \subseteq r2$	Non
13. Equivalence de relation		SamePropertyAs $r1 \equiv r2$	Ex13	equivalentProperty	Non
14. inverse	inverse	InverseOf $r1 \equiv r2^{-1}$	Ex12	InverseOf	Non
15. transitivité	Non	transitive	Ex11	TranstitiveProperty	Non
16. symétrie	Non	symmetric		SymmetricProperty	Non
17. fonctionalité	Non	functional		FunctionalProperty	Non
18. injection	Non	Non		InverseFunctionalProperty	Non
19. object equivalence	Non	sameIndividualAs $\{x1\} \equiv \{x2\}$		sameIndividualAs	Non
20. object differentation	Non	differentIndividualFrom $\{x1\} \subseteq \neg \{x2\}$		differentIndividualFrom	Non
21. Règle		Non	Ex15	Non	Carin-rule
22. Métaclasse	Oui	Non	Ex18	Non	Non

Tableau 2: Axiomes des différents langages pour les ontologies

Pour répondre aux besoins du domaine médical pour un Web Sémantique, il sera sûrement nécessaire de combiner les avantages des formalismes de représentation de connaissances structurées comme les logiques de description avec des règles. La question ouverte est de déterminer quel langage serait plus adapté : un langage comme CARIN–$\mathcal{ALN}$ basé sur la logique de description $\mathcal{ALN}$ qui offre moins de constructeurs, mais dont le pouvoir d'expression est étendu par des règles, et qui reste de complexité polynomiale pour la subsomption, ou OWL-DL successeur de DAML+OIL, basé sur la logique de description plus expressive $\mathcal{SHIQ}$ mais où la satisfiabilité et la subsomption sont de complexité exponentielle. CARIN–$\mathcal{ALN}$, a un atout majeur, puisqu'il peut servir de langage de requêtes pour interroger des sources d'information hétérogènes et réparties via des médiateurs comme le permet l'environnement PICSEL [Rousset et al.]. Pour OWL un tel environnement, s'il est à l'étude, n'est pas encore disponible. Mais par contre, OWL-DL offre des constructeurs et des axiomes intéressants qui n'existent pas dans CARIN–$\mathcal{ALN}$. Toutefois les propriétés des algorithmes sont directement liées à l'expressivité des langages. Ainsi comme le souligne [Rousset et al.] « le choix de ne pas offrir de constructeur de disjonction ou d'exprimer des relations inverses dans la logique CARIN–$\mathcal{ALN}$ est justifié par la volonté de garantir la polynomialité du test de subsomption », de même que sa restriction pour le langage de vues de PICSEL a été nécessaire pour la décidabilité de la construction des plans de requêtes. Il en résulte plusieurs questions : serait-il possible d'autoriser d'autres axiomes dans CARIN–$\mathcal{ALN}$ sans modifier les propriétés du langage ? La polynomialité de la subsomption est-elle indispensable pour les services attendus par les communautés médicales d'un Web Sémantique et pour quel service? Un langage de requêtes n'est-il pas crucial pour un Web Sémantique Médical ? Pourrait-on éviter les métaclasses ? Une direction de recherche importante serait d'identifier un noyau de primitives *minimal* (constructeurs et axiomes), qui réponde au mieux aux principaux besoins du Web Sémantique de cette communauté, recherche et partage d'informations biomédicales, et de déterminer si OWL complété par une couche règles conviendrait, ou s'il faudrait plutôt un langage homogène unifiant logique de description et règles, comme CARIN–$\mathcal{ALN}$ [Rousset et al.]. L'enjeu est de contribuer à la mise à disposition d'un langage apportant des solutions adéquates en vue d'un Web Sémantique Médical. Il s'agit d'un problème difficile qui nécessite une étude approfondie pour identifier précisément les besoins du monde médical, l'expressivité et services souhaités, d'où les exigences à satisfaire par les langages d'ontologies du Web. Une collaboration pluridisciplinaire

étroite entre spécialistes des différentes disciplines, informatique médicale, informatique, médecine serait souhaitable à cette fin.

6 Conclusion

Il est admis aujourd'hui que pour aller vers un Web Sémantique, il est nécessaire de disposer d'un langage d'ontologie ayant une sémantique formelle bien définie. Ce langage doit offrir d'une part un pouvoir d'expression suffisant pour représenter finement de vastes quantités de connaissances médicales complexes, et d'autre part des mécanismes efficaces pour raisonner sur ces ontologies : classification automatique, vérification formelle de cohérence, services de réponse à des requêtes complexes mettant en jeu des informations hétérogènes et réparties, mécanismes de modularité et de réutilisation pour assembler des ontologies locales construites séparément. Cette expérimentation a montré qu'il est indispensable de pouvoir représenter, en plus de connaissances taxonomiques, des connaissances déductives. OWL est donc un bon candidat, mais à condition d'être complété par une couche de règles. Et puisque expressivité et « tractabilité » sont opposées, il faudra trouver un juste équilibre permettant d'exprimer les connaissances les plus importantes pour satisfaire les principaux usages du Web attendus de la communauté biomédicale : recherche d'informations et partage d'informations biomédicales hétérogènes pour l'aide à la décision. Idéalement, il faudrait certainement avoir un langage unifié combinant un sous-langage de OWL avec un langage de règles (version appropriée de RuleML).

7 Références

[Bechhofer et al.] Bechhofer S., Horrocks I., Goble C., Stevens R. OILEd: a Reason-able Ontology Editor for the Semantic Web. Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence, September 19-21, Vienna. Springer-Verlag LNAI Vol. 2174, pp 396--408. 2001.

[Bowden et al.] Bowden, DM and Martin, RF, NeuroNames Brain Hierarchy, Neuroimage, 1995, 2, 63-83

[Brinkley et al.] Brinkley J.F. and Rosse C. Imaging informatics and the Human Brain Project : the role of structure, Yearbook of Medical Informatics 2002, 131-148, 2002.

[DAML+OIL] DAML+OIL (March 2001) Reference Description. D. Connolly, F. van Harmelen, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. Andrea Stein. W3C Note 18 December 2001. Latest version available at http://www.w3.org/TR/daml+oil-reference.

[Dameron et al.] Dameron O., Burgun A., Morandi X., Gibaud, B. Modelling dependencies between relations to insure consistency of a cerebral cortex anatomy knowledge base. Proc. Medical Informatics in Europe 2003.

[Galen] Galen RAL and Nowlan, WA and the GALEN Consortium, The GALEN Project Computer Methods and Programs in Biomedicine, 1993, 45, 75-78

[Hendler et al.] Hendler, J. and McGuinness, D.L. (2000). The DARPA Agent Markup Language.*IEEE Intelligent Systems* 16(6): 67-73.

[Lindberg et al.] Lindberg D.A, Humphreys, B.L. McCray AT. (1993).The Unified Medical Language System. Meth. Inf Med Aug;32(4):281-91

[Mendonca et al.] Mendonca EA, Cimino JJ, Johnson SB, Seol YH. Accessing heterogeneous sources of evidence to answer clinical questions. J Biomed Inform. 2001 Apr;34(2):85-98]

[Noy et al.] Noy N. F. Sintek M, Decker S.,. Crubezy M, Fergerson R. W., & Musen M. A. Creating Semantic Web Contents with Protege-2000. IEEE Intelligent Systems 16(2):60-71, 2001. The Protégé Project. http://protege.stanford.edu

[Noy, Musen et al.] Noy N F., Musen M. A.,. Mejino J. L.V, Rosse C. Pushing the Envelope: Challenges in a Frame-Based Representation of Human Anatomy, SMI2002-092 report, 2002.

[Ono et al.] Ono, M, Kubik, S and Abernathey, Atlas of the Cerebral Sulci, Thieme Medical Publishers, Inc,1990

[OIL] Dieter F., van Harmelen F., Horrocks I., McGuinness D.L., and Patel-Schneider P. F. OIL: An ontology infrastructure for the semantic web. IEEE Intelligent Systems, 16(2):38-45, 2001.

[OWL] OWL Web Ontology Language 1.0 Reference. Mike Dean, Dan Connolly, Frank van Harmelen, James Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. W3C Working Draft 12 November 2002. Latest version is available at http://www.w3.org/TR/owl-ref/.

[RDF] RDF/XML Syntax Specification (Revised) Dave Beckett, ed. W3C Working Draft 23 January 2003. Latest version is available at http://www.w3.org/TR/rdf-syntax-grammar/.

[Rosse et al.] Rosse C, Mejino JL, Modayur BR, Jakobovits R, Hinshaw KP, Brinkley JF. Motivation and organizational principles for anatomical knowledge representation: the digital anatomist symbolic knowledge base.J Am Med Inform Assoc. 1998 5(1):17-40.

[Rousset et al.] Rousset M-C, Bidault A, Froidevaux C, Gagliardi H, Goasdoué F, Reynaud C, Safar B. Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL, Revue I3 : Information - Interaction - Intelligence, 2002.

[Sakai] Sakai Y. Metadata for evidence based medicine resources, Proc. Int'l Conf. On Dublin Core and Metadata Applications 2001, 81-85, 2001.

[Toga] Toga A.W. Neuroimage databases : the good, the bad and the ugly, Nature reviews neuroscience vol 3 302-309, 2002.

[XML] Extensible Markup Language (XML) 1.0 (Second Edition). Tim Bray, Jean Paoli, C. M. Sperberg-McQueen, and Eve Maler, eds. W3C Recommendation 6 October 2000. http://www.w3.org/TR/REC-xml.

[XML Schema] Paul V. Biron and Ashok Malhotra, eds. XML Schema Part 2: Datatypes. W3C Recommendation 02 May 2000. Latest version is available at http://www.w3.org/TR/xmlschema-2/.

Partage des modèles XML : une solution pour les échanges électroniques professionnels

A. MKADMI, N. BOUHAÏ
Laboratoire PARAGRAPHE, Université Paris8,
2, rue de la Liberté
93526 - SAINT-DENIS cedex 02,
Mail : mkadmi@hymedia.univ-paris8.fr ; bouhai.nacer@free.fr
Tél : +33 1 49 40 67 58 Fax : +33 1 49 40 67 83

M. LANGLOIS
EDIFRANCE
13, rue Camille Desmoulins
92441 Issy-les-Moulineaux Cedex
Mail : langloismarc@compuserve.com
Tél : +33 1 58 04 25 32 Fax : +33 1 58 04 23 00

Résumé

Nous présentons dans cet article le projet de création d' un répertoire de modèles XML que nous avons réalisé dans le cadre d'une coopération entre le laboratoire Paragraphe, Edifrance, Mutu-XML, GFII, FING et UIC. Nous présentons également le contexte et l' intérêt général du projet. Une description de l' application avec ses différentes fonctionnalités concernant l' accès, la recherche, le partage et la révision des modèles sera présentée. Enfin, nous présenterons les résultats et perspectives de ce projet.

Abstract

We present in this article the project of creation of repository of XML models which we carried out within the framework of co-operation between laboratory Paragraphe, Edifrance, Mutu-XML, GFII, FING and UIC. We present also general context and interest of the projet. A description of the application with its various functionalities relating to access, research, division and revision of the models will be presented. Lastly, we will present the results and perspective of this project.

1 Introduction

Comme toute technologie émergente, il aura fallu attendre plusieurs années avant que le format XML (eXtensible Markup Language) ne s' impose réellement comme étant le format privilégié de l'échange de documents - et plus généralement d'informations - en milieu ouvert. En dissociant le contenu de la publication, XML est apparu en réponse au besoin d'interaction et de coopération entre des systèmes d'information hétérogènes utilisant jusqu'alors des structures de données largement incompatibles entre elles. En effet, les travaux de normalisation menés par le W3C, les développements des éditeurs de logiciels et les préconisations de différents groupes et consortiums (ebXML, OASIS) conjuguent leurs efforts pour définir, promouvoir et utiliser XML dans différentes situations. Ce méta-langage est utilisé aujourd' hui par tous: les fournisseurs d' ERP, les éditeurs de middleware, les fournisseurs de bases de données, etc. Les raisons de ce consensus sont à chercher du côté de la simplicité et de la richesse d' expression d' XML.

Cependant, pour que les objectifs d' XML, en l' occurrence permettre l' échange généralisé intersectoriel quel que soit le type d' acteur, soient réellement atteints, il paraît très intéressant que les modèles de ces documents structurés échangés, ainsi que toutes les informations associées soient partagés. Ces informations associées servent à mieux comprendre les modèles pour pouvoir les exploiter de façon rapide et efficace. Il devient alors possible de recevoir n'importe quel document issu d'un modèle particulier afin d'être en mesure de l'exploiter avec des logiciels génériques. On pourra, par exemple, éditer un document avec des environnements standard et disponibles, mais aussi le visualiser sur un système de consultation standard du Web.

2 Problématique

Associant aux données une structure sémantique (sous forme d' éléments et attributs) et permettant de séparer cette structure du contenu, ainsi que la présentation de ce contenu, XML a été retenu comme le langage d' avenir pour la génération des échanges électroniques, que ce soit entre les grandes entreprises, entre les PME ou entre les grandes entreprises et les PME. Comparé à l' EDI conventionnel, XML offre la capacité de:

- afficher les données sous une forme humaine par l' utilisation des feuilles de style…;
- convertir facilement une structure de message en une autre structure de message (ce qui facilite l' intégration des données dans des applications existantes..) [GENCOD EAN, 2002]

Cependant, pour permettre un traitement automatique de documents XML provenant d' autres partenaires, il est nécessaire que les différents acteurs impliqués se mettent d' accord sur un formalisme de structuration des informations. Cette structuration est définie selon un modèle sous formats de schémas ou de DTD qui donne les règles d' assemblage et d' ordonnancement des données.

De ce fait, il devient alors intéressant de pouvoir partager ces modèles. C' est dans ce contexte que notre projet de création d' un répertoire de modèles XML a vu le jour pour pouvoir identifier, partager et réutiliser les modèles de documents pour les différentes applications XML.

3 Partage des modèles de documents : état de l' art

Le partage et la réutilisation des modèles de documents étaient depuis longtemps le souci des professionnels, et si le déploiement de l'Internet décuple les possibilités de partage

d'information, les syntaxes utilisées jusqu' alors (HTML notamment) limitent les types de réutilisation possibles. Le développement de XML rend possible le partage de données structurées. Ceci correspond à une mutation profonde de la façon dont les applications informatiques vont partager des référentiels sémantiques, avec des conséquences importantes en terme d'infrastructures et d'outils disponibles. Nous présentons en ce qui suit quelques exemples de modèles de documents partagés utilisés dans le domaine public.

3-1 Répertoire de l'OASIS[1]

Le répertoire de l'OASIS (OASIS registry/repository) est un système ayant, comme son nom anglais l' indique, deux composants, un « registre » et un « dépôt ». On parle donc des objets enregistrés et des entrées de registre. Les objets enregistrés représentent tout ce qu' un auteur mette à la disposition du public pour être utilisé par un client. Ils sont sauvegardés dans le « dépôt ». Quant aux entrées de registre, elles représentent les métadonnées (informations décrivant les objets enregistrés), elles sont stockées dans le registre et servent de référence pour accéder aux objets.

Ce répertoire ouvert pour XML et SGML DTDs et schémas représente un pas pour déployer des répertoires XML interopérables sur Internet. Il offre un forum indépendant des éditeurs pour les programmeurs et organismes de standardisation pour soumettre publiquement, publier et échanger des spécifications et vocabulaires XML. Il est conçu pour servir de modèle pour un réseau extensible de registres et répertoires XML distribués sur Internet. Les spécifications du registre ebXML de l'OASIS sont disponibles à http://www.oasis-open.org/committees/regrep/documents/2.0/specs.

Cependant ce répertoire, malgré l' intérêt qu' il présente dans l' orientation vers une plus grande utilisation du langage XML, son utilisation n' est pas à la portée de tout le monde (qui n' est pas forcément spécialiste en XML). De plus son orientation anglophone peut être une contrainte pour d' autres acteurs (notamment francophones) au niveau de son utilisation.

3-2 Répertoire de l'ATICA[2]

Dans le domaine de l' administration, il existe aussi un projet similaire mené par l' ATICA visant à publier tous les schémas et DTD issus des domaines documentaire et juridique présentant un intérêt général dans un répertoire. Ce répertoire représente un outil mutualisé pour favoriser les échanges au sein des administrations et avec leurs partenaires.

Ce projet a été déclenché par une Circulaire du 21 janvier 2002 relative à la mise en œuvre d'un cadre commun d'interopérabilité pour les échanges et la compatibilité des systèmes d'information des administrations.

> *« Enfin, il sera bon que chaque nouveau projet de système comportant des échanges d'informations (au sein de l'administration ou avec les tiers) soit l'occasion de poursuivre, et même d'intensifier, l'élaboration de schémas XML, dont on connaît l'importance pour faciliter les échanges. Ils seront conçus de manière à faire clairement apparaître leur définition, ainsi que celle des éléments qui les composent, par application de la méthode dite des « espaces nominatifs », conforme aux standards de l'Internet. Ils seront publiés, d'abord à l'état de projet, puis sous leur forme définitive, dans le répertoire des schémas XML de l'administration. »* [Jospin, 2002].

[1] OASIS Registry/Repository : technical Specification, Working Draft 1.1 december 20, 2000/ OASIS-Organization for the Advancement of Structured Information Systems.
[2] ATICA : Agence pour les Technologies de l'Information et de la Communication dans l'Administration

Cependant, ce projet, malgré l'intérêt qu'il présente quant à la dématérialisation et l'interopérabilité des échanges au sein de l'administration et entre l'administration et ses partenaires, il n'est pas précisé jusqu'à aujourd'hui comment les modèles seront présentés au sein du répertoire, ni quelles sont les documentations qui doivent accompagner ces modèles. Ces lacunes n'ont pas encouragé les administrations à se lancer dans ce projet pour enrichir le répertoire par leurs modèles.

Ces deux exemples de projets (qui ne sont pas exclusifs) montrent bien l'intérêt de la création d'un répertoire de schémas XML. En effet, Le fait de partager des modèles participe à l'évidence à l'acceptation de ceux-ci et représente, de plus, un facteur de montée en compétences des organisations confrontées à l'explosion des applications XML. Plus les modèles sont accessibles de manière facile, plus l'échange entre partenaires devient aisé.

4 Intérêt des modèles de documents

Un modèle de document est une structure permettant de donner les règles d'assemblage des données. Cette structure permet à un système d'information de comprendre que <Code_Postal>96100</Code_Postal> représente bien un code postal d'une ville et non pas une quelconque suite de chiffres. Ce système peut aussi comprendre, d'un point de vuestructurelle, que ce code fait partie d'une adresse d'un client. À défaut de tel modèle, ce numéro peut être codé de différentes manières dans des différents documents, exemple : <codepostal>, <CodePostal>, …, et par conséquent, il peut être interprété comme une suite de chiffres.

Un modèle de documents sert donc dans un premier temps à définir tous les éléments utilisés dans un document, et deuxièmement pour définir les relations et l'ordonnancement entre ces éléments.

Pour créer des modèles de documents, deux recommandations existent aujourd'hui. Les DTDs et les schémas XML. La première sur les DTDs est historique et est devenue reconnue dans tous les domaines applicatifs. La deuxième est récente et vient palier aux déficiences de la première se rapportant notamment au typage de données, au langage utilisé et au support des espaces de noms. En effet, XML Schema est un nouveau langage proposé par le W3C qui propose, en plus des fonctionnalités fournies par les DTD, plusieurs nouveautés à savoir :

- un grand nombre de types de données intégrées comme les booléens, les entiers, les intervalles de temps, etc. De plus, il est possible de créer de nouveaux types par ajout de contraintes sur un type existant ;
- des types de données utilisateurs qui nous permettent de créer notre propre type de données nommé ;
- la notion d'héritage : Les éléments peuvent hériter du contenu et des attributs d'un autre élément. C'est sans aucun doute l'innovation la plus intéressante de XML Schema ;
- le support des espaces de nom ;
- les indicateurs d'occurrences des éléments peuvent être tout nombre non négatif ;
- une grande facilité de conception modulaire de schémas.

Les modèles de documents servent donc à définir la cohérence d'un ensemble de documents, lesquels peuvent être utilisés par n'importe quelle application informatique en ne se définissant que par rapport au modèle sous-tendu. Ceci permet évidemment de gagner beaucoup de temps, d'argent et de fiabilité dans les travaux coopératifs.

5 Présentation du répertoire de modèles XML

Le répertoire de modèles XML est une base de données permettant à tout utilisateur de prendre connaissance des modèles existants dans un domaine d' activité particulier pour un besoin particulier, ainsi que des modèles permettant d' échanger avec ses partenaires. Ce répertoire favorise l' échange ouvert entre professionnels qui, avec cet outil, seront capables de se définir par rapport à l' existant et d' avoir une cohérence des méthodes de travail dans leurs domaines d' activités.

Le répertoire de modèles XML doit alors proposer des accès à un ensemble de modèles bien documentés donnant une idée sur les initiatives normatives prises dans un domaine d' activité bien déterminé. Il doit représenter aussi un espace de travail collaboratif dans le cadre des échanges électroniques professionnels.

5-1 Principes de base

Le répertoire de modèles est accessible librement à tout le monde et il n' y a aucune limite dans la consultation des schémas et des DTDs qui y sont stockés. Quant à la sécurité, il n' y a pour le moment aucune notion de confidentialité au niveau accès des données. La recherche des modèles, comme nous allons la présenter par la suite, se fait à travers plusieurs critères définissant le contexte et l' appartenance de chaque modèle. Quant à la modification et la mise à jour, elles se font par le propriétaire des modèles, seuls responsables des structures de leurs documents.

Cependant, pour la publication de ces modèles, elle se passe par un « comité éditorial » qui vérifie la forme de ce qui est proposé à publication, c' est à dire la cohérence des données contextuelles, la pertinence de la définition par rapport au contexte, la bonne syntaxe des schémas et des DTDs, ainsi que la cohérence des documents d' exemple au regard des modèles.

Pour cela, trois acteurs au moins se distinguent : l' utilisateur, le participant au groupe de travail et l' administrateur. L' utilisateur devra avoir accès aux modèles par le biais d' une interface Web à partir de n' importe quel navigateur. Un système de session personnel est mis en place. Cet utilisateur pourra être n' importe qui, mais il pourra être aussi un soumissionnaire d' un modèle dans la base. Le participant au groupe de travail est une personne qui se dote d' un mot de passe pour accéder à un espace de travail bien particulier. L' administrateur devra pouvoir gérer cette base de données (ajouter, modifier ou supprimer un champ), mais il devra aussi pouvoir modifier les différents modules du serveur.

Parmi les contraintes définies, on peut noter :

- aucun utilisateur et/ou propriétaire du modèle ne peut modifier un modèle qui ne lui appartient pas ;
- les propriétaires des modèles doivent être munis des mots de passe pour accéder à leurs espaces de travail ;
- les modèles déclarés doivent subir une opération de révision et de contrôle de la part d' un comité spécial pour assurer la cohérence des données;
- l' application doit être ouverte et évolutive: facile à mettre à jour et à y ajouter d' autres fonctionnalités répondant à des nouveaux besoins ;
- ni le logiciel serveur, ni le logiciel client ne doit être d' une technologie propriétaire.

5-2 Structure des données

Avant de concevoir la structure de l' application, nous avons essayé de définir la structure des données qui permet de préciser tous les éléments, les entités, les attributs, les relations entre

eux, ainsi que les différentes caractéristiques de ceux-ci. Cependant, vu la nature de notre projet qui se veut à la fois générique touchant le maximum possible de domaines et à la fois spécifique à l' échange électronique des données d' affaires (commerce électronique), nous avons opté pour deux structures de données. La première répondant au premier objectif, c' est à dire une structure générale permettant à n' importe qui de l' utiliser pour déclarer son modèle de données, et la deuxième structure est spécifique à la déclaration des « core components[3] ebXML» (composants élémentaires ebXML) [UN/CEFACT, 2001]. La définition de ces deux structures s' est fait en utilisant les schémas XML[Langlois, 2002].

La première structure se présente brièvement comme suit :

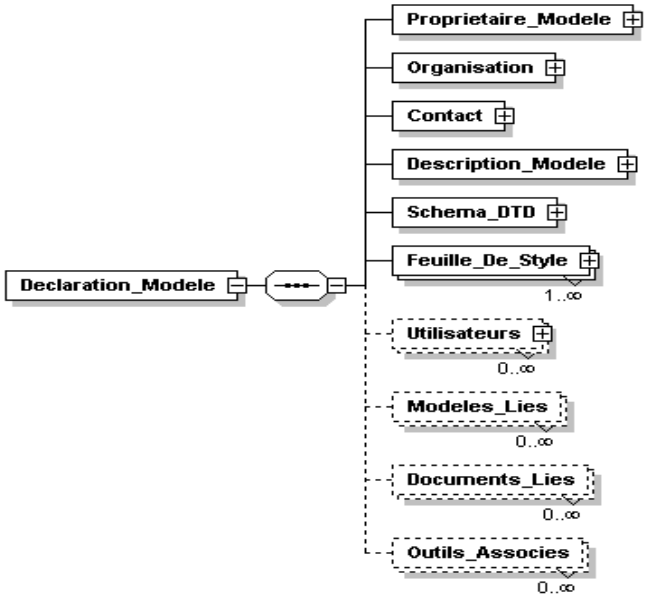


Figure 1 : Structure de données simple et générique

Quant à la deuxième structure, elle contient toutes les données de la première structure, mais elle englobe d' autres données liées à la définition d' un « core component ». Ces données sont extraites de la spécification ebXML [OASIS, 2002].

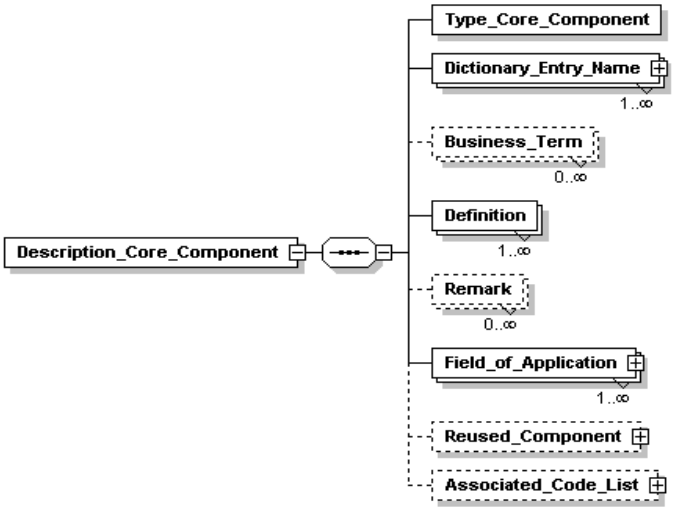


Figure 2 : Structure de données liée aux « core components ebXML »

[3] Traduit en français *composants élémentaires*, les « core components » sont des pièces d' assemblage, ayant chacune une définition unique en sémantique d' affaire. Pour plus d' informations, veuillez voir : http://www.autoroute.gouv.qc.ca/publica/normes/norme111.htm

5-3 Structure de l'application

L' application est structurée en six parties:

- une partie contenant les éléments de définition du modèle ;
- une partie définissant toutes les Informations associées au modèle ;
- un module définissant tous les critères de recherche ;
- un module pour gérer le contrôle et la révision des modèles proposés à publication ;
- un module pour la consultation des données du répertoire ;
- et un module pour la soumission des modèles et documents XML.

Cette structure a été définie, après une étude de besoins qui a touché les partenaires principaux de ce projet à savoir Edifrance (Association pour le développement des échanges électroniques professionnels), MUTU-XML et GFII (Groupement Français de l' Industrie de l' Information), FING (Fondation Internet Nouvelle Génération), ainsi que l' UIC (Union Internationale de Chemins de Fer) à Bruxelles.

Cette structure peut être schématisée comme suit :

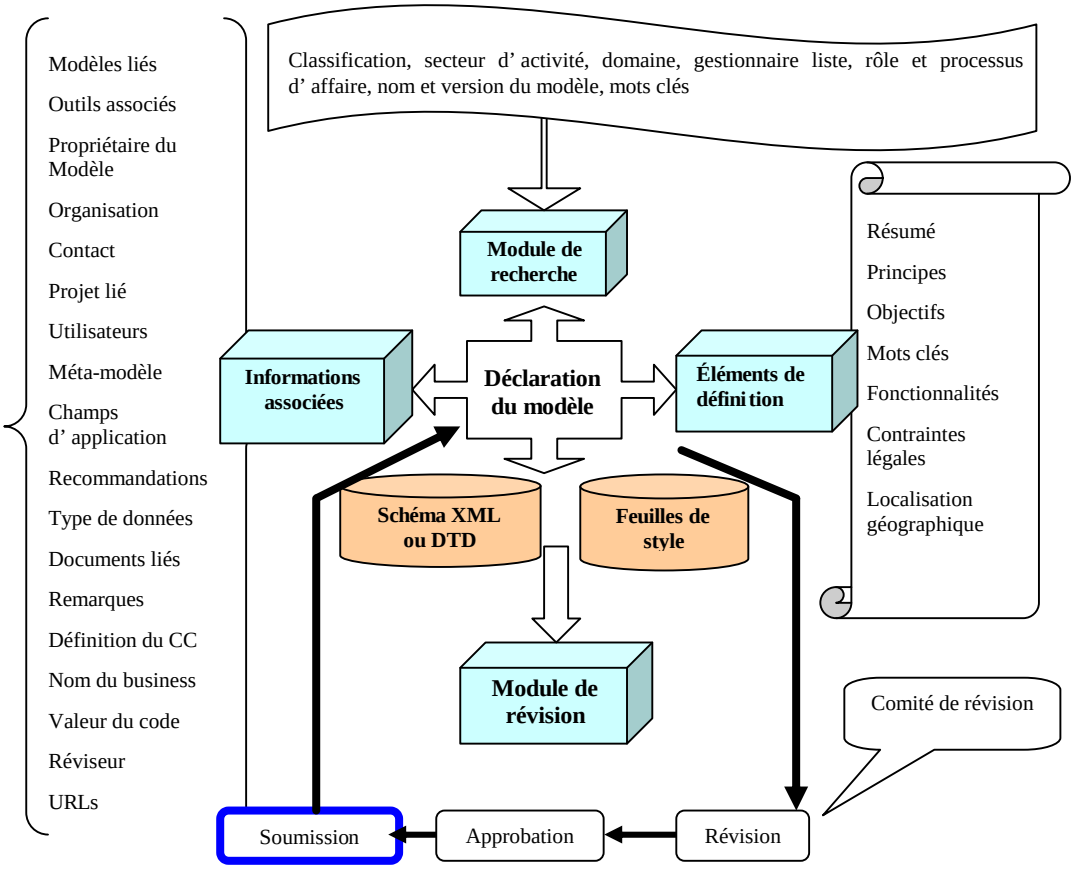


Figure 3 : Structure de la base de donnée du répertoire de modèles XML

Le module de recherche peut utiliser plusieurs critères : secteur d'activité, domaine d' usage, rôle dans le processus d' affaire, classification du produit, processus d' affaire, nom du modèle et mots clés. Ces critères peuvent être utilisés conjointement avec les opérateurs logiques "et" et "ou". Le module de recherche est accessible par consultation à distance. Les trois premiers critères constituent une liste fermée maintenue par le comité éditorial. Les autres sont ouverts.

Le module de consultation donne accès, à partir d'une recherche, aux données du répertoire. Il peut, soit permettre une consultation sur navigateur, soit permettre un téléchargement du résultat de recherche.

Quant au module de révision, il représente une antichambre pour assurer la meilleure diffusion des modèles XML et, donc, de favoriser la meilleure compréhension possible de l'objectif et des conditions d'utilisation d'un schéma particulier.

Le module de soumission d'une nouvelle entrée utilise le même modèle que celui qui sert à la restitution, à partir du module de recherche. La soumission est stockée en attente d'approbation. L'approbation permet de mettre à jour le répertoire.

En ce qui concerne la partie se rapportant aux éléments de définition du modèle, elle permet de définir le modèle à travers le résumé, les objectifs, les principes de base, les différentes fonctionnalités et les mots clés. Elle permet de définir aussi les caractéristiques des situations d' affaires influencées par des exigences légales ou réglementaires (lois, règlements, conventions, traités, …).

L' autre partie permet, quant à elle, de présenter toutes les informations associées au modèle à savoir son propriétaire, l' organisation responsable, les modèles et documents liés, les outils utilisés, les principaux utilisateurs, etc.

5-4 Modes d'utilisation

L' utilisation de l' application « répertoire de modèles XML » diffère selon qu' on est soumissionnaire de modèle, utilisateur ou réviseur.

Pour un soumissionnaire, il doit passer par une identification (nom, prénom, Login et mot de passe, etc.) s'il est nouveau, si non, il n' entre que son login et son mot de passe pour accéder à une interface qui lui permet de choisir le type de modèle (modèle simple ou modèle des « core components ») qu' il souhaite enregistrer, ainsi que le mode de soumission (manuelle ou par exportation[4]), et il reçoit ainsi un formulaire qui lui permet de remplir toutes les données y compris le schéma XML. Ce système d' identification nous permet d' une part d' éviter les redondances des entrées au niveau propriétaires de modèles et d' autre part de contrôler et d' assurer que la modification d' un modèle ne se fait que par son propriétaire.

Pour un utilisateur du grand public qui veut consulter un modèle ou tous les modèles existants dans la base, il lance sa requête à partir de l' interface de recherche et reçoit une liste de modèles, avec seulement le nom du modèle et un lien vers son contenu global.

Quant au réviseur, il passe lui aussi par un mot de passe pour accéder aux modèles soumis, et après vérification de leurs contenus, il note ses commentaires et remarques sur chaque modèle pour son propriétaire si le modèle manque d' informations ou s' il ne répond pas aux exigences prédéfinies. Si non, il valide le modèle pour qu' il devienne accessible au grand public.

La première page de l' application nous permet donc de consulter la liste des modèles, de chercher un modèle par plusieurs critères ou de soumettre et/ou modifier un modèle en passant par une identification. Cette dernière utilisation (soumission et/modification) donne à l' utilsateur accès à une autre page sommaire qui lui propose quatre fonctionnalités (figure 4).

[4] Lors de la soumission, on peut, au lieu de remplir toutes les données manuellement, les exporter à partir d' un fichier Excel, et ceci représente un besoin déclaré par les partenaires à l' état actuel qui travaillent beaucoup avec des tableurs Excel pour stocker leurs « core components ». Lors de la consultation et téléchargement, on peut faire l' opération inverse (importer les données de la base dans un fichier Excel).

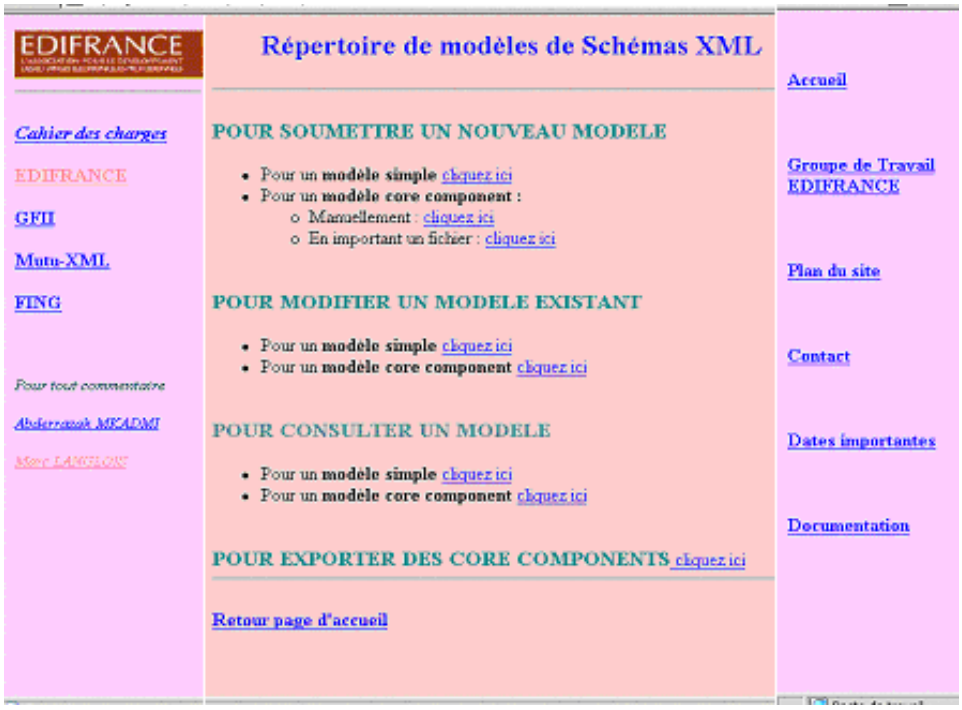


Figure 4 : Page sommaire du répertoire de modèles XML

Comme le présente cette figure qui représente une des pages d' accès au répertoire de modèles XML, les principales fonctionnalités sont : la soumission, la consultation, la modification et l' exportation des modèles.

6 Résultats et perspectives

Le répertoire a été testé et est utilisé actuellement en local par les différents partenaires du projet notamment par Edifrance pour la publication des « core components ». Il sera bientôt hébergé et publié par l' UIC (Union Internationale des Chemins de fer) quand il aura évolué son serveur pour supporter les fonctionnalités des outils utilisés dans le développement de cette application. Ceci permettra aux différents partenaires du projet de publier leurs modèles et d' en profiter de l' existence d' autres modèles liés à différentes activités. Cette application permet également d'éviter certainement des travaux redondants d' une part, et encouragera d' autres à utiliser la technologie XML, tout en s' inspirant des modèles déjà publiés. Il offre, par ailleurs, un service d' aide à l' élaboration des schémas XML.L' évolution prévue de cette application consiste en premier temps à ajouter une fonctionnalité d' exporter les modèles de la base de données en format XML et de proposer après l' outil à différents acteurs utilisant XML tout en l' adaptant à leurs besoins.

7 Conclusion

Nous avons présenté dans cet article le répertoire de modèles de schémas XML. Le contexte de création et les structures de données et de l' application ont également été présentés, ainsi que quelques intérêts de cet outil. Cependant, d' autres intérêts peuvent être évoqués à travers cette application, à savoir le travail collaboratif, la recherche sémantique…, qui s' intègrent dans une problématique très large dépassant le cadre de cet article. Toutefois, nous pouvons dire que le répertoire de modèles XML représente un pas exceptionnel dans le domaine de l' EDI. Un tel projet était depuis un bon moment, (en tous cas, depuis le développement et l' introduction du métalangage XML dans le domaine du commerce électronique et dans

l' administration française) le souci de tous les organismes travaillant dans le domaine de la normalisation et la standardisation.

8 Références

[Attar, 2002] ATTAR, Pierre. - Documentation Technique du modèle permettant l' échange de descriptions de modèles**.** - Version 0.1, 1er Janvier 2002

[Chauvet, 2002] CHAUVET, Jean-Marie. - Services Web avec SOAP, WSDL, UDDI, ebXML. – Paris : Eyrolles, 2002

[GENCOD EAN, 2002a] GENCOD EAN & Edifrance. - Le commerce électronique pour l'entreprise : guide de mise en œuvre des échanges électroniques professionnels. – France, janvier 2002

[GENCOD EAN, 2002b] GENCOD EAN & Edifrance.- comprendre XML pour les échanges électroniques professionnels (cd-rom). – France, mai 2002

[Jospin, 2002] JOSPIN, Lionel. - Circulaire du 21 janvier 2002 relative à la mise en œuvre d'un cadre commun d'interopérabilité pour les échanges et la compatibilité des systèmes d'information des administrations, J.O n° 30 du 5 février 2002 page 2335, NOR: PRMX0205357C, Paris, 2002. – http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=PRMX0205357C

[Langlois, 2002] LANGLOIS, Marc ; MKADMI, Abderrazak. - Création d' un répertoire de modèles XML : cahier des charges, EDIFRANCE, 2002. - http://www.edifrance.org/accueil_schemaxml.php

[OASIS, 2002] OASIS. - OASIS Registry/Repository Technical Specification, Working Draft 1.1 December 20, 2000.

[Parent, 2002] PARENT, Richard. - Normes ouvertes en technologies de l'information : Spécification technique des composants élémentaires, Partie 1. - http://www.autoroute.gouv.qc.ca/publica/normes/liste.htm

[UN/CEFACT, 2001] UN/CEFACT. – Core Components Technical Specifications, Part 1// Spécification technique des composants élémentaires, Partie 1, traduit par Richard Parent. - http://www.autoroute.gouv.qc.ca/publica/normes/norme111.htm

[Van Der List, 2001] VAN DER VLIST, Eric.- Le répertoire de schémas XML de la MTIC ouvre ses portes. - http://xmlfr.org/actualites/decid/010413-0002

Sites Web

W3C : http://www.w3.org/
XML : http://www.w3.org/XML/
Schémas XML : http://www.w3.org/XML/Schema
ebXML : http://www.ebxml.org/.
OASIS : http://www.oasis-open.org/
Mutualiser l' effort de montée en compétences sur XML: http://www.mutu-xml.org/
EDIFRANCE : http://www.edifrance.org
GFII : http://www.gfii.asso.fr/dep.htm
FING : http://www.fing.org/index.php?rubrique=lafing
XML SPY : http://www.xmlspy.com/

Journées Francophones de la Toile - JFT'2003

Services Web

Implémenter des composants actifs dans le web sémantique

N. Sabouret

LIMSI-CNRS, BP 133, 91403 Orsay Cedex, FRANCE.
Email : Nicolas.Sabouret@limsi.fr
Tél : +33 1 69 85 81 04 Fax : +33 1 69 85 80 88

Résumé

Les modèles actuels du web sémantique permettent de munir les pages d'une représentation *structurelle* sémantisée pour répondre à des requêtes sur la signification de l'information qu'ils manipulent. Au contraire, les composants actifs dans le web (services, agent, *etc*) avec lesquels les utilisateurs ordinaires vont interagir sont caractérisés par un *fonctionnement*. Dans cet article, nous montrons qu'il est possible d'utiliser XML comme langage de programmation des composanst actifs, pour unifier la description de la partie structurelle et celle du fonctionnement. Nous présentons un modèle d'activité et un langage de programmation de composants actifs qui permettent d'avoir accès en cours d'exécution à une description sémantisée de leur fonctionnement pour pouvoir répondre à des requêtes sur leurs actions.

Abstract

In current models of the semantic web, web pages are provided with a *structural* semantic representation to answer requests about the meaning of the processed information. On the contrary, users interact with active components (services, agents...) which are characterised by a *functioning*. In this paper, we show that it is both required and possible to use XML as a programming language for such active components, thus integrating the action description with the structural semantic representation. We present an execution framework and a programming language for active components that allows us to have access at runtime to a semantic description of their actions, in order to answer requests about their activity and behaviour.

1 Contexte de l'étude

1.1 Les composants actifs dans le web sémantique

Les recherches actuelles sur le web sémantique proposent de s'appuyer sur des techniques de représentation des connaissances [Charlet et al., 2000] pour munir l'information contenue dans les pages web d'une sémantique. L'émergence de standards comme DAML-OIL pour la représentation de l'information *structurelle* contenue dans les pages web permet de répondre à requêtes portant sur la signification de cette information [Reynaud et al., 2001, Andreasen et al., 2000]. Cependant, les utilisateurs du web sémantique n'accèdent pas directement à cette information. Au contraire, ils interagissent avec des composants actifs (services, agents, *etc*), c'est-à-dire des composants logiciels qui manipulent cette information en fonction de l'interaction avec l'utilisateur. Ces composants actifs sont caractérisés par un *fonctionnement* [Guessoum and Briot, 1999] et, contrairement au « script CGI » qui ne fait que traiter, manipuler ou produire l'information contenue sur le web, ce fonctionnement fait *partie intégrante* de l'information. Les utilisateurs de tels composants actifs doivent alors pouvoir poser des questions non seulement sur les données qu'ils manipulent, mais aussi sur les actions qu'ils effectuent.

Pour ce faire, le web sémantique doit permettre de représenter et de manipuler des connaissances non seulement structurelles, mais aussi de nature *fonctionnelle*, c'est-à-dire de représenter le fonctionnement des composants actifs qui y agissent. Plusieurs modèles de description de services web ont été proposés (WSDL [Christensen et al., 2001], WSMF [Fensel and Bussler, 2002] ou DAML-S [DAML-S Coalition, 2002]) afin de permettre aux agents de trouver, invoquer, surveiller ou composer en services plus complexes des ressources Web offrant tels services et vérifiant telles propriétés. Cependant, dans tous ces modèles, la description du service est séparée de l'information manipulée. Au contraire, dans un composant actif, la structure, les données manipulées et le fonctionnement doivent être intégrés au sein d'une même représentation pour que le composant puisse interagir avec l'utilisateur aussi bien à propos de la structure [Andreasen et al., 2000] de l'information que du fonctionnement [Sabouret, 2002a, Sabouret, 2002c]. Il faut donc pouvoir *programmer* directement en XML les composants actifs, afin qu'ils soient capables de raisonner sur *leur propre code*.

1.2 Le projet *InterViews*

Dans le projet *InterViews* [Sansonnet, 1999], nous étudions la problématique de l'interaction entre un utilisateur humain en situation de demande d'aide et un agent conversationnel. Nous essayons de modéliser et d'implémenter de tels agents qui seraient capables de se représenter leur fonctionnement, de raisonner dessus et d'interagir en langue naturelle avec l'utilisateur pour répondre *en cours d'exécution* à un large éventail de questions portant sur leurs actions et leur exécution. Par exemple : « *Qu'est-ce que tu fais ?* », « *Comment faire pour... ?* », « *Pourquoi tu as fait ça ?* », etc. L'étude du problème nous a amené à définir un langage de programmation *spécifique* pour les composants actifs

[Sabouret, 2002a, Sabouret, 2002c], appelé VDL pour ***V****iew* ***D****esign* ***L****anguage*, qui vérifie les propriétés suivantes :

1. Il permet d'accéder *en cours d'exécution* à la description des actions des composants, afin de produire des explications sur le fonctionnement en réponse à des questions de l'utilisateur et en s'appuyant sur la sémantique opérationnelle du langage de description.
2. Les descriptions du fonctionnement et de la partie structurelle sont *intégrées* au sein d'une même représentation.

La description d'un composant en VDL est un arbre dont les noeuds, appelés *concepts*, sont réduits à de simples chaînes de caractère (le langage VDL n'est pas structuré). L'exécution d'un composant consiste en la réécriture de l'arbre à chaque *cycle* en fonction de concepts spécifiques, comme dans les algèbres évoluantes [Gurevich, 1995] ou *Maude* [Meseguer, 2000].

Notre modèle propose une sémantique opérationnelle pour la description de composants actifs. Il est Turing-complet (c'est-à-dire qu'il est possible de modéliser n'importe quelle machine de Turing en VDL et donc, théoriquement, d'écrire n'importe quel programme en VDL) et linéaire en temps et en espace par rapport à la machine de Turing équivalente. Nous avons défini complètement la sémantique opérationnelle de ce langage dans [Sabouret, 2002c] et nous l'avons implémenté en Java 1.1.7.[1].

1.3 Plan de l'article

Nous avons montré récemment [Sabouret, 2002b] que le langage VDL pouvait constituer un fondement théorique pour intégrer, au sein de pages XML, la représentation de la structure (bien définie dans les modèles actuels du web sémantique) et celle du fonctionnement des composants actifs. Dans cet article, nous présentons plus précisément la sémantique opérationnelle du langage et nous l'illustrons sur un exemple simple.

2 Principe général

2.1 Notion d'observateur

Le modèle d'exécution que nous proposons pour les composants actifs décrits en XML s'appuie sur la notion d'*observateur*, inspirée des SGML. Un observateur est un composant logiciel externe qui recherche dans un document XML des éléments spécifiques, conformes à un schéma de description ou à une DTD, et les interprète pour effectuer des actions particulières. Un observateur est donc caractérisé par :

[1]Des démonstrations sont disponibles sur notre page web : `http://www.limsi.fr/Individu/nico/exemples`

- Une DTD, méta-DTD ou un schéma auquel le document doit se conformer pour pouvoir être interprété par cet observateur. Elle décrit l'ensemble des éléments qui sont connus de cet observateur (ceux qu'il va interpréter) et la *syntaxe* à respecter ;
- Une fonction d'interprétation φ pour les éléments définis dans ce schéma, implémentée dans un langage de programmation utilisant l'API « *DOM* » [W3C, 2002]. Elle définit les actions effectuées sur les éléments reconnus par l'observateur, c'est-à-dire la *sémantique* de ces éléments.

Du point de vue de notre étude, nous pouvons mettre en évidences deux types d'observateurs :

1. **Les observateurs statiques**, qui ne modifient pas le document de manière proactive. Par exemple, l'observateur structurel qui répond aux requêtes sémantisées portant sur les informations contenues dans le composant, l'observateur d'interface qui gère l'interface graphique utilisateur du composant (modifiée par effet de bord), *etc.*
2. **L'observateur dynamique** qui considère le document passé en paramètre comme la description du composant à l'instant t et qui construit sa description à l'instant suivant $t + 1$ en fonction des éléments spécifiques décrivant le fonctionnement du composant actif.

L'exécution d'un composant actif dans ce schéma consiste donc en la réécriture du document XML. La description du composant en XML est sérialisée à chaque cycle d'exécution et chaque observateur (l'ensemble des observateurs statiques puis à nouveau l'observateur dynamique) interprète le nouveau document. De plus, les composants sont proactifs : ils peuvent s'exécuter en dehors de toute interaction avec l'utilisateur.

Dans cet article, nous nous intéressons uniquement à l'observateur d'exécution, que nous noterons *vdl* (par référence aux travaux effectués dans le projet *InterViews*). Dans ce modèle d'exécution basé sur la réécriture d'arbre, le fonctionnement est décrit explicitement (à la fois dans l'abstraction XML et dans le document sérialisé à chaque cycle) et peut être manipulé pour répondre à des requêtes sur le fonctionnement, en utilisant la sémantique opérationnelle du langage implémentée dans l'API « *DOM* ».

2.2 Partie structurelle

Le schéma de l'observateur d'exécution *vdl*, qui en définit la syntaxe, est disponible sur notre page web[2]. Nous en présentons ici les principaux éléments.

2.2.1 Principe général

Un composant actif est décrit dans notre modèle par un arbre XML de racine *view* :

```
<xsd:element name="view">
  <xsd:complexType content="elementOnly"/>
</xsd:element>
```

[2] http://www.limsi.fr/Individu/nico/xml/vdl.xsd

Le contenu d'un élément *view* est laissé libre, afin de permettre l'inclusion d'éléments issus d'autres DTD ou schémas, mais il doit nécessairement s'agir d'un arbre (en effet, le modèle VDL ne travaille que sur des arbres).

2.2.2 Notion de valeur

Les « valeurs » des variables modifiées sont, par définition, les données contenues dans les éléments (c'est-à-dire ce qui est entre les balises), conformément aux recommandations du W3C. Au contraire, les noms de balises, les attributs et les valeurs d'attributs sont du ressort de la définition de la structure et ne peuvent pas être modifiées (de même qu'un programme ne change pas le type d'une variable en C). Les actions élémentaires de modification du document XML ne travailleront donc qu'au niveau des contenus.

Dans notre modèle, pour des raisons de simplification, nous ne prenons pas en compte le type des contenus manipulés. Pour donner des valeurs dans la description du fonctionnement, nous les englobons dans des éléments **value** :

```
<xsd:element name="value" type="xsd:anyType"/>
```

Le contenu d'un élément *value* n'est donc pas précisé, pour permettre l'utilisation soit directement de valeurs (sérialisées par un contenu de type *string*), soit d'éléments structurés définis dans un schéma ou une DTD spécifique au composant. Notre langage de programmation est donc *non typé*. Comme nous le verrons, il peut être étendu pour prendre en compte les types primitifs des schémas et même les types plus riches définis dans les schémas structurels spécifiques à chaque composant actif.

2.2.3 Remarque

Parce que les éléments *view* et *value* peuvent contenir des éléments quelconques, en particulier des éléments qui ne sont pas définis dans le schéma *vdl.xsd*, notre modèle permet d'intégrer la description du fonctionnement au sein de n'importe lequel de ces documents structurés pour en faire un composant actif. Le document initial doit être englobé dans l'élément *view* et les éléments modifiables (ceux manipulés par le composant) doivent être englobés dans les éléments *value*. Soulignons aussi que, puisque le fonctionnement est décrit dans la même représentation que la structure, il est lui-même manipulable, comme dans les langages de programmation orientée Intelligence Artificielle comme Lisp ou Prolog.

3 Partie fonctionnelle

3.1 Références

Notre objectif est de pouvoir décrire des actions qui modifient les contenus des éléments (ces contenus peuvent être eux-mêmes des éléments). Dans ce cadre, les actions doivent pouvoir faire référence aux éléments pour les modifier ou pour en extraire leur valeur. Pour

ce faire, nous proposons d'utiliser la syntaxe *XPath* [W3C, 1999] définie dans la norme XML. Les éléments permettant de faire des références dans notre langage sont **path** et **get**. L'élément *path* permet de faire référence directement à un élément du composant pour le modifier. Il sera utilisé dans les *actions élémentaires* (section 3.3). Au contraire, l'élément *get* permet de faire référence à un élément pour extraire sa « valeur ». Il est utilisé, par exemple, dans les *opérations arithmétiques* (section 3.2).

3.2 Opérateurs arithmétiques

Dans le langage VDL, les éléments suivants sont utilisés pour les opérations arithmétiques de base : **plus**, **times**, **minus** et **inverse** pour les réels ; **equals** et **"greater than"** pour les relations ; **and**, **or**, **not**, **true** et **false** pour les booléens. Dans notre schéma *vdl.xsd*, l'ensemble de ces opérations arithmétiques, l'élément de calcul de valeur d'un terme *get* et l'élément de valeur terminale quelconque *value* sont regroupées dans le groupe *operation*.

Nous avons défini dans [Sabouret, 2002c] une interprétation canonique $\zeta : \Upsilon \longrightarrow \Upsilon$ sur les termes. L'interprétation canonique d'une valeur (*value*) est l'interprétation canonique de son contenu (par exemple, 2 est interprété comme le nombre entier deux) ; l'interprétation canonique des éléments représentant des opérateurs arithmétiques est l'opération représentée ; l'interprétation canonique d'un élément *get* est le contenu (non interprété) de l'élément référencé dans le *get*. Le calcul du résultat de la référence se fait suivant l'algorithme donné dans les recommandations du W3C pour *XPath* [W3C, 1999].

Ainsi, l'opération $1 + \frac{1}{2}$ est représentée dans notre langage par l'élément :

```
<plus>
  <value>1</value>
  <inverse><value>2</value></inverse>
</plus>
```

Cet élément sera évalué par ζ en :

```
<value>1.5</value>
```

A titre d'exemple, l'algorithme utilisé pour l'élément *plus* dans la fonction récursive ζ est le suivant :

> Soit $\mathbb{R}_{vdl}$ l'ensemble des nombres dans leur représentation usuelle sous la forme d'une chaîne de caractères (par exemple : -2.63).
>
> Soit $e_1, ...e_n$ les sous-éléments de l'élément *plus*.
>
> $\forall i \in [1, n]$, soit $\widetilde{e_i} = \zeta(e_i)$ l'interprétation canonique de e_i.
>
> Si $\forall i \in [1, n]$, $\widetilde{e_i} \in \mathbb{R}_{vdl}$ alors l'interprétation canonique de l'élément *plus* est égale à $\sum_{i=1}^{n} \widetilde{e_i}$, représenté dans la notation usuelle $\mathbb{R}_{vdl}$.

3.3 Actions

3.3.1 Syntaxe

Nous avons montré dans [Sabouret, 2002c] qu'il existe trois actions élémentaires de modification d'un arbre formant une base (chaque action élémentaire n'est pas exprimable en fonction des deux autres), correspondant chacune à une balise XML dans notre schéma : **add** pour ajouter des fils à un élément, **put** pour remplacer tout le contenu (tous les fils) d'un élément, **del** pour supprimer un élément et tous ses fils. Les actions élémentaires doivent nécessairement contenir au moins un élément *path*. De plus, les éléments *add* et *put* doivent contenir au moins un élément du groupe *operation* donnant les attributs à ajouter.

Chaque action élémentaire peut être munie de deux sortes de préconditions :

- Des gardes booléennes, définies dans un élément **guard** qui contient nécessairement des éléments du groupe *operation* pouvant être évalués à *true* ou $false$;
- Des événements externes, utilisés pour modéliser l'interaction avec un utilisateur humain ou avec d'autres composants, définis dans un élément **event**. Un événement externe est un élément XML qui peut être envoyé au composant au cours de son exécution. L'élément *event* permet au programmeur de définir dans le composant quels événements doivent être traités et comment ils sont utilisés.

Les actions sont les éléments contenant au moins une action élémentaire, une précondition ou une sous-action. Ils sont matérialisés par l'élément **action**.

3.3.2 Exemple

```
<action><name>action1</name>
  <guard><get>//variable[@name="test"]</get></guard>
  <action><name>action2</name>
    <event><ok/></event>
    <del><path>//un-terme</path></del>
  </action>
  <put>
    <path>//variable[@name="compteur"]</path>
    <plus><get>//variable[@name="compteur"]</get><value>1</value></plus>
  </put>
</action>
```

L'action *action*1 est effectuée dès lors que la variable « test » est évaluée à *true*. Elle incrémente la valeur de la variable « compteur » (action élémentaire *put*). De plus, si l'utilisateur a envoyé un événement externe « ok », la sous-action *action*2 supprime l'élément « un-terme » du document.

3.3.3 Valeur de vérité des préconditions

L'ensemble des *préconditions* d'une action élémentaire a est l'ensemble des éléments de balise *guard* ou *event* dont l'élément parent de type *action* est aussi un parent de l'action a. Autrement dit, les préconditions d'une action élémentaire sont ses frères, oncles, grands-oncles, etc. de type *guard* ou *event*.

Pour les événements externes. Soit e un élément, $e_1, ..., e_n$ ses sous-éléments et $i_0 \in [1, n]$ tel que e_{i_0} est un élément *event*. Soit $e_{n+1}, ..., e_m$ les sous-éléments de e_{i_0}. Alors les éléments $(e_i)_{i\in[1,n],i\neq i_0}$ ne sont interprétés que si la liste des événements pour le cycle courant de l'exécution du composant contient l'un des événements $e_{n+1}, ..., e_m$.

Pour les gardes booléennes. Soit e un élément, $e_1, ..., e_n$ ses sous-éléments et $i_0 \in [1, n]$ tel que e_{i_0} est un élément *guard*. Soit $e_{n+1}, ..., e_m$ les sous-éléments de e_{i_0}. Alors les éléments $(e_i)_{i\in[1,n],i\neq i_0}$ ne sont interprétés que si $\forall i \in [n+1, m], \zeta(e_i) = true$.

3.3.4 Exécution des actions

L'exécution des composants s'effectue en deux temps : la fonction φ de l'observateur d'exécution parcourt le document pour y rechercher les actions élémentaires dont elle calcule la valeur de vérité. L'algorithme est une recherche en profondeur dans laquelle les branches correspondant à des préconditions non vérifiées sont ignorées. Le résultat est un ensemble d'actions de modification du document pour ce cycle d'exécution. Plusieurs actions élémentaires peuvent alors entrer en concurrence pour la modification d'un même élément. Notre mécanisme de gestion de la concurrence utilise un algorithme déterministe pour élire les actions devant être effectuées. Il s'inspire du *non déterminisme local* des algèbres évoluantes [Gurevich, 1995].

Pour chaque action élémentaire élue, la fonction φ calcule l'ensemble P des éléments qu'elle modifie (ce sont ceux référencés dans un élément *path* de l'action) et l'ensemble Q des interprétations canoniques des autres sous-éléments de l'action. Par exemple, pour l'action *put* de l'exemple donné en section 3.3.2, P contient la variable « compteur » et Q contient un nombre, égal à « compteur+1 ».

Si l'action devant être effectuée est un élément *add*, la fonction φ ajoute alors dans les éléments de P une copie des éléments de Q. Pour une action *put*, il supprime d'abord tous les sous-éléments des éléments de P et pour une action *del*, il supprime simplement les éléments de P.

4 Conclusion et discussion

Le web sémantique aujourd'hui s'est beaucoup orienté sur la description de l'information structurelle sémantisée et prend peu en compte l'étude du fonctionnement des services et, plus généralement, des composants actifs interagissant avec les utilisateurs.

Ces composants sont aujourd'hui généralement « **enchassés** » dans les documents sous la forme d'applets, et non **intégrés** au contenu structurel du document. C'est la raison pour laquelle nous proposons d'utiliser le formalisme VDL [Sabouret, 2002c] comme fondement théorique pour l'intégration de la structure et du fonctionnement des composants actifs au sein de pages XML.

Les composants actifs ainsi décrits ont accès *en cours d'exécution* à une description de leur fonctionnement et de leur structure. Il leur est alors possible de raisonner dessus, par exemple pour répondre à des questions d'un utilisateur en situation de demande d'aide, à travers un observateur de traitement de requêtes sur les actions. Nous avons développé ces travaux dans [Sabouret, 2002a].

Dans notre étude, nous nous sommes intéressés tout particulièrement à la représentation du fonctionnement dans un arbre de réécriture. Cela nous a conduit à proposer un langage de programmation spécifique qui permet d'attacher, sur les aspects structurels riches, des aspects liés aux fonctionnement, afin de décrire et d'exécuter les composants actifs sous la forme de pages web en XML. Le modèle proposé reste relativement simple et devra être étendu par la suite mais, en l'état, il est suffisamment générique pour permettre de représenter une large classe de composants actifs. Plusieurs exemples ont été implémentés en VDL et sont visibles sur notre page web[3].

Notre langage n'utilise pas les opérateurs usuels des langages de programmation, comme if, $while$, *etc*, issus du modèle *implémentatoire* classique « instruction-valeur ». Au contraire, comme les modèles de description et de composition de services web (WSFL[4], WSMF [Fensel and Bussler, 2002], DAML-S [DAML-S Coalition, 2002]...), il se situe *au niveau conceptuel*, dans lequel la sémantique des opérations effectuées est matérialisée par leurs *préconditions*. Il s'agit certes d'une conception différente de la programmation d'un composant actif, cela n'en fait pas moins un véritable langage de programmation. Nous avons montré dans [Sabouret, 2002c] que le langage VDL (et donc notre modèle procédural en XML) est aussi expressif que CSP [Hoare, 1986], c'est-à-dire qu'il permet de représenter simplement tous les opérateurs de combinaison d'instructions rencontrés dans les modèles de processus communicants classiques. En particulier, notre modèle de description de composant permet de décrire toutes les constructions de process proposées par DAML-S ($Sequence$, $Concurrent$, $Split$, $Split + Join$, $Unordered$, $Choice$, $If - Then - Else$, $Repeat - Until$ et $Repeat - While$). La différence fondamentale entre VDL et DAML-S est que VDL permet d'intégrer la description du fonctionnement au sein même d'un document structuré existant, et non pas comme une couche englobante qui verrait l'exécution comme une « boite noire » spécifiée *a priori* par un expert.

Références

[Andreasen et al., 2000] Andreasen, T., Nilsson, J., and Thomsen, H. (2000). Ontology-Based Querying. In *Proc. 4th Intl. Conf. on Flexible Query Answering Systems*

[3] `http://www.limsi.fr/Individu/nico/exemples`
[4] `http://www.ebpml.org/wsfl.htm`

(FQAS'2000), pages 15–26.

[Charlet et al., 2000] Charlet, J., Zacklad, M., Kassel, G., and Bourigault, D., editors (2000). *Ingénierie des connaissances : Évolutions et récents défis*. Eyrolles.

[Christensen et al., 2001] Christensen, E., Curbera, F., Meredith, G., and Weerawarana, S. (2001). Web services description language (wsdl). http ://www.w3.org/TR/wsdl.

[DAML-S Coalition, 2002] DAML-S Coalition (2002). *DAML-S : Web Service Description for the Semantic Web*.

[Fensel and Bussler, 2002] Fensel, D. and Bussler, C. (2002). The Web Services Modeling Framework WSMF. In *1st meeting of the "Semantic Web enabled Web Services workgroup"*.

[Guessoum and Briot, 1999] Guessoum, Z. and Briot, J. (1999). From Active Objects to Autonomous Agents. *IEEE Concurrency*, 7(3) :68–76.

[Gurevich, 1995] Gurevich, Y. (1995). Evolving Algebra Lipari Guide. In Börger, E., editor, *Specification and Validation Methods*, pages 9–36. Oxford University Press.

[Hoare, 1986] Hoare, C. (1986). *Communicating Sequential Processes*. Prentice Hall.

[Meseguer, 2000] Meseguer, J. (2000). Rewriting Logic and Maude : Concepts and Applications. In *Proc. RTA 2000*, pages 1–26.

[Reynaud et al., 2001] Reynaud, C., Safar, B., and Gagliardi, H. (2001). Une expérience de représenattion d'une ontologie dans le médiateur PICSEL. In *Proc. IC'2001*, pages 329–348.

[Sabouret, 2002a] Sabouret, N. (2002a). A model of requests about actions for active components in the semantic web. In *Proc. STAIRS 2002*, pages 11–20.

[Sabouret, 2002b] Sabouret, N. (2002b). Programmer des composants actifs dans le web sémantique. In *Journées de l'Action Spécifique STIC « Web Sémantique »*.

[Sabouret, 2002c] Sabouret, N. (2002c). *Étude de modèles de représentation, de requêtes et de raisonnement sur le fonctionnement des composants actifs pour l'interaction homme-machine*. PhD thesis, Université Paris-Sud.

[Sansonnet, 1999] Sansonnet, J. (1999). Description Scientifique du Projet InterViews — The InterViews Project. Technical Report 99-01, LIMSI-CNRS.

[W3C, 1999] W3C (1999). XML Path Language (XPath). http ://www.w3.org/TR/xpath.

[W3C, 2002] W3C (2002). Document Object Model (DOM) core specifications. http ://www.w3.org/TR/xslt.

Médiation de services sur le Web

C. REYNAUD, G. GIRALDO
Université Paris-Sud
CNRS (L.R.I.) & INRIA (Futurs)
L.R.I., Bâtiment 490
91405 Orsay cedex, France
Mail :{cr,giraldo}@lri.fr
Tél : +33 1 69 15 66 45 Fax : +33 1 69 15 65 86

Résumé

Cet article porte sur la médiation de services sur le Web de façon à les rendre plus facilement accessibles à des utilisateurs finaux. Leur diversité, leur hétérogénéité, la nécessité de les combiner, rendent nécessaire d'aider l'utilisateur final avec une interface proposant un accès convivial à un système unique, centralisé et homogène. Cet article porte sur l'utilisation de l'approche médiateur pour concevoir une telle interface.

Après avoir décrit l'approche médiateur, nous décrivons les problèmes spécifiques à la médiation de services sur le Web puis les principes à respecter pour que l'approche médiateur soit applicable. Deux types de principes sont donnés. Le premier concerne l'intégration sémantique des services. Il exploite les travaux portant sur la standardisation des transactions inter-entreprises propres à des domaines d'application. Le second type de principe concerne le passage à l'échelle du Web des systèmes médiateurs. Il repose sur deux idées : l'indépendance des modules composant l'architecture médiateur et une forte médiation. Dans une dernière section, nous décrivons comment ces principes sont mis en œuvre au sein du projet PICSEL 2, en présentant une expérimentation d'intégration de services appliquée au commerce électronique dans le domaine du tourisme.

Abstract

This document is about mediation systems integrating services on the Web to make them usable to final users. As services on the Web are numerous and heterogeneous, as they must sometimes be combined to satisfy final users requirements, interfaces giving the illusion of a convivial, unique, centralized and homogeneous access are necessary. In this paper, we propose to use a mediator approach to design such an interface.

First, we give an overview of the mediator approach. Then section 3 is about problems specific to mediation systems integrating services on the Web. We explain which problems arise and we give principles to make the mediator approach applicable in this context. The first principle is a solution to semantic heterogeneity. It exploits results in standardization relative to business transactions. The second principle concerns the scalability of the mediator approach. It emphasizes two points: de-coupling of the various components and strong mediation. In the last section, we present an architecture of a mediation system integrating services, designed in the setting of the PICSEL 2 project. We illustrate with an application of e-commerce in the tourism domain.

1 Introduction

Le commerce électronique est un domaine d'application en plein essor. Dans ce contexte existent des systèmes centralisés jouant le rôle d'intermédiaires entre entreprises, des prestataires de services d'un côté, des revendeurs, principalement des agences de voyage, de l'autre. Les échanges inter-entreprises effectués dans ce cadre, de type Business to Business (B2B), sont actuellement l'objet de travaux importants de standardisation visant à normaliser le contenu des messages échangés. Notre objectif, dans le cadre du projet PICSEL2[1], consiste à exploiter les résultats de ces travaux de normalisation pour automatiser la construction de systèmes médiateurs intégrant des services et permettre leur utilisation à l'échelle du Web.

Aujourd'hui, un utilisateur a, en effet, des difficultés pour trouver sur le Web tous les services répondant précisément à un besoin. Imaginons, par exemple, qu'il recherche un service de réservation de voyages lui permettant de réserver à la fois un hôtel dans un lieu précis et un vol pour se rendre dans ce lieu. Sur le Web, les services de réservation de ce type sont multiples et très hétérogènes. Chaque service a été conçu différemment. Il est accessible selon ses propres modalités. Il utilise un vocabulaire propre. Par ailleurs, les services accessibles via le Web diffèrent par rapport à la façon dont ils vont pouvoir satisfaire l'utilisateur. Certains offriront l'intégralité du service recherché. D'autres n'offriront qu'une partie du service mais, combinés à d'autres, pourront permettre d'obtenir pleinement satisfaction. Cela signifie qu'il sera nécessaire, dans ce cas, de faire plusieurs accès (par exemple l'un permettant de réserver un vol, l'autre permettant de réserver des hôtels). Ce processus est complexe. Les utilisateurs ont besoin d'être guidés dans leur démarche. Il faut, pour cela, leur proposer une interface simulant un système plus convivial, unique, centralisé et homogène. Les systèmes médiateurs sont une des solutions. Les résultats à ce jour concernent l'intégration de sources de données hétérogènes et, dans ce cadre, ils ont fait preuve de leur faisabilité et de leur utilité. Leur aide peut recouvrir différentes formes : découvrir les sources pertinentes étant donné un besoin, aider à accéder à ces sources pertinentes, évitant à l'utilisateur d'accéder lui-même à chacune d'elles selon ses propres modalités et son propre vocabulaire, enfin combiner automatiquement les réponses partielles obtenues de plusieurs sources de façon à délivrer une réponse globale. Ils offrent à l'utilisateur une vue uniforme et centralisée des données distribuées, cette vue pouvant aussi correspondre à une vision plus abstraite, condensée, qualitative des données et donc, plus signifiante pour l'utilisateur. Ces systèmes de médiation sont aussi très utiles en présence de données hétérogènes, car ils donnent l'impression d'utiliser un système homogène. Les problèmes d'intégration de services auxquels nous sommes confrontés ressemblent ainsi beaucoup aux problèmes d'intégration de sources. Nous proposons alors dans cet article d'utiliser l'approche médiateur pour intégrer des services.

Les services que nous souhaitons intégrer sont des services accessibles via le Web. Le contexte de déploiement nécessite que l'approche médiateur soit applicable à l'échelle du Web, ce qui signifie, en particulier, faciliter la construction des systèmes de médiation en l'*automatisant* le plus possible et concevoir ces systèmes comme des systèmes « ouverts » capables d'intégrer très facilement de nouveaux services.

Nous proposons ainsi une approche d'automatisation de la construction d'une l'ontologie, point clé de la mise en œuvre de tels systèmes d'intégration, basée sur l'exploitation de transactions standards inter-entreprises établies pour un domaine d'application donné. Nous proposons également d'automatiser la description des fonctionnalités de services respectant les standards utilisés pour construire l'ontologie, ces descriptions étant nécessaires au système médiateur pour raisonner. Une telle approche permet de ne pas faire dépendre la conception de l'ontologie directement des services intégrés. Le système médiateur ainsi conçu peut être qualifié de système « ouvert ». Il est naturellement « prêt » à intégrer de nouveaux services quels qu'ils soient, à partir du moment où ceux-ci respectent les standards sur lesquels repose la conception du système.

Dans une première partie, nous présentons l'approche médiateur appliquée à des sources de données hétérogènes. Nous décrivons ensuite les problèmes que pose la médiation de services sur le Web et

[1] Le projet PICSEL 2 est financé par France Telecom R&D dans le cadre d'une CTI. Il fait suite au projet PICSEL 1. (http://www.lri.fr/~picsel)

énonçons quelques principes devant faciliter son applicabilité. En section 4, nous décrivons comment ces principes sont mis en œuvre au sein du projet PICSEL 2 en présentant l'architecture d'un système de médiation de services dans le domaine du tourisme.

2 L'approche médiateur

L'approche médiateur a été jusqu'alors exploitée pour intégrer des sources de données hétérogènes, distribuées et autonomes. Nous décrivons ses principes. Nous citons ensuite les domaines sur lesquels des résultats ont été obtenus.

2.1 Présentation générale

L'approche médiateur [Wiederhold 1992] appliquée à un ensemble de sources hétérogènes et distribuées consiste à définir une interface entre l'agent (humain ou logiciel) qui pose une requête et l'ensemble des sources accessibles potentiellement pertinentes pour répondre. L'objectif est de donner l'impression d'interroger un système centralisé et homogène alors que les sources interrogées sont réparties, autonomes et hétérogènes.

Un médiateur est spécifique à un domaine d'application donné. Il comprend un schéma global, ou ontologie, dont le rôle est central. C'est un modèle du domaine d'application du système. Il fournit un vocabulaire structuré servant de support à l'expression des requêtes. Par ailleurs, il établit une connexion entre les différentes sources accessibles. En effet, dans cette approche, l'intégration d'information est fondée sur l'exploitation de vues abstraites décrivant de façon homogène et uniforme le contenu des sources d'information dans les termes de l'ontologie. Les sources d'information pertinentes, pour répondre à une requête, sont calculées par réécriture de la requête en termes de ces vues. Le problème consiste à trouver une requête qui, selon le choix de conception du médiateur, est équivalente ou implique logiquement, la requête de l'utilisateur, mais n'utilise que des vues. Les réponses à la requête posée sont ensuite obtenues en évaluant les réécritures de cette requête sur les extensions des vues.

L'approche médiateur présente l'intérêt de pouvoir construire un système d'interrogation de sources de données sans toucher aux données qui restent stockées dans leurs sources d'origine. Ainsi, le médiateur ne peut pas évaluer directement les requêtes qui lui sont posées car il ne contient pas de données, ces dernières étant stockées de façon distribuée dans des sources indépendantes. L'interrogation effective des sources se fait via des adaptateurs qui traduisent les requêtes, réécrites en termes de vues, dans le langage accepté par chaque source.

2.2 Problèmes étudiés

Les travaux réalisés jusqu'alors dans le domaine des systèmes médiateurs se situent principalement dans le contexte d'une médiation centralisée [Chawathe et al. 1994], [Genesereth et al. 1997], [Kirk et al. 1995], [Rousset et al. 2002].

Dans ce cadre, des études ont porté sur les langages pour modéliser le schéma global, pour représenter les vues sur les sources à intégrer et pour exprimer les requêtes provenant des utilisateurs humains ou d'entités informatiques [Goasdoue et al. 2000].

Des travaux ont porté sur la conception et la mise en œuvre d'algorithmes de réécriture de requêtes en termes de vues sur les sources de données pertinentes, celles-ci pouvant être connectées directement ou indirectement aux sources du serveur interrogé. Le problème à ce niveau peut consister à générer des expressions de calcul permettant de définir tous les objets du niveau global à partir des sources existantes. Le calcul de ces expressions nécessite la connaissance de l'ensemble des sources utiles à sa dérivation [Goasdoue 2001].

Enfin, plus récemment, certains travaux portent sur la conception d'interfaces intelligentes assistant l'utilisateur dans la formulation de requêtes, l'aidant à affiner une requête en cas d'absence de réponses ou de réponses beaucoup trop nombreuses [Bidault et al. 2000].

3 Médiation de services et Web

Utiliser une approche médiateur pour intégrer des services nécessite d'adapter les systèmes existants à ce nouveau contexte applicatif. Il s'agit principalement d'adapter le contenu de l'ontologie et de transformer les vues décrivant le contenu des sources accessibles en vues décrivant les fonctionnalités des services exécutables. Ces aspects seront abordés en section 4. Dans cette partie, nous abordons les difficultés de mise en œuvre de l'approche médiateur à l'échelle du Web. Nous décrivons en section 3.1 les problèmes à résoudre. Nous proposons en section 3.2 quelques principes à respecter pour résoudre ces problèmes.

3.1 Problèmes à résoudre pour utiliser l'approche médiateur dans le contexte du Web

Utiliser l'approche médiateur dans le contexte du Web nécessite de lever un certain nombre d'obstacles.

Un des obstacles concerne la construction d'ontologies comme support pour la recherche de services multiples, distribués et hétérogènes. La construction d'ontologies est centrale dans le développement de systèmes médiateurs. La construction manuelle d'une ontologie, même assistée par des outils conviviaux, est un travail de modélisation long et difficile. Il est absolument nécessaire de mettre en œuvre des approches permettant d'*automatiser* leur construction.

Il doit être également possible d'intégrer un très grand nombre de services, et donc de faciliter la description des fonctionnalités offertes par chacun d'eux. Là encore, une voie possible serait d'automatiser la construction des vues sur les services.

Enfin, le contexte de déploiement visé nécessite de concevoir des systèmes médiateurs « ouverts », capables d'intégrer très facilement de nouveaux services propres à un domaine, sans remettre en cause l'ontologie. Cela signifie que l'automatisation de la construction de l'ontologie ne doit pas reposer sur l'exploitation des services intégrés au sein du système. L'idéal serait de la rendre la plus indépendante possible des services qu'il est possible d'intégrer.

3.2 Principes pour un passage à l'échelle du Web

3.2.1 Principe pour l'intégration « sémantique » des services

Les travaux réalisés dans le cadre du Web en matière de standardisation étant importants, nous proposons d'exploiter les résultats obtenus en la matière, ceux concernant l'infrastructure nécessaire à la mise en œuvre des services Web (ebXML, SOAP) mais, surtout, les résultats des travaux portant sur la standardisation des transactions propres à des domaines d'application. Les organismes qui travaillent sur ces aspects [OTA, 2002], [UN/EDIFACT, 2002] normalisent le vocabulaire utilisé dans les échanges inter-entreprises. Issus d'un consensus entre professionnels d'un domaine, ces résultats sont précieux. Nous proposons de les exploiter dans la construction de l'ontologie d'un système médiateur intégrant des services du domaine concerné.

3.2.2 Principe pour le passage à l'échelle des systèmes médiateurs

Principe d'indépendance

Rendre, le plus possible, les différents modules composant l'architecture médiateur indépendants ne peut être que favorable au contexte de déploiement visé. Nous proposons ainsi que la construction du système médiateur, notamment de l'ontologie, ne dépende pas directement des services qui seront intégrés. Nous préconisons ensuite que les services intégrés soient les plus indépendants possibles du médiateur, en distinguant une partie privée, propre à chacun d'eux, considérée comme une boite noire du point de vue du médiateur, et une partie publique permettant au médiateur de communiquer. Ainsi, un service ne sera vu par un médiateur qu'au travers de sa partie publique, jouant le rôle d'interface.

Principe de forte médiation

Le système médiateur doit permettre à un utilisateur d'accéder facilement à un maximum de services d'un domaine donné, le nombre de services intégrés pouvant devenir très important sans pour cela remettre en cause la conception du système.

4 Application de ces principes dans le cadre de PICSEL 2

Les principes précédemment décrits sont mis en œuvre dans le projet PICSEL 2. L'architecture du système en cours de développement permet d'intégrer des services variés, distribués sur le Web, spécifiques à un domaine d'application donné, et dont les messages traités sont conformes aux spécifications définies par un organisme de normalisation (cf. figure n°1). A chacun des services intégrés sont ainsi associées des interfaces (publiques) décrivant les messages traités dans les termes définis par l'organisme de standardisation. Nous réutilisons le moteur de requêtes développé dans le cadre du projet PICSEL 1. Le système de médiation comporte une ontologie des services du domaine d'étude. C'est un ensemble de prédicats modélisant le domaine de services auxquels on s'intéresse. Cette ontologie est un support à l'expression de la demande de services. Elle permet aussi de connecter et de combiner les différents services accessibles. La construction de l'ontologie est basée sur l'exploitation des contenus des messages spécifiés par l'organisme de standardisation.

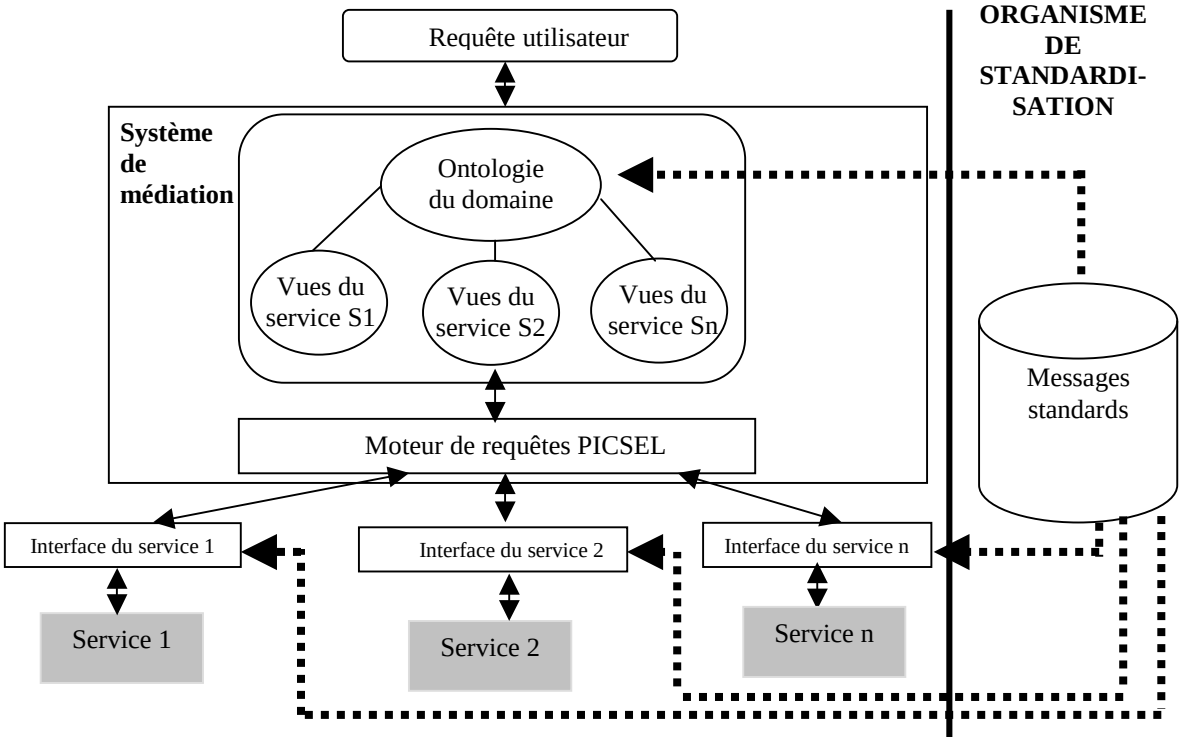


Figure n°1 : Architecture d'un système de médiation de services appliqué au commerce électronique

4.1 Exploitation de résultats de normalisation définis par l'OTA

L'application que nous avons développée a pris appui sur le travail de standardisation effectué par l'OTA [OTA 2002] dans le domaine du tourisme. L'OTA est un consortium qui regroupe plus de 150 organisations relatives à l'industrie du voyage : des agences de voyage, des hôtels, des agences de location de voitures, des compagnies aériennes, etc. En association avec la DISA [DISA 2002], cet organisme a développé des standards de communication basés sur XML pour faciliter l'emploi du commerce électronique entre entreprises.

Pour l'application réalisée, nous avons exploité 115 XML-Schemas (cf. figure n°2) décrivant le contenu des messages relatifs aux demandes de disponibilité, de réservation ou d'annulation de produits du tourisme : vols, chambres d'hôtels, locations de voitures, assurances de voyage, cours de golf, excursions. Les données mises à disposition par l'OTA sont de deux types : un premier type que nous appellerons « données métiers » qui décrivent des messages relatifs à des sous-domaines particuliers de l'industrie du tourisme (messages concernant la réservation de vols, d'hôtels, etc.) et un second type décrivant des messages plus techniques, nécessaires pour garantir une communication

fiable et sûre. Dans l'application réalisée, nous n'avons retenu que les descriptions du premier type. Ces descriptions contiennent le vocabulaire à présenter à l'utilisateur lorsqu'il désire exprimer une requête. Nous avons éliminé les descriptions de messages techniques contenant des termes que nous avons jugés non pertinents pour l'expression des requêtes.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema targetNamespace="http://www.opentravel.org/OTA/2002/08"
xmlns="http://www.opentravel.org/OTA/2002/08" xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified" version="2002A">
...
<xs:element name="OTA_AirAvailRQ">
    <xs:complexType>
        <xs:sequence>
            <xs:element name="OriginDestinationInformation" type="OriginDestinationInformationType"/>
            ...
        </xs:sequence>
        <xs:complexType name="OriginDestinationInformationType">
                <xs:complexContent>
                    <xs:extension base="TravelDateTimeType">
                        <xs:sequence>
                            <xs:element name="OriginLocation" type="LocationType">
                            </xs:element>
                            <xs:element name="DestinationLocation" type="LocationType">
                            </xs:element>
                            ...
                        </xs:sequence>
                    </xs:extension>
                </xs:complexContent>
        </xs:complexType>
    ...
    </xs:complexType>
</xs:element>
</xs:schema>
```

Figure n° 2 : Extrait d'un XML-Schema décrivant le message AirAvailabilityRequirement

4.2 Construction semi-automatique de l'ontologie du système de médiation

La construction de l'ontologie du système de médiation est automatisée. Le processus mis en œuvre exploite les données de l'OTA décrites précédemment, ce qui ne le rend pas directement dépendant des services particuliers intégrés.

L'ontologie est construite en 2 étapes. Une version initiale très simple est tout d'abord construite à la main. Ce processus de construction manuel est guidé par les données de l'OTA. Il s'agit de classer les messages pour lesquels l'OTA a défini un contenu standard (exemple : AirAvailabilityRequirement, AirBookRequirement, ...) en les regroupant par catégorie (exemple : AirBookingService). On obtient les deux premiers niveaux d'une hiérarchie de classes (cf. figure n°3). Les classes du niveau 2 correspondent aux noms de messages standards de l'OTA, les classes du niveau 1 sont des noms de catégories de messages, le nom de la classe racine précise le domaine des services auxquels on s'intéresse (Tourism Service). Cette hiérarchie de classes servira de support à l'utilisateur pour rechercher un service. En effet, les classes du niveau 2 correspondent à des services recherchés. Grâce à cette classification, l'utilisateur n'aura pas à sélectionner les services qui l'intéressent parmi un vaste ensemble de services diversifiés mais pourra faire cette sélection parmi les services appartenant à des catégories.

La hiérarchie initiale est ensuite enrichie semi-automatiquement à partir des descriptions des contenus des messages tels que les a spécifiés l'OTA. L'approche retenue comporte 3 phases : (1) une phase d'extraction d'éléments utiles pour compléter l'ontologie initiale (classes, propriétés, relations), exploitant la structure des messages standards de l'OTA, (2) une phase d'organisation des éléments acquis, (3) une phase de représentation des connaissances préalablement acquises dans un formalisme à base de classes interprétable par le moteur de requête PICSEL.

La phase d'extraction est semi-automatique. Le résultat de cette phase est un ensemble de classes, de propriétés caractérisant des classes, et de relations entre classes. Le processus d'extraction mis en œuvre est décrit dans [Giraldo et al. 2002].

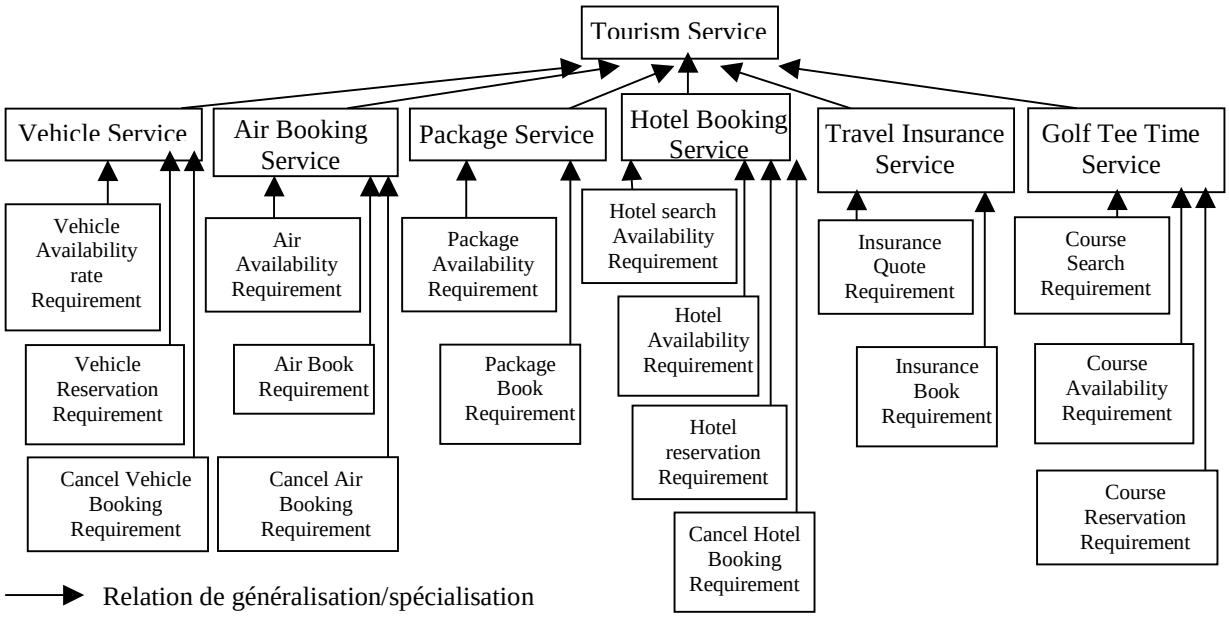


Figure n° 3 : Hiérarchie initiale de concepts

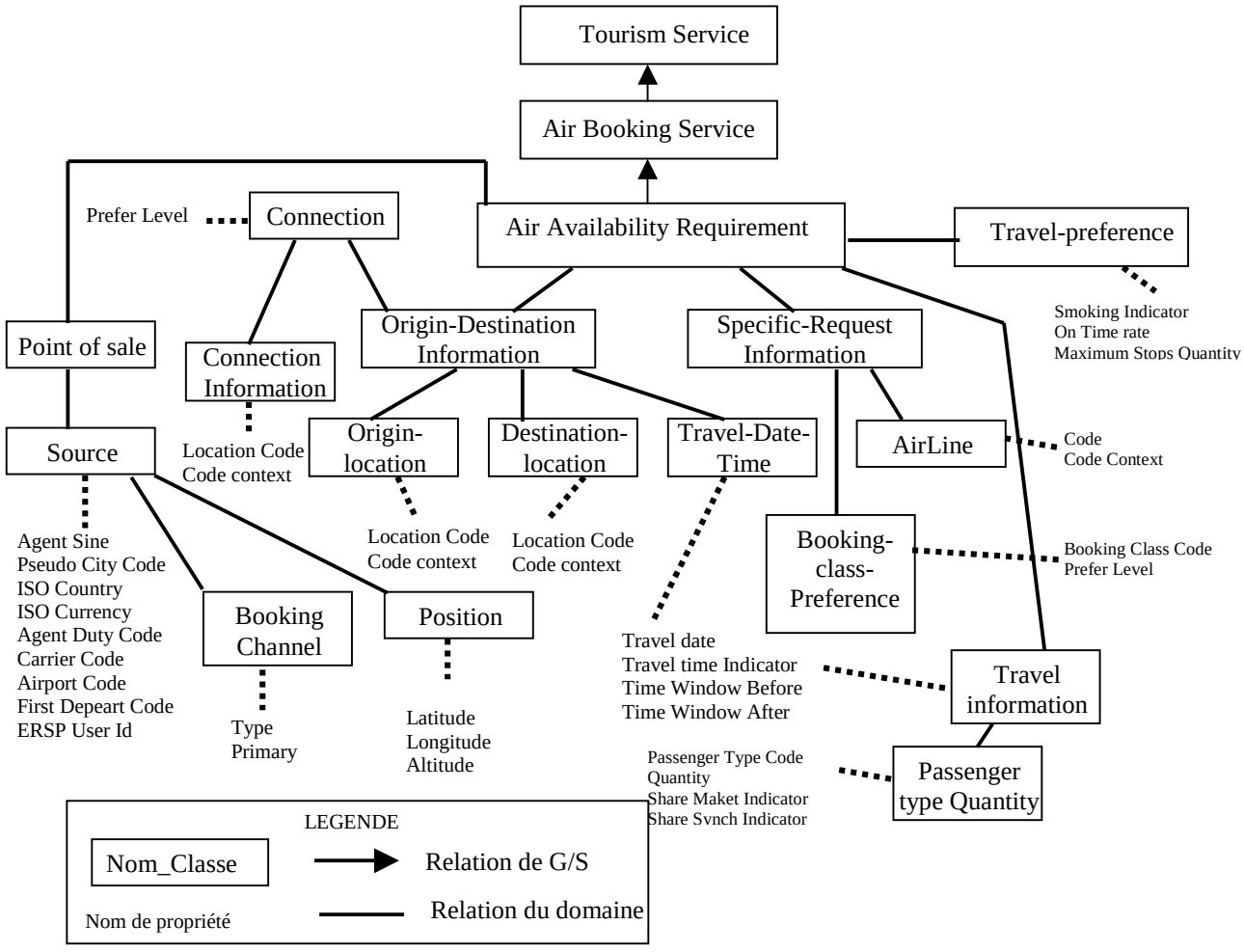


Figure n°4 : Partie de l'ontologie initiale enrichie avec des classes, des propriétés et des relations

La phase de structuration est entièrement automatique. Comme les noms des messages standards de l'OTA font partie des composants extraits automatiquement, pour lesquels des propriétés et des relations avec d'autres classes ont été trouvées, et qu'ils correspondent aussi aux classes du niveau 2 de la hiérarchie initiale construite manuellement, cette phase de structuration est facilitée. Elle conduit à un réseau de relations liant les classes les unes aux autres (cf. figure n°4).

L'ontologie est ensuite représentée à l'aide de la composante terminologique de CARIN-ALN[2] [Goasdoue et al 2000], comprenant des définitions et des inclusions de concepts. Ce processus de traduction est entièrement automatique.

L'ontologie obtenue comprend une classification des messages standards de l'OTA, chaque message étant précisément défini. Ainsi, le message AirAvailabilityRequirement est un message qui appartient à la catégorie AirBookingService. Il est défini comme une classe à laquelle sont associées des informations portant sur le départ et l'arrivée, en l'occurrence, un lieu de départ, un lieu d'arrivée, une date et une heure, des informations plus spécifiques portant sur la compagnie aérienne, la classe, le nombre de passagers, etc. Tous ces termes sont utiles à l'utilisateur. Ils lui permettent de renseigner les données contenues dans les messages, ces données étant nécessaires pour que les services puissent faire le traitement correspondant.

4.3 Génération automatisée des vues décrivant les fonctionnalités des services intégrés

La description du contenu d'une source en termes de vues, dans PICSEL, consiste principalement à définir un ensemble d'implications logiques de la forme $v_i(X) \Rightarrow p(X)$ reliant chaque vue v_i à la relation du domaine p dans l'ontologie dont elle peut fournir des instances.

Dans le contexte dans lequel nous nous plaçons, les services intégrés sont des services capables de traiter des messages définis par l'OTA. Si nous utilisons PICSEL pour intégrer les services qui adhèrent aux standards de l'OTA, il faut alors définir autant de vues que de messages gérés par ces services (cf. exemple ci-dessous). Les services déclarent ces messages dans leur partie publique (cf . figure n°1 – Interface du service). Ces définitions de vues peuvent dont être générées automatiquement à partir des noms de messages extraits de la partie publique des services intégrés.

Exemple :

VOL_v1 ⇒ AirAvailabilityRequirement (1')
VOL_v2 ⇒ AirBookRequirement (1'')

v1 et v2 sont des noms de vues respectivement associées aux termes de l'ontologie AirAvailabilityRequirement et AirBookRequirement, deux classes correspondant à des noms de messages de l'OTA que le service VOL déclare être capable de traiter. Ainsi, ces implications expriment que le service VOL permet d'une part de rechercher s'il existe des places disponibles sur des vols (1'), d'autre part de réserver des places sur des vols (1'').

Ces déclarations de vues ne suffisent pas. Un utilisateur doit pouvoir décrire précisément le service qu'il recherche. Il doit par exemple pouvoir indiquer qu'il recherche un service de réservation de vols *au départ de Roissy-CDG*. Exprimer une recherche de services à ce niveau de détail est tout à fait possible. Il suffit d'exploiter la partie de l'ontologie définissant les messages de l'OTA. Toutefois, pour que cette information puisse être prise en compte par le médiateur, que celui-ci l'intègre dans son raisonnement, il est nécessaire de définir des vues associées à l'ensemble des relations de l'ontologie utilisées à cette fin. Ceci revient à spécifier les données des messages traitées par les services.

Exemple :

VOL_v3 ⇒ OriginDestinationInformation (2')
VOL_v4 ⇒ OriginDestinationInformationAssociate (2'')

v3 et v4 sont des noms de vues respectivement associées à la classe de l'ontologie

[2] Langage exploité par le moteur de requêtes PICSEL.

OriginDestinationInformation et à la relation OriginDestinationInformationAssociate (nom de rôle en CARIN-ALN). Ces données sont utiles pour préciser, au service VOL, les informations relatives au départ et à l'arrivée du vol pour lequel l'utilisateur recherche l'existence de places disponibles.

Les services intégrés au sein du médiateur étant des services ayant adhéré aux standards de l'OTA, la composition des messages traités est celle qui est spécifiée par l'OTA. Les données qui en font partie sont représentées dans l'ontologie (cf. figure n°4). Lors de la traduction de l'ontologie en CARIN-ALN, ces données ont été représentées sous forme de relations, les classes ont été représentées par des relations unaires, les rôles (relations du domaine ou propriétés) ont été représentés par des relations binaires. Il est ainsi tout à fait envisageable de générer automatiquement les vues associées à l'ensemble des relations liées directement ou indirectement à une classe correspondant à un certain message (classe du niveau 2 de la hiérarchie initiale) à partir de l'ontologie.

4.4 Génération automatique des messages à traiter en réponse à une recherche de services

Soit la requête Q exprimant la recherche de services informant sur la disponibilité de vols et de chambres d'hôtels, pour un vol partant de l'aéroport Roissy-CDG le 1er mai 2003 à destination d'Athènes :

Q(x) : VOL+HOTELAvailabilityRequirement (x).

Afin d'éviter à l'utilisateur d'avoir à parcourir lui-même l'ontologie pour sélectionner les termes traduisant au mieux sa demande, nous utilisons le module d'aide à la formulation des requêtes de PICSEL [Reynaud et al. 2002]. Ce module propose à l'utilisateur des requêtes prédéfinies et le guide ensuite dans l'expression des contraintes à satisfaire. Q(x) correspond à l'une de ces requêtes prédéfinies. Les contraintes, dans ce cas précis, portent sur le lieu où se trouve l'hôtel, les lieux de départ, d'arrivée et la date de départ du vol. Elles sont précisées au cours d'un dialogue proposé par le module d'aide.

Si PICSEL comprend la description de 3 services : S_1 traitant uniquement des messages du type AirAvailabilityRequirement, S_2 traitant uniquement des messages du type HotelAvailabilityRequirement, S_3 traitant les 2 sortes de messages, les réécritures fournies en réponse à la requête Q(X) signifient que la réponse complète à la requête peut être obtenue de quatre façons différentes : en composant S_1 et S_2, S_1 et S_3, S_2 et S_3 ou en accédant uniquement à S_3. Lorsque plusieurs services sont nécessaires à l'exécution complète du service recherché, le moteur indique les conditions selon lesquelles cette composition doit être effectuée, en l'occurrence ici il indique que la ville de destination du vol recherché lors de la recherche de disponibilité d'un vol doit être identique au lieu où se trouve l'hôtel lors de la recherche de disponibilité de chambres d'hôtels.

Une fois le résultat du médiateur connu, il est possible de générer automatiquement le contenu des messages (document XML cf. figure n°5) que doivent exécuter les services auxquels le médiateur a accès. La structure des ces messages est donnée par l'OTA sous forme de XML-Schema. Le plan de requêtes fourni par PICSEL indique les valeurs des attributs à renseigner dans le document (en gras sur la figure n°5).

```
<AirAvailabilityRequirement>
    ...
    <OriginDestinationInformation>
        <OriginLocation locationCode= 'cdg ' />
        ...
        <TravelDateTime traveldate='01/05/2003' />
    </OriginDestinationInformation>
    ...
</AirAvailabilityRequirement>
```

Figure n°5 : Extrait d'un document XML décrivant le message AirAvailabilityRequirement à envoyer au service S_1 à l'issue de l'exécution de PICSEL

5 Conclusion

Nous avons présenté une architecture de système médiateur « ouvert » intégrant des services de commerce électronique dans le domaine du tourisme. L'approche est une première voie de solution pour le passage à l'échelle du Web des approches médiateurs. Le système présenté en partie 4 est en cours d'implémentation. La construction de l'ontologie et sa traduction en CARIN sont terminées. Les travaux en cours concernent la génération automatisée des vues décrivant les fonctionnalités des services et des messages à traiter en réponse à une recherche de services. Ce travail sera poursuivi pour intégrer une modélisation des différents utilisateurs.

Références

[Bidault et al. 2000] Bidault A., Froidevaux C., Safar B. (2000). Repairing queries in a mediator approach. In 14th European Conference on Artificial Intelligence, p. 406-410, Berlin.

[Chawathe et al. 1994] Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J. D., Widom J. (1994). The TSIMMIS project: Integration of heterogeneous information sources. In 16th Meeting of the Information Processing Society of Japan, p. 7-18, Tokyo.

[DISA 2002] Data InterChange Standards Association. (2002). http://www.disa.org

[Genesereth et al. 1997] Genesereth M. R., Keller A. M., Duschka O. M., (1997). Infomaster: an information integration system. In Joan M. Peckman, editor, proceedings, ACM SIGMOD International Conference on Management of Data: SIGMOD 1997: May 13-15, 1997, Tucson, Arizona, USA, volume 26(2) of SIGMOD Record (ACM Special Interest Group on Management of Data), p. 539-542, New-York, NY 10036, USA. ACM Press.

[Giraldo et al. 2002] Giraldo G., Reynaud C., (2002). Construction semi-automatique d'ontologies à partir de DTDs relatifs à un même domaine, 13èmes journées francophones d'Ingénierie des Connaissances, Rouen, 28-30 mai.

[Goasdoue 2001] Goasdoue F. (2001). Réécriture de requêtes en termes de vues dans CARIN et intégration d'informations. Université de Paris XI – Orsay, novembre 2001. Thèse de doctorat.

[Goasdoue et al. 2000] Goasdoue F., Lattes V., Rousset M.-C. (2000). The use of CARIN language and algorithms for Integration Information: the PICSEL system, International Journal of Cooperative Information Systems, 9(4):383-40.1

[Kirk et al. 1995] Kirk T., Levy A. Y., Sagiv Y., Srivastava D. (1995). The Information Manifold. In C. Knoblock and A. Levy, editors, Information Gathering from Heterogeneous, Distributed Environments, AAAI Spring Symposium Series, Stanford University, Stanford, California.

[OTA, 2002] Open Travel Alliance. (2002). www.opentravel.org.

[Reynaud et al. 2002] Reynaud C., Safar B. (2002). Aide à la formulation de requêtes dans un médiateur. 13ème congrès francophone AFRIF-AFIA de Reconnaissance des Formes et Intelligence Artificielle, 8-10 janvier 2002, Angers.

[Rousset et al. 2002] Rousset M.-C., Bidault A., Froidevaux C., Gagliardi H., Goasdoue F., Reynaud Ch., Safar B., (2002). Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL, Revue I3, Volume 2, Numéro 1, pages 9-59.

[UN/EDIFACT, 2002] United Nations Directories for Electronic Data Interchange for Administration, Commerce and Transport. (2002). United Nations Economic Commission for Europe, www.unece.org/trade/untdid.

[Wiederhold 1992] Wiederhold G. (1992). Mediators in the architecture of future information systems, Computer, 25(3):p.38-49.

Meilleures interfaces entre services et utilisateurs

ISABELLE BERRIEN ET FRANCOIS LABURTHE
e-lab BOUYGUES SA
1, avenue Eugène Freyssinet,
78061 Saint Quentin en Yvelines, FRANCE
Mail :iberrien@bouygues.com
Tél : +33 1 30 60 53 66 Fax : +33 1 30 60 22 15

Résumé

Cette étude présente une méthode pour améliorer le dialogue d'un site avec un utilisateur quand celui-ci navigue pour chercher un contenu particulier. Nous proposons de concevoir un service s'inspirant du bon sens humain qui nous permet de nous adapter à nos interlocuteurs. Nous montrons ainsi comment passer continûment d'un échange d'informations générales à des informations de plus en plus précises. Ce dialogue continu peut se décliner comme raffinement progressif d'une question trop floue, ou comme la recherche d'un compromis pour une question trop exigeante. Nous détaillons en particulier la prise en compte des annotations sémantiques sur le contenu proposé par le site tout au cours de ces échanges. La pertinence d'une structuration hiérarchique de valeurs sémantiques tant au niveau de la saisie (lorsque l'utilisateur exprime son souhait par exemple) qu'au niveau des réponses qui lui sont fournies est démontrée tout au long du processus d'échange et reproduit la souplesse qu'on peut retrouver lors d'un échange entre 2 personnes. Elle permet d'éviter quel que soit l'état de l'utilisateur (indécis ou au contraire déterminé) des situations d'impasse qui non seulement ralentissent le processus de requête mais nuisent indubitablement au site et diminuent les chances que l'utilisateur revienne.

Abstract

This study presents a method to improve dialog between a web site and an internaut searching for a particular content. We propose a service inspired from the human capability of adjusting to speakers. We show herein how to slide smoothly from a general information exchange to a more and more precise one. This continuous dialog can be viewed as a progressive refinement of a too fuzzy request, or as a seek for compromises for a too demanding one. We relate in particular the importance of semantic annotations of the content given by the site during the exchange process. The pertinence of a hierarchical structure adapted to semantic values either on the form stage (when the user expresses his wish for instance) or on the answers is clearly demonstrated and reflects the flexibility you can find when people talk to each other. It prevents whatever the user's mind state is (undecided or rather demanding) from falling into deadlocks which not only slow down the request process but also decrease the probability the user comes back.

1. Des sites désagréables à utiliser

La « recherche » d'information, de liens, de données est une des fonctionnalités majeures des sites web mis à notre disposition. Le plus souvent, les moyens mis en œuvre lorsque l'utilisateur arrive sur un site et qu'il recherche quelque chose sont soit un champ de saisie de mots clés, soit des liens qui permettent l'accès par clics successifs en général arborescents. Parce que nous avons remarqué que ce sont des situations dans lesquelles l'utilisateur est souvent mis en difficulté, nous avons cherché à trouver une solution afin de créer autour de lui un climat de confiance. Dans un premier temps, nous recensons quelques défauts communs de sites et services, tant d'information que de e-commerce[1,2]. Il est en fait question de tous les sites d'exploration d'un catalogue de produits ou d'une masse de contenu.
Dans un deuxième temps, nous présentons une méthode de conception de sites et d'outils d'aide à la navigation pour pallier aux défauts identifiés.

1.1 Défauts des sites actuels

Des sites complexes et longs à utiliser : l'utilisateur est souvent confronté à de multiples écrans intermédiaires avant d'obtenir l'écran de requête souhaité. Dans les interfaces de saisie classiques (formulaires multiples), aussi bien la saisie que les renvois d'informations sont très souvent présentés dans un cadre ne faisant pas apparaître de liens logiques ou conceptuels entre eux.
Par exemple, dans un site de navigation musical, où le raffinement des styles musicaux est très important, on peut accéder *via* une requête par mots clés à l'artiste désiré, mais il est impossible de savoir dans quelle partie de l'arborescence des styles il se situe. L'utilisateur a alors une impression de confusion [LARDY *et al*, 1992].
De plus, la navigation hypertextuelle sur le WWW tend à créer un phénomène de désorientation. En navigant de lien en lien, le visiteur éprouve quelquefois des difficultés à se situer spatialement et logiquement dans l'environnement web qu'il explore. Le mécanisme mis en œuvre pour atteindre son but est souvent altéré.
Des sites peuvent contenir un grand nombre d'informations fiables et à jour mais néanmoins avoir une organisation mal perçue par l'utilisateur. Si celui-ci tourne en rond ou se perd dans le site à la recherche de son information, il n'est pas plus avancé... Ceci entraîne la plupart du temps de nombreux retours en arrière pour sortir de l'impasse et il en ressort une confusion générale. Il est important que l'utilisateur sache exactement où il en est. Ce défaut est souvent constaté dans des sites d'achats diversifiés tel que http://www.lacentrale.fr **ou** http://www.ebay.fr.
Un site où l'information est bien organisée sera plus facile à consulter. Une segmentation logique de l'information et des titres clairs aideront les visiteurs à s'orienter à l'intérieur même du site.

Des sites rigides : la quasi-totalité des formulaires disponibles sur le net présentent des champs obligatoires. Ceci met à nouveau l'utilisateur en situation de difficulté soit parce qu'il ne comprend pas ou tout simplement parce qu'il ne sait pas répondre à une question.
L'utilisateur est forcé de formuler quelque chose qu'il n'a pas forcément en tête. Il entre dans un échange qui n'est plus en adéquation avec sa logique et se trouve acculé dans des situations parfois qui ne le satisfont pas et donc nuisent au site.

Des sites irritants : la plupart des sites, face à une réponse trop vaste vont soit fournir des pages entières de réponses ou un écran « pas de solution ».

Avec tous ces défauts, il est fréquent que l'utilisateur abandonne[3] [Axance, 2001]. Environ 65 % des acheteurs en ligne abandonnent avant l'achat, ce qui un pourcentage conséquent.

---

[1] *Evaluation d'un site Web (Bibliothèque des sciences et de la santé, Université de Montréal)* http://www.bib.umontreal.ca/SA/caps31.htm (02-10-2000)
[2] *Evaluation de la recherche d'information Herschaft, L. (URFIST de Lyon)* http://www.urfist.cict.fr/lettres/lettre24/lettre24-44.html
[3] http://www.nlc-bnc.ca/publications/1/p1-256-f.html; http://www.axance.com/

Il semblerait donc qu'en rendant l'interface plus **compréhensibl**e, les informations (offertes ou renvoyées) plus **exploitables et pertinentes** et en offrant à l'utilisateur une **saisie neutre** de ses critères, l'utilisateur se sentirait plus à son aise, trouverait plus vite ce qu'il recherche et la mission du site serait atteinte.
Une interface plus compréhensible signifie qu'elle doit permettre à l'utilisateur de converger vers la réponse dans un temps suffisamment court au delà duquel il se déconnecte ou change de site. Avoir accès à des informations plus exploitables et pertinentes sous-entend par exemple au niveau du formulaire de saisie, avoir des champs de saisie en adéquation avec le contexte de requête et fidèles à la logique de la base de données. Enfin, une saisie neutre des critères d'un formulaire veut simplement dire qu'aucun de ses champ n'est obligatoire.

1.2. Proposition et principes de la plateforme Wishbone

Nous décrivons là en réponse à tous les problèmes évoqués ci-dessus une méthode reposant sur une plateforme logicielle Wishbone (« a Backbone for Your Wishes » [Caseau and Laburthe, 2002], [Berrien *et al*, 2003]), qui permet de concevoir des sites et services dont l'interface est simple, neutre et adaptative [Höök, 1997]. Bref, nous proposons une solution qui permet de faire des sites efficaces et attachants.

Notre travail se situe en aval de la tâche de structuration du contenu. Nous partons du postulat qu'une source de données déjà structurées est disponible. Bien entendu, le travail d'acquisition, indexation, structuration, et organisation est primordial. Et la méthode que nous proposons est d'autant plus efficace que les données sont riches et bien décrites.
Cet outil prend en entrée un catalogue de ressources, décrites en XML ainsi qu'un modèle de données, construit à partir des types de bases (*bool, string, num*) et de types extensibles (*symbol* parmi différentes ontologies prédéfinies, *object* dans une hiérarchie de classes).
Notre démarche est très générale et s'applique à de nombreux types de sites différents. A titre d'illustration, notre plateforme logicielle a été mise en œuvre entre autres sur :

- un site Web de e-commerce BtoC
- un Intranet recueillant l'ensemble des documents de travail d'une SSII
- une base de FAQ utilisées dans un centre d'appel pour assistance clientèle
- un Intranet recueillant 15 ans d'historique de fiches projets, partenaires

Enfin, notre approche permet de mettre en œuvre un même type d'échange, quelque soit le terminal utilisé (Web, Wap, i-mode, Voix, …).

2. Principes d'un dialogue plus humain

2.1 Expérience personnelle

Nous avons tous en tête des situations où, en tant que consommateur, nous avons été contents de côtoyer un vendeur compétent dans son domaine, et avenant de surcroit.
Par exemple, prenons l'achat d'une maison dans un point de vente immobilier. Le vendeur nous pose deux ou trois questions pour situer notre demande dans son contexte, décrit alors très rapidement l'offre (les maisons individuelles, assez chères, et pour ceux qui veulent faire des économies des maisons mitoyennes, meilleur marché), laisse le client choisir sa catégorie avant de présenter ses produits plus précisément. A part quelques approches de ce système d'interaction [Pu et al, 2000], peu de sites permettent de refléter ce type de comportement, et de reproduire le même cheminement.

2.2. Principes clés pour un dialogue réussi

La plateforme logicielle que nous présentons, Wishbone, est basée sur ce processus : accompagner le client en lui présentant des vues synthétiques de groupes d'objets. Un groupe d'objets se décrit par ses

caractéristiques communes (sa catégorie, sa fonction, ses thèmes, sa marque, sa plage de prix, ….). Choisir un groupe d'objet revient à poser une question plus précise.
Toute la partie de la discussion qui reste au niveau des groupes d'objets sert à cerner l'intention et les besoins de l'utilisateur.
Tout l'art de la vente consiste à amener cette discussion sur les besoins jusqu'à une demande précise et satisfiable. L'étape ultime ne consiste alors plus qu'à présenter les quelques solutions correspondant à cette demande.

Cette recherche d'un dialogue à un niveau abstrait (sur l'intention de l'utilisateur) est implémentée dans cette plateforme logicielle. Elle permet de remplir les objectifs suivants en offrant une interface:

- adaptative : être capable de percevoir le cadre de référence de l'utilisateur
- informative : pouvoir lui fournir les renseignements qu'il faut au bon moment, bien triées
- continue : ne pas rompre le dialogue soit en offrant des vues synthétiques de groupe d'objets en donnant un aperçu pertinent de chaque groupe (projections, voir ci-dessous), soit de générer un compromis (éviter à tout prix la sensation d'échec)

Le service joue ainsi un rôle d'intermédiaire entre les souhaits de l'utilisateur et les disponibilités d'un catalogue. Il adopte un point de vue dirigé par les données pour répondre aux demandes, ce qui assure de toujours faire converger le dialogue vers une solution à la recherche.

3. Scénarios d'interaction

Le dialogue entre l'utilisateur et le moteur est fondé sur un échange. L'information échangée se situe à un niveau abstrait : l'utilisateur soumet sa demande ; le service l'assiste pour préciser sa demande jusqu'à une formulation précise et satisfiable. Ces deux messages peuvent être décrits dans un formalisme commun puisqu'ils décrivent des contraintes sur les caractéristiques du contenu recherché.

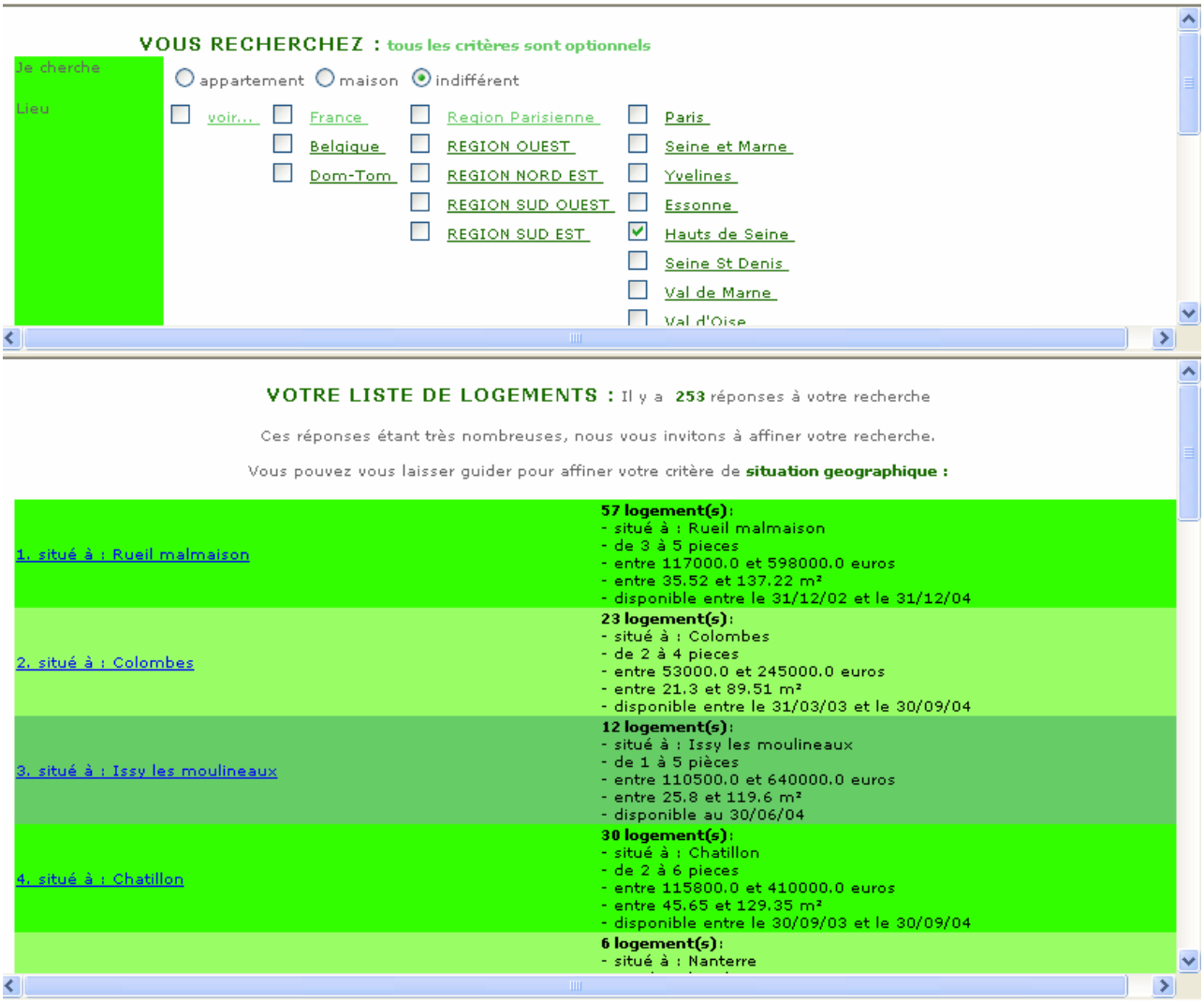
Plus précisément, la réponse fournie par le moteur est typée différemment selon les circonstances :
Lorsque le volume de réponses est trop grand pour qu'une présentation à plat de toutes les solutions soit facilement exploitable par l'utilisateur, la plateforme Wishbone présente à l'internaute toutes les solutions groupées par paquets. C'est ce que nous avons défini comme approche « gros grain ». Le moteur pratique alors une segmentation des solutions par groupes d'objets similaires. Chaque groupe est présenté en décrivant les caractéristiques qui lui sont communes (toutes les maison individuelles de plus de 350000 euros ont un jardin).
Lorsqu'il n'y a pas de réponses, cette même approche de survol est utilisée. Il propose alors à nouveau des groupes de réponses les plus proches de la requête initiale, chaque groupe étant le plus proche soit des critères conservés, soit des critères relâchés.
Lorsque le volume de réponses est suffisamment faible pour pouvoir les présenter à l'utilisateur sans créer une sensation de « surdosage », alors l'application affiche les solutions. C'est exactement le comportement standard qu'on a actuellement avec les moteurs de recherche ou les formulaires de requête actuels. Il s'agit de la version « grain fin » du dialogue, c'est typiquement le cas où le conseiller de vente vous emmène voir les 3 plans de maisons qui sont sensés correspondre à votre souhait.

4. Utilisation des annotations sémantiques

La capacité de raisonner à différents niveaux de précision est justement un des atouts des travaux associés à la prise en compte de dimensions sémantiques des données. C'est ce qui permet entre autres de recréer une atmosphère d'échange humain. Par exemple, une hiérarchie de concepts permet de savoir que Boulogne est en Ile de France, qui est en France, qui est en Europe…
Nous modélisons nos données à l'aide d'un catalogue avec des champs symboliques ainsi que des ontologies pour organiser les symboles en différentes hiérarchies. Ces ontologies sont utilisées soit de manière statique, dans un formulaire, pour recueillir le souhait de l'utilisateur, soit de manière dynamique, pour affiner ou relâcher ce souhait. Il peut comme il le désire formuler une requête très

précise en indiquant qu'il recherche un appartement à Boulogne, ou de manière plus floue en indiquant juste Région Parisienne. Ou encore, l'application peut lui proposer parmi la Région Parisienne le choix entre Paris et Seine-et-Marne…. Elle peut transformer une demande « Boulogne » en demande « Hauts-de-Seine » s'il n'y a pas de réponse.
L'avantage majeur des ontologies utilisées dans ce contexte est de permettre d'atteindre différents niveaux de précision avec un seul critère de saisie et d'induire une « distance sémantique » associée. Ceci évite la présence de plusieurs champs. Pour saisir le lieu par exemple, l'utilisateur devra remplir soit la région, le département ou la commune car ces critères ne sont pas liés entre eux. Alors qu'il y a clairement une notion de spécialisation entre ces valeurs. L'avantage d'une telle structure est que ce souhait sera interprété de manière dynamique par le moteur grâce à cette hiérarchisation des concepts, lui inférant une souplesse supplémentaire .

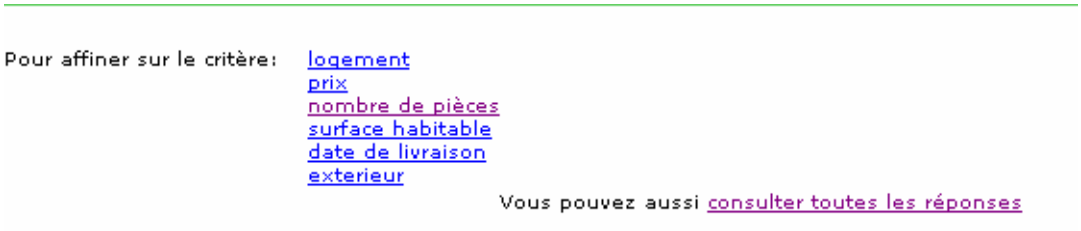
Imaginons le cas où l'utilisateur est moins fixé et qu'il veut juste un logement dans les Hauts de Seine (Figure 1).



-Figure 1-

Il y a 253 réponses à cette requête, ce qui est sans aucun doute un chiffre trop élevé pour espérer satisfaire l'internaute en lui présentant la liste complète.

Le moteur présente donc ces réponses triées selon un critère qu'il juge le plus pertinent pour le regroupement par paquet (ici, le lieu, c'est-à-dire toutes les communes de Hauts de Seine). L'utilisateur clique alors sur le groupe de son choix et raffine ainsi rapidement sa recherche.
Egalement, à tout moment de la navigation, lorsqu'il est encore un stade d'abstractions, l'utilisateur peut décider soit de voir les solutions, soit de trier les réponses selon le critère de son choix (Figure 2):

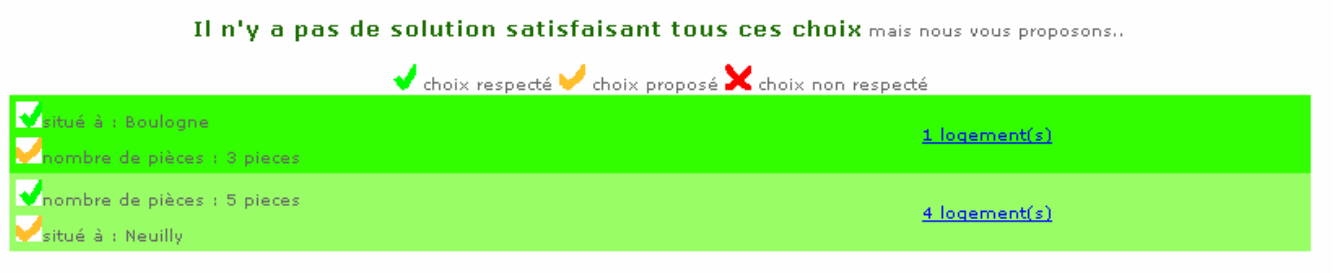


-Figure 2-

Prenons maintenant le cas où l'utilisateur recherche un appartement de 5 pièces à Boulogne (Figure 3).

-Figure 3-

La base de données n'a pas d'appartement satisfaisant la requête de l'internaute, et fournit la réponse suivante (Figure 4) :



-Figure 4-

On distingue 2 groupes de réponses, l'un où c'est le lieu qui a été respecté (Boulogne), l'autre où c'est le nombre de pièces (5 pièces).
Il faut noter que à chaque fois, c'est le groupe de solutions avec la plus grande similarité qui est fournit.
On remarque donc que pour un appartement à Boulogne, c'est un 3 pièces qui est proposé, car c'est le critère le plus similaire qui permet de fournir une réponse.
De même, pour avoir un 5 pièces, il faut aller à Neuilly, qui est bien dans les Hauts de Seine. On voit en encore bien là apparaître la proximité des réponses fournies en alternative, où toute la pertinence de la réponse repose sur l'adéquation entre relâchement de la contrainte et le nombre de solutions.

5. Principes de fonctionnement du moteur

Son principe de fonctionnement est d'être un moteur de requêtes dirigé par les données. On entend par le terme « dirigé par les données » le fait que les contraintes ne sont pas prises en compte directement. Une fois l'ensemble de solutions calculé à partir des contraintes envoyées, le moteur les reformule, les analyse et les retourne à nouveau sous forme de contraintes le plus proche de la description des objets de la base de données. C'est pourquoi toute la gestion des contraintes réalisée par le moteur est relative à l'état courant de la base de données.

5.1 Principes généraux

Lorsqu'une requête est envoyée au moteur, le premier traitement consiste à calculer l'ensemble des solutions. Le traitement algorithmique diffère suivant que l'on en trouve ou non.

5.1.1. Cas où il y a des solutions

Le traitement applique trois familles d'algorithmes : des algorithmes de projection, qui projettent l'ensemble des solutions sur un critère donné, des algorithmes de séparation qui partitionnent l'ensemble des solutions en groupes aussi denses que possibles et enfin des algorithmes de traitement de contraintes pour fournir, pour chaque caractéristique, l'expression la plus adaptée au contexte donné.
Projection : L'ensemble des solutions à la requête est un ensemble d'instances dans une hiérarchie de classes. Le plus petit ancêtre commun à toutes ces classes est calculé : c'est le support sur lequel se projette l'ensemble des solutions. Puis, pour chacun des champs de ce support, une image est calculée : il peut s'agir d'un intervalle de valeurs (comme des nombres), d'un ensemble de valeurs (des clés parmi une énumération), d'un type d'objets, d'une union de sous-arbres dans une ontologie, etc … Cet algorithme de projection est appliqué tant pour décrire l'ensemble global des solutions à la requête de l'utilisateur que chacun des groupes dans la partition.
Partition : Lorsque l'ensemble des solutions est trop grand, le moteur essaye de séparer les solutions par groupes d'objets similaires. On cherche ainsi, pour chaque dimension de l'espace (chaque champ

défini sur la classe support), une séparation en groupes. Ces groupes doivent être peu nombreux (pour proposer une alternative à quelques branches), de taille comparable (pour garantir une convergence rapide de la navigation), et denses (regrouper des objets dont la valeur du champ étudié soit comparable). Dans le cas de séparation suivant un champ numérique, on peut toujours séparer les objets en classes disjointes, dans le cas de valeurs dans une ontologie, les classes peuvent se recouvrir.
Simplification : L'information retournée par le moteur est une description de groupes d'objets exprimée sous forme de contraintes. Ces contraintes sont simplifiées pour ne présenter que l'information strictement pertinente. Par exemple, on évite de présenter de contraintes déjà connues au moment précis du dialogue : chacun des groupes d'objets dans une partition n'est décrit que par l'information qui lui est spécifique (si parmi les 120 appartements possibles, le groupe des 35 moins chers est en banlieue ouest, l'information de localisation géographique n'est présentée que si ce n'était pas le cas de l'ensemble des 120 appartements). Par ailleurs, les algorithmes de partition numérique sont biaisés pour fournir des groupes dont les bornes des intervalles caractéristiques sont autant que possible des « chiffres rond » (et correspondent ainsi à des catégories significatives pour l'utilisateur).

5.1.1. Cas où il n'y a pas de solutions

Dans le cas où aucune solution n'est trouvée à la requête de l'utilisateur, le moteur essaie de transformer celle-ci en une requête un peu différente, et qui admette des solutions. Pour cela, on cherche à garder une partie des contraintes, en relâcher d'autres (en élargissant des bornes numériques par exemple), et en oublier complètement d'autres. Le choix des contraintes gardées, relâchées et oubliées traduit un ordre de préférence sur les différents critères qui ont été contraints.
Le serveur calcule différentes propositions correspondant à différents ordres de préférence : pour chaque proposition, un sous-ensemble maximal de contraintes « dures » est gardé, un sous-ensemble minimal de contraintes est oublié, et les contraintes restantes sont relaxées aussi peu que possible, de sorte à ce que quelques solutions existent. Cette génération de compromis correspond à un calcul dirigé par les données puisque l'on assure que chaque proposition admet des solutions. Par ailleurs, la proposition explicite de différentes manières de trouver un compromis permet de ne faire aucune supposition sur les préférences de l'utilisateur, et ainsi, de s'adapter à ses goûts, pour autant que les données disponibles le permettent.

5.2. Cas particuliers des annotations sémantiques

5.2.1. Etude des algorithmes de clustering

Lorsqu'on soumet une requête au moteur et que l'ensemble des solutions est trop grand pour qu'une présentation brute de toutes les réponses soit convenable, le moteur partitionne cet ensemble de solutions. Les algorithmes utilisés font du clustering unidimensionnel (c'est à dire qu'ils partitionnent l'ensemble des objets solution selon les valeurs d'un champ donné) et des projections (fourniture de plages de valeurs communes à l'ensemble des objets du groupe). On distingue 2 cas possibles :

- ➔ champ multivalué mais pointant sur une seule ontologie
- ➔ Certains champs des instances de la base de données peuvent pointer sur plusieurs ontologies.

5.2.2.Etude des algorithmes de projection

Dans le cas où les réponses fournies sont trop nombreuses, le moteur fournit donc une réponse organisée en groupe de solutions. Afin de les caractériser et donc de les rendre plus lisibles aux yeux de l'utilisateur, ces clusters sont décrits grâce à des caractéristiques communes calculées par le moteur. Il est possible de définir préalablement les caractéristiques que l'on ne veut jamais voir apparaître parce qu'elles ne sont pas intéressantes dans le processus d'exploration ou de recherche. D'autre part, les critères se voient attribuer initialement une valeur d'intérêt par défaut qui évolue selon le cheminement réalisé par l'utilisateur.

5.2.3. Etude de l'algorithme de relaxation.

Lorsque la requête posée ne génère aucune solution, le moteur propose des alternatives qui relâchent en fait les contraintes posées afin de fournir des solutions les plus proches de celles conformes à la requête.

6. Conclusion

Nous avons présenté dans cet article une méthode nouvelle d'échange entre un utilisateur en quête d'information (que ce soit dans le but d'explorer sans idée arrêtée ou de rechercher un contenu précis) . Cette méthode cherche à mettre en œuvre quelques caractéristiques de dialogues humain :

- **Souplesse de la requête** (aucun champ obligatoire). Tel un vendeur, Wishbone ne va pas demander systématiquement le prix du logement que l'acheteur est prêt à investir s'il n'est pas fixé, et sera cependant capable de lui fournir une réponse.
- **Simplicité de l'interface** (l'accès à des champs plus précis se fait dynamiquement, conformément au modèle objet de la base de données). Lorsqu'un acheteur vient pour se renseigner au sujet d'une maison, le conseiller de clientèle ne l'envahit pas tout de suite avec tous les différents types de revêtement muraux ou marque de robinets.
- **Fiabilité de la réponse quelque soit la requête.** Quelle que soit la requête, trop floue ou trop exigeante, Wishbone ne renvoie pas « d'écran vide » qui entraîne l'utilisateur dans une voie sans issue, et organise de façon synthétique sa réponse. L'acheteur qui rentre dans un cabinet de vente immobilière en n'ayant aucune idée du logement qui l'intéresse, obtiendra un aperçu clair de l'offre, même si celle-ci englobe tout le parc immobilier. De même, s'il se montre trop exigeant, un bon vendeur saura le re-aiguiller ; c'est ce que réalise Wishbone en regroupant des réponses trop nombreuses, ou en proposant des compromis lorsqu'il ne trouve pas de solutions. Nous avons donc permis de retranscrire la modification du comportement (synthétique, précis ou mode compromis) que peut avoir l'interlocuteur face à une demande qui ne correspond pas forcément au cas le plus facile, c'est-à-dire le cas où il n'y a qu'un petit nombre de réponses facilement exploitables.

Ce travail est un travail en cours, et un véritable protocole expérimental de mesure de satisfaction des internautes et de son impact sur le site reste à faire.

Références

[LARDY et al, 1992] LARDY J.-P., HERZHAFT L. (8-10 December 1992). *Bibliographic treatments according to bibliographic errors and data heterogeneity: the end-user point of view*. In Online Information 92. 17th International Online Information Meeting,. London: Learned Information, 1992. p. 547-556

[Höök 1997]. Höök, K. (March 1997) *Steps to take before IUI becomes real* The reality of intelligent interface technology, Edinburgh.

[Axance, 2001] Axance (june 2001), *The Web as a pre-buying tool*, study on e-commerce

[Pu et al, 2000] P. Pu, B. Faltings (2000), *Enriching buyers' experiences : the SmartClient approach,* CHI :289-296, 2000.

[Caseau and Laburthe, 2002] Caseau, Y. and Laburthe, F. Bouygues S.A. (2002) *Method for the data-driven adjustment of queries over a database and system for implementing the method*, European patent request 02290920.4

[Berrien *et al*, 2003] Berrien I., Laburthe F and Ruvini J. D. (2003). *Interaction et navigation pour une base documentaire : étude d'un cas concret,* JFT 2003 (submitted).

Journées Francophones de la Toile - JFT'2003

Propriétés du Web

Inscription sociale des outils d'observation massive des mots et des liens : quelques pistes.

ALAIN LELU
INRA / Unité Mathématiques, Informatique, Génomique
Université de Franche-Comté / LASELDI
alain.lelu@univ-fcomte.fr

Résumé

Les outils d'observation massive des mots et des liens sur Internet, mal connus en définitive, et en plein développement, commencent à trouver leurs formes d'inscription dans la réalité sociale. Nous dressons un rapide panorama de ces " optiques pour photographier le virtuel ", et livrons quelques pistes de réflexion quant aux évolutions possibles de leur inscription sociale.

Abstract

Mass observation tools for word associations and Web links are in full expansion and not so well-understood. To our eyes, a clear view of this growing field is essential for thinking about their present and future social impacts : we first set out a state-of-the-art of these "lens for viewing the virtual world", then we sketch a few directions about the process of their social rooting.

1 Introduction

Chercheur dans le domaine de la fouille de textes [Lelu, 1994] et concepteur du logiciel NeuroNav [Lelu et Aubin, 2001], nous avons publié en 2002 un état de l'art sur les filtrages et synthèses d'information à grande échelle qu'autorise désormais l'existence d'Internet [Lelu, 2002] qui nous a suggéré quelques réflexions prospectives sur leurs impacts sociaux ; nous livrons ici de façon condensée cet état de l'art, à nos yeux préalable à toute réflexion sérieuse, et quelques pistes dans cette direction.

Le succès d'Internet est dû, pour beaucoup, à ce qu'il permet tout à la fois la communication en temps réel, en temps différé, et l'édition des écrits, du son et de l'image. C'est donc la capitalisation de ces formes d'expression et leur mise à disposition du public qui se trouvent désormais à la portée de (presque) tous. L'exploitation de ces flux et de ces stocks d'information le plus souvent très informelle a attiré en premier lieu l'attention des services de renseignements, en particulier de la NSA (National Security Agency), l'agence de renseignement militaire américaine. Le financement d'équipes de recherche par cet organisme, la révélation du réseau d'écoutes planétaire Echelon, et de la puissance informatique concentrée à Fort Meade pour le traitement de ces flux, ont montré que beaucoup de techniques de filtrage et de synthèse d'informations, dont certaines auraient pu relever il y a peu de la science-fiction, sont d'ores et déjà mises en œuvre. L'utilisation de telles techniques constitue un enjeu considérable pour les institutions civiles, les entreprises et les particuliers, que ce soit pour chercher des informations sur un sujet précis, pour préciser ce qui est connu, ou pour dégager ce que l'on ne connaît pas, ou mal, sur des thèmes généraux. Le passage du filtrage de corpus textuels volumineux par mots-clés (« word spotting ») au filtrage par thèmes (« topic spotting »), améliore le recueil d'informations bien ciblées. Les progrès de la détection automatique de thèmes, cette fois en vue d'en

tirer des synthèses interprétables, est gros de conséquences culturelles, sociales, politiques, que nous entrevoyons à peine, mais qui accompagneront de façon incontournable, pour le meilleur et pour le pire, le monde « branché » qui se dessine sous nos yeux.

2 Exploiter les informations structurées explicites

2.1 Le filtrage par les « en-têtes »

Les informations circulant sur Internet sont constituées de parties non structurées (texte libre, images ou sons, programmes compilés) insérées dans une enveloppe structurée : adresses de départ et de destination, dates, volume, type physique… Exploiter ces données d'« en-têtes » est le B-A-BA de l'organisation et de la recherche d'information, ce que font tous les logiciels de messagerie et de participation à des forums (« newsgroups ») en nous proposant des listes de messages triées par date, par auteur, par destinataire, etc.

Les services de renseignement ne manquent pas d'exploiter cette ressource et, compte tenu de l'énormité des flux à surveiller, se basent en premier lieu sur la bonne vieille méthode d'écoute des adresses suspectes : Echelon semble intercepter la totalité des messages transitant par satellites - Internet et téléphone confondus - et la lecture ou l'écoute directe, humaine, des communications suspectes est la façon la plus fiable de recueillir de l'information. Le tri à réaliser pour rester dans des limites budgétaires supportables semble au plus de 1 pour 1 million de messages. D'où l'intérêt d'autres méthodes de filtrage, plus automatisées et plus adaptées à un sondage exhaustif de flux.

Une autre stratégie tentée par le FBI consiste à rendre obligatoire l'installation, chez chaque prestataire de messagerie aux USA, du logiciel Carnivore permettant de mettre sur écoute certains comptes, sur décision judiciaire – idée qui a provoqué une forte émotion devant les abus possibles et le manque de contrôle effectif du dispositif.

2.2 Exploiter les liens entre pages Web

De quoi la Toile est-elle tissée ? Le vocable « la Toile » (*The Web*) provient des liens dont sont pourvues les pages, liens qui les rattachent à d'autres pages ou à d'autres points d'ancrage dans ces mêmes pages. L'ensemble de ces liens forme un enchevêtrement, un feutre plus qu'un tissu régulier, comparable aux liens de citation qu'on trouve dans les articles scientifiques et qui traduisent la reconnaissance par l'auteur d'un certain nombre de travaux antérieurs comme références appropriées - que ce soit pour les citer comme étapes de son cheminement ou pour les critiquer.

Dès lors que l'informatique permet de collationner « quelque part » - par exemple dans un serveur documentaire de *pre-prints* ou dans un moteur de recherche Web - tous les liens issus d'un grand nombre d'unités documentaires (textes et descriptions documentaires, pages Web…), il devient possible d'enrichir le point de vue du surfeur à la recherche d'informations : sur le moteur Google, tout ensemble de pages obtenu à partir d'une requête classique par mot(s) présente en tête de liste les pages les plus « populaires » ou reconnues, vers lesquelles pointent le plus de liens. De même le serveur de pre-prints et publications auto-archivées [CiteSeer] donne accès, pour chaque article, à de nombreux liens directs ou calculés : articles cités, mais aussi liens calculés à partir des articles citants, des articles partageant les mêmes citations (« co-citations »), ou à partir des mots ou expressions communs dans les résumés (cf. plus bas), …

Ces possibilités d'exploiter de façon « panoptique » les liens hypertextuels nous montrent qu'il est faux de considérer le World Wide Web comme un support d'édition libre, soustrait à tout phénomène de validation par les pairs : n'importe qui peut publier à peu près n'importe quoi, mais peu de gens y accèderont si cette œuvre n'est pas référée dans les pages émanant au moins d'un micro-milieu partageant un micro-point de vue sur un micro-sujet ! « Se faire référer », et pas seulement par les moteurs Web, est un enjeu essentiel pour qui publie un site ; la marche initiale à franchir est moins élevée que sur support papier, mais la logique éditoriale de la circulation de la reconnaissance entre pairs y joue tout autant. Le Web est ainsi propice à propice à l'éclosion et au renforcement d'innombrables communautés d'intérêts, passionnés de numismatique comme militants rassemblés par un événement politique. Il paraît évident que les services de renseignements et de police utilisent les liens pour cerner les contours de « cliques » qui peuvent se référencer mutuellement sur des sujets

sensibles, tout en employant un langage codé (hackers, terrorisme biologique ou nucléaire, nazis, pédophiles…).

Les liens hypertextes forment le tissu direct et explicite de la Toile ; mais bien d'autres types de liens peuvent être mis à jour pour qui dispose de la capacité de traiter le contenu d'un grand nombre de textes : ce sont les liens calculés, ou déduits, implicites.

3 Calculer des liens à partir du contenu des textes sur Internet

Pour peu qu'on se trouve en mesure de définir un indice numérique de ressemblance entre deux documents, une génération automatique de liens hypertextuels s'en déduit : il suffit de calculer les valeurs de cet indice entre un document donné et tous les autres, et de décider d'un seuil de valeur au-dessus duquel on considérera que des liens existent. En général on présente à l'utilisateur la liste, par ordre de valeurs de liens décroissantes, des documents liés, et celui-ci décide de les explorer, ou pas, en progressant plus ou moins dans cette liste.

3.1 Liens calculés à partir du vocabulaire partagé par chaque paire de pages

Compter le nombre de mots communs à deux textes est une façon aisée de les rapprocher globalement : s'il n'y en a aucun, ces textes n'ont clairement rien à voir entre eux, s'ils y en a beaucoup, alors ils parlent de sujets voisins. A partir de cette idée simple, voire simpliste, un bon nombre d'indices de ressemblance entre textes ont été proposés pour compenser les inévitables inconvénients de cette idée de base. Le « cosinus TF-IDF » de Salton en est le plus répandu, et répond à deux exigences :1) Donner moins d'importance aux mots fréquents, communs à la majorité des documents, et qui n'apportent quasiment aucune discrimination entre ceux-ci. 2) Ne pas dépendre de la taille de chaque texte : l'indicateur de Salton est normalisé, entre 0% et 100%, pour ne pas favoriser les grands textes au détriment des petits. Un tel indicateur demande des comptages et des calculs numériques qui le lient au modèle vectoriel en documentation, dont Gerald Salton fut le héraut depuis les années 1960.

Un certain nombre de moteurs de recherche documentaire et sur le Web incluent cette fonction de recherche de documents similaires (*similarity ranking*), parfois aussi dénommée *relevance feed-back* (retour de pertinence) dans la mesure où l'utilisateur, confonté à une liste de documents issus d'une requête précédente, en choisit un qu'il considère comme pertinent, et dont les proches risquent fort de l'être également [ex. : MEDLINE]. En d'autres termes, on fait suivre une requête booléenne, nécessairement limitée, par une « requête-document » beaucoup plus riche en mots potentiellement pertinents.

Plus radicalement, dès lors qu'on est libéré des liens directs, explicites, entre documents, rien n'empêche de considérer un ensemble de textes indexés (c'est à dire décrits par des mots) comme un réseau de mots : chaque mot étant caractérisé par ses fréquences dans tous les documents où il est présent, le même type d'indicateur de ressemblance vu plus haut peut être défini, mais cette fois entre deux mots, créant de cette façon des « couronnes » de mots proches d'un mot donné. Cette opération, possible dans quelques systèmes documentaires (elle s'appelle parfois *Zoom*), a été implantée dans le passé sur les moteurs Excite et AltaVista, aujourd'hui sur <*www.influo.com*> ; elle a le mérite de suggérer à l'utilisateur des mots pertinents présents dans le corpus qui ne lui venaient pas à l'esprit – ce problème du décalage entre vocabulaire de l'utilisateur et vocabulaire du corpus est reconnu comme une des difficultés récurrentes de la recherche d'information[i].

A noter que la qualité, la pertinence, de ces liens calculés dépend en ligne directe de la qualité de l'indexation : si celle-ci se contente d'élever chaque chaîne de caractères à la dignité de mot-clé (indexation *full text*, en texte intégral), beaucoup de « bruit » viendra polluer les exploitations ultérieures. La qualité de celles-ci sera améliorée par une analyse morpho-syntaxique du corpus permettant de dégager les formes normalisées des mots, ou *lemmes* (verbes rapportés à l'infinitif, adjectifs au masculin singulier, …), et surtout de dégager les termes composés, véritables atomes sémantiques, les plus proches d'une indexation idéale par d'authentiques concepts. Pour un exemple de moteur Web utilisant un traitement linguistique, voir [PERTIMM].

3.2 Autres types de liens calculés

Caractériser un texte par un profil de fréquences de mots n'est pas la seule façon de le décrire : plusieurs travaux [Damashek, 1995][Lelu, 1998] ont montré qu'il était aussi possible de le caractériser par un profil de fréquences de n-grammes[ii], ce qui rend cette technique indépendante de la langue et du type de codage des caractères, donc particulièrement appropriée pour les langues asiatiques, sans séparateurs de mots. La qualité des représentations synthétiques obtenues dépend de la qualité des filtrages statistiques réalisés en amont, question toujours ouverte qui seule permettra de rivaliser avec la qualité obtenue à partir de termes lemmatisés et filtrés. On dégage alors par des « surlignages flous » des termes simples et composés candidats au statut de mots d'index, caractéristiques de chaque texte, qui en présentent, en quelque sorte, un résumé thématique.

Les pages multimédias peuvent aussi être à l'origine d'autres liens calculés : par exemple, ceux déduits sur la base d'indicateurs numériques de couleurs, de textures ou de formes sur les images [ex. : LTUtech].

4 Filtrer l'information

Le schéma dominant de la recherche d'information sur Internet est celui de la requête, c'est à dire d'une question que l'on pose, qu'un système automatique « comprend » et à laquelle il fournit une réponse. Le schème alternatif du butinage (*browsing*) au gré des liens entre pages, bien que séduisant au départ et homologue à la démarche d'associations d'idées prônée par le pionnier de l'hypertexte Vannevar Bush, donne lieu à en pratique à des séances de dérive étourdissantes, vertigineuses, dont on ressort l'esprit broyé, malaxé, avec un sentiment de difficulté à capitaliser, à faire un bilan ; l'internaute de base a du mal à passer maître dans le maniement des fragiles traces collectées par les navigateurs, et dans la gestion des listes de favoris, toujours à la merci d'un reparamétrage intempestif, d'une réinstallation de logiciels ou d'un changement de configuration ! En l'absence de repères solides pour répondre aux questions - Où suis-je ? Saurais-je plus tard revenir à tel endroit où je me trouvais tout à l'heure ? Vers où me diriger ? – et en attendant la généralisation des moteurs et métamoteurs cartographiques comme <www.kartoo.com>, <www.mapstan.com>, la moins mauvaise métaphore reste peut-être aujourd'hui celle du dialogue avec la machine, du couple requête/ réponse au moyen de mots répertoriés par la machine et signifiants pour l'homme. On se souvient qu'en posant telle requête sur tel moteur on retrouvera sans coup férir ce qui nous avait intéressé il y a quelque temps.

Le filtrage est la généralisation de ce schéma à un flux de données : une fois une requête mise au point, on l'applique à intervalle régulier au nouvel état d'un stock d'informations, ou à un flux de messages. Les documentalistes appellent cela depuis longtemps « diffusion sélective d'information » , et les informaticiens l'ont remis au goût du jour sous le vocable d'agents un peu abusivement qualifiés d'intelligents.

4.1 Filtrage par localisation de mots (*word-spotting*)

Le but du filtrage est de recueillir un maximum de documents sémantiquement homogènes, répondant à une problématique claire du point de vue de la compréhension humaine. Or le langage est truffé de pièges sémantiques – polysémies, homonymies, synonymies – qui ne sont pas des exceptions, mais la règle dans toutes les langues. C'est en effet le contexte sémantique qui est l'élément de base, commun et fédérateur, dans la communication humaine, et qui permet de résoudre sans problème ces difficultés, qui n'en sont que pour l'ordinateur, voué par construction à ne traiter que des codes, ou chaînes de caractères, définis de façon univoque, et indifférent à la notion de contexte. En informatique, le chemin est long de la forme graphique d'un mot à son sens ; il est jalonné par plusieurs générations de techniques, utilisées pour commencer à prendre en compte ces questions.

4.1.1 Techniques basées sur les chaînes de caractères

Historiquement, l'assimilation abusive 1 mot = 1 chaîne de caractères a été la première façon et la plus simple de prendre en compte le texte intégral sur un ordinateur. Elle le reste aujourd'hui dans l'écrasante majorité des moteurs de recherche Web ; parmi ceux-ci, seuls ceux dotés de capacités de reconnaissance de la langue des pages sont en mesure d'éliminer les mots grammaticaux comme *le* ou

the, au moyen d'anti-dictionnaires. Grâce à des opérateurs de troncature permettant d'ignorer les terminaisons, il est possible de composer des requêtes booléennes et d'adjacence (comme : *bases NEAR données*) qui reportent sur l'auteur de la requête le poids de la prise en considération de la variation syntaxique, et des compromis bruit documentaire / silence qui en résulteront. Il semble que le réseau Echelon fasse un usage intensif de ces techniques pour traiter les flux de messages interceptés.

4.1.2 Traitements linguistiques

La raison pour laquelle le repérage « bête » de chaînes de caractères reste prédominant est que la majorité des pages Web et des messages électroniques ne sont pas de vrais textes, conformes, même de loin, aux canons de la langue, sans parler de la multiplicité des langues autres que l'anglais, de plus en plus présentes sur ce média. Par contre les dépêches d'agence, les articles de journaux, les résumés bibliographiques - qu'on trouve aussi sur Internet -, s'en rapprochent davantage, et il est alors possible d'en faire l'analyse morpho-syntaxique pour étiqueter grammaticalement chaque mot, le ramener à sa forme normalisée (lemme), et repérer sur une base statistique ou linguistique les termes composés. Certains moteurs de recherche documentaires dédiés à la veille en entreprise (veille stratégique, technique, concurrencielle, d'image) et quelques rares moteurs Web [NOMINO][PERTIMM] font appel à de telles techniques, à des degrés très disparates d'élaboration et de fiabilité. Elles accroissent sans conteste la qualité et la précision de l'indexation, mais peuvent aussi engendrer un bruit de nature différente de celui de l'indexation *full text*. Là aussi le compromis est difficile entre 1) la tolérance aux fautes d'orthographe et de syntaxe, aux néologismes, 2) la fiabilité de l'analyse syntaxique (beaucoup de cas ne se résolvent que par le contexte sémantique, généralement hors de portée), et 3) la puissance de calcul nécessaire. Le filtrage est alors plus simple à définir et de meilleure qualité que sur l'indexation en texte intégral, mais ici aussi le mot n'est pas le concept, et bien des « faux négatifs » passent entre les mailles, alors que beaucoup de « faux positifs » sont recueillis.

4.1.3 Traitements linguistiques + sémantiques

Les chercheurs en intelligence artificielle symbolique n'ont jamais cessé de caresser le rève de traduire le langage naturel en une langue « idéale » où les mots auraient un sens univoque et parviendraient à décrire les concepts et relations entre concepts sous-jacents à tout texte ; en d'autres termes, de passer des données à la connaissance. Parmi les réalisations, celle qui semble à notre connaissance la plus opérationnelle à grande échelle est due à une équipe française, celle de Christian Krumeich, à l'origine chez Thomson, qui a mis au point pour le renseignement militaire français (DGSE et DRM) le logiciel Taïga (Traitement automatique de l'information géopolitique d'actualité), conçu à l'origine pour dépouiller les bases documentaires russes au moment de la Pérestroïka [Krumeich, 1994]. Quelques dizaines de postes de travail sont en fonctionnement, pour dépouiller des sources structurées, homogènes et rédigées en toutes langues, après un travail considérable de formalisation des connaissances, pour alimenter son moteur sémantique, dans des domaines bien définis comme la construction aéronautique ou la prospection pétrolière. Outre ce travail important que peu d'entreprises ou d'institutions peuvent se permettre de financer, cette technique semble peu compatible avec le filtrage à grande échelle de messages informels, comme le réalise Echelon.

4.2 Filtrage par localisation de contextes sémantiques pré-définis (*topic-spotting*)

Dès lors que le sens d'un mot dépend de son contexte, l'objectif de localisation de chaque atome de sens derrière chaque mot paraît peu réaliste dans l'état de l'art actuel. Il paraît plus raisonnable, et de toute façon satisfaisant pour une large palette d'applications, de se rabattre sur l'objectif, tout compte fait moins ambitieux, de repérage de *contextes* sémantiques, c'est à dire d'ensemble de textes qui en gros parlent de la même chose, du même sujet ; ce sujet peut impliquer un grand nombre de mots qui chacun peuvent intervenir dans d'autres contextes, mais que seul le contexte de ce sujet précis met en relation, fait intervenir ensemble. C'est ainsi qu'on peut « piéger » un contexte sémantique, et c'est plus facile qu'il n'y paraît : il suffit de rassembler un nombre suffisant de textes en rapport, de *notre* point de vue humain, avec ce sujet, et qui en présentent un échantillon significatif des expressions possibles. A partir de là, plusieurs approches sont réalisables, selon la façon dont on représente les textes, dont on les abstrait :

- On peut les représenter par des profils de mots. A notre connaissance, les applications de cette approche concernent le problème du classement (en anglais : *classification*) de documents textuels, c'est à dire de l'attribution automatique à tout nouveau document répertorié d'une catégorie prise dans une liste pré-définie, par exemple l'attribution à une page Web d'une rubrique dans l'arborescence d'un guide d'information en ligne.

De nombreuses méthodes statistiques, souvent formalisées en tant que modèles neuronaux, sont disponibles pour résoudre ce problème dit de « discrimination », linéaire ou non (on parle aussi d'*apprentissage supervisé*). C'est ainsi qu'à partir d'une base initiale de pages Web classées manuellement dans les catégories de son guide en ligne, le moteur de recherche Voilà fait appel à une technique issue de France Télécom Développement (ex-CNET) pour assurer le classement du flux des nouvelles pages répertoriées, plutôt que de confier cette tâche, sous le contrôle indispensable – et coûteux - de son personnel documentaliste, aux éditeurs de sites désirant se faire répertorier comme le font la plupart des autres guides en ligne.

- On peut représenter les textes par leurs profils de n-grammes (cf. note plus haut).

Nous avons connaissance d'une seule application de ce type, mais elle est importante puisqu'elle semble mise en œuvre au sein du réseau Echelon. Elle consiste à faire définir par un analyste le sujet qui l'intéresse sous forme d'un ensemble de textes, puis à comparer au profil global de n-grammes de cet ensemble les profils de tous les textes balayés, pour ne retenir que ceux dont l'indicateur de similarité est supérieur à un certain seuil. Cette technique a été publiée en 1995 [DAM 95] et un brevet a été pris par la NSA. Elle présente l'avantage considérable de ne dépendre ni de la langue, ni de l'écriture, et d'être robuste face aux fautes diverses, coquilles et autres scories de débalisage des textes : après constitution dans une langue donnée d'une base d'exemples par un expert, les textes japonais ou chinois sont filtrés aussi efficacement que les textes anglais.

Le succès des approches supervisées de détection de thèmes que nous venons de passer en revue repose sur la qualité du travail de catégorisation opéré par les analystes qui les initialisent, par regroupement de documents autour d'un thème ou mise à jour d'un dictionnaire sémantique. Mais ces opérateurs peuvent 1) faire une mauvaise analyse, c'est à dire vouloir retrouver dans les données des concepts vagues ou à la mode qui n'y sont pas, 2) faire une bonne analyse, mais disposer d'un ensemble de textes insuffisant qualitativement ou quantitativement pour en rendre compte. D'où l'intérêt, pour qui ne s'intéresse pas à des thématiques trop ponctuelles, d'une approche *non supervisée*, où un processus automatique repère les groupements « objectifs » de documents, les zones de forte densité de l'espace des données[iii].

5 Faire émerger l'essentiel, sans idées préconçues, à partir d'un enchevêtrement de liens

Quand on s'intéresse aux liens *explicites* entre pages Web, et qu'on s'aperçoit que la structure des liens sur un site de taille moyenne devient vite inextricable, la première idée qui vient à l'esprit pour en offrir une vue d'ensemble est de dessiner pour l'utilisateur, sur le plan de l'écran, un graphe représentant les pages (« nœuds ») par des ronds ou des rectangles, et les liens par des arcs entre ces nœuds. Très vite se posent des problèmes de présentation optimale, comme disposer les nœuds de façon à ce que les arcs évitent au maximum de se croiser, ou disposer les libellés des nœuds en évitant au maximum qu'ils ne se recouvrent. Mais dès que le nombre de nœuds dépasse quelques dizaines, malgré les trésors de « design » dépensés jusqu'ici (nœuds et libellés déplaçables, utilisation de la couleur, de la perspective, …), le graphe devient illisible et d'autres techniques de visualisation doivent prendre la relève, comme la possibilité de centrer une vue « fish-eye » sur un nœud particulier quand on clique dessus – les relations de voisinage de ce nœud apparaissent alors clairement, tandis que les relations entre voisins de plus en plus éloignés se trouvent " tassées " à la périphérie du graphe.

Représenter de façon accessible des graphes immenses comme ceux qu'on peut définir sur le Web est un domaine de recherche en soi, mais se heurte toujours aux limites physiques de l'écran : une solution pour en sortir est de passer à un niveau supérieur de synthèse et d'abstraction, c'est à dire de grouper les nœuds par paquets, ou classes, de rôles semblables ; les nœuds d'une classe pointant vers (ou étant pointés par) *grosso modo* le même ensemble de voisins. En termes plus formels, ceci revient

à classer – sans superviseur, c'est à dire sans idée préconçue – les éléments dont les relations sont décrites par un tableau carré croisant ces éléments avec eux-mêmes ; la force de chaque relation est alors traduite par un nombre à l'intersection d'une ligne et d'une colonne, par exemple un « 1 » si telle page pointe vers telle autre, un « zéro » sinon.

Peu de travaux sur de telles représentations de faisceaux de liens explicites sur le Web semblent avoir débouché de façon opérationnelle ; par contre rien n'empêche de les envisager quand le tableau carré évoqué ci-dessus traduit des liens *calculés* : par exemple, à partir d'un tableau de co-occurrences de mots dans un ensemble de pages on peut tirer des classes de mots homogènes, traduisant les grands thèmes dont il est question dans l'ensemble des pages. La méthode des mots associés [Courtial 1996] qui fonde le logiciel Sampler, ou a fondé la fonction AltaVista / Refine, part d'un tel principe ; elle place les classes de mots les unes par rapport aux autres sur un graphe global, lisible et aéré, et les mots eux-mêmes sont les noeuds du graphe local de chaque classe ; l'épaisseur des traits entre classes ou mots traduit la force du lien d'apparition dans les mêmes pages qui les unit.

Bien d'autres méthodes traitent de tels tableaux carrés de co-occurrence de mots. Mais l'information qu'ils véhiculent peut être extraite directement du tableau rectangulaire brut (documents × mots) répertoriant la présence ou la fréquence de chaque mot dans chaque document, qui a servi à le construire. Ce tableau sert de matière première à deux grandes familles de méthodes, dites directes, de synthèse d'informations, ainsi qu'à une famille hybride :

- *Méthodes factorielles* : un ensemble de textes décrits par des fréquences de mots (ou de n-grammes) peut être considéré techniquement, même si c'est inaccessible à notre intuition, comme un nuage de points, représentant chacun un texte, dans un espace comportant autant de dimensions que de mots (ou de n-grammes) différents répertoriés. Le principe de l'analyse factorielle consiste à réduire ce nombre de dimensions en « photographiant » le nuage de façon optimale, c'est à dire en perdant le moins d'information possible quant à la forme générale du nuage, en respectant au mieux les distances originelles entre points. Pour prendre une analogie triviale, une girafe, « objet » défini dans 3 dimensions, sera mieux caractérisée sur une photo, c'est à dire en deux dimensions, si on la saisit de profil plutôt qu'en vue de dessus ou de devant, et pour réduire la surface de cette photo, en découpant celle-ci parallèlement à « l'axe principal » de l'animal, à savoir son cou… L'inconvénient principal de la méthode découle de cette représentation : dès que l'on a plusieurs milliers de documents et de mots, ce qui arrive très vite sur Internet, les cartes deviennent illisibles. Une autre façon d'opérer est de ne pas chercher à interpréter les nouveaux axes de coordonnées obtenus, et de se contenter d'utiliser la propriété de réduction du nombre de dimensions, afin de se situer dans un espace sémantiquement significatif et dégagé de la part de bruit et d'arbitraire inhérente au langage naturel : c'est ce que réalise la méthode Latent Semantic Analysis <http://lsa.colorado.edu>.

- *Méthodes de classification automatique* : les méthodes de classification automatiques dégagent d'elles-mêmes, sans qu'on ait besoin de leur injecter de connaissances sur les « bonnes » classes à détecter, des ensembles de documents homogènes, en maximisant l'homogénéïté interne des classes et l'hétérogénéïté des classes entre elles, du point de vue du vocabulaire employé. Le modèle neuronal dit carte auto-organisatrice (Self-Organizing Map) de T. Kohonen <http://websom.hut.fi/websom> permet de réaliser simultanément une classification des documents et un placement des classes sur une grille représentant un pavage régulier de « neurones » incarnant les classes. Pour une application à la navigation dans une base bibliographique sur Internet, voir le site < http://simbad.u-strasbg.fr/A+A/map.pl>.

- *Méthodes mixtes* : dans notre propre ligne de recherche, nous avons développé des méthodes intermédiaires entre l'analyse factorielle et la classification automatique, qu'on peut décrire comme des méthodes de classification floue et recouvrante, où chaque thème regroupe des documents homogènes dotés chacun d'une valeur de « typicité », de centralité dans ce thème, ainsi que de mots caractéristiques de ce thème, eux aussi avec une valeur de centralité. Cette représentation est apte à la traduction d'effets de contexte (cf. www.upmf-grenoble.fr/adest/seminaires/diatopie.ppt).

6 Des synthèses pour quoi faire ?

Pour qui a les moyens de « visualiser les masses de liens », des réseaux sociaux deviennent lisibles sur le Net ; à la différence des réseaux de la socialité habituelle, la proximité physique, géographique peut y être remplacée - ou traduite - par celle directement exprimée par les liens hypertextes, mais aussi par la (ou les) proximité(s) déduite(s) par ces observatoires du virtuel que sont les moteurs de recherche, qui se dotent petit à petit d' « instruments d'optique » de plus en plus puissants : Google donne les meilleures pages de référence sur un sujet donné, Kartoo < www.kartoo.com > nous rend évidents, par paquets réduits d'une quinzaine de pages, les liens d'association de vocabulaire au sein de chaque paquet. Mais des dizaines de prototypes en gestation dans les laboratoires sont en train de définir les contours des « optiques » de la prochaine génération, celles qui nous permettront de combiner les bons filtres, le bon tissu à observer (liens explicites ou implicites ?), et le bon grossissement, tous paramètres que nous apprendrons à adapter itérativement à nos objectifs de recherche d'information.

Une autre voie possible pour l'exploitation de ce tissu de liens nous est suggérée par le développement de la scientométrie, ou étude de la vie des courants scientifiques et techniques à partir des fiches bibliographiques stockées dans les grandes bases documentaires mondiales (articles scientifiques, brevets…) : depuis les 2 ou 3 décennies que ces bases existent sur support électronique, elles constituent la matière première de l'observation du mouvement vivant des sciences et techniques par des sociologues [Polanco, 1996] et des chargés d'études auprès des institutions où se décident (ou s'infléchissent) les politiques scientifiques (Ministères, grands organismes de recherche…). Pour les sociologues du Centre de Sociologie de l'Innovation [Callon, 1989] les articles scientifiques et les brevets représentent bien plus que des traces de l'activité scientifique et technique : ils en sont les objets-acteurs principaux - les citations et le vocabulaire employé, les « mots-bannières », créent la dynamique d'alliance et d'exclusion entre « acteurs-réseaux » humains et non-humains. C'est dans et pour ce milieu que sont apparues en premier les méthodes de synthèse d'informations, de représentations cartographiques, appliquées à la même époque aux USA aux liens de co-citation entre articles, et en France puis aux Pays-Bas aux liens calculés à partir du vocabulaire d'indexation [Courtial, 1996][cf. aussi <www.cwts.nl/ed>].

Si l'on extrapole à la Toile, la situation nouvelle dans laquelle il sera possible de repérer et d'explorer les zones de forte densité d'entrelacs de liens, liens tant explicites que calculés – service fourni a priori par les moteurs ou méta-moteurs de recherche -, sera largement inédite : en effet ces zones, aujourd'hui largement invisibles, marquent la condensation de processus collectifs, mis de façon volontaire et en connaissance de cause sur la vaste place publique que constitue le Web par les individus sociaux que nous sommes. Les outils que nous avons brièvement mentionnés permettraient alors l'exploration et le parcours de cette « géographie virtuelle » et mouvante (pour une veille sur ce domaine, voir le site < www.cybergeography.org>).

Une réflexion voisine, bien que distincte par son objet, est menée actuellement par les milieux qui explorent le concept de « mémétique », ou métaphore de l'évolution génétique appliquée au domaine des idées [www.cpm.mmu.ac.uk/jom-emit]. Le concept de « Global Brain » issu de ce courant, propose la métaphore de l'entrelac des liens du Web comparé à celui des neurones, liés dans la matière grise par les axones et leurs ramifications dendritiques. Ces pistes sont intéressantes, et sans doute riches de progrès possibles pour les sciences humaines - contribueront-elles à dégager des « observables » dont l'analyse puisse être susceptible de réfutation ? Elles débouchent parfois sur la généralisation, prématurée à notre avis, des conséquences de certains concepts considérés comme acquis, souvent dans une optique de critique sociale : ainsi [Bollen et Heylighen, 1996] voient dans le Global Brain la naissance d'une pensée conformiste à l'échelle planétaire – alors que d'autres, y compris parmi les proches de la mémétique, y voient au contraire un émiettement en de multiples particularismes s'ignorant les uns des autres, et le renforcement d'une tribalisation générale.

Loin de ces spéculations à long terme, où les auteurs de science-fiction nous paraissent les mieux placés pour forcer le trait dans la description, apocalyptique ou non, de mondes possibles [ex. : Truong, 2001], nous nous contenterons de questions plus circonscrites. En effet, au-delà des opérations ponctuelles de « Web-mining », la typologie des pages Web [Borzic, 1998], leur

catégorisation automatique à partir de leurs liens ou de leurs contenus par des neurones artificiels sont des opérations à visée « panoramique » qui seront bientôt possibles dans la pratique quotidienne du Web, ne nécessitant qu'un changement d'échelle par rapport aux prototypes des laboratoires. Sous quelles formes se réalisera leur assimilation sociale ? Serviront-elles en premier lieu, comme d'autres avant elles, aux objectifs des services de renseignements, si tant est que ceux-ci se dotent de la culture permettant d'en tirer parti ? Ou bien, dans le prolongement de la scientométrie actuelle, fourniront-elles d'abord des techniques de « photographie du virtuel », matière première pour des analyses (enfin) réfutables, et bases d'une capitalisation d'acquis à la fois théoriques et empiriques dans les sciences humaines et sociales ? Apparaîtront-elles publiquement en priorité en tant qu'outils pour la consultation de collections homogènes et spécialisées sur le Web, comme les revues électroniques ou les bases de *preprints*, ou encore les interfaces vers les bases de données du Web invisible, comme Simbad mentionnée plus haut ? Ou bien leur utilisation par monsieur Tout-le-monde dans des moteurs de recherche, qui sont actuellement les mieux placés pour les proposer, sera-t-il au contraire l'événement déclencheur des autres usages, comme pourrait le préfigurer le succès de Kartoo (cartographie à partir des mots à l'échelle restreinte d'une quinzaine de pages-réponses) ? Autant de questions auxquelles nous ne nous hasarderons pas à répondre, échaudés par le constat de la lenteur de pénétration de ces techniques de *text-mining* (ou fouille de textes) dans les entreprises depuis une quinzaine d'années, malgré la généralisation de l'information sous forme électronique, malgré les discours incessants sur l'excès d'informations, sur la nécessité d'extraire et gérer la connaissance, sur la mémoire et les savoir-faire d'entreprise à préserver… Processus jalonné de multiples faillites ou difficultés chroniques des entreprises actuellement sur ce créneau de logiciels et services, processus incertain posant peut-être question de la place de la réflexion au-delà du court terme et des sciences sociales dans l'entreprise : la pente naturelle y est d'attendre des synthèses presse-bouton, des réponses aux questions posées dans la langue de bois à la mode à l'instant donné (bien décrite dans [Volle, 2000]), toutes choses antinomiques d'un processus d'aller-retour entre élaboration des données et des problématiques, paramétrages des analyses, et interprétation des résultats pour éclairer l'action.

On peut voir dans les perspectives de catégorisation et de cartographie du Web une opportunité fantastique pour le développement des sciences sociales, qui y trouveront à la fois un gisement d'observations inespéré, autant que d'expérimentations participantes, sur les traces de la scientométrie, d'un impact sans doute considérable tant sur leurs théories que sur leurs pratiques.

On peut y voir aussi le danger de telles méthodes appliquées de façon perverse aux flux de messages personnels, pour l'instant textuels, plus tard téléphoniques et visiophoniques, dans des buts de contrôle social. L'espionnage « ponctuel » d'aujourd'hui y serait remplacé par l'espionnage des âmes… Mais l'histoire n'est jamais écrite à l'avance : toute action suscite une réaction, immédiate ou latente ; d'ajustements en ajustements, de crises en crises, les choses vont leur cours, et nul ne peut sérieusement se targuer « d'avoir tout dit » à l'avance sur l'assimilation sociale d'une technologie de l'information et de la communication.

On peut y voir enfin le support d'une boucle d'auto-régulation sociale d'un type nouveau, où les acteurs sociaux disposeraient en temps réel de cartographies de leur identité et de leurs relations, ce qui ne manquerait pas d'influer en retour sur cette identité et sur ces relations, et sur la dialectique transparence/opacité à l'oeuvre dans toute communication humaine... Mais ceci relève pour l'heure, où une faible proportion de relations sociales sont concernées par Internet, malgré son développement, de la science (sociale)-fiction !

Références

[Bollen et Heylighen, 1996] Bollen J. & Heylighen F., "Algorithms for the self-organisation of distributed, multi-user networks. Possible application to the future World Wide Web ", *Cybernetics and Systems '96*, R. Trappl ed., 1996, p. 911-916 : http://pcp.vub.ac.be/papers/SelfOrganWWW.html

[Borzic, 1998] Borzic B.,"Un modèle de gestionnaire iteratif de flux informationnel sur Internet", Thèse de doctorat, Information Scientifique et Technique, Paris, Avril 1998, CNAM/CNRS.

[Callon, 1989] Callon M., *La science et ses réseaux*, Paris, La Découverte, 1989

[CiteSeer] Base de publications en libre accès CiteSeer : http://citeseer.nj.nec.com

[Courtial 1996] J.P. Courtial, Construction des connaissances scientifiques, construction de soi et communication sociale, revue en ligne Solaris N°2, 1996 : www.info.unicaen.fr/bnum/jelec/Solaris/d02/2courtial.htm

[Damashek, 1995] Damashek M., " Gauging Similarity with n-grams: Language-Independent Categorization of Text ", *Science*, vol.267, pp.843-848, 1995.

[Krumeich, 1994] C. Krumeich, Intervention au 3e Forum de l'Intelligence Economique et Concurrentielle, Sophia-Antipolis, 8 Déc. 1994 : www.scipfrance.org/documents.htm

[Lelu, 1994] Lelu A., "Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets", *New Approaches in Classification and Data Analysis,* E. Diday, Y. Lechevallier & al. eds., pp.241-248, Springer-Verlag, Berlin, 1994

[Lelu et al., 1998] Lelu A., M. Hallab, B. Delprat, "Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de n-grammes", *Actes de JADT'98*, coord. S. Mellet, UPRESA "Bases, corpus, langages", Nice, 1998. : www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt1998/lelu.htm

[Lelu et Aubin, 2001] Lelu, A., Aubin, S., Vers un environnement complet de synthèse statistique de contenus textuels, séminaire ADEST du 13/11/2001 : www.upmf-grenoble.fr/adest/seminaires/lelu02/ADEST2001_SA_AL.htm

[Lelu, 2002] Lelu, A. – Filtrages et synthèses de masse sur Internet : état de l'art et prospective – Les Cahiers du Numérique, vol.3, N°1, pp. 171-196, Hermès, Paris, 2002

[LTUtech] Moteur de recherche d'images Image-Seeker de LTUtech SA : http://corbis.ltutech.com

[MEDLINE] Base bibliographique MEDLINE PubMed, fonction " Related articles " : http://www.ncbi.nlm.nih.gov/entrez

[NOMINO] Moteur Nomino : www.gouv.qc.ca/gouv/gouvqc/index.asp

[PERTIMM] logiciel PERTIMM, de SYSTAL S.A. : www.systal.com ; démo (12 ans de Journal Officiel) http://pertimm.ensmp.fr

[Polanco 1996] Polanco X., " Aux sources de la scientométrie ", revue en ligne Solaris N°2, 1996 www.info.unicaen.fr/bnum/jelec/Solaris/d02/2polanco.htm

[Truong, 2001] Truong, J.M. *Totalement inhumaine*. Les empêcheurs de penser en rond, Paris, 2001

[Volle, 2000] Volle M., *E-conomie*, Paris, Economica, 2000 : www.volle.com/e-conomie/table.htm

[i] Le logiciel Neuronav, décrit dans [JADT 02] et [ADE 01] généralise ces opérations vectorielles : de façon itérative, un paquet de documents pertinent est caractérisé par ses mots les plus typiques, puis une sélection parmi les mots proches d'un ou plusieurs mots intéressants élargit le cercle des mots pertinents, qui appellent alors davantage de documents à retenir, etc., jusqu'à ce que l'utilisateur ait le sentiment d'avoir " fait le tour du problème ", sans pour autant avoir consulté des milliers de pages.

[ii] Les n-grammes sont les suites de caractères de longueur fixe N (par ex., 2, 3, …) obtenues en promenant sur le texte une fenêtre de N caractères. Les premiers bigrammes de la présente note sont : *Le, es, s_, _n,* … On collationne tous les n-grammes différents dans tous les textes, et chaque texte est caractérisé par le profil des fréquences de ses n-grammes. Cette technique est très utilisée pour l'étude des " textes " génomiques (alphabet de 4 lettres) et protéomiques (alphabet de 20 lettres).

[iii] Le but étant d'extraire des thèmes " parlants " pour l'intellect, cette démarche nécessite un travail d'interprétation et de va-et-vient entre les regroupements obtenus, la lecture des plus typiques des textes regroupés, et le re-paramétrage de l'analyse, que ce soit pour en ajuster la finesse (le " grossissement ") ou pour rectifier les contours du corpus analysé – l'inverse d'une démarche " presse-bouton "…

Un modèle gravitationnel du Web

Toufik Bennouas, Mohamed Bouklit et Fabien de Montgolfier

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier,
161 rue Ada, 34392 Montpellier Cedex 5, France.
Email : {bennouas,bouklit,montgolfier}@lirmm.fr
Tél : +33 4 67 41 85 78 Fax : +33 4 67 41 85 00

Résumé

Cet article fournit un nouveau modèle du Web, permettant de détecter les cybercommunautés, de visualiser l'ensemble des pages hypertextes, et d'avoir une mesure d'audience. Il s'inspire du modèle PageRank. Les pages sont modélisées comme des particules massives, et les liens hypertextes comme des forces gravitationnelles. Nous obtenons des galaxies correspondant à des cybercommunautés dont les pages principales sont au centre.

Abstract

This article provides a new model of the Web allowing to detect cybercommunities, to visualize the set of the hypertext pages, and to provide an audiance measure. It derives from the PageRank model. The Web pages act as massive particles, while the hypertext links are modelized as gravitational forces. We obtain galaxy structures, associated to cybercommunities, which have the authoritative pages in the center.

1 Introduction

La croissance exponentielle du Web rend problématique l'appréhension de sa structure globale. Pourtant, une connaissance du contenu et de la structure du Web est indispensable pour réaliser de nombreuses tâches essentielles à la vie de l'internaute, telles que la **recherche d'information** (où trouver une page sur tel sujet ?) ou la **mesure d'audience** (ma page est-elle populaire ?).

Ces problèmes ont conduit les chercheurs à élaborer des modèles et outils divers. L'un d'eux, simple et directement inspiré de la structure hypertexte, consiste à modéliser le Web comme un graphe orienté formé par les pages Web et les liens hypertextes qui les relient [Kumar et al., 2000, Kleinberg et al., 1999]. L'analyse de ce **graphe du Web** a permis d'améliorer la performance des moteurs de recherche. Le modèle Pagerank, utilisé par Google.com, tire parti des propriétés spectrales de ce graphe pour offrir une mesure de popularité efficace (cf. section 2).

Cette étude propose un outil contribuant à la solution de trois problèmes :

- identifier les **cybercommunautés**, groupes de pages partageant le même centre d'intérêt. Il existe en effet des définitions concurrentes et plus ou moins empiriques, basées sur la sémantique, la co-citation ou des sous-graphes particuliers [Kleinberg, 1998, Gibson et al., 1998, Efe et al., 2000].
- fournir un outil de **visualisation** de la structure du Web. Appréhender un aussi vaste objet est une gageure !
- offrir une mesure d'**audience** des pages Web.

Pour ce faire, nous proposons un modèle **particulaire** : les pages Web deviennent des *particules* évoluant dans un espace tridimensionnel. Les liens hypertextes se traduisent en **forces** gravitationelles s'exerçant sur ces pages ; ainsi le mouvement de l'ensemble est-il induit par sa structure hypertexte. Enfin, l'audience d'une page donne une **masse** à la particule.

Nous nous sommes inspirés du modèle cosmologique du Big Bang [Hawking, 1988], qui décrit comment la matière, uniformément répartie dans l'univers à son commencement, a été façonnée en galaxie par deux actions : la **gravitation** et l'**expansion**. La première tend à **regrouper** les particules qu'elle lie, tandis que la seconde, dilatation de l'espace qui *diminue* à mesure que l'univers vieillit, tend à **écarter** les particules sans relation. Notre modèle adapte ces deux phénomènes au Web. Ils agissent au cours du temps et isolent les ensembles densément liés de pages, qui conservent la somme de leurs masses et se regroupent en globules. Ils nous permettent de proposer une nouvelle définition, *par émergence*, des cybercommunautés. Celles-ci sont des groupes de pages de même sujet, densément hyperliées entre elles, que leur mutuelle attraction regroupe.

L'autorité des particules fournit une autre analogie avec la masse : une page de référence se comportera comme un soleil, immobile autour d'un nuage de planètes ayant trait au même sujet. À notre connaissance, il n'existe pas de travaux antérieurs présentant le problème sous cet angle.

2 Le modèle PageRank

Lancé en 1998, le moteur de recherche Google est devenu un des plus utilisés. Une des clefs de son efficacité est le facteur *PageRank* [Page et al., 1998], un indice numérique (le «rang») qui est attribué à chaque page et reflète sa popularité. Mais comment connaître l'audience d'une page sans avoir de mesure d'accès réelle (comme des compteurs) à sa disposition ? On peut tirer parti

de la *structure hypertexte* du Web. Un lien hypertexte vers une page est interprété comme un **vote positif** en faveur de la page pointée. Cette sémantique est vraie pour une grande majorité des liens hypertextes inter-sites. Parmi plusieurs pages traitant d'un même sujet, celle ayant le plus de *liens entrants* est donc supposée être le choix des internautes rédacteurs, et *a fortiori* des internautes surfeurs. On peut ordonner les résultats d'un moteur de recherche selon le degré entrant décroissant.

Mais un tel indice est faible. Il est par exemple facile de rendre une page Web intéressante en créant plusieurs pages fictives qui pointent vers elle. Dans [Page et al., 1998], les auteurs proposent un modèle de *conservation du rang* et un algorithme permettant son calcul : PageRank. Il modélise le comportement d'un surfeur aléatoire, passant d'une page à l'autre au gré des liens hypertextes. Tous les liens sortants d'une page sont supposés équiprobables (cet axiome est discutable et nous ne l'utilisons pas). Le Web devient alors une chaîne de Markov, dont le vecteur stationnaire est la probabilité de présence de l'internaute probabiliste sur une page donnée. Cette mesure est assimilée à la *popularité* de la page ; elle est en tous cas une bonne mesure de son *accessibilité*. Elle est *robuste* aux changements temporels locaux du Web et aux tentatives de *spamming*.

Soit A la matrice telle que $A[p,q] = \frac{1}{d^+(p)}$ s'il existe un lien hypertexte dans la page p vers la page q, et 0 sinon. A est une matrice sous-stochastique que l'on nommera *matrice du Web*. Soit n la dimension de A (nombre de pages Web) et $\overrightarrow{E}$ le vecteur $(\frac{1}{n}, \frac{1}{n}, \ldots \frac{1}{n})^t$. Le vecteur PageRank $\overrightarrow{R}$ est la solution (unique) de l'équation $\overrightarrow{R} = d\ A^t\ \overrightarrow{R} + (1-d)\ \overrightarrow{E}$.

Le terme $(1-d)\ \overrightarrow{E}$, appelé ici *facteur zappe* (car il représente la probabilité, pour notre surfeur aléatoire, de «zapper» vers une page équiprobablement choisie du Web), est ajouté pour assurer l'existence d'une solution (que les pages sans successeurs menaceraient sinon). Il est pris égal à 0.85 par [Page et al., 1998] afin d'accélérer la convergence de l'algorithme. On peut critiquer le fait que le *zappe* suive une loi de distribution uniforme [Bouklit and Jean-Marie, 2002]. PageRank peut être vu comme une *distribution* à chaque étape du rang d'une page à toutes les pages qu'elle pointe (cf. figure 1).

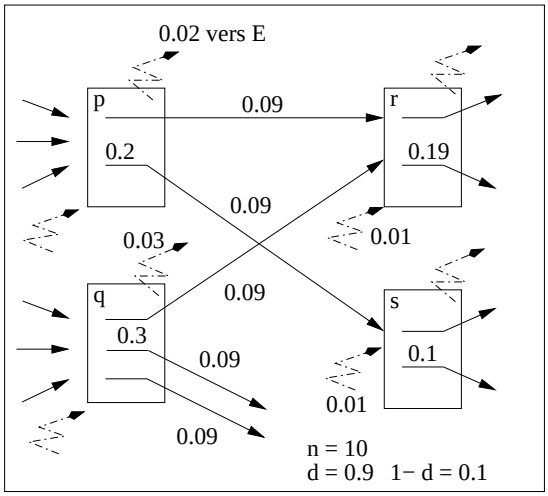


FIG. 1 - Propagation de rang

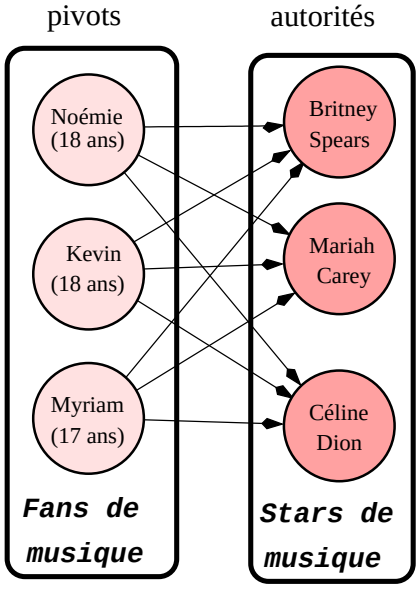


FIG. 2 - Exemple de communauté

3 Cybercommunautés

Dans [Gibson et al., 1998], Gibson *et al.* utilisent la méthode **HITS** de [Kleinberg, 1998] pour détecter des communautés hyperliées sur le Web. Une de leurs communautés correspond à un ensemble de pages *pivots* qui pointent sur un ensemble de pages faisant *autorité* pour un sujet donné. Le sous-graphe induit est biparti (figure 2).

Dans [Efe et al., 2000], les auteurs proposent un état de l'art intéressant des méthodes permettant d'extraire sur le Web des cybercommunautés [Kumar et al., 1999, Flake et al., 2000, Dean and Henzinger, 1999, Adamic, 1999].

4 Description du modèle

4.1 Modélisation de l'Univers

Notre modélisation du Web distingue deux entités. La première est le graphe du Web $G = (P, \mathcal{H})$ où P désigne l'ensemble des pages hypertextes. $(p, p') \in \mathcal{H}$ si et seulement si il existe un lien hypertexte dans $p \in P$ pointant $p' \in P$. Ce graphe est la donnée du problème. La deuxième entité est l'**espace** et le **temps** au sein desquels évolue le système. Pour obtenir un modèle se rapprochant le plus possible de la physique, nous avons pris l'espace euclidien[1] $\mathcal{E} = \mathbb{R}^3$, mais les nécessités de l'algorithmique nous ont fait choisir un temps discret $\mathcal{T} = \mathbb{N}$. Les pages y sont présentes en tant que particules massives.

4.2 Modélisation des actions

Les forces mettent en mouvement les pages/particules. Une interaction gravitationelle n'a lieu qu'entre pages unies par un lien hypertexte. Cette interaction respecte le principe galiléen *d'action et de réaction* : la force subie par la page pointée est la même que celle subie par la page qui pointe. Le sens de l'hyperlien ne compte que pour le transfert de masse (*cf. infra*). Nous avons utilisé simplement la force de gravitation newtonienne :

$$F_{pq} = \mathcal{G}\frac{m(p).m(q)}{dist(p,q)^2}$$

L'autre action subie par les particules est l'**expansion** de l'univers, qui les sépare au commencement. Nous avons pris une définition où l'univers a un centre O. Un point P de l'espace est translaté en un instant t en suivant

$$\overrightarrow{OP}_{t+1} = (1 + \lambda e^{-\alpha t}).\overrightarrow{OP}_t$$

L'expansion s'arrête asymptotiquement (assez vite, car $\sum e^{-\alpha t}$ converge).

4.3 Modélisation des transferts de masse

La masse représente l'*autorité* d'une page. Elle varie au cours du temps, ce qui viole le bon sens physique. Les transferts de masse s'inspirent du modèle PageRank (voir section 2), en y ajoutant

[1] augmenter le nombre de dimensions sépare davantage les pages, mais augmente l'espace mémoire requis, aussi pensons-nous que trois est déjà suffisant

la notion de distance. À chaque étape, une page **répartit** la totalité de sa masse entre les pages qu'elle pointe. La masse circule donc le long des liens hypertextes. Cette circulation respecte la loi des nœuds pour chaque page ; cela pose un problème pour les pages sans successeur. Une page transfère préférentiellement sa masse à ses proches voisines (renforcement local) ; la proportion de masse transférée est asymptotiquement nulle avec la distance. Enfin, la masse totale du système se conserve. Le transfert de masse blesse l'intuition du physicien dans la mesure où l'énergie ne se conserve pas. Mais il contribue au renforcement mutuel des pages spatialement proches en cybercommunautés.

La masse d'une page p à l'étape t est définie comme suit :

$$m_t(p) = d\Big(\sum_{q \text{ pointant } p} \frac{1}{dist_t(p,q)^\delta} \frac{m_{t-1}(q)}{S_t(q)} \Big) + (1-d) \sum_{r \in P} m_{t-1}(r)$$

Cette loi se ramène à celle de PageRank pour $\delta = 0$, la masse étant alors équitablement répartie entre les successeurs de la page, dont $S_t(q)$ est le degré sortant. Nous choisissons une distribution de la masse selon la distance euclidienne des pages : les pages proches reçoivent plus de masse que les pages lointaines. Nous utilisons $\delta = 2$ par cohérence avec la force gravitationelle.

Le facteur S_t assure la conservation de la masse et vaut

$$S_t(q) = \sum_{q \text{ pointant } r} \frac{1}{dist_t(q,r)^\delta}$$

Suivant [Page et al., 1998], nous prenons $d = 0.85$. Signalons enfin que les pages sans successeur et les erreurs d'arrondi font perdre de la masse au système. La masse est donc renormalisée après chaque itération afin qu'elle se conserve au total.

5 Implémentation du modèle

5.1 Implémentation en machine

Le calcul d'une itération (passage de l'instant t à $t+1$) se fait en temps linéaire par rapport au nombre de sommets et d'arcs du graphe. Les liens n'ont pas besoin d'être en mémoire : une seule passe le long du fichier des listes d'adjacence suffit à faire les calculs. Le facteur limitant est la *mémoire vive* plus que le temps, car chaque sommet occupe 32 octets (position, vitesse et masse), limitant à quelques dizaines de millions de sommets les expérimentations. Le programme pourrait facilement être parallélisé pour vaincre cette barrière. Le choix des constantes $\mathcal{G}$, λ, α et d se fait empiriquement.

5.2 Graphes utilisés

Nous avons utilisé deux sortes de jeux de données : tout d'abord des graphes artificiels. Nous avons en particulier testé des **graphes petits mondes**
[Watts and Strogatz, 1998, Newman, 1999, Adamic, 1999] qui nous ont permis de vérifier que ces derniers se regroupaient bien en galaxies. Pour ce faire, nous avons utilisé un algorithme de réorientation aléatoire des arêtes proposé par Watts et Strogatz [Watts and Strogatz, 1998] permettant de générer un graphe intermédiaire entre un graphe régulier et un graphe aléatoire sans altérer le nombre de sommets dans le graphe. Partant d'un graphe k-régulier à n sommets

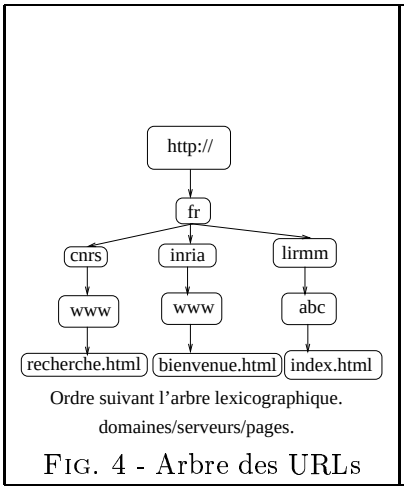


FIG. 4 - Arbre des URLs

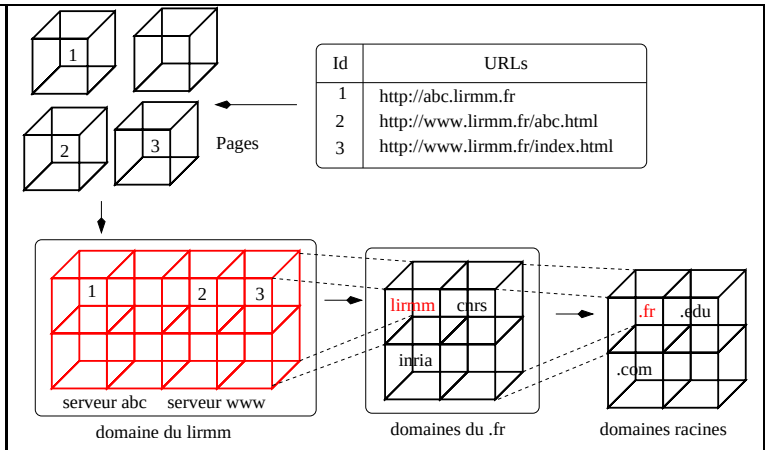


FIG. 5 - Représentation en cubes

disposé en anneau, l'algorithme réoriente chaque arête avec une probabilité p. Leur construction leur permet de générer un graphe *petits mondes* intermédiaire entre régularité ($p = 0$) et désordre ($p = 1$) (figure 3).

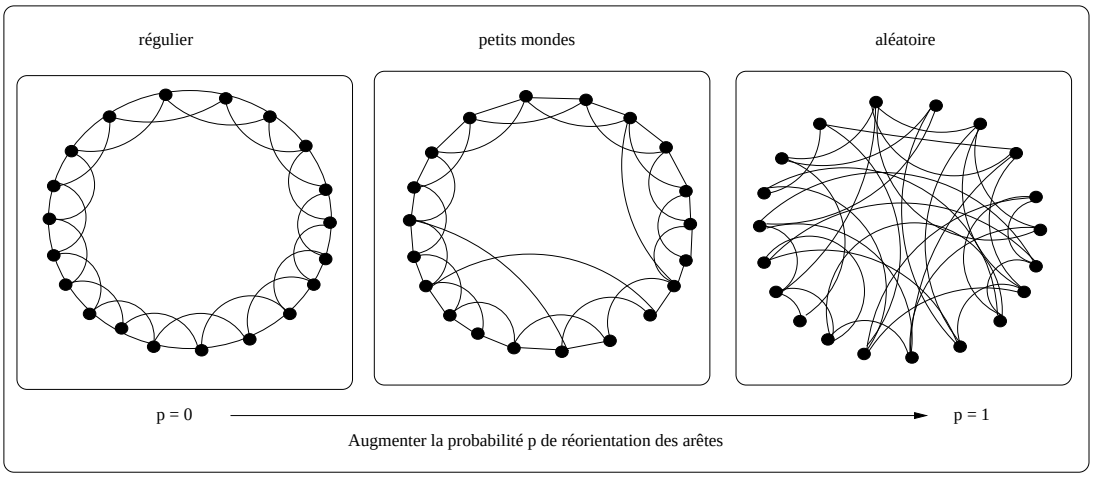


FIG. 3 - Les graphes *petits mondes* entre régularité et désordre

Nous avons également utilisé des *crawls*, parcours réels d'une partie du Web par des robots [Ailleret, 2000], images forcément incomplètes du Web mais qui en donnent une bonne idée. Le graphe «théorique» et instantané diffère nécessairement des différents avatars que peut en fournir un crawler ; l'existence des pages dynamiques le rend potentiellement infini.

5.3 Conditions initiales

Nous avons initialement placé les pages aléatoirement dans l'espace. Nous avons observé que dans ces conditions la formation des communautés s'effectuait lentement. C'est pourquoi nous avons pris le choix de faire la répartition initiale selon les sites. Cela permet entre autres de détecter des *micro-cybercommunautés* à l'intérieur d'un même site. Les pages Web du crawl sont

d'abord regroupées en un arbre (domaines/serveurs/pages) (figure 4). Puis cet arbre est parcouru en largeur. Un nœud à f fils donne naissance à un *cube* dans l'espace, dans lequel chacun de ses fils prend place comme cube de côté $\sqrt[3]{f}$, jusqu'aux feuilles qui donnent les points (figure 5). Paradoxalement, rien ne lie donc les sommets proches initialement : chacun est libre de migrer vers sa cybercommunauté. Par ailleurs, l'analyse de notre base de données révèle que 95 % des liens sont navigationnels. Les liens navigationels, qui sont entre pages spatialement proches, risquent de biaiser le modèle en captant tous les transferts de masse : pour cette raison, ils sont supprimés. Les sites peuvent aussi être regroupés en une seule particule. Tout cela demande une bonne définition des sites [Mathieu and Viennot, 2003].

5.4 Extraction des résultats

Un algorithme implémentant notre modèle fournit facilement une animation vidéo, donnant un résultat assez esthétique. Il est en revanche plus dur de détecter automatiquement les cybercommunautés. Il existe plusieurs méthodes de *clustérisation*, prenant en entrée le nuage de points avec ses positions, et fournissant en sortie les amas. Il en existe essentiellement deux types :

- *la k-clustérisation*, qui exhibe une partition de k éléments minimisant un certain critère, tel que la somme des carrés des distances. Le problème est NP-complet, mais de bonnes heuristiquues existent, basées sur les k-moyennes [MacQueen, 1967] ou les k-medioïdes [Kaufman and Rousseeuw, 1990].
- *la clustérisation hiérarchique* : il s'agit de produire un arbre dont les feuilles sont les éléments. Un algorithme ascendant, à chacune de ses itérations, lie deux sous-forêts minimisant un certain critère : distance moyenne, distance entre les feuilles minimale (*single linkage*) ou maximale (*complete linkage*), médiane, médioïde, etc. [Lance and Williams, 1967] étudie ces techniques dans le cadre (entre autres) de la distance euclidienne.

[Fasulo, 1999] étudie les différentes techniques. Nous n'avons implémenté que la clustérisation ascendate avec minimisation de la distance entre barycentres. Nous somme en train d'en implémenter d'autres, notamment les méthodes dynamiques de [Gibson et al., 1997].

6 Résultats

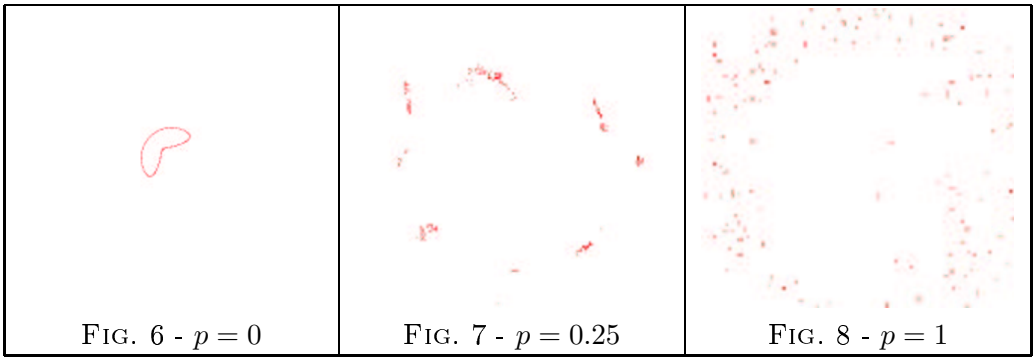


FIG. 6 - $p = 0$ FIG. 7 - $p = 0.25$ FIG. 8 - $p = 1$

Comme on peut le voir à figure 7, les graphes *petits mondes* se regroupent bien en communautés dans notre modèle. En revanche, les graphes totalement aléatoires se dissolvent rapidement dans l'espace (figure 8) tandis les graphes réguliers mettent en évidence une communauté de

voisinage compte tenu de leur structure bien ordonnée (figure 6). Enfin, les figures 9 à 12 représentent un crawl de 8 millions de pages. En quelques itérations, nous voyons se former à l'écran des cybercommunautés. Par ailleurs, nous avons constaté que 80 % des pages ont tendance à quitter leur emplacement d'origine (site) pour migrer vers une cybercommunauté.

Des résulats exhaustifs (listes de communautés) seraient fastidieux à fournir. Le problème pour valider le modèle est, étant donnée un (longue) liste d'URL, de vérifier si toutes sont sémantiquement liées. Nous ne le faisons actuellement que manuellement et songeons à mécaniser le prodécé (par des techniques d'I.A. qui restent à définir). Souvent les communautés ont en commun plus qu'un simple mot-clef : ainsi les pages Web des élèves de l'ENS d'Ulm forment un ensemble très compact et aisément identifiable !

7 Conclusion et perspectives

Partant d'une idée originale de représentation du Web, nous avons aussi trouvé une définition alternative de la notion de cybercommunauté, qui donne des résultats honorables. On a beaucoup dit que le graphe du Web possède une structure de *petits mondes* : notre modèle en fournit une preuve *visuelle*, les mondes en question se formant et tournant effectivement à l'écran !

Le Web est un objet très dynamique [Brewington and Cybenko, 2000], et notre modèle se prête particulièrement bien à l'évolution des liens et des pages. Sur le plan algorithmique, ce n'est en revanche pas aussi simple ; il serait intéressant de travailler dans ce sens.

Un des défis actuels est le **stockage** du graphe du Web, sa structure hypertexte ayant des milliards de liens [Bharat et al., 1998]. Les codages à base d'arbre [Guillaume et al., 2002] sont une bonne solution. Or, notre méthode de clusterisation fournit un arbre, différent de l'arbre des sites, et encore plus dense en liens : la probabilité d'un lien non-navigationnel entre deux feuilles proches est grande, tandis qu'elle est faible dans l'arbre des sites. Bien sûr, pour les liens navigationnels, c'est l'inverse. Mais une combinaison des deux représentations fournirait assurément un codage hybride d'une grande puissance.

Un autre problème est la construction de **crawlers intelligents** qui parcourent le Web en économisant la bande passante, donc en sélectionnant des pages *a priori* meilleures [Cho et al., 1998, Najork and Wiener, 2001]. Toute mesure d'audience permet de faire un choix : par définition, les successeurs d'une page au fort PageRank auront un fort PageRank ; ils doivent être retrouvés prioritairement. Notre modèle propose un autre critère : on peut chercher à obtenir une image d'une *région* du Web, en se concentrant sur les pages proches dans l'espace. À cause de la densité de liens, on obtiendra ainsi rapidement des pages pertinentes.

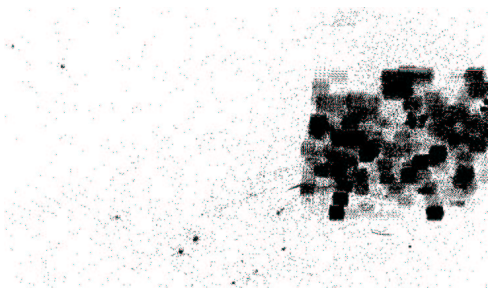


FIG. 9 - crawl de 8000000 pages

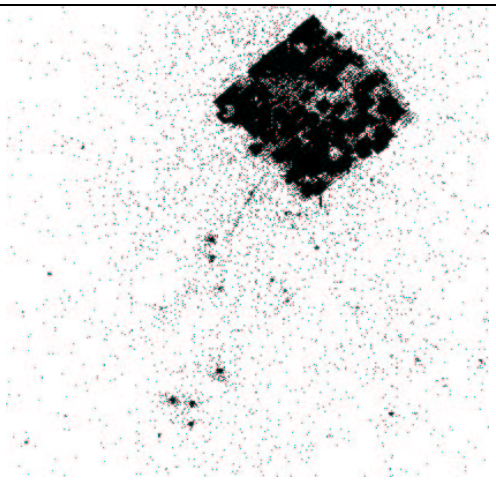


FIG. 10 - autre vue du même instant

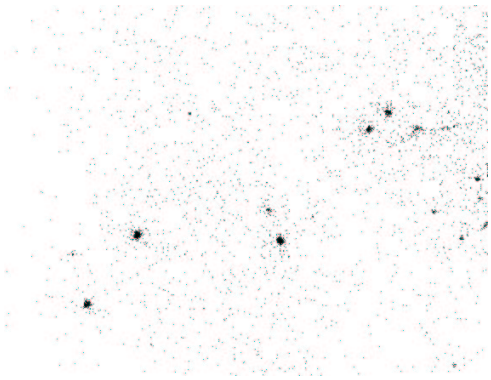


FIG. 11 - Quelques «galaxies»

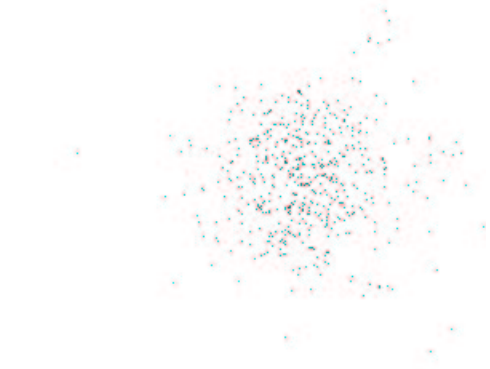


FIG. 12 - Zoom sur un «sytème solaire»

Références

[Adamic, 1999] Adamic, L. A. (1999). The small world web. In Abiteboul, S. and Vercoustre, A.-M., editors, *Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL*, volume 1696, pages 443–452. Springer-Verlag.

[Ailleret, 2000] Ailleret, S. (2000). Crawler larbin : http ://larbin.sourceforge.net/.

[Bharat et al., 1998] Bharat, K., Broder, A., Henzinger, M., Kumar, P., and Venkatasubramanian, S. (1998). The connectivity server : Fast access to linkage information on the web. In *Proceedings of the 7th International World Wide Web Conference(WWW7)*, Brisbane, Australia.

[Bouklit and Jean-Marie, 2002] Bouklit, M. and Jean-Marie, A. (2002). Une analyse de pagerank, une mesure de popularité des pages web. In *Proceedings ALGOTEL'02*, Mèze, France.

[Brewington and Cybenko, 2000] Brewington, B. E. and Cybenko, G. (2000). How dynamic is the Web ? *Computer Networks (Amsterdam, Netherlands : 1999)*, 33(1–6) :257–276.

[Cho et al., 1998] Cho, J., García-Molina, H., and Page, L. (1998). Efficient crawling through URL ordering. *Computer Networks and ISDN Systems*, 30(1–7) :161–172.

[Dean and Henzinger, 1999] Dean, J. and Henzinger, M. R. (1999). Finding related pages in the World Wide Web. *Computer Networks (Amsterdam, Netherlands : 1999)*, 31(11–16) :1467–1479.

[Efe et al., 2000] Efe, K., Raghavan, V., Chu, C. H., Broadwater, A. L., Bolelli, L., and Ertekin, S. (2000). The shape of the Web and its implications for searching the Web.

[Fasulo, 1999] Fasulo, D. (1999). An analysis of recent work on clustering algorithms.

[Flake et al., 2000] Flake, G., Lawrence, S., and Giles, C. L. (2000). Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA.

[Gibson et al., 1997] Gibson, D., Kleinberg, J., and Raghavan, P. (1997). Two algorithms for nearest-neighbor search in high dimensions. In *Proceedings of the 29th ACM STOCS*.

[Gibson et al., 1998] Gibson, D., Kleinberg, J. M., and Raghavan, P. (1998). Inferring web communities from link topology. In *UK Conference on Hypertext*, pages 225–234.

[Guillaume et al., 2002] Guillaume, J., Latapy, M., and Viennot, L. (2002). Efficient and simple encodings for the web graph. In *Proceedings of the 11-th international conference on the World Wide Web*.

[Hawking, 1988] Hawking, S. W. (1988). *A Brief History of Time*. Bantam, NY.

[Kaufman and Rousseeuw, 1990] Kaufman, L. and Rousseeuw, P. (1990). *Finding groups in data : an Introduction to Cluster Analysis*. John Wiley & Sons, Inc.

[Kleinberg, 1998] Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California.

[Kleinberg et al., 1999] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The Web as a graph : Measurements, models and methods. *Lecture Notes in Computer Science*, 1627 :1–? ?

[Kumar et al., 2000] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. (2000). The Web as a graph. In *Proc. 19th ACM SIGACT-SIGMOD-AIGART Symp. Principles of Database Systems, PODS*, pages 1–10. ACM Press.

[Kumar et al., 1999] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands : 1999)*, 31(11–16) :1481–1493.

[Lance and Williams, 1967] Lance, G. N. and Williams, W. T. (1967). A general theory of classification sorting strategies. *Computer Journal*, 9 :373–380.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. of fifth Berkeley Symposium*, pages 281–297.

[Mathieu and Viennot, 2003] Mathieu, F. and Viennot, L. (2003). Aspects locaux de l'importance globale des pages web. In *Algotel, 5ème Rencontres Francophones sur les aspects Algorithmiques des Télécommunications*.

[Najork and Wiener, 2001] Najork, M. and Wiener, J. L. (2001). Breadth-First Crawling Yields High-Quality Pages. In *Proceedings of the 10th International World Wide Web Conference*, pages 114–118, Hong Kong. Elsevier Science.

[Newman, 1999] Newman, M. (1999). Small worlds : The structure of social networks. cond-mat/0001118.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking : Bringing Order to the Web. Technical report, Computer Science Department, Stanford University.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of small-world networks. *Nature*, 393(1–7) :440–442.

Effet de la touche *Back* dans un modèle de surfeur aléatoire : application à PageRank

M. Bouklit, F. Mathieu

L.I.R.M.M.,
161, rue Ada,
34392 Montpellier Cedex 5, FRANCE.
Email : bouklit@lirmm.fr,fmathieu@clipper.ens.fr
Tél : +33 4 67 41 85 85 Fax : +33 4 67 41 85 00

Résumé

L'analyse du graphe formé par les pages Web et les liens hypertextes qui les relient, communément appelé graphe du Web, a permis d'améliorer la performance des moteurs de recherche actuels. Ainsi, lancé en 1998, le moteur de recherche Google classe les pages grâce à la combinaison de plusieurs facteurs dont le principal porte le nom de PageRank. Ce dernier est un indice numérique qui utilise le nombre de liens pointant sur les pages.

Dans cet article, nous analysons les différentes variantes du modèle PageRank proposée par [Page et al., 1998, Brin and Page, 1998, Kleinberg, 1998, Haveliwala, 1999, Abiteboul, 2001] que nous justifions, quand c'est possible, à l'aide d'un modèle probabiliste de l'internaute. Nous proposons d'affiner ces modèles en simulant l'effet de la touche *Back* du navigateur. Enfin, nous nous intéressons à l'analyse de ces différents algorithmes.

Abstract

Theorical analysis of the web graph is often used to improve the efficience of search engines. The PageRank algorithm, proposed by [Page et al., 1998], is used by the Google search engine [Google, 1998] to improve the results of requests. The purpose of this article is to describe the PageRank algorithm, its parameters and the probabilistic model that hides behind, then to propose an enhanced version using a model for the *Back* button.

1 Introduction

Les moteurs de recherche ont développé des méthodes de tri automatique des résultats. Leur but est d'afficher dans les dix à vingt premières réponses les documents répondant le mieux à la question. Dans la pratique, aucune méthode de tri n'est parfaite, d'autant plus que la question de la justesse d'un classement est en grande partie subjective. Un classement est justifié au mieux par un sondage, le plus souvent au jugement du lecteur. Cependant, la variété des méthodes offre à l'usager la possibilité de traquer l'information de différentes manières: cette variété augmente donc ses chances d'améliorer ses recherches.

Ainsi, lancé en 1998, le moteur de recherche Google[Google, 1998] classe les pages grâce à la combinaison de plusieurs facteurs dont le principal porte le nom de *PageRank*[Page et al., 1998]. Plus précisément, le classement des pages est fait en utilisant un indice numérique (le «rang») calculé pour chaque page.

En amont du processus, il y a tout d'abord les robots qui chalutent continuellement le Web dans l'intention de découvrir de nouvelles pages et à défaut de mettre à jour les anciennes. Ces pages sont stockées dans un entrepôt de données. Viennent ensuite les hyperliens qui sont stockés séparément pour former au final un sous-graphe du Web. Ce graphe est alors utilisé pour le calcul des rangs de page. Le rang d'une page permettra en particulier d'ordonner les résultats d'une requête d'un usager. Dans [Page et al., 1998], les auteurs proposent un modèle de *conservation du rang* et un algorithme permettant son calcul : l'algorithme PageRank.

Dans la section 2, nous analysons les différentes variantes du modèle PageRank proposée par [Page et al., 1998, Brin and Page, 1998, Kleinberg, 1998, Haveliwala, 1999] et [Abiteboul, 2001] que nous justifions, quand c'est possible, à l'aide d'un modèle probabiliste de l'internaute. Nous proposons également d'affiner ces modèles en simulant l'effet de la touche *Back* du navigateur.

Enfin, la section 3 s'intéresse au coût et à la convergence des algorithmes issus des différents modèles.

2 Modélisations probabilistes de l'internaute

Dans cette section, nous présenterons le concept du *surfeur aléatoire*, puis nous verrons comment la description du comportement de ce surfeur entraîne des équations de conservation locale pouvant s'exprimer sous forme vectorielle.

Nous appellerons $G = (V,E)$ le graphe orienté formé par les pages web V et les liens hypertextes qui les relient E. En pratique, G est principalement obtenu par une succession de parcours du Web (*crawls*).

2.1 Le surfeur aléatoire

L'axiome caché derrière l'algorithme de PageRank est assez étrange, voire peu flatteur pour les internautes. Il dit que les pages les plus intéressantes sont celles sur lesquelles on a le plus de chance de tomber en cliquant au hasard. Exprimé autrement, le cerveau est un outil secondaire quand il s'agit de trouver des «bonnes» pages web. Toutes les variantes de PageRank ont donc comme point de départ un *surfeur aléatoire*, censé modéliser un internaute lambda, dont le comportement, bien qu'aléatoire, est soumis à certaines règles qui définissent la variante. Le plus souvent, ces règles se traduisent par

un processus stochastique de type markovien. À partir d'une distribution initiale de probabilité sur l'ensemble des pages web, le processus est itéré et, sous réserves de garanties de convergence et d'unicité de la limite, tend vers une distribution de probabilité qui est par définition le PageRank de la variante en question.

2.2 Modèle initial

Le niveau zéro du *surfeur aléatoire*, proposé par [Page et al., 1998], suppose que notre internaute, quand il est sur une page donnée, va ensuite cliquer de manière équiprobable sur un des liens sortants. Si R_n représente le vecteur des probabilités de présence de notre surfeur à l'instant n, l'équation de propagation est :

$$R_{n+1}(p) = c_n \sum_{q \to p} \frac{R_n(q)}{d^+(q)} \tag{1}$$

$d^+(q)$ étant le degré externe de q et c_n étant un facteur de normalisation choisi de telle sorte que $\forall n$, $||R_n||_1 = 1$.

Vectoriellement, si on appelle M la matrice d'adjacence de G, et $A = (\frac{1}{d^+(i)} M_{i,j})$ (par convention, $\frac{0}{0} = 0$), l'équation de propagation est $R_{n+1} = c_n A^t R_n$.

La présence du facteur de normalisation c_n assure la convergence vers un vecteur non nul. En effet, en l'absence de normalisation, le processus itératif (1) n'est stable pour la norme ℓ_1 que si A est stochastique, c'est-à-dire si le graphe est sans feuille. Par défaut, ce n'est pas le cas du graphe du web.

2.3 Le facteur *zap*

Comme nous le verrons dans la section 3, l'unicité de la convergence n'est assurée qu'en cas de forte connexité. Un graphe du web ne l'est en général pas. En plus des *puits de rang* (les feuilles), il existe une partie non négligeable du web dont on ne peut pas sortir en cliquant [Broder et al., 2000]. Pour échapper aux circuits sans issue et aux *puits de rang*, il est nécessaire «de temps en temps» de sauter aléatoirement vers une page quelconque du Web.

2.3.1 Méthode du rang initial

Pour modéliser les sauts aléatoires, [Page et al., 1998] proposent de doter chaque page d'un rang initial (*source de rang*). Ainsi chaque sommet p se voit attribuer un rang de $S(p) > 0$. (1) devient alors :

$$R_{n+1}(p) = c_n \left(\sum_{q \to p} \frac{R_n(q)}{d^+(q)} + S(p) \right) \tag{2}$$

L'écriture vectorielle de cette équation est $R_{n+1} = c_n(A^t R_n + S)$.

Quand $||S||_1 = 1$, S représente une loi de distribution sur l'ensemble des pages de E. Le plus souvent, on choisit pour S une loi de distribution uniforme: $\forall p \in E, S(p) = \frac{1}{|E|}$. Mais il a été proposé que cette distribution puisse être «personnalisée» [Brin et al., 1998].

2.3.2 Facteur d'amortissement

L'équation (2) n'admet pas d'interprétation probabiliste directe. [Brin and Page, 1998], puis [Kleinberg, 1998] propose une variante empirique du modèle en introduisant un facteur de pondération $d \in [0,1]$, ce qui donne:

$$R_{n+1}(p) = c_n \left(d\times \left(\sum_{q\to p} \frac{R_n(q)}{d^+(q)} \right) + (1-d)S(p) \right) \tag{3}$$

avec pour condition $\sum_{p\in V} S(p) = 1$ et $\forall n, \sum_{p\in V} R_n(p) = 1$.

L'écriture vectorielle de cette équation est :

$$R_{n+1} = c_n(dA^t R_n + (1-d)S) \tag{4}$$

L'initialisation R_0 du processus est égal à la distribution S «par défaut» mais peut être choisi autrement. Par exemple, prendre le résultat d'un calcul précédent peut souvent accélérer la convergence. Cette remarque s'applique à l'algorithme précédent, même si S y a l'interprétation de «rang initial», voir (2).

On espère approcher asymptotiquement des valeurs (R,c) vérifiant :

$$R = c(dA^t R + (1-d)S) \tag{5}$$

Remarque : L'équation (5) peut être reformulée comme suit : $R = c(d\ A^t + (1-d)\ E\times\mathbf{1})$R où $\mathbf{1}$ désigne le vecteur ligne ne contenant que des 1. En effet, comme $\sum_{i=1}^{n} R_i = 1$, on obtient alors : $\mathbf{1}\times R = 1$. Il en découle que : $E = E\times\mathbf{1}\times R$ et par conséquent R est un vecteur propre de la matrice $d\ A^t + (1-d)\ E\times\mathbf{1}$ pour la valeur propre $1/c$. Cette propriété sera exploitée au paragraphe 3.3.

La définition de l'équation (4) peut s'interpréter, à normalisation près, ainsi : à chaque page, notre internaute lambda a la possibilité

- soit de *cliquer* sur l'un des liens sortants $(A^t\ R)$ avec la probabilité d.
- soit de *zapper*, avec la probabilité $1-d$ cette fois, sur une page choisie aléatoirement selon la distribution de E.

Ce modèle est repris dans [Haveliwala, 1999]. Selon cet auteur, l'introduction du paramètre d'amortissement d est destinée à améliorer la «qualité» du PageRank en garantissant la convergence vers un unique vecteur rang. La matrice A est explicitement supposée stochastique en éliminant itérativement les pages sans liens. La figure 1 illustre une propagation de rang d'une paire de pages à l'autre. On y suppose $d = 0.9$ et $n = 10$. En observant la page p sur cette figure, nous remarquons que:

- d =90 % de son rang (soit 0.18) est redistribué équitablement sur ses liens sortants (soit $\frac{0.18}{2} = 0.09$) affectant ainsi le rang des pages pointés par p.
- $1-d$ =10 % de son rang (soit 0.02) est dissipée au profit d'une répartition globale sur l'ensemble du graphe contribuant ainsi à alimenter chaque page d'un rang égal à $\frac{1-d}{n} = \frac{0.1}{10} = 0.01$.

Nous pouvons vérifier par exemple que le rang de la page r est bien 0.19:

$$\begin{aligned} R(r) &= d\times\frac{R(p)}{d^+(p)} + d\times\frac{R(q)}{d^+(q)} + \frac{1-d}{n} = 0.9\times\frac{0.2}{2} + 0.9\times\frac{0.3}{3} + \frac{0.1}{10} \\ &= 0.19 \end{aligned}$$

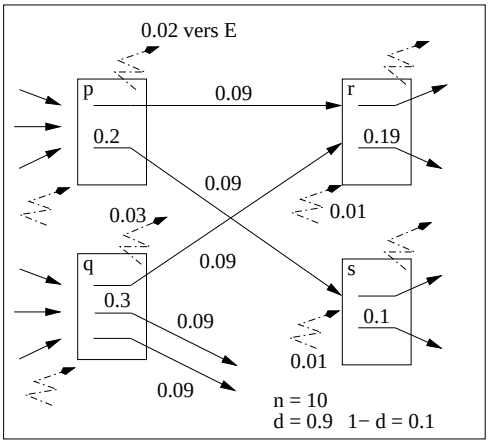


FIG. 1 – *propagation de rang*

2.3.3 Page virtuelle de *zap*

La présence de *puits de rang* rend toujours la normalisation obligatoire. Une troisième façon de modéliser le *zap* a été proposée par Serge Abiteboul dans [Abiteboul, 2001] : elle consiste à rajouter à G un sommet fictif π, pointé et pointant sur tous les autres sommets. On simule alors le *zap* en appliquant l'équation (1) sur $G' = (V \cup \pi, E \cup (E,\pi) \cup (\pi,E))$. L'introduction du facteur d'amortissement est toujours possible (il suffit de pondérer les transitions vers π) et il n'est plus nécessaire de normaliser à chaque itération [1].

2.4 Modélisation de la touche *Back*

Nous nous proposons d'affiner notre modèle de l'internaute en incorporant, en plus de *zapper* et *cliquer*, la possibilité de *revenir* [2]. Nous travaillerons ici dans l'hypothèse où l'historique du navigateur se limite à une page. En effet, la théorie des chaînes de Markov nous apprend que pour modéliser un processus à $|V|$ états et à mémoire m, il faut travailler sur le sous-ensemble de V^{m+1} des chemins de longueur m possibles. Pour $m = 1$, cela correspond à l'ensemble des hyperliens, ce qui est encore faisable, au-delà, la taille du graphe du web rend la tâche impossible.

Dans un premier temps, par simplicité, nous verrons comment modéliser le retour sans amortissement.

2.4.1 *Back* réversible

Dans ce modèle, nous supposerons qu'à chaque étape , notre internaute peut cliquer de manière équiprobable sur les liens ou la touche *Back* (la touche *Back* est donc considérée comme un lien sortant comme les autres). Si la touche *Back* vient d'être utilisée, elle reste active et correspond à l'annulation du *Back* précédent. Rappelons que nous interprétons $R(p)$ comme la probabilité de présence en un sommet p. Appelons $R^{Br}(q,p)$ la probabilité d'être en p en ayant été en q à l'instant d'avant. On peut alors écrire les lois de conservation du rang. La probabilité d'être en p à l'instant n est la somme des probabilités d'être en p à l'instant n à partir de l'ensemble des pages q possibles :

1. Une fois la convergence obtenue, il suffit de diviser les rangs de V par $(1 - R(\pi))$.

2. Cette amélioration a déjà été proposée au niveau théorique dans [Fagin et al.,]. Notre approche est légèrement différente, plus basée sur le surcoût du modèle.

$$R_n(p) = \sum_{q \leftrightarrow p} R_n^{Br}(q,p)$$

On travaille dans ce cas sur le graphe non-orienté induit par G. En effet, à cause de la touche *Back*, on peut venir sur une page p par un lien entrant ou sortant. Sur le même principe, on déduit que l'équation donnant $R^{Br}(q,p)$ est :

$$\begin{aligned} R_{n+1}^{Br}(q,p) &= \frac{1}{d^+(q)+1}\left(\sum_{q \leftrightarrow p} R_n^{Br}(q',q) + R_n^{Br}(p,q)\right) &(6)\\ &= \frac{1}{d^+(q)+1}\left(R_n(q) + R_n^{Br}(p,q)\right) &(7)\end{aligned}$$

2.4.2 *Back* irréversible

Nous allons maintenant envisager le cas où la touche *Back* ne peut pas être appelée deux fois consécutivement. Ce modèle, *a priori* plus complexe que le précédent, a cependant trois avantages qui méritent notre attention :

- Il réduit de manière significative le PageRank accumulé par «effet de serre» dans les feuilles.
- Il est plus adapté à l'introduction d'un facteur d'amortissement (voir 2.4.3).
- Contrairement à ce que l'on pourrait penser, il est beaucoup moins gourmand en ressources.

Appelons $R^{Bi}(q,p)$ la probabilité de se trouver en p en ayant cliqué à partir de q (sans la touche *Back*), et $R^{Bi}(p)$ celle de se trouver en p sans pouvoir utiliser la touche *Back*. La probabilité de se trouver en p grâce à la touche *Back* est déduite des probabilités de se trouver sur une page q en venant de p :

$$R_{n+1}^{Bi}(p) = \sum_{q \leftarrow p} \frac{R_n^{Bi}(p,q)}{d^+(q)+1}$$

La probabilité $R^{Bi}(q,p)$ dépend quant à elle des différents $R^{Bi}(q',q)$ ainsi que de $R^{Bi}(q)$:

$$R_{n+1}^{Bi}(q,p) = \sum_{q' \to q} \frac{R_n^{Bi}(q',q)}{d^+(q)+1} + \frac{R_n^{Bi}(q)}{d^+(q)}$$

On peut constater que, contrairement au cas réversible, $R^{Bi}(q,p)$ ne dépend pas du sommet d'arrivée p. Il n'est donc plus nécessaire de stocker une information par lien, mais une par page, que nous noterons R^B.

On obtient alors un système itératif plus simple :

$$\begin{aligned} R_{n+1}^{Bi}(p) &= R_n^B(p)\sum_{q \leftarrow p} \frac{1}{d^+(q)+1} &(8)\\ R_{n+1}^{B}(p) &= \sum_{q \to p} \frac{R_n^B(q)}{d^+(p)+1} + \frac{R_n^{Bi}(p)}{d^+(p)} &(9)\\ & &(10)\end{aligned}$$

2.4.3 Touche *Back* et *zap*

L'introduction de la touche *Back* éliminant les puits de rang, nous n'allons pas utiliser le rajout d'une page virtuelle, mais plus simplement le facteur d'amortissement. Celui-ci autorise en toute rigueur toutes les transitions de $V \times V$. Ceci représentant un espace trop grand en pratique, nous allons devoir désactiver la touche *Back* lors du *zap*. La modélisation *Back* irréversible donc pour cette raison supplémentaire. On obtient alors le système de propagation du rang suivant :

$$R^{Bi}_{n+1}(p) = dR^{B}_{n}(p)\left(\sum_{q\leftarrow p}\frac{1}{d^{+}(q)+1}\right)+(1-d)S(p) \quad (11)$$

$$R^{B}_{n+1}(p) = d\left(\sum_{q\rightarrow p}\frac{R^{B}_{n}(q)}{d^{+}(p)+1}+\frac{R^{Bi}_{n}(p)}{d^{+}(p)}\right) \quad (12)$$

3 Analyse

Les résultats empiriques rapportés dans [Page et al., 1998] indiquent une convergence rapide de l'algorithme en pratique : en quelques dizaines d'itérations, une approximation raisonnable du PageRank est atteinte sur un graphe de 322 millions de liens. Les auteurs suggèrent que l'explication pourrait provenir d'une propriété d'*expansion* du graphe du Web, et font référence à [Motwani and Raghavan, 1995]. En effet, pour un graphe expansif, on sait donner une borne supérieure pour les valeurs propres de A^t différentes de la valeur propre principale. En fait, ces résultats ne sont démontrés que pour des graphes non orientés de degré constant, donc pas directement au problème des pages du Web. Mais il est certain que les relations entre la topologie du graphe et la vitesse de convergence de l'algorithme restent à explorer. Il existe des outils analytiques et géométriques permettant d'attaquer le problème en dehors du cadre des graphes expansifs [Saloff-Coste, 1996].

3.1 Coût de la touche *Back*

La taille exacte du graphe du web est difficile à estimer, ne serait-ce qu'à cause de la difficulté d'avoir une définition précise : pages dynamiques, documents multimedia, pages dédoublées, *frames*, intranets... Qu'est-ce qui fait partie du graphe du web ? De plus, même en précisant l'ensemble des données que l'on considère comme formant le graphe du web, il est physiquement impossible de tout parcourir. Cependant, pour donner un ordre de grandeur, il est bon de rappeler que *Google* affirme avoir indexé plus de trois milliards de documents.

Cette quantité phénomale de données en expansion permanente impose déjà une contrainte, qui est que tout algorithme doit être linéaire. Plus précisément, à l'heure actuelle, l'application d'algorithmes comme *PageRank* est principalement limitée par le nombre d'accès mémoire. Les versions classiques de *PageRank* nécessitent à chaque itération $|V|$ écritures et $|E|$ lectures des données de *PageRank*. La version que nous avons proposé au paragraphe 2.4.3 nécessite quant à elle $2\,|V|$ écritures pour $|E|+2\,|V|$ lectures.

Notre modèle est plus fin que celui du *PageRank* classique, mais son coût est plus important. On constate cependant qu'il est au plus deux fois plus coûteux en ressources, ce qui reste raisonnable, alors qu'on aurait pu s'attendre à un facteur de l'ordre $\frac{|E|}{|V|}$ (ce qui se serait produit si on avait implémenté la touche *Back* réversible).

3.2 Convergence : cadre théorique

Par souci de simplicité, nous étudierons ici l'équation (5). L'implémentation de la touche *Back* ne change fondamentalement pas les résultats, mais complique les notations, c'est pourquoi nous ne l'analyserons pas ici.

Du point de vue théorique, les premières questions qui se posent sont :

1. existe-t-il un vecteur de rangs R, et une constante c solution de l'équation de conservation ?
2. cette solution est-elle bien définie (unique)?
3. l'algorithme itératif correspondant converge-t-il vers cette solution ?

La réponse à ces trois questions est apportée par l'analyse des valeurs propres de la matrice. On observe en effet qu'il est possible de reformuler (5) en:

$$\left(d\ A^t + (1-d)\ S{\times}\mathbf{1}\right)\ R\ =\ \frac{1}{c}\ R\ . \tag{13}$$

Le problème est donc: existe-t-il un vecteur positif R qui soit le vecteur propre de la matrice $d\ A^t + (1-d)\ S{\times}\mathbf{1}$, pour une valeur propre $1/c > 0$?

Le Théorème de Perron-Frobenius permet de répondre. Nous incluons ci-après une version de ce théorème (voir par exemple [Horn and Johnson, 1985]). On rappelle qu'une matrice est *irréductible* si son graphe est fortement connexe, et *apériodique* si le p.g.c.d. des longueurs des circuits est 1.

Théorème 1 (Perron-Frobenius) *Soit A une matrice positive, irréductible et apériodique. Alors il existe une valeur propre r de A telle que:*

a) $r > 0$;

b) r est associée à un vecteur propre gauche $x > 0$ et un vecteur propre droit $y > 0$;

c) pour toute autre valeur propre λ de A, $|\lambda| < r$;

d) soit la matrice $L = yx/xy$: alors $(A/r)^n \to L$ quand $n \to \infty$, et pour tout $\phi > \frac{1}{r}max\{|\lambda|,\ \lambda$ valeur propre de $A \neq r\}$, il existe C tel que pour tout n,

$$\|(A/r)^n - L\|_1\ \leq\ C\ \phi^n\ . \tag{14}$$

La valeur propre r est dite valeur propre principale (ou: de Perron-Frobenius) de A.

3.3 Existence de solutions

Le Théorème de Perron-Frobenius s'applique très facilement dès que l'on suppose A stochastique. Comme nous l'avons vu, il existe de nombreuses méthodes pour arriver à ce résultat : page virtuelle de zap, modélisation de la touche *Back* ou encore *zap* à probabilité 1 sur les feuilles. Dans tous ces cas, il est possible de transformer A en une matrice A' stochastique.

La matrice $B = dA' + (1-d)\mathbf{1}^t{\times}S^t$ est aussi stochastique pour tout $d \in [0,1]$, de plus elle est apériodique et irréductible si $0 \leq d < 1$. En effet, nous avons supposé que $S > 0$, donc $B_{pq} = dA_{pq} + (1-d)S(p) > 0$. D'après le théorème de 1, il existe donc une solution unique $R > 0$ au système (13), avec pour constante c l'inverse de la valeur propre principale de B, à savoir 1.

Le système (13) a alors pour solution $c = 1$ et R s'interprète comme le vecteur de *probabilité stationnaires* de la chaîne de Markov dont la matrice de transition est B (voir par exemple [Kemeny and Snell, 1960]). Cette chaîne de Markov est précisément celle parcourue par un internaute ayant le comportement décrit au paragraphe 2.1.

3.4 Convergence du processus itératif

Soit $c = 1/r$, r la valeur propre principale de B. La récurrence se résout en:

$$R_n = (c_{n-1}/c)(cB^t)R_{n-1} = \prod_{m=0}^{n-1}(c_m/c)(cB^t)^n R_0 = \gamma_n (cB^t)^n R_0,$$

où γ_n est la constante telle que $||R_n||_1 = 1$. Appliquons la partie d) du Théorème 1. Nous choisissons comme vecteur propre droit de B^t le vecteur R tel que $||R||_1 = 1$. Il lui correspond un vecteur propre gauche S tel que $SR = 1$. Nous avons alors: $(cB^t)^n = RS + O(\phi^n)$. Par conséquent,

$$R_n = \gamma_n \ (RS \ + \ O(\phi^n)) \ R_0 = \gamma_n \ R \ (SR_0) \ + \ O(\phi^n) = \delta_n \ R \ + \ O(\phi^n).$$

Comme $||R_n||_1 = 1$, il vient: $\delta_n = 1 + O(\phi^n)$. Finalement, nous avons:

$$R_n \ = \ R \ + \ O(\phi^n) \ . \tag{15}$$

Le processus itératif converge donc vers R, et sa vitesse de convergence est géométrique. Le facteur de convergence ϕ dépend du rapport des deux plus grandes valeurs propres de B. Le nombre d'itérations nécéssaires pour approcher R à une distance de ϵ est de l'ordre de $\log \epsilon / \log \phi$.

4 Conclusions

Nous avons décrit les principes de la méthode PageRank, et montré ses liens avec la théorie des chaînes de Markov. Nous avons également décrit des algorithmes pour calculer le rang des pages, et expliqué leur convergence. De nombreuses questions restent en suspens.

D'un point de vue pratique, la méthode fait intervenir deux paramètres *a priori*: le facteur de pondération d et la distribution S. Or, il est clair que la valeur du rang obtenue dépend de ces deux paramètres. Les résultats empiriques s'accordent sur une valeur $d = 0.85$, en utilisant pour S la distribution uniforme[Brin and Page, 1998]. Peut-on justifier, voire améliorer ce choix?

D'un point de vue théorique, il reste à expliquer la convergence rapide des algorithmes, c'est-à-dire la petitesse de la valeur ϕ dans (15). Parmi les outils abordant cette question, nous avons mentionné les graphes expansifs [Motwani and Raghavan, 1995], et les approches analytiques ou géométriques du calcul du «trou spectral» [Saloff-Coste, 1996]. Un autre sujet intéressant serait de trouver comment exprimer le rang comme fonction plus explicite du graphe du Web. Par exemple, certains travaux font apparaître que les distributions du rang et du degré des pages suivent toutes deux une loi de puissance[Pandurangan et al.,].

Du point de vue algorithmique, la question est d'augmenter encore la vitesse de calcul du rang, sachant que le graphe est énorme et dynamique. Une idée serait d'utiliser les résultats précédemment calculés (incrémentalité), mais aussi de décomposer le calcul en «blocs» selon les vitesses d'évolution respectives (parallélisme, asynchronisme). Ce type d'approche dynamique du problème est actuellement en cours d'étude.

Références

[Abiteboul, 2001] Abiteboul, S. (2001). Page rank incremental. personnal communication with L. Viennot.

[Brin et al., 1998] Brin, S., Motwani, R., Page, L., and Winograd, T. (1998). What can you do with a Web in your Pocket? *Data Engineering Bulletin*, 21(2):37–47.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

[Broder et al., 2000] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. In *Proc. 9th International World Wide Web Conference*, pages 309–320.

[Fagin et al.,] Fagin, R., Karlin, A. R., Kleinberg, J., Raghavan, P., Rajagopalan, S., Rubinfeld, R., Sudan, M., and Tomkins, A. Random walks with back buttons.

[Google, 1998] Google (1998). http://www.google.com/.

[Haveliwala, 1999] Haveliwala, T. (1999). Efficient computation of PageRank. Technical report, Computer Science Department, Stanford University.

[Horn and Johnson, 1985] Horn, R. and Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press.

[Kemeny and Snell, 1960] Kemeny, J. and Snell, J. (1960). *Finite Markov Chains*. Springer Verlag.

[Kleinberg, 1998] Kleinberg, J. (1998). Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, San Francisco, California.

[Motwani and Raghavan, 1995] Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Computer Science Department, Stanford University.

[Pandurangan et al.,] Pandurangan, G., Raghavan, P., and Upfal, E. Using PageRank to Characterize Web Structure. *Manuscrit*.

[Saloff-Coste, 1996] Saloff-Coste, L. (1996). Lectures on finite Markov chains. In E. Giné, G. G. and Saloff-Coste, L., editors, *Lecture Notes on Probability Theory and Statistics*, number 1665 in LNM, pages 301–413. Springer Verlag.

Journées Francophones de la Toile - JFT'2003

Navigation et comportement utilisateur

Le prétraitement des fichiers logs Web dans le "Web Usage Mining" multi-sites

Doru Tanasa, Brigitte Trousse

Équipe AxIS, INRIA Sophia Antipolis,
2004, Route des Lucioles,
06902 Sophia Antipolis Cedex, FRANCE.
Email : {Doru.Tanasa,Brigitte.Trousse}@inria.fr
Tél : +33 4 92 38 78 36 Fax : +33 4 92 38 76 69

Résumé

Cet article concerne l'étape de prétraitement du "Web Usage Mining", étape reconnue comme importante et nécessaire avant d'appliquer les techniques de fouille de données. Nous présentons en détail une méthode générale de prétraitement des logs HTTP que nous avons utilisée lors d'une expérimentation en 2002. Cette expérimentation a été effectuée sur les fichiers logs HTTP de quatre serveurs Web de l'Inria. L'originalité de notre méthode est le fait qu'elle prenne en compte l'aspect multi-sites, indispensable pour mieux appréhender les pratiques des internautes et qu'elle intègre les principaux travaux existants sur le thème.

Abstract

In this document we present the preprocessing step of the Web Usage Mining (WUM), well known as an important and necessary step before applying the data mining techniques. We detail a general technique for preprocessing the HTTP logs that we used within an experiment done in 2002. This experiment was done on the HTTP log files extracted from four of the Inria's Web servers. The originality of our method relies first in the fact that we take into account the multi-site feature, which is crucial for a better understanding of the internauts' habits, and secondly in the integration of the main existing works in WUM.

1 Introduction

Avec plus de 3 milliards de documents en ligne (cf. `http://www.google.com/`) et 20 millions de nouvelles pages Web publiées chaque jour [Vie02], le Web deviendra bientôt la principale source d'information. Par conséquent, à l'avenir, on recherchera l'information sur l'Internet plutôt que dans une bibliothèque. Les créateurs des sites Internet intéressés à attirer et à garder les nouveaux visiteurs doivent savoir que leur offrir plus d'information ne constitue pas toujours une solution. En effet, les utilisateurs d'un site Web apprécieront davantage la manière dont cette information est présentée au sein du site. Par conséquent, l'analyse du comportement des utilisateurs (enregistré dans les fichiers de type log) est une tâche importante dans la reconception des sites Web.

Le "Web Usage Mining" (WUM) est défini comme l'application d'un processus d'Extraction

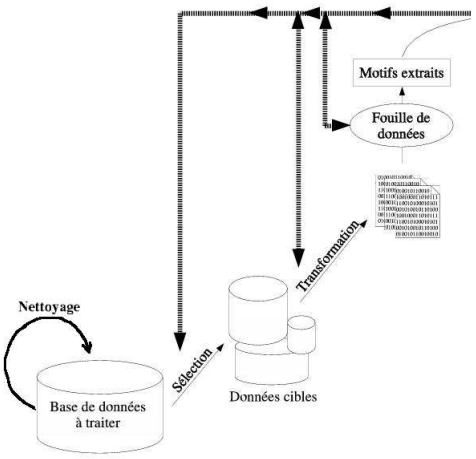


FIG. 1 – *Le schéma du processus ECD*

des Connaissances à partir de bases de Données (ou ECD, cf. figure 1) aux données logs HTTP dans le but d'extraire des patrons d'accès Web en vue d'une amélioration ou personnalisation du site. Un processus WUM comporte trois étapes principales : prétraitement, fouille de données et analyse de motifs extraits. Dans cet article, nous nous concentrerons seulement sur la première étape, le prétraitement du WUM (nettoyer, sélectionner et transformer les données), qui est un processus fastidieux et complexe dû principalement à la grande quantité de données (les fichiers logs Web) et à la faible qualité de l'information qu'on trouve dans les fichiers logs Web.

Dans cette première étape, plusieurs tâches doivent être accomplies, comme : le nettoyage des données, l'identification des utilisateurs, l'identification des vues de pages ("page view") et l'identification des visites et épisodes. Une description détaillée de cette étape peut être trouvée à la section 3 de l'article.

Récemment, beaucoup de travaux en WUM sont apparus, mais seuls quelques-uns d'eux se sont concentrés vraiment sur l'étape de prétraitement. La plupart des auteurs travaillant sur des techniques de fouille de données les ont appliquées sur les fichiers logs ayant peu subi de transformations [CPY96, PHMaZ00, MTP01]. Cependant, nous pouvons noter quelques exceptions accordant plus d'importance à l'étape de prétraitement [Coo00, Yal02].

Un autre point que nous traitons dans cet article est le WUM multi-sites. Prenons, une organisation importante, comme l'Inria, qui a six unités de recherche dans toute la France. Il existe un site/serveur Web au niveau national et un pour chaque unité de recherche[1]. Pour notre expérimentation, nous avons pris les logs HTTP du serveur Web national et celles du serveur Web

1. Inria niveau national: `http://www.inria.fr/`. Par example l'unité de recherche de Sophia Antipolis: `http://www-sop.inria.fr/`

d'Inria Sophia Antipolis. En plus, comme les deux sites avaient chacun leur propre moteur de recherche (du type Altavista), nous avons pris les logs de ces deux moteurs de recherche: un utilisateur qui recherche de l'information, "navigue" parmi tous ces serveurs d'une façon relativement transparente car les pages de différents serveurs Web sont fortement liées entre elles. Il y a de fortes chances que le visiteur ne remarque même pas que le serveur Web a changé (pour voir ceci il doit regarder l'adresse dans la barre d'adresses). Au contraire, pour l'analyste WUM, ce changement est très important. Pour lui permettre d'analyser le comportement de l'utilisateur dans sa recherche de l'information et de valider par exemple le rôle de routage de certaines pages Web du site national vis à vis des sites des unités de recherche, il doit reconstituer le chemin suivi par l'utilisateur sur les différents serveurs à partir de toutes les entrées des fichiers logs des divers serveurs sur lesquels l'utilisateur a navigué (il y a un fichier log Web par serveur Web). Notre solution est de fusionner tous ces fichiers logs Web puis de reconstituer les visites des internautes. Nous détaillerons ce procédé dans la section 3.2.
Concernant le WUM multi-sites, on note aucun travail spécifique à ce jour. Citons, cependant le travail de [MTP01] qui a comme objectif l'étude des comportements des utilisateurs visitant des sites Web partenaires. La différence de leur travail avec le notre est que, dans notre cas, les sites Web sont issus d'une même organisation ou entreprise et sont fortement inter-liées), alors que dans leur cas il s'agit de sites indépendants. Leur méthode d'analyse est aussi différente de la notre car ils analysent l'usage d'un site Web à la fois (de manière incrémentale) et non le comportement des internautes sur plusieurs sites dans le même temps.
La structure de cet article est la suivante. La section 2 présente les principales notions ainsi que leur formalisation utilisée dans cet article. La section 3 détaille les différentes étapes de prétraitement. La section 4 illustre notre méthode sur une expérimentation que nous avons réalisée en 2002. Pour conclure, la section 5 présente certaines remarques et perspectives de recherche.

2 Définitions et formalisation du problème

Dans cette section, nous présentons les définitions des principales notions (utilisées dans le domaine du WUM), définitions inspirées du travail de W3C sur la terminologie de caractérisation du Web [LN99].

2.1 Définitions

Ressource - d'après la spécification de W3C pour URI[2], une ressource R peut être "tout objet ayant une identité"[LN99]. Comme exemples de ressources, nous pouvons citer: un fichier**html**, une image ou un service Web.
Ressource Web - une ressource accessible par une version du protocole HTTP ou un protocole similaire (ex. HTTP-NG).
Serveur Web - un serveur qui donne accès à des ressources Web.
Requête Web - une requête pour une ressource Web, faite par un client (navigateur Web) à un serveur Web.
Page Web - ensemble des informations, consistant en une (ou plusieurs) ressource(s) Web, identifiée(s) par un seul URI. Exemple: un fichier HTML, un fichier image et un applet Java accessibles par un seul URI constituent une page Web.
Vue de page (*page view*) - le fait d'afficher une page Web dans l'environnement visuel client à un moment précis en temps. Une vue de page (ou page) peut être composée de plusieurs pages Web et ressources Web comme, par exemple, dans le cas des pages incluant des "cadres".
Navigateur Web (*Browser*) - logiciel de type client chargé d'afficher des pages à l'utilisateur et de faire des requêtes HTTP au serveur Web.

2. Uniform Resource Identifier - "une chaîne de caractères utilisée pour identifier une ressource abstraite ou physique"[LN99]

Utilisateur - personne qui utilise un navigateur Web.
Session utilisateur - un ensemble délimité des clics utilisateurs sur un (ou plusieurs) serveur(s) Web.
Visite(s) - L'ensemble des clics utilisateur sur un seul serveur Web (ou sur plusieurs lorsque on a fusionné leurs fichiers logs) pendant une session utilisateur. Les clics de l'utilisateur peuvent être décomposés dans plusieurs visites en calculant la distance temporelle entre deux requêtes HTTP consécutives et si cette distance excède un certain seuil, une nouvelle visite commence.
Episode - un sous-ensemble de clics liés qui sont présents dans une session utilisateur.*Exemple: pendant une session sur* `yahoo.com`*, l'utilisateur a vérifié son e-mail, a regardé les valeurs des actions en Bourse et a cherché des photos de la Ferrari F60.*Il s'agit dans ce cas de trois épisodes distincts.
Chaque demande d'affichage d'une page Web, de la part d'un utilisateur, peut générer plusieurs requêtes. Des informations sur ces requêtes (notamment les noms des ressources demandées et les réponses du serveur Web) sont stockées dans le fichier log du serveur Web. Il existe plusieurs formats pour les fichiers logs Web, mais le plus courant est le CLF (Common Logfile Format) [Luo95]. Selon ce format six informations sont stockées: 1) le nom ou l'adresse IP de la machine appelante, 2) le nom et le login HTTP de l'utilisateur, 3) la date de la requête, 4) la méthode utilisée dans la requête (GET, POST, etc.) et le nom de la ressource Web demandée, 5) le statut de la requête et 6) la taille du fichier envoyé. Le format ECLF, qui représente une version plus complète du CLF, contient en plus le nom du navigateur Web et le système d'exploitation (cf. "User Agent") et l'adresse de la page où se trouvait avant l'utilisateur, lorsqu'il a lancé la requête (cf. "referrer"). Une ligne d'un log ECLF est présenté ci-dessous:

```
138.96.69.7 - - [30/May/2003:16:43:58 +0200] "GET /axis/people.shtml HTTP/1.0" 200 10677
"http://www-sop.inria.fr/axis/table.html" "Mozilla/4.76 [en] (X11; U; Linux 2.4.20 i686)"
```

2.2 Formalisation du problème

Cette section contient la formalisation du fichier log, du site Web, de la visite d'un utilisateur et de la session serveur ou multi-serveur de l'utilisateur.

- Soit R l'ensemble $\{r_1,r_2,...,r_{n_R}\}$ de toutes les ressources Web d'un site Web.

Si $U = \{u_1,u_2,...,u_{n_U}\}$ est l'ensemble de tous les utilisateurs ayant accédé au site dans une période de temps donnée, alors nous pouvons définir une entrée dans le fichier log Webl_i ansi:
$l_i =< u_i,t,s,r_i,[ref_i] >$, où: $u_i \in U$, t représente la date et l'heure de la requête, s représente le statut[3] de la requête, $r_i,ref_i \in R$, (ref_i est optionnel et il n'est pas présent dans les fichiers logs CLF).
Un **log Web** est défini comme l'ensemble ordonné croissant par la valeur du temps del_i, $L = \{l_1,l_2,...,l_{n_L}\}$.
$\mathcal{L}ogs = \{L_1,L_2,...,L_N\}$ est l'ensemble des logs Web dans le cas de N serveurs Web $\{L_i\}$,$1 \leq i \leq N$.

- Comme mentionné avant, la carte du site peut être utile dans l'étape de prétraitement pour l'identification de pages, visites et épisodes. Dans [Coo00], la carte d'un site Web est formalisée comme suit:

$\mathcal{M} = [< \mathcal{F}_1; ...; \mathcal{F}_n >]$ - la carte du site Web
$\mathcal{F} = \{h_f,\mathcal{L}_1,...,\mathcal{L}_m\}$ - un cadre
$\mathcal{L} =< r,(h_1,g_1)|...|(h_p,g_p) >$ - la liste des liens.
Chaque cadre est constitué d'un fichier h_f, une liste $\mathcal{L}_i$ de liens associés (h_i) et de destinations (g_i). r représente le lien qui décrit la méthode avec laquelle la page peut être demandée (GET, POST etc.).

- Dans l'étape de prétraitement, les entrées du log sont d'abord groupées par vues de pages

3. Par exemple un statut égal à 200 signifie une requête reussie, alors que la valeur 404 du statut signifie que la ressource Web n'a pas été trouvée

en utilisant la carte du site, si disponible.
Une vue de page (ou page) p_i est composée de $p_i = \{r_{i_1}, r_{i_2}, ..., r_{i_P}\}$ où $r_{i_j} \in R$.
L'entrée log devient alors $lp_i =< u_i, t, p_i, [ref_i] >$, où u_i, t, ref_i ont la même signification que dans la définition de l_i et $p_i \in P = \{p_1, p_2, ..., p_{n_P}\}$.
Etant donnée un intervalle de temps Δt, la visite v_{ij} de l'utilisateur u_i est définie:
$v_{ij} =< u_i, t, pv_i >$, où $pv_i =< (t_1, p_1), (t_2, p_2)..., (t_n, p_n) >$, $t_{i+1} \geq t_i$ et $t_{i+1} - t_i < \Delta t$, $i = \overline{1..n-1}$.

- La session serveur ou multi-serveur s_i de l'utilisateur u_i est: $s_i = \{v_{ij}\}$, où v_{ij} est une visite de l'utilisateur u_i.

Formalisation de l'énoncé du problème (pour le prétraitement) Etant donné $\mathcal{L}ogs$ un ensemble de fichiers logs pour un site Web d'une organisation ou d'une entreprise (multi-serveur) et la carte du site $\mathcal{M}$, extraire les visites d'utilisateurs du site pour l'intervalle donné Δt.

3 Méthode de prétraitement pour le WUM multi-sites

La méthode de prétraitement que nous présentons contient dix étapes distinctes dont quatre de nettoyage et six de transformation de données. Cette méthode est supportée par un ensemble de programmes écrits en Perl, dont seules les étapes T4 et T6 n'ont pas encore été implémentées. La succession chronologique de ces étapes est donnée dans le tableau 1.

Unité de traitement	requête				session		visite		
Nettoyage des données		N1	N2a, N2b					N2c	
Transformation des données	T1			T2	T3	(T4)	T5		(T6)

TAB. 1 – *Les étapes du prétraitement dans le WUM multi-sites*

3.1 Nettoyage des données

Le nettoyage des données pour les fichiers logs Web consiste à supprimer les requêtes pour les ressources Web qui ne font pas l'objet de l'analyse (étape N1) et les requêtes ou visites provenant des robots Web (étapes N2).
Pour les portails Web et les sites Web très populaires la dimension de fichiers logs Web est comptée en gigabytes par heure. Par exemple, Yahoo^{TM}, le plus populaire site Web à la date de notre expérimentation [Nie02], collectait presque 100 GB de données logs Web par heure en Mars 2002[4]. Même avec les systèmes et les logiciels de nos jours, manipuler des fichiers de telles dimensions devient très compliqué. Pour cette raison, bien nettoyer ces données avant toute analyse est crucial dans le WUM. Par le filtrage des données inutiles, non seulement on gagne de l'espace disque, mais, dans le même temps on rend plus efficaces les tâches qui suivent dans le processus de WUM. Par exemple, dans le cas des sites Web de l'Inria, en supprimant les requêtes pour les images et les fichiers multimédia, nous avons réduit les dimensions de fichiers logs à 40%-50% de la dimension initiale (cf. tab. 2).

Suppression des requêtes pour les ressource Web non-analysées (N1)
Etant donné l'objectif de notre analyse: étudier le comportement des utilisateurs sur les sites Web d'Inria, nous avons gardé seulement une requête par page visitée (en général pour la page .html) et nous avons supprimé les requêtes auxiliaires comme celles pour les images ou les fichiers multimédia. Quand l'analyste a l'intention de trouver les failles de la structure du site Web ou d'offrir des liens dynamiques personnalisés aux visiteurs du site Web, les informations sur des vues de page sont suffisantes. Par contre, quand l'objectif de l'analyse est le "Web caching" ou le "Web pre-fetching", il ne devrait pas supprimer des telles requêtes du fichier log. Lors de la

4. Nous avons calculé cette valeur en utilisant le nombre de vues de pages pour le mois de mars 2002, 1.62 milliards (cf. [Yah02]), pour le mois de septembre 2000, 760 millions, et la dimension du fichier log pour une heure en septembre 2000, 48GB (cf. [SK02]).

suppression des requêtes pour des images, la carte du site Web peut être employée car, dans certains cas, ces images ne sont pas incluses dans les fichiers HTML. On peut avoir une image qui nécessite de cliquer sur un lien pour l'afficher. Dans un tel cas nous devons maintenir la requête pour cette image dans le fichier log car elle indique une action de l'utilisateur.

Suppression des requêtes et visites provenant des robots Web (N2a,b,c)
Les robots Web (WRs) sont des logiciels utilisés pour balayer un site Web, afin d'extraire son contenu. Ils suivent automatiquement tous les liens d'une page Web. Les moteurs de recherche, comme Google, envoient régulièrement[5] leurs robots pour extraire toutes les pages d'un site Web afin de mettre à jour leurs index de recherche. Le nombre de demandes d'un WR est en général supérieur au nombre de demandes d'un utilisateur normal. En mode paradoxal, si le site Web n'attire pas beaucoup de visiteurs, les demandes faites par tous les WR qui l'ont visité peuvent être supérieures aux demandes faites par des humains. Dans le cas de l'Inria, la taille des requêtes de robots Web a représenté 46.5% de la taille du fichier obtenu après avoir supprimé les requêtes pour les images (cf. tab. 3).
La suppression des entrées dans le fichier log produites par WRs simplifie la tâche de fouille de données qui suivra et permet également de supprimer les sessions non-intéressantes, en particulier en cas de reconception de site. Habituellement, un WR s'identifie en employant le champ "User Agent" dans les fichiers logs. Cependant, aujourd'hui, il est presque impossible de connaître tous les agents qui représentent un WR car chaque jour apparaissent des nouveaux WRs et ceci rend cette tâche très difficile.
Nous avons utilisé trois heuristiques pour identifier les requêtes (N2a, b) ou visites (N2c) issues des WRs:

a. Identifier les IP qui ont fait une requête pour la page `\robots.txt`'. (**N2a**)
b. Utiliser des listes de "User agents" connus comme étant des WR[6]. (**N2b**)
c. Utiliser un seuil pour "la vitesse de navigation" BS ("Browsing Speed"), qui est égale au rapport: $BS(v_{ij}) = |v_{ij}|/(Durée\ de\ la\ visite v_{ij})$. Si $BS(v_{ij}) > 2$ pages/seconde et $|v_{ij}| > 15$ pages, alors la visite v_{ij} vient d'un WR. (**N2c**)

L'étape N2c doit être exécutée après l'étape T5 car, pour calculerBS, il faut que les requêtes soit groupées en visites. Une fois que toutes les requêtes/visites venant des WRs ont été identifiées, nous pouvons procéder à leur suppression.

3.2 Transformation des données

Dans l'étape de transformation des données, nous avons fusionné des fichiers logs Web (**T1**), rendus annonymes les IP (ou les noms des domaines) dans le fichier log obtenu (**T2**) et nous avons groupé les requêtes par sessions (même IP, même User Agent) (**T3**). Ensuite, les sessions ont été divisées en visites en choisissant un $\Delta t = 30min$ (**T5**). Nous avons élaboré deux autres étapes: l'identification des vues des pages (**T4**) et l'identification des épisodes (**T6**) qui sont en cours d'implémentation.

Fusionner les fichiers logs ensemble (T1)
Avant même de commencer le processus de nettoyage, nous avons dû fusionner les différentes fichiers logs de $\mathcal{L}ogs$. Les requêtes de tous les fichiers logs L_i ont été mises ensemble dans un seul fichier L^J. Au préalable e nom du serveur Web de la requête a été ajouté au nom de la ressource Web demandée. Un algorithme qui accomplit cette tâche est présenté dans la suite:

$\mathcal{L}ogs = \{L_1, L_2, ..., L_N\}$
L^J - le fichier log final
$L_i.c$ - un curseur sur les requêtes de L_i

5. Pour Google la période de temps est de quatre semaines.
6. Nous avons utilisé deux liste: la première est disponible sur`http://www.robotstxt.org/wc/robots.html` et la deuxième nous a été gracieusement offerte par le site Web`www.abcinteractive.com`

$L_i.l$ - l'entrée courante dans le fichier log L_i (indiquée par le curseur $L_i.c$)
$T = \{(id,t) | id = \overline{1,N}, t - temps\}$ - une liste ordonnée qui contient les IDs de fichiers logs et le temps de l'entrée courante dans le fichier log
$S = (w_1,w_2,...,w_N)$ - un tableau avec les noms des serveurs Web
$S[i]$ - le nom du serveur Web pour le log i

```
Procedure JoinLogs (Logs)
    T = φ
    for i = 1,N
        L_i.c = 1
        t = L_i.l.time
        InsertT(i,t)
    while T.lenght > 1
        while T[1].t > T[2].t
            id = T[1].id
            l' = S[id] + L_id.l
            L^J = L^J ∪ l'
            L_id.c = L_id.c + 1
            if EOF(L_id)
                RemoveT(T[1])
                break
            end if
        end while
        if not EOF(L_id)
            OrderT(1,T[1])
        end if
    end while
    while not EOF(L_T[1].id)
        id = T[1].id
            l' = S[id] + L_id.l
        L^J = L^J ∪ L_id.l
        L_id.c = L_id.c + 1
    end while
End JoinLogs
```

Rendre anonymes les IP des utilisateurs (T2)

Pour des raisons de confidentialité[7], nous avons remplacé le nom original ou l'adresse IP de la machine appelante avec un identificateur. Toutefois dans le codage de l'identificateur, nous gardons l'information sur l'extension du domaine (pays ou type d'organisation: .com, .org, .edu, etc.). Pour les machines appelantes de l'Inria, nous avons gardé certaines informations comme: le nom de l'unité de recherche et un identifiant pour les équipes de recherche ou les services pour une analyse sur l'usage du site par le personnel de l'Inria.

Identification de l'utilisateur (T3)

L'identification des utilisateurs à partir du fichier log n'est pas une tâche simple en raison de plusieurs facteurs comme: les serveurs proxy, les adresses dynamiques, le cas d'utilisateurs utilisant le même ordinateur (dans une bibliothèque, club Internet, etc.) ou celui d'un même utilisateur utilisant plus d'un navigateur Web ou plus d'un ordinateur. En effet, en employant le fichier log, nous connaissons seulement l'adresse de l'ordinateur (IP) et "l'agent" de l'utilisateur. Il existe d'autres méthodes qui fournissent plus d'informations. Les plus utilisées sont: les "cookies", les pages dynamiques Web (avec un identifiant de session dans l'adresse URL), les utilisateurs enregistrés, les navigateurs modifiés etc. Dans [Coo00], l'auteur utilise les historiques de navigations pour faire la distinction entre deux utilisateurs.
Pour les fichiers logs de l'Inria, nous avons utilisé le couple `(Host, User Agent)` pour l'identification de l'utilisateur. Pour obtenir la session multi-serveur de chaque utilisateur, nous avons ordonné le fichier log par le couple `(Host, User Agent)` et ensuite par le temps. Cette session multi-serveur contient toutes les requêtes de l'utilisateur dans la période analysée. Ensuite, nous avons divisé la session en visites (cf. T5).

Identification des vues de page (T4)

On peut grouper les requêtes par vues de page en utilisant la carte du site Web$\mathcal{M}$. Nous distinguons deux cas:

1. La requête pour la vue de page p_i est présente dans le fichier log. Dans ce cas, les entrées dans le fichier log qui correspond aux ressources contenues en p_i doivent être supprimées

7. Dans le cas où il faut distribuer les fichiers logs pour l'analyse ou pour publier des résultats. Pour coder les noms de machines dans les fichiers logs nous avons utilisé une version du script Lire `-www.logreport.org`

du fichier log et seulement la requête pour p_i est gardée.

2. La requête pour la page p_i manque (à cause du cache du navigateur Web ou d'un serveur de cache) et seulement quelques unes des entrées pour les ressources inclues en p_i sont présentes. Les entrées qui correspondent aux ressources doivent être remplacées par une requête pour p_i. Le temps de cette requêtes est mis égal à $t_i = min\{time(l_i)\}$, où l_i est l'entrée dans le fichier log qui correspond à la ressource r_i.

Après l'exécution de cette procédure, le fichier log contient une seule requête pour une action utilisateur.
Comme la structure du site Web n'était pas disponible au moment de l'expérimentation nous n'avons pas identifié les vues de pages. Depuis nous avons implémenté une méthode pour représenter la structure d'un site Web, $\mathcal{M}$ [ST03], en utilisant XGMML [PK01].

Identification des visites (T5)
Jusqu'à présent, pour un utilisateur u_i nous avons obtenu une séquence de vues de pages $< p_{ij} >$. Ceci représente sa séquence de clics sur un site Web (session serveur) dans une certaine période. Par exemple, considérons l'utilisateur u qui a généré la session serveur suivante:
$v = \{u,16:09:10, < (A,16:09:10),(B,16:09:43),(C,16:12:02),(A,18:32:02),(C,18:33:05),(E,18:47:12),(C,18:48:20),(H,19:15:49),(C,19:51:32) >\}$
En considérant $\Delta t = 30$ *minutes*, largement utilisé comme un seuil temporel standard[8] nous obtenons les trois visites suivantes:
$v_1 = \{u,16:09:10, < (A,16:09:10),(B,16:09:43),(C,16:12:02) >\}$,
$v_2 = \{u,18:32:02, < (A,18:32:02),(C,18:33:05),(E,18:47:12),(C,18:48:20),(H,19:15:49) >\}$ et
$v_3 = \{u,19:51:32, < (C,19:51:32) >\}$.

Identification de l'épisode (T6)
Il existe trois méthodes pour identifier les épisodes: la référence-avant maximale (MF - "Maximal Forward") [CPY96], le typage des pages et la longueur de la référence [Coo00].
Dans la méthode MF, les auteurs ne considèrent pas une deuxième fois les pages qui ont été traversées par l'utilisateur dans sa visite. En utilisant cette méthode, la visite v_2 aura la forme:
$v_2 = \{u,18:32:02, < (A,18:32:02),(C,18:33:05),(E,18:47:12),(H,19:15:49) >\}$.
Cette méthode a un désavantage dans le fait que, pour certaines classes d'applications, il est important de prédire même ces types de référence en arrière.
Les deux méthodes de [Coo00], le typage de pages et la longueur de la référence, sont basées sur la classification de pages. Les pages sont classifiées en "type média" ou en "type auxiliaire". La différence entre les deux méthodes consiste dans l'algorithme de classification. Cet algorithme est basé sur *les données d'usage* pour la méthode longueur de référence et sur *le contenu de la page* pour la méthode typage de page.
Même si, avec ces deux méthodes, l'auteur obtient de meilleures résultats qu'avec la méthode MF, les deux se basent sur des heuristiques pour déterminer la définition sémantique du type de la page.

4 Application dans le contexte de l'Inria

Dans cette section, nous détaillons l'expérimentation que nous avons effectuée pour le site Web de l'Inria en début 2002. Dans cet article, nous ne détaillons pas les étapes de fouille de données et d'analyse de motifs extraits: pour obtenir plus de détails sur ces deux aspects, nous invitons le lecteur intéressé à se rapporter à [TT01].
Comme mentionné avant, le site Web de l'Inria est composé de plusieurs serveurs Web. Les pages

8. Il a été démontré par [CP95] que une valeur de 25.5 minutes est nécessaire pour déterminer les limites d'une visite

servies par ces serveurs sont fortement inter-liées. L'objectif de l'expérimentation était de découvrir les patrons d'accès fréquents entre les pages de tous ces sites. Un patron d'accès fréquent est un motif séquentiel dans un large ensemble de fichiers logs Web, que les utilisateurs ont suivi fréquemment [PHMaZ00].
La machine utilisée pour l'expérimentation est équipée de deux processeurs Pentium III de 1000MHz et dispose de 1GB de mémoire vive.
Nettoyage de données: Les essais ont été effectués sur un ensemble de données consistant en fichiers logs des sites Web d'Inria (`www.inria.fr`, `www-sop.inria.fr` et deux moteurs de recherche). La période analysée s'étend de Novembre 2001 à Janvier 2002. La dimension de l'ensemble de données est de 4432 MB. Dans le tableau 2, nous montrons qu'après l'étape de nettoyage, les fichiers logs sont réduits de moitié par rapport à leur dimension initiale.
Suite à la suppression des requêtes qui proviennent des WRs la dimension des fichiers logs Web sera encore réduite de moitié (cf. tab. 3).

Log \ Mois	**Nov 2001**	**Dec 2001**	**Jan 2002**	**Total**
Logs `www.inria.fr`	897	896	1069	2862
Logs `www-sop.inria.fr`	507	449	598	1554
Logs **moteurs de recherche**	3	7	6	16
Total	1407	1352	1673	4432
Nettoyé	602	680	760	2042
% dimension initiale	42.8	50.3	45.4	46.1

TAB. 2 – *Dimensions de fichiers logs Web de sites d'Inria (en MB)*

Identification d'utilisateurs (sessions) et visites Sur les 10 141 961 requêtes que nous avons eu pour les trois mois analysés, nous avons identifié 569 464 utilisateurs (sessions) différents (couple IP, "User agent"). Les 13 268 IP des WRs représentent seulement 2.33% de ces utilisateurs mais leurs requêtes comptent pour 21.41% du nombre total de requêtes.
Nous avons identifié un total de 971 755 visites. 48.59% de ce visites étaient de longueur 1 et n'ont pas été utilisées pour l'analyse: seules celles de longueur superieure à 2 ont été gardées.

Méthode \ Mois	**Nov 2001**	**Déc 2001**	**Jan 2002**	**Total**	**Total(%)**
Dimension initiale	602	680	760	2402	100
N2a - robots.txt	416	430	518	1364	66.9
N2b - "User agent" connu	535	590	691	1816	88.9
N2c - BS	529	605	691	1825	89.4
Les trois	314	339	440	1093	53.5

TAB. 3 – *Dimensions des fichiers logs Web après la suppression des WRs (MB)*

5 Travaux similaires et conclusions

Pour conclure, la méthode de prétraitement proposée vise à préparer les données pour l'étape de fouille de données qui suit. Les principaux problèmes posés par les fichiers logs Web sont la grande taille de ces fichiers et la faible qualité de ces données. Nous avons vu dans la section 3 comment réduire la taille en éliminant les données inutiles, non-objet de notre analyse (par exemple: les requêtes pour les images et les requêtes des WRs). Aussi, le fait de structurer ces requêtes par sessions, visites et épisodes nous permet d'avoir une meilleure vision de ces données lors de l'analyse.
Il existe beaucoup de travaux récents dans le domaine du WUM. Cependant, à notre connaissance,

très peu sont dédiés à l'étape de prétraitement: [Tan02, Coo00, Yal02]. L'originalité de notre méthode consiste dans l'utilisation de plusieurs heuristiques pour la détection des WRs et surtout dans la prise en compte de l'aspect multi-serveurs dans l'analyse des comportements utilisateurs. Ainsi notre méthode se veut plus globale pour l'analyse de l'usage d'un site (basé sur plusieurs serveurs) d'une organisation ou d'une entreprise.
Dans le futur, nous envisageons poursuivre nos recherches visant la conception d'un véritable entrepôt de données pour le WUM multi-sites intégrant les fichiers logs Web, les informations sur les structure des sites et sur leurs utilisateurs.

Références

[Coo00] R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, May 2000.

[CP95] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995.

[CPY96] M. S. Chen, J. S. Park, and P. S. Yu. Data mining for path traversal patterns in a web environment. In *Sixteenth International Conference on Distributed Computing Systems*, pages 385–392, 1996.

[LN99] Brian Lavoie and Henrik Frystyk Nielsen. Web characterization terminology & definitions sheet. http://www.w3c.org/1999/05/WCA-terms/, May 1999.

[Luo95] A. Luotonen. The common log file format. http://www.w3.org/pub/WWW/, 1995.

[MTP01] Florent Masseglia, Maguelonne Teisseire, and Pascal Poncelet. Web usage mining inter-sites : Analyse du comportement des utilisateurs à impact immédiat. In *17^{ièmes} Journèes Bases de Données Avancées (BDA 2001)*, Agadir, Maroc, October 2001.

[Nie02] Search Engines, Portals and Communities are the Most Popular Online Category Worldwide, According to Nielsen//NetRatings Global Index. http://www.nielsen-netratings.com/pr/pr_020228_eratings.pdf, March 2002.

[PHMaZ00] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu. Mining access patterns efficiently from web logs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* pages 396–407, 2000.

[PK01] John Punin and Mukkai Krishnamoorthy. XGMML (eXtensible Graph Markup and Modeling Language). http://www.cs.rpi.edu/~puninj/XGMML/draft-xgmml.html, June 2001.

[SK02] Cyrus Shahabi and Farnoush Banaei Kashani. A framework for efficient and anonymous web usage mining based on client-side tracking. In *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001, Revised Papers*, pages 113–144. Springer, 2002.

[ST03] A. Santoni and D. Tanasa. Construction de la structure logique d'un site web. Technical report, Inria Sophia Antipolis, Equipe AxIS, 2003.

[Tan02] D. Tanasa. Lessons from a web usage mining intersites experiment. In *First International Workshop on Data Cleaning and Preprocessing*, Maebashi, Japon, 9 December 2002. ICDM02.

[TT01] Doru Tanasa and Brigitte Trousse. Web access pattern discovery and analysis based on page classification and on indexing sessions with a generalised suffix tree. In *3rd International Workshop on Symbolic and Numeric Algorithms for Scientific Computing* pages 62–72, Timisoara, Romania, October 2001.

[Vie02] The one-stop portal. http://www.line56.com/articles/default.asp?ArticleID=4075, October 8, 2002.

[Yah02] Yahoo! financial metrics. http://docs.yahoo.com/info/pr/investor_metrics.html, March 2002.

[Yal02] Berkan Yalçinkaya. Predicting next page access by time length reference in the scope of effective use of resources. Master's thesis, Bilkent University, Department of Computer Engineering, September 2002.

SurfMiner : Associer des données personnelles à des navigations sur un site

K. CHEVALIER
France Telecom R&D, DMI/GRI,
2 Av. Pierre Marzin,
22300 Lannion, FRANCE
LIP6,
Université Paris VI,
Paris, France,
Mail : karine.chevalier@francetelecom.com
karine.chevalier@lip6.fr

C. BOTHOREL
France Telecom R&D, DMI/GRI,
2 Av. Pierre Marzin,
22300 Lannion, FRANCE
Mail : cecile.bothorel@francetelecom.com

V. CORRUBLE
LIP6,
Université Paris VI,
Paris, France,
Mail : vincent.corruble@lip6.fr

Résumé

Chercher à caractériser les internautes fréquentant son site devient une pratique de plus en plus courante et indispensable pour offrir des services personnalisés et ainsi fidéliser ses visiteurs.

Dans cet article, nous décrivons une nouvelle méthode (SurfMiner) d'acquisition de connaissances sur les utilisateurs d'un site donné. Notre méthode repose sur l'idée que la population d'un site peut être divisée en groupes d'utilisateurs, chaque groupe ayant des pratiques différentes sur le site et ayant des caractéristiques personnelles différentes, ces deux aspects étant corrélés dans une certaine mesure.

Notre méthode consiste à identifier et qualifier des groupes d'utilisateurs par rapport à des usages sur un site et par rapport à des traits socio-démographiques ou des traits représentant des centres d'intérêts.

Abstract

In order to offer a personalized service, it is now crucial to be able to characterize the users of a web site.

In this paper, we describe a new method (SurfMiner) for the discovery of knowledge on the users of a site. Our method is based on the idea that the population of a site can be divided in a number of coherent users groups, and that each group shows a distinct behaviour on the site, has distinct personal characteristics, and that the two are correlated in some way.

Our method consists in creating a number of user categories based on rich descriptions and site usage.

1 Introduction

Les internautes sont au cœur des préoccupations des concepteurs de sites web : il est important de savoir quels documents ou pages ils consultent sur le site et qui ils sont, afin de mieux construire le site et d'adapter son contenu.

Il existe de nombreuses méthodologies pour mettre en évidence la fréquentation d'un site et qualifier son audience. On peut distinguer deux grandes approches qui répondent de façon complémentaire à ce problème : une approche centrée sur les utilisateurs et une approche centrée sur le site.

L'**approche centrée sur les utilisateurs** désigne les méthodes analysant les activités d'un échantillon d'utilisateurs sur Internet pour comprendre les usages de l'Internet. Ces analyses sont effectuées par quelques sociétés (Media Metrix [Media Metrix] et NetValue [NetValue]). Ces techniques procèdent de la manière suivante : un panel d'utilisateurs est créé reflétant au mieux la population des internautes ; d'un côté, des données personnelles sont recueillies auprès de chaque panéliste (âge, sexe, ancienneté sur le web...) ; de l'autre côté, toutes les activités sur Internet des panélistes sont suivies et capturées à l'aide d'un logiciel implanté dans leur ordinateur ; enfin, une analyse est opérée sur ces données afin de décrire les usages d'Internet par rapport aux profils de ces internautes. Les observations faites sur ce panel sont extrapolées à l'ensemble des internautes.

Du point de vue des sites, cette approche offre dans le meilleur des cas une description de leur audience et leur donne des moyens de comparaison avec d'autres sites. Mais seuls quelques sites peuvent bénéficier de cette connaissance : en effet ces méthodes donnent une mesure peu fiable pour les sites peu fréquentés, leurs utilisateurs ayant peu de chances d'être représentés dans le panel. L'intérêt de telles méthodes est avant tout de montrer les tendances sur Internet : les usages et les thèmes à la mode.

L'**approche centrée sur le site** désigne les méthodes d'analyse d'usage obtenue à partir des logs d'un site (quelques outils industriels : [WebTrends], [Fireclick]). Elle consiste à collecter sur un site les navigations de tous les utilisateurs, puis à analyser ces données. Elle permet de quantifier la fréquentation des pages d'un site et de mettre en évidence des parcours de navigation typiques sur le site. Les connaissances déduites de cette méthode restent tout de même assez pauvres, en particulier peu d'informations sont obtenues sur les utilisateurs. Nous détaillerons dans la section 2 quelques méthodes d'analyse de log.

Dans cet article, nous présentons la méthode SurfMiner qui étudie la possibilité de corréler des informations de type profil avec des navigations utilisateurs sur site. SurfMiner permet de découvrir des connaissances sur les utilisateurs d'un site qualifiant à la fois leurs usages et leur profil. Cette méthode combine les deux approches citées précédemment :

- A la manière de l'approche centrée sur les utilisateurs, les données personnelles proviennent d'un groupe d'internautes mais ici ce groupe est limité à des utilisateurs d'un site et le suivi se fait seulement sur ce site.
- A la manière de l'approche centrée sur le site, l'usage du site est extrait en analysant les navigations effectuées sur le site, mais ici les logs analysés sont limités au groupe d'utilisateurs pris en référence.

Dans la prochaine section, nous revenons sur les techniques d'extraction des usages à partir des sessions. La troisième section décrit la méthode SurfMiner. La quatrième partie présente les premiers résultats obtenus et les premières évaluations de la méthode.

2 Extraire des usages à partir des navigations sur un site

De nombreux travaux se sont intéressés à l'extraction des usages à partir d'un ensemble de sessions sur un site donné. Nous définissons une session comme étant une visite sur un site effectuée par un utilisateur. Plus simplement, une session est une suite de pages ordonnée dans le temps. Ces sessions sont extraites des fichiers logs contenant les traces de navigation effectuées sur un site. L'article [Cooley et al. 1999(a)] fait une revue des problèmes et des techniques pour convertir les fichiers logs en sessions.

Dans [Borges et Levene, 1998] et [Borges et Levene, 1999], les auteurs proposent une méthode pour découvrir des chemins de navigation dans un site. Un graphe orienté et valué est construit à partir

des sessions des utilisateurs : un arc correspond à un lien entre deux pages, et le poids associé à un arc correspond à la probabilité que ce lien soit suivi. Les chemins de navigation "préférés" sont les chemins ayant une forte probabilité. De même, l'algorithme WebSPADE [Demiriz 2002] (adaptation de l'algorithme SPADE [Zaki2001]) permet d'extraire les séquences fréquentes de pages dans un jeu de sessions. Dans les deux dernières méthodes, le chemin ou la séquence est pertinent parce qu'il est fréquent et représente un usage général du web ou du site. Ces méthodes peuvent donc passer à coté d'informations : un type de navigation particulier peut être noyé dans toutes les navigations et ne pas être mis en évidence.

Ils existent d'autres approches qui ne reposent pas seulement sur la fréquence. Dans [Cooley et al., 1999(b)], les auteurs décrivent l'algorithme BME (Beliefs Mined Evidence) qui s'appuie sur le fait que tous les ensembles fréquents de pages co-occurrentes ne sont pas tous pertinents. Un ensemble de pages est considéré comme intéressant s'il contient des pages qui ne sont pas connectées directement l'une à l'autre (pas de lien HTML entre elles et pas de contenus similaires). Dans un premier temps tous les motifs de navigation sont extraits des sessions des utilisateurs. L'algorithme BME ne garde que les motifs qui ne sont pas évidents (pas de contenu commun et pas de relation structurelle entre ces pages). Cette méthode sélectionne les ensembles fréquents les plus pertinents selon leur stratégie de non-évidence, cependant elle ne permet pas de retrouver des usages particuliers du site. L' analyse reste encore trop générale car elle s' appuie sur des connaissances issues de méthodes basées sur la fréquence. L' usage spécifique à un petit groupe d' utilisateurs n' est pas reconnu dans ces analyses.

Spiliopoulou et al. [Spiliopoulou et al. 1998] ont créé un outil supervisé WUM (Web Utilization Miner) d'aide à l'analyse d'usage et à la découverte de motifs de navigation. Dans un premier temps un arbre est construit à partir des sessions des utilisateurs : les chemins ayant le même préfixe sont regroupés sous le même nœud. Les motifs de navigation sont repérés par l'outil suivant les critères entrés par un expert du site grâce au langage MINT d'interrogation. WUM est donc un outil qui permet à un expert de faire une lecture de l'arbre. Ici, tous les usages du site sont repérés mais seul un expert peut déterminer les usages pertinents.

Une autre approche ([Hay et al. 2001]) consiste à appliquer un clustering sur les sessions des utilisateurs afin de découvrir des groupes de sessions. Une session est considérée comme une suite de pages ordonnées dans le temps. La similarité entre deux sessions est évaluée grâce à une distance d'édition. La distance d'édition entre deux sessions correspond à la quantité de travail nécessaire pour convertir une session en une autre. Le clustering est appliqué sur l'ensemble des sessions en utilisant la distance d'édition pour évaluer une similarité entre deux sessions. Ces groupes de sessions montrent chacun un usage du site. Il suffit ensuite de caractériser ([Hay 2001]) les clusters de session obtenus en terme de motifs de navigation pour obtenir ces usages. La précision de la description des usages dépend dans cette approche du nombre de clusters créés. Si le nombre de clusters est insuffisant, nous retombons sur les problèmes des méthodes basées sur la fréquence : l' analyse d' usage ne peut extraire que les usages généraux du site.

Toutes ces méthodes sont fondées sur l'analyse des sessions des utilisateurs sur un site. L'usage du site n'est décrit qu'en terme de chemins de navigation ou de pages consultées, on ne sait pas quels types d'utilisateurs fréquentent les différentes parties du site. De plus les motifs de navigation découverts par ces méthodes sont généraux et ne reflètent alors que des usages généraux lissés, ils ne tiennent pas compte des particularités.

3 Méthode SurfMiner

La méthode SurfMiner propose une analyse mettant en évidence les usages du site associés à des descriptions d'utilisateurs. Cette méthode repose sur l'hypothèse (forte) qu'il existe des corrélations entre la façon de naviguer sur un site et des traits décrivant les utilisateurs. Nous pouvons décomposer cette hypothèse en deux sous-hypothèses :

- Si deux utilisateurs sont similaires du point de vue socio-démographique ou du point de vue de leurs centres d'intérêts, leurs navigations sur un site sont similaires.
- Si deux utilisateurs ont des navigations similaires alors ils sont similaires du point de vue socio-démographique ou du point de vue de leurs centres d'intérêts.

SurfMiner cherche à extraire deux types de connaissances :

- Des motifs de navigation[1] enrichis de traits d'utilisateurs. Par exemple : 'la plupart des utilisateurs consultant la page A puis la page D du site sont des cadres supérieurs'. Ces motifs enrichis peuvent être vus comme des règles du type : "motif -> trait d'utilisateur", règles auxquelles sont associés un support et une confiance.
- Des motifs de navigation spécifiques à un groupe d'utilisateurs. 'Les femmes ont tendance à visiter la page E et la page F du site'. Ces motifs spécifiques peuvent être représentés par des règles du type "trait d'utilisateur -> motif" associées à un support et une confiance.

L'apprentissage de SurfMiner s'effectue sur un jeu de données dédiées à la méthode, des données plus riches que de simples logs d' utilisateurs. Ce jeu de données est constitué,comme pour les méthodes citées en partie 2, de l'ensemble des sessions sur un site donné d'un groupe d'utilisateurs. Par contre, SurfMiner dispose en plus de la description de chacun de ces utilisateurs. Chaque utilisateur est ainsi décrit par un ensemble de données personnelles (25 ans, profession, ..) et par un ensemble de sessions.

L'apprentissage de SurfMiner se fait de deux manières complémentaires :

- En extrayant les motifs de navigations de l'ensemble des sessions puis en cherchant à associer aux motifs découverts des traits d'utilisateurs. Nous avons intitulé ce premier mode d'apprentissage "SurfMiner *pattern2profile*".
- De manière symétrique, en regroupant les utilisateurs par rapport à leurs descriptions puis en découvrant pour chaque groupe les motifs fréquents de navigation présents dans les sessions des utilisateurs du groupe. Nous avons intitulé ce second mode d'apprentissage "SurfMiner *profile2pattern*".

Les connaissances acquises par ces deux processus sont dédiées au site sur lequel a été élaboré le jeu de données. Nous détaillons les deux modes d'apprentissage de SurfMiner dans les deux sections suivantes.

3.1 SurfMiner pattern2profile

SurfMiner *pattern2profile* commence par extraire les motifs de navigation de l'ensemble des sessions du jeu de données puis cherche à associer aux motifs découverts des traits d'utilisateurs.

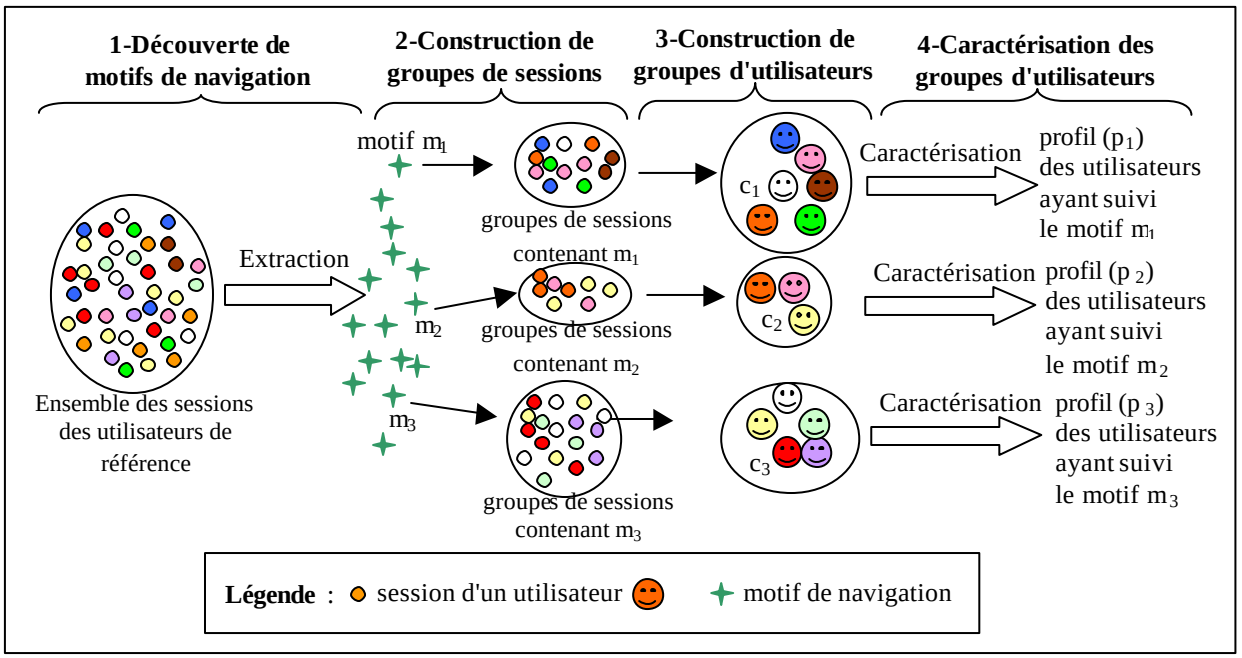


Figure 1. SurfMiner pattern2profile

[1] Nous appelons motif de navigation : un ensemble fréquent ou une séquence fréquente de pages. Un ensemble fréquent de pages est un groupe de pages apparaissant ensemble chez plus d'un nombre minimum d'utilisateurs.

Nous décomposons ce processus en quatre phases :
- Découverte de motifs fréquents de navigation
- Construction de groupes de sessions relativement aux motifs découverts
- Construction de groupes d'utilisateurs relativement aux groupes de sessions précédents.
- Caractérisation des groupes d'utilisateurs grâce aux descriptions de ces utilisateurs fournies par le jeu de données.

La figure 1 décrit les quatre étapes de SurfMiner *pattern2profile*. Nous détaillerons les différentes phases dans les paragraphes suivants.

3.1.1 Etape 1 : Découverte de motifs de navigation

La première étape consiste à mettre en évidence les motifs de navigation présents dans l'ensemble des sessions. Les motifs de navigation trouvés représentent l'usage du site de tous les utilisateurs de référence. Les motifs de navigation recherchés correspondent à des ensembles fréquents de pages et à des séquences fréquentes de pages.

Nous utilisons un algorithme du type Apriori [Agrawal et Srikant, 1994] pour retrouver les ensembles fréquents de pages. Cet algorithme repose sur l'idée suivante : "un ensemble fréquent est composé de sous-ensembles fréquents". Nous avons choisi d'associer à chaque ensemble de pages la liste des sessions (sous la forme d'une liste d'identifiants) dans lesquelles elle apparaît, afin d'éviter de parcourir le jeu de sessions à chaque calcul d'occurrence (FreeSpan [Han 2000], PrefixSpan [Pei 2001], Spade [Zaki 2001], et WebSpade[Demiriz 2002]). Le principe est le suivant : (1) construire pour chaque page une liste des identifiants de sessions dans lequel elle apparaît ; (2) ne garder que les pages présentes chez plus de *minOccurrence* utilisateurs ; (3) construire à partir de ces pages fréquentes, les ensembles candidats de deux pages, la liste des identifiants d'un ensemble-candidant correspond à l'intersection des listes des deux pages fréquentes le composant ; (4) ne garder que les ensembles apparaissant chez plus de *minOccurrence* utilisateurs ; (5)…

Nous avons préféré calculer la fréquence d'un ensemble de pages par rapport au nombre d'utilisateurs (plutôt qu'au nombre de sessions) afin d'éviter que les pratiques d'un utilisateur cachent les pratiques des autres utilisateurs.

De la même manière, nous utilisons un algorithme similaire pour extraire les séquences fréquentes, celui-ci prend en compte l'ordre des pages dans le calcul.

3.1.2 Etape 2 : Construction de groupes de sessions

Nous construisons ensuite les groupes de sessions relativement à un motif de navigation. Un groupe de sessions contient toutes les sessions dans lesquelles apparaît le motif.

3.1.3 Etape 3 : Construction de groupes d'utilisateurs

La troisième étape consiste à créer, à partir de ces groupes de sessions, des groupes d'utilisateurs. Pour chaque groupe de sessions, un groupe d'utilisateur est construit contenant tous les utilisateurs possédant au moins une session dans ce groupe de session. A la fin de cette étape, nous obtenons des groupes d'utilisateurs réunis autour d'un même motif de navigation.

3.1.4 Etape 4 : Caractérisation des groupes d'utilisateurs

La quatrième phase (enrichissement d'un motif par des traits d'utilisateurs) a pour but de caractériser les groupes d'utilisateurs (réunis autour d'un motif) par des traits personnels. Elle consiste pour un groupe d'utilisateurs donné à découvrir des traits d'utilisateurs partagés par un nombre suffisant d'utilisateurs de ce groupe. Ce nombre d'utilisateurs représente la confiance de la caractérisation.

Cette étape peut générer beaucoup de caractérisations peu pertinentes comme des caractérisations qu'on pouvait déjà observer dans le groupe initial des utilisateurs. Autrement dit, ces caractérisations n'apportent pas d'information par rapport aux connaissances directement déductibles de la répartition de la population de notre panel selon les différentes valeurs des traits. Ces caractérisations sont donc

inutiles à produire, car auront un apport nul. Il est donc nécessaire d'utiliser un critère de pertinence pour écarter les caractérisations non significatives.

La pertinence d'une caractérisation est évaluée en comparant la proportion des utilisateurs ayant ce trait dans le groupe d'utilisateurs (que l'on cherche à caractériser) et la proportion d'utilisateurs ayant ce trait dans la population totale des utilisateurs. Si les deux proportions sont significativement différentes, alors la caractérisation est pertinente. Pour cela, nous calculons l'écart réduit (**e**) entre les deux proportions. L'écart entre les deux proportions est significatif quand, pour un risque a= 0,05 alors e>= 1,96.

$$e = \frac{p_o - p_t}{\sqrt{\frac{p_t(1-p_t)}{n}}} \quad \text{avec} \begin{cases} e : \text{écart réduit entre la proportion observée et la proportion théorique} \\ p_o : \text{proportion observée (proportion du trait dans le groupe d'utilisateurs)} \\ p_t : \text{proportion théorique (proportion du trait dans la population totale)} \\ n : \text{nombre d'utilisateurs chez le groupe observé} \end{cases}$$

Un groupe d'utilisateurs réunis autour d'un motif de navigation peut être caractérisé de multiples manières. A la fin de cette phase, un groupe d'utilisateurs est décrit par un motif de navigation et par des traits d'utilisateurs.

Cette méthode produit une liste de motifs de navigation enrichis de traits d'utilisateurs, c'est-à-dire une liste d'associations entre des motifs de navigation et des profils d'utilisateurs.

3.1.5 Conclusion

Le mode d'apprentissage *pattern2profile* de la méthode SurfMiner a pour principal objectif d'enrichir des motifs de navigation extraits de l'ensemble des sessions par des traits d'utilisateurs. SurfMiner *pattern2profile* est une prolongation naturelle de l'analyse d'usage d'un site à partir des logs : les méthodes classiques permettent de regrouper des utilisateurs par rapport aux motifs de navigation, SurfMiner *pattern2profile* enrichit cette découverte en caractérisant ces groupes par des traits d'utilisateurs. Nous obtenons à la fin de cette étape des règles d'association entre des motifs et des traits d'utilisateurs : "motif -> trait de l'utilisateur"

Les motifs de navigation sont sélectionnés par SurfMiner *pattern2profile* parce qu'ils apparaissent fréquemment. Les motifs peu répandus mais significatifs pour un groupe d'utilisateurs ne peuvent pas être récupérés avec ce mode d'apprentissage (même défaut que les techniques exposées en partie 2). Le mode d'apprentissage *profile2pattern* de la méthode SurfMiner (décrit dans la section suivante) a pour rôle entre autres de capturer ces motifs particuliers à des groupes d'utilisateurs.

3.2 SurfMiner profile2pattern

Nous décrivons dans cette section le deuxième mode d'apprentissage de SurfMiner : *profile2pattern*. La description est abrégée, nous supposons que les concepts importants ont été abordés dans la section précédente.

SurfMiner *profile2pattern* commence par regrouper les utilisateurs de référence par rapport à des traits d'utilisateurs puis cherche à mettre en évidence, pour chaque groupe d'utilisateurs de référence, les motifs de navigation présents dans leurs sessions. Ce processus se déroule en trois phases :

- Construire des clusters d'utilisateurs par rapport à leurs données personnelles : plusieurs partitions doivent être effectuées car nous ne savons pas quels traits (ou combinaisons de traits) sont les plus pertinents pour regrouper des utilisateurs de référence par rapport à l'ensemble des descriptions d'utilisateurs et l'usage d'un site. Nous avons choisi de créer une partition des utilisateurs pour chaque attribut décrivant les utilisateurs. Dans le cas des attributs de type nominal (ou ordinal), l'algorithme se contente de construire pour chaque modalité de l'attribut le groupe des utilisateurs ayant comme valeur d'attribut cette modalité. Dans le cas des attributs de type continu, la stratégie est d'appliquer un algorithme de type Kmeans afin de discrétiser les valeurs de l'attribut. Les groupes des utilisateurs générés à partir d' un attribut sont filtrés afin de ne retenir que les groupes contenant au moins le seuil minimum d' utilisateurs.

- Construire des groupes de sessions : un groupe de session est associé à chaque cluster d'utilisateurs; il correspond à l'union de l'ensemble des sessions de tous les utilisateurs appartenant à ce cluster. A la fin de cette étape, un cluster d'utilisateurs est décrit par un profil d'utilisateur et un ensemble de sessions lui est associé.
- Découvrir des motifs de navigations dans chaque groupe de sessions : nous utilisons les algorithmes décrits dans la section 3.2.1 appliqués sur chaque groupe de session. Ces motifs caractérisent chacun des groupes. Cette étape peut générer beaucoup de caractérisations peu pertinentes comme des motifs qu'on pouvait déjà observer dans le groupe initial des utilisateurs. Ces caractérisations sont donc inutiles à produire. Il est donc nécessaire d'utiliser un critère de pertinence pour écarter les caractérisations non significatives.(cf section 3.1.4)

La figure 2 décrit les trois étapes de SurfMiner *profile2pattern*.

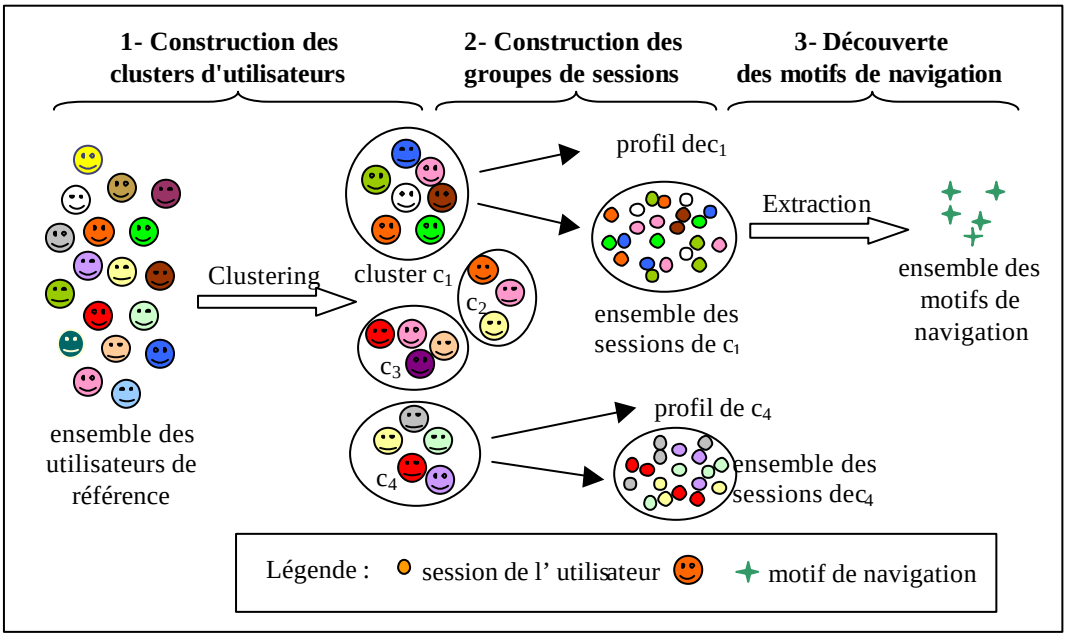


Figure 2. SurfMiner profile2pattern

En fin de traitement, chaque cluster d'utilisateurs est décrit par un profil d'utilisateur et un ensemble de motifs de navigation. Cette méthode génère une liste d'associations entre des profils d'utilisateurs et des motifs de navigation.

Le mode d'apprentissage *profile2pattern* de la méthode SurfMiner cherche à mettre en évidence des motifs de navigation invisibles pour des analyses classiques d'usage à partir de logs parce que ces motifs de navigation ne sont pas assez fréquents. La stratégie de SurfMiner *profile2pattern* est de diviser l'espace de recherche des motifs de navigation pour appliquer l'extraction sur de plus petits ensembles de sessions. Cette partition se fait par rapport aux traits d'utilisateurs. Cette division de l'espace de recherche n'est valide que si on fait l'hypothèse 1 : 'Si deux utilisateurs sont similaires du point de vue socio-démographique ou du point de vue de leurs centres d'intérêts, leurs navigations sur un site sont similaires'.

4 Résultats et Evaluation de la méthode SurfMiner

Nous expérimentons et évaluons la méthode SurfMiner sur des données d'usages d'Internet. Celles-ci proviennent des traces de trafic Internet d'une cohorte d'un millier d'internautes extraite du panel NetValue[2] en 2000, exploitées et enrichies dans le cadre d'un partenariat entre France Télécom R&D et NetValue sur les usages d'Internet [Beaudouin, Assadi et al., 2002].

[2] NetValue est une société de mesure d'audience sur Internet.

Les utilisateurs sont décrits par 35 traits : des traits fournis directement par les utilisateurs comme l'âge, l'occupation professionnelle, la taille (en nombre d'habitants) de leur ville, ou encore la date de leur première connexion à Internet; des traits déduits de leurs activités sur Internet comme le nombre total de sessions Internet effectuées durant l'année 2000, le nombre de requêtes moteur faites en 2000 ou encore leur type du point de vue des services de communication Internet qu'ils utilisent.

Les données de navigation ont été antérieurement prétraitées afin que chaque requête http soit enrichie d'un identifiant de session et l'identifiant de l'utilisateur. Elle comprend plusieurs champs comme la date, l'url demandée …L'ensemble des sessions est donc facilement déductible de ces données. Une session est une suite d'urls ordonnée dans le temps (cf. [Beaudouin, Assadi et al., 2002] pour une description détaillée des méthodes et outils utilisés pour réaliser ces traitements).

Nous avons construit 5 jeux de données correspondant à l'ensemble des sessions effectuées en 2000 sur 5 sites différents (anpe.fr, boursorama.com, liberation.fr, mp3.com et voila.fr) et l'ensemble des descriptions des utilisateurs, auteurs de ces sessions. Nous avons décomposé les données en deux ensembles distincts :

- Un ensemble d'apprentissage utilisé pour faire l'apprentissage de motifs de navigation enrichis de descriptions d'utilisateurs et de motifs de navigation spécifiques à un groupe d'utilisateur. Cet ensemble contient les descriptions et les traces de navigation d'à peu près 80% des panélistes choisis de façon aléatoire.
- Un ensemble de test permettant d'estimer la validité des motifs trouvés. Cet ensemble contient les descriptions et les traces de navigation des 20% de panélistes restants.

4.1 Application de SurfMiner sur les jeux de données

Nous avons fixé le support minimum à 5% et la confiance minimum à 50%. Le tableau suivant montre le nombre de règles obtenues pour chaque site.

	anpe.fr (140 utilisateurs, 1179 sessions)	**boursorama.com** (191 utilisateurs, 6557 sessions)	**liberation.fr** (146 utilisateurs, 901 sessions)	**mp3.com** (155 utilisateurs, 364 sessions)	**voila.fr** (657 utilisateurs, 13797 sessions)
motif -> trait	77074	14773	450	724	483
trait -> motif	1432	26	5	7	10

Tableau 1. Quantité de règles obtenues pour chaque site

Le tableau suivant présente le nombre de règles de type "motif -> trait" générées par rapport à un support minimum donné et une confiance minimum donnée, sur le jeu de données concernant le site mp3.com.

mp3.com motif -> trait		**Confiance**			
		50	**60**	**70**	**80**
Support	**5**	724	522	431	430
	10	157	115	81	78
	15	61	40	27	26
	20	55	34	23	21

Tableau 2. nombre de règles "motif ->trait" générées pour le site mp3.com

L'encadré ci-dessous montre une règle du type "motif -> trait" associé à une confiance donnée.

ensemble{*www.anpe.fr/recherch/index.htm*,*www.anpe.fr/regions/accueil.htm* } Þ {SEXE=femme}
avec confiance=0.87

Ce type de règles peut être interprété de la manière suivante : les utilisateurs pour lesquels le motif est observé possèdent la valeur du trait indiquée dans la règle avec la confiance donnée. Par exemple, la règle suivante signifie que les utilisateurs qui consultent les pages "*www.anpe.fr/recherch/index.htm*" et "*www.anpe.fr/regions/accueil.htm*" sont à 87% des femmes.

{AGER=35-49 ans} Þ séquence{www.anpe.fr -> www.anpe.fr/accht.htm -> www.anpe.fr/offremp/index.htm}
avec confiance=0.55

Inversement, les règles du type "trait -> motif" avec une confiance donnée peuvent être interprétées comme suit : les utilisateurs possédant la valeur du trait indiquée dans la règle suivent le motif avec la confiance donnée. Par exemple, la règle dans l'encadré suivant signifie que les utilisateurs appartenant à la tranche d'âge 35-49 ans consultent dans 55% des cas la séquence de pages *www.anpe.fr , www.anpe.fr/accht.htm* et *www.anpe.fr/offremp/index.htm*.

4.2 Evaluation des règles motif -> trait

Nous présentons dans ce paragraphe une première évaluation des connaissances générées par la méthode SurfMiner. Nous cherchons à évaluer la prédiction des règles du type "motif→trait". Nous avons testé le pourcentage de bonnes prédictions des règles sur l'ensemble de test de chaque jeu de données. Ce test consiste à utiliser les règles issues de SurfMiner sur chaque session (et à chaque moment de la session) pour prédire le profil de l'utilisateur, auteur de la session. Une règle est activée lorsque le motif peut être observé à un moment donné de la session. Nous comparons ces chiffres avec une prédiction aléatoire tenant compte de la répartition des individus par rapport aux différents traits. (prédiction de référence, calculable sans SurfMiner). Lorsqu'une règle SurfMiner est activée donnant lieu à une prédiction sur la valeur d'un trait alors une prédiction aléatoire est également générée.

	% prédictions correctes avec les règles issues de SurfMiner	Nombre total de prédictions avec les règles issues de SurfMiner	% prédictions aléatoires correctes	Nombre total de prédictions aléatoires
Mp3.com	44.78%	5062	23,40%	5025
Libération	42,61%	1217	27,32%	12036
Voilà.fr	65,60%	44575	23,43%	44554
Boursorama.fr	55,39%	533926	21,26%	528406
Anpe.fr	48,63%	918657	27,20%	808293

Tableau 3. Comparaison entre les prédictions des règles "motif ->trait" de SurfMiner et des prédictions aléatoires

Ce tableau montre que sur tous les jeux de données testés, les règles issues de SurfMiner donnent de meilleures prédictions que le hasard.

4.3 Conclusion

Nous sommes au début des évaluations de la méthode SurfMiner. Il nous reste à évaluer la qualité des règles issues de SurfMiner : sont-elle pertinentes ? Quel est l'apport de ces connaissances par rapport à des motifs classiques de navigation et à la connaissance des profils des utilisateurs d'un site ?

5 Conclusion et perspectives

Cet article présente une nouvelle méthode pour obtenir des connaissances sur les utilisateurs d'un site donné. La méthode SurfMiner cherche à mettre en évidence les usages du site associés à des descriptions d'utilisateurs. Le principal problème de cette méthode réside dans le fait d'enrichir des navigations grâce à des données personnelles : en effet, ces deux types d'information ne sont pas, de

façon évidente, "liables". Les données personnelles d'un utilisateur sont des données externes et par conséquent décorrelées du contenu d'un site (par exemple le sexe, la catégorie socio-professionnelle, etc.). On peut alors se demander s'il y a adéquation entre le profil des utilisateurs et leur parcours sur un site.

L'originalité de la méthode par rapport à des analyses classiques d'usage est donc d'introduire des descriptions d'utilisateurs dans le processus d'acquisition des motifs de navigation. Elle s'appuie sur l'hypothèse qu'il existe des corrélations entre les pratiques sur un site (mises en évidence par les motifs de navigation) et des traits d'utilisateur. SurfMiner extrait des motifs de navigation enrichis de traits d'utilisateurs et des motifs spécifiques à des groupes d'utilisateurs (partageant des traits communs). Les données utilisées pour évaluer la méthode SurfMiner proviennent de données centrées utilisateur du type panel. La principale difficulté de cette méthode est l'obtention ou la constitution de données aussi riches. Dans le cas d'une utilisation "réelle", il faudra réfléchir à l'acquisition des données.

SurfMiner est en cours d'évaluation, nous projetons notamment d'évaluer l'apport, dans un système de recommandation, des motifs de navigation enrichis et spécifiques issus de SurfMiner par rapport à une extraction classique de motifs de navigation.

Si notre approche obtient de bons résultats, de nombreuses applications sont possibles dans des systèmes de personnalisation, en particulier pour des outils d'aide à la navigation, des recommandations de pages ou de services. La personnalisation ne nécessitera pas d'informations personnelles de la part des utilisateurs, puisqu'elles seront déduites de sa navigation sur le site. Un visiteur anonyme pourra avoir un site adapté à sa navigation et bénéficier des expériences des internautes déjà passés sur ce site.

Références

[Beaudouin V., Assadi H. et al. 2002] Beaudouin V., Assadi H., Beauvisage T., Lelong B., Licoppe C., Ziemalicki C., Arbues L., Lendrevie J. (2002). Parcours sur Internet : analyse des traces d'usage. *Rapport RP/FTR&D/7495, France Telecom R&D, Net Value, HEC.*

[Borges et al., 1999] Borges, J., Levene, M. (1999) Data Mining of User Navigation Patterns. *In Proceedings of the Workshop on Web Usage Analysis and User Profiling, pages 31-36. August 15,1999, San Diego, CA.*

[Cooley et al. 1999(a)] Cooley, R., Mobasher, B., Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *In Knowledge and Information System, 1(1):5-32, 1999.*

[Cooley et al., 1999(b)] Cooley, R., Tan, P., Srivastava, J. (1999). WebSIFT: The Web Site Information Filter System. *In Proceedings of the Web Usage Analysis and User Profiling Workshop, August 1999.*

[Demiriz et al., 2002]Demiriz, A., Zaki, M. (2002). webSPADE : A Parallel Sequence Mining Algorithm to Analyze the Web Log Data. *Submitted to KDD'02.*

[Fireclick] httt://www.fireclick.com/

[Hay et al., 2001] Hay, B., Wets, G., Vanhoof, K. (2001) Clustering navigation patterns on a website using a Sequence Alignment Method. *In proceedings of IJCAI's Workshop on Intelligent Techniques for Web Personnalisation, Seattle, Washington 4-6 August 2001.*

[Media Metrix] Understanding Measurement of the Internet and Digital Media Landscape. Source : Media Metrix, http://us.mediametrix.com/products/us__methodology_long.pdf http://www.mediametrix.com/ ou http://fr.jupitermmxi.com/home.jsp

[NetValue] http://www.netvalue.fr/

[Spiliopoulou et al., 1998] Spiliopoulou, M., Faulstich, L.: WUM (1998). Web Utilization Miner. *In Workshop on the Web and Data Bases (WebDB98) pages 109-115, 1998.*

[WebTrends] http://www.webtrends.com/

[Zaki 2001] Zaki, M. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *In Machine Learning, pp 31-60, vol. 42 nos. ½, jan/Feb 2001.*

Web-R : pour la mémoire exhaustive de ma Toile

A. LIFCHITZ et J.D. KANT

CNRS - LIP6 (Laboratoire d'Informatique - Université Paris 6),
8, rue du Capitaine Scott,
75015 Paris, FRANCE
Mails : alain.lifchitz@lip6.fr , jean-daniel.kant@lip6.fr
Téls. : +33 1 44 27 43 32, +33 1 44 27 88 05, Fax : +33 1 44 27 70 00

Résumé

Dans cet article, nous présentons le système *Web-R*, un outil non-intrusif et rapide d'enregistrement qui réalise un stockage complet et systématique des pages web de l'utilisateur et de sa navigation. Il sauvegarde tous les composants qui seront nécessaires et suffisants à restituer hors-ligne la page de la même façon qu'elle l'a été vue en-ligne. Nous montrons que cet enregistrement systématique du Web personnel est non seulement techniquement réalisable, mais aussi réaliste car ne demandant qu'une faible fraction du volume de stockage des disques durs actuels. De surcroît, *Web-R* fournit un mécanisme de gestion de cet espace de stockage local incluant la comparaison du contenu de pages, évitant une inutile redondance. Enfin, puisque l'ensemble du contenu des pages visitées est stocké, *Web-R* permet à l'utilisateur d'avoir une vue globale de sa navigation personnelle par des statistiques, des outils de tri et de filtrage, pour sa réutilisation et plus tard sa structuration semi-automatique.

Abstract

In this paper, we present the *Web-R* system, a non-intrusive and fast recording tool that performs a systematic and complete storage of user's web pages and navigation. It saves all the components that are necessary and sufficient to visualize offline the page the same way it was displayed online. We show that a systematic storage of the personal web is non only technically feasible but also realistic since it requires only a small fraction of the storage space available in current hard disks. Moreover, *Web-R* also provides a way to manage this storage space and integrates a page comparison mechanism to avoid unnecessary redundancy. Finally, since the full content of visited pages is stored, *Web-R* allows the user to have a global view on his/her personal navigation by incorporating statistics, sorting and filtering tools and later on its semi-automated structuring.

1 Introduction

1.1 Problématique

L'expansion croissante du *World Wide Web* (WWW) a conduit récemment à une forte demande pour une interaction plus spécifique et plus personnalisée. Nous ne sommes plus

confrontés à la difficulté d'accéder à l'information en quantité, mais plutôt d'être capable d'atteindre l'information pertinente basée sur nos buts et besoins du moment. Pour faire face à cette demande, et pour éviter à l'utilisateur d'être submergé et perdu sur le WWW, un important effort a été consenti pour concevoir des applications comme les moteurs de recherche *(e.g.* [*Google*]), les services de catégorisation (*e.g.* [*Yahoo*]) ou plus généralement les outils d'aide à la navigation (*e.g. WebScout* [Milic-Frayling et al., 2002]). Pour pouvoir répondre aux rapides changements du contenu du WWW, ces applications habituellement construisent un cache local ou un entrepôt de pages web qui contribue à maintenir la cohérence des traitements suivants. Ainsi, dans chacune de ces situations, nous avons besoin d'outils spécifiques pour enregistrer les pages web qui alimenteront un traitement de plus haut niveau.

Ce travail est une partie d'un projetrécent plus large, le projet [*SpiderMem*], actuellement en développement. *SpiderMem* adopte délibérément une approche centrée sur l'utilisateur, qui l'aide à concevoir et structurer son propre *Web*. Ceci signifie que ce qui peut être organisé est restreint aux seuls sites déjà visités, de telle façon qu' une cartographie de l'ensemble, ou d'une partie seulement de la navigation passée, puisse être dessinée. De plus, une gestion interactive (assistée / automatisée) de l'information, des connaissances, dont les signets, est proposée. Au cœur de l' application *SpiderMem*, l'outil *Web-R* permet à l'utilisateur d'enregistrer et de rejouer toutes séquences de ses navigations passées.

1.2 Les outils d'enregistrement

Cette tâche d'enregistrement semble à première vue presque triviale… mais ne l'est pas en réalité. Il y a de nombreuses familles d'outils d'enregistrement et de nombreux outils appartiennent à plus d'une famille. Pour pouvoir mieux situer la problématique de cet article, nous allons effectuer un tour rapide des outils d'enregistrement du *Web*.

Le « *Suivi d'URL*» est le mécanisme qui sous-tend les fonctionnalités de signets (favoris) et d'historiques des navigateurs actuels. Celasignifie que dans chacun de ces cas, la seule information enregistrée est l'adresse de la page, complétée le plus souvent par l'horodatage du chargement. Par exemple, c'est le choix de conception fait dans l'outil d'aide à la navigation [*Internet Cartographer*], bien qu'il analyse au vol (au chargement) les mots-clés contenus dans les métadata de la page web. L'URL (*Uniform Resource Locator*) est une information pertinente et compacte, aisée à stocker, mais une page web est parfois fugace, et quand nous marquons son URL grâce au signet, nous ne pouvons assurer, spécialement en raison de la nature dynamique croissante du *Web* vivant, que cette page sera disponible dans le futur et/ou que son contenu sera identique à ce qu'il est en cet instant. Pour prendre en compte pleinement cette situation notre outil doit non seulement stocker l'URL mais aussi l'intégralité du contenu d'une page ou plus exactement de ses composants.

Chaque navigateur web possède une fonctionnalité « *Enregistrer sous…* » qui devrait réaliser cette tâche. Cependant la sauvegarde des pages doit être automatisée, car bien sûr il est trop fastidieux pour l'utilisateur cliquer sur « *Enregistrer sous…* » chaque fois qu'il parcourt une page particulière... De plus, la nécessaire redirection vers les fichiers locaux pour la commutation du contexte en-ligne / hors-ligne est rarement complètement et correctement réalisée. Par exemple, quand nous essayons de sauvegarder une page qui inclut une animation *Flash Macromedia®*, avec la version la plus récente (6) de *Microsoft Internet Explorer®* - après avoir pris la précaution de vider les caches mémoire et disque – nous obtenons un rectangle blanc frustrant à la place de l'animation attendue lorsque nous affichons la page enregistrée sans connexion Internet. Cela est dû au fait que des hyperliens pointent encore sans objet vers le *WWW* au lieu d'être des chemins valides vers des fichiers locaux existant effectivement.

Un *butineur* web (*crawler* appelé aussi *webbot* ou *spider*) est un programme configuré pour parcourir et traiter automatiquement, à un moment prédéterminé, les pages, en traversant

systématiquement une topologie web statique. Clairement, ni un *butineur* web personnel (*e.g.* [*WebSPHINX*], [Miller and Bharat, 1998]) ni son pendant global (*e.g.* [*Googlebot*], [*Slurp*]) sont les outils que nous recherchons puisque leur but est de recueillir l' intégralité des fichiers de tout ou partie du *Web* visible. Au contraire, nous souhaitons seulement recueillir les pages visitées par l'utilisateur actuel.

L'*aspirateur* de site (*site ripper*) est utilisé pour dupliquer (aspirer) un site entier sur le disque local. Il s'active à la demande spécifique d'un utilisateur. Certains *aspirateurs* (*e.g.* [*BlackWidow*], [*Wget*]) intègrent des fonctionnalités de visualisation / parcours du web et de gestion. Par rapport aux visées de *Web-R*, il manque à ces outils de l'automatisation et de ne pas être « *centré utilisateur* » (pas d'adaptation, ne fonctionnent qu'à la demande explicite).

Les « *navigateurs hors-ligne* » (*offline browsers*) sont apparus dans les années 1990 pour réduire / rentabiliser la durée de connexion et économiser la bande passante, denrée rare à l' époque. Les versions les plus récentes de ces « *navigateurs hors-ligne* » intègrent d' intéressants outils de parcours et de gestion (*e.g.* [*SurfSaver*]) mais le plus souvent la redirection *WWW* / locale est imparfaite et rarement automatisée.

La dernière famille d' outils d'enregistrement est celle que nous avons dénommée la « famille *Web-R* ». Nous avons découvert deux très récents systèmes, qui ont des caractéristiques proches de celles que nous souhaitons pour la conception de *Web-R* : [*Keepoint*] et *WebScout* [Milic-Frayling et al., 2002]. Comme *Web-R*, *Keepoint* offre le stockage automatique et complet des pages web parcourues au travers d'*Internet Explorer* (*IE*). Il intègre aussi des outils de gestion qui vont au-delà de ce que nous proposons dans la version courante de *Web-R*, puisqu' un utilisateur de *Keepoint* peut notamment annoter les pages stockées et les classer automatiquement par domaines, organisations ou mots-clés. En revanche, son moteur d'enregistrement souffre – du moins dans la version actuelle (1.0) – de deux inconvénients majeurs qui épargnent *Web-R*. D'une part, le mode d' enregistrement automatique réclame encore une intervention systématique de l' utilisateur : celui-ci doit spécifier au système si la page visitée courante, qui partage le même URL qu' une page déjà stockée, doit-être ajoutée à la base de données ou bien substituée à celle-ci. Si l'utilisateur choisi l'automatisation complète, cette décision (duplication ou substitution) sera alors systématiquement appliquée quelle que soit la situation rencontrée pour les pages suivantes avec le risque de redondance ou à l'opposé la perte de pages importantes. La raison profonde de cette difficulté est que *Keepoint*, au contraire de *Web-R*, n'intègre pas un mécanisme de comparaison de pages. D'autre part, le second inconvénient – moins fondamental parce qu'il peut être aisément contourné (cf. 2.3.1) – est que les pages en mode sécurisé (`https://…`) soient incomplètement stockées, ce qui n'est pas le cas de notre système.

Le projet *Microsoft WebScout* semble être très ambitieux et est en fait plus proche de *SpiderMem* que de *Web-R*, puisqu' il vise à intégrer l'annotation de page, le traitement du langage naturel, l'analyse d'image, l'indexation et la recherche d'information textuelle et visuelle, ainsi que la visualisation de données. Jusque-là, trois composants ont été conçus, parmi lesquels le *History Explorer* qui partage les mêmes buts que *Web-R*. Cependant, nous n'avons aucun détail sur cette implantation, et il est donc difficile de la comparer à la solution que nous proposons.

1.3 Objectifs

Pour résumer, nous souhaitons réaliser une application compagne du navigateur, intégrant navigation, enregistrement de celle-ci et des contenus, visualisation, et qui puisse fonctionner aussi bien hors-ligne qu'en-ligne. Nous proposons donc de sauvegarder automatiquement l'intégralité des pages que l'utilisateur parcourt pour lui permettre d'organiser à tout moment son *Web* personnel. Pour être concrètement utilisable cette application devra être non-intrusive, rapide et capable de fournir un service de gestion de l'espace pour éviter une éventuelle saturation du disque local. Elle devra aussi fournir un stockage, précis et complet,

des pages web vues par l' utilisateur,en sauvegardant seulement les composants nécessaires et suffisants à la restitution hors-ligne des pages conforme à ce qu' elles étaient en-ligne.

Cet article est organisé comme suit. Nous décrivons les bases principales de conception du système *Web-R*, puis détaillerons son implémentation en section 2. Les travaux à poursuivre dans un futur proche et moyen, en section 3, concluront ce document.

2 Conception et implémentation

2.1 Communication / contrôle concernant le navigateur

Afin de pouvoir stocker les pages web, systématiquement et automatiquement, pendant que l' utilisateur navigue, nous avons à choisir le mode d'interaction le plus approprié avec le navigateur envisagé. Cette interaction devra, non seulement ne pas consommer trop de temps et autres ressources, mais aussi être suffisamment efficace pour supporter des pages complexes, comme celles comportant du *JavaScript* ou des *frames*. Il y a au moins, de base, quatre façons d' interagir avec un navigateur : Le*proxy*, l'*applet*, le *plugin* et l' API.

On peut, par exemple, prévoir un serveur *proxy* entre le *WWW* et le navigateur de l' utilisateur, fonctionnant comme un cache pour les pages entrantes. Nous rejetons cette architecture « serveur », parce qu'elle est très lourde et peut conduire notamment à des failles concernant la confidentialité des données personnelles.

Les trois solutions restantes sont des architectures « client », bien plus adaptées à l'outil personnel que nous souhaitons concevoir. Ces architectures mettent à profit les fonctionnalités du [*Document Object Model*] (DOM). La première solution « client » est d'utiliser une *applet Java* conjointement avec le navigateur. Dans le système *WebVCR* [Anupam et al., 2000], cette solution s'est montrée efficace pour enregistrer les URLs visités et les informations associées. Une deuxième solution est de réaliser un *plugin* qui prend le contrôle interne du navigateur utilisé. Enfin, une troisième solution « client » est de prendre le contrôle externe du navigateur : sous les environnements *Microsoft Windows*®, nous pouvons utiliser *OLE Automation*® pour contrôler et interagir avec *IE*[1]. C' est la solution que nous avons adoptée pour *Web-R*, puisque nous souhaitions doter cet outil de fenêtres et d' interfaces spécifiques alors qu' un plugin ne peut qu'ouvrir des fenêtres du navigateur sous contrôle.

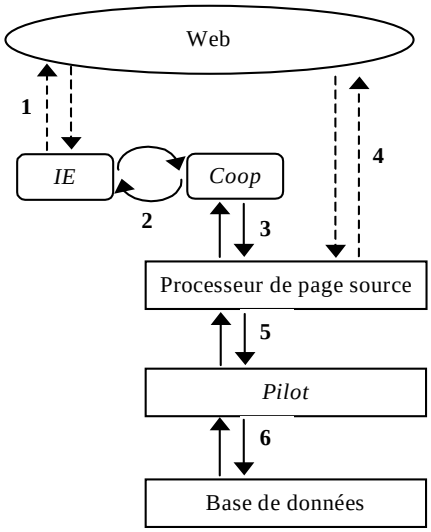
2.2 Architecture

Web-R est une application qui prend le contrôle de son instance d'*IE* exécutée enfant. Seules des instances qui ont été créées au travers de cette interface racine de *Web-R*, seront sous contrôle. Cette interface utilisateur, de niveau le plus élevé, est représentée Figure 1, ci-dessous.

Figure 1. L' interface racine de *Web-R* (échelle réelle).

L'enregistrement des pages est réalisé comme représenté Figure 2 ci-dessous. L' instance d'*IE* obtient la page à partir du *Web* (1), et interprète son code pour affichage. Le chargement de la page par *IE*, est sous le contrôle du module *Coop* qui attend qu'*IE* ait fini son traitement (2). *Coop* récupère le résultat de ce traitement (*i.e.* la page traitée) et l'envoie au module *Processeur de page source* (PPS) (3) qui peut profiter du pré-traitement d'*IE* au travers de DOM.

[1] Nous avons dû choisir pour *Web-R* un développement spécifique à ce navigateur pour deux raisons : 1) *OLE Automation* était disponible et bien documenté ; 2) Il est de loin le navigateur le plus utilisé.



Le but de ce module est de réaliser tout le traitement nécessaire pour s'assurer que le stockage local de la page courante soit « parfait ». Par stockage parfait, nous entendons sauvegarde de toutes les données nécessaires et suffisantes (composants et chemins) pour ré-afficher la page hors-ligne conformément à ce qu'elle était en-ligne. Ainsi, le PPS devra collecter depuis le *Web* (4) les fichiers additionnels nécessaires (*e.g.* images, sons, animations *Macromedia Flash*®, code *Javascript*, etc.). Une fois le traitement de la page réalisé exhaustivement, celle-ci et ses composants sont transmis au module *Pilot* (5) pour le stockage de tous ces fichiers et de leur description dans la *Base de données* (6).

Figure 2. Organigramme de l'enregistrement des pages.

Un des points essentiels de la conception réside dans le module *Coop* qui doit gérer les événements *IE*. *IE* active des évènements de la même façon que tout objet *Component Object Model* (COM). Chaque fois que *IE* souhaite fournir une information à ses clients sur son activité courante, il active un évènement par son point de connexion. Pour l'essentiel, le module *Coop* est un *handler* d'événement de *IE*. Il fournit l'information sur la progression du traitement de la page réalisé par *IE*.

2.3 Processeur de page source

Du fait du changement de contexte, du *Web* en-ligne au stockage local, les sources de pages doivent être éditées : c'est le rôle attribué au *Processeur de page source*.

2.3.1 Chemins de fichiers dans les tags HTML

La fonction essentielle du PPS est de prendre en charge correctement la redirection des noms de fichiers des composants de page, du *Web* en-ligne au disque dur local. Pour pouvoir déterminer où ces liens doivent être modifiés dans le code HTML de la page, il faut intervenir sur les libellés des attributs pertinents en localisant les tags potentiellement concernés, c'est-à-dire ceux qui peuvent comporter un lien vers un fichier externe. Ces tags sont localisés dans le code source en utilisant DOM. Dans sa version courante, *Web-R* prend en charge la redirection locale automatique des couples `<tags> attributs` (HTML 4.0 & 4.01) suivants : `<BODY> BACKGROUND` (URL d'image), `<FRAME>` et `<IFRAME> SRC` (URL de *frame*), `<IMG> SRC` (URL d'image), `<INPUT> SRC` (URL d'image), `<LINK> HREF` (URL de fichier ressource), `<EMBED> SRC` (URL de données de *plug-in*), `<SCRIPT> SRC` (URL de fichier script)[2]. Il faut remarquer que les pages sécurisées sont traitées de la même façon en redirigeant simplement toutes les entrées `"https://…"` du code HTML vers les fichiers locaux appropriés.

2.3.2 Frames

Un *frame* inclus dans une page déclenche un traitement spécifique, puisque la page maîtresse est en fait une page composite de pages filles. Cette imbrication de pages filles au sein d'une page maîtresse devra être suivie récursivement sur plusieurs niveaux, puisqu'une page fille d'un *frame* peut elle-même comporter un *frame*, et ainsi de suite. Par conséquent, quand la présence d'un *frame* a été détectée, le PPS démarre le traitement des pages filles de chaque *frame*, et le poursuit pour tous les *frames* englobés, récursivement, jusqu'à ce

[2] Il n'y a aucun obstacle conceptuel à prendre en charge de la même façon le code XML (XHTML, SVG) des pages du *Web* les plus récentes. Cela n'est pas encore réalisé faute de moyens humains/matériels plus importants…

qu' aucun *frame* ne soit plus détecté. Parallèlement, les différents composants sont stockés localement jusqu' à ce que le niveau le plus profond soit atteint. Une des difficultés rencontrées dans le traitement des *frames* provient du mécanisme de sécurité implanté dans *IE*, où l' accès à un*frame* fils, dont le domaine diffère du *frame* père, n' est pas évident.

2.3.3 Pages dynamiques

Nous avons conçu le PPS pour qu' il prenne encharge, aussi correctement que possible, les pages dynamiques. Le premier cas de pages locales dynamiques sont les *scripts* notamment *JavaScript*. Quand ces scripts sont des fichiers externes, nous les traitons par une redirection locale décrite en 2.3.1 ci-dessus.

Avec le code embarqué (*embedded*) / externe, deux situations fréquentes sont actuellement prises en charge :

- Le script modifie la page à l'affichage initial. Dans ce cas, *IE* exécutera le script, modifiera la page en conséquence, et alors transformera en commentaire le script pour éviter sa re-exécution intempestive. Par conséquent, aucun traitement additionnel n'est nécessaire puisque nous récupérons la page déjà pré-traitée par *IE*.
- Le script est interactif et s' active lorsque certains évènements se produisent (*e.g. OnMouseOver*, *OnMouseOut*). Il s'exécutera correctement hors-ligne, pour autant que tous les hyperliens soient relatifs à la page maîtresse, et là encore aucun traitement additionnel n' est requis.

Un autre cas de page locale dynamique est l'objet animation (*e.g. Flash Macromedia*®). Il sera traité par une simple redirection de lien vers un fichier local dans le tag `EMBED`, comme décrit dans 2.3.1 ci-dessus.

Le cas des pages dynamiques coté serveur (*e.g.* ASP, PHP, CGI, etc.) ne requièrent aucun traitement additionnel, puisque instanciées par *IE* comme des pages HTML statiques relevant du traitement générique de ces pages.

2.4 Stockage

2.4.1 Contenu de page

Le contenu de page est stocké dans un sous-répertoire d' un répertoire unique dédié à l' utilisateur courant. Ce sous-répertoire contient toutes les pages qui ont été visitées, par l' utilisateur, ce jour courant. Ce contenu est constitué du fichier source HTML d' origine (celui reçu du *WWW* avant tout traitement, pour mémoire), du fichier source HTML transformé (sortie du PPS), et de tous les composants de page nécessaires et suffisants pour pouvoir rejouer la page hors-ligne de façon fiable (fichiers images, fichiers sons, frames, fichiers embarqués, fichiers scripts, fichiers de style (CSS), etc.).

2.4.2 Visite de page

Pour refléter la visite d'une page, des informations additionnelles sont stockées dans une base de données. La version courante de *Web-R* utilise le moteur *Microsoft Jet*® pour cela. Nous stockons pour chaque page, un identifiant numérique unique (*id*), l'emplacement des composants sur le disque local, le nombre d'accès, la durée cumulée de visite, la taille logique totale et la taille physique totale (sur le disque). Pour la dernière visite de chaque page, nous stockons la date et l'heure de début de visite, la durée de celle-ci, la taille et la position de la fenêtre *IE*. Les composants ont aussi chacun un descripteur stocké dans la base, qui comporte l'identifiant de la page à laquelle ils appartiennent, un identifiant de composant, la localisation sur le disque et les tailles logiques / physiques. Pour finir, nous ajoutons un champ de validité (si une erreur s'est produite pendant le chargement de page et dans ce cas le code d' erreur) pour chaque entrée (aussi bien pour la page que pour le composant).

2.5 Au-delà de l'enregistrement

2.5.1 Redondance (contenus identiques)

Pour éviter l' essentiel de la redondance, nous avons implanté un algorithme simple de comparaison de surface pour déterminer si la page courante est déjà stockée. Cet algorithme se déroule en 3 étapes pour déterminer si la page *P* a besoin d'être stockée :

1. Soit {*P*} un ensemble de pages stockées avec le même URL que *P*. Si {*P*} est vide, stocker *P*;
2. Sinon, rechercher une page $P' \in \{P\}$ avec le même code HTML que *P*. Si une telle page n' existe pas, stocker*P*;
3. Sinon, comparer les tailles individuelles respectives des composants homologues dans *P'* et *P*. S' ils diffèrent, stocker*P*; sinon ne pas stocker *P*.

Cet algorithme est rapide et fonctionne dans la très grande majorité des situations concrètes rencontrées dans une navigation web. Une variante maîtrise la prolifération de pages non-identiques, par le contenu, avec le même URL (ce qui peut se produire aisément dans le cas d'accès répétitifs au même site, *e.g.* dans le suivi en temps réel des cours de bourse).

2.5.2 Gestion de l' espace de stockage

Puisque toutes les pages visitées seront stockées sur le disque local durant une session *Web-R*, on peut se préoccuper de l' évolution de l'espace occupé. Nous pouvons prévoir que la disponibilité et la banalisation rapide de très grand volume de stockage ainsi que de la puissance de traitement ne pourront qu'aider...mais bien sur ce n'est pas suffisant.

En fait, nous estimons que la taille totale du contenu web personnel ne dépassera pas 5 Go / an pour un utilisateur typique. Des études récentes confortent cette estimation.

Une première expérience conduite par [Huberman et al, 1998] sur un échantillon de 23962 utilisateurs AOL donne un total de 3247054 pages visitées qui conduit à 147 pages visitées / utilisateur / jour. Cela donne un total d'approximativement 50000 pages visitées / utilisateur / an. De plus, l' index [Nielsen//NetRatings, 2002] qui mesure chaque semaine et mensuellement l' audence *Web* sur plusieurs pays montre un maximum de l' ordre de 150 sites visités par mois (*e.g.* pour octobre 2002 : 147 aux USA, 52 en Angleterre, 59 en France, 82 en Allemagne). Le nombre observé de pages visitées par site étant approximativement de 20 (il était de 16 pour l'utilisation privée aux USA en juin 2002, où 735 pages ont été visitées pour un total de 47 sites). Ainsi, les statistiques de l'index Nielsen//NetRatings donnent un maximum de 150x20=3000 pages visitées mensuellement, soit 36000 pages visitées annuellement par utilisateur. Étant donné que la taille moyenne de la page ne doit pas excéder 100 Ko (la troisième *State of Web Survey* [SOW Survey, 1999] rapporte une taille moyenne de la page web de 60 Ko en 1999, images et tous autres composants compris), nous arrivons à la conclusion que la taille maximum web personnel annuel est compris entre 3,6 Go (d' après Nielsen//NetRatings) et 5 Go (d' après l'étude de Huberman *et al.*). Dans tous les cas, cette taille est largement compatible avec les capacités courantes des disques durs actuels et leurs coûts relativement faibles.

Quoi qu' il en soit, nous devons fournir un outil degestion de l' espace de stockage comme dans tout système de cache. Bien sûr, l'algorithme de maîtrise de la redondance que nous avons présenté en section 2.5.1 ci-dessus ne peut qu'aider. Cependant, l' utilisateur de*Web-R* accède à un mécanisme de limitation de la taille réservée au stockage de la navigation passée, au moyen de l'interface de configuration : Cette limite peut être inexistante, fixe ou un pourcentage de la taille totale du disque local. Quand cette limite est atteinte, un mécanisme de purge est déclenché, qui peut être manuel ou bien automatique.

Dans le cas de la purge automatique, les pages destinées à disparaître sont automatiquement sélectionnées, sur la base d'une combinaison de trois critères : taille physique, date de dernière visite et nombre de visites. Ce calcul de score vise à effacer en

priorité les pages qui auront la plus grande taille, la date de dernière visite la plus ancienne et le moins grand nombre de visites. L'importance accordée à chaque critère est fixée par l' utilisateur pour obtenir le score global. Pour la purge manuelle, l'utilisateur sélectionne, depuis la liste précédente, les pages qu'il choisit d'effacer du disque local.

2.5.3 Gestion d'utilisateurs multiples

Bien que *Web-R* soit un outil délibérément personnel, il peut être utile qu' il soit mis à la disposition, successivement, de plusieurs utilisateurs sur la même machine. Pour ce faire, *Web-R* est capable de gérer plusieurs utilisateurs, en limitant le risque d'interférence.

Cependant, on peut souhaiter fusionner les navigations partielles provenant d' utilisateurs différents en une unique base de données. C'est le cas, par exemple, lorsqu' on veut incrémenter une navigation personnelle par celle d' un autre utilisateur, ou reporter sur l' ordinateur du domicile les pages issues de l' ordinateur du bureau. Un autre exemple est celui où l' on souhaite conduire une étude sur l' usage du *Web* : on a un *panel* d' utilisateurs et l'on souhaite collationner toutes les pages de leur navigation.

Pour répondre à ces besoins, nous avons prévu une fonctionnalité d'importation/exportation de navigation. Un utilisateur peut exporter toute sélection de navigation en un unique fichier d'échange. Ce fichier d' échange est une archive compressée (*zip*) qui contient la base de données, tous les répertoires et leurs fichiers de pages visitées. Quand l'utilisateur choisit d'importer un fichier d'échange, sa base de données personnelle est fusionnée avec celle contenue dans le fichier d'échange, les répertoires et les fichiers de pages sont ajoutés à ceux du répertoire de données *Web-R*, et les statistiques mises à jour.

2.6 Gestion du *Web* personnel

Dans sa version actuelle, *Web-R* offre aussi un outil de base pour le visionnage et la gestion des pages stockées.

Quand le Gestionnaire / Visionneuse de Pages est activé, il affiche une fenêtre comportant une liste multi-colonnes qui contient les descripteurs de toutes les pages stockées, comme on le voit sur la Figure 3 ci-dessous :

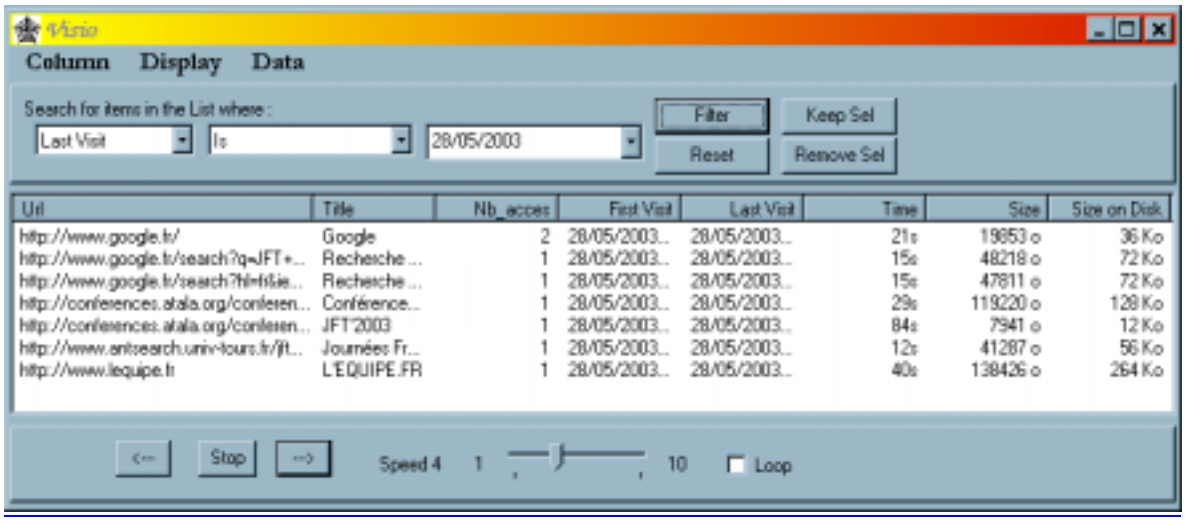


Figure 3. Visionneuse de pages avec filtre.

Ces entrées peuvent être classées selon les attributs suivants : URL, titre, nombre d'accès, date et heure de la première visite, date et heure de la dernière visite, taille logique / physique. De plus, en haut de la fenêtre du Gestionnaire / Visionneuse de Pages, un filtre est prévu, pour ne lister que des ensembles de pages satisfaisant certaines combinaisons de valeur des attributs ci-dessus. Par exemple, on peut sélectionner les pages dont l' URL contient la chaîne « 2003 » postérieures à juin 2002. Remarquons que chaque attribut sélectionne

automatiquement son type de filtre (e.g. pour une valeur numérique, le filtre propose : égal / moins que / plus grand que / différent). Cet outil de gestion permet aussi de re-visionner chaque page stockée. Son contenu (hors-ligne) sera affiché quand l'utilisateur double-clique sur un item de la liste, comme on le voit sur la Figure 4 ci-dessous :

Figure 4. Visualisation hors-ligne d' une page enre gistrée.

3 Conclusion et projets futurs

Dans cet article, nous avons montré que le stockage systématique du *Web* personnel est non seulement techniquement réalisable, mais réaliste avec un espace de stockage maîtrisé. Nous avons présenté l'outil *Web-R*[3], un moteur d' enregistrement non-intrusif, rapide et fiable qui peut sauvegarder systématiquement, au cours du temps, le contenu de l' ensemble des pages visitées par un utilisateur. De plus, *Web-R* fournit aussi des informations sur la navigation de l' utilisateur, etintègre des outils pour gérer cette navigation web stockée. Ces pages stockées peuvent être visionnées hors-ligne, en étant reproduites conformément à ce qu' elles étaient, en-ligne,lors de leur visite.

Web-R est en cours de développement, ce qui explique certaines de ses limitations[2]. Plusieurs questions techniques doivent cependant être approfondies. A court terme, nous devons consolider le *Processeur de page source* pour le rendre plus robuste. De plus, nous souhaitons faciliter la navigation hors-ligne, en distinguant les liens vers les pages déjà stockées (disque local) et liens vers des pages non-stockées (*WWW*). Dans ce dernier cas, notre système devra demander à l' utilisateur s'il souhaite stocker cette page externe. Nous projetons aussi de faire un meilleur usage des ressources du navigateur, comme l'utilisation de son cache pour éviter de multiples - et donc inutiles - accès aux mêmes composants du

[3] Une version de démonstration de *Web-R* est disponible à l' adresse :http://webia.lip6.fr/~lifchitz/SpiderMem/download
Une version en-ligne de ce document est disponible à l' adresse :http://webia.lip6.fr/~lifchitz/SpiderMem/papers

Web en-ligne. Nous devrions aussi utiliser un algorithme de compression pour les fichiers de contenu stockés pour optimiser l' espace de stockage, et encrypter toutes les données pour garantir une totale confidentialité. Enfin, pour améliorer l'ergonomie de notre logiciel, nous envisageons une hybridation des techniques de contrôle d'*IE*, où un plugin et une application externe se partageront le contrôle du navigateur.

Enfin l' aspect validation/expérimentation auprès des utilisateurs ne pourra être considéré que sur une version suffisamment stable et ergonomique de cet outil.

4 Remerciements

Les auteurs aimeraient remercier le LIP6 pour son soutien financier à ce projet. Nous remercions aussi Nicolas Limare, Arnaud Joly et Violaine Ruffié pour leur participation au développement de la version préliminaire du logiciel *Web-R*.

Références

[Anupam et al., 2000] Anupam, V., Freire, J., Kumar, B., and Lieuwen, D. (15-19 May 2000). Automating Web Navigation with the WebVCR. In *Proceedings of WWW9*, Amsterdam (The Netherlands). http://www9.org/w9cdrom/208/208.html

[*BlackWidow*] *BlackWidow*. http://sbl.net/BlackWidow

[*Document Object Model*] *Document Object Model*. http://www.w3.org/DOM/

[*Google*] *Google*. http://google.com/

[*Googlebot*] *Googlebot*. http://www.googlebot.com/bot.html

[Huberman et al, 1998] Huberman, B., Pirolli, P., Pitkow, J., and Lukose, R. (3 April 1998). Strong regularities in World Wide Web Surfing. *Science*, 280(5360): 95–97. http://external.nj.nec.com/~giles/huberman/98.huberman.pdf

[*Internet Cartographer*] *Internet Cartographer*. http://www.inventix.com/

[*Keepoint*] *Keepoint*. http://www.keepoint.com/

[Milic-Frayling et al., 2002] Milic-Frayling, N., Sommerer, R., and Tucker, R. (7-11 May 2002). MS WebScout: Web Navigation Aid and Personal Web History Explorer. In *Proceedings of WWW2002*, Honolulu, (Hawaii, USA). http://www2002.org/CDROM/poster/170/index.html

[Miller and Bharat, 1998] Miller, R., and Bharat, K. (14-18 April 1998). SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers. In *Proceedings of WWW7*, Brisbane (Australia). In *Computer Network and ISDN Systems* 30: 119-130. http://www7.scu.edu.au/programme/fullpapers/1875/com1875.htm

[Nielsen//NetRatings, 2002] http://www.nielsen-netratings.com/hot_off_the_net.jsp

[*Slurp*] *Slurp*. http://www.inktomi.com/slurp.html

[SOW Survey, 1999] SOW Survey. http://www.pantos.org/atw/35654.html

[*SpiderMem*] *SpiderMem*. http://www-connex.lip6.fr/~lifchitz/SpiderMem

[*SurfSaver*] *SurfSaver*. http://www.surfsaver.com/

[*WebSPHINX*] *WebSPHINX*. http://www-2.cs.cmu.edu/~rcm/websphinx

[*Wget*] *Wget*. http://www.gnu.org/directory/wget.html

[*Yahoo*] *Yahoo*. http://yahoo.com/

Modèles pour l'intégration de l'accès progressif dans les systèmes d'information sur le Web

M. Villanova-Oliver, J. Gensel, H. Martin

Laboratoire LSR-IMAG
BP 72, 38402 St Martin d' Hères
Grenoble, FRANCE
Mail : {villanova, gensel, martin}@imag.fr
Tél : (+33) 4 76 82 72 80 Fax : (+33) 4 76 82 72 07

Résumé

L'importante quantité d'information gérée par les Systèmes d'Information sur le Web (SIW) peut être à l'origine de syndromes de désorientation et de surcharge cognitive chez les utilisateurs. Afin d'atténuer ces effets, nous avons introduit le concept d'accès progressif dont le but est d'offrir aux utilisateurs de SIW un accès graduel, personnalisé et flexible à l'information. L'accès progressif repose sur une stratification de l'information décrite par un modèle central appelé, Modèle d'Accès Progressif (MAP). Nous présentons ici le MAP et ses connections avec quatre autres modèles (modèle du domaine, modèle des fonctionnalités, modèle de l'hypermédia, modèle des utilisateurs). Nous proposons également une méthodologie de conception exploitant ces modèles. Une fois instanciés, ces modèles permettent de générer des SIW proposant un accès progressif à l'information.

Abstract

Because of the large amount of information managed by Web-based Information Systems (WIS), their users often experience some disorientation and cognitive overload syndromes. In order to attenuate this negative effect, we introduce the concept of Progressive Access which aims at giving WIS users a flexible and personalized access to data. Progressive access requires to stratify the information space. These stratifications are described through a central model called the Progressive Access Model (PAM). We present here the PAM and its connections to four other models (domain model, functionality model, hypermedia model, user model). We propose then a WIS design methodology which exploits these models. Instantiating these models leads to the generation of WIS which integrate the progressive access approach.

1 Introduction

Les Systèmes d'Information sur le Web (SIW), comme tout autre Système d'Information traditionnel, permettent d'acquérir, de structurer, de stocker, de gérer, et de diffuser de l'information, mais en s'appuyant sur une infrastructure Web. La principale différence entre les SIW et les autres applications Web (notamment les sites Web) réside dans la complexité des services offerts [Mecca *et al.*, 1999][Conallen, 2000]. Alors que les services sont presque inexistants sur les sites Web, les SIW proposent des fonctionnalités complexes activables dans une interface multimédia à travers un navigateur Web. Nous considérons ici que les SIW sont des applications réunissant les caractéristiques des Systèmes d'Information et des Hypermédias basés sur le Web.

Basé sur le Web et sur une structure hypermédia, un SIW permet donc aux utilisateurs de naviguer dans un large espace d'information. Cette caractéristique est souvent à l'origine des syndromes de désorientation et de surcharge cognitive [Conklin, 1987]. Le syndrome de désorientation est ressenti par les utilisateurs qui naviguent dans un espace d'information dont la structure est complexe, et se sentent rapidement distraits ou perdus. Egalement, les utilisateurs peuvent subir une surcharge cognitive lorsqu'ils sont confrontés à une masse d'information trop importante, difficile à comprendre ou à filtrer, voire inutile. Afin de limiter ces deux effets préjudiciables à la pérennité même du SIW, nous pensons que l'information délivrée doit être organisée, contrôlée, accessible progressivement, et présentée de façon personnalisée. De plus, nous proposons que ces deux risques soient prévenus dès la phase de conception du SIW.

Depuis quinze ans, plusieurs méthodes ont été proposées pour la conception d'applications s'étendant des hypermédias (basés ou non sur le Web) aux sites Web, jusqu'aux SIW. Les méthodes les plus récentes (WSDM [De Troyer *et al.*, 1998], UWE [Koch, 2000], WebML [Ceri *et al.*, 2000], AWIS-M [Gnaho, 2001], OOHDM [Rossi *et al.*, 2001]) privilégient davantage les aspects fonctionnels que ne le faisaient les plus anciennes (HDM [Garzotto *et al.*, 1993], RMM [Isakowitz *et al.*, 1995]). La littérature du domaine (voir, par exemple, [Isakowitz *et al.*, 1998][Fraternali, 1999][Baresi *et al.*, 2000]) montre que la spécification et la conception d'une application Web est une activité multi-facettes qui nécessite la description du domaine d'application, la composition et la définition de l'apparence des pages Web générées, la spécification de mécanismes de navigation, jusqu'à une interaction entre le contenu et les fonctionnalités [Ceri *et al.*,2000], voire des capacités d'adaptation aux utilisateurs [Brusilovsky, 1998][Stephanidis *et al.*, 1998]. En fait, le consensus se fait autour de la nécessité de disposer d'au moins trois modèles *i)* un *modèle de données* qui permet de décrire les concepts du domaine d'application, *ii)* un *modèle* chargé de la *structuration* et de la *navigation* dans l'hypermédia, et *iii)* un *modèle de présentation* chargé de l'apparence des pages Web. De plus, pour l'adaptation aux utilisateurs, deux autres modèles sont proposés : un *modèle des utilisateurs* qui décrit les caractéristiques et/ou les préférences des utilisateurs [Gnaho, 2001][Frasincar *et al.*, 2001], et un *modèle d'adaptation* qui permet d'exprimer des expressions algébriques [Gnaho, 2001] ou des règles [Koch, 2000][Frasincar *et al.*, 2001] définissant l'adaptation.

A notre connaissance, il n'existe pas de méthode de conception de SIW dont le but soit de réduire les risques de désorientation et de surcharge cognitive évoqués plus haut. C'est pourquoi, nous avons proposé la notion d'*accès progressif* [Villanova *et al.*, 2001] qui repose sur une organisation de l'information telle que l'espace d'information d'un SIW soit structuré en différents niveaux de détail. Une telle organisation est appelée *stratification.* L'idée est qu'un utilisateur accède, dans cette stratification, en premier lieu à une information jugée essentielle au regard de ses besoins, puis, progressivement, à une information jugée complémentaire mais moins importante. Par cette gradation, les risques de désorientation et de surcharge cognitive sont limités. Afin de stratifier un espace d'information, nous avons proposé [Villanova, 2002] un modèle appelé *Modèle d'Accès Progressif* (MAP).

Dans cet article, après l'avoir formalisé, nous décrivons comment le concept d'accès progressif est intégré dans une approche de conception de SIW reposant sur cinq modèles. Le MAP constitue le modèle de référence dans notre approche. Nous décrivons quatre autres modèles fortement liés au MAP, chargés respectivement du contenu informationnel, des aspects fonctionnels, des caractéristiques hypermédia, et des profils des utilisateurs. Nous proposons alors une méthodologie

basée sur sept étapes qui exploite ces modèles et permet de concevoir des SIW intégrant l'accès progressif.

Le plan de l'article est le suivant. Dans la section 2, nous formalisons le concept d'accès progressif. Dans la section 3, nous présentons les cinq modèles pour la conception de SIW intégrant l'accès progressif. Enfin, dans la section 4, une méthodologie de conception est proposée.

2 Définitions pour l'accès progressif

L'idée centrale derrière la notion d'*accès progressif* est que l'utilisateur d'un SIW n'a pas besoin d'accéder *tout* le temps à *toute* l'information disponible. Notre objectif est de construire des SIW qui ont la capacité de délivrer progressivement à leurs utilisateurs une information personnalisée. L'information considérée comme essentielle doit être immédiatement accessible, puis, au besoin, des informations supplémentaires peuvent être atteintes par une navigation guidée. Ce double objectif nécessite à la fois une organisation de l'espace d'information permettant l'accès progressif, et la gestion de profils afin que le contenu et la présentation de l'information soit adaptés aux besoins et préférences des utilisateurs. Nous nous intéressons au premier point en formalisant ici la notion d'accès progressif.

L'accès progressif est lié à la notion d'*Entité Masquable*. Une *Entité Masquable* (EM) est un ensemble d'au moins deux éléments (i.e. $|EM| \geq 2$) sur lequel un accès progressif peut être mis en œuvre. L'accès progressif à une EM repose sur la définition de *Représentations d'Entité Masquable* (REM) associées à cette EM. Les REM d'une EM sont des sous-ensembles d'éléments de l'EM, ordonnés par la relation d'inclusion ensembliste. Deux types de REM, extensionnelles ou intensionnelles, sont distingués. Les *REM extensionnelles* sont construites à partir de l'*extension* (i.e. l'ensemble des éléments) d'une EM. Lorsque cette EM est un ensemble de données structurées de même type, des REM *intensionnelles* peuvent être construites à partir de l'*intension* de l'EM. L'*intension* d'une EM structurée est définie ici comme l'ensemble des descriptions de variables (attributs ou champs) qui compose la structure de l'EM. Chaque REM (extensionnelle ou intensionnelle) d'une EM possède un *niveau de détail* associé. Nous appelons REM_i, la REM d'une EM correspondant au niveau de détail *i*, avec $1 \leq i \leq max$, où *max* est le plus grand niveau de détail disponible pour cette EM. Nous donnons ci-dessous les règles de définition d'une REM valide.

– **Règles pour une REM extensionnelle** : soit $\mathbf{M}$ une EM dont l'extension est définie par $E(\mathbf{M})=\{e1,e2,\ldots,en\}$, avec $e_k \in \{e_1,\ldots,e_n\}$ élément de $\mathbf{M}$.

- toute REM extensionnelle REM_{EXTi} de $\mathbf{M}$ est non vide (i.e. $REM_{EXTi} \neq \emptyset$), $\forall i \in [1,max]$
- $REM_{EXT1}=\{e_m,\ldots,e_p\}$ avec $\{e_m,\ldots,e_p\} \subset E(\mathbf{M})$ (i.e. $REM_{EXT1} \neq E(\mathbf{M})$)
- pour toutes REM extensionnelles REM_{EXTi} et REM_{EXTj} définies pour $\mathbf{M}$ et telles que $j=i+1$, $\forall i \in [1,max-1]$, $REM_{EXTj}=REM_{EXTi} \cup \{e_r,\ldots,e_s\}$, avec $\{e_r,\ldots,e_s\} \subseteq (E(\mathbf{M}) \setminus REM_{EXTi})$

– **Règles pour une REM intensionnelle** : soit $\mathbf{M}$ une EM dont l'intension est définie par $I(\mathbf{M})=\{a_1:t_1,a_2:t_2,\ldots,a_n:t_n\}$ avec $a_k \in \{a_1,\ldots,a_n\}$ variable (attribut ou champ) de la structure de chaque élément de $\mathbf{M}$, et $t_k \in \{t_1,\ldots,t_n\}$ type de la variable a_k.

- toute REM intensionnelle REM_{INTi} de $\mathbf{M}$ est non vide (i.e. $REM_{INTi} \neq \emptyset$)
- $REM_{INT1}=\{a_m:t_m,\ldots,a_p:t_p\}$ avec $\{a_m:t_m,\ldots,a_p:t_p\} \subset I(\mathbf{M})$ (i.e. $REM_{INT1} \neq I(\mathbf{M})$)
- pour toutes REM intensionnelles REM_{INTi} et REM_{INTj} définies pour $\mathbf{M}$ et telles que $j=i+1$, $\forall i \in [1,max-1]$, $REM_{INTj}=REM_{INTi} \cup \{a_r:t_r,\ldots,a_s:t_s\}$, avec $\{a_r:t_r,\ldots,a_s:t_s\} \subseteq (I(\mathbf{M}) \setminus REM_{INTi})$

Ces règles imposent que la REM REM_{i+1} (extensionnelle ou intensionnelle) associée au niveau de détail $i+1$ ($1 \leq i \leq max-1$) contienne au moins un élément de plus que la REM REM_i (voir Figure 1). Les REM extensionnelles peuvent être vues comme des masques ordonnés sur l'extension de l'EM associée, alors que les REM intensionnelles peuvent être vues comme des masques ordonnés sur l'intension de l'EM associée.

Nous définissons à présent deux fonctions pour l'accès progressif permettant de passer d'une REM (extensionnelle ou intensionnelle) à une autre :

- la fonction *masquage* permet de passer de la REM REM_i, de niveau de détail i, à la REM REM_{i-1} de niveau de détail $i-1$: $masquage(REM_i)=REM_{i-1}$, avec $2 \leq i \leq max$.
- la fonction *dévoilement* permet de passer de la REM REM_i, de niveau de détail i, à la REM REM_{i-1} de niveau de détail $i+1$: $dévoilement(REM_i)=REM_{i+1}$, avec $1 \leq i \leq max-1$.

Nous appelons *stratification* (extensionnelle ou intensionnelle) d'une EM, une séquence de REM (extensionnelles ou intensionnelles) de cette EM ordonnées par niveau de détail croissant (donc, par inclusion ensembliste). Définir une stratification sur une EM permet un accès progressif aux différentes REM de cette EM, le passage d'un niveau de détail à un autre (d'une REM à une autre) se faisant grâce aux fonctions de masquage et de dévoilement.

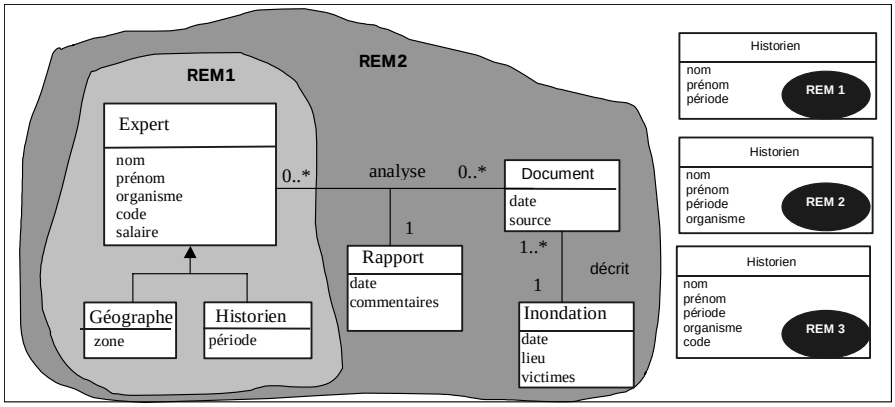


Figure 1. Exemple de deux entités masquables : à gauche, un diagramme de classes stratifié en 2 REM; à droite une classe stratifiée en 3 REM.

La Figure 1 montre une application de l'accès progressif sur un extrait du diagramme de classes UML [OMG, 2001] associé à un système d'information dans le domaine des inondations. A gauche, l'entité masquable considérée est le schéma lui-même. Deux REM extensionnelles lui sont associées. Au premier niveau de détail (REM 1), l'utilisateur a accès aux informations jugées essentielles, c'est-à-dire les instances des classes "*Expert*", "*Géographe*" et "*Historien*". Ensuite, il peut accéder au second niveau (REM 2), aux autres classes, ainsi qu'aux associations "*analyse*" et "*décrit*". A droite, l'entité masquable est la classe "*Historien*", sur laquelle une stratification intensionnelle de 3 REM est définie. Au premier niveau de détail, sur chaque instance de la classe "*Historien*", 3 variables sont visibles ("nom", "prénom" et "période"). Au deuxième niveau, la variable "organisme" devient visible. Au dernier niveau, toutes les variables de la classe sont visibles. Il faut remarquer que la variable "salaire" demeure cachée à l'utilisateur.

Comme le montre cet exemple, l'accès progressif est une notion générique qui peut s'appliquer à divers niveaux de structuration de l'information et peut être appliqué à des fins de confidentialité.

3 Modèles pour l'accès progressif

Nous décrivons ici cinq modèles qui permettent de concevoir des SIW mettant en œuvre un accès progressif personnalisé à l'information.

3.1 Le modèle d'accès progressif

Le Modèle d'Accès Progressif (MAP) permet de spécifier des stratifications validant les règles énoncées plus haut. Le MAP exploite des stéréotypes UML afin d'introduire des éléments de modélisation spécifiques à l'accès progressif. La Figure 2 montre les principales classes du MAP et les associations existant entre elles. Ce modèle générique est valide que la stratification soit extensionnelle ou intensionnelle. Le modèle présenté sur cette figure est simplifié, notamment les variables et méthodes des classes sont cachées.

Chaque notion présentée dans la section 2 est implémentée par une classe stéréotypée. Une instance de la classe "*Entité Masquable*" est composée d'au moins deux instances de la classe "*Elément d'Entité Masquable*". Une instance de la classe "*Stratification*" est représentée comme une agrégation d'au moins deux instances de la classe "*Représentation d'Entité Masquable*", ordonnées par l'inclusion ensembliste. Une instance de cette classe est liée par l'association "*contient*" à une ou plusieurs instances de la classe "*Elément d'Entité Masquable*" qui sont les éléments de l'EM ajoutés par la REM REM_i, au niveau de détail *i*. La relation de dépendance ({*inclusion*}) assure que les éléments ajoutés appartiennent à l'EM.

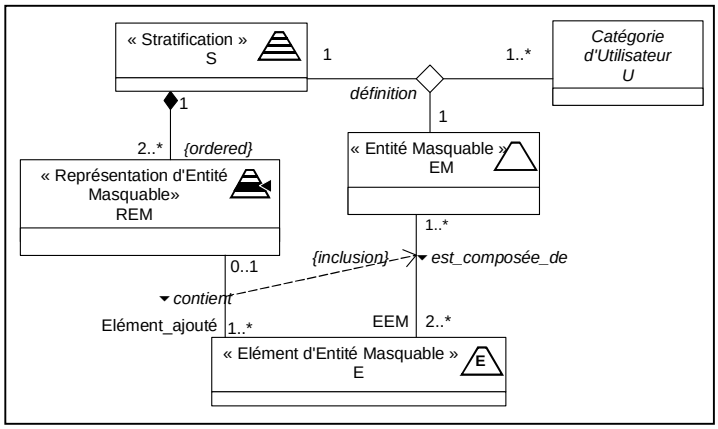


Figure 2. Le MAP décrit par des stéréotypes UML.

Une caractéristique importante du MAP est l'association ternaire "*définition*" qui lie les classes "*Stratification*" (S), "*Entité Masquable*" (EM) et "*Catégorie d'Utilisateur*" (U). Cette dernière est une classe abstraite qui doit être spécialisée (voir section 3.5) de façon à maintenir des informations (profils, besoins, préférences..) concernant les util isateurs et pertinentes au regard des objectifs du SIW. Cette classe permet d'introduire les caractéristiques dédiées à l'adaptation. L'association "*définition*" permet au concepteur de spécifier autant de stratifications que nécessaire pour une entité masquable donnée. Ainsi, chaque utilisateur du SIW peut bénéficier d'un accès progressif personnalisé. Les multiplicités définies pour cette association garantissent que :

- une seule stratification est définie pour un (groupe d')utilisateur(s) et une entité masquable,
- une seule entité masquable est associé à un (groupe d')utilisateur(s) et à une stratification,
- plusieurs (groupe d')utilisateurs peuvent partager la même stratification définie pour une entité masquable.

Les quatre modèles présentés ci-dessous sont étroitement liés (cf. Figure 3) au MAP afin que les principes de l'accès progressif soit répercutés dans chaque dimension (contenu, fonctionnalités, hypermédia) d'un SIW.

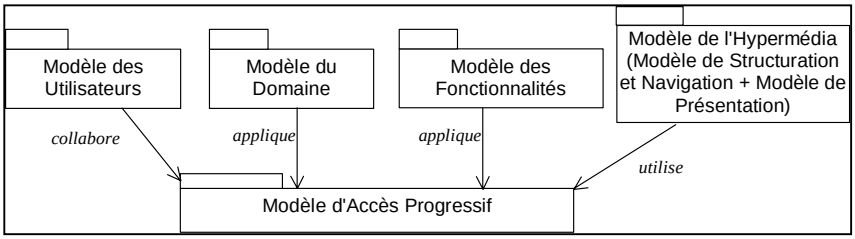


Figure 3. Liens entre le MAP et les 4 autres modèles.

3.2 Le modèle du domaine

Le *Modèle du Domaine* d'un SIW vise à représenter les objets du monde réel qui définissent le domaine d'application. Ce modèle conceptuel met l'accent sur les concepts du domaine d'application et sur les relations entre ces concepts.

Le Modèle du Domaine peut être décrit par un diagramme de classes UML. Il faut noter cependant que l'accès progressif s'applique à un modèle de données quelconques : non structurées, semi-structurées (XML), ou structurées (objet ou relationnel). Par exemple, nous avons montré [Villanova *et al.*, 2001] comment coupler le MAP avec un formalisme de représentation de connaissances par objets et bénéficier d'un accès progressif sur les objets d'une base de connaissances.

Dans le Modèle du Domaine choisi, chaque constituant (classe, association ou classe-association) peut être considéré comme une entité masquable (cf. Figure 1) et peut donc être stratifié. Au niveau du MAP, cela signifie que la classe "Entité Masquable" fait référence au composant correspondant du Modèle du Domaine. Dès lors, les principes de l'accès progressif via le MAP peuvent être appliqués au

Modèle du Domaine (cf. Figure 3). Nous soulignons le fait que la conception d'un Modèle du Domaine doit s'appuyer sur une analyse des besoins des utilisateurs du SIW qui peut être décrite en utilisant les cas d'utilisation d'UML. De même, ces cas d'utilisation constituent un point de départ dans la conception du Modèle des Fonctionnalités présenté ci-dessous.

3.3 Le modèle des fonctionnalités

Le *Modèle des Fonctionnalités* décrit les tâches qu'un utilisateur peut réaliser avec le SIW. Ce modèle repose sur une organisation en trois niveaux de granularité des concepts d'*espace fonctionnel*, de *rôle fonctionnel* et de *fonctionnalité*. La conception du Modèle des Fonctionnalités est guidée par les besoins des utilisateurs et peut donc s'appuyer sur la description des cas d'utilisation du SIW. Au niveau le plus bas, les *fonctionnalités* attendues du SIW sont décrites. Elles sont identifiées à partir de cas d'utilisation où les acteurs sont des personnes utilisant le système (par opposition aux cas d'utilisation où les acteurs sont, par exemple, des composants de l'application, d'autres systèmes, etc.). Au niveau intermédiaire, les fonctionnalités définies pour un même acteur sont regroupées dans un *rôle fonctionnel*. Au niveau, le plus haut, un *espace fonctionnel* est associé à chaque utilisateur et rassemble les divers rôles fonctionnels que cet utilisateur peut jouer. Un utilisateur peut remplir les rôles fonctionnels de plusieurs acteurs.

Chacun des trois concepts de l'organisation peut être considéré comme une entité masquable et, par conséquent, être référencé par le MAP. En effet, un espace fonctionnel est un ensemble de rôles fonctionnels, un rôle fonctionnel est un ensemble de fonctionnalités, et le résultat d'une requête correspondant à une fonctionnalité est un ensemble d'éléments d'information. Ceci signifie que l'on peut disposer dans un SIW d'un accès progressif à l'information (les données) mais également sur chacun des trois niveaux de l'organisation de la dimension fonctionnelle de ce SIW. Cette propriété est exploitée par le Modèle de l'Hypermédia afin de présenter de manière graduelle à un utilisateur, l'ensemble des fonctionnalités que le SIW met à sa disposition.

3.4 Le modèle de l'hypermédia

Le *Modèle de l'Hypermédia* est composé de deux sous-modèles. Le premier, appelé *Modèle de Structuration et de Navigation*, décrit la structuration de l'hypermédia et les chemins de navigation empruntables dans le SIW, en accord avec l'accès progressif établi. Ce modèle qui utilise les concepts décrits par le MAP (cf. Figure 3), constitue une première étape vers l'implémentation de la logique de l'accès progressif dans un hypermédia. Le second, appelé *Modèle de Présentation*, est chargé de la composition et de l'apparence des pages Web dont l'organisation est décrite par le Modèle de Structuration et de Navigation. Par manque de place, nous ne présentons ici que le Modèle de Structuration et de Navigation (voir [Villanova, 2002] pour une description du Modèle de Présentation).

La structure de l'hypermédia (la partie visible d'un SIW) est dérivée du Modèle des Fonctionnalités par le Modèle de Structuration et de Navigation. Cette structure est constituée de *contextes* et de *nœuds*. Les contextes correspondent aux trois types de concepts du Modèle des Fonctionnalités (espace fonctionnel, rôle fonctionnel, et fonctionnalité) qui peuvent être vus comme des entités masquables et donc être stratifiés. Chaque contexte est composé de trois types de nœuds : un *nœud stratification* qui représente la stratification du contexte en tant qu'EM, des *nœuds REM* qui représentent les REM de la stratification, et des *nœuds éléments* qui représentent pour chaque REM de la stratification, l'ensemble des éléments qui la composent. La navigation dans le SIW peut s'effectuer, soit *à l'intérieur* d'un contexte, en utilisant les mécanismes de masquage et de dévoilement (on navigue ici à l'intérieur d'un contexte, entre les REM, on change donc de niveau de détail), soit *entre* deux contextes à la condition que le contexte de départ soit à la fois un élément de REM et une EM (on navigue ici entre deux contextes, on change donc niveau de granularité et de stratification).

En tant qu'EM, un espace fonctionnel est stratifié en REM qui sont des ensembles de rôles fonctionnels. De même, un rôle fonctionnel peut être stratifié en REM qui sont des ensembles de fonctionnalités. Enfin, une fonctionnalité peut être stratifiée en REM qui sont, par exemple, des ensembles de variables impliquées dans le résultat de la requête associée à cette fonctionnalité. Sur la Figure 4, les flèches grises correspondent aux liens de navigation de masquage et de dévoilement à l'intérieur d'un contexte. Sur la gauche est montré l'effet de l'activation de ces liens dans le contexte d'un espace fonctionnel. Ces liens permettent d'accéder progressivement aux différents rôles

fonctionnels (un par groupe) que l'utilisateur peut remplir. Les flèches noires symbolisent un changement de contexte. Une fois que l'utilisateur a choisi un rôle fonctionnel, il accède à l'ensemble des fonctionnalités qui constitue ce rôle. Il peut ensuite choisir la fonctionnalité à activer en naviguant progressivement dans cet ensemble Lorsque la fonctionnalité cible est atteinte, il peut utiliser les mécanismes de masquage et de dévoilement pour obtenir plus ou moins de détails dans le résultat affiché par l'exécution de la requête associée à la fonctionnalité.

Le résultat d'une requête est vu comme un ensemble d'instances (ou objets) de classes, d'associations ou de classes-associations du Modèle du Domaine. Par exemple, dans le cas où la fonctionnalité consiste simplement à visualiser l'ensemble des instances d'une classe et si une stratification intensionnelle est définie pour cette classe, alors l'utilisateur peut accéder graduellement à plus ou moins d'informations (caractérisées ici par les variables) sur les instances de cette classe à travers les mécanismes de masquage et de dévoilement.

L'organisation des pages Web du SIW est dictée par le Modèle de Structuration et de Navigation tel que décrit à la Figure 4, alors que les liens de navigation (affichage d'une nouvelle page) sont modélisés par les changements de contexte (flèches noires) et, éventuellement, par les fonctions de masquage et dévoilement.

Il est également possible de définir des contextes additionnels et des nœuds (par exemple, le contexte Login dans la Figure 4) ou de lier deux contextes qui correspondent à des éléments de la même REM, par exemple deux fonctionnalités dont les résultats sont simultanément affichés dans la même page Web.

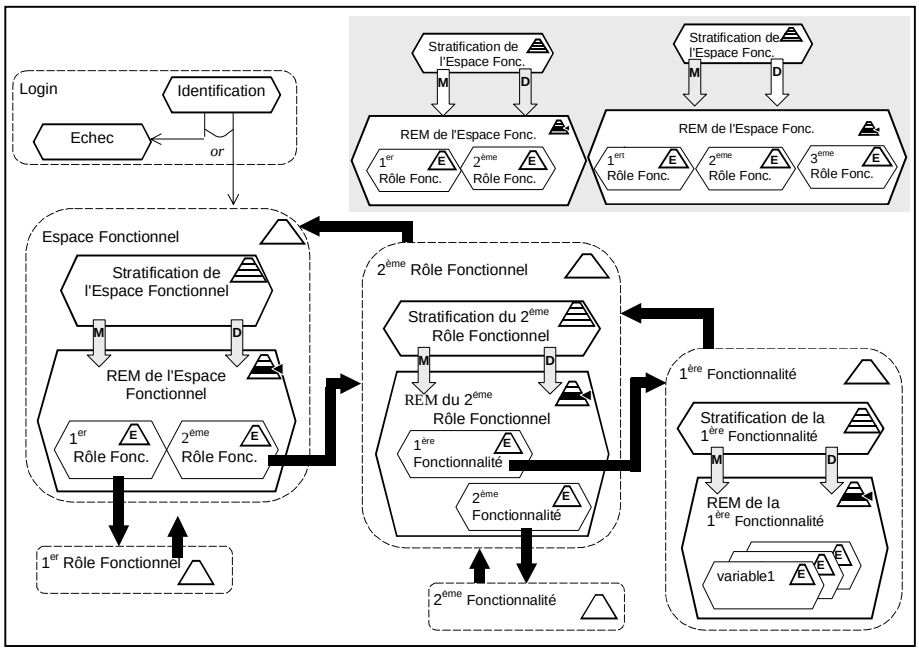


Figure 4. Le Modèle de Structuration et Navigation dérivé du Modèle des Fonctionnalités.

3.5 Le modèle des utilisateurs

Dans l'approche de conception basée sur UML, l'utilisateur est considéré comme un acteur, une entité externe qui interagit avec le système mais qui n'a pas de représentation interne. S'agissant d'adaptation, le système a besoin de stocker des informations sur les utilisateurs. C'est pourquoi certaines méthodes de conception d'application Web proposent une représentation interne des utilisateurs [Gnaho, 2001][Koch, 2000][Ceri *et al.*, 2000]. Nous adoptons également cette approche. Notre *Modèle des Utilisateurs* peut être représenté par un diagramme de classes UML. Nous distinguons les *groupes d'utilisateurs* des *utilisateurs individuels.* Une classe "*Groupe*" (resp. "*Utilisateur*") décrit les profils des groupes d'utilisateurs (resp. des utilisateurs individuels). Ces deux classes sont des sous-classes de la classe "*Catégorie d'Utilisateur*" introduite dans le MAP (cf. Figure 2). La relation entre le MAP et le Modèle des Utilisateurs peut être vue comme une collaboration. D'un côté, il est possible d'étendre un profil d'utilisateur en prenant en compte les besoins et

préférences en matière de contenu et de présentation de l'information. D'un autre côté, le Modèle des Utilisateurs permet de personnaliser chaque stratification. Cela augmente la flexibilité et la portée du MAP. Un groupe d'utilisateurs rassemble l'information concernant des utilisateurs qui jouent le même rôle fonctionnel. Un utilisateur peut appartenir à un ou plusieurs groupes. Concernant la description des profils, en plus des variables classiques qui fournissent le nom et la description d'un groupe, ou les paramètres d'identification et d'authentification pour un utilisateur, trois profils sont considérés en tant que sous-classes de la classe "*Groupe*" (resp. "*Utilisateur*") :

- *G_AccèsProgressif* (resp. *U_ AccèsProgressif*), qui informe sur les stratifications établies pour ce groupe;

- *G_Préférences* (resp. *U_ Préférences*), qui décrit les préférences du groupe (resp. utilisateur) en termes de présentation de l'hypermédia;

- *G_Comportement* (resp. *U_ Comportement*) qui décrit le comportement du groupe (resp. utilisateur) lors de l'utilisation du SIW et peut être utilisé pour une adaptation dynamique du contenu et de la présentation.

Chacune de ces trois classes représente un profil utilisé pour l'adaptation. Le profil d'accès progressif est exploité par le Modèle des Fonctionnalités et, par dérivation, par le Modèle de Structuration et de Navigation, pour construire la structure hypermédia des pages Web présentées par le SIW. Les préférences sont exploitées par le Modèle de l'Hypermédia pour construire les pages personnalisées à partir des chartes graphiques (i.e. des spécifications imposées pour la présentation) associées au groupe ou à l'utilisateur. Le profil comportemental est d'abord exploité pour stocker les actions réalisées par le groupe ou l'utilisateur sur le SIW. De telles informations peuvent être utilisées par le MAP afin de réorganiser les stratifications associées au groupe ou à l'utilisateur, en utilisant les primitives décrites dans [Villanova, 2002]. Il devient alors possible pour le SIW de s'adapter de manière dynamique à ses utilisateurs.

4 Méthodologie de conception

Les relations de dépendance (cf. Figure 3) entre les cinq modèles de notre proposition déterminent partiellement les grandes lignes du guide méthodologique de conception que nous présentons dans cette section. Ces dépendances résultent du fait que la définition d'un modèle requiert que la spécification d'un ou plusieurs autres modèles soit achevée.

En amont du processus de conception, une analyse des besoins fonctionnels doit être menée. Cette étape préliminaire conduit à l'élaboration de cas d'utilisation qui permettent d'identifier l'information essentielle à la spécification du Modèles du Domaine et du Modèle des Fonctionnalités. En aval, la génération du SIW termine ce processus. Entre ces deux phases, s'intercalent les sept étapes de conception suivantes :

- *Etape 1 : Définition du Modèle du Domaine.* Elle consiste à créer les classes, associations et classes-associations qui représentent le domaine d'application.

- *Etape 2 : Définition du Modèle des Fonctionnalités.* Cette étape requiert la création des Fonctionnalités du SIW qui doivent être identifiées durant la phase d'analyse des besoins et peuvent être décrites par des cas d'utilisation. Une fonctionnalité est représentée par un nom et par la description de la tâche qu'elle permet d'effectuer. Cette étape inclut également la spécification de la requête sur le Modèle du Domaine qui correspond à la fonctionnalité. Les rôles fonctionnels sont alors décrits (par un nom et une description) et chacun d'eux est lié à un ensemble de fonctionnalités.

- *Etape 3 : Définition du Modèle des Utilisateurs (1/3).* La définition du Modèle des Utilisateurs est une étape particulière divisée en trois sous-étapes. Dans cette première sous-étape, les groupes sont identifiés et un rôle fonctionnel est associé à chacun d'eux. Les utilisateurs individuels identifiés par le concepteur du SIW sont créés et déclarés membres des groupes adéquats.

- *Etape 4 : Identification des Besoins d'Accès Progressif.* Afin de faciliter la description des stratifications, nous proposons des notations spécifiques qui permettent d'exprimer des besoins en termes d'accès progressif. Ces notations sont basées sur des cas d'utilisation et sur des diagrammes de séquence UML. Elles sont décrites, ainsi que les directives spécifiques qui accompagnent leur utilisation, dans [Villanova, 2002].

- *Etape 5 : Définition du Modèle des Utilisateurs (2/3).* Dans cette deuxième sous-étape, les profils d'accès progressif sont définis pour les groupes et les utilisateurs individuels. Les stratifications des rôles fonctionnels et des fonctionnalités sont créées en utilisant les spécifications de l'étape 4. Ces

stratifications sont référencées par chaque profil. Les stratifications des espaces fonctionnels des utilisateurs connus sont à leur tour définies.

- *Etape 6 : Définition du Modèle de l'Hypermédia.* Cette étape débute par la dérivation du Modèle de Structuration et de Navigation (une par groupe). Elle donne donc le squelette de l'hypermédia qui sera proposé à chaque groupe. Chacun de ces modèles est traduit en un Modèle de Présentation en appliquant une charte de composition et une charte graphique par défaut. Le concepteur peut également spécifier une charte de composition et une charte graphique particulière.

- *Etape 7 : Définition du Modèle des Utilisateurs (3/3).* Cette dernière sous-étape consiste à mettre à jour les profils de préférences. Le concepteur référence dans chaque profil de groupe et, au besoin, dans chaque profil d'utilisateur individuel, la charte de composition et la charte graphique à appliquer.

5 Conclusion et perspectives

L'accès progressif est une contribution à l'adaptabilité dans les Systèmes d'Information sur le Web (SIW) qui permet de limiter la surcharge cognitive et la désorientation que peut éprouver un utilisateur lorsqu'il navigue dans l'espace d'information d'un tel SIW. Pour cela, l'espace d'information, qu'il s'agisse des données ou des fonctionnalités du SIW, est stratifié en différents niveaux de détail. Nous avons décrit dans cet article cinq modèles qui permettent d'intégrer l'accès progressif dans un SIW. Le Modèle d'Accès Progressif (MAP) décrit des stratifications sur des entités masquables qui peuvent dès lors être graduellement accédées. Le Modèle du Domaine décrit le domaine d'application. Le Modèle des Fonctionnalités décrit les tâches que les utilisateurs peuvent effectuer sur le SIW. Il possède trois niveaux de granularité (espace fonctionnel, rôle fonctionnel, fonctionnalité) qui sont des entités masquables et qui peuvent donc être également accédées progressivement par l'utilisateur. A partir des spécifications du Modèle des Fonctionnalités, une représentation logique de l'hypermédia du SIW est dérivée par le Modèle de Structuration et de Navigation qui constitue la première partie du Modèle de l'Hypermédia. Le Modèle des Présentations complète le modèle Hypermédia et décrit l'apparence des pages Web affichées par le SIW. Nous avons ensuite donné un guide méthodologique permettant en sept étapes de concevoir un SIW intégrant l'accès progressif à l'information.

Nous avons implémenté cette approche (les cinq modèles et les sept étapes de conception) dans KIWIS [Villanova *et al.*, 2002], une plate-forme pour la conception et la génération de SIW adaptables et proposant un accès progressif à l'information. KIWIS se présente comme un serveur WEB qui permet soit de créer un SIW, soit d'accéder à un SIW existant et créé avec KIWIS. En mode conception, KIWIS assiste le concepteur du SIW dans sa tâche en décomposant le processus de conception en les sept étapes présentées et lui permet d'instancier les cinq modèles présentés. Une fois ces étapes achevées, KIWIS génère le SIW en le mettant à la disposition des utilisateurs autorisés. En mode utilisation, KIWIS constitue l'environnement d'exécution permettant d'activer les fonctionnalités du SIW généré. KIWIS est utilisé notamment dans le projet européen SPHERE [Davoine *et al.*, 2001] qui est un Système d'Information dédié aux données géographiques et historiques sur les inondations. Les stratifications permettent à différentes catégories d'utilisateurs (experts en hydrologie, employés de mairie, élus..) de consulter des données personnalisées (en contenu et présentation) sur le même thème (inondations) mais avec des niveaux de détail différents et des centres d'intérêt différents.

Les travaux futurs concernent l'extension de l'accès progressif vers l'adaptabilité dynamique. A partir d'une description du comportement des utilisateurs et l'établissement de règles de décision, le SIW pourra dynamiquement (à l'exécution) ré-organiser les stratifications, s'il apparaît que les niveaux de détail existants ne sont pas satisfaisants. C'est par exemple le cas lorsque le système détecte que l'utilisateur accède toujours à l'information située au troisième niveau de détail sans utiliser les deux premiers.

6 Bibliographie

[Baresi et al., 2000] Baresi L., Garzotto F., Paolini P., From Web Sites to Web Applications: New Issues for Conceptual Modeling. Proc. of the *International Workshop on The World Wide Web and Conceptual Modeling*, co-located with the *19th International Conference on Conceptual Modeling*, Salt Lake City (USA), October 2000.

[Brusilovsky, 1998] Brusilovsky P., Methods and Techniques of Adaptive Hypermedia. In Brusilovsky P., Kobsa A. & Vassileva J. (eds.), Adaptive Hypertext and Hypermedia, Kluwer Academic Publishers, 1998, pp. 1-43.

[Ceri et al., 2000] Ceri S., Fraternali P. & Bongio A., Web Modeling Language (WebML): a modeling language for designing Web sites, Proc. of the *9th International World Wide Web Conference (WWW9)*, Amsterdam, Netherlands, May 15 - 19, 2000.
[Ceri et al., 2000] Ceri S., Fraternali P., Bongio A., Maurino A., Modeling data entry and operations in WebML, *WebDB 2000*, Dallas, USA, 2000.
[Conallen, 2000] Conallen J., *Building Web applications with UML*, Addison-Wesley Longman, 2000.
[Conklin , 1987] Conklin J., Hypertext: An introduction and survey. *IEEE Computer 20(9)*, 1987, pp. 17-41.
[Davoine et al., 2001] Davoine, P.A., Martin, H., Trouillon, A., Cœur, D., Lang, M., Bariendos M., Llasat C.: Historical Flood Database for the European SPHERE Project: modelling of historical information, Proc. of the *21th General Assembly of the European Geophysical Society*, Nice, 2001.
[De Troyer et al., 1998] De Troyer O.M.F., Lejeune C.J., WSDM : a User-Centered Design Method for Web Sites, *7th Int. World Wide Web Conference (WWW7)*, April 1998.
[Frasincar et al., 2001] Frasincar F., Houben G-J., Vdovjak R., An RMM-Based Methodology for Hypermedia Presentation Design, Proc. of the *5th East European Conference on Advances in Databases and Information Systems (ADBIS 2001)*, LNCS 2151, Vilnius, Lithuania, September 25-28, 2001, pp. 323-337.
[Fraternali, 1999] Fraternali P., Tools and Approaches for Developing Data-Intensive Web Applications: A Survey, *ACM Computing Surveys*, 31(3), September 1999, pp. 227-263.
[Garzotto et al., 1993] Garzotto F., Mainetti L., Paolini P. HDM2: Extending the E-R Approach to Hypermedia Application Design, in Elmasri R., Kouramajian V., & Thalheim B. (Eds), Proc. of the *12th Int. Conference on the Entity-Relation Approach (ER' 93)* Arlington, TX, December 1993, pp. 178-189.
[Gnaho, 2001] Gnaho C., Web-based Informations Systems Development – A User Centered Engineering Approach, in Murugesan S. & Deshpande Y. (Eds.), *Web Engineering: Managing Diversity and Complexity of Web Application Development*, LNCS 2016, 2001, pp. 105-118.
[Isakowitz et al., 1995] Isakowitz T., Stohr A., Balasubramanian E., RMM : A methodology for structured hypermedia design. *Communications of the ACM 38(8)*, pp. 34-44, August 1995.
[Isakowitz et al., 1998] Isakovitz T., Bieber M., Vitali F., Web Information Systems, *Communications of the ACM*, 41(7), July 1998, pp. 78-80.
[Koch, 2000] Koch N., Software Engineering for Adaptative Hypermedia Systems – Reference Model, Modelling Techniques and Development Process, Ph.D Thesis, Fakultät der Mathematik und Informatik, Ludwig-Maximilians-Universität München, December 2000.
[Mecca et al., 1999] Mecca G., Merialdo P., Atzeni P., Crescenzi V., The ARANEUS Guide to Web-Site Development, ARANEUS Project Working Report, AWR-1-99, March 1999.
[OMG, 2001] Object Management Group, Unified Modeling Language Specifications, Version 1.4, Sept. 2001. http://www.omg.org
[Rossi et al., 2001] Rossi G., Schwabe D., Guimarães R., Designing Personalized Web Applications, Proc. of the 10th Int. *World Wide Web Conference (WWW10)*, May 1-5, Hong Kong, 2001.
[Stephanidis et al., 1998] Stephanidis C., Paramythis A., Akoumianakis D., Sfyrakis M., Self-Adapting Web-based Systems: Towards Universal Accessibility, Proc. of the *4th Workshop on User Interface For All*, Stockholm, Sweden, October, 1998.
[Villanova et al., 2001] Villanova M., Gensel J., Martin H., Progressive Access to Knowledge in Web Information Systems through Zooms, Proc. of the *7th International Conference on Object Oriented Information Systems (OOIS2001)*, Calgary, Canada, August 27-29, 2001, pp 467-476.
[Villanova et al., 2002] Villanova-Oliver M., Gensel J., Martin H., Erb C., Design and Generation of Adaptable Web Information Systems with KIWIS, Proc. of the *3rd IEEE International Conference on Information Technology (ITCC2002)*, Las Vegas, USA, april 8-10, 2002.
[Villanova, 2002] Villanova-Oliver M., MAP : Un modèle pour l'accès progressif à l'information Systems, Proc. du *XXème Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2002)*, Juin 4-7, 2002, Nantes, France.
[Villanova, 2002] Villanova-Oliver M., Adaptabilité dans les systèmes d' Information sur le Web : Modélisation et mise en œuvre de l' accès progressif, Thèse de Doctorat, Institut National Polytechnique de Grenoble, décembre 2002.

Serveur dynamique de cartes géostatistiques[1]

E.EDI, S. OULAHAL, JM. VINCENT

Laboratoire d'Informatique et Distribution,
Projet APACHE,
ZIRST, Antenne ENSIMAG,
51 avenue Jean KUNTZMANN,
38330 Montbonnot Saint Martin, FRANCE.
Email : {Euloge.Edi,Said.Oulahal,Jean-Marc.Vincent}@imag.fr
Tél : +33 4 76 61 20 39 Fax : +33 4 76 61 20 99

Résumé

Fournir sur le web à des utilisateurs distants, des outils efficaces d'aide à la décision est une des opportunités de l'internet. Dans ce but, nous développons un serveur intéractif de cartes géographiques, dont l'efficacité (temps de réponse, robustesse) est améliorée grâce à l'utilisation de la programmation parallèle couplée avec des *caches de données* partiels. L'objectif de cet article est de décrire les relations entre les structures de données utilisées et le modèle des requêtes qui permettent une parallélisation et l'utilisation de caches.

Abstract

Providing tools on internet for decision in social or economical contexts is of great interest. A new web-platform has been developped to generate maps on-demand representing social phenomena on geographical areas. Quality of service such as interactivity, robustness is improved by coupling the web server to a computing cluster of PCs. Taking into account the structure of the problem, we propose a method to parallelize this kind of application and manage data caches.

[1]Ce travail est en partie soutenu par le projet européen **ORATE-EPSON** en association avec le LSR-IMAG, l'UMS RIATE et l'UMR GEOGRAPHIE-CITE

1 Introduction

Internet est le lieu par excellence de présentation des personnes et des entreprises diverses. Les applications accessibles via le web sont de plus en plus nombreuses et consommatrices de ressources (cpu, bande réseau, ...). Notre objectif concret est de construire un serveur web qui fournisse en temps réel, des cartes géostatistiques. Ce serveur, outil d'aide à la décision, doit pouvoir générer des informations simulant plusieurs "stratégies". Par exemple, on pourrait étudier l'impact du passage de l'Europe des 15 à l'Europe des 25 ; ou encore analyser ce que pourrait être la contribution de la Suisse ou de la Norvège si elles rejoignaient l'Europe.

Nous avons en entrée des fichiers de données représentant le découpage de l'Europe en communes (nom, frontiere, ...) et des valeurs obtenues par recensement (comme le PNB/habitant, la natalité, la population, ...) sur l'ensemble de ces communes. Chaque commune avec l'ensemble de ses données constitue une unité élémentaire de traitement. Les cartes fournies par le serveur sont le résultat d'un traitement personnalisé, effectué à la demande sur l'ensemble des données. L'activité de notre serveur est basée sur des agrégations et désagrégations d'unités, suivies d'une génération à la volée des cartes résultant du traitement effectué. Une requête client consiste à demander un traitement (ex : calcul des disparités) en spécifiant le niveau de représentation souhaité (ex : l'europe des 15), l'espace de référence (ex : découpage en communes) et la donnée statistique à visualiser (ex : population, richesse, ...)(Figures 1 & 2)[2]. L'objectif scientifique de ce travail est de développer, en relation avec le mode de représentation des données et le modèle des requêtes servies, une méthode de décomposition parallèle de l'application pour accélerer son exécution et utiliser efficacement des caches de données. En effet, des observations sur des séquences de requêtes montrent qu'une même partie de calcul peut être partagée par différentes requêtes.

2 Architecture logicielle du serveur

2.1 Représentation des données

Chaque unité élémentaire correspond au plus petit découpage de l'espace considéré (ex : découpage de l'Europe sur l'ensemble des communes (niveau NUTS-5). Les fichiers de données, base du traitement, sont très grands (ex : 791 Mo pour un recensement de 116000 unités) et peuvent évoluer dans le temps. Définir une architecture logicielle qui résiste à l'évolution des données mais sur laquelle s'appuie le traitement évite de reconstruire le logiciel à chaque modification.
On construit la structure de donnée globale décomposée en tables et relations sur lesquelles on distingue les attributs caractéristiques des unités, des valeurs qui leurs sont associées (figure 3).

[2]Cette interface est développée par le LSR-IMAG, Philippe MARTIN au sein du projet ORATE-ESPON

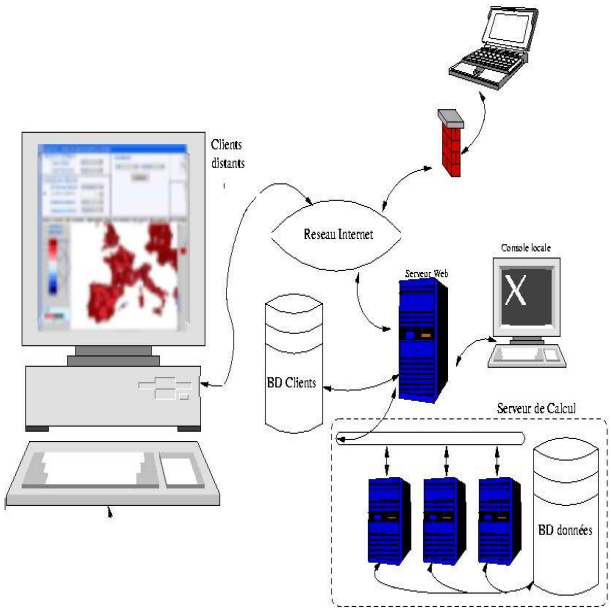


FIG. 1 – Architecture générale du serveur.

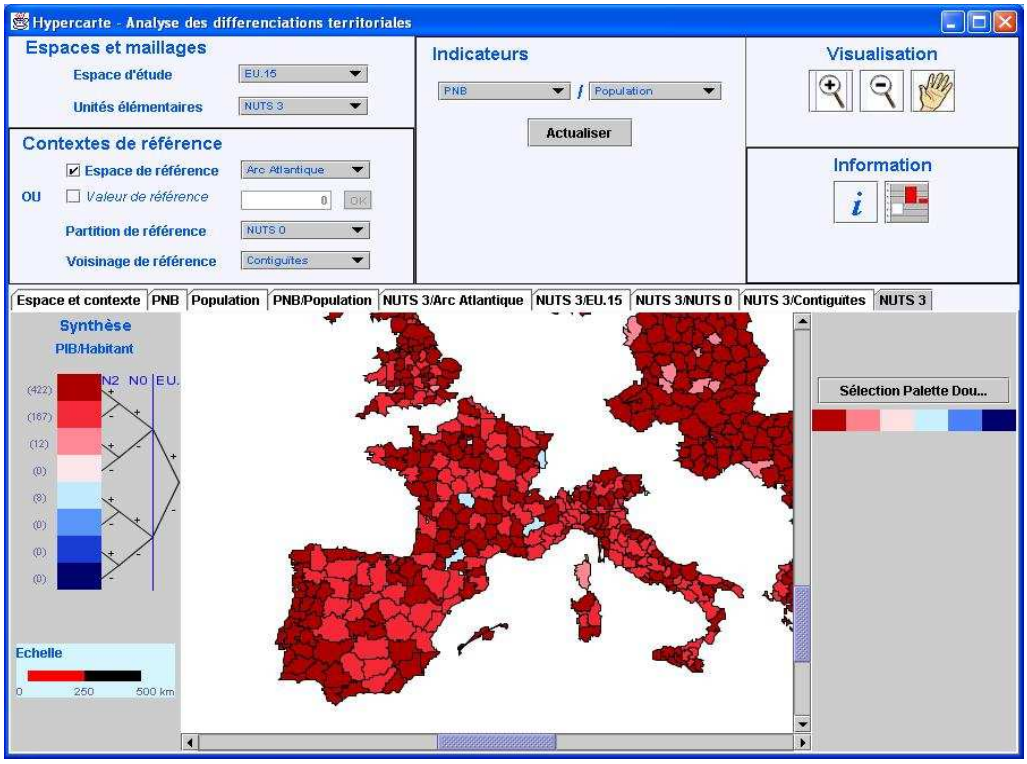


FIG. 2 – Interface utilisateur : on y retrouve les paramètres de la requête à soumettre au serveur web : espace et maillage, contexte d'analyse, variables statistiques et paramètres de visualisation

Les attributs caractéristiques d'une unité (nom, code, unités supérieures) permettent de la situer dans la hiérarchie des unités (dans l'ensemble du modèle) et servent à y accéder lors des traitements. Les valeurs associées (frontière, ressources), sont celles sur lesquelles sont concentrés les calculs (somme et division pour les variables statistiques ; comparaison, ajout et suppression d'arcs pour les frontières). La structure de donnée présente des redondances d'informations qui diminuent les temps d'exécution car :

- Le temps de parcours de la structure est diminué ;
- Le temps de recherche est faible (recherche ciblée sur un attribut particulier d'un tableau) ;
- Les transformations sur un attribut particulier se font sans avoir à gérer toutes les données (travailler séparement sur les frontières et les variables).

Les requêtes portent essentiellement sur des agrégations/desagrégations d'unités ce qui modifie la géométrie des unités et les variables statistiques qui leur sont associées.

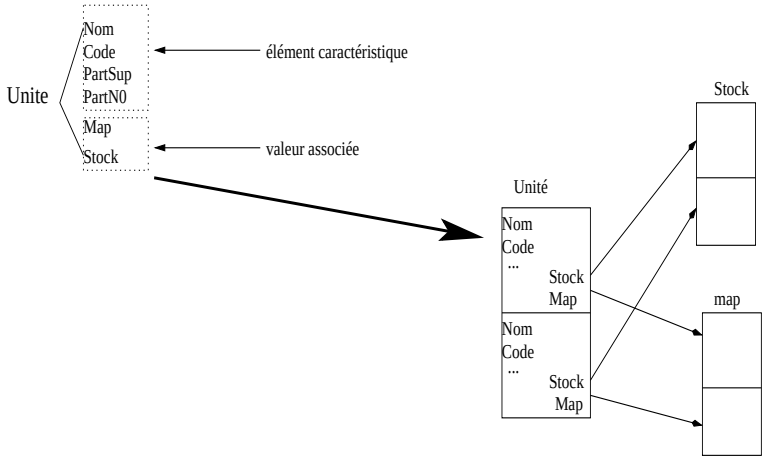


FIG. 3 – On peut modifier le découpage en frontière ou mettre à jour les données de récensement sans modifier la structure.

Par exemple, le calcul des disparités régionales dans un ensemble contenant N unités élémentaires se compose de :

- Calcul des relations de voisinage : complexité en $O(N)$ (le nombre de voisins sur chaque carte géographique est en général majoré par 10) ;
- Comparaison des valeurs de ressources : complexité en $O(N)$.

Au vu de la taille des fichiers en entrée et les calculs à effectuer, une telle application est impossible à réaliser de manière interactive sur une plate-forme monoprocesseur et nécessite donc une parallélisation.

2.2 Architecture

Les études précédentes sur les serveurs de requêtes dynamiques ont montré que le facteur influençant le temps de réponse utilisateur est le temps de calcul du résultat de la requête [1, 7]. L'architecture cible [4], est un serveur dont le module de calcul est une grappe de PCs. Le frontal est la première structure du serveur qui reçoit les requêtes. Il contient un cache qui lui permet de garder les traces des requêtes déjà exécutées et de conserver des résultats intermédiaires acquis. Le frontal transmet les requêtes non présentes dans son cache au serveur de calcul après les avoir transformées en graphe de tâches pour un traitement en parallèle. On utilise Athapacan-1 [3], un langage de programmation parallèle qui supporte un modèle de programmation haut niveau, au sens où il est indépendant de l'architecture de la machine cible. Athapascan construit automatiquement le graphe de tâches en utilisant la granularité fixée par l'utilisateur et les relations de précédences imposées par le programme. Il ordonnance dynamiquement ce graphe sur la machine cible, dans notre cas une grappe de PC (de l'ordre de 60 noeuds).

2.3 Calcul

Dans notre serveur, tout traitement se fait sur un espace de référence prédéfini, qui consiste à choisir les pays formant l'entité Europe, et le niveau de découpage interne de chacun de ces pays, unité élémentaire de recensement des variables. Par exemple, pour de la prospective on veut représenter l'Europe des 27 (l'Europe considéré avec 27 pays) sur l'ensemble des ses régions. Précalculer l'ensemble des espaces de référence permet de définir plusieurs niveaux hiérarchiques qui s'agrègent comme le permettent les unités initiales (Figure 4). Ces résultats sont conservés dans le cache du serveur de calcul. Ainsi, tout ou partie de chaque requête peut être exécutée à différents niveaux du système client/serveur. Exemple : représenter l'Europe des 15 sur l'ensemble des pays suivant la population est une opération simple qui peut être faite sur la machine cliente. L'affichage après calcul peut donc être exporté chez le client.

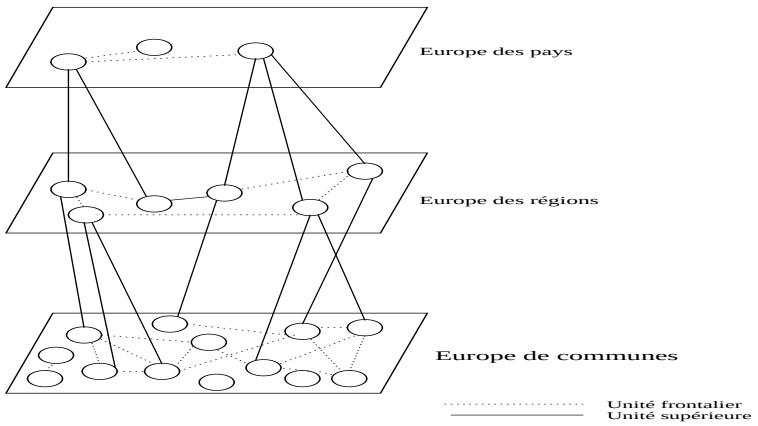


FIG. 4 – Hiérarchie des découpages prédéfinis.

L'ensemble des traitements se compose d'agrégations et de désagrégations d'unités suivies de mises à jour des valeurs associées (variables statistiques et frontières) aux nouvelles unités formées. On utilise des algorithmes d'*Union-Fusion* sur des structures arborescentes. Les niveaux hiérarchiques définis plus haut correspondent bien à la base de traitement de ces algorithmes. Les traitements sont faits au niveau le plus bas pour être ensuite remontés aux niveaux correspondants aux résultats demandés par l'utilisateur. L'amélioration du traitement sur le serveur se fait de deux façons :

- Parallélisation du code ;
- Utilisation d'un cache web.

La parallélisation du code d'une requête se fait après une recherche infructueuse de sa réponse dans le cache du frontal. Dans ce cas, on développe le graphe de tâche associé. Comme pour des requêtes semblables, la structure du graphe de tâche est identique, seules quelques tâches s'exécutent avec des jeux de paramètres différents. Par exemple, calculer les disparités régionales suivant le PNB ou suivant la natalité sur l'Europe génère deux graphes de tâches pour lesquels les calculs sur les frontières sont identiques. L'originalité de ce travail repose donc sur les choix de construction du graphe de tâche en se préoccupant de la complexité des tâches et de leur fréquence dans les flots de requêtes.

Notre serveur contient donc deux caches :

- L'un, sur le frontal dont le but est de conserver des résultats complets des requêtes déjà exécutées ;
- L'autre, sur la grappe de calcul qui conserve des résultats de traitements intermédiaires pour permettre à d'autres requêtes utilisant ceux-ci d'être exécutées beaucoup plus rapidement.

La politique de gestion du cache privilégie les tâches indépendantes. La cohérence des reponses contenues dans le cache avec celles que l'on peut calculer est assurée par le fait qu'une modification des fichiers de données élémentaires invalide toutes les données correspondantes dans le cache.

2.4 Décomposition d'une session

On distingue plusieurs profils utilisateurs travaillant sur les mêmes données. L'objectif principal est de décomposer la session cliente en tâches élémentaires de manière à avoir un traitement récurrent qui peut alors être mis en cache. On peut résumer une session utilisateur en un graphe défini comme suit :

Ce graphe représente les tâches qui peuvent être exécutées dans un ordre quelconque compatible avec les dépendances de données. Le calcul de nouvelles unités consiste à créer de nouveaux ensembles en associant ou en dissociant des unités d'un ensemble initial. On répercute les modifications apportées à la nouvelle structure d'ensemble. Les disparités entre voisins se calculent en définissant le modèle de voisinage. Dans un premier temps, on travaille sur le voisinage géographique.

On distingue donc à travers ce graphe, des actions qui peuvent être repétées lors d'une session utilisateur ou en comparant un ensemble de sessions différentes.

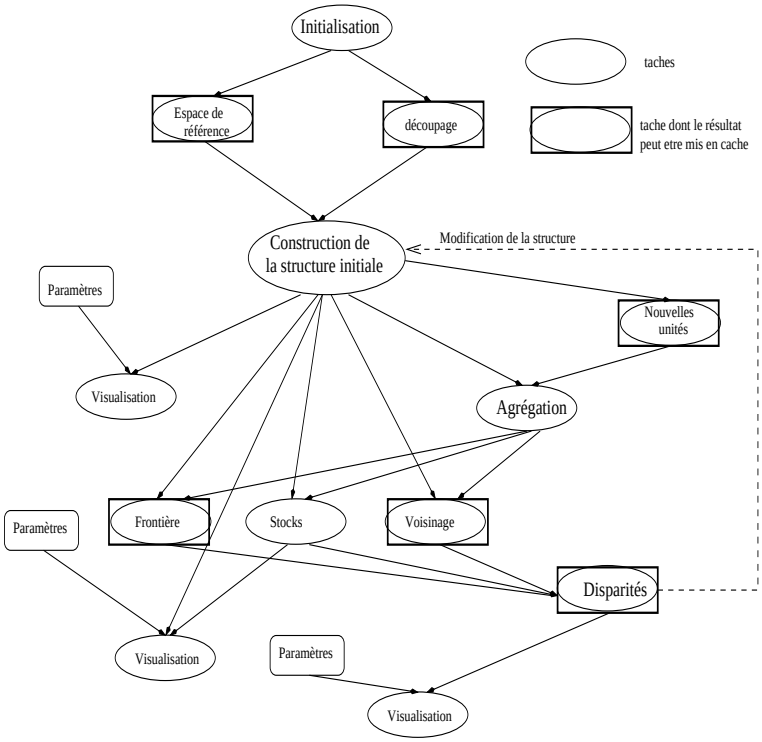


FIG. 5 – Graphe de tâches lors d'une session utilisateur

- Lors d'une session utilisateur, deux traitements sont susceptibles de générer du calcul redondant :
  - La demande de pages précédentes ;
  - Le traitement sur un ensemble de données avec des paramètres modifiés.
- Lors de session utilisateurs différentes, on peut :
  - Avoir le même parcours sur le graphe ci-dessus. Exemple : repondre à un appel d'offre nécessite un type d'étude, qui peut être réalisé par plusieurs clients ;
  - Avoir à faire du traitement sur un ensemble de données avec des paramètres modifiés.

L'analyse de ce graphe permet donc de définir des tâches intermédiaires lors d'une session cliente. L'efficacité du modèle consiste à déterminer parmi ces états élémentaires générés, lesquels sont les plus susceptibles de resservir dans le futur. Dans notre cas ce sont les tâches surchargés par un rectangle. On décompose alors le traitement de sorte à avoir des tâches de granularité les plus grandes possibles et les plus indépendantes des paramètres les plus sensibles aux variations. On recupère donc ces résultats intermédiaires dans des fichiers solutions auxquels on attribue les valeurs :

- Volume de stockage ;
- Temps de calcul (complexité de la solution) ;
- Utilité de la réponse (liée au nombre de rappels de cet élément).

Ces fichiers constituent les éléments de notre cache, rangés par ordre de priorité et hiérar-

chisés. Ils sont identifiés par un nom représentatif de leur contenu, et rangés à différents niveaux du cache selon leur poids évalué par une fonction de coût. La politique de gestion du cache est donc fonction des pondérations associées aux fichiers conservés.

3 Implantation Client-Serveur

Pour définir l'architecture, il est impératif de tenir compte de certaines contraintes :
- Sécurité : Il faut protéger les données de calcul, ainsi que l'accès à la grappe de PC (serveur de calcul).
- Evolution : Les interfaces de communication doivent être évolutives, pour offrir une interconnexion avec d'autres services (facturation, ...). L'architecture doit être indépendante du type de calcul pour faciliter l'intégration de différents services.
- Latence : Il faut réduire au minimum l'échange de données pour ne pas ralentir l'accès au serveur. Pour optimiser la bande passante, certains calculs peuvent être exécutés directement chez le client, après une première phase de traitement au niveau du serveur, et ce sans donner la possibilité au client de les enregistrer (protection des données).
- Il ne doit pas y avoir d'interférences entre les différents clients, surtout dans le cas où ils travaillent sur leurs propres données.

Les contraintes de notre architecture sont d'offrir un service accessible depuis différentes plate-formes, mais aussi depuis tout point géographique via le web. Pour cela l'interface repose sur une Applet Java. L'Applet permet une exécution multi plate-forme sans passer par une phase d'installation de logiciel, si ce n'est la présence d'une JVM compatible (disponible avec les navigateurs). Elle ne permet pas d'écriture sur le poste client, d'où une garantie pour l'utilisateur (confinement) assurant que celui-ci ne puisse pas sauvegarder les données temporaires (respect de la propriété des données statistiques). La seconde contrainte de l'architecture est d'offrir un accès en tout point du réseau Internet. Par mésure de sécurité, certains ports réseaux sont verrouillés par les administrateurs réseaux. Un seul port est au moins toujours disponible en sortie, c'est le port 80 qui correspond au protocole HTTP. L'avantage de ce dernier est de nous permettre de nous affranchir du développement d'un protocole spécifique, coûteux et sans garantie (tunneling http). Cette solution permet aussi d'éviter les contraintes dues aux firewalls, puisqu'au niveau réseau tout se passe comme si l'utilisateur était en train de consulter un site web. L'interface de connexion au niveau serveur est de type cgi-bin (figure 6), qui est l'interface standard sur les serveurs webs.

Dans notre cas particulier, une requête est une trame suivant le format HTTP 1.1 en spécifiant le type MIME correspondant aux images (voir fig 6). Pour optimiser les trames de retour, on n'envoie qu'une image (brouillon). Les résultats finaux sont envoyés par mail à la demande du client sous forme de fichiers textuels contenant des données détaillées prêtes à être intégrées dans un système d'information géographique. L'émission des mails peut être différée en fonction de la charge de travail du serveur. Un des points à améliorer au niveau de ce mode d'échange est l'intégration de ce service au sein d'une autre appli-

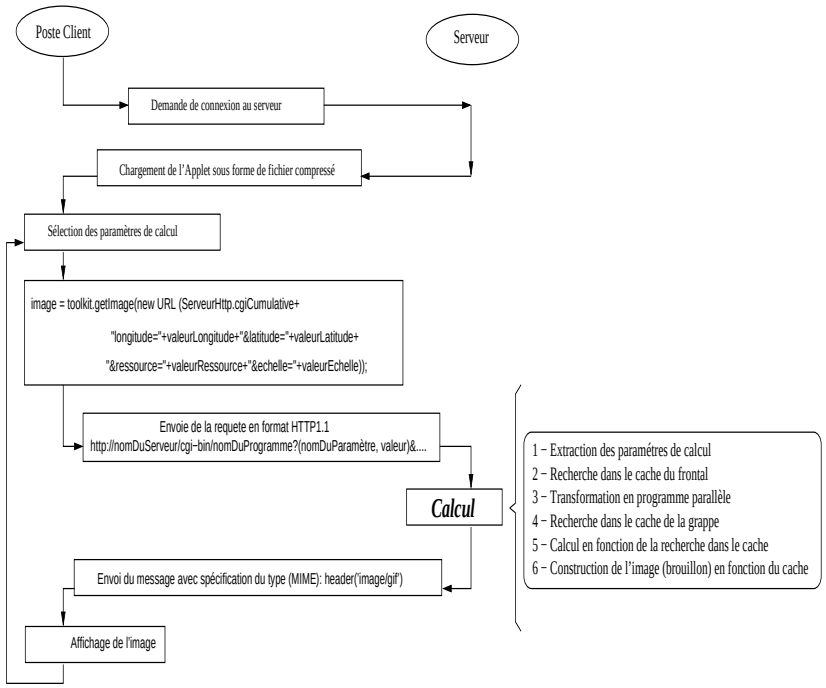


FIG. 6 – Echanges client/serveur.

cation. Nous sommes actuellement obligés de spécifier les paramètres des requêtes sous forme textuelle. Une des évolutions est d'établir des connexions de type Web Services via le protocole SOAP. Dans SOAP, les requêtes sont sous forme HTTP-POST, Ce protocole définit la structure d'élément simple indépendamment du langage ou de la plate-forme. Il reprend le format de structure XML et le principe des RPC. Cette solution a l'avantage d'avoir un format standard pour la liaison de différentes applications multi-plate-formes et multi-langages. Ce protocole entraîne un surcoût en terme d'analyse pour extraire l'information (analyse du fichier XML). Mais, il permet de faire abstraction des types de paramètres.

4 Conclusion

Fournir une application impossible à réaliser sur des machines personnelles par manque de puissance de calcul est le but de ce projet. En utilisant les caractéristiques particulières au modèle des requêtes de notre serveur, on peut améliorer le temps de réponse aux utilisateurs avec du parallélisme et des caches webs. La granularité des tâches du programme est importante car l'indépendance des tâches vis à vis des paramètres donne de l'efficacité aux caches.

Notre travail est dirigé par la structure des applications et les profils des usagers de l'application. Une plate-forme est actuellement en test, elle devrait être intégrée dans le programme Européen ORATE-ESPON. Des utilisateurs pourront ainsi valider ce travail

et nous disposerons de traces d'exécution permettant de valider nos stratégies d'optimisation.

Par la suite, nous envisageons de généraliser cette approche à d'autres applications gourmandes en temps calcul et en interactivité via le réseau Internet (synthèse d'image, traitements numériques, contrôle en robotique,...).

Références

[1] T. YANG B. SMITH, A. ACHARYA. Exploiting result equivalence in caching dynamic web content. *In USENIX Symposium on Internet Technologies and Systems*, 1999.

[2] J-M VINCENT C. GRASLAND, H. MATHIAN. Multiscalar analysis and map generalization of discrete social phenomena : Statistical problems and political consequences statistical. *Journal of the United Nations Economic Commission for Europe*, pages p157–188, Vol 17 no.2 2000.

[3] M. DOREILLE. *Athapascan-1 : Vers un Modèle de Programmation Parallèle adapté au Calcul Scientifique*. PhD thesis, Institut National Polytechnique de Grenoble, 1999.

[4] JM. VINCENT E.EDI, D.TRYSTRAM. Amélioration de performance de serveur web de requêtes dynamiques. In *Colloque Africain sur la recherche en informatique, CARI'02*, 2002.

[5] T. YANG H. ZHU, B. SMITH. Hierarchical resource management for web servers cluster web servers. *In ACM SIGMETRICS*, 1999.

[6] M. DIAS M. COLAJANI, P. S. YU. Analysis of tasks assignement policies in scalable web-server system. *IEEE TPDS, volume 9, no. 6*, 1998.

[7] T. YANG V. HOLMEDHAL, B. SMITH. Cooperative caching of dynamic content on a distributed web server. *In the 7th IEEE International HPDC-7, Chicago*, pages 28–31, IL USA july 1998.

Navigation et recherche par catégorisation floue des pages HTML

F. PAPY, N. BOUHAÏ
Laboratoire Paragraphe,
Université Paris VIII
2, rue de la Liberté ,
93526 Saint-Denis, FRANCE
Mail :fabrice.papy@univ-paris8.fr
nasreddine.bouhai@univ-paris8.fr
Tél : +33 1 49 40 67 58 Fax : +33 1 49 40 67 83

Résumé

Dans le but d'améliorer la recherche et la navigation sur le Web, nous proposons une approche inédite fondée sur la classification de pages. Nous prenons en considération les balisages utilisés dans les pages Web pour élaborer des profils.. Pour établir cette catégorisation de classification automatique des pages, nous nous sommes appuyés sur les travaux d'Alain Lelu en utilisant l'algorithme de K-means axiales. Utilisée dans le moteur de recherches NeuroWeb, cette méthode de sélection automatique de pages a été transposée dans notre dispositif de construction d'espaces de connaissances HyWebMap.

Abstract

In order to improve search and navigation through the Web, we propose an original approach based upon pages classification. We use tags embedded in web pages to build specific profiles. To produce this automatic categorization, we were inspired by Alain Lelu' s research on K-means axial algorithm. Initially used into the search engine Neuroweb, this method was adapted for our knowledge buildin tool ; HyWebMap.

1 Introduction

Internet est une source d'informations stratégiques que l'on peut aujourd'hui difficilement nier. Les entreprises, les organisations gouvernementales ou non sont ainsi confrontées à la pléthore d'informations, à l'abondance des processus nouveaux et à leur rapide obsolescence. Dans ce contexte, s'impose la nécessité d'une collecte et d'une transmission sélective de l'information, de dispositifs permettant de la trouver, de la filtrer et de la traiter automatiquement. Car, le meilleur côtoie aujourd'hui le pire : sites institutionnels des gouvernements, des universités, des centres de recherches sont proposés dans les listes d'URL des moteurs de recherches au même titre que les pages personnelles ou les sites commerciaux. Séparer le bon grain de l'ivraie est une tâche de plus en plus difficile malgré les fonctionnalités de plus en plus sophistiquées des systèmes de recherches d'informations en ligne qui outre le volume, les redondances…doivent répondre aux problèmes du multilinguisme. A l'heure de la surinformation sur les réseaux, la nécessité d'opérer une sélection s'impose donc comme un enjeu déterminant du traitement de l'information.

Les technologies de filtrage et de résumé apportent une réponse aux professionnels de l'information, cyber-documentalistes, courtiers, veilleurs, webmasters, responsables de portails d'information, etc. dans des secteurs aussi différents que le knowledge management, le e-commerce et l'intelligence économique.

La recherche d'informations repose assurément sur une relation étroite entre les possibilités opératoires des outils d'extraction de données et la capacité de l'utilisateur à s'impliquer dans sa recherche. Cette implication sous-entend une volonté de chercher qui s'exprime au travers de stratégies de recherches visant à évaluer, comparer et confronter les résultats.

Développé dans le cadre du laboratoire Paragraphe de l'Université Paris 8, le projet de moteur de recherches NeuroWeb *(http://neuroweb.univ-paris8.fr)* avait pour objectif un dispositif d'exploration et de recherche fine sur le Web en couplant une approche multilingue (utilisant les N-grammes) avec une approche linguistique / sémantique.

Ce projet, retenu à la suite d'un appel d'offres lancé par le Ministère de l'Éducation Nationale et l'Agence Nationale pour la Recherche Technologique, se proposait d'intégrer dans un prototype de moteur de recherche permettant requêtes fines et cartographies à la demande, deux approches complémentaires ; la première exploitant les méthodes d'exploration de corpus à partir de cartographies textuelles utilisant les N-grammes, la seconde utilisant la lemmatisation de textes et les réseaux sémantiques.

Ce moteur de recherche alimenté par un robot de téléchargement de sites développé spécifiquement, est doté d'une série de modules déclenchés circonstanciellement en fonction des actions opératoires lancées par l'utilisateur à partir du navigateur (texte intégral, approximation lexicale, interrogation sur lemme, expansion de requêtes lemme $\Rightarrow$ lemmes et document $\Rightarrow$ documents, N-grammes dotés de fonctions de proximité lexicale, cartographie dynamique)

Dans le cadre de ce projet NeuroWeb, nous avons sélectionné les sites d'information devant alimenter le moteur en utilisant une méthode originale pour réduire l'espace de recherche sur le web en classifiant automatiquement les pages HTML. Nous proposons de prendre en considération les balisages utilisés dans les pages pour construire les profils des pages Web. Cette approche est fondée sur les caractéristiques de pages HTML. Cette catégorisation permet alors :

- d'améliorer les navigations en réduisant l'espace de recherche en montrant seulement les pages pertinentes par rapport aux souhaits de l'usager,
- d'éviter la situation de surcharge cognitive à laquelle l'usager est souvent confronté au fil de ses lectures,
- de signaler à l'usager les types de pages auxquels aboutit sa requête,

- de donner des possibilités à l'usager de filtrer et de choisir les types de pages qu'il désire consulter.

Utilisée par les agents arpenteurs afin de collecter des sites "homogènes", nous avons transposé cette méthode dans notre logiciel de construction d'espaces de connaissances virtuels HyWebMap.

2 Type d'informations sur le Web

Il existe plusieurs approches pour aider l'utilisateur à naviguer sur le Web mais aucune ne prend en considération la notion de profil syntaxique des documents. Pourtant ces profils permettent d'identifier les types de données qu'ils contiennent. Les balisages utilisés dans les documents écrits par exemple en HTML, fournissent explicitement ces types de données.

Nous proposons de prendre en considération les balises des pages pour construire les profils des pages Web. Pour établir une catégorisation de classification automatique des pages Web, nous nous sommes appuyés sur les travaux d'Alain Lelu en utilisant l'algorithme de K-means axiales [Lelu 99], [Balpe & al. 96].

Les documents sur le Web sont hétérogènes (sites commerciaux, pages personnelles, livres, articles, annuaires), ne possèdent aucune véritable structure. Les sources d'informations sont diverses, ainsi que leurs types. [Bélisle et al., 99] distinguent plusieurs grands types d'information :

· Information publique de référence, provenant des gouvernements, d'organismes professionnels, de bibliothèques, d'associations, ou de sociétés privées.

· Information scientifique et éducative (disciplinaire), dont les banques de données traditionnelles, provenant de laboratoires de recherche, d'universités, ou de sociétés de services.

· Information publicitaire, visée commerciale provenant des entreprises.

· Information médiatique, provenant des organismes des presses.

· Information personnelle, provenant des individus ayant leur propre site.

Cette distinction est floue car certains sites proposent plusieurs types d'informations. Les contenus des sites peuvent varier d'un site à un autre par rapport aux objectifs de chaque site. Nous distinguons trois catégories de sites Web par rapport à leurs contenus :

· Les sites textuels privilégient les contenus textuels avec plusieurs liens internes et des liens externes car leur objectif est de diffuser les informations auprès des utilisateurs (les sites institutionnels, bibliothèques, universitaires, entreprises). Dans ceux-ci, les images ou les illustrations offrent des informations complémentaires et n'interviennent le plus souvent qu'à un deuxième niveau de recherche.

· Les sites visuels : privilégient les contenus visuels (images, graphiques d'illustration, etc.). Ainsi, ils intègrent souvent des formulaires (champs de saisies), par exemple les sites commerciaux, publicitaires, commerces électroniques, musées. L'image joue un rôle important, elle participe à l'attractivité du site et pour les commerciaux, elle est une valeur ajoutée indispensable. Pour les sites "plus techniques", l'image a une fonction différente. Elle permet à l'utilisateur de mettre rapidement ses attentes en correspondance avec l'information présentée. Dans ces sites les textes offrent des informations complémentaires et n'interviennent qu'à un deuxième niveau de recherche.

· Les sites portails (annuaires) : privilégient plutôt les liens externes.

3 Les attentes des utilisateurs

Le Web est un service d'information et de communication d'un contenu selon certaines modalités. L'un et l'autre répondent à des besoins précis des utilisateurs, dans un contexte

donné. La qualité du service dépend de façon cruciale de l'identification correcte de ces besoins qui doivent demeurer centraux.

Les besoins changent par rapport aux objectifs de chaque utilisateur, certains souhaitent visiter des pages contenant seulement des images lorsqu'ils visitent un site de musée, ou un catalogue de produit, d'autres souhaitent visiter des pages textuelles lorsqu'ils visitent un site institutionnel, universitaire, ou d'autres souhaitent visiter des pages contenant textes et images. Cela peut se comprendre comme une demande de maîtrise de la recherche de contenus plus qu'une demande brute portant sur le simple accès à l'information.

3.1 Typologies des utilisateurs sur le Web

Après des données recueillies et analysées au centre Georgie Institut of technology les chercheurs Catledge et Pitlow [Catledge & Pitlow 95] proposent trois classes d'utilisateurs, cette analyse illustre la tension entre recherche d'information (query en anglais) et navigation :

· Les utilisateurs appelés "*searchers*" reprennent épisodiquement des séquences courtes mais s'engagent souvent dans des séquences longues.

· Les "*general purpose browser*" n'ont en moyenne qu'une probabilité sur 4 de répéter des séquences complexes (inertie moyenne des usagers).

· Les "*serendipitous browser*" évitent systématiquement de s'engager dans de longues séquences de navigation, ils visitent superficiellement les sites.

Cette analyse montre d'une part que l'utilisateur, explore une zone restreinte à l'intérieur d'un site visité et d'autre part que les utilisateurs ne traversent que rarement plus de deux couches d'hypertexte avant de retourner à leur point d'entrée.

Cette étude suggère que les pages personnelles sont utilisées préférentiellement comme relais dans la navigation et jouent ainsi un rôle d'index vers les autres sites. Ainsi cela illustre la nécessité de bien connaître et comprendre les stratégies navigationnelles des utilisateurs comme base pour la conception de nouveaux logiciels navigateurs.

4 Classification automatique : choix algorithmique

Historiquement, la classification de documents a été utilisée pour améliorer les performances des systèmes de recherche d'informations. L'hypothèse formulée par [Rijsbergen 79] est la suivante : "*...closely associated documents tend to be relevant to the same requests...*". Au lieu de comparer une requête à chaque document d'une base documentaire, le système ne compare la requête qu'avec le modèle représentant chacun des groupes de documents. Les algorithmes actuels autorisent la comparaison de la requête avec chaque document. Aujourd'hui, les méthodes de classification automatique sont utilisées dans plusieurs domaines pour visualiser un ensemble de documents retournés par un moteur de recherche Web.

Les deux méthodes les plus utilisées pour classer les documents sont :

1. **La classification hiérarchique ascendante** : elle possède deux propriétés intéressantes. Tout d'abord, l'utilisateur doit définir le nombre de groupes à obtenir. Ensuite, la méthode induit naturellement une hiérarchie entre les groupes de documents.

Cette propriété est intéressante à condition que la hiérarchie ne soit pas trop profonde. Une hiérarchie trop profonde nuit en effet à la recherche d'informations. [Maarek & Schaul 96] a proposé de simplifier la hiérarchie en la coupant arbitrairement à des seuils de similarité de 10%, assurant donc une profondeur maximale de 10. Une autre propriété importante concerne la représentation des groupes de documents [Cutting 92]. Les noms de groupe permettent en effet aux utilisateurs de décider quelle branche de classification explorer. Généralement, le nom d'un groupe reflète le contenu thématique des documents qu'il contient. L'approche

traditionnelle consiste à sélectionner quelques mots dont l'importance est calculée en combinant la fréquence des mots et le nombre de documents où ils apparaissent. Lorsque le texte intégral est utilisé, des étapes de filtrage sont nécessaires pour choisir un nombre réduit de termes et les présenter à l'utilisateur.

2. **La classification des K-means vers les K-means axiales** : Cette méthode de classification automatique est très ancienne. Connue sous d'autres noms (quantification vectorielle, méthode des centres mobiles, etc.), souvent réinventée et dotée de nombreuses variantes (méthode des nuées dynamiques, ISODATA, etc.), elle est de fonctionnement très intuitif :

· On choisit le nombre K de classes désirées.

· On sème au hasard K points représentatifs des futures classes dans l'espace où sont représentés les objets à classer (par exemple, on détermine au hasard K profils de fréquences de mots, s'il s'agit de classer des documents).

· On présente le premier objet à classer - géométriquement, c'est un point dans cet espace - et on détermine à partir de ses distances aux K centres de classes quel centre est le plus proche : comme la classe que représente ce centre ne contient aucun élément, ce centre est mis à la place du premier objet et symbolise la classe 1.

· On présente le deuxième objet, et on calcule encore ses distances aux K classes. Si la classe la plus proche est la classe 1, le point représentatif de celle-ci est déplacé, dans la direction de l'objet, de la moitié de la distance à celui-ci (la classe contient déjà un élément). Si la classe la plus proche n'est pas la classe 1, le deuxième objet devient le premier élément d'une nouvelle classe.

· On procède de la même façon pour tous les autres objets : si la classe la plus proche d'un objet en contient déjà n, elle est rapprochée en direction de celui-ci d'un n-ième de sa distance [Lelu 93].

En fin de compte, après épuisement des objets à classer, chaque centre de classe représente la position moyenne des points affectés à cette classe. D'où le nom de K-means, signifiant moyennes en anglais.

Alain Lelu [Lelu 93] propose une extension de cette méthode, appelé K-means axiales, le principe est très proche de celui des K-means. La différence réside dans le fait que les K-means axiales contraignent les objets et les centres de classes à se trouver sur une hypersphère de rayon 1, c'est-à-dire à être représentés par des vecteurs normalisés, de longueur 1. Tout se passe comme si on obtenait des axes de classes, et non des points représentatifs de ces classes, au moyen de corrections angulaires successives. En définitive, on obtient un axe par classe, sur lequel on projette les éléments de cette classe : les éléments les plus élevés, dont les projections sont les plus proches de la valeur 1, sont les plus centraux et typiques de la classe. Il est possible de projeter également les éléments qui n'en font pas partie : c'est de cette façon que cette méthode de classification stricte au départ peut nous fournir la représentation " floue" qu'illustre "l'allumage" nuancé des diverses classes par un objet présenté au système.

Autres différences avec les K-means originelles :

· Les K-means axiales représentent les K classes par leurs positions sur une carte globale. Celle-ci est obtenue par une analyse des données "au deuxième degré", par exemple par une analyse en composantes principales sur le nuage des K points représentatifs des classes.

· Des résultats plus stables et indépendants de l'ordre d'entrée des données sont obtenus par des variantes itératives, non adaptatives, des K-means. La méthode K-means axiales comporte en option une telle variante. Pour la description précise de cet algorithme, nous renvoyons le lecteur à [Balpe & al. 96, annexe 2].

Nous nous sommes arrêtés sur la méthode K-means axiales pour la catégorisation automatique des pages Web parce que :

· de par sa variante, elle présente l'avantage de cumuler - pour une fois - une exécution très rapide avec une occupation très faible d'espace mémoire, ce qui la rend apte à analyser nos colossales bases documentaires avec nos moyens de calcul actuels.

· elle fournit une représentation floue de classes (par exemple, une page Web peut-être à la fois, textuelle et image, etc.).

5 Objectifs

Nous avons défini comme objectif de :

· Permettre une catégorisation rapide d'un ensemble de documents Web issus des sites en ligne.

· Donner la possibilité à un utilisateur lors de la consultation d'un document Web de connaître sa catégorie (texte, graphique, navigation, etc.).

· Mettre en place des outils graphiques de consultation et de navigation permettant d'exploiter cette catégorisation.

5.1 Les étapes méthodologiques

Notre démarche consiste à retenir préalablement des sources d'informations (sites ou pages Web). Nous constituons alors un échantillon de documents représentatifs (articles de journaux, publications, documents techniques, documents amateurs, etc.).

La présence d'un nombre important d'outils de recherche francophones (annuaires et moteurs de recherche) sur Internet rend la tâche plus délicate dans la mesure où les uns et les autres ne proposent pas les mêmes fonctionnalités. Notre choix s'est porté sur l'annuaire Lokace (désormais Nomade), un choix basé sur les données statistiques annoncées par ce dernier.

5.1.1 Recherche et sélection des sources

Le guide francophone Lokace/Nomade proposait 12 catégories générales (principales) au départ, elles-même décomposées en sous-catégories. Ces sous-catégories sont découpées en un nombre variable de sous-catégories. Le chiffre annoncé était de 3000. Le nombre des sites pointés dépasse 20000, couvrant les grands domaines.

L'utilisation de la méthode de classification K-means axiales avec le choix de K classes désirées, nécessite l'utilisation d'une quantité de données très importante pour obtenir un résultat "interprétable". A partir de cette collecte, nous nous sommes fixé de réunir 20000 profils de pages sur l'ensemble des catégories du guide (3000). Pour arriver à ce résultat, nous nous sommes basés sur les données annoncées (nombre de catégories et nombre de sites du guide). Une moyenne de 6 sites par catégorie (20000 sites/3000 catégories), cette moyenne a été vérifiée sur l'une des catégories.

Le choix aléatoire d'une catégorie sur six nous donne environ 500 catégories. En prenant aléatoirement un site sur six par catégorie, on aura environ 500 sites. Il faut prendre en compte un certain nombre d'URLs obsolètes. Le nombre moyen de pages par site est de 70.

5.1.2 Le recueil des données

Les premières données à recueillir sont les URLs des pages HTML qui vont constituer l'échantillonnage documentaire. Cette opération consiste à effectuer un sondage de l'annuaire retenu en appliquant la stratégie de sélection expliquée auparavant. Nous avons utilisé un

agent Web pour explorer la structure catégorique de l'annuaire Lokace/Nomade. Les points de départ de cette exploration sont les URLs des 12 catégories principales de l'annuaire. L'agent Web devait sonder la totalité de la structure arborescente.

6 Profil de documents Web

HTML définit un ensemble de balises de base. On cite les balises de structure, puis celles qui permettent d'agencer et de composer du texte. L'autre catégorie de balises est celle qui permet de mettre en place des hyperliens. Une page Web peut être définie par un ensemble de caractéristiques (domaine du site, structure (frames, etc.), liens internes, liens externes, quantité et poids des images intégrées, rapport balise/contenu, …)

On part de l'idée qu'une page HTML peut être intéressante par sa forme descriptive et par son aspect. Celle-ci est intéressante si elle contient des liens vers le site lui même, des liens externes vers d'autres serveurs. Une page Web peut contenir des formulaires ce qui permet de comprendre qu'il s'agit d'une interface de saisie.

Il est aussi important de signaler que le poids d'une page est un élément très significatif car il peut permettre de déduire l'importance du contenu de la page quantitativement. La présence d'images dans une page est un élément qui permet aussi de dégager une idée sur la dimension esthétique de la page.

Une page HTML peut être considérée du point de vue de son contenu réel (contenu et balises HTML) ou de son rendu-écran (dans un navigateur).

En partant des caractéristiques citées auparavant et en observant une page Web sous ces deux angles, Il est possible d'établir le profil d'une page HTML en constituant un vecteur d'informations.

Le profil est construit par une analyse et un traitement statistique de balises HTML. Nous avons sélectionné les données les plus significatives (cf. tableau 6.1.) obtenues à partir de notre échantillon documentaire initial.

Liens locaux	Liens internes	Liens externes	Taille Hors balises	images	tableaux	formul.	mail
0	12	5	24096	7	1	1	0

Tableau 6.1. Exemple de vecteur de données (profil de page Web)

7 Processus de catégorisation

7.1 Catégorisation de base

Les indicateurs quantitatifs recueillis par l'agent Web sont stockés sous forme de matrice (la relation Profil) (cf. tableau 6.1.). Le processus débute par une catégorisation de l'échantillon de documents collectés auparavant, cela passe d'abord par l'analyse et l'extraction des profils des documents. Les profils sont stockés sous forme de matrice, chaque ligne correspond à un document et chaque colonne correspond à l'un des attributs cités précédemment.

La méthode des classification K-means axiales nécessite le choix de K classes de sorties, pour cela nous avons effectué une catégorisation avec 10 classes. Les résultats obtenus étaient difficilement interprétables. Cette difficulté résulte d'un éparpillement sur plusieurs classes dont certaines sont très proches. L'autre choix consiste à donner 5 classes. Ce choix se justifie par les résultats de la catégorisation manuelle [Borzic 98] sur un échantillon d'une centaine de documents.

Nous avons validé les résultats obtenus lors de cette deuxième catégorisation par une vérification manuelle sur le Web. Une centaine de documents apparus au début, au milieu et à la fin de chaque classe ont été vérifiés, la quantité importante des documents ne permettant pas une vérification complète. Les données statistiques obtenues avancent un pourcentage de 72% de classification pertinente.

L'algorithme des K-means axiales [Lelu 93] permet une projection des éléments d'une classe sur un axe, les éléments les plus élevés, dont les projections sont les plus proches de la valeur 1, sont les plus centraux et typiques de la classe. Concrètement, on obtient une matrice de 5 colonnes, chacune de ces colonnes correspond à une classe. L'identification des classes (texte, graphique, etc.) se fait par l'étude des premières valeurs et les documents qui leurs sont associés.

La figure 5.1 présente le fonctionnement de l'agent Web et du module de classification. Cinq types de pages ont été ainsi distingués automatiquement, et leur degré de typicité visualisé par une échelle à trois degrés(*, **, ***). En effet, ces cinq catégories constituent des pôles flous, plus que des classes bien distinctes :

· **Page informative textuelle** (représenté par la lettre T) : Le contenu de la page est un texte.

· **Page informative avec texte illustré** (représenté par les lettres I) : Le contenu de la page est une illustration visuelle, ce peut être des images, des figures, des boutons, etc.

· **Page carrefour interne au site** (représenté par les lettres CI) : le contenu de la page est un ensemble de liens internes au site.

· **Page carrefour externe au site** (représenté par les lettres CE) : le contenu de la page est un ensemble des liens externes au site.

· **Page interface à la saisie** (représenté par la lettre S) : le contenu de la page est un ensemble de champs de saisie.

On peut constater que cette classification est floue car on peut avoir une page entrant dans deux ou trois catégories. Par exemple une page peut être Page informative textuelle alors elle est représentée par l'échelle T*, ou bien une Page informative textuelle avec des liens externes, alors elle est représentée par TCI*, etc.

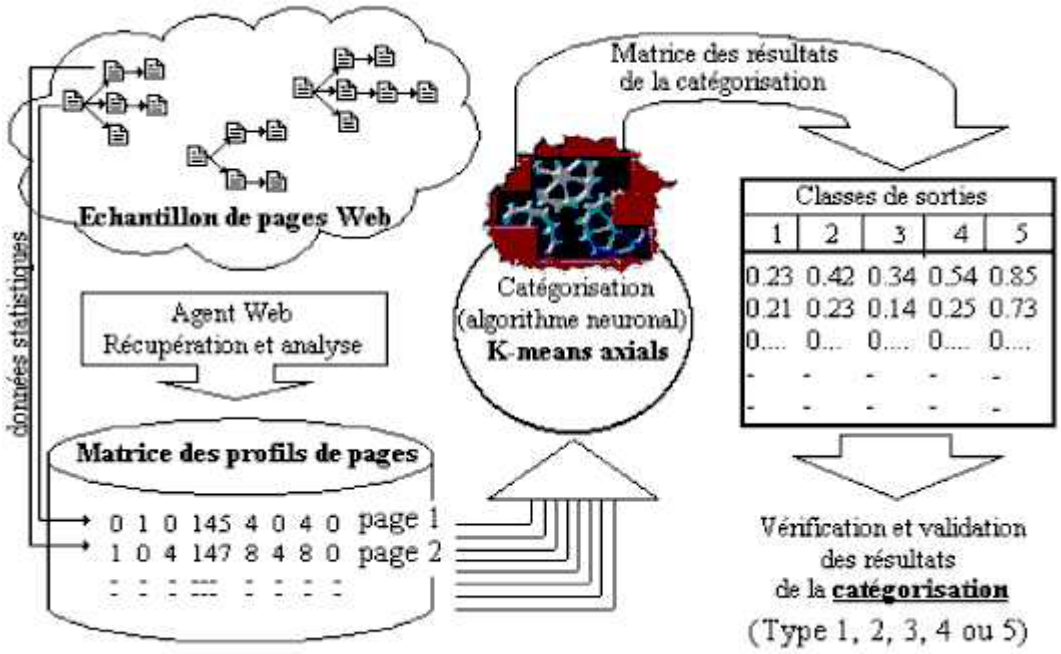


Figure 5.1 : Catégorisation automatique d'un échantillon de pages Web

8 Usage et exploitation des résultats

8.1 Comme module HyWebMap

La catégorisation automatique des pages Web au niveau du système HyWebMap permet à l'utilisateur de filtrer les pages Web référencées par les différents nœuds d'un réseau HyWebMap. Cela lui permettra de naviguer par catégorie donc par type de pages. Il est possible d'indiquer la ou les classe(s) à visualiser.

8.2 Au sein d'un moteur de recherche

En tant que filtre des pages HTML entre l'utilisateur et les moteurs de recherches sur le Web : la figure 5.4. présente l'interface d'un prototype de méta-chercheur, l'utilisateur choisit un ou plusieurs mots-clés, ainsi que la catégorie de pages qu'il souhaite visiter.

Les agents Web sollicitent le moteur de recherche Altavista , et les résultats obtenus suite à la requête sont analysés et un profil de chaque document est construit.. Le module de catégorisation utilise une matrice intermédiaire pour insérer le nouveau profil dans la catégorisation de base validée. De cette manière, il est plus facile de dresser une typologie d'un document Web sans le visualiser.

Dans l'exemple de la figure 5.2, l'utilisateur ne souhaite visiter que les pages de catégorie texte, il sélectionne alors la catégorie Texte, où se trouve les mots-clés "guide Internet". Après récupération, analyse, débalisage et comparaison, les agents affichent les résultats avec un degré d'échelle : T* (une page contient du texte), T** (une page moyennement textuelle), T***(une page faiblement textuelle). En tant que module de filtrage pour un moteur de recherche : cette classification peut être exploitée dans le cadre d'un moteur de recherche pour éviter l'indexation de pages qui contiennent peu de texte et qui sont donc peu significatives.

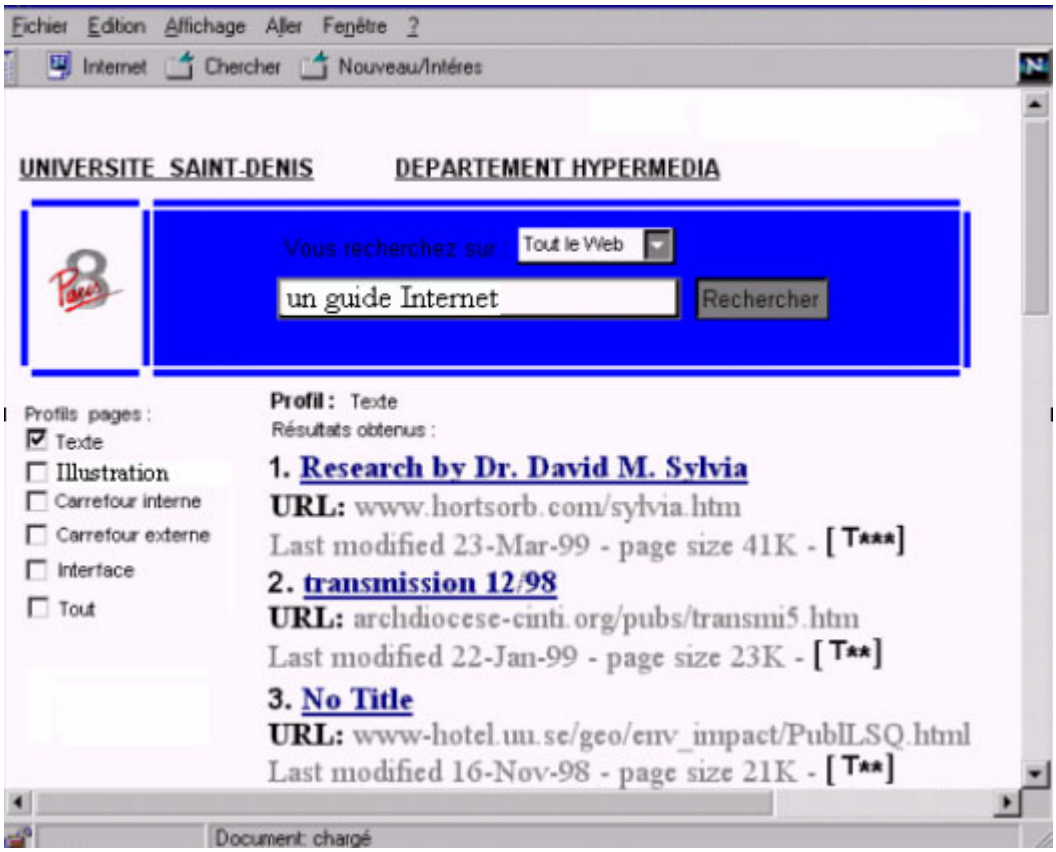


Figure 5.23 : Utilisation du module de catégorisation au sein d'un moteur de recherche

9 Conclusion

Les moteurs de recherches, les agents intelligents, les portails, les solutions d'E-learning, etc…au-delà de leur spécificités fonctionnelles, témoignent de la présence d'informations conçues et élaborées pour le Web. Ce sont désormais des informations ciblées, que ces dispositifs grâce à leurs puissants mécanismes de filtrage transmettent aux internautes. Ceux-ci ne sont plus confrontés à la recherche de «LA» bonne information mais à la nécessité de traiter les multiples bonnes réponses qui leur sont envoyées. Compiler les URL, les colationner, les ré-organiser fréquemment, les compléter le cas échéant représente un mode de fonctionnement (presque) banalisé chez les internautes les plus concernés par l'exploration du Web. La classification automatique de pages HTML que nous avons présentée, s'insère dans cette logique d'assistance à la recherche et la navigation dans l'espace virtuel du Web. Cette technique algorithmique est opérationnelle sous la forme d'un module utilisable au sein d'un environnement comme le système HyWebMap ou comme un module spécifique de moteur de recherche pour le filtrage de document à destination d'environnement d'indexation ou comme indicateur informatif sur la typologie de documents Web. Il reste à étudier l'amélioration de l'ergonomie des interfaces d'affichage : les techniques 3D peuvent être utiles pour présenter les résultats et s'y mouvoir.

Références

[Balpe et al. 96] Balpe J.P., Lelu A., Saleh I., Papy F., "Techniques avancées pour l'hypertexte", Editions Hermès, 1996.

[Bélisle et al. 99] Bélisle C., Zeiliger R., Cerratto T., "S'orienter sur le Web en construisant des cartes interactives : le navigateur NESTOR", in Hypertextes hypermedias et Internet H2PTM'99 Balpe, Natkin, Lelu, Saleh, Hermes Science Publications, Paris, pp. 101-117, 1999.

[Borzic 98] Borzic B., "Un modèle de gestionnaire itératif de flux informationnel sur Internet", Thèse de doctorat, Information Scientifique et Technique, CNAM, Paris, mars 1998.

[Catledge 95] Catledge L.D., Pikow J.E., "Characterizing browsing startegies in the World Wide Web", Proc. of the 3th International Conference on the World Wide Web, Darmstadt, Germany, 1995.

[Cutting 92] Cutting D.R. et al., "Scatter/Gather : A cluster-based approch to browsing large document collections", Actes de la 15th Conférence ACM/SIGIR Research and developpement in information retrievel, Danemark, 1992.

[Lelu & al 99] Lelu A., Hallab M., Rhissassi H., Papy F., Bouyahi S., Bouhaï N., He H., Qi C., Saleh I., "Projet NeuroWeb : un moteur de recherche multilingue et cartographique", 5e conf. Int. H2PTM'99, Paris, France, septembre 1999.

[Lelu 93] Lelu A., "Modèles neuronaux pour l'analyse de données documentaires et textuelles", Doctorat de l'Université Paris VI, mars 1993.

[Maarek & Schaul 96] Maarek Y.S., Schaul I., "Automatically organizing bookmarks per content", 5th International World Wide Web Conference, Paris, 1996.

[Rijsbergen 79] Rijsbergen C.J., "Information retrievel", Butterwords, London, 1979.

Construction de Classes de Documents Web

Benjamin Nguyen

INRIA FUTURS
4 Rue J. Monod
ZAC des Vignes
91893 Orsay CEDEX
FRANCE
Email : Benjamin.Nguyen@inria.fr
Tél : 33 1 72 92 59 31 Fax :

Iraklis Varlamis, Maria Halkidi, Michalis Vazirgiannis

Department of Informatics, Athens University of Economics and Buisiness
76, Patision Street, Athens 10434
GREECE
Email : {varlamis,mhalk,mvazirg}@aueb.gr
Tél : (301) 8203 519 Fax :

Résumé

Dans cet article, nous proposons une nouvelle mesure de similarité entre des documents caractérisés par un ensemble succint de concepts d'une hiérarchie (ontologie), ce qui est le cas de documents web. Cette mesure permet d'exploiter un algorithme de regroupement par densité (DB-SCAN), issu du domaine des bases de données spatiales, que nous avons adapté à nos besoins, afin de permettre la construction de classes de documents. Nous donnons également des résultats expérimentaux qui montrent la pertinence de notre approche.

Abstract

In this paper, we introduce a novel similarity measure between documents, such as web documents, that can be characterized by a small set of concepts, that each belong to a hierarchy (ontology). We use this measure with a slightly modified DB-SCAN clustering algorithm in order to construct clusters of semantically related documents. We also give experimental results that illustrate the quality of our approach.

1 Introduction

En recherche d'information, on caractérise traditionnellement un document par son contenu. [PW00] montrent que dans le cas de pages web il est en fait possible de construire des ensembles succints de l'ordre de 5 à 10 mots, qui caractérisent de manière pertinente chaque page. Nous avons dans nos travaux précédents étendu cette approche pour caractériser chaque document web par un ensemble de concepts d'une ontologie. Pour nous, une ontologie est simplement un arbre (IS-A), mais les résultats de cet article s'appliquent tel quels à un arbre avec d'autres relations entre les termes, et peuvent se généraliser avec des modifications minimes dans le cas où l'ontologie serait un treillis ou un graphe acyclique orienté. Nous supposons **donnés** les ensembles de mots-clés et concepts de l'ontologie qui caractérisent chaque document, et ne discuterons pas ici de la manière dont ces ensembles sont générés. Pour plus d'informations sur la manière de construire ces ensembles, le lecteur pourra se reporter à [VNVA02, HNVV02], qui détaillent également l'architecture globale de notre projet, appellé THESUS. L'objectif de cet article est de tenter de trouver une méthode efficace pour résoudre le problème suivant :

> Étant donné un ensemble de documents, dont chacun est caractérisé par un (petit) ensemble de concepts d'une ontologie, trouver une méthode pour regrouper en classes les documents qui ont une sémantique proche.

La réponse classique à un problème de ce type serait d'utiliser des techniques d'IR (Information Retrieval), comme celles décrites dans [SM83]. Toutefois, ces techniques présentent l'inconvénient de se baser sur un appariement **exact** des mot clés. En effet, ces méthodes ne permettent pas de prendre en compte le fait que certains mots peuvent avoir une *proximité sémantique* entre eux. Par exemple, imaginons un document qui serait caractérisé par les mots "serpent" et "désert" et un autre document caractérisé par les mots "aspic" et "Sahara". En utilisant les techniques traditionnelles, ces documents ne seraient en aucun cas considérés comme étant proches. Toutefois, en utilisant une ontologie pour mieux *comprendre* ce que signifie chaque terme, nous pouvons proposer une mesure de similarité bien plus pertinente, et par conséquent l'appliquer à un algorithme de regroupement par densité de manière efficace.

Plan de l'article : Dans la section suivante, nous présentons un court état de l'art, centré sur les mesures de similarité. Nous présentons notre mesure dans la section 3, et l'algorithme de regroupement dans la section 4, puis dans la section 5 nous donnons des résultats expérimentaux, qui témoignent de la pertinence de notre approche.

2 État de l'art

Notre but ici n'est pas de dresser un état de l'art complet sur le domaine de la classification de documents. De nombreux travaux ont déjà été réalisés, nous citerons notamment [Fis87, AGY99]. Le problème que nous considérons est un peu plus spécifique, puisque nous nous intéressons à des documents Web [ZE98], et nous prenons en compte les liens vers ces documents, comme le détaille [Kle99]. En effet, les liens simplifient entre autres la manière de caractériser une page par un nombre restreint de mots-clés, et ce sont notamment sur ces résultats que nous nous basons pour affirmer qu'il est possible de construire un ensemble concis de termes qui vont correctement caractériser un document web. Nous détaillons dans nos travaux précédents [HNVV02] comment en utilisant WordNet [Wor] nous sommes en mesure de construire à partir d'un ensemble de

mots clés caractérisant une page, cet ensemble de concepts (5 à 10 environ) d'une ontologie d'un domaine particulier. Les détails complets de notre système sont donnés dans [HNVV02]. D'autres systèmes de meta-moteurs de recherche s'intéressent aux liens : Kartoo, [kar] représente graphiquement les liens entre sites, Vivissimo, [Viv] construit des classes de documents en temps réel à partir du résultat de requêtes sur des moteurs de recherche classiques.

Les travaux les plus proches des notres sont ceux de [DJ01] qui ont construit une distance sur une ontologie afin de pouvoir indexer des sites web, mais nous n'avons trouvé aucun travail qui porte précisément sur le thème des mesures de similarité entre ensembles d'éléments d'une hiérarchie. Nous détaillons dans les paragraphes suivants les travaux sur les mesures de similarité qui présentent néanmoins une certaine pertinence.

2.1 Notions générales sur les mesures de similarité

Notre but est de construire une mesure de similarité entre documents, ce qui se réduit dans notre cas à trouver une mesure de similarité entre ensembles d'éléments appartenant à une hiérarchie. Nous partageons les mêmes intuitions que [Lin98] en ce qui concerne les propriétés que devrait posséder une telle mesure.

Pour deux ensembles A et B :

- La similarité entre A et B est fonction de ce qu'ils ont en commun. Plus ils ont d'éléments en commun, plus leur similarité sera élevée.
- La similarité entre A et B est fonction de leurs différences. Plus ils ont de différences, plus leur similarité sera faible.
- La valeur de similarité maximale est obtenue lorsque A et B sont identiques, quelque soit le nombre d'éléments qu'ils ont en commun.

2.2 Mesures de similarité entre ensembles

Il existe un certain nombre de mesures de similarité entre ensembles. La plus utilisée, le coefficient de Jaccard est très simple. Soient A et B deux ensembles finis d'éléments. La similarité entre A et B se définit comme : $S_J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. Cette mesure respecte bien les intuitions du paragraphe précédent, mais ne prend en compte que l'appariement exact de termes ; or nous voulons aller plus loin en introduisant une mesure qui prenne en compte la proximité des mots, plutôt qu'une comparaison binaire. [EM97] font une revue de diverses mesures de similarité entre deux ensembles finis de points dans un espace métrique et montrent que toutes ces mesures ont une complexité polynômiale en fonction des éléments de l'ensemble. [GHOS96] s'intéressent également à une dizaine de mesures de similarités entre ensembles, et expliquent leur sémantique, mais toutes sont basées sur le coefficient de Jaccard. Les problèmes d'indexation entre ensembles est également traité dans les travaux de [GGK01], où les auteurs montrent comment grouper des ensembles d'éléments, mais toujours en utilisant le coefficient de Jaccard. Pour finir, notons que la mesure *cosinus* traditionnelle [SM83] de l'IR est aussi une application directe du coefficient de Jaccard. Dans tous ces travaux, l'ensemble de valeurs sur lesquelles la mesure de similarité est calculée est un ensemble **plat**, c'est-à-dire que tous les éléments de l'ensemble sont indépendants les uns des autres, et ceci nous a encouragés à nous efforcer à prendre en compte les relations sémantiques qui pouvaient exister entre les concepts d'une ontologie, tout comme [BFS02] qui s'intéressent dans leur approche médiateur à la similarité entre concepts appartenant à la même hiérarchie.

2.3 Similarité entre deux éléments d'une ontologie

[RSM, Res95, Res99] proposent diverses méthodes pour calculer la similarité entre deux concepts d'une ontologie, par exemple WordNet, et [Lin98] a effectué une comparaison entre ces méthodes. Il en ressort que la mesure de Wu et Palmer [WP94] est la plus rapide à calculer, tout en restant aussi expressive que les autres, d'où notre choix de cette mesure comme fondement de nos travaux. Sa définition est la suivante :

Étant donné un arbre, et deux noeuds a et b de cet arbre, soit c l'ancêtre commun le plus profond (sachant que la racine est de profondeur 1). La mesure de similarité entre a et b s'exprime alors : $S_{WP}(a,b) = \frac{2 \times Profondeur(c)}{Profondeur(a)+Profondeur(b)}$

Il est immédiat de vérifier que cette mesure respecte les critères de 2.1. Il est possible de définir la distance canonique associée à cette mesure de similarité de la manière suivante : $D_{WP}(a,b) = 1 - S_{WP}(a,b)$. Il est possible de vérifier que D_{WP} est bien une distance, mais nous ne le détaillons pas ici.

3 Une nouvelle mesure de similarité entre ensembles de concepts

Nous avons vu dans la section précédente que les mesures de similarité existantes ne prennent en compte qu'un appariement exact de termes. Dans cette section, nous montrons comment étendre les idées de Wu et Palmer pour proposer une mesure de similarité sur des ensembles de termes d'une ontologie. D'après les études de [EM97, Nii87], aucune mesure n'a été proposée sur le calcul de la similarité entre des ensembles d'éléments d'un espace métrique. Notre cas est même un peu plus général, puisqu'il ne s'agit pas d'un espace métrique, mais simplement d'un espace sur lequel il existe une mesure de similarité.

L'idée fondamentale réside dans l'utilisation d'une mesure de similarité entre les éléments individuels de chaque ensemble, pour estimer de manière adéquate la similarité entre les ensembles eux mêmes. Il est important de souligner que si nous utilisons la mesure de Wu et Palmer comme mesure de base, il est possible d'en utiliser une autre, à condition que la mesure ainsi définie soit une généralisation de la mesure de base.

3.1 Wu et Palmer généralisé aux ensembles

Dans les paragraphes qui suivent, nous allons définir formellement la généralisation de la mesure de Wu et Palmer que nous proposons. Nous commençons par donner les intuitions de cette mesure, qui doit respecter les idées générales sur la similarité présentées en 2.1.

Intuition : Chaque document est représenté par un ensemble de concepts de l'ontologie. Si les ensembles contiennent beaucoup de termes semblables, voire identiques, alors ces ensembles seront très similaires. Au contraire, si ils possèdent beaucoup de termes très différents (très distants dans l'ontologie), leur similarité devra être faible. Dans un premier temps, on cherche pour chaque concept du premier ensemble A, de quel concept du second ensemble B il est le plus proche. Ainsi, même si l'ensemble A est composé de concepts très différents, tant que pour chaque concept de A on trouve un concept de B qui est proche, la similarité va être grande. Dans un deuxième temps, on effectue la même opération pour B, ce qui permet de voir s'il n'y a pas dans B des concepts qui sont très différents de ceux de A, et qui donc feront chuter la similarité, ou bien si tous les concepts de B trouvent un bon appariement avec ceux de A.

Formalisme : Soit Ω une ontologie (dans les exemples qui suivent, nous utilisons WordNet). Ω est un ensemble fini, dont chacun des éléments est un concept. On note en minuscules (a,b) les éléments de Ω (concepts). On note en majuscules (A,B) les sous ensembles de Ω. On note $|A|$ le cardinal de l'ensemble A. Soit $S_{WP}(a,b)$ la mesure de similarité de Wu et Palmer entre deux concepts a et b de Ω. On définit :

$$\zeta(A,B) = \frac{1}{2}(\frac{1}{|A|}\sum_{a\in A}\max_{b\in B}(S_{WP}(a,b)) + \frac{1}{|B|}\sum_{b\in B}\max_{a\in A}(S_{WP}(a,b))) \quad (1)$$

Vérifions les propriétés de $\zeta(A,B)$.

Propriétés : Il est immédiat de vérifier que $\zeta(A,B)$ est une mesure de similarité : i. $\zeta(A,B) = 1$ ssi $A = B$, ii. par construction, $\zeta(A,B) = \zeta(B,A)$. De plus, cette mesure de similarité étend bien la mesure de Wu et Palmer. Si les ensembles sont réduits à un élément, par exemple $|A| = \{a\}$ et $|B| = \{b\}$, alors $|A| = |B| = 1$ et $\zeta(A,B) = S_{WP}(a,b)$. Si on considère que la complexité pour calculer $S_{WP}(a,b) = O(1)$, la complexité de cette mesure est égale à $O(|A| \times |B|)$. En réalité, le coût pour calculer S_{WP} dépend de l'ontologie, et dans le pire des cas est égale à $O(h)$ où h représente la profondeur maximale de l'ontologie. Heureusement, en règle générale, les ontologies sont assez peu profondes, mais très larges. Toutefois, dans tous les cas la complexité S_{WP} est indépendante du nombre d'éléments dans A ou B, et le temps de son calcul peut être sorti comme une constante. Pour des ontologies de taille raisonnable (< 5000 concepts) on peut tout simplement pré-calculer toutes les similarités entre chaque paire de concepts. Notre mesure est donc tout à fait performante, vis-à-vis des mesures proposées dans [EM97] où les meilleures complexités théoriques sont polynômiales dans les nombres d'éléments des deux ensembles.

Exemple : Calculons la similarité entre les ensembles suivants : $A = \{chat, CD\}$ et $B = \{felin, disque, moto\}$. Les similarités de Wu et Palmer sont les suivantes : $S_{WP}(chat, felin) = 0.95$; $S_{WP}(chat, moto) = 0.15$; $S_{WP}(chat, disque) = 0.19$; $S_{WP}(CD, felin) = 0.13$; $S_{WP}(CD, disque) = 0.83$; $S_{WP}(CD, moto) = 0.28$. Les éléments les plus proches de ceux de A sont $felin$ pour $chat$ et $disque$ pour CD. Réciproquement, les éléments les plus proches de B sont $chat$ pour $felin$, CD pour $disque$ et CD pour $moto$. Notons ici qu'un document qui parlerait juste de $felin$ et $disque$ serait plus proche du document A, puisqu'ici le fait de parler de $moto$ est bien entendu considéré comme moins pertinent.

On a donc :

$$\zeta(A,B) = \frac{1}{2} \times (\frac{1}{|A|} \times (S_{WP}(chat, felin) + S_{WP}(CD, disque)) +$$
$$\frac{1}{|B|} \times (S_{WP}(felin, chat) + S_{WP}(disque, CD) + S_{WP}(moto, CD)))$$

L'application numérique donne dans ce cas $\zeta(A,B) = 0.78$. Nous donnons une petite table avec quelques exemples. Il est à noter que bien entendu ces valeurs sont modulées selon l'ontologie que l'on utilise. Dans la table 1, nous rappelons que nous avons utilisé WordNet pour calculer les similarités, et non une ontologie d'un domaine bien particulier.

4 Application à un Algorithme de Regroupement par Densité

Dans cette section, nous expliquons comment regrouper dans des classes des documents caractérisés chacun par un ensemble de concepts issus d'une même ontologie, en fonction de

A	B	$\zeta(A,B)$
Chat, CD	Chat, CD	1.0
Chat, Tigre, Lynx	Félin	0.95
Chat, CD	Tigre, Disque, Félin	0.89
Chat, CD	Moto, Disque, Félin	0.78
Microprocesseur	Technologie, Electronique	0.70
Cléopâtre	Reine, Egypte	0.60
Espion, Microfilm	Révolution Française, Guillotine	0.11

TAB. 1 – *Similarité entre Ensembles*

leur proximité. Nous montrons que le problème peut se rapprocher, grâce à la définition d'une mesure de similarité, du cas classique de regroupement dans un espace métrique, traité par l'algorithme d'Ester, Kriegel *et al.* DB-SCAN [EKSX96]. Nous donnons une explication intuitive de l'algorithme, mais nous n'avons malheureusement pas la place d'expliciter ici l'algorithme complet.

4.1 Explication Intuitive de l'algorithme

Pour faire fonctionner cet algorithme, il est nécessaire de pouvoir évaluer la similarité entre deux documents. Nous avons montré dans la section précédente comment construire une mesure de similarité entre deux documents caractérisés par des ensembles de termes d'une ontologie. On définit deux seuils, $MinSim \in [0;1]$ et $MinDocs > 1$, l'un mesurant la similarité minimale entre documents pour qu'ils soient considérés comme voisins, et l'autre mesurant le nombre minimal de documents par classe. L'algorithme commence par prendre un document au hasard et à regarder combien de documents sont 'proches' de lui, c'est-à-dire combien de documents ont une similarité supérieure à $MinSim$ avec lui. Si plus de $MinDocs$ documents sont proches de lui, alors l'algorithme groupe tous ces documents ensemble, marque le premier document, puis passe à un autre document de ce groupe et effectue la même recherche, en aggrégeant les résultats à la classe qui est en train d'être construite. Lorsqu'il est impossible de progresser (lorsqu'il n'y a plus dans la classe traitée de documents similaires en nombre supérieur à $MinDocs$) la classe est considérée comme achevée, et l'algorithme continue en prenant un autre document (au hasard) qui ne fait pas partie d'une classe et qui n'a pas déjà été traité. L'algorithme se termine lorsque tous les documents ont été traités. Le nom 'regroupement par densité' vient du fait que les classes représentent des zones où la densité de documents est suffisante.

4.2 Modifications de l'algorithme

Un contexte très différent : À l'origine, l'algorithme a été utilisé dans le cadre des bases de données spatiales, où les éléments à regrouper sont des points d'un espace métrique. La mesure utilisée pour évaluer la distance est la mesure Euclidienne. Pour calculer les points qui sont dans le voisinage les uns des autres, les points sont tous stockés dans un R*-Tree [BKSS90]. Il est intéressant de remarquer que le temps de calcul du R*-Tree n'est pas pris en compte dans la complexité de DB-SCAN par ses auteurs. Or, dans notre cas, nous ne sommes pas dans un espace métrique dont on connaîtrait la dimension. Nous ne pouvons donc pas utiliser de R*-Tree. A la place nous pré-calculons la similarité entre les n documents à classer et sauvegardons ces

valeurs dans n listes de longueur n que nous trions en utilisant QuickSort. La complexité pour trier une liste de longueur n étant $O(n \log n)$, la complexité pour trier n listes est $O(n^2 \log n)$. Une fois cette phase préliminaire achevée, on peut faire tourner l'algorithme en remplaçant le R*-Tree par notre liste triée.

Complexité : La complexité est fonction de la complexité moyenne pour définir des documents voisins d'un autre. Dans notre système, vu que la similarité d'un document avec tous les autres est pré-calculée et présentée dans une liste ordonnée, le temps moyen de parcours pour trouver tous les documents qui ont une similarité $\zeta > MinSim$ est $O(\log n)$ en utilisant une méthode dichotomique, si les similarités entre documents sont stockées en mémoire vive, c'est-à-dire la même complexité que pour un R*-Tree. L'algorithme répète ensuite ceci pour tous les documents, ce qui donne une complexité finale en $O(n \log n)$ pour la seule phase de clustering.

Etiquettage : Bien que grouper des documents entre eux soit une tâche qui contribue à simplifier la recherche en leur sein, il est encore plus intéressant de mettre une étiquette sur chaque classe. Une méthode simple et qui donne néanmoins de bons résultats est la suivante :

- Pour chaque classe, nous construisons U l'union de tous les concepts qui appartiennent au moins à un document de cette classe.
- Pour chaque concept $k_i \in U$ nous calculons la proportion de documents de la classe considérée auxquels il appartient.
- Nous considérons comme pertinent les concepts qui apparaissent dans une proportion supérieure à un certain seuil. Dans nos expériences, nous avons fixé arbitrairement ce seuil à 51%, il conviendrait de faire plus d'expériences pour mesurer l'impact de cette valeur.

5 Expérimentation

Dans cette section, nous présentons les résultats fournis par le système, en utilisant une méthode de test 'aveugle'. Le paradigme ici est que si les résultats de la classification sont bons, alors cela signifie que la mesure de similarité est correcte.

5.1 Similarité

Protocole : Le protocole expérimental est le suivant : parmi tous nos documents (environ 40 000) nous avons sélectionné 20 paires de documents qui pouvaient ou non faire partie de la même classe, et nous avons présenté cette liste à 10 testeurs. Chaque testeur devait dire s'il grouperait ensemble les deux documents ou non, en leur attribuant une valeur allant de 1(pas de rapport) à 5 (quasi identiques). Nous avons calculé pour chaque paire la valeur moyenne de similarité 'humaine' en faisant la moyenne des scores donnés par les différents testeurs. Par exemple, si tous les testeurs donnent une valeur de 5 à la paire, alors la similarité 'humaine' est de 1. Un testeur qui donnerait 4 et tous les autres qui donneraient 5 ferait une similarité 'humaine' de 0.98. Cette valeur est attribuée à chaque document. Nous fixons ensuite le seuil $MinSim$. Si la similarité humaine pour une paire de documents est supérieure à ce seuil, la paire est admise comme 'pertinente'. Si la valeur est inférieure à ce seuil, cela signifie que les êtres humains ne jugent pas ces documents sémantiquement proches. Nous avons ensuite comparé les résultats avec notre système en jouant sur la valeur $MinSim$ de l'algorithme de clustering. Pour notre système, THESUS, les paires de documents pertinentes sont celles où les documents ont été affectés à la même classe, avec le même paramètre $MinSim$. Les résultats montrent le taux de

MinSim = 0.80
Taux de rappel Humain / THESUS : 77%
Taux de précision Humain / THESUS : 80%
Corrélation Humain / Humain : 81%

TAB. 2 – *Résultats pour THESUS (MinSim = 0.80)*

MinSim = 0.90
Taux de rappel Humain / THESUS : 83%
Taux de précision Humain / THESUS : 80%
Corrélation Humain / Humain : 81%

TAB. 3 – *Résultats pour THESUS (MinSim = 0.90)*

rappel et de précision[1] entre les résultats humains, et ceux du système. La corrélation entre les humains eux-mêmes est obtenue en calculant la dispersion des résultats entre testeurs, et peut être comparée aux deux taux précédents.

Résultats : Les résultats dans les tables 2 et 3 montrent que le système est tout aussi fiable qu'un testeur. Ceci montre que la mesure de similarité et l'algorithme de clustering donnent des résultats qui ont du sens. Nous avons effectué la même manipulation en utilisant le système Vivissimo, et donnons les résultats dans la table 4. Pour que la comparaison soit possible, nous avons considéré que Vivissimo donnait comme classés ensemble des documents qu'il plaçait dans un même chemin dans son arbre de classification. Nous voyons que sur ce jeu de documents différents, la corrélation entre testeurs humains est également aux alentours de 80%. Les résultats donnés par Vivissimo sont en revanche bien moins bons que ceux de notre système, il suffit de comparer les taux de rappel et de précision de la Table 4.

Et Jaccard ? Nous ne donnons pas ici de table comparative avec les coefficient de Jaccard car parmi les paires de documents choisies, pratiquement aucune n'avait de terme en commun. Ainsi, la similarité pour chacune des paires de documents en utilisant Jaccard aurait été nulle. C'est en partie cette constatation qui a encouragé nos travaux.

5.2 Etiquetage

Protocole : Nous présentons à chaque testeur les étiquettes (un ensemble de concepts) des classes que notre système a découvertes. Pour un ensemble de documents (environ 50 par testeur) nous lui demandons de mettre le document dans une des catégories générées automatiquement par le système, ou bien de le mettre dans une catégorie *bruit* si il estime qu'aucune étiquette ne correspond. En quelque sorte nous posons au testeur la question, "si vous aviez à mettre une étiquette sur ce document, laquelle choisiriez vous, sachant que vous pouvez choisir de ne pas en mettre du tout si aucune ne vous satisfait". Nous indiquons dans la table 5 le pourcentage de documents qu'un testeur humain 'moyen' (c'est-à-dire en réalité une moyenne sur les testeurs) a mis dans la même catégorie que le système, pour diverses valeurs du paramètre $MinSim$ de l'algorithme. Une corrélation de 80% signifie que sur les 50 documents, 40 auront été classés dans

1. Le *recall* et *precision* de la littérature anglophone sur Information Retrieval. On considère ici que les résultats à ramener sont les n documents au dessus du seuil humain, et que le système retourne m documents, dont p sont corrects. Le *recall* est donné par $\frac{p}{n}$ et la *précision* par $\frac{p}{m}$

MinSim = 0.80
Taux de rappel Humain / Vivissimo : 33%
Taux de précision Humain / Vivissimo : 40%
Corrélation Humain / Humain : 79%

TAB. 4 – *Résultats pour Vivissimo ($MinSim = 0.80$)*

MinSim	0.6	0.7	0.8	0.9
Corrélation humain/THESUS	75%	82%	80%	68%

TAB. 5 – *Etiquetage*

la même classe par l'humain 'moyen' et le système. Ainsi cette méthode nous permet de vérifier la capacité de notre système à générer une étiquette pertinente. En revanche, nous n'avons pas fait d'expérience sur la façon dont les testeurs humains auraient pu regrouper ensemble ces mêmes documents.

Résultats : Les résultats sont donnés dans la Table 5. Soulignons que seuls environ 5% des documents n'ont pas été mis dans une catégorie par notre système. Nous estimons que cette valeur très faible de bruit, couplée à une forte corrélation entre les humains et le système concourent à prouver que l'etiquetage est performant. La perte de performance lorsque $MinSim$ est très élevé s'explique par le fait que l'algorithme commence à rater des classes. Pour information, le temps de calcul pour faire tourner l'algorithme sur 39000 document sur un Pentium III, 450MHz, 512 MB RAM est de 45 secondes. Ce temps prend en compte à la fois la création des classes, ainsi que leur étiquetage, mais le précalcul des distances entre les termes de l'ontologie n'est pas compris.

6 Conclusion

Dans cet article, nous avons présenté une nouvelle mesure de similarité entre ensembles de concepts d'une hiérarchie, et avons proposé d'utiliser cette mesure avec l'algorithme DBSCAN. Les résultats obtenus sont probants, et justifient notre approche. Un point important que nous n'avons pas abordé ici est l'hypothèse de base, qui est que les documents web peuvent être représentés par un ensemble concis de concepts. Il est intéressant de noter que dans la construction de cet ensemble, nous nous servons notamment de la mesure définie ici, que nous appliquons sur WordNet, en la considérant comme une *ontologie*.

Remerciements: G. Cobena et S. Abiteboul pour les discussions sur SPIN, un entrepôt de données thématique. Ch. Froidevaux, C. Nicaud, K. Nørvåg, B. Safar et L. Segoufin (mesures de similarité) J.P. Sirot (ontologies) P. Rigaux et P. Veltri (R*-Tree). Nous remercions nos testeurs : Stratis, Yannis, Magdalini, Georges, Christos, Omar, Mathieu, Julien, Sandrine et Antoinette pour leur aide. Enfin, un grand merci à C.Jacquin et E.Desmontils pour avoir suggéré l'écriture de cet article, et B. Safar pour ses précieux conseils concernant la version française.

Références

[AGY99] Charu Aggarwal, Stephen Gates, and Philip Yu. On the merits of building categorization systems by supervised clustering. In *Proceedings of the ACM-SIGKDD*, 1999.

[BFS02] A. Bidault, Ch. Froidevaux, and B. Safar. Proximité entre requêtes dans un contexte médiateur. In *RFIA*, 2002.

[BKSS90] N. Beckmann, H.P. Kriegel, R. Schneider, and B. Seeger. The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the ACM-SIGMOD*, 1990.

[DJ01] E. Desmontils and C. Jacquin. Des ontologies pour indexer un site web. In *Journées Francophones d'Ingénierie des Connaissances*, 2001.

[EKSX96] Martin Ester, Hans-Peter Kriegel, Jord Sander, and Xiaowei Xu. A density based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM-SIGKDD*, 1996.

[EM97] Thomas Eiter and Heikki Mannila. Distance measures for point sets and their computation. *Acta Informatica Journal*, 34:109–133, 1997.

[Fis87] Douglas Fischer. Knowledge acquisition via incremental conceptual clustering. *Machine Learning Journal*, 2:139–172, 1987.

[GGK01] Aristides Gionis, Dimitrios Gunopulos, and Nick Koudras. Efficient and tunable similar set retrieval. In *Proceedings of the ACM-SIGMOD*, 2001.

[GHOS96] J. Green, N. Horne, E. Orlowska, and P. Siemens. A rough set model of information retrieval. *Theoretica Informaticae*, 28:273–298, 1996.

[HNVV02] M. Halkidi, B. Nguyen, I. Varlamis, and M. Vazirgiannis. Thesus: Organising web document collections based on semantics and clustering. Technical report, Verso Technical Report, 2002.

[kar] http://www.kartoo.com/.

[Kle99] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[Lin98] D. Lin. An information theoretic definition of similarity. In *Proceedings of the 15th ICML*, 1998.

[Nii87] I. Niiniluoto. Truthlikeness. 1987.

[PW00] T. Phelps and R. Wilensky. Robust hyperlinks cost just five words each. Technical Report UCB//CSD-00-1091, UC Berkeley Computer Science, 2000.

[Res95] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, 1995.

[Res99] P. Resnik. Semantic similarity in a taxonomy: An information based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11, 1999.

[RSM] R. Richardson, A.F. Smeaton, and J. Murphy. Using wordnet as a knowledge base for measuring semantic similarity between words. Technical report.

[SM83] Gerard Salton and Michael McGill. *An Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[Viv] http://www.vivissimo.com/.

[VNVA02] I. Varlamis, B. Nguyen, M. Vazirgiannis, and S. Abiteboul. Effective thematic selection in the www based on link semantics. Technical report, 2002.

[Wor] http://www.cogsci.princeton.edu/~ wn.

[WP94] Z. Wu and M. Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Associations for Computational Linguistics*, 1994.

[ZE98] O. Zaimir and O. Etzioni. Web document clustering, a feasibility demonstration. In *Proceedings of the ACM-SIGIR*, 1998.

Recherche d'informations

Une Terminologie Orientée Ontologie pour la Recherche d' Information sur la Toile

LF. SOUALMIA ET SJ. DARMONI
Laboratoire PSI, FRE 2645 CNRS, INSA de Rouen
Place Emile Blondel, BP 08
76131 Mont Saint Aigan, FRANCE
&
DIR, CHU de Rouen
1, rue de Germont
76031 Rouen Cedex, FRANCE
Mail :{Lina.Soualmia, Stefan.Darmoni}@chu-rouen.fr
Tel : + 33 2 32 88 88 00 Fax : + 33 2 32 88 88 32

Résumé

L' information accessible sur la Toile (ou Web) est disponible en quantité importante et elle ne cesse de croître. La recherche d' information demeure problématique malgré l' existence de nombreux moteurs de recherche et de sites catalogues en ligne. Le Web doit faire face aux problèmes d' exhaustivité et de précision en recherche d' information. Le projet CISMeF (Catalogue et Index de Sites Médicaux Francophones) a été développé afin de faciliter l' accès à l' information de santé disponible sur l' Internet. La problématique d' aujourd' hui se veut aussi être une *recherche d' information intelligente* dans l' infrastructure du Web Sémantique, une extension du web actuel qui permettrait de rendre interprétable le contenu des ressources par les hommes mais aussi par les machines, grâce à des ontologies et des méta-données. La recherche d' information dans CISMeF se fait à l' aide d' une terminologie semblable à une ontologie et un ensemble de méta-données qui nous permettent de placer le projet à cheval entre le Web actuel qui est informel, et le Web Sémantique de demain.

Abstract

There is a large amount of accessible data on the Web. Information retrieval remains problematic in spite of the numerous existing online catalogues and search engines. The Web must face exhaustivity and precision problems of information retrieval. The CISMeF project (acronym of Catalogue and Index of French-speaking Medical Sites) was developed in order to facilitate the access to the health information available on Internet. Today the problem is also the *intelligent information retrieval* in the Semantic Web infrastructure, an extension of the current Web in which the resources' content would be interpretable not only by humans, but also by machines thanks to ontologies and metadata. Information retrieval in the CISMeF catalogue is done with a terminology that is similar to ontology and a set of metadata. This allows us to place the project at an overlap between the actual Web, which is informal, and the forthcoming Semantic Web.

1 Introduction

La quantité d'information disponible sur le Web est importante et elle ne cesse de croître. La recherche d'information demeure problématique: il est en effet difficile de trouver ce que l'on recherche malgré l'existence de sites catalogues (comme par exemple Yahoo, http://www.yahoo.fr) et de moteurs de recherche (par exemple Google, http://ww.google.com). Les moteurs de recherche par mots clés renvoient généralement comme réponse à une requête un grand nombre de pages à consulter, ce qui demande à l'utilisateur de faire lui-même le tri dans cette masse d'information. Les résultats ne sont pas tous pertinents et l'information retrouvée n'est pas complète. La recherche plein texte n'est pas toujours efficace : la page recherchée peut utiliser un terme sémantiquement proche mais syntaxiquement différent de celui de la requête; les fautes de frappe et les variantes lexicales sont généralement considérées comme étant des termes différents ; les moteurs de recherche actuels ne peuvent pas traiter *intelligemment* les pages HTML, langage le plus répandu sur le Web. Dans les catalogues comme Yahoo les ressources sont indexées et classées manuellement en fonction de catégories qui sont d'ordre trop général pour répondre à des requêtes spécifiques: en effet il y a souvent un recoupement entre les différentes catégories et une certaine ambiguïté quant à leur étendue [Risden, 1999]. Ceci peut dérouter l'utilisateur qui ne sait pas trop dans quelle thématique rechercher son information. Par exemple dans Yahoo, la catégorie *Actualité et Médias* est à la fois une catégorie principale mais aussi une sous-catégorie de *Santé*, *Sport et Loisir* ou encore *Enseignement et Formation.*

Aujourd'hui, la problématique qui se pose est celle d'une *recherche d'information intelligente* sur le Web. Le Web Sémantique [Berners-Lee, 2001] est un espace d'échange qui reste à construire. Un de ses intérêts est d'une part d'apporter suffisamment de renseignements sur les ressources, en ajoutant des annotations sous la forme de *méta-données* et d'autre part, de décrire leur contenu de manière à la fois formelle et signifiante à l'aide d'une *ontologie* pour être interprétables aussi bien par les humains que par les machines. Cet espace doit être formalisé, le Web actuel étant informel. En effet, il est composé principalement de pages HTML écrites à la main ou générées automatiquement pour un traitement humain. Les ontologies et les méta-données sont donc deux éléments principaux pour la construction de l'infrastructure du Web Sémantique [Laublet et al., 2002]. Une ontologie est une modélisation partagée d'un domaine pour améliorer la communication et éliminer les ambiguïtés entre personnes, entre personnes et applications ou entre applications. Elle est composée d'une hiérarchie de concepts[1], de relations entre concepts et d'un ensemble de règles ou de contraintes. Les méta-données font référence à une information descriptive des ressources du Web. Leur première utilité est la recherche d'information.

L'objectif de cet article est en premier lieu de présenter le projet CISMeF (Catalogue et Index de Sites Médicaux Francophones http://www.chu-rouen.fr/cismef/) [Darmoni et al., 2001], dont la structure nous permet de le placer à cheval entre le web informel d'aujourd'hui et le Web Sémantique de demain. Nous proposons ensuite l'application de méthodes issues de domaines différents pour améliorer la recherche d'information dans le catalogue. Tout d'abord nous décrivons en section (2) la modélisation des méta-données et de la terminologie CISMeF qui est semblable à une ontologie du domaine médical, sans être aussi exhaustive et cohérente qu'une ontologie formelle. Nous décrivons en section (3) l'exploitation de cette ontologie dans le site Web du catalogue et proposons en section (4) différentes méthodes pour optimiser la recherche d'information. Enfin, en conclusion, nous décrivons quelques perspectives quant à la mise en œuvre de ces méthodes.

2 Vers un web sémantique médical

Le catalogue CISMeF a été développé en 1995 pour assister les professionnels de santé, les étudiants et le grand public dans leur recherche d'information de santé sur le Web. CISMeF et Doc'CISMeF, le moteur de recherche associé, prennent en compte la diversité des utilisateurs et leur permettent de trouver des documents de qualité qui répondent à un besoin précis. Un grand nombre de ressources (n=12,131) sont sélectionnées en fonction de critères stricts par une équipe de documentalistes et sont répertoriées selon une méthodologie de mise à jour du catalogue. Une

[1] Une terminologie rassemble l'ensemble des termes d'un domaine, l'ontologie en désigne les concepts (le ou les sens des termes) [Fortier, 2001].

ressource peut être un site Web, une page Web, un document, un rapport : tout support qui contient des informations relatives à la santé. La description de ces ressources se fait à l' aide de *notices* en se basant sur un ensemble de méta-données et une terminologie structurée semblable à une ontologie documentaire du domaine médical.

2.1 Les méta-données de CISMeF

La recherche d' information est la première utilité des méta-données [Laublet et al., 2002]. Ce sont des données concernant les données elles-mêmes. Les ressources répertoriées dans CISMeF sont décrites par onze des éléments du Dublin Core (*auteur*, *date*, *description*, *format*, *identifiant*, *langage*, *éditeur*, *type de ressource*, *droits*, *sujet* et *titre*) ainsi que huit autres éléments spécifiques à CISMeF (*institution*, *ville*, *province ou département*, *pays*, *public ciblé*, *type d' accès*, *coût* et *parrainage* de la ressource). Le type d' utilisateur est également pris en compte. Pour les ressources destinées aux professionnels de la santé (les lignes directrices et consensus de bonne pratique clinique) deux champs supplémentaires sont définis par CISMeF : *indication du niveau de preuve* et la *méthode* utilisée pour le déterminer. Pour les ressources pédagogiques ce sont onze éléments de la catégorie « Educational » du standard IEEE 1484 qui sont rajoutés. Le format de ces méta-données est passé du langage HTML en 1995 à XML en 2000 pour des soucis d' interopérabilité et depuis décembre 2002 à RDF [Lassila and Swick, 1999] dans le cadre du projet européen MedCIRCLE [Mayer et al., 2003]. Le vocabulaire des méta-données HIDDEL[2] est contenu dans une ontologie (représentée à l' aide du langage RDFS) et les ressources sont désormais décrites en RDF à partir des éléments de l' ontologie HIDDEL.

2.2 La terminologie CISMeF

Les ressources sont indexées en fonction de la terminologie CISMeF. Celle-ci a été construite à partir des concepts du thésaurus MeSH[3] (Medical Subject Headings développé depuis 1960) et de sa traduction en français fournie par l' INSERM[4] (Institut National de la Santé et de la Recherche Médicale). Le MeSH dans sa version 2003 est composé de 22,012 *mots clés* (exemple : *abdomen*, *hépatite*) et 84 *qualificatifs* (exemple : *diagnostic*, *complications*, *thérapeutique*) regroupés sous la forme d' arborescences de concepts médicaux. Les mots clés sont organisés sous la forme de hiérarchie à 9 niveaux allant du terme le plus général en haut de la hiérarchie aux termes les plus spécifiques en bas de la hiérarchie. Par exemple le mot clé *aberration chromosomique* est plus général que le mot clé *trisomie*. Les qualificatifs, organisés également en hiérarchie, permettent de préciser le sens des mots clés en limitant leur étendue à certains aspects. Par exemple l' association du mot clé *lombalgie* et du qualificatif *diagnostic* (notée *lombalgie/diagnostic*) permet de restreindre la *lombalgie* au seul aspect *diagnostic*. Bien qu' il existe des ontologies médicales comme GALEN [Rodrigues et al., 1998] ou MENELAS, concernant les coronaropathies [Bouaud et al., 1995], c' est le MeSH qui a été choisi car il correspond aux attentes des documentalistes et il est connu des professionnels de santé.

Les mots clés ont été regroupés dans CISMeF en fonction de spécialités médicales (n=66) intitulés *métatermes* (exemple : Cancérologie). Ce sont des super-concepts qui permettent une vision plus globale concernant une spécialité en offrant un niveau supplémentaire d' abstraction. Les métatermes permettent de connaître l' ensemble des termes MeSH qui sont répartis dans plusieurs arborescences mais qui concernent une même spécialité. Une hiérarchie de *types de ressources* (n=115) a été modélisée et elle permet de décrire la nature de la ressource (exemple : *cours*, *information patient*). A partir du MeSH fournit sous la forme de fichiers texte (Figure 1) seules les relations du type '*est-un*' et '*partie-de*' sont utilisées pour définir des liens de subsomption dans la hiérarchie des mots clés CISMeF (n=7,435 soit ~34% du MeSH). Ces liens de subsomption sont notamment exploités pour la recherche d' information. Par exemple le concept *Oreille* initialement défini comme étant *partie-de* du concept *Tête*, est défini dans la terminologie CISMeF par *Oreille* est *subsumé* par le concept *Tête*. Les fichiers MeSH sont traités automatiquement pour renseigner la terminologie CISMeF afin qu' elle soit exploitable au niveau du site de CISMeF.

[2] Health Information Description Description Evaluation Language, http://www.medcircle.org
[3] Le MeSH est produit par la US-National Library of Medicine pour la base documentaire Medline.
[4] http://dicdoc.kb.inserm.fr:2010/basimesh/mesh.html

```
LT = POMPES IONIQUES
UF = POMPES A IONS
NT = ANTIPORTEURS
NT = PROTEINES DE TRANSPORT ANIONS
NT = PROTEINES DE TRANSPORT CATIONS
NT = SYMPORTEURS
RT = CANAL MEMBRANAIRE
RT = TRANSPORT BIOLOGIQUE ACTIF
RT = TRANSPORT IONIQUE
```

```
LT= Terme principal
UF= Synonyme
NT= terme spécifique
RT= Voir Aussi
```

Figure 1. Exemple de termes et de relations dans les fichiers texte du MeSH

Schématiquement, la terminologie a la structure suivante (Figure 2) :

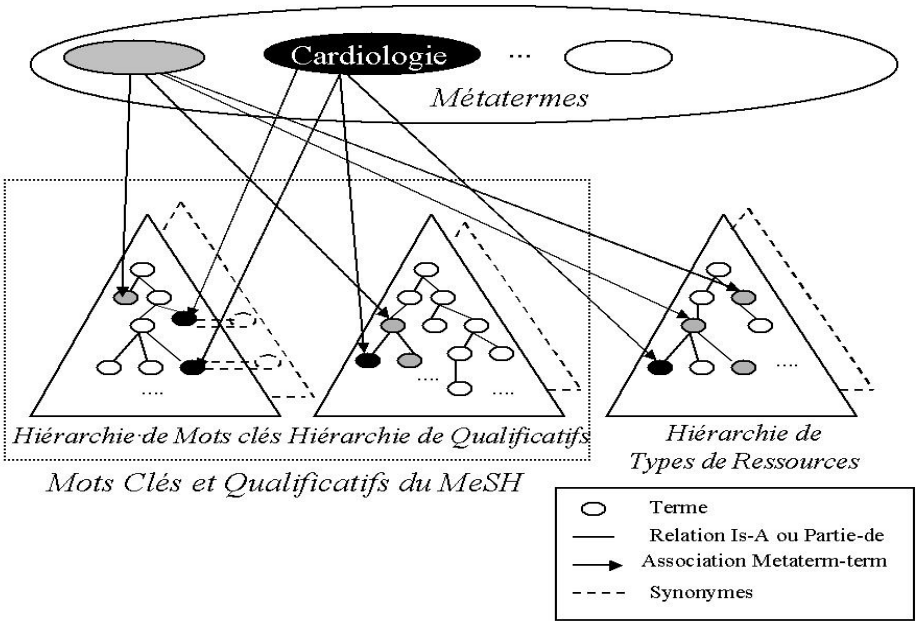


Figure 2. Structure de la terminologie CISMeF.

La terminologie CISMeF a une structure semblable une ontologie terminologique [Sowa, 2000] :

- ✓ Le vocabulaire est bien connu des documentalistes et des professionnels de la santé et il correspond à celui du domaine médical.
- ✓ Chaque concept (Figure 3) a :
  - un terme préférentiel (`Descripteur`) pour l' exprimer en langue naturelle
  - un ensemble de propriétés
  - une définition en langage naturel pour quelquefois le différencier des concepts le subsumant et de ceux qu' il subsume
  - un ensemble de synonymes
  - un ensemble de règles et de contraintes (Figure 4)
- ✓ Les concepts sont organisés selon une relation de subsomption allant du concept le plus général au plus spécifique.

D' après l' exemple de définition de la Figure 3, le terme associé au concept ayant l' identifiant (unique) `D006521` est ***`Hépatite Chronique`***. Le `code cat.MeSH` indique à quel niveau ce concept est situé dans la hiérarchie : on peut déduire que ***`Hépatite Chronique`*** (`C06.552.380.350`) est subsumé par ***`Hépatite`*** (`C06.552.380`). La Figure 4 est un exemple de contraintes sous la forme de règles à appliquer sur les concepts. Par exemple l' association `hépatite/induit chimiquement` est équivalente (=) au concept `hépatite, toxique`. Toutes ces informations ainsi que les notices descriptives des ressources sont stockées dans une base de données relationnelle qui est exploitée par le serveur du site Web de CISMeF.

```
Descripteur Francais: HEPATITE CHRONIQUE
Descripteur Americain: Hepatitis, Chronic
Code Cat MESH: C06.552.380.350
Synonymes Français: HEPATITE CHRONIQUE ACTIVE
Synonymes Américains: Chronic Hepatitis
                      Cryptogenic Chronic Hepatitis
                      Hepatitis, Chronic, Cryptogenic
Derives Americains: Hepatitis, Chronic Active
                    Active Hepatitides, Chronic
                    Active Hepatitis, Chronic
                    Chronic Active Hepatitides
                    Chronic Active Hepatitis
                    Chronic Hepatitides
                    Chronic Hepatitides, Cryptogenic
                    Chronic Hepatitis, Cryptogenic
                    Cryptogenic Chronic Hepatitides
                    Hepatitides, Chronic
                    Hepatitides, Chronic Active
                    Hepatitides, Cryptogenic Chronic
                    Hepatitis, Cryptogenic Chronic
MESH définition: A collective term for a clinical and pathological syndrome which has several causes
and is characterized by varying degrees of hepatocellular necrosis and inflammation. Specific forms of
chronic hepatitis include autoimmune hepatitis (HEPATITIS, AUTOIMMUNE), chronic hepatitis B;
(HEPATITIS B, CHRONIC), chronic                    hepatitis C; (HEPATITIS C, CHRONIC), chronic
hepatitis D; (HEPATITIS D, CHRONIC), indeterminate chronic viral hepatitis, cryptogenic chronic
hepatitis and drug-related chronic hepatitis (HEPATITIS, CHRONIC, DRUG-INDUCED).
Numero NLM: D006521
```

Figure 3. Exemple de définition de concept.

```
Hepatitis : C06.552.380+
Viral Hepatitis = Hepatitis, Viral Human and Hepatitis, Viral Animal
/chemically induced = Hepatitis, Toxic
/veterinary = Hepatitis, Animal or Hepatitis, Viral, Animal
hepatitis parenterally transmitted= Hepatitis C
hepatitis enterally transmitted = Hepatitis E
not specified as parenteral or enteral = probably Hepatitis, Viral, Human
Non-A, Non-B hepatitis = probably Hepatiis C
```

Figure 4. Exemple de contraintes sur les concepts.

3 Exploitation de l' ontologie documentaire

La structure de l' ontologie est exploitée pour l' indexation des documents, la visualisation etla navigation dans les hiérarchies des concepts du domaine et la recherche de ressources par le moteur Doc' CISMeF. Chaque ressource est indexée à l' aide des termes de l' ontologie (mots clés, qualificatifs et types de ressources). A partir d' un ensemble d' heuristiques et d' un algorithmde classification, les spécialités (métatermes) auxquelles se rattachent les ressources sont déduites en utilisant les différents liens existants entre (métaterme-mot clé), (métaterme-qualificatif) et (métaterme-type de ressource) et classées en fonction de leur niveau d' importance.

La navigation dans l' ontologie, grâce à un index thématique et alphabétique, permet à l' utilisateur d' appréhender les termes du domaine et leurs relations en affichant les différentes hiérarchies auxquelles il appartient. Chaque terme a sa propre page associée et des liens qui permettent de rechercher toutes les ressources qui y sont rattachées, ou encore restreindre la recherche en fonction de l' utilisateur (ressources destinées aux professionnels, aux étudiants ou aux patientset au grand public).

La dernière utilité de l' ontologie est son exploitation par le moteur de recherche. Différents modes sont possibles. La recherche dite *simple* permet à l' utilisateur de saisir une requête en texte libre en français ou en anglais avec ou sans accent en majuscule ou en minuscule. La recherche dite *avancée* engage des recherches plus pointues à l' aide d' un formulaire contenant des listes déroulantes et permet de combiner plusieurs champs (mots clés, titre, année…etc.) avec des opérateurs bo oléens (ET , OU, SAUF). La recherche *logique* s' effectue à l' aide d' un langage de requêtes associé, des opérateurs booléens et des caractères spéciaux.

La recherche « simple » telle qu' en place aujourd' hui se base sur les relations de subsomption. Si le terme (un mot ou une expression) saisi par l' utilisateur est un terme existant dans l' ontologie, le résultat de la requête est l' union de toutes les ressources instances du terme et des ressources instances des termes qu' il subsume, directement ou indirectement, et ce dans toutes les hiérarchies dans lesquelles il peut se trouver. Par exemple une requête sur le terme *tumeur* va renvoyer comme réponse l' ensemble des ressources rattachées à *tumeur* mais également celles rattachées à *tumeur colon*, *tumeur rectum*…etc. De même qu' une requête sur *tête* va renvoyer les ressources rattachées à *tête* mais également à *oreille, nez*…etc. et c' est d' ailleurs pour cette raison que nous considérons les liens *partie-de* du MeSH comme des liens de subsomption dans CISMeF. Si le terme saisi par l' utilisateur n' est pas un terme réservé, une recherche sur tous les autres champs de méta-données est effectuée,

voire plein texte sur tous les documents indexés. Ce type de recherche simple nécessite donc une bonne connaissance des termes de CISMeF, ce qui n' est pas évident pour un utilisateur novice.

4 Améliorer la recherche d' information

La ou les requêtes saisies par l'utilisateur correspondent rarement à la formulation exacte effectivement utilisée pour l' indexation. Nous avons extrait les requêtes des utilisateurs à partir des logs du serveur http du moteur Doc' CISMeF et déterminé le type de requête employé ainsi que le nombre de réponses obtenu entre le 15/08/2002 et le 06/02/2003 (Tableau 1).

	Requêtes		**Requêtes Nulles**	
Type de Requête	**Nombre**	**Pourcentage**	**Nombre**	**Pourcentage**
Simple	892 591	58.62 %	365 688	40.97 %
Autre	630 175	41.38 %	144 790	22.97 %
Total	1 522 776		510 478	

Tableau 1. Analyse des requêtes des utilisateurs du 15/08/2002 au 6/02/2003.

Une analyse plus fine des requêtes simples (Tableau 2) nous a permis de déduire que 12.01% des réponses sont nulles non pas du fait qu' elles correspondent à des requêtes erronées (ce sont bien des termes réservés), mais du fait qu' aucune ressource ne leur est rattachée.

	Nombre	**Pourcentage**
Expression reconnue	43 922	12.01 %
Expression non reconnue	321 766	87.99%
Total	365 688	

Tableau 2. Répartition des requêtes simples à 0 réponse.

Pour améliorer cette recherche, nous proposons d' appliquer trois méthodes issues de domaines différents. Nous proposons d' évaluer l' apport de chacune d' elles en terme de recherche d' information dans le cadre du projet CISMeF et d' étudier leur complémentarité. Nous détaillons pour chacune d' elles les prétraitements des données qui sont nécessaires à leur application.

4.1 Traitement Automatique du Langage Naturel

Afin de maximiser les chances de retrouver l'information souhaitée, nous nous appuyons sur une analyse morpho-syntaxique. Nous avons analysé au préalable la structure de la terminologie de CISMeF (Tableau 3) relative la composition des termes réservés.

Nombre de mots	**Mots Clés**	**Qualificatifs**	**Types de Ressources**	**Termes**
1	1 437	55	28	1 520
2	1 706	10	42	1 758
3	612	11	39	662
4	148	3	12	163
5	40	--	4	44
6	8	--	2	10
7	2	--	--	2
TOTAL	**3953**	**79**	**127**	**4 159**

Tableau 3. Structure des termes utilisés pour l' indexation

Une étude préliminaire [Zweigenbaum et al., 2001] a déjà été réalisée sur un ensemble de requêtes lancées sur Doc' CISMeF. Les résultats ont montré que l' utilisation de connaissances morphologiques amélioraient sensiblement les résultats des requêtes en diminuant le nombre de réponses nulles. La base de connaissances morphologique a été construite automatiquement dans des travaux antérieurs et l' algorithme proposé consiste à corriger la requête de l' utilisateur (dans le cas de non réponse seulement) en éliminant les mots vides (*comment, alors, du*…etc.) et en remplaçant chaque terme de la requête par une disjonction de tous les termes de sa famille morphologique. Une famille morphologique est composée de flexions, dérivations et compositions. Par exemple le terme *Cœur* a comme flexion *Cœurs*, comme dérivation *Cardiaque*, et comme composition *Cardiovasculaire*. Si l' utilisateur saisit la requête `interaction entre médicaments et alimentation` l' algorithme permet de retrouver le mot clé *interaction aliment médicament*.

La base de connaissances morphologiques n' a pas été spécifiquement construite par rapport aux termes de CISMeF et elle contient 6,312 couples (mot | adjectif). Après comparaison (Tableau 4) avec un sous-ensemble des termes de CISMeF utilisés pour l' indexation des ressources, nous avons obtenu 568 termes dérivés qui couvrent « exactement » 516 Mots Clés.

	Mots Clés	Qualificatifs	Types de Ressources	Termes
Nb identifiés	516	39	13	568
Couverture 1 mot	35.91%	70.91%	46.42%	37.37%
Couverture Total	13.05%	49.37%	10.24%	13.66%

Tableau 4. Couverture exacte à l' aide de la base initiale créée automatiquement

Dans un second temps nous avons complété nos données grâce à la ressource terminologique Lexique [New et al., 2001]. Elle comporte tout le lexique du français contemporain déduit à partir d' un corpus de textes, écrits entre les années 1950 et 2000, en se basant sur des calculs de fréquences d' apparition des mots contenus dans des pages du Web. Il n' existe aujourd' hui aucune ressource lexicale en langue française concernant la terminologie médicale. Ce module s' inscrit dans le projet UMLF [Zweigenbaum et al., 2003]. Toutes les variantes lexicales possibles, y compris les verbes et subjonctifs, sont présentes dans Lexique et la liste est relativement complète.

L' analyse des termes composés de 2 ou plusieurs mots nous a permis de déduire que 1,935 termes étaient ' semi-couverts' (1,899 mots clés; 8 qualificatifs; 28 types de ressources). Par exemple *accident circulation* est un mot clé composé d' un terme dérivé du mot clé *accidents* qui a comme famille : *{accident, accidents, accidenté, accidentés, accidentées, accidentel, accidentels, accidentelle, accidentelles, accidentellement, accidenter}*. Celle-ci existant déjà dans la base grâce à l' étape de reconnaissance des termes, on considère que le terme *accident circulation* est semi-couvert. Il reste donc à compléter ces connaissances.

	Couverture	Mots Clés	Qualificatifs	Types de Ressources	Termes
Lexique	**Nb identifiés**	943	48	26	1 017
	Couverture 1 mot	65.62%	87.27%	92.85%	66.91%
	Couverture Totale	23.86%	60.76%	20.47%	24%
Lexique & Base Initiale	**Nb identifiés**	1 207	55	28	1292
	Couverture 1 mot	83.99%	100%	100%	85%
	Couverture Totale	30.53%	69.62%	22.04%	31.06%

Tableau 5. Récapitulatif des couvertures exactes de la terminologie

Les résultats obtenus dans [Grabar et al., 2003] montrent qu' une normalisation des requêtes et de la terminologie augmente la couverture de la terminologie : en effet si le mot clé est *accidenté* la requête *Accidentè* sera nulle. Nous avons donc désaccentué et mis en minuscule tous les termes dérivés obtenus. La gestion des accents et minuscules est également effectuée sur les requêtes au niveau du prototype de recherche développé pour la réalisation des différents tests. A présent, l' algorithme développé permet de déduire à partir de la requête simple *douleurs dorsales* la requête logique `douleur.mc ET dorsalgie.mc`, avec `mc` indiquant que le terme considéré a été reconnu par l' algorithme de recherche comme étant un mot clé.

La base de données Lexique nous a permis par la même occasion d' extraire tous les termes qui pouvaient correspondre à des mots vides. Les mots vides considérés sont tous les adjectifs possessifs (*mon*), les conjonctions (*mais*), les déterminants (*du*), les interjections (*diantre*), les prépositions (*durant*), les pronoms personnels (*il*), les pronoms possessifs (*leur*) et les pronoms relationnels (*auquel*). Nous avons déterminé ainsi 873 mots vides supplémentaires aux 473 initiaux, nous donnant un total de 1,346 mots vides. Ce nombre est élevé vu que des termes comme *boum, bye, bravo* ou encore *sniff* sont considérés comme vides. La requête *douleur du bas du dos* est ainsi transformée en `douleur.mc ET dos.mc`.

En plus de connaissances morphologiques, des connaissances sémantiques sont nécessaires. En effet, les termes synonymes du MeSH ne correspondent pas aux termes en usage courant et sont plutôt des réécritures de termes (Figure 3). Par exemple le terme médical correspondant à *fausse couche* est *avortement spontané*. Nous étudions actuellement les logs des utilisateurs et collaborons avec des associations de patients et la Ligue Nationale contre le Cancer pour compléter la liste des synonymes CISMeF.

4.2 Découverte de connaissances dans les bases de données

Nous souhaitons découvrir de nouvelles connaissances à partir de la base de données CISMeF (en particulier à partir des notices et des termes) qui seront exploitées dans le processus de recherche d'information. Nous appliquons une technique de Data Mining appelée *Règles d'Association* dans le but d'extraire des associations intéressantes, non triviales, précédemment inconnues à partir de la base. Les règles d'association ont été initialement utilisées en analyse des données puis en fouille de données dans les bases de données relationnelles de grande taille [Agrawal and Srikant, 1994]. Nous nous intéressons à la découverte de règles d'association booléennes. Une règle d'association booléenne RA est de la forme :

$$RA : a_1 \wedge a_2 \wedge \ldots \wedge a_i \Rightarrow a_{i+1} \wedge \ldots \wedge a_n \quad (1)$$

Elle s'interprète intuitivement de la manière suivante: si un objet possède les attributs $\{a_1, \ldots, a_i\}$ alors il a tendance à posséder également les attributs $\{a_{i+1}, \ldots, a_n\}$. Le *support* d'une règle représente son utilité. Cette mesure correspond à la proportion d'objets qui contiennent à la fois l'antécédent et le conséquent de la règle. La *confiance* représente sa précision. Cette mesure correspond à la proportion d'objets contenant le conséquent de la règle parmi ceux contenant l'antécédent. Le processus d'extraction de connaissances est composé de plusieurs phases: la préparation des données et du contexte (sélection des objets et des attributs), l'extraction des ensembles fréquents d'attributs (*itemsets* fréquents par rapport à un seuil de support minimum), la génération des règles d'association les plus informatives à l'aide d'un algorithme de Data Mining (par rapport à un seuil de confiance minimum) et enfin l'interprétation des résultats (ou déduction de nouvelles connaissances).

Notre contexte d'extraction est le triplet $C=(O,A,R)$ avec O l'ensemble des objets, A l'ensemble des items, R une relation binaire entre O et A. Les objets sont les notices utilisées pour décrire les ressources indexées. Elles ont un identifiant unique. La relation R correspond à la relation d'indexation entre un objet et un item. Nous considérons pour l'instant deux cas différents pour les items: A={Mot Clé} l'ensemble des Mots clés; A={(Mot Clé, Qualificatif)} l'ensemble des couples (Mot Clé, Qualificatif).

Un itemset est fréquent dans son contexte C si son support est supérieur à un seuil minimal défini au préalable. Le problème de l'extraction des itemsets fréquents est de complexité exponentielle dans la taille n de l'ensemble d'items, le nombre d'itemsets fréquents potentiels étant 2^n. Dans le cas a) nous avons n=7,435. Les itemsets forment un treillis [Davey and Priestley, 1994]. Plusieurs algorithmes de découverte d'itemsets fréquents ont été proposés. Le plus connu est l'algorithme Apriori [Agrawal and Srikant, 1994]. Nous utilisons l'algorithme A-Close [Pasquier, 2000]. L'extraction se fait par le calcul des itemsets *fermés* fréquents avec l'opérateur de fermeture de la connexion de Galois d'une relation binaire finie [Ganter and Wille, 1999]. L'espace des itemsets à étudier est ainsi réduit. L'algorithme détermine aussi les générateurs des itemsets fermés fréquents. Les générateurs d'un itemset fermé I_f sont les itemsets de taille maximale dont la fermeture est égale à I_f. Les bases pour les règles d'association sont déterminées à partir des itemsets fermés fréquents et de leurs générateurs. L'union de ces bases est un ensemble de générateurs non redondants de toutes les règles d'association non redondantes, d'antécédents minimaux et de conséquences maximales qui ne représentent aucune perte d'information. Ce sont les règles les plus utiles et les plus pertinentes.

Pour tester l'algorithme nous avons fixé le support à 5 documents et la confiance à 100%. La première étape de l'algorithme (itemsets de taille 2) nous a permis de trouver les règles suivantes :

`hépatite C` $\Rightarrow$ `sida` ; support = 14 (cas a)).

`sida/prévention et contrôle` $\Rightarrow$ `condom` ; support = 6 (cas b)).

La seconde étape de cette étude est de déterminer toutes les autres règles d'association qui seront exploitées dans le processus de recherche d'information par expansion de requêtes.

4.3 Raisonnement sur les ontologies

Afin de compléter l'ontologie CISMeF nous envisageons d'exploiter le réseau sémantique de l'UMLS [Lindberg et al., 1993]. Le réseau sémantique est composé de concepts médicaux (n=134) et de relations (n=54) entre les concepts. Elles sont de la forme *Complications (Hépatite, Cirrhose)* indiquant que le concept *Hépatite* est relié au concept *Cirrhose* par la relation *Complications*. En

analysant de plus près ces relations on remarque qu' elles correspondent aux qualificatifs du MeSH et que les concepts sont les mots clés du MeSH. Comme ce type de relation ne sera pas utilisé pour annoter les ressources et qu' elles ne contribuent pas à la définition de concepts, elles seront traduites sous la forme de règles d' inférence. Ces règles permettent entre autres d' enrichir l' ontologie car la seule information qui est disponible est que les concepts *Cirrhose* et *Hépatite* sont tous deux subsumés par le concept *Maladies du Foie.*

Pour un raisonnement sur le contenu des ressources, nous traduirons une partie de l' ontologie CISMeF dans le langage RDFS en transformant les mots clés et types de ressource en concepts et les qualificatifs en relations entre concepts. RDFS permet de définir des hiérarchies de classes et des propriétés mais il n' intègre pas de capacités de raisonnement, comme ceux qu' offrent les systèmes basés sur des langages formels comme les Logiques de Description. La cohérence (pas de contradiction) et la complétude (tous les concepts utiles doivent se trouver dans l' ontologie) d' une ontologie décrite en langage formel sont facilement vérifiables. L' écriture des règles d' inférence n' étant pas possible en RDFS, nous utiliserons les fonctionnalités de l' outil TRIPLE [Sintek and Decker, 2001] qui a été développé pour une recherche information intelligente basée sur les connaissances. Il permet de réaliser des raisonnements complexes sur des annotations RDF. L' outil traduit RDFS en Horn-Logic mais aussi en DAML+OIL[5]. En utilisant des classifieurs comme RACER [Haarslev and Möller, 2001] (basé sur les Logiques de Description) il permet également de vérifier la cohérence de l' ontologie. Nous espérons que le moteur de d' inférences de TRIPLE permettra de réaliser des requêtes d' un niveau supérieur dans CISMeF.

5 Perspectives et Conclusion

Nous avons abordé dans cet article la problématique de la recherche d' information sur la Toile. Nous avons présenté certains aspects du projet CISMeF qui a été créé pour assister les professionnels de santé, les étudiants en médecine ainsi que les patients et le grand public dans leur recherche. Il faut noter que CISMeF est le seul projet de ce type en France. La modélisation des informations et des ressources (utilisation de méta-données et d' une ontologie de domaine) est similaire à celle de projets comme KA² [Benjamins and Fensel, 1998] (à la différence que ce sont les auteurs qui annotent leurs documents), PME [Kassel et al., 2000] ou encore CoMMA [Gandon and Dieng-Kuntz, 2000]. Les « raisonnements » dans CISMeF exploitent les liens de subsomption entre concepts pour la recherche de documents ainsi que leur classification en fonction de spécialités, mais ils ne permettent pas de déduire de nouvelles connaissances. Nous avons également proposé différentes méthodes pour améliorer la recherche d' information. Les techniques de traitement automatique du langage naturel ont permis de construire une base de connaissances morphologique. Le data mining permettra de découvrir des règles d' association entre concepts et de revoir l' indexation des ressources. Enfin le raisonnement sur les ontologies offrira un niveau supérieur tant au niveau de l' ontologie (vérification de la consistance et cohérence, exploitation du réseau sémantique de l' UMLS) qu' au niveau recherche d' information. Nous pensons que la plus value se situera dans la combinaison de ces différentes techniques. L' évaluation de l' apport de chacune des méthodes se fera de deux manières. Tout d' abord par une expansion automatique (enrichissement) des requêtes pour élargir le champ de la recherche en utilisant chacune des ressources (base morphologique, règles d' association, et ontologie formelle) séparément puis conjointement. Les requêtes considérées sont celles du fichier log dont le nombre de réponses est nul. Ensuite par une expansion interactive : nous demanderons à un échantillon d' utilisateurs abonnés au site d' évaluer, pour chaque requête qu' ils poseront, la « utilité » des différentes suggestions de requêtes enrichies apportées par les différentes méthodes. Cette évaluation (expansion automatique ou interactive) à échelle réelle permettra d' établir une base de règles, ou un protocole pour l' application des méthodes en fonction du type de requête posé.

Références

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant , R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings VLDB Conference*, p. 478-499.

[Benjamins and Fensel, 1998] Benjamins, VR. and Fensel, D. (1998) The Ontological Engineering Initiative (KA)². *Proceedings International Conf. on Formal Ontologies in Informations Systems.*

[5] DAML+OIL joint committee. http://www.daml.org/2001/03/daml+oil-index.html.

[Berners-Lee et al., 2001] Berners-Lee, T., Heudler, J. and Lassila, O. (2001). The Semantic Web. *Scientific American.*p. 35-43.

[Bouaud et al., 1995] Bouaud, J., Bachimont, B., Charlet, J. and Zweigenbaum, P. (1995) Methodological Principles for Structuring an «Ontology ». *Proceedings of IJCAI conference*.

[Darmoni et al., 2001] Darmoni, SJ., Thirion, B., Leroy, JP., Douyère, M., Lacoste, B., Godard, G., Rigolle I., Brisou, M., Videau, S., Goupy, E., Piot, J., Quéré, M., Ouazir, S. and Abdulrab, H. (2001). A Search Tool based on ' Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, 26 (3) :165-178.

[Davey and Priestley, 1994] Davey, BA. and Priestley, HA.(1994) Introduction to Lattices and Order. *Cambridge University Press*.

[Fortier, 2001] Fortier, JY. (2001). Construction d'une ontologie pour gérer la documentation d'un centre de recherche. *Rapport Interne LaRIA*.

[Gandon et al., 2002] Gandon, F., Dieng-Kuntz, R., Corby, O. and Giboin, A. (2002) Web Sémantique et Approche Multi-Agents pour la Gestion d' une Mémoire Organisationnelle Distribuée. *Actes Journées Francophones d' Ingénierie des Connaissances*,p.15-26.

[Ganter and Wille, 1999] Ganter, B. and Wille R. (1999) Formal Concept Analysis : Mathematical Foundations. *Springer-Verlag*.

[Grabar et al., 2003] Grabar, N., Zweigenbaum, P., Soualmia, LF., and Darmoni SJ.(2003) Matching Controlled Vocabulary. *Medical Informatics Europe* (in press).

[Haarslev and Möller, 2001] Haarslev, V. and Möller, R. (2001) Description of the RACER System and its Applications. *Proceedings Internationl Worrkshop on Description Logics.*

[Kassel et al., 2000] Kassel, G., Abel, MH., Barry, C., Boulitreau, P., Irastorza, C. and Perpette, S. (2000) Construction et Exploitation d' une Ontologie pour la Gestion des Connaissances d' une Equipe de recherche. *Actes Journées Francophones d' Ingénierie des Connaissances.*

[Lassila and Swick, 1999] Lassila, O. and Swick, R. (1999) Resource Description Framework (RDF) Model and Syntax Specification. *W3C Candidate Recommendation 1999.*

[Laublet et al., 2002] Laublet, P., Reynaud, C. and Charlet, J. (2002). Sur Quelques Aspects du Web Sémantique. *Actes des deuxièmes assises nationales du GdRI3*, p.59-78.

[Lindberg et al., 1993] Lindgberg, DAB, Humphreys, BL and McCray, AT. (1993) The Unified Medical Language System. *Methods of Information in Medicine.*

[Mayer et al., 2003] Mayer, MA., Darmoni, SJ., Fiene, M., Köhler, C., Roth-Berghofer, T., and Eysenbach, G. (2003) MedCIRCLE - Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. *Medical Informatics Europe* p. 667-672.

[New et al., 2001] New, B., Pallier, C., Ferrand, L. and Matos R. (2001) Une Base de Données Lexicales du Français Contemporain sur Internet: LEXIQUE, *L'Année Psychologique*, p. 447-462.

[Pasquier, 2000] Pasquier N. (2000) Data Mining, : Algorithmes d' Extraction et de Réduction des Règles d' Association dans les Bases de Données.*Thèse de doctorat,* Univ. de Clermont-Ferrand II.

[Risden, 1999] Risden, K. (1999). Toward Usable Browse Hierarchies for the Web. Bullinger and Zieder (eds). *Human Computer Interaction : Ergonomics and User Interfaces*. 1 :1098-1102.

[Rodrigues et al., 1998] Rodrigues, JM., Trombert-Paviot, B., Baud, R., Wagner, J. and Meusinet-Carriot, F.(1998) GALEN-In-Use : using Artificial Intelligence Terminology Tools to Improve the Linguistic Coherence of a National Coding System for Surgical Procedures. Cesnik et al. (eds). *MedInfo' 1998*

[Sintek and Decker, 2001] Sintek, M. and Decker, S. (2001) TRIPLE- An RDF Query, Inference and Transformation Language. *Proceedings of Deductive Databases and Knowledge Management Workshop.*

[Sowa, 2000] Sowa, JF. (2000) Ontology, Metadata and Semiotics. *ICCS' 2000*

[Zweigenbaum et al., 2001] Zweigenbaum, P., Grabar, N., and Darmoni, SJ. (2001). Apport de Connaissances Morphologiques pour la Projection de Requêtes sur une Terminologie Normalisée. *Actes du Congrès Traitement Automatique du Langage Naturel.* p. 403-408.

[Zweigenbaum et al., 2003] Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Le Duff, F., Thirion, B. and Darmoni, SJ.(2003) Towards a Unified Medical Lexicon for French. *Medical Informatics Europe*. p 415-420.

Apport d'une ontologie du domaine pour affiner une requête à l'aide d'un treillis de Galois

B. SAFAR, H. KEFI

Université Paris-Sud,
CNRS (LRI) & INRIA (Futurs)
LRI, Bâtiment 490
91405 Orsay cedex, France
Email : {safar,kefi}@lri.fr
Tél : + 33 1 69 15 64 94 Fax : + 33 1 69 15 65 86

Résumé

Dans cet article, nous étudions comment aider un utilisateur qui effectue une recherche dans un entrepôt thématique à affiner sa requête quand celle-ci retourne trop de réponses. En utilisant une ontologie du domaine et un ensemble de ressources annotées avec les termes de cette ontologie, nous montrons comment utiliser un treillis de Galois pour élaborer, en interaction avec l'utilisateur, une requête plus précise qui réponde mieux à ses attentes.

Abstract

In this paper we study how to help user to refine his query when the search for documents in ressources has produced too many answers. Using a domain ontology and a set of ressources indexed with terms of this ontology, we show how we can use the immediate readability and intelligibility of the Galois Lattice clustering structure without paying the cost of its whole building. We also show how to build a refined query in interaction with the user.

1 Introduction

Quand il effectue une recherche d'information sur le web ou dans une base de données, l'utilisateur d'un moteur de recherche est bien souvent submergé par la masse de réponses retournées. Pour répondre à ce problème, de nombreux travaux ont portés sur le classement des réponses retournées en fonction de leur pertinence supposée ou sur leur regroupement en classes de réponses proches (clusters).

Dans cet article, nous étudions comment affiner la requête de l'utilisateur quand celle-ci obtient trop de réponses. Nous voulons l'aider à trouver les termes pertinents à rajouter à sa requête pour que celle-ci obtienne un nombre de réponses plus acceptable. Nous présentons pour cela OntoRefiner, un système qui crée des classes à partir des réponses obtenues et permet ensuite à l'utilisateur d'affiner ces classes de façon interactive. Ce système l'aide ainsi à se focaliser sur le sous-ensemble de réponses le plus pertinent et lui permet d'affiner sa requête initiale.

OntoRefiner est utilisé dans un entrepôt de données thématique. Le thème de l'entrepôt est décrit dans une ontologie du domaine composée d'une hiérarchie de termes ou mot-clés. L'entrepôt contient un ensemble de ressources (documents, services ou bases de données) relatives au domaine et chaque ressource est décrite par un ensemble de termes issus de l'ontologie. De même, l'utilisateur exprime ses requêtes au moteur de recherche associé à l'entrepôt en utilisant les termes de l'ontologie. Les ressources retournées par le moteur sont celles dont la description vérifie les termes de la requête posée.

Les entrées de l'algorithme de construction de classes sont l'ontologie du domaine, la requête posée et les descriptions des réponses retournées par le moteur. A partir de ces entrées, OntoRefiner construit des sous-ensembles grossiers et identifie dans chaque sous-ensemble des directions possibles de spécialisation. L'utilisateur peut alors choisir le sous-ensemble le plus intéressant ainsi qu'une direction de spécialisation. Le processus de construction de classes reprend sur le sous-ensemble choisi et se poursuit jusqu'à ce que l'utilisateur soit satisfait du résultat.

Le plan de l'article est le suivant. Nous présentons tout d'abord en section 2 l'ontologie du domaine et la façon dont nous l'utilisons pour faire apparaître de multiples points de vue dans une description de ressource. La méthode de classification par construction de treillis de Galois sur laquelle notre algorithme est basé et son intérêt pour l'affinement de requête sont présentés en section 3. Nous montrons ensuite en section 4 comment notre algorithme utilise l'ontologie du domaine pour ne construire que des fractions du treillis de Galois sans perdre son intérêt.

2 Ontologie et description des ressources

L'ontologie du domaine est décrite par une hiérarchie de noms de classe (noté $\mathcal{D}_h$) et représentée par un ensemble de règles de la forme $C_1 \rightarrow C_2$, où C_1 et C_2 sont des noms de classe ou termes. Par exemple dans le domaine des produits de tourisme, nous utilisons des termes comme AccomodationPlace, Localization, Equipment, Activities et la figure 1 montre un fragment de la hiérarchie décrivant le terme Localization selon différents points de vue, comme les aspects physiques ou géographiques.

Les deux définitions suivantes établissent la notion de généralisant d'un terme.

Definition 2.1: *Un* **généralisant direct** *d'un terme c est un terme c' tel que la règle c $\rightarrow$ c' appartient à $\mathcal{D}_h$. c est une spécialisation directe de c'.*

Un terme peut avoir plusieurs généralisants et plusieurs spécialisations.

Definition 2.2: *c' est un* **généralisant** *d'un terme c si c' est généralisant direct de c ou s'il existe c" tel que c' est un généralisant direct de c" et c" est un généralisant de c.*

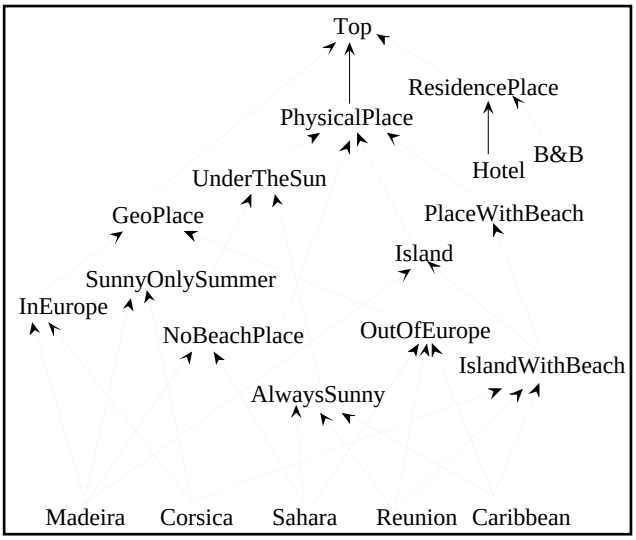


FIG. 1 – *Une fraction de la hiérarchie de termes*

Exemple 1 *Les généralisants directs de Reunion sont AlwaysSunny, OutOffEurope et IslandWithBeach alors que ses autres généralisants sont PlaceWithBeach, Island, GeoPlace, etc. Les spécialisations directes de UnderTheSun sont SunnyOnlySummer et AlwaysSunny et ses autres spécialisations sont Corsica, Sahara, etc.*

Les ressources sont supposées être indexées [1] par les termes les plus précis de l'ontologie du domaine qui les caractérisent. En utilisant la hiérarchie de la fi gure 1, une ressource portant sur un hôtel situé à la Réunion aura pour description l'ensemble de termes suivant {Hotel, Located, Reunion}.

Avant d'effectuer la classifi cation des ressources retournées, la description de chacune de ces ressources est enrichie, en utilisant la hiérarchie, de tous les généralisants des termes qui la composent. Ce mécanisme, appelé **saturation**, est défi ni comme suit :

Definition 2.3: *Sat(C) est le* **saturé** *de l'ensemble de termes C, si pour chaque terme c de C, tous les généralisants de c sont dans Sat(C).*

Exemple 2 *Avec la hiérarchie de la fig.1, pour C = {Hotel, Located, Reunion}, Sat(C) = {Hotel, ResidencePlace, Located, Reunion, AlwaysSunny, UnderTheSun, OutOfEurope, IslandWithBeach, PlaceWithBeach, Island, GeoPlace, PhysicalPlace}.*

Réciproquement, nous défi nissons la **désaturation** d'un ensemble C comme le processus qui retire de cet ensemble tous les généralisants dont les spécialisations sont dans C.

Definition 2.4: *DeSat(C) est le* **désaturé** *de l'ensemble de termes C, si pour chaque terme c de DeSat(C), c est dans C et il n'existe pas de c" dans C tel que c"soit une spécialisation de c.*

L'objectif de cette étape de saturation est de faire apparaître des éléments communs entre des descriptions qui pourraient sinon ne pas en avoir. Ainsi dans une comparaison terme à terme, l'intersection des deux descriptions suivantes {Madere} et {Réunion} est vide alors que travailler sur les saturés fait émerger les généralisants communs {UnderTheSun, Island}. La saturation permet aussi de faire apparaître des similarités à un niveau d'abstraction plus élevé qu'entre les descriptions initiales (en utilisant les termes les plus généraux des descriptions) et/ou selon les différents point de vue pertinents pour le thème que recouvre implicitement un terme et qui sont explicités dans l'ontologie. Remarquons en effet que l'ontologie est orientée : nous nous intéressons au domaine du tourisme, et

1. Nous ne traitons pas ici le problème de l'indexation du contenu de chaque ressource. Nous supposons que cette tâche a déjà été réalisée par ailleurs, automatiquement ou manuellement.

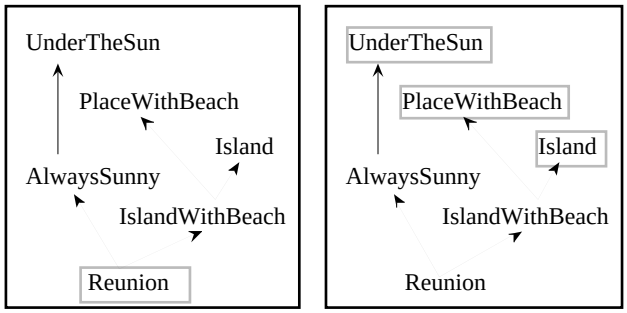


FIG. 2 – *fig.2a : Desat(C), fig.2b : mgt(C)*

les caractéristiques explicitées pour chaque terme sont celles qui sont pertinentes pour ce domaine, en terme de loisir (ici, pour les lieux, ce qui fait qu'on peut pratiquer ou pas des activités nautiques, se détendre au soleil, ...). Dans un autre contexte on pourrait s'intéresser aux points de vue historique, géologique, politique.

Dans la recherche des termes partagés par plusieurs descriptions, nous travaillerons donc sur les descriptions non désaturées et nous nous focaliserons sur les termes les plus généraux car ces termes ont plus de chances d'être communs à plusieurs descriptions.

Definition 2.5: *mgt(C) est l'ensemble des* **termes les plus généraux** *de l'ensemble de termes C si pour chaque terme t de mgt(C), t est dans C et il n'existe pas de t' dans C tel que t' soit un généralisant de t.*

Exemple 3 *Dans la figure 2, en utilisant la hiérarchie fig.1, pour C = {Reunion, AlwaysSunny, UnderTheSun, Island-WithBeach, PlaceWithBeach, Island}, DeSat(C) = {Reunion} et mgt(C) = {UnderTheSun, PlaceWithBeach, Island}. Les termes de DeSat(C) et de mgt(C) apparaissent respectivement encadrés dans les figures fig. 2a et fig. 2b.*

3 Méthode de construction de classes

Notre méthode de construction de classes est basée sur la notion de treillis de Galois [10], qui a déjà été utilisée pour faire du raffinement de requêtes [4, 3]. Etant donnés deux ensembles finis D (un ensemble de documents) et T (un ensemble de termes), et une relation binaire $\mathcal{R}$ entre ces deux ensembles, le treillis de Galois est un ensemble particulier de classes, dans lequel chaque classe est un couple, composé d'un sous-ensemble de documents $D' \subseteq D$, appelé **extension** du couple, et un sous-ensemble de termes $T' \subseteq T$, appelé **intention** du couple. Chaque couple (D', T') doit être un couple **complet** pour $\mathcal{R}$, ce qui signifie que T' doit seulement contenir les termes partagés par tous les documents de D', et symétriquement, les documents de D' doivent précisement être ceux qui partagent tous les termes de T'.

$T' = f_1(D')$ où $f_1(D') = \{t' \in T' \mid \forall d' \in D', d'\mathcal{R}t'\}$
$D' = f_2(T')$ où $f_2(T') = \{d' \in D' \mid \forall t' \in T', d'\mathcal{R}t'\}$

Exemple 4 *La figure 3 présente un exemple de 5 documents annotés par 7 termes, {a,b,c, d, e, f, g}. Le document 1 est annoté par les termes {a, c, f}. Le terme g est présent dans les descriptions des documents 2 et 3.*
f_1({1,4}) = {c, f} et f_2({c, f}) = {1,4, 5}. ({1,4},{c, f}) n'est donc pas un couple complet (puisque f_2({c, f}) $\neq$ {1,4}) alors que le couple ({2, 3},{a, g}) est complet car f_1({2, 3}) = {a, g} et f_2({a, g}) = {2, 3}.

Soient deux couples $C_1 = (D_1, T_1)$ et $C_2 = (D_2, T_2)$, une relation d'ordre partiel ($<$) est définie entre les couples par : $C_1 < C_2 \Leftrightarrow T_1 \subseteq T_2$ et $C_1 < C_2 \Leftrightarrow D_2 \subseteq D_1$.

Cet ordre partiel est utilisé pour générer le graphe du treillis de la façon suivante : il existe un arc (C_1, C_2) si $C_1 < C_2$ et s'il n'existe pas d'autre couple C_3 dans le treillis tel que $C_1 < C_3 < C_2$. C_1

	a	b	c	d	e	f	g
1	•		•			•	
2	•		•				•
3	•			•			•
4		•	•			•	
5		•	•		•	•	

FIG. 3 – *Un exemple de relation entre 5 documents et 7 termes*

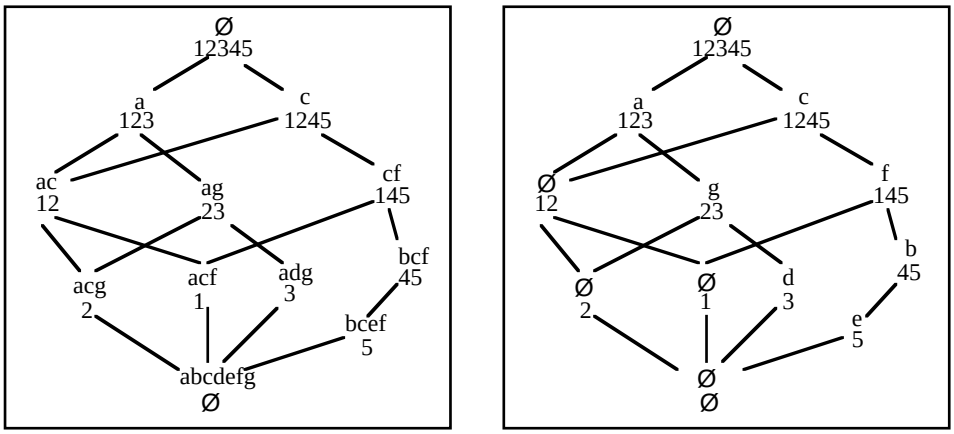


FIG. 4 – *Graphe du treillis de Galois et $T' - ICL$ correspondant*

est dit le **père** de C_2 et C_2 le **fils** de C_1. Le graphe révèle les relations de généralisation/spécialisation entre les couples où $C_1 < C_2$ représente le fait que C_1 est plus général que C_2.

Exemple 5 *Dans la figure 4a qui présente le graphe du treillis généré pour la relation de la figure 3, on voit que le couple ({a},{1,2,3}) est plus général que le couple ({a, g},{2,3}). Un ensemble de documents plus grand a moins de termes en commun qu'un ensemble plus restreint.*

Un treillis de Galois est une représentation qui comporte beaucoup de redondances. Pour un couple $C = (D', T')$, D' est présent dans chaque ancêtre de C et réciproquement, T' apparaît dans chacun de ses descendants. Godin et al. ont montré dans [5] que cette redondance peut être éliminée sans perdre d'informations si le graphe est maintenu. Pour un couple $C = (D', T')$, soit T'' l'ensemble des termes non redondants de T', $T'' = \{t' \in T' | t' \in f_1(D')$ et tel qu'il n'existe pas d'autre couple $C' = (X, Y) < C$ tel que $t' \in Y\}$.

Godin défi nit le $T'-$ *inheritance terme lattice* ($T' - ICL$) qui utilise les ensembles de couples (D', T''), et montre que pour un couple donné C la valeur correspondante de T' peut être calculée en faisant l'union des ensembles T'' des ancêtres de C y compris C lui-même. Les éléments de T' qui ne sont pas dans T'' sont ainsi hérités de ses ancêtres. Dans ce papier, nous utiliserons pour notre algorithme des couples de la forme (D', T'').

Exemple 6 *En comparant les figures 4a et 4b, on voit que le couple ({a, c},{1,2}) de la figure 4a n'a plus pour intention dans la figure 4b que ∅. Son intention peut être recalculée en prenant l'union des intentions de ses ancêtres, les couples ({a},{1,2,3}), ({c},{1,2,4,5}) et (∅, {1,2,3,4,5}).*

La structure des treillis de Galois est très intéressante dans le contexte de la construction de classes car le contenu d'une classe (D', T') est directement étiqueté par son intention T' qui est un simple ensemble de termes immédiatement lisible et compréhensible par l'utilisateur. De plus, cette structure de treillis autorise le recouvrement des classes : un document peut apparaître dans différentes classes s'il possède différents descripteurs.

Godin[4] et Carpineto[3] ont utilisé ce type de structure pour indexer une base de documents décrits par un ensemble de termes. Une fois construit le treillis entier caractérisant la totalité de la base, les requêtes des utilisateurs qui interrogent la base sont rapprochées des intentions des classes du treillis. En effet, pour chaque classe, son intention peut être vue comme une requête conjonctive et son extension, comme les documents répondant à cette requête. Répondre à une requête de l'utilisateur consiste donc, dans ce contexte, à rechercher la classe la plus générale dont l'intention contient tous les termes de la requête considérée. Une fois cette classe identifiée,[2] l'utilisateur peut naviguer dans le treillis et explorer les classes voisines : les fils de la classe correspondent à un pas de raffinement de la requête et ses parents à un pas d'élargissement.

Dans notre contexte où l'ensemble des documents (des ressources) à classer n'est pas un ensemble déterminé à l'avance mais est généré dynamiquement à partir de la requête de l'utilisateur, la construction du treillis complet serait un processus trop couteux. Pour l'éviter, nous ne construisons que les premiers nœuds du treillis (son premier niveau) et nous identifions dans chaque nœud des axes de spécialisations potentielles de ce nœud. Ainsi l'interaction avec l'utilisateur lui permet de choisir un nœud et un axe et de relancer la construction d'un nouveau niveau du treillis en n'utilisant que les documents apparaissant dans l'extension du nœud sélectionné. Remarquons que comme nous ne construisons donc pas réellement un treillis, et que certains documents ne sont plus pris en compte, nous autorisons l'utilisateur à revenir sur ses choix pour explorer d'autres nœuds.

4 Algorithme

Etant donnée une requête de l'utilisateur, les entrées de l'algorithme sont :

• l'ensemble D des identifiants des ressources dont la description s'apparie avec tous les termes de la requête,

• pour chaque identifiant d_i de D, un ensemble de termes appelé **description courante** Cdesc(d_i), qui est initialisé par le saturé de l'ensemble des termes décrivant la ressource.

Exemple 7 *Supposons que la requête de l'utilisateur soit Q = ResidencePlace. Supposons aussi (pour construire un tout petit exemple) que le moteur de recherche ne trouve que quatre ressources correspondant à cette requête et que les ensembles de termes décrivant ces ressources (ou descripteurs) soient respectivement :*

- d_1: {Hotel, Located, Reunion},
- d_2: {Hotel, Located, Carribean},
- d_3: {B&B, Located, Sahara},
- d_4: {B&B, Located, Corsica}.

Après calcul du saturé de l'ensemble des termes décrivant la ressource d_1, sa description courante est alors initialisée comme suit : Cdesc(d_1) = {Hotel, ResidencePlace, Located, Reunion, AlwaysSunny, UnderTheSun, OutOfEurope, IslandWithBeach, PlaceWithBeach, Island, GeoPlace, PhysicalPlace}.

La sortie de l'algorithme est une spécialisation de la requête initiale.

L'algorithme fonctionne comme suit : à chaque itération, (chaque niveau du treillis), il réalise trois tâches :

1- la construction de la classe racine du niveau,

2- la construction des classes du niveau,

3- l'identification des axes d'affinement possibles de chaque classe,

puis il passe la main à l'utilisateur qui choisit une classe, puis décide soit d'itérer le processus en sélectionnant un axe de spécialisation soit de l'arrêter.

2. La classe recherchée peut ne pas exister si la requête posée n'a pas de réponse dans la base. Le problème des réactions coopératives à des réponses vides n'est pas étudié ici et le lecteur intéressé par ce problème pourra se référer à [3], [2] ou [8].

4.1 Construction de la classe racine du niveau

Les descripteurs partagés par toutes les ressources de D doivent être identifiés comme étant l'intention de la classe racine du niveau. En effet, les ressources apparaissant dans D sont celles retournées par le moteur de recherche, donc les descriptions courantes de chacune de ces ressources partagent au moins les descripteurs qui vérifient les termes de la requête de l'utilisateur. Dans les itérations suivantes, un certain nombre de descripteurs aussi seront communs à toutes les ressources considérées. Cet ensemble de descripteurs communs est retiré de la description courante de chaque ressource et, une fois désaturé, cet ensemble est conservé comme intention de la classe dont l'extension est l'ensemble de ressources considérées dans le niveau. Ainsi, nous éliminons les termes redondants de T' comme dans les T'-ICL de Godin [5].

Exemple 8 *Les descriptions courantes des quatres ressources de l'exemple partagent les descripteurs suivants : {ResidencePlace, Located, UnderTheSun, GeoPlace, PhysicalPlace}. La classe racine construite avec le désaturé de cet ensemble est (1234, {ResidencePlace, Located, UnderTheSun}) puisque* GeoPlace *et* PhysicalPlace *sont des généralisants de* UnderTheSun.

4.2 Construction des classes du niveau

Dans cette étape, nous travaillons sur les descriptions courantes de chaque ressource dont les termes partagés identifiés à l'étape précédente ont été retirés.

Nous commencons par identifier l'ensemble des **termes les plus généraux** du niveau, Mgt, qui est l'union des termes les plus généraux apparaissant dans la description courante de chaque ressource du niveau, Mgt = $\bigcup_{i\in[1\ldots n]}$ mgt(Cdesc(d_i)).

Exemple 9 *Pour la description courante de d_1, Cdesc(d_1) = {Hotel, Reunion, AlwaysSunny, OutOfEurope, IslandWithBeach, PlaceWithBeach, Island}, mgt(Cdesc(d_1)) = {Hotel, AlwaysSunny, OutOfEurope, PlaceWithBeach, Island}.*

Pour les quatre ressources considérées, Mgt = {Hotel, B&B, AlwaysSunny, SunnyOnlySummer, OutOfEurope, InEurope, PlaceWithBeach, NoBeachPlace, Island}.

Puis, pour chaque terme t de Mgt, l'algorithme construit le couple complet (D', T') centré autour de t, i.e., tel que les ressources de D' sont celles dont la description courante contient t et tel que T' est l'ensemble des termes communs à toutes les descriptions courantes des ressources de D'. Une fois construits les différents couples correspondant à chaque t de Mgt, nous ne conservons que les couples les plus généraux, i.e., ceux dont l'extension n'est pas incluse dans l'extension d'un autre couple complet.

Exemple 10 *Le couple complet centré autour de AlwaysSunny est : (123, {AlwaysSunny, OutOfEurope}).*
Il est plus général que celui centré autour de Hotel, qui est : (12, {Hotel, AlwaysSunny, OutOfEurope, PlaceWithBeach}).
Seul le premier couple est conservé.

Les couples les plus généraux obtenus ici sont les mêmes que ceux calculés avec un algorithme classique de construction descendante de treillis de Gallois, mais ils sont moins couteux à obtenir.

Un algorithme classique descendant construit les couples complets en calculant la fermeture de chaque terme apparaissant dans une description courante, alors que nous ne calculons que la fermeture de chaque $t \in$ Mgt. Supposons qu'il existe un terme t tel que $t \notin$ Mgt et qu'il existe un couple complet (X, Y) centré autour de t qui soit construit par l'algorithme classique. Si $t \notin$ Mgt cela signifie que dans chaque description où t apparaît il existe un terme t' tel que t' est un généralisant de t et $t' \in$ Mgt. Toute les descriptions des ressources de X contenant t, elles contiennent aussi t'. Cela signifie que $t' \in Y$ et que le couple complet (X, Y) a aussi été construit par notre algorithme.

L'intention de chacun des couples les plus généraux construits sera présentée à l'utilisateur, lui permettant ainsi de choisir la classe qui lui semblera la plus pertinente.

Exemple 11 *Trois classes sont finalement conservées à cette étape pour être présentées à l'utilisateur :*
- (34, {B&B})
- (123, {AlwaysSunny(y), OutOfEurope})
- (124, {PlaceWithBeach, Island, IslandWithBeach})

4.3 Identification des axes possibles d'affinement des classes

Pour aider l'utilisateur à affiner la classe qu'il choisira, nous voulons lui proposer, associé à chaque classe, un ensemble d'axes possibles d'affinement, c.à.d, un ensemble de termes parmi lesquels il pourra choisir celui qu'il veut voir apparaître dans la future requête. Dans chaque classe, l'ensemble des descripteurs qui seront proposés comme axes sont sélectionnés en recherchant à nouveau les termes les plus généraux parmi les descripteurs des ressources de la classe, une fois retirés bien sûr, les descripteurs apparaissant dans l'intention de la classe. L'identification des axes associés à chaque classe implique donc tout d'abord de mettre à jour, dans chaque classe, les descriptions courantes de chaque ressource.

Dans ce but, dans chaque classe, des copies des descriptions courantes de chaque ressource de la classe sont tout d'abord créées, liées à la classe et mises à jour : les descripteurs apparaissant dans l'intention de la classe sont retirés de la description courante de chaque ressource.

Exemple 12 *A l'étape précédente, la description courante de d_1 était Cdesc(d_1) = {Hotel, Reunion, AlwaysSunny, OutOfEurope, IslandWithBeach, PlaceWithBeach, Island}. Après copie et mise à jour dans les 2 classes où cette ressource apparaît, ses nouvelles descriptions courantes sont maintenant :*
- pour la classe (123, {AlwaysSunny, OutOfEurope}),
Cdesc(d_1) = {Hotel, Reunion, IslandWithBeach, PlaceWithBeach, Island}.
- pour la classe (124, {PlaceWithBeach, Island, IslandWithBeach}),
Cdesc(d_1) = {Hotel, Reunion, AlwaysSunny, OutOfEurope}).

Puis, à nouveau, nous recherchons l'ensemble des termes les plus généraux issus de chaque description courante Cdesc(d'_i), Mgt = $\bigcup_{i\in[1\ldots n]}$ mgt(Cdesc(d'_i)), et ce, dans chaque classe.
Le Mgt de chaque classe est présenté, avec celle-ci, comme l'ensemble des axes de spécialisation possibles de la classe. Dans l'affichage à l'utilisateur, chaque axe de spécialisation est suivi par le nombre de ressources de la classe qui le vérifient.

Exemple 13 *Pour la classe (123, {AlwaysSunny, OutOfEurope}), mgt(Cdesc(d_1)) = mgt(Cdesc(d_2)) ={Hotel, PlaceWithBeach, Island} et mgt(Cdesc(d_3)) = {B&B, NoBeachPlace}.*
Le Mgt de la classe est {Hotel, B&B, PlaceWithBeach, Island, NoBeachPlace} et l'ensemble des axes affiché est {Hotel 2, B&B 1, PlaceWithBeach 2, Island 2, NoBeachPlace 1}.

4.4 Interaction avec l'utilisateur

A ce stade, le système a affiché les différentes classes et leurs axes de spécialisation possibles. L'utilisateur doit choisir une classe (D', T'), puis il peut soit :

• choisir un axe de spécialisation t parmi les termes proposés avec la classe sélectionnée. L'algorithme identifie alors le sous-ensemble de ressources de D' dont la description courante contient ce terme t, et réitère l'algorithme en reprenant à la tâche 1, avec ce sous-ensemble de ressources.

Exemple 14 *Supposons que l'utilisateur choisisse la classe (123, {AlwaysSunny, OutOfEurope}) et le terme* PlaceWithBeach *comme axe d'affinement. Parmi les 3 ressources de la classe, seules les ressources 1 et 2 vérifient ce terme, et c'est sur elles et leur description courante actuelle que l'algorithme reprend à l'étape 1.*

• signifier que la classe sélectionnée est satisfaisante. La requête affinée correspondant à cette classe est alors construite en faisant l'union des intentions des classes successivement choisies par l'utilisateur, puis en calculant le désaturé de l'ensemble de descripteurs ainsi constitué.

Exemple 15 *Supposons ici aussi que l'utilisateur choisisse la classe (123, {AlwaysSunny, OutOfEurope}) mais qu'il dise que cela lui convient.*
La requête affinée est composée de l'union des ensembles suivants : {ResidencePlace, Located, UnderTheSun} qui est l'intention de la classe racine, calculée dans l'exemple 8 et {AlwaysSunny, OutOfEurope} qui est celle de la classe qui vient d'être choisie. Après désaturation, la requête résultante est {ResidencePlace, Located, AlwaysSunny, OutOfEurope}.

Bien évidemment si les classes construites lui déplaisent, l'utilisateur peut toujours remonter à un choix précédent dans l'arbre construit et explorer une autre branche. Le mécanisme proposé lui permet ainsi d'effectuer pas à pas les étapes d'affinement comme il aurait pu le faire en naviguant dans un vrai treillis de Galois : en effet, pour une classe, chaque mgt proposé comme axe de spécialisation correspond à une des caractéristiques que présenteraient certains des fils de cette classe dans le vrai treillis de Galois. Mais le temps de construction de notre arbre est moindre ou du moins fractionné si l'utilisateur veut finalement explorer toutes les branches de l'arbre.

5 Travaux proches et conclusion

De nombreux travaux se sont déjà attaqué au problème de l'organisation des documents retournés par un moteur de recherche en réponse à une requête. Pour éviter d'imposer à l'utilisateur de parcourir une longue liste de documents classés selon leur pertinence supposée, certains chercheurs ont déjà proposé de regrouper ces documents en classes cohérentes par des techniques dites de **classifications éphémères** [7, 11]. Les exigences de ce type de classification ont été définies par Maarek et al.[7] : il nécessite d'une part une grande précision car l'utilisateur qui n'est la plupart du temps pas un expert d'un domaine est peu tolérant aux erreurs et d'autre part une forme de présentation qui lui permette de parcourir aisément les classes, avec des techniques de visualisation et/ou d'étiquettage automatique des classes construites.

Pour augmenter la précision, Maarek et al. utilisent une méthode de Construction Hiérarchique par Agglomération (HAC) et définissent une mesure de similarité entre documents basée sur une unité d'indexation qui consiste en des paires de mots reliés par affinité lexicale. Pour améliorer l'étiquettage automatique des classes construites et pour en présenter une description concise et précise, Zamir et Etzioni [11] effectuent leur classification en construisant un arbre de suffixes extraits des phrases issues des résumés associés aux documents retournés par le moteur de recherche. Mais ces travaux n'acceptent pas les réactions de l'utilisateur et ne l'aident pas à reformuler sa requête initiale.

Pour augmenter la recherche interactive et prendre en compte les retours de l'utilisateur Leuski et Allan [6] combinent liste ordonnée et construction de classes : les documents sont ordonnés suivant leur probable pertinence et l'utilisateur est supposé suivre cet ordre jusqu'à ce qu'il trouve le premier document réellement intéressant. Ce premier document sert alors à initialiser la construction de classes avec les documents non encore examinés.

Berenci et al. aident l'utilisateur à se focaliser sur les termes pertinents et à reformuler sa requête en visualisant les documents retournés regroupés par paquets suivants les différents sous-ensembles de termes de la requête initiale qu'ils vérifient [1].

Les travaux basés sur les treillis de Galois [4, 3] présentés en section 3 sont les plus proches des notres. En effet cette méthode ne nécessite ni l'utilisation de mesure de similarité comme [7, 6] ni l'analyse de phrases complètes comme [11]. Ses avantages sont de deux sortes. D'une part, l'étiquettage d'une classe est fait automatiquement : c'est l'intention de la classe et elle recouvre l'ensemble

des termes partagés par tous les documents de la classe. D'autre part, le treillis permet d'explorer facilement les spécialisations d'une requête en explorant les voisins de la classe représentant la requête dans le treillis. Le principal problème de cette méthode de construction de classe est son coût : elle ne peut être appliquée en temps réel sur de trop gros ensemble de documents.

Dans notre contexte d'entrepôt thématique, le nombre de documents retournés, même s'il est trop important pour être géré facilement par l'utilisateur, reste raisonnable. De plus, notre algorithme ne nécessite pas de construire le treillis dans sa totalité : cette méthode de construction de classes retrouve ainsi toutes ses qualités. Un des aspects importants de notre travail est l'utilisation d'une ontologie du domaine dans laquelle chaque terme a de multiples généralisants et spécialisations. L'usage de cette ontologie permet d'enrichir la description d'une ressource par de multiples points de vue qui sont utilisés dans la phase de construction des classes et augmentent ainsi sa précision. L'algorithme présenté ici est implémenté en Java et actuellement expérimenté dans le médiateur PICSEL [9] dans le cadre d'un contrat avec France-Télécom R&D.

Références

[1] Ezio Berenci, Claudio Carpineto and Vittorio Giannini. Improving The Effectiveness Of Web Search Engines Using Selectable Views Of Retrieval Results. *Journal of Universal Computer Science*, 4, 737-747, 1998.

[2] A. Bidault, Ch. Froidevaux and B. Safar, Repairing Queries in a Mediator approach. In *Proc. of ECAI 2000*, pp. 406-410, Berlin, August 2000.

[3] Claudio Carpineto and Giovani Romano. Information Retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45, 553-578, 1996.

[4] Robert Godin, Rokia Missaoui and Alain April. Experimental comparison of navigation in a Gallois lattice with conventional information retrieval methods. *International Journal of Man-Machine Studies*, 38, 747-767, 1993.

[5] Robert Godin, Rokia Missaoui and H. Alaoui. Incremental Concept Formation Algotithms Based on Galois (Concept) Lattices. *Computational Intelligence*, 11(2), 246-267, 1995.

[6] Anton Leuski and James Allan. Improving Interactive Retrieval by Combining Ranked Lists and Clustering. In *Proceedings of RIAO'2000*. 665-681, April 2000.

[7] Yoelle Maarek, Ronald Fagin, Israel Ben-Shaul, and Dan Pelleg. Ephemeral Document Clustering for Web Applications. IBM Research Report RJ 10186, April 2000.

[8] T. Gaasterland, P. Godfrey, and J. Minker. An overview of Cooperative Answering. *Journal of Intelligent Information Systems*, vol.1, pp. 123-157, 1992.

[9] M.-Ch. Rousset, A. Bidault, Ch. Froidevaux, H. Gagliardi, F. Goasdoué, C. Reynaud et B. Safar, Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL, Dans *Revue I3*, Vol 2,(1), pp. 9-59, Septembre 2002.

[10] R. Wille. Concept Lattices and Conceptual Knowledge Systems. *Computers and Mathematics with Applications* 23, 403-515, 1992.

[11] Oren Zamir and Oren Etzioni. Web document Clustering: A Feasibility Demonstration. In *Proceedings of SIGIR'98*, 46-54, Melbourne, August 1998.

L'interdisciplinarité et la terminologie mathématique: les termes migrateurs

A. TOMA

Département de linguistique

A 314, Place de l'Université 3,

1211 Genève 4, SUISSE

Mail: toma1@etu.unige.ch

Tél: +41 7 28 73 19 / +41 22 705 73 33

Résumé

Les recherches scientifiques actuelles font intervenir plusieurs disciplines pour étudier un même objet afin d'obtenir les résultats souhaités. L'interface linguistique de cette réalité scientifique est le transfert des termes d'un domaine à l'autre. On reconnaît d'amblé, d'une part, l'importance de la communication « sans reste » des connaissances, d'autre part, le fait que le discours mathématique est capable, par sa forme, d'interdire le refus de son contenu. De ce fait, on se pose la question si le terme mathématique a un accès particulier au phénomène de migration.

Abstract

Currently scientific researches make use of several disciplines to study the same object in order to obtain the desired results. The linguistic interface of the scientific reality is the transfer of the terms from one field to another. It is well known among the scientists that it is important that, on the one hand, a scientific communication is done "without remainder" ok knowledge and, on the other hand, that the very form of mathematical discourse prevents the content from being rejected. It is probably why mathematical terms migrate easily from one field of knowledge to another.

1 Introduction

1.1 Deux modèles d'étude des termes interdisciplinaires

Louis Guilbert étudie la naissance du vocabulaire de l'aviation, en suivant le processus même de la constitution de ce nouveau domaine scientifique: « L'activité nouvelle se constitue par un double processus: d'une part, des notions scientifiques se dissocient de leur sphère primitive pour s'intégrer dans le champ de la nouvelle science; d'autre part, les notions théoriques tendent à se traduire en applications pratiques et à passer au stade de la technique. Un système de vocabulaire se crée pour exprimer l'évolution en cours sur le plan des signifiés; il prend corps par le passage d'unités de signification du vocabulaire scientifique proprement dit au vocabulaire technique de l'aéronautique ou de l'aviation » [Guilbert, 1965: 140].

La situation extralinguistique, un corps plus lourd que l'air muni de sa force motrice suspendu dans l'atmosphère par la vertu de son mouvement, se traduit en plan proprement linguistique « par le transfert d'un certain nombre d'unités de signification dans le champ sémantique en voie de formation »[Guilbert, 1965: 140]. Étant donné cette situation extralinguistique, certains domaines sont sources favorables au transfert des termes (l'aéronautique, la physique, l'astronomie, les mathématiques et la mécanique). Mais « il apparaît que les notions constitutives ne sont pas groupées en autant de champs notionnels qu'il y a de sciences distinctes et que l'on n'assiste pas à une juxtaposition systématique et ordonnée d'autant de groupes d'éléments constitutifs dans le domaine de l'aéronautique; par contre-coup, les signes correspondants à ces notions ne s'intègrent pas dans le champ sémantique de l'aéronautique par strates nettement différenciées.»[Guilbert, 1965: 172].

Le nouveau domaine, un domaine technique, est constitué par l'application pratique des autres domaines, des domaines scientifiques. Le vocabulaire de ce nouveau domaine est considéré formé quand il est stable, c'est-à-dire, il est constamment utilisé à l'intérieur d'un groupe social. Au niveau linguistique, le passage du niveau scientifique au niveau technique est reflété par la préférence du nouveau vocabulaire pour le « syntagme complexe autonome » [Guilbert, 1965: 173]. Son élément de base exprime les notions théoriques qui reçoivent une expansion adjectivale ou nominale pour exprimer les données pratiques. Cet élément de base est considéré comme un « néologisme d'emploi ou néologisme sémantique » [Guilbert, 1965: 197]. Le second élément est l'instrument linguistique principal du transfert.

Les études plus récentes [Losee 1995] analysent en termes de fréquence le rapport entre les domaines[1] « source » et les domaines « cible » des termes migrateurs. Le terme apte de transfert est celui qui décrit un phénomène général susceptible de transgresser son domaine d'origine. Le plus souvent le domaine « source » est une science « dure », tandis que le domaine « cible » est une science « faible ». La voie de passage d'un domaine à l'autre est soit le scientiste qui passe d'un domaine à l'autre, soit le scientiste qui, étant compétent dans deux domaines, utilise les moyens d'un premier pour résoudre les problèmes de l'autre.

L'étude de la fréquence des termes à l'intérieur d'une même sous-langage montre l'aspect « original » ou « d'emprunt » d'un terme. Une fréquence élevée d'un terme montre sa spécificité pour un domaine, tandis que son fréquence basse correspond à une utilisation non-spécifique, fait qui indique que le terme est emprunté, le plus souvent avec une modification de sens.

Selon la force de la fréquence de premiers dix termes, Losee établit les huit premiers domaines: l'électronique, la biologie, la physique, la psychologie, les mathématiques, l'économie, la sociologie et l'histoire. Elles sont censées être disciplines - « source ». L'étude chronologique de la fréquence des termes peut fournir l'image de l'évolution de la migration d'un terme (l'apparition, le développement, la disparition), mais aussi l'image de l'évolution du terme même.

[1] "An academic *discipline* or *field* is a large group of individuals within academia or the professions who are working on a broad range of related research or professional problems. Those within these fields use a *sublanguage*, incorporating the general language of the larger society, as well as a grammer and vocabulary used in a discipline specific manner." [Guilbert, 1965: 267).

L'étude de la migration et l'identification des domaines « source » pour un certain domaine « cible » peut être bénéfique pour l'orientation des scientifiques vers une formation qui leur permet de maîtriser le transfert des termes.

1.2 Une étude statique des termes migrateurs

Il existe – d'une parte – des champs d'étude qui font l'objet de la recherche de plusieurs domaines scientifiques et qui permettent l'apparition des termes scientifiques interdisciplinaires. D'autre part, les sciences, pour perfectionner leurs instruments d'investigation, s'approprient des méthodologies qui appartiennent à d'autres sciences. Dans le cas des mathématiques il s'agit de ce qu'on appelle la *mathématisation* des sciences (par exemple, des sciences du langage). Voilà deux voies de migration de termes scientifiques. On va examiner les termes scientifiques interdisciplinaires ayant comme trait commun le fait que tous appartiennent au domaine des mathématiques[2]. On va établir quels sont les domaines préférés par les termes mathématiques et quels sont les domaines dis-préférés par ceux-ci.

On va distinguer les contacts profonds entre les domaines, qui se reflètent au niveau du lexique dans l'existence des termes interdisciplinaires mono- sémantiques (phénomène qu'on va appeler *interdisciplinarité* ou *migration totale*) et les contacts superficiels entre différents domaines, fait qui renvoie aux termes dont la caractérisation de mono- sémantique est difficile à défendre (phénomène qu'on va appeler *interférence* ou *migration partielle*)[3]. On se propose aussi de relever les phénomènes linguistiques qui accompagnent le transfert d'un terme d'un domaine à l'autre (la sémantique et la morphologies des termes). L'article s'appuie sur un corpus résulté de la consultation d'un dictionnaire de spécialité et de plusieurs dictionnaires généraux. Il ne s'intéresse pas de la direction de la migration des termes, mais il se contente d'analyser les propriétés des termes mathématiques présents dans plusieurs domaines.

On va présenter, tour à tour, le corpus ; les domaines que les termes mathématiques traversent et, finalement, les particularités sémantiques et co-textuelles des termes migrateurs.

2 Les termes migrateurs et les domaines de leur circulation

Le corpus est construit suite à la consultation des dictionnaires généraux dans lesquels on repère les indicateurs du domaine qui sont soit explicitement présents, soit récupérés à partir du sens (définition lexicographique) enregistré dans l'ouvrage consulté.[4] Le mot accepté dans la base de données à analyser doit appartenir au moins à deux domaines ; ou, autrement dit, une entré lexicale constitue un terme interdisciplinaire ou migrateur si et seulement si le mot comprend un sens à double indicateur de domaine (enregistré ou récupéré). La double ou multiple appartenance est le résultat du passage d'un terme d'un domaine vers l'autre ; on précise, juste en passant, que le mouvement est orienté principalement en fonction de divers facteurs de nature extra- linguistique parmi lesquels l'intérêt d'une certaine discipline pour un objet d'étude d'une autre discipline ou le transfère méthodologique d'un domaine à l'autre. Mais il existe aussi des facteurs de nature linguistiques qui peuvent déterminer la multiplication d'usage lexicale des termes mathématiques, comme : la précision du sens (mono-sémantisme), la sémiotique iconique associée à un terme mathématique. Si les facteurs d'environnement cognitif sont chronologiques ou successifs, les facteurs linguistiques impliquent des chaînes causales qu'on peut décrire en amont du phénomène étudié ou envisager en aval de celui-ci.

Le phénomène de migration des termes fait partie du phénomène plus général d'hétérogénéité et de transversalité du discours scientifique « circulant »[5] qui convient aux hypothèses initiales de Michel Foucault, qui écrivait : « Au lieu d'être une chose dite une fois pour toute […] l'énoncé, en même temps qu'il surgit dans sa matérialité, apparaît avec un statut, entre dans des réseaux, se place dans des

[2] La raison pour laquelle on a choisit les mathématiques est double : d'une part, parce qu'elle s'avèrent très productives pour l'étude qu'on se propose, d'autre part, parce qu'elles représentent une discipline qu'on maîtrise bien.
[3] v. Bidu-Vrănceanu, Angela; Toma, Alice (2001) – *Lexic ştiintific interdisciplinar*, EUB, Bucureşti
[4] Moingeon, Marc; Berthelot, Jacques (dir.) (1990). *Le dictionnaire ne notre temps*, Hachette, Paris ; Péchoin, Daniel ; Demay, François (1994). *Le petit Larousse*, Larousse, Paris.
[5] v. Moirand, Sophie (2001). *Les politiques médias et les discours de spécialité*, Cediscor, Paris.

champs d'utilisation, s'offre à des transferts, à des modifications possibles, d'intègre à des opérations et à des stratégies où son identité se maintient ou s'efface. » [Foucault, 1969 : 138][6].

En particulier, par le fait qu'on part dans notre analyse des termes mathématiques, le phénomène de migration est d'autant plus intéressant parce qu'il contredirait l'intuition conforme à laquelle les mathématiques, sont considérées, en général, comme science « fermée, ésotérique » [Candel, 1998 :46], comme science moins ouverte à la vulgarisation : « La communauté scientifique mathématique pourrait constituer un bon exemple de communauté à faible impact discursif externe » [Beacco, 2001 :20]. Notre analyse de l'impact terminologique interdisciplinaire des termes mathématiques se réalise en deux temps : dans un premier temps, on retient toutes les entrées lexicales qui contiennent plusieurs indicateurs de domaines dans leurs définitions, entrées susceptibles d'être termes interdisciplinaires (migrateurs) ; dans un deuxième temps, on distingue les termes vraiment migrateurs des termes appelés partiellement migrateurs.

L'interdisciplinarité des termes mathématiques – dans un premier temps – s'avère riche aussi par le nombre des termes interdisciplinaires, que par le nombre des domaines dans lesquels on retrouve des termes mathématiques.

Domaine[7] est le concept par rapport auquel on définit le terme interdisciplinaire ou migrateur. Dans un sens large, *domaine* représente le champ du savoir humain ; en fonction de l'objet de connaissance qui a ou n'a pas de relation de continuité avec l'homme en tant que corps, en tant qu'objet de la nature, il existe la distinction entre deux *domaines* : les sciences de la nature et les sciences exactes, d'un côté, et les sciences humaines, de l'autre. Dans un sens restreint, un *domaine* constitue une discipline ou une science particulière (caractérisée par un objet et une méthodologie spécifique d'étude) et son lexique. Cette dernière acception est retenue dans cette étude ; le domaine lui- même est parfois divisé en sous-domaines ; on utilise la délimitation en domaines offerte dans les ouvrages lexicographiques par le biais d'indications pour les langues de spécialité auxquelles on apporte quelques précisions. Les indicateurs de domaines précisent le champ de l'usage d'un terme, mais ils mélangent les domaines et les sous- domaines et le phénomène de migration, qui présuppose la transgression d'une limite de domaine, est estompé, il devient difficile à saisir. Ainsi, dans le cas des indicateurs qui correspondent aux domaines strictement délimités, la migration d'un terme est facile à démontrée par l'intermède de la lecture de ces indicateurs divers qui accompagnent un même terme ; dans le cas d'indicateurs qui permettent l'inclusion d'un domaine dans l'autre ou l'intersection des domaines différents, le phénomène de migration devient difficile à saisir à partir de l'information des indicateurs lexicographiques et, par conséquent, l'interprétation de ceux-ci doit se réaliser à l'aide d'une information supplémentaire tirée de l'analyse du sens du terme. Il faut établir s'il est le cas d'une migration d'un sous- domaine à l'autre à l'intérieur du domaine auquel l'indicateur lexicographique correspond ou s'il n'est que le cas d'une information moins forte que l'ouvrage lexicographique nous donne pour le terme analysé dont l'appartenance à un sous domaine pourrait être strictement précisée.

[6] Foucault, Michael (1969). *L'archéologie du savoir*, Gallimard, Paris.

[7] v. Rastier, François; Cavazza, Marc; Abéillé, Anne (1994). *Sémantique pour l'analyse de la linguistique à l'informatique*, Masson, Paris, p. 61-64. Rastier considère le *domaine* un type de classe lexicale située, en descendant, après le *taxème*, mais avant le *champ* et la *dimension.* « Chaque domaine est lié à un certaine type de pratique sociale déterminée. Les indicateurs lexicographiques, comme *chim* (chimie), ou *mar* (marine) sont des indicateurs de domaine ». Il y a deux teste pour identifier un domaine : 1) dans un domaine la polysémie lexicale est absente ; 2) entre les unités d'un même domaine il n'y a pas de lien métaphorique. v. aussi F. Mazière (1981 – II) – L*e dictionnaire et les termes*, in « Cahiers de lexicologie », vol. XXXIX , p. 79-101.

En effet, la délimitation des domaines est très importante pour décrire ensuite la migration des termes, leur interdisciplinarité. L'organigramme suivant nous montre la répartition aléatoire des indicateurs lexicographiques par rapport aux domaines :

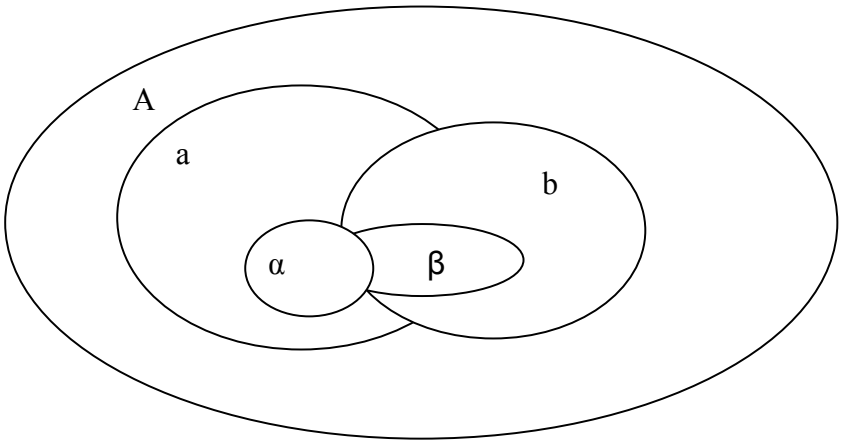


Fig. 1

Ainsi, si *A* est le domaine du savoir ; *a, b, ...* représentent les domaines et *α, β, ...*sont des sous-domaines, les indicateurs « de domaine » se placent au niveau des ensembles *a, b, ...,* et aussi au niveau des ensembles *α, β, ...* Cette répartition dispersée des indicateurs empêche une bonne description de la migration, étant donné qu'un terme *x* est migrateur ou interdisciplinaire s'il respecte une des formules suivantes :

i) $\forall$ *a, b, ... domaines,* $a \neq b$, $\exists$ *x migrateur ssi* $M = a \cap b (\cap c ...) \neq \varnothing$ *et* $x \in M$;

Quels que soient deux ou plusieurs domaines distincts, un terme est migrateur s'il appartient à l'intersection d'au moins deux d'entre eux

ii) $\forall$ *a, b, ... domaines,* $a \neq b$ *et* $\forall \alpha, \beta, ...$ *sous- domaines,* $\alpha \neq \beta$, $\exists$ *x migrateur ssi* $N = a \cap \alpha (\cap b ...) \neq \varnothing$, $\{\alpha, \beta, ...\} \not\subset \{a, b, ...\}$ *et* $x \in N$.

Quels que soient deux ou plusieurs domaines ou sous- domaines distincts, un terme est migrateur s'il appartient à l'intersection d'au moins deux d'entre eux et si le sous- domaine n'est pas inclus dans le domaine d'intersection

Parfois, les indications de langue de spécialité comprennent le domaine et quelques sous-domaines : GRAM $\subset$ LING ou PHON $\subset$ LING (il manque ici des sous- domaines comme : sémantique, lexicologie, pragmatique, stylistique, etc.). Le même mélange de domaines et sous-domaines est enregistré dans les séries :

-ANAT – BIOL – BOT – EMBRYOL – GENET;

-COMPTA – ECON – FIN.

Si un terme comme *temps* a l'indicateur LING et comme l'on sait que la linguistique comprend plusieurs sous- domaines, une partie d'entre eux comptés comme indicateurs lexicographiques, on a de bonnes raisons de se demander si le comportement de ce terme à l'intérieur du domaine linguistique est ponctuel (pourquoi il n'a pas un indicateur de sous- domaine comme GRAM, présent d'ailleurs parmi les indicateurs ?) ou multiple, récursif d'un sous- domaine à l'autre. Évidemment, le terme linguistique *temps* n'a qu'un seul sous- domaine où il est utilisé, mais l'indicateur lexicographique permet quand même une telle question, d'autant plus si l'on trouve à côté de lui des termes dont le comportement répond positivement à la même question : *système, unité, valeur*.

Les domaines en tant que macro- contexte détermine le phénomène de migration (v. 2.).

3 Domaines scientifiques et migration des termes mathématiques

Il suffit de regarder la liste des domaines dans lesquels on rencontre des termes mathématiques pour se rendre compte de leur omniprésence. Pourquoi les mathématiques ? Parce que les mathématiques, en tant que discours, mais aussi en tant que méthode de recherche assurent l'exactitude des résultats et constituent un modèle pour les autres sciences. (N'oublions pas que ce fait n'est qu'une hypothèse.)

En décrivant la réalité, les sciences s'approprient des objets communs, mais la perspective d'étude est différente. Au niveau du lexique, cet aspect se reflète dans l'existence des termes dont la référence unique est reliée aux différents domaines ; autrement dit : les termes interdisciplinaires sont le résultat d'une double ou multiple connexion à la même référence :

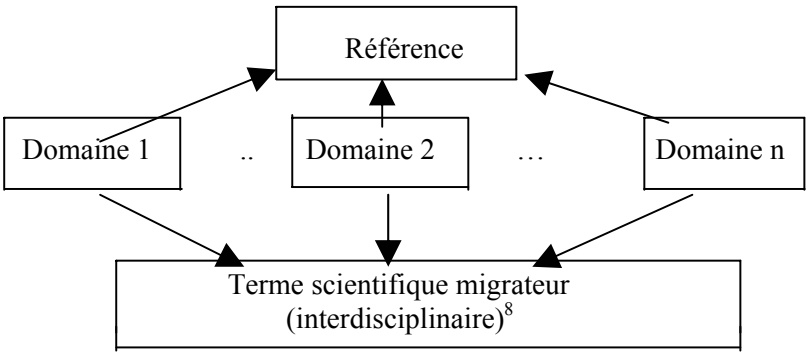


Fig. 2

soit s'il est le cas d'un terme dénominateur d'un objet, soit s'il est le cas d'un terme dénominateur d'un instrument d'étude. La relation linguistique – extra- linguistique est inhérente au niveau du lexique[9]. Si l'objet et les instruments font lieu commun des deux sciences (pour notre analyse, les mathématiques et un autre domaine), alors l'ensemble des termes communs est plus nombreux :

- physique – 80 termes communs avec les mathématiques ;
- technique – 46 termes communs avec les mathématiques ;
- économie – 30 termes communs avec les mathématiques ;
- astronomie – 28 termes communs avec les mathématiques ;
- logique – 27 termes communs avec les mathématiques ;
- chimie – 26 termes communs avec les mathématiques ;
- géographie – 26 termes communs avec les mathématiques ;
- médecine – 23 termes communs avec les mathématiques ;
- biologie – 22 termes communs avec les mathématiques ;
- droit – 20 termes communs avec les mathématiques ;
- politique – 18 termes communs avec les mathématiques ;
- militaire – 17 termes communs avec les mathématiques ;
- linguistique – 16 termes communs avec les mathématiques ;
- musique – 15 termes communs avec les mathématiques ;

[8] Il faut préciser que les deux concepts qu'on utilise, *interdisciplinaire* et *migrateur*, ont la signification générale, « qui appartient à deux ou plusieurs domaine" , mais la différence est donnée par l'orientation de la circulation des termes , l'aspect statique de l'appartenance des ter mes scientifiques aux différents domaines ou, respectivement, l'aspect dynamique du passage des termes d'un domaine à l'autre.

[9] Kleiber, Georges (1999). *Problèmes de sémantique. La polysémie en questions*, PUS, Septentrion.

- anatomie – 14 termes communs avec les mathématiques ;
- art – 14 termes communs avec les mathématiques ;
- géologie – 9 termes communs avec les mathématiques ;

Si ce qui vient dans l'intersection des deux sciences fait partie seulement du savoir commun de celles-ci (exceptant leur méthodologie), alors le nombre des termes communs est petit :

- biochimie – anaglyphe ;
- embryologie – disque ;
- finance – valeur ;
- génétique – condition ;
- histologie – disque ;
- industrie - disque ;
- médecine – analyse ;
- mécanique – dynamique ;
- phonétique – fréquence ;
- transport – adhérence ;
- urbanisme – construction ;
- zoologie – couronne .

Une explication linguistique du phénomène de la migration est donnée par la sous- catégorisation sémantique des termes mathématiques interdisciplinaires : *des objets concrets, des objets abstraits, des processus, des phénomènes, des états*. Les domaines qui combinent plusieurs sous- catégories sémantiques des termes, contiennent plusieurs éléments communs avec un autre domaine, tandis que pour les domaines qui prennent des termes appartenant à quelques sous- catégories sémantiques, les éléments de leur intersection sont moins nombreux. Le critère quantitatif (type de la catégorie) se combine avec le critère quantitatif (nombre des catégories) pour déterminer la richesse des termes de l'intersection des deux ou plusieurs domaines.

La sous- catégorie sémantique la plus productive pour les termes mathématiques interdisciplinaires est *l'objet concret* et, parmi les objets concrets, surtout ceux dont la sémiotique relève des aspects « dimensionnels ».

Le domaine dont l'intersection avec les mathématiques est la plus large est la physique. Les termes que les deux sciences possèdent en commun dénomment :

- des objets concrets : arc, axe, champ, corde, corps, parallélogramme, plan, prisme ;
- des objets abstraits : équations, énergie, facteur ;
- des processus et des phénomènes : accélération, amplification, analyse, oscillation, transformation ;
- des états : équilibre, stabilité, cohérence.

Le fait que les mathématiques et la physique ont beaucoup de termes communs s'explique au niveau extra- linguistique par un champ d'analyse commun et au niveau linguistique par la variété des sous- catégories sémantiques. Quant à l'origine des termes, l'accessibilité sémantique des termes permettrait d'affirmer qu'il y a des termes qui partent de la langue commune (ceux qui appartiennent à la première sous- catégorie sémantique), des termes qui ont comme origine les mathématiques (*les objets abstraits*) ou la physique (les termes des deux dernières sous- catégories sémantiques.

Les termes qui appartiennent en même temps aux mathématiques et à la linguistique ont une distribution pareille aux termes communs aux mathématiques et à la physique en classes sémantiques :

- des objets concrets : centre, champ, point, racine ;
- des objets abstrait : complément, conjonction, déterminant, hyperbole, unité ;
- des processus e t des phénomènes : analyse, relation ;
- des états : cohérence, transitivité.

Il parait que la sous- catégorie sémantique est plus forte comme facteur qui détermine la richesse d'une intersection entre deux domaines par rapport au type de domaines concernés.

Les deux premières classes sémantiques de termes se retrouvent aussi dans les domaines qui contiennent un nombre moins nombreux de termes communs avec les mathématiques : géographie (*axe, centre, cercle, champ, facteur,forme*) ou musique (*forme, imagine, espace, temps, valeur*).

Quand l'intersection des deux domaines se réduit à un seul élément, celui-ci peut être placé arbitrairement dans une des classes sémantiques mentionnées : *des objets concrets (disque, cylindre, couronne, corde) ; des objets abstraits (condition) ; des processus, des phénomènes, des états (analyse, adhérence, dispersion)*.

4 Macro- interdisciplinarité et micro- interdisciplinarité

Il y a deux façons de concevoir l'interdisciplinarité : soit à partir d'un domaine pour lequel on cherche les contacts avec d'autre domaines (*macro- interdisciplinarité* ou *migration globale*), soit à partir d'un termes pour lequel on regarde le parcours d'un domaine à l'autre(*micro- interdisciplinarité* ou *migration locale*).

En analysant des termes qui passent d'un domaine à l'autre on obtient une échelle qui contient les termes à partir des ceux les plus migrateurs jusqu'aux ceux moins migrateurs :

- 15 domaines : corps ;
- 14 domaines : analyse, centre ;
- 10 domaines : base, courbe ;
- 9 domaines : disque, forme ;
- 8 domaines : cône, contrôle, couronne, unité ;
- 7 domaines : aire, cercle, condition, valeur ;
- 6 domaines : cylindre, champ, classe, corde, direction, distance, pyramide, place, vitesse ;
- 5 domaines : équilibre, élément, facteur, fluxe, image, espace ;
- 4 domaines : absorption, axe, code, complément, construction, continuité ;
- 3 domaines : amplification, anaglyphe, arc, argument, conclusion, conjonction ;
- 2 domaines : affixe, algorithme, axiome, calcul, cohérence.

Du point de vue sous- catégoriel sémantique les dix premiers termes ont une répartition non-homogène : *des objets concrets (corps, centre, base, courbe, cône, couronne) ; des objets abstraits (contrôle, unité) ; des processus, des phénomènes, des états (analyse, dynamique).*

Au niveau de la macro- interdisciplinarité mais aussi au niveau de la micro- interdisciplinarité, les termes qui occupent une position scalaire plus haute sont ceux qui appartiennent à la sous- catégorie sémantique des objets concrets, contrairement à l'hypothèse intuitive qui pourrait soutenir une degré plus élevé de circulation des termes abstraits dans les sciences ; on constate une préférence accentuée pour les termes de la langue commune et non pas pour ceux dont l'origine est strictement scientifique.

5 La migration des termes scientifiques et le co- texte

L'actualisation du sens des termes mathématiques interdisciplinaires est, dans certains cas, indépendante du co- texte (linguistique) : il suffit d'avoir entendu en parlant des mathématiques un terme comme *algorithme, cône, fascicule, variable* pour saisir le concept correspondent.

La migration des termes indépendants est restreinte, à quelques exceptions près : *dynamique* (10 domaines) ; *disque* (9 domaines) ; *densité* (6 domaines), *image* (6 domaines), fait du a une spécialisation forte des termes qui ne permet l'utilisation de ces termes que dans le domaine des mathématiques. Si le terme mathématique indépendant transgresse son domaine, il doit se contenter, en général, dans un autre domaine, d'une utilisation dont le sens est différent (il s'agit de la polysémie ou même de l'homonymie) : *absorption, calcul, groupe, rapport, segment, unité*. Par exemple, *calcul (1)* (« mise en œuvre des règles élémentaires d'opération (addition, soustraction, multiplication, division) sur les nombres. ») est homonyme avec *calcul (2)* (« MED. concrétion pierreuse qui se forme dans divers organes (vessie, reins, vésicule biliaire, etc. »). Il n'y a que quelques termes interdisciplinaires qui gardent leur sens indépendant dans le passage d'un domaine à l'autre et ce fait n'est valable que pour quelques domaines (logique, linguistique, philosophie, physique, astronomie).

Pour certains termes l'existence d'un co-texte syntagmatique pour la récupération du concept est nécessaire (*analyse mathématique, forme trigonométrique, système d'équations, structure algébrique, unité imaginaire, valeur numérique*) ou facultative (*complément (d'un angle), dimension (d'un espace linéaire)*).

L'absence du mono- sémantisme[10] du mot – base de la syntagme déclanche la nécessité d'actualisation du syntagme entier pour récupérer le concept. Parmi les termes qui peuvent être utilisés soit indépendamment, soit dans des syntagmes, la plus grande partie garde un sens tout le longue de leur utilisation, les autres change leur sens dans le passage de l'indépendance à la dépendance co-textuelle. Par exemple, le terme *équation*, qui a le sens « égalité qui n'est vérifiée que par certaines valeurs attribuées aux inconnues » (DNT) garde le sens dans les syntagmes , les mots ajoutés ne réalisent qu'une classification, une typologie des équations : *équation différentielle, équation d'une courbe, équation du temps, équation personnelle*.

Les termes qui font partie de la sous- catégorie sémantique des objets concrets et qui garde leur sens de l'utilisation individuelle sont susceptibles d'une production riche de syntagmes : *fonction (MATH – fonction réelle d'une variable réelle, fonction complexe d'une variable réelle, fonction algébrique, fonction numérique, fonction du premier degré, fonction du deuxième degré, fonction logarithmique, fonction transcendante ; (AD M) – fonction publique, fonction publique territoriale ; PHYS – fonction de nutrition, fonction de reproduction,*

fonctions digestives ; CHI M – fonction acide, LING – fonction dénotative ; GRAM – fonction syntaxique ; LOG – fonction propositionnelle ; ECON – fonction de production, fonction commerciale) ; cercle (MATH – grand cercle d'une sphère, petit cercle, cercle d'Euler ; ASTRO – cercle de hauteur, cercle horaire d'un astre, cercle méridien ;BOT – cercle annuel ; PHYS – cercle oculaire ; TECH – vin en cercles ; LOG – cercle vicieux ; (SOCIOL) – cercle de famille, cercle d'études ; (SPORT) – cercle sportif ; (POL) – cercle politique). Les termes qui appartiennent aux autres catégories sémantiques donnent naissance à un nombre réduit de syntagmes : *algorithme (algorithme d'Euclide).*

L'absence de l'homonymie[11] dans le langage de spécialité, est à l'origine de la manifestation rare du phénomène de changement de sens du terme- base d'un syntagme : *base, champ, classe, complément, corps, facteur*. Par exemple, le sens du terme seul *facteur :* « chacun des nombres figurant dans un produit » est différent du sens qu'il prend dans le syntagme *facteur premier d'un nombre :* « nombres premiers, distincts ou non, dont le produit est égal à ce nombre. (Un nombre admet une décomposition unique en facteurs premiers). ».

Quant au phénomène de migration des syntagmes, aucun syntagme mathématiques n'est pas utilisé tel quel dans un autre domaine. Le terme- base peut émigrer, mais il devient base pour de nouveaux syntagmes. Par exemple, *facteur* aide à la construction des syntagmes : *facteur de puissance* (physique) ou *facteur général* (psychologie).

[10] Le postulat de la terminologie de Wűster à Lerat et Cabré est le mono- sémantisme des termes de spécialité.
[11] Marcus, Solomon (1970) – *Poetica matematică*, EARSR, Bucureşti, p. 34-35.

6 Conclusions

Contrairement à l'hypothèse initiale les termes qui réalisent une macro- interdisciplinarité ou une micro- interdisciplinarité plus riche sont des termes qui dénotent des objets concrets et non pas des objets abstraits.

Entre les divers domaines il y a beaucoup de contacts, mais l'interdisciplinarité ou la migration totale est assez réduite, le passage d'un terme d'un domaine à l'autre est en général accompagné soit d'un changement partiel de sens (polysémie), soit d'un nouveau co- texte, l'ancien terme devient la base d'un nouvel syntagme. Parfois le changement de sens est radicale (on se pose même la question s'il s'agit toujours d'un même terme) et une seule forme lexicale couvre deux concepts totalement différents, distincts (homonymes).

Références

Béjoin, H. et Thoiron, Ph. (dir.) (2000). *Le sens en* termonilogie, Presses Universitaitre de Lyon, Lyon.

Bidu-Vranceanu, A. (2000). ***Terminologiile ştiinţifice din perspectivă interdisciplinară***, AUB,Bucureşti.

Bidu-Vranceanu, A.; Toma, A. (2000). ***Lexic comun, lexic specializat***, EUB, Bucureşti.

Bidu Vranceanu, A.; Toma, A. (2001). ***Lexic ştiinţific interdisciplinar,*** EUB, Bucureşti.

Candel, D. et Leujeune, D. (1998). **Définir en mathématiques. Regards lexicographiques sur des textes de mathématiques**, in *Cahiers de lexicologie, 7:*43-60.

Foucault, M. (1969). ***L'archéologie du savoir***, Gallimard, Paris.

Gaudin, F. (2003). *Socioterminologie. Une approche sociolinguistique de la terminologie*, Éditions Duculot, Bruxelles.

Guilbert, L. (1965). *La formation du vocabulaire de l'aviation*, Librairie Larousse, Paris.

Kleiber, G.(1999). ***Problèmes de sémantique. La polysémie en questions***, PUS.

Kocourek, R. (1991). ***La langue française de la technique et de la science***, 2 ed.,, Brandstetter Verlag/ *La documentation française, Wiesbasen/ Paris* .

Lerat P.(1995). ***Les langues spécialisées***, PUF, Paris.

Losee M., R. (1995). The Developement and Migration of Concepts from Donor to Borrower Disciplines: Sublanguage Term Use un Hard & Soft Sciences, in *Proceedings of the Fifth International Conference on Scientometrics and Infometrics*, Chicago, June 1995: 265-274.

Marcus, S. (1970). ***Poetica matematică***, EŞ, Bucureşti.

Rey, A.(1979).***La terminologie***, PUF, Paris.

Roulet, E. ; Filliettaz, L.; Grobet, A. (2001). ***Un modèle et un instrument d'analyse de l'organisation du discours***, Peter Lang, Éditions scientifiques européennes.

Stoichiţoiu, A. (1990). ***Sens şi definiţie în limbajul juridic***, în SCL, XVI, nr.4.

Stoichiţoiu, A. (2001). ***Semiotoca discursului juridic***, EUB, Bucureşti.

Un système de calcul des thèmes de l'actualité à partir des sites de presse de l'internet

JACQUES VERGNE
GREYC - UMR 6072
campus II - BP 5186
Université de Caen
14000 Caen, FRANCE
mail :Jacques.Vergne@info.unicaen.fr
tél. : 02 31 56 73 36 fax : 02 31 56 73 30

Résumé

Dans cet article, nous présentons un système de constitution de revue de presse à partir des sites de presse présents sur l'internet[1]. Il s'agit de répondre à des questions telles que : "de qui, de quoi est-il question aujourd'hui dans la presse de tel espace géographique ou linguistique ?". L'utilisateur, qu'il soit un journaliste qui prépare sa revue de presse, ou simplement une personne intéressée par l'actualité, définit en entrée l'espace de recherche qui l'intéresse. Ce système inverse la problématique des moteurs de recherche : au lieu de rechercher des documents à partir de mots-clés qui représentent des thèmes, il s'agit de produire en sortie les thèmes principaux de l'actualité, et de donner accès aux articles concernés par ces thèmes. Les thèmes d'actualité sont capturés en relevant les termes récurrents dans les "textes" d'hyperliens des "Unes" des sites de presse. Le système calcule un graphe de termes dans lequel les nœuds sont les termes et les arcs sont les relations entre termes, relations définies par la co-occurrence de deux termes dans un "texte" d'hyperlien. L'interface exploite ce graphe en permettant à l'utilisateur de naviguer parmi les termes et d'avoir accès aux articles contenant ces termes[2].

Mots-clés : hypertextes, web, internet, documents électroniques, web mining, recherche d'informations, veille stratégique, fouille de textes.

Abstract

In this paper, we present a system for building a news review, from news sites on the web. We want to be able to answer questions as : "who, what are papers speaking about today in the news of a given geographic or linguistic search space". The user, a journalist preparing his news review, or somebody interested in news, defines as input the search space he is interested in. This system reverses the issues of search engines : in spite of searching documents from key-words which represents topics, we want to produce as output the main topics of the news, and to give access to related papers. News topics are captured while computing recurrent terms in hyperlinks texts of front-pages of news sites. The system computes a graph in which nodes are terms and arcs are links between terms; a link is defined as a co-occurrence of two terms in a same link text. The interface is based on this graph as the user can browse through the terms and have access to papers containing these terms.

[1] Une démonstration est accessible sur :
http://www.info.unicaen.fr/~jvergne/demoRevueDePresse/index.html

[2] Le système présenté a des analogies avec celui de Google News (http://news.google.fr), mais Google News n'a pas encore publié sur son processus de traitement.

Key-words : hypertexts, web, internet, electronic documents, web mining, information retrieval, strategic watching, text mining.

1 Introduction

Le système que nous présentons comporte en donnée une liste aussi large que possible des URL des sites de presse du monde entier, qui constituent les points d'entrée possibles. L'utilisateur définit en entrée un espace de recherche géographique et/ou linguistique sous la forme d'un sous-ensemble de cette liste. Sa requête implicite est : "de qui, de quoi est-il question aujourd'hui dans la presse de cet espace ?". Le système fournit en sortie un graphe de termes valués, reliés par des relations valuées. Un terme est valué par trois grandeurs : le nombre de termes auxquels il est relié, le nombre de sites sur lesquels il a été trouvé, et le nombre de textes de liens dans lesquels il a été trouvé (ce qui correspond au nombre d'articles concernés par le terme). La relation entre deux termes est définie par leur co-occurrence dans un même "texte" d'hyperlien. Une relation est valuée par le nombre de textes de liens où les deux termes sont présents. Ces valeurs attribuées aux termes et à leurs relations permettent de les classer pour les présenter à l'utilisateur par ordre de présence décroissante dans l'actualité du jour dans l'espace de recherche défini en entrée. Elles permettent aussi des traitements de graphe particuliers. L'utilisateur prend connaissance des résultats en naviguant dans le graphe : il choisit un terme, puis des termes liés, et a accès à tout moment aux documents concernés par les termes et leurs liens.

Nous présentons d'abord les principes de fonctionnement, puis ensuite le processus général de calcul et ses étapes successives.

2 Principes de fonctionnement

Pour chaque site, **un seul** document est téléchargé : le document du point d'entrée, c'est-à-dire la "Une" du site de presse. De ce document, sont extraits les hyperliens : les URL et le code source des "textes" de liens. On observe que ces codes source de "texte" de liens sont composés de titres ou de résumés d'articles (avec leur mise en forme), et d'URL vers des images ou des photographies. Dans une même Une, une même URL peut apparaître plusieurs fois. Les URL de photographies permettent de les montrer à l'utilisateur en sortie.

C'est dans les "textes" de liens (leur code source débalisé) que sont extraits les termes. Les URL des articles ne serviront qu'en sortie, pour donner accès à un article, si l'utilisateur le décide. Le système ne se sert pas des articles eux-mêmes. Cette économie de traitement s'appuie sur le fait que la rédaction d'un texte de lien est un choix éditorial des journalistes des sites de presse.

Un point délicat est la **méthode d'extraction des termes** à partir du corpus des textes de liens débalisés. La tâche est relativement simple : il s'agit de trouver les motifs répétés, dans un corpus thématiquement varié, relativement petit (environ 90 à 160 Ko, 17000 à 25 000 mots), tout en repérant les mots grammaticaux (motifs de hautes fréquences), pour éviter d'en faire des termes. Mais nous avons des contraintes particulières sur la méthode : elle doit être robuste et indépendante des langues, ce qui est une nécessité pour un logiciel de traitement d'informations sur l'internet, caractérisé par la multiplicité des langues; la méthode ne doit pas utiliser de ressources propres à une langue, car il n'est pas question de faire un travail de préparation de ressources linguistiques à chaque nouvelle langue traitée.

Nous n'avons donc pas besoin de méthode lourde avec analyse syntaxique et mises en relation, telle que celle de Didier Bourigault qui utilise l'analyseur syntaxique en dépendance SYNTEX pour construire une ontologie à partir d'un corpus vaste et très cohérent thématiquement (voir [Bourigault, 2000] et [Bourigault, 2002]).

Les méthodes d'André Salem (voir [Salem, 1987]) et d'Helena Ahonen (voir [Ahonen-Myka, 1999] et [Ahonen-Myka, 2002]) recherchent les motifs répétés en utilisant des algorithmes extrapolés de l'algorithme glouton (recherche des n-grammes à partir des n-1-grammes), mais ces méthodes utilisent en entrée les mots grammaticaux de la langue traitée pour éviter de les prendre comme termes (stopword-list).

Plusieurs méthodes ont été explorées, en nous imposant la contrainte de trouver une méthode n'utilisant pas de ressources linguistiques, pour rester robuste et indépendant des langues : la recherche des motifs répétés par l'algorithme glouton a été expérimentée, d'abord avec puis sans l'utilisation des majuscules[3], avec exclusion des mots grammaticaux par leur fréquence (test de Zipf). Puis une méthode tout à fait originale fondée sur la périodicité des longueurs de mots a été mise au point. Elle permet de calculer si un mot est grammatical ou lexical sans autre ressource que le corpus traité lui-même (une méthode dite "endogène", en reprenant le terme de Didier Bourigault).

Un fois les termes calculés, on leur associe la liste des sites où on les a trouvé, et on ne garde que les termes trouvés sur au moins deux sites, donc dans des textes de liens vers deux articles différents de deux sites différents. C'est cette opération d'intersection qui permet d'exclure les termes particuliers à un site unique (menus, publicités, etc.).

On associe ensuite à chaque terme restant sa liste de liens vers les articles.

Les **relations entre termes** peuvent ensuite être calculées : il existe une relation entre deux termes s'ils sont co-occurrents dans un même texte de lien, ou, ce qui est équivalent, si l'intersection des deux listes d'articles n'est pas vide. Dans cet étape, il s'agit simplement de calculer les intersections de listes de liens des termes deux à deux.

Étant donné que plusieurs centaines de termes sont extraits, il faut, pour faciliter l'accès à l'utilisateur, les lui présenter par **groupes de termes fortement reliés** : les groupes calculés par le système à partir de propriétés du graphe, sont interprétés par l'utilisateur comme des groupes de termes thématiquement reliés (cf. le "grouping process" de Google News[4]). Une idée simple et classique est de calculer les composantes connexes du graphe de termes (c'est-à-dire ses sous-graphes non connexes). Cette solution a été expérimentée, puis abandonnée, car le graphe est composé d'une très grosse composante (presque tous les thèmes de l'actualité sont liés), et de beaucoup de petites composantes (un thème a donné 2 termes reliés ou même un seul terme). Donc le problème à résoudre est de segmenter la composante connexe principale en ensembles de nœuds fortement reliés. Là encore, une solution classique se présente : la recherche des cliques maximales[5]. Cette solution a aussi été expérimentée, et aussi abandonnée : l'algorithme glouton (recherche des cliques de n nœuds à partir des cliques de n-1 nœuds) donne un résultat combinatoire qui pose le problème supplémentaire de choisir entre des cliques équivalentes. Mais cette méthode a surtout un intérêt sur les graphes non valués. Or nous sommes en présence de graphes valués, d'où la recherche d'une méthode de groupage des termes fondée sur les valeurs des liens et sur les valeurs des nœuds : les groupes sont constitués à partir des nombres de co-occurrences des termes et de leurs nombres de sites.

Il s'agit enfin de présenter à l'utilisateur une **interface d'accès au graphe des termes et aux articles**. Une solution envisageable est un graphe cliquable, sous-graphe du graphe des termes, à la manière de Kartoo (`www.kartoo.com`). Nous avons choisi une solution plus simple à mettre en œuvre (Kartoo utilise la technologie flash) : le graphe des termes est transposé dans un graphe de documents html (un document html par nœud-terme) reliés par des hyperliens dans les deux sens (2 hyperliens réciproques instancient un arc non orienté entre 2 termes). L'utilisateur peut ainsi naviguer dans le graphe des termes.

Ces principes de fonctionnement ont permis de construire un système léger, robuste, sans ressources linguistiques, indépendant des langues, qui utilise des propriétés très générales des langues.

3 Le processus général et ses étapes

Voici les étapes du processus général des traitements :

- phase préparatoire manuelle : collecter les URL des sites de presse

[3] Pour pouvoir traiter des langues sans majuscules, comme l'arabe par exemple.

[4] Voir sur `http://news.google.com/help/about_news_search.html` : "an automated grouping process for Google News that pulls together related headlines and photos from thousands of sources worldwide".

[5] Une clique (ou graphe complet) est un ensemble de nœuds où tout nœud est relié à tous les autres.

- traitements sur l'ensemble des sites :
  - télécharger et analyser **la Une de tous les sites**
  - relever **les termes** dans les textes de liens
  - pour chaque terme, calculer sa liste de sites et sa liste d'articles

- entrer l'espace de recherche de l'utilisateur

- traitements sur les sites choisis par l'utilisateur :
  - calculer **les relations entre termes**
  - regrouper les termes fortement reliés
  - sortir les résultats = calculer l'interface

3.1 Collecter les URL des sites de presse (phase préparatoire)

Les URL des sites de presse ont été collectées sur les sites suivants :

- le Courrier International propose une liste commentée d'environ 800 sites de presse du monde entier présentée par continent (sauf la France) :
  `http://www.courrierinternational.com/kiosk/`
- Google News donne une centaine de ses sources sur les 4000 revendiquées :
  `http://news.google.fr/news/`
- le site de NewsLink donne aussi de nombreux sites de presse :
  `http://newslink.org/`

Notre système est testé quotidiennement sur une quarantaine de sites : 22 de la presse française nationale et régionale, 17 de la presse européenne (Suisse, Belgique, Allemagne, Italie, Espagne, UK, Irlande), et 4 sites de presse nord-américaine (espace 1 dans la suite de l'article). Des tests de vitesse d'exécution sont aussi fait sur les 100 sites de Google News, environ la moitié sont des sites nord-américains (espace 2 dans la suite de l'article).

Cette phase s'apparente au "sourcing" des sociétés de veille technologique.

3.2 Traitements sur l'ensemble des sites

3.2.1 Télécharger et analyser la Une de tous les sites

Le téléchargement des Unes est fait périodiquement (une fois par jour, par exemple), pour l'ensemble des sites de la base de sites. Le téléchargement des Unes et le relevé des termes dans les textes de liens sont dissociés de l'interrogation par l'utilisateur. Les phases suivantes, à commencer par le calcul des relations entre termes, sont dépendantes de l'espace de recherche défini par l'utilisateur.

Après téléchargement, le code source de la page est analysé. Pour chaque lien, les URL et le code source des textes de liens sont extraits[6]. Les URL relatives sont converties en URL absolues. Comme une même URL peut apparaître plusieurs fois, les textes de liens sont concaténés, et les URL d'image sont aussi converties en URL absolues. Après ces conversions, les URL et les codes sources des liens sont mémorisés dans une base de données.

3.2.2 Relever les termes dans les textes de liens

Cette étape consiste à extraire des termes d'un corpus : le corpus des textes de liens concaténés, obtenus par débalisage des codes source des liens, entre les balises `<a>` et `</a>`. Pour fixer les idées, ce corpus est de l'ordre de 90 Ko pour les 43 sites de l'espace de recherche 1.

Une première question se pose au sujet de **l'espace de constitution du corpus**, c'est-à-dire l'espace dans lequel on va rechercher des motifs répétés. Plusieurs solutions sont possibles : un corpus pour tous les sites, un corpus par langue (si on décidait de se servir de la langue de chaque site), ou un corpus par site. Nous avons choisi un corpus pour tous les sites, ce qui permet dès cette étape, de ne considérer que les termes répétés (au moins 2 occurrences) dans l'ensemble du corpus, qui pourront

[6] Il est relativement fréquent que les balises `<a>` ne soient pas fermées par une balise `</a>`. Dans ce cas, on utilise une fermeture implicite par la prochaine balise `</p>` ou `</td>`. De plus, pour pallier les liens quasi-vides (tels que "*Lire*"), on concatène au texte de lien le texte du `<td>` englobant le lien.

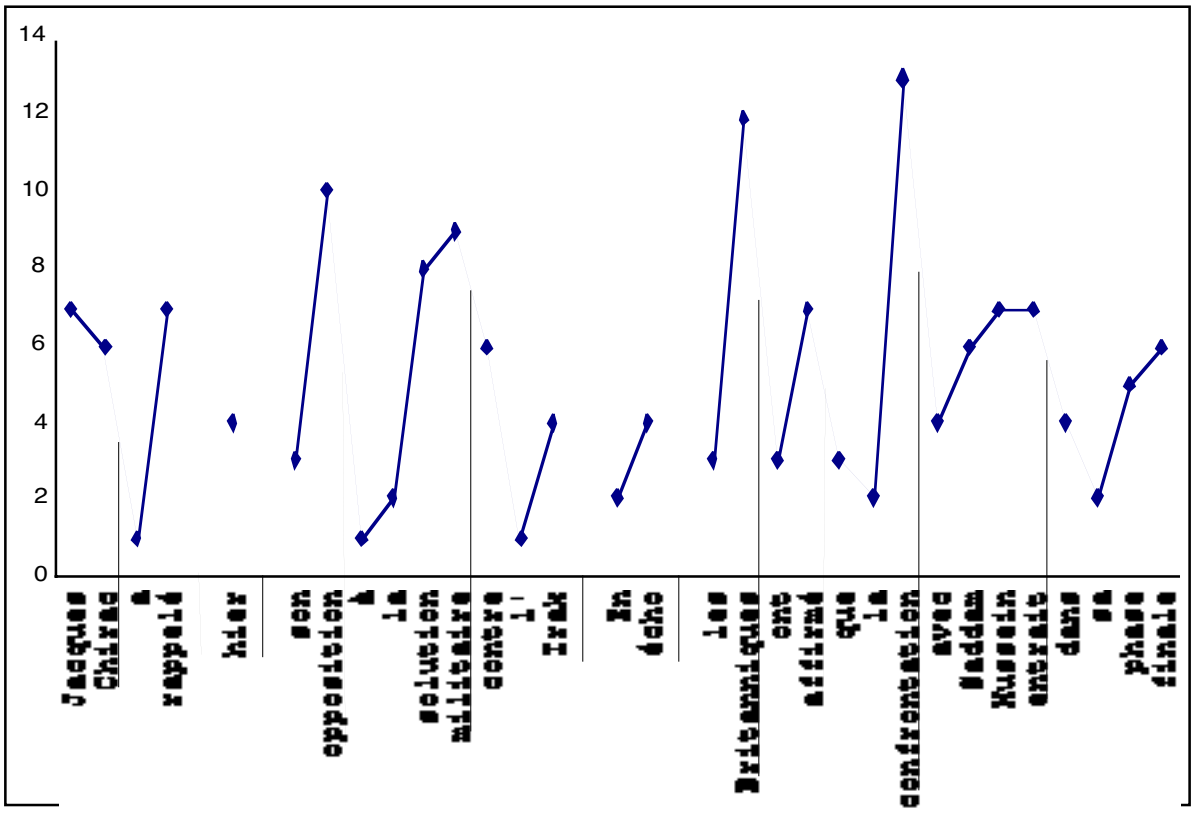
éventuellement se trouver sur deux sites différents; alors que le calcul des termes par site oblige à garder temporairement les termes hapax, car deux termes hapax dans deux sites différents peuvent être identiques, et donc répétés sur l'ensemble des sites; or garder les termes hapax temporairement alourdit inutilement les calculs.

Une deuxième question est celle de la **méthode d'extraction des termes** à partir d'un corpus.

Dans un premier temps, nous avons expérimenté l'algorithme glouton, associé avec un test de Zipf pour éliminer les mots grammaticaux des termes, sur un simple critère de fréquence. Les résultats étaient corrects, mais le départage entre les mots grammaticaux et les mots lexicaux très fréquents était difficile (tel que "guerre" aujourd'hui).

D'où une autre manière de poser le problème : comment distinguer les mots lexicaux et les mots grammaticaux à partir du corpus ? Une direction de travail est d'utiliser à la fois les formes et leurs positions (une constante du Groupe Syntaxe du GREYC, dont Hervé Déjean [Déjean, 1998], Nadine Lucas [Lucas, 2001]), alors que le test de Zipf n'utilise que les fréquences des formes, sans exploiter leurs positions relatives. Or une observation fondamentale de Zipf est que les mots grammaticaux sont fréquents et courts et que les mots lexicaux sont plus rares et plus longs (ce qui est d'usage fréquent est court : c'est la loi de l'économie d'effort dans l'usage d'un code, caractérisée par Zipf, et visible aussi dans les langages de programmation). D'où l'idée d'utiliser non plus les différences de fréquence, mais les différences de longueur. Si on associe les longueurs des mots (un type de forme) avec les positions relatives des mots, on est tout naturellement conduit à s'intéresser aux variations des longueurs au fil du texte. Voici en figure 1 un exemple de graphe de la fonction :

Figure 1 : Graphe de la fonction :
nombre de lettres d'un mot = f(position du mot au fil du texte)
(exemple extrait de Ouest-France du 22 février 2003)



Le texte est segmenté sur les ponctuations, d'où des segments physiques que nous nommons "virgulots". Dans chaque virgulot, on observe quelques périodes d'**un signal périodique**, avec des alternances court-long avec le plus souvent un début de virgulot court; on observe aussi que la segmentation en chunks est manifestée par un début court, soit en début de virgulot, soit après un mot

long. Sur ces propriétés, nous avons bâti un algorithme de chunking approché et de catégorisation des mots en trois catégories : les mots courts, les mots longs, et les mots indéterminés. Cet algorithme consiste à couper aux endroits où la pente est négative, ce qui revient à garder les parties monotones croissantes (motif court*-long+), qui correspondent à un chunk, c'est-à-dire à un groupe accentuel : on obtient un segment caractérisé par sa prosodie, propriété linguistique indépendante des langues (voir la présentation prosodique du chunk chez Abney, au début de [Abney, 1991]).

Puis, à l'intérieur de chaque chunk, l'algorithme consiste à catégoriser chaque mot : court ou long, suivant sa longueur et sa position. Puis les déductions locales aux chunks sont consolidées par une étude globale sur tout le corpus : ne sont retenus que les mots de nombre d'occurrences supérieur à 1 (les seuls qui pourront devenir des termes) et de catégorie stable sur le corpus (la "stabilité" se caractérisant par une majorité pour une catégorie). À partir de cette catégorisation, deux motifs de termes sont possibles (avec le langage des expressions régulières, c pour **c**ourt, L pour **L**ong) :

`c* L+` *(la) [violence] routière*

`c* L+ cL+` *(l)'[intervention] (de) [Jean-Pierre Raffarin]*

Le terme (`L+` ou `L+cL+`) est mémorisé avec tous ses contextes (leurs chunks), pour pouvoir les fournir à l'utilisateur en sortie (par coloriage).

Enfin, certains termes sont fusionnés sur la ressemblance de leurs graphies :

ONU ≡ Onu, Côte d'Ivoire ≡ Côte-d'Ivoire

Cette méthode a de nombreux avantages : pas de ressources externes, pas de diagnostic de langue, exploitation de propriétés linguistiques indépendantes des langues, robustesse, insensibilité aux déséquilibres entre langues dans le corpus, étant donné qu'on n'utilise pas de test de Zipf [7].

3.2.3 Pour chaque terme, calcul de sa liste de sites et de sa liste d'articles

Pour chaque terme, on calcule sa liste de sites en vérifiant sa présence dans le corpus de chaque site. À ce stade, les termes présents sur un seul site sont éliminés.

Pour chaque terme, pour chaque site où est ce terme, on calcule enfin sa liste d'articles en vérifiant sa présence dans chaque texte de liens de ce site.

3.3 Entrer l'espace de recherche de l'utilisateur

L'utilisateur définit interactivement par un formulaire son espace de recherche, c'est-à-dire un sous-ensemble de l'ensemble des sites de presse entrés dans la phase préparatoire. Les critères sont linguistiques et/ou géographiques. Chaque site doit être complété de ces deux informations, qui ne servent qu'à cette étape.

3.4 Traitements sur les sites choisis par l'utilisateur

Au moment de l'entrée de l'espace de recherche de l'utilisateur, la liste d'articles de chaque terme est filtrée pour ne retenir que les sites choisis.

3.4.1 Calculer les relations entre termes

À partir de cette étape, les calculs sont dépendants de l'espace de recherche défini par l'utilisateur.

Il existe une relation entre deux termes si l'intersection des deux listes d'articles n'est pas vide. La valeur d'une relation est définie par le cardinal de l'intersection.

Dans cette étape, il s'agit simplement de calculer les intersections de listes d'articles deux à deux, ce qui est possible à l'aide de requêtes sur la base de données. On obtient le graphe (non orienté) des termes.

On peut donner ici quelques informations quantitatives sur les graphes de termes obtenus : 400 à 700 termes-nœuds, 1600 à 2500 relations-arcs, d'où des densités (rapport entre le nombre d'arcs réels et le nombre d'arcs possibles) de l'ordre de 1%.

[7] Voir aussi dans [Vergne, 2003], une extension de cette méthode, avec utilisation des différences de longueur **et** d'effectif dans le corpus des "textes" de liens.

3.4.2 Regrouper les termes fortement reliés

Pour regrouper les termes, on parcourt la liste des couples de termes liés par valeurs décroissantes des relations (jusqu'à une valeur 2 : on néglige à ce stade les relations trop faibles), et on place les 2 termes a et b dans un groupe avec les règles suivantes :

- si un groupe contient a ou b ou les 2, placer a et b dans ce groupe
- si aucun groupe ne contient a ou b, créer un nouveau groupe et y mettre a et b
- si a et b sont déjà dans 2 groupes différents, mémoriser ce couple.

Les couples ainsi mémorisés sont des relations passerelles entre groupes, à partir desquelles on fusionne quelques groupes fortement liés, en fusionnant d'abord des groupes de tailles très différentes (un petit rejoint un gros). Le problème est de définir un critère d'arrêt, pour éviter de remettre trop de termes ensemble. Le critère actuel est un seuil du rapport de taille des groupes, la taille d'un groupe étant évaluée par la somme des valeurs des relations des couples qui le constituent. Le fonctionnement actuel est correct, mais demande encore à être travaillé.

3.4.3 Sortir les résultats : interface

L'interface permet à l'utilisateur de naviguer dans le graphe des termes et d'accéder aux articles. Le système produit en sortie un document html par terme, avec 2 hyperliens réciproques par relation entre termes.

L'interface est constituée de 2 sous-fenêtres :

- la première propose un choix entre les termes, présentés par groupe, par ordre décroissant d'importance (somme des valeurs des relations), et dans chaque groupe, les termes par nombre de sites décroissants; l'utilisateur choisit un terme en cliquant dessus, et le terme apparaît dans la deuxième sous-fenêtre;
- la deuxième sous-fenêtre permet de naviguer dans le graphe à partir du terme choisi : on y voit ce terme, avec ses termes liés présentés par valeur décroissante de la relation; pour chaque terme lié, sont présentés les articles dans lesquels les deux termes sont co-occurrents; pour chaque article, l'utilisateur voit le lien sur l'article, le texte du lien avec son éventuelle photographie associée et sa mise en forme originale (le terme choisi et ses termes liés sont coloriés pour faciliter leur repérage en lecture rapide); l'utilisateur peut choisir de cliquer sur un terme pour continuer sa navigation dans le graphe des termes, ou il clique sur un lien d'article, ce qui fait apparaître l'article dans une nouvelle fenêtre.

Quelques résultats intermédiaires des calculs et une copie d'écran de l'interface sont donnés en annexe à la fin de l'article.

4 Conclusion

La collecte d'informations sur l'internet, la synthèse des informations collectées, et leur mise à la disposition des utilisateurs, sont des tâches très intéressantes car, tout en étant un enjeu opératoire et social, elles posent aussi de nouveaux problèmes de traitement du matériau linguistique : sur la toile, les langues sont multiples, le lexique est ouvert; la variété et l'immensité du matériau, rendant illusoire l'accumulation de ressources linguistiques propres à des langues particulières, nous conduisent à l'exploitation de propriétés linguistiques de plus en plus générales, et de plus en plus abstraites.

Dans le cas particulier du système présenté dans cet article, nous nous sommes servi (implicitement) de propriétés du groupe accentuel pour ne retenir que des mots lexicaux dans les termes; nous avons beaucoup utilisé la co-occurrence, l'unique ne nous intéresse pas, nous recherchons le multiple : co-occurrence entre 2 sites, puis entre 2 articles; la co-occurrence est une manière d'exprimer la relation : principalement la relation entre termes, qui se traduit dans notre système par un arc du graphe des termes.

Quels sont les acquis ? Le système fonctionne très bien et donne quotidiennement satisfaction à l'auteur comme utilisateur intéressé par l'actualité mondiale. Il permet de mettre à l'épreuve notre nouvelle méthode d'extraction de termes à partir de corpus brut multilingue. Il est frappant de constater combien la recherche multisite est facile, alors que la même tâche en monosite est très difficile (nous nous y sommes attaqué précédemment); c'est une bonne illustration de l'efficacité de la co-occurrence intersite.

Le système présenté nous invite à poursuivre nos recherches dans plusieurs directions : notre méthode d'extraction de termes de corpus brut est à consolider sur des corpus plus importants et sur une plus grande variété de langues (actuellement : français, anglais, allemand, italien, espagnol), et sa robustesse doit être étendue : le choix actuel est de tenter d'annuler le bruit, au prix d'un certain silence, ce qui n'est pas un désavantage dans la tâche actuelle, qui consiste à trouver des termes fréquents, et à ignorer les hapax. De manière plus précise, le chunking sur corpus brut multilingue et sans ressources devra être approfondi. Le calcul des intersections des ensembles de document devra être amélioré car, du point de vue calculatoire, c'est le point faible de la chaîne des traitements : il consiste en un parcours de la demi-matrice carrée des quelque 500 termes (moins la diagonale), soit de l'ordre de 500*500 / 2 = 125 000 calculs d'intersection. Le groupage de termes, qui ne donne pas encore entière satisfaction, est un problème difficile, qui demandera l'exploration de nouvelles solutions. Enfin, l'interface doit évoluer, en interaction avec des utilisateurs.

Références

[Abney, 1991] Abney Steven (1991). "Parsing By Chunks". In: Robert Berwick, Steven Abney and Carol Tenny (eds.), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.
http://www.sfs.nphil.uni-tuebingen.de/~abney/Abney_90e.ps.gz

[Ahonen-Myka, 2002] Ahonen-Myka Helena (2002). Discovery of frequent word sequences in text. *The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, Imperial College, London, 16-19 September.
http://www.cs.helsinki.fi/u/hahonen/ahonenmyka_patws02.ps

[Ahonen-Myka, 1999] Ahonen-Myka Helena (1999). Finding All Frequent Maximal Sequences in Text. *Proceedings of the 16th International Conference on Machine Learning ICML-99 Workshop on Machine Learning in Text Data Analysis*, eds. D. Mladenic and M. Grobelnik, p. 11-17, J. Stefan Institute, Ljubljana .
http://www.cs.helsinki.fi/u/hahonen/ham_icml99.ps

[Bourigault, 2002] Bourigault, Didier (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus, *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*, Nancy, pp. 75-84.
http://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc

[Bourigault, 2000] Bourigault Didier & Slodzian Monique (2000). Pour une terminologie textuelle, *Terminologies Nouvelles*, n° 19., pp. 29-32.
http://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TN-Bourigault-Slodzian.rtf

[Déjean, 1998] Déjean Hervé (1998). "Concepts et algorithmes pour la découverte des structures formelles des langues", spécialité : informatique, thèse de l'Université de Caen.

[Lucas, 2001] Lucas Nadine (2001). Étude et modélisation de l'explication dans les textes, *Colloque "L'explication: enjeux cognitifs et communicationnels"*, Paris, 30 novembre - 1er décembre.

[Salem, 1987] Salem André (1987), *Pratique des segments répétés*, Publications de l'INaLF, collection "St.Cloud", Paris, Klincksieck, 333p.

[Vergne, 2000] Vergne Jacques (2000). *Trends in Robust Parsing*, tutoriel du CoLing 2000, Nancy, Sarrebrück.
http://www.info.unicaen.fr/~jvergne/tutorialColing2000.html

[Vergne, 2001] Vergne Jacques (2001). Analyse syntaxique automatique de langues : du combinatoire au calculatoire (communication invitée), *Actes de TALN 2001*, 15-29.
http://www.info.unicaen.fr/~jvergne/TALN2001_JV.ppt.zip

[Vergne, 2002] Vergne Jacques (2002). Une méthode pour l'analyse descendante et calculatoire de corpus multilingues : application au calcul des relations sujet-verbe, *Actes de TALN 2002*, 63-74.
http://www.info.unicaen.fr/~jvergne/TALN_2002/TALN2002_JVergne.doc.pdf

[Vergne, 2003] Vergne Jacques (2003). Un outil d'extraction terminologique endogène et multilingue, *Actes de TALN 2003*, à paraître.

[Zipf, 1949] George Kingsley Zipf (1949), *Human Behavior and the Principle of Least Effort*, Harper, New York, réédition 1966.

Annexe 1 : quelques résultats intermédiaires

Voici les termes les plus fréquents collectés le 26 février 2003 dans chacun des deux espaces définis en 3.1, classés par ordre de nombre de sites décroissants, avec pour chaque terme son nombre de sites et son nombre d'articles :

espace 1 (22 sites de la presse française, 17 de la presse européenne, et 4 sites de presse nord-américaine)			espace 2 (100 sites de Google News)		
Irak (14-29)	jour (7-8)	ONU (6-7)	Iraq (25-48)	Health (13-14)	North (10-11)
France (14-25)	Economie (7-7)	Saddam Hussein (6-7)	2003 (23-28)	Home (12-18)	Update (10-11)
Saddam (13-15)	politique (7-7)	Sports (6-7)	U.S. (20-51)	Terms (12-13)	Council (10-10)
2002 (12-17)	site (7-7)	intermittents (6-7)	Bush (20-30)	Special (11-27)	Services (9-13)
2003 (12-14)	article (6-51)	public (6-7)	Privacy (19-21)	House (11-19)	Information (9-12)
Monde (11-15)	Jacques Chirac (6-11)	semaine (6-7)	Saddam (17-23)	White (11-17)	Media (9-12)
Europe (10-11)	Business (6-9)	Blix (6-6)	Business (16-21)	National (11-16)	Sports (9-12)
Sport (10-11)	News (6-9)	Milan (6-6)	International (15-20)	York (11-15)	America (9-11)
Iraq (9-19)	gouvernement (6-9)	Washington (6-6)	more (14-36)	*this* (11-15)	Travel (9-11)
Bush (9-16)	mort (6-9)	dossiers (6-6)	Security (14-23)	Privacy Policy (11-11)	Europe (9-10)
Chirac (9-16)	Bernard Loiseau (6-8)	emploi (6-6)	Search (14-22)	South (10-14)	
guerre (9-14)	Blair (6-8)		news (14-20)	UN (10-14)	
mois (9-12)	America (6-7)		Press (13-21)	war (10-14)	
Jacques (7-14)			Help (13-15)	Your (10-13)	
Raffarin (7-12)				Online (10-12)	
Cinéma (7-9)					

On peut observer l'abondance de mots avec initiale majuscule, alors que ces mots ne sont pas favorisés par l'algorithme, ils sont simplement plus fréquents dans ce type de corpus. On observe aussi que l'élimination des mots grammaticaux n'est pas parfaite.

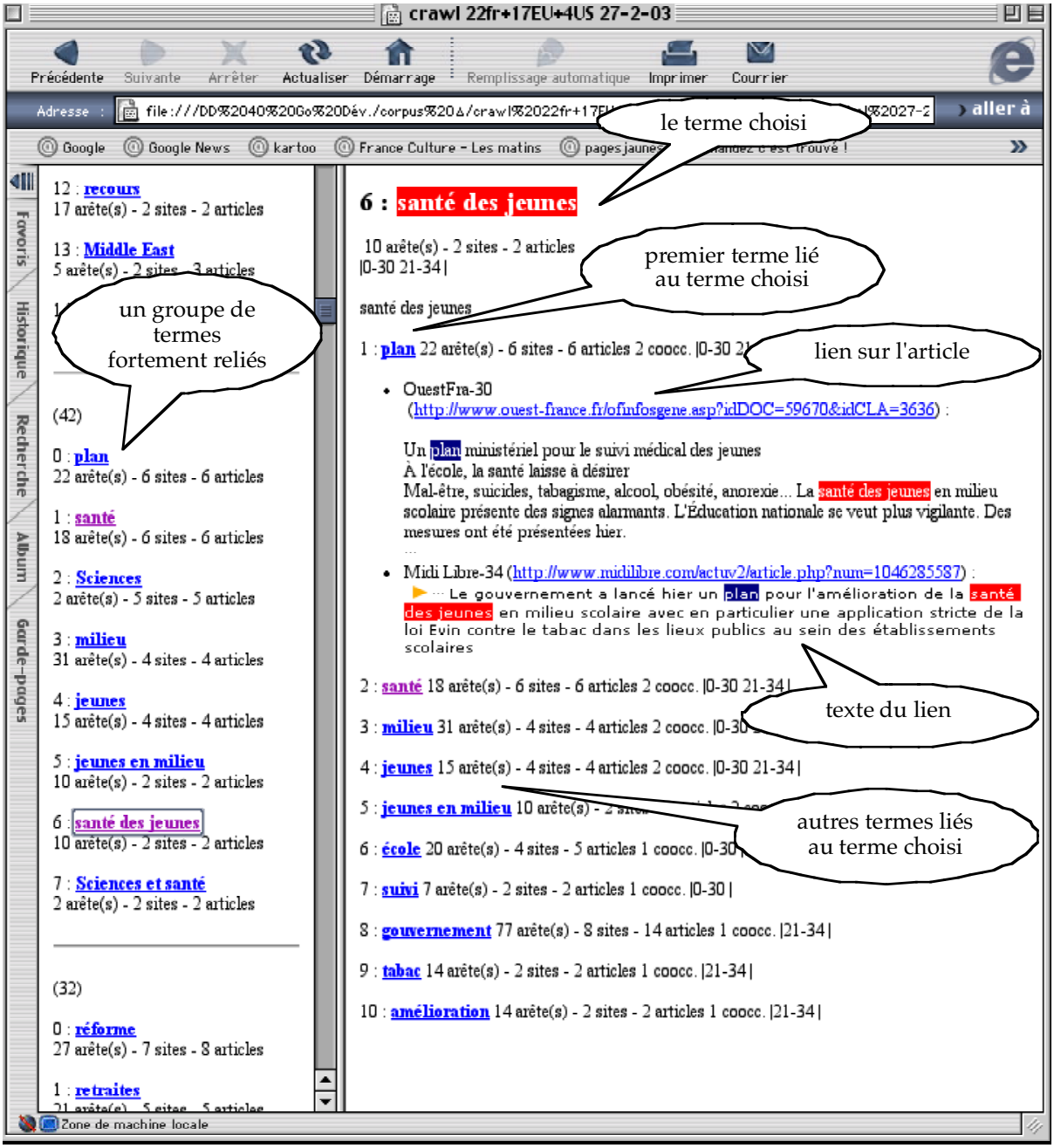
Voici les 22 liens les plus forts entre 2 termes, avec leur nombre de co-occurrences :

espace 1	espace 2
Chirac <—11—> Jacques	Privacy <—11—> Privacy Policy
Chirac <—11—> Jacques Chirac	House <—10—> White
Jacques <—11—> Jacques Chirac	White <—10—> White House
article <—11—> Lire	House <—10—> White House
article <—11—> Lire l'article	Iraq <—8—> U.S.
Lire <—11—> Lire l'article	North <—7—> North Korea
Saddam <—7—> Saddam Hussein	Security <—6—> Council
Irak <—6—> Chirac	Security <—6—> Security Council
Irak <—5—> Saddam	Council <—6—> Security Council
Chirac <—5—> article	Iraq <—5—> UN
Irak <—4—> Jacques	Iraq <—5—> Council
Irak <—4—> Jacques Chirac	U.S. <—5—> Turkey
Jacques <—4—> article	2003 <—5—> TM
article <—4—> Jacques Chirac	U.S. <—5—> U.S. Troops
Raffarin <—4—> Jean-Pierre Raffarin	Saddam <—5—> Hussein
America <—4—> Cup	Saddam <—5—> Saddam Hussein
Côte <—4—> Ivoire	Security <—5—> Homeland Security
America <—4—> America's Cup	Hussein <—5—> Saddam Hussein
Côte <—4—> Côte d'Ivoire	Iraq <—4—> Security
Cup <—4—> America's Cup	Iraq <—4—> House

Ivoire <—4—> Côte d'Ivoire Bové <—4—> José Bové	U.S. <—4—> House Iraq <—4—> Blix

On observe une redondance entre termes simples et termes multiples. En prenant l'exemple de *Jacques Chirac*, cela est dû aux faits qu'il existe d'autres *Jacques* et que *Chirac* existe seul sans le prénom.

Annexe 2 : interface



Le Web et la question-réponse : transformer une question en réponse

Luc Plamondon

RALI/DIRO, Université de Montréal
CP 6128, Succ. Centre-Ville
Montréal (Québec) Canada, H3C 3J7
Email : plamondl@iro.umontreal.ca
Tél : +1 514 343-6111 #3507 Fax : +1 514 343-2496

Leila Kosseim

CLAC Laboratory, Concordia University
1455 de Maisonneuve Blvd. West
Montréal (Québec) Canada, H3G 1M8
Email : kosseim@cs.concordia.ca
Tél : +1 514 848-3074 Fax : +1 514 848-2830

Résumé

De plus en plus de systèmes de question-réponse comptent le Web parmi les outils qu'ils utilisent pour trouver une réponse courte et précise à une question posée en langue naturelle. Dans cet article, nous présentons comment des règles simples de transformation de questions permettent de générer des contextes de réponse suffisamment restrictifs pour repérer des réponses sur le Web. La recherche de ces contextes (seuls ou en conjonction) alliée à une vérification sémantique de base des réponses extraites nous a permis de trouver la bonne réponse à 25 % des questions des campagnes d'évaluation TREC-10 et TREC-11.

Abstract

An increasing number of Question Answering systems use the Web to find a short and precise answer to a natural language question. In this paper, we present how simple question re-write rules allow for the generation of answer contexts that are restrictive enough for finding answers on the Web. Searching for such contexts (alone or in conjunction) and then performing simple semantics checks on the extracted answers leads to a correct answer for 25 % of a set of questions from the TREC-10 and TREC-11 conferences.

1 Introduction

Les développements récents dans le domaine de la question-réponse (*question answering*) rendent maintenant possible la recherche, dans un ensemble de textes, de réponses précises à des questions posées en langue naturelle. Par exemple, un utilisateur qui interrogerait un système de question-réponse (QR) avec la question *Qui était le premier ministre du Canada en 1873 ?* se verrait proposer une réponse spécifique, telle que *Alexandre Mackenzie*, plutôt qu'un document entier dans lequel il devrait localiser lui-même la réponse. À cet égard, la QR peut être vue comme la prochaine génération d'outils de recherche dans d'imposantes collections de textes telles que le Web.

Les campagnes TREC (*Text Retrieval Conference*), sous les auspices de l'institut américain des standards et de la technologie (NIST), ont donné le coup d'envoi aux recherches en QR en 1999 en lançant la première campagne d'évaluation de systèmes de QR. Depuis, chaque année, NIST met à la disposition des équipes participantes un ensemble de questions et une collection de textes afin qu'elles mettent leur système à l'épreuve. Les réponses sont ensuite évaluées par NIST suivant une méthodologie standard. Les questions et les textes fournis par NIST sont exclusivement en anglais, c'est pourquoi des campagnes d'évaluation de systèmes traitant d'autres langues ont récemment vu le jour : EQueR (dans le cadre du projet EVALDA de Technolangue/ELDA) pour la question-réponse en français et CLEF pour la question-réponse multilingue (textes en anglais et questions en d'autres langues, dont le français).

À ce jour, les campagnes TREC continuent d'être le foyer des principaux avancements en QR. On y a vu l'apparition de systèmes dérivés des moteurs de recherche classiques, de systèmes à apprentissage statistique, de systèmes faisant de l'analyse linguistique de surface et de systèmes faisant appel à des bases de connaissances et à des réseaux sémantiques tels que Wordnet [Miller, 1995]. Bien que les réponses suggérées par les systèmes doivent être présentes dans la collection de textes standard à TREC, de plus en plus de systèmes consultent aussi le Web : parfois pour départager plusieurs réponses prometteuses trouvées dans la collection de textes [Magnini et al., 2002], d'autres fois pour y puiser directement des réponses (quitte à les apparier ensuite avec des réponses similaires présentes dans la collection) [Brill et al., 2001, Brill et al., 2002, Clarke et al., 2001].

Afin d'améliorer notre système de QR, Quantum [Plamondon et al., 2002], nous avons privilégié la deuxième approche, c'est-à-dire utiliser le Web comme source primaire d'information. Quantum tente de prévoir, à partir de la question, le contexte (c'est-à-dire la séquence des mots environnants) dans lequel la réponse est susceptible d'être rencontrée, dans un texte quelconque. Il interroge ensuite un moteur de recherche Web avec ce contexte de réponse et, si le contexte est trouvé, il peut procéder à l'identification de la réponse exacte. Nous verrons qu'il est possible d'atteindre des résultats intéressants même avec un nombre limité de règles permettant de prévoir le contexte d'une réponse.

Deux stratégies de recherche sur le Web s'offrent à nous. La première vise un rappel élevé : elle consiste à obtenir une longue liste de candidats et à utiliser ensuite une méthode efficace d'évaluation des candidats afin de sélectionner les meilleurs. Cette stratégie est suivie notamment par les systèmes qui s'appuient sur la *redondance des réponses* et qui

considèrent que plus une réponse est fréquente, plus elle a de chances d'être la bonne. Nous avons plutôt opté pour la stratégie inverse : viser une précision élevée aux dépens du rappel. Il nous apparaissait en effet préférable, pour les besoins de Quantum, d'obtenir peu de candidats du Web mais d'avoir une confiance élevée en leur qualité. Pour ce faire, vu la quantité astronomique de textes accessibles sur le Web et leur impressionnante variété, nous avons choisi d'utiliser des critères de recherche très serrés qui garantiraient autant que possible la validité des réponses trouvées. Ces critères ne sont nul autres que le contexte exact de la réponse, que nous générons automatiquement à partir de la question.

2 Travaux antérieurs

Une première tentative de formulation automatique de contextes de réponse a été faite à TREC-10, où [Clarke et al., 2001] et [Brill et al., 2001] ont vu le Web comme une source gigantesque de textes supplémentaires pour améliorer l'extraction de réponses. En particulier, le système présenté par Microsoft à TREC-10 [Brill et al., 2001, Brill et al., 2002] recherche sur le Web une liste de formulations possibles d'une réponse, produites en permutant les mots de la question. Autrement dit, étant donné une question de la forme : *Who is* w_1 w_2 w_3 ... w_n *?*, le système cherche :

"w_1 `is` w_2 w_3 ... w_n"
"w_1 w_2 `is` w_3 ... w_n"
"w_1 w_2 w_3 `is` ... w_n"
...

Par exemple, étant donné la question : *Who is the world's richest man married to ?*, les requêtes suivantes sont construites :

```
"the is world's richest man married to"
"the world's is richest man married to"
"the world's richest man is married to"
...
```

Avec un peu de chance, au moins une des expressions (vraisemblablement la dernière de l'exemple ci-haut) sera trouvée dans une page Web et permettra d'identifier la réponse correcte. Bien que simple, cette stratégie s'avère très efficace. En utilisant cette méthode, [Brill et al., 2001] se sont classés 9^e sur 37 équipes à la campagne TREC-10. Suite à ces résultats, le Web est devenu une ressource presque standard pour la QR [Duclaye et al., 2002].

D'un autre côté, dans les travaux de [Agichtein and Gravano, 2000] et de [Lawrence and Giles, 1998], les formulations de réponses sont produites dans le but spécifique d'améliorer la recherche d'information sur le Web. Les formulations produites sont

précises mais elles sont employées pour l'expansion de requêtes, et non pas pour la recherche de réponses exactes. Tandis que [Lawrence and Giles, 1998] établissent à la main des règles de transformation pour générer des contextes de réponse tels que "`NASDAQ stands for`" et "`NASDAQ means`" à partir de questions comme *What does NASDAQ stand for ?*, [Agichtein and Gravano, 2000] utilisent de l'apprentissage automatique.

3 La génération et la recherche Web de contextes de réponse

Dans Quantum, la génération d'un contexte de réponse se fait en transformant la question en sa forme déclarative à l'aide de règles établies manuellement. Par exemple, la question *#1697 – Where is the Statue of Liberty ?*[1] est reformulée "`the Statue of Liberty is` <LIEU>". Nous nous sommes inspirés de 198 questions de TREC-8 et de 682 questions de TREC-9 pour bâtir les règles de transformation et nous avons utilisé 447 questions de TREC-10 et 454 de TREC-11 pour les tester. Nous avons éliminé de notre étude 112 questions pour lesquelles NIST ne fournit pas de réponse, soit parce qu'elles ont été retirées par NIST pour des raisons techniques, soit parce qu'elles n'ont pas de réponse dans la collection de textes de TREC.

Avant de procéder à la transformation, la forme grammaticale de la question est normalisée afin de limiter le nombre de cas à traiter. Par exemple, une question commençant par *What's ...* est normalisée en *What is ...* , *What is the name of ...* est changée en *Name ...* Au total, 17 règles grammaticales sont utilisées pour la normalisation.

Le tableau 1 illustre un exemple de règle de transformation. La règle est formée de 2 catégories de patrons : des patrons qui déterminent à quelle forme de question s'applique la règle, et des patrons qui définissent la forme de la réponse à chercher. Les patrons mettent en jeu des mots spécifiques (ex. `when`), des chaînes de caractères quelconques (représentées dans le tableau par SYNTAGME) et des étiquettes grammaticales (ex. VERBE-SIMPLE). Les patrons de réponse, que nous appelons *contextes de réponse*, utilisent les mêmes types de traits, en plus d'une spécification de la classe sémantique de la réponse (ex. <TEMPS>). Les classes sémantiques sont employées plus tard, après l'extraction des réponses, afin de baisser le poids de celles qui ne correspondent pas au type de réponse attendu. Dans l'exemple du tableau 1, la réponse doit être, de péférence, une expression de <TEMPS>, ce qui élimine des réponses comme "`The Jurassic period ended` **the same way the Tiassic period did**" (classe sémantique différente) ou "`The Jurassic period ended` **a long time ago**" (réponse trop générale). La version actuelle de Quantum comprend 10 classes sémantiques.

La transposition d'un verbe de sa forme simple (lorsqu'il est utilisé dans une question avec l'auxiliaire du passé *did*) à sa forme passée (dans la réponse, sans auxiliaire de temps) est un phénomène particulier de l'anglais qui ne peut être traité par la seule application de patrons. Afin d'effectuer cette correspondance rapidement, nous avons au préalable extrait

[1] Les questions précédées d'un numéro proviennent des campagnes TREC.

Règle de transformation	**Exemple**
Q : `When did` SYNTAGME VERBE-SIMPLE `?`	**Q :** *#22 – When did the Jurassic Period end ?*
R : SYNTAGME VERBE-PASSÉ <TEMPS>	**R :** `the Jurassic Period ended` <TEMPS>
R : <TEMPS> SYNTAGME VERBE-PASSÉ	**R :** <TEMPS> `the Jurassic Period ended`
R : <TEMPS>`,` SYNTAGME VERBE-PASSÉ	**R :** <TEMPS>`, the Jurassic Period ended`

TAB. 1 – Exemple de règle de transformation

tous les verbes répertoriés par le réseau sémantique Wordnet et nous avons construit une table de hachage permettant de faire correspondre la forme simple d'un verbe à sa forme passée.

Afin d'augmenter les chances de succès, Quantum peut chercher des contextes en *conjonction*. Par exemple, la réponse à la question *#670 – What type of currency is used in Australia ?* doit apparaître à la fois dans "SYNTAGME `is used in Australia`" et dans "SYNTAGME `is a type of currency`", peu importe si ces deux contextes sont présents dans des pages Web différentes. Les conjonctions sont introduites dans le cas de questions syntaxiquement plus complexes qui génèreraient des contextes plus longs ayant peu de chances d'être trouvés sur le Web. Dans la version actuelle du système, les conjonctions sont générées pour des patrons de questions pré-identifiés. Cependant, la décision de générer des conjonctions de contextes pour une question donnée pourrait être dynamique, par exemple en fonction de la longueur de la question ou du contexte.

Au total, 76 patrons de question ont été élaborés. Ils sont essayés l'un après l'autre sur une même question et tous ceux qui s'appliquent déclenchent la création des contextes de réponses correspondants. Le tableau 2 montre le nombre de patrons de question par type de question. Nous y apprenons que, par exemple, 6 patrons peuvent être employés pour transformer les questions de type *when* et que le nombre de contextes de réponse élaborés pour ce type de question s'élève à 9 (un patron de question de ce type engendre en moyenne 1,5 contexte de réponse). Les 76 patrons de question couvrent 93 % des 198 questions de TREC-8 et 89 % des 682 questions de TREC-9. Par couverture, nous entendons qu'au moins un patron de question est applicable. 409 contextes de réponse ont été produits suite à l'analyse des 185 questions de TREC-8 couvertes par les patrons de question, alors que 1209 contextes de réponse ont été produits suite à l'analyse des 610 questions de TREC-9 couvertes. Ainsi, en moyenne, nous obtenons 2 contextes de réponse par question.

Une fois les contextes de réponse générés, nous collectons les pages Web qui les contiennent à l'aide du moteur de recherche *Yahoo!*. Il suffit ensuite d'unifier l'inconnue d'un contexte avec un court extrait du texte. Quantum effectue ensuite des tests simples pour vérifier que le candidat ainsi extrait est du type sémantique recherché : par exemple, il met de côté les candidats qui ne débutent pas par une majuscule si le type de la réponse doit être un <LIEU>.

Afin de déterminer le meilleur de tous les candidats trouvés, un système de pointage a

Type de question	Patrons de question	Contextes de réponse	Moyenne
when	6	9	1.5
where	9	12	1.3
how many	12	13	1.1
how much	5	5	1.0
how (autres)	10	14	1.4
what	21	21	1.0
which	2	2	1.0
who	7	9	1.3
why	2	2	1.0
name	2	2	1.0
Total	**76**	**89**	**1.2**

TAB. 2 – Nombre de patrons de question et de réponse par type de question, et moyenne de contextes de réponse par patron de question

été mis en place. Un candidat qui est du type sémantique voulu débute avec un score de 0.6. Pour chaque occurrence additionnelle de ce candidat trouvée ailleurs dans un autre contexte de réponse, le score est augmenté ainsi :

$$score_{i+1} = score_i + \frac{1 - score_i}{2} \quad (1)$$

Ainsi, une deuxième occurrence du candidat fait grimper son score de 0.6 à 0.8, une troisième occurrence le fait grimper à 0.9, et ainsi de suite. Quant aux candidats qui échouent le test sémantique, ils débutent avec un score de 0.1. Ce score est augmenté de 0.1 à chaque nouvelle occurrence du candidat, et ce tant que le score n'excède pas 0.6 :

$$score_{i+1} = min(0.6, score_i + 0.1) \quad (2)$$

De cette façon, un candidat qui est extrait à l'aide d'un contexte de réponse mais qui ne semble pas être du type voulu n'a jamais un score plus élevé qu'un candidat sémantiquement valide, si un tel candidat est trouvé par le système.

4 Évaluation et discussion

La recherche de contextes de réponse sur le Web est un ajout récent au système de QR Quantum. Ce dernier s'appuie aussi sur d'autres techniques telles que l'analyse syntaxique de surface des questions, la consultation du réseau sémantique Wordnet, l'extraction d'entités nommées à l'aide du module *NE Transducer* de GATE [Cunningham et al., 2002] et l'apprentissage du poids de chacun de ces traits [Plamondon et al., 2002, Plamondon and Kosseim, 2002]. Cependant, nous nous concentrons ici sur l'évaluation du module Web utilisé seul.

Nous avons utilisé les questions de TREC-10 et de TREC-11 pour lesquelles nous disposions de réponses fournies par NIST afin de tester l'efficacité de notre approche de

Corpus	Nombre de questions	Couverture	Nombre de contextes générés
TREC-8 (développement)	192	93 %	409
TREC-9 (développement)	682	89 %	1209
TREC-10 (test)	443	91 %	664
TREC-11 (test)	454	87 %	722

TAB. 3 – Couverture des patrons de question pour le corpus de développement et le corpus de test, et nombre de contextes de réponse générés

recherche de réponses sur le Web. Les patrons de question couvrent 89 % des 897 questions du corpus de test avec pour effet d'engendrer un total de 1386 contextes de réponse à chercher sur le Web, soit 1.7 contexte de réponse par question en moyenne (tableau 3). La couverture obtenue avec le corpus de test est conforme aux prévisions basées sur le corpus de développement. Elle est particulièrement satisfaisante, considérant que seulement 76 patrons de question ont suffi à couvrir presque 90 % de toutes les questions qu'un utilisateur potentiel pourrait se poser. Notons à ce sujet que les questions de TREC sont courtes (jamais plus d'une phrase), qu'elles sont la plupart du temps exemptes d'erreurs syntaxiques et orthographiques, qu'elles attendent une réponse factuelle d'ordre général, que la majorité d'entre elles ont été réellement posées par des internautes sur des sites Web mais que NIST s'est réservé le droit de les filtrer et de les adapter aux besoins de TREC (voir [Voorhees, 2002] pour une description détaillée de la plus récente piste QR). Hormis ces précisions, aucune contrainte imposée aux questions n'aurait permis de prévoir la teneur et la structure de celles du corpus de test à partir du corpus de développement.

Si nous effectuons la recherche des contextes de réponse ainsi générés dans la collection standard de textes de TREC, qui fait environ 3 gigaoctets, nous trouvons au moins un contexte pour seulement 83 (10 %) des 800 questions du corpus de test couvertes par les patrons de question. Cependant, si nous effectuons plutôt la recherche sur le Web (un corpus nettement plus vaste), nous trouvons au moins une occurrence d'un contexte de réponse pour 354 questions (44 % des questions couvertes). Déjà, le potentiel du Web est évident. Le système trouve au moins une réponse correcte pour 227 (64 %) de ces 354 questions. En moyenne, 15 réponses par question sont extraites et lorsqu'elles sont ordonnées par le système selon leur qualité présumée, une réponse correcte ne se trouve au premier rang que pour 108 (49 %) de ces 227 questions. En clair, si on ne permet au système qu'une seule suggestion de réponse, elle est correcte dans 12 % des cas, et cela simplement en cherchant des contextes de réponse sur le Web sans aucun traitement supplémentaire (hormis quelques tests de validité sémantique de base). Le tableau 4 indique la performance du système à chacune des étapes du processus.

Il est difficile de cibler quelle étape du processus complet il conviendrait d'améliorer en priorité car chacune diminue les possibilités de succès d'environ 50 %. Il est cependant clair que de nombreuses améliorations peuvent être apportées à la version actuelle de Quantum à tous les niveaux. Tout d'abord, l'élaboration de plus de contextes de réponse et l'utilisation de davantage de conjonctions de contextes de réponse permettraient d'aug-

Type de question	Nombre de questions (corpus test)	Couverture des règles (a)	Au moins un contexte trouvé (b)	Au moins une réponse correcte (c)	Réponse correcte en 1^{er} (d)
when	93 (100 %)	83 %	26 %	18 %	14 %
where	63 (100 %)	87 %	41 %	33 %	25 %
how many	9 (100 %)	56 %	33 %	33 %	22 %
how much	12 (100 %)	33 %	8 %	8 %	8 %
how (autres)	48 (100 %)	73 %	27 %	10 %	6 %
what	548 (100 %)	93 %	41 %	25 %	10 %
which	25 (100 %)	88 %	0 %	0 %	0 %
who	93 (100 %)	98 %	65 %	45 %	20 %
why	4 (100 %)	0 %	0 %	0 %	0 %
name	2 (100 %)	100 %	50 %	50 %	0 %
Total	**897 (100 %)**	**89 %**	**39 %**	**25 %**	**12 %**

TAB. 4 – Pourcentage de succès après chaque étape du processus d'extraction de réponse, par type de question. Les pourcentages représentent la proportion des questions du corpus de test, par rapport à toutes celles d'un type donné, qui satisfont de façon cumulative aux conditions suivantes : (a) il existe au moins un patron de question qui s'applique et qui génère un contexte de réponse (b) au moins un de ces contextes a été trouvé sur le Web (c) au moins une réponse correcte a été trouvée sur le Web (d) une fois les réponses ordonnées, la première suggérée s'avère être correcte

menter le nombre de candidats trouvés (colonne b du tableau 4) et de mieux identifier les candidats hautement pertinents (colonne d). Aussi, il faudrait mesurer précisément l'apport des conjonctions de contextes. Les conjonctions ont été introduites de façon intuitive pour scinder les contextes qui nous paraissaient trop longs pour être trouvés sur le Web. Il serait intéressant de mesurer formellement l'apport de ces conjonctions en fonction de la longueur du contexte original.

Il nous semble très important d'améliorer aussi l'algorithme de pointage pour discerner les bons candidats des mauvais (colonne d). Ceci inclut l'amélioration des tests sémantiques et possiblement l'introduction de tests syntaxiques séparés. En effet, les tests devraient être capables de faire une analyse syntaxique et sémantique plus fine de la phrase dans laquelle le contexte est trouvé (colonne b). Par exemple, pour répondre à la question *#1697 – Where is the Statue of Liberty ?*, on cherche le contexte *the Statue of Liberty is <LIEU>* sur le Web et on peut trouver :

1. The arm of *the Statue of Liberty is* **42 feet long**
2. *The Statue of Liberty is* **recognized** as a symbol of freedom throughout the world
3. *The Statue of Liberty is* **a huge sculpture** that is located on Liberty Island.

Bien que les 3 phrases contiennent toutes le contexte de réponse cherché, aucun des candidats qui suivent le contexte ne constitue une réponse correcte. Le candidat de l'exemple 1 ne constitue pas une réponse appropriée car la tête du groupe nominal sujet n'est pas *Statue of Liberty* mais *arm*. Dans l'exemple 2, le type syntaxique du candidat ne correspond

pas au type syntaxique recherché : le candidat fait partie d'un groupe verbal alors que la réponse doit être un groupe nominal. Quant au candidat de l'exemple 3, c'est son type sémantique qui ne convient pas. Ces exemples montrent que la présence du contexte cherché dans la collection de textes ne garantit pas la validité du candidat, ce qui se traduit par une importante baisse de score entre les colonnes b et c du tableau 4.

La vérification de contraintes linguitisques est nécessaire pour bien distinguer les bons des mauvais candidats et mérite que l'on s'y attarde. En effet, lorsque l'on demande au système de choisir une seule réponse (colonne d) parmi tous les candidats qu'il a trouvés, il suggère une réponse correcte pour seulement 12 % des questions alors qu'il dispose d'au moins un candidat correct pour 25 % des questions.

Nous n'avons pas mesuré l'importance relative de chacune des sources d'erreurs mentionnées ci-haut, à savoir une mauvaise affectation ou vérificantion de la classe sémantique, la non-vérification de la classe syntaxique et de la tête du groupe nominal sujet. Cependant, l'affectation de la classe sémantique étant faite à l'aide de patrons développés manuellement, il nous semble que les erreurs proviennent surtout de la vérification *ad hoc*.

5 Conclusion

Nos expérimentations nous ont permis d'entrevoir le potentiel du Web pour la question-réponse. En effet, nous avons constaté que la taille gigantesque du Web rend possible la recherche de contextes de réponse très restrictifs, ce qui augmente d'autant les chances que la chaîne de caractères trouvée soit bel et bien la réponse à la question. Dans un corpus relativement petit comme celui de TREC ($\approx$ 3 Go), cette stratégie ne pourrait porter fruit car la probabilité de trouver un contexte particulier est nettement trop inférieure.

Nous avons généré les contextes de réponse en transformant la question en sa forme déclarative à l'aide de règles de réécriture simples. Les quelque 76 patrons de question et 89 contextes de réponse ont suffi à couvrir 89 % d'un ensemble de 897 questions issues des campagnes d'évaluation TREC. Ils ont permis de trouver une bonne réponse à 25 % d'entre elles ; cependant, l'algorithme de pointage des candidats doit être amélioré car la bonne réponse n'est suggérée en première position que dans 12 % des cas.

Les questions de TREC étant d'ordre général, le Web constitue une base documentaire tout indiquée pour les questions sur lesquelles a porté notre étude. De plus, elles sont courtes et exemptes d'erreur d'orthographe et de syntaxe, donc plus faciles à analyser mais pas nécessairement représentatives des questions posées par un utilisateur moyen. Il serait intéressant de tester la robustesse de notre système de question-réponse sur des questions authentiques. Le fait que les réponses aux questions de TREC soient factuelles joue aussi en notre faveur car les *comment ?* et les *pourquoi ?* sont exclus, les réponses sont courtes (rarement plus longues qu'un syntagme nominal) et elles ne nécessitent pas de regrouper ni de synthétiser des éléments tirés de plusieurs documents.

Remerciements

Nous tenons à remercier Louis-Julien Guillemette pour sa programmation et les relecteurs de

JFT pour leur commentaires. Ce projet a été financièrement soutenu par les Laboratoires Universitaires Bell et le Conseil de recherche en sciences naturelles et en génie du Canada (CRSNG).

Références

[Agichtein and Gravano, 2000] Agichtein, E. and Gravano, L. (2000). Snowball : Extracting Relations from Large Plain-Text Collections. In *Proceedings of the 5th ACM International Conference on Digital Libraries*, San Antonio, Texas.

[Brill et al., 2002] Brill, E., Dumais, S., and Banko, M. (2002). An Analysis of the AskMSR Question-Answering System. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, Philadelphie, Pennsylvanie.

[Brill et al., 2001] Brill, E., Lin, J., Banko, M., Dumais, S., and Ng, A. (2001). Data-Intensive Question Answering. In *Proceedings of The Tenth Text Retrieval Conference (TREC-10)*, pages 393–400, Gaithersburg, Maryland.

[Clarke et al., 2001] Clarke, C. L. A., Cormack, G. V., Lynam, T. R., Li, C. M., and McLearn, G. L. (2001). Web Reinforced Question Answering (MultiText Experiments for TREC 2001). In *Proceedings of The Tenth Text Retrieval Conference (TREC-10)*, pages 673–679, Gaithersburg, Maryland.

[Cunningham et al., 2002] Cunningham, H., Maynard, D., Bontcheva, K., and Tablan, V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphie, Pennsylvanie.

[Duclaye et al., 2002] Duclaye, F., Yvon, F., and Collin, O. (2002). Using the Web as a Linguistics Resource for Learning Reformulations Automatically. In *Proceedings of LREC*, pages 390–396, Las Palmas, Espagne.

[Lawrence and Giles, 1998] Lawrence, S. and Giles, C. L. (1998). Context and Page Analysis for Improved Web Search. *IEEE Internet Computing*, 2(4) :38–46.

[Magnini et al., 2002] Magnini, B., Negri, M., Prevete, R., and Tanev, H. (2002). Is It the Right Answer ? Exploiting Web Redundancy for Answer Validation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 425–432, Philadelphie, Pennsylvanie.

[Miller, 1995] Miller, G. (1995). WordNet : a Lexical Database for English. *Communications of the ACM*, 38(1) :39–41.

[Plamondon and Kosseim, 2002] Plamondon, L. and Kosseim, L. (2002). QUANTUM : A Function-Based Question Answering System. In *Proceedings of the 15th Conference of the Canadian Society for Computational Studies of Intelligence (AI 2002)*, pages 281–292, Calgary, Canada.

[Plamondon et al., 2002] Plamondon, L., Lapalme, G., and Kosseim, L. (2002). The QUANTUM Question Answering System at TREC-11. In *Notebook Proceedings of The Eleventh Text Retrieval Conference (TREC-11)*, pages 157–165, Gaithersburg, Maryland.

[Voorhees, 2002] Voorhees, E. M. (2002). Overview of the TREC 2002 Question Answering Track. In *Notebook Proceedings of The Eleventh Text Retrieval Conference (TREC-11)*, pages 115–123, Gaithersburg, Maryland.

Journées Francophones de la Toile - JFT'2003

Hypertextes et documents pour le web

Les hypermédias graphiques explorateurs

Les hypermédias cartographiques

D. BIHANIC
Laboratoire CERAP
(Centre d'Etude et de recherche en Arts Plastiques),
Université de Paris I Panthéon-Sorbonne
27-31 rue Lombard,
92260 Fontenay, FRANCE
Mail :root@fxdesignstudio.com
Tél : +33 2 99 32 32 23

Résumé

La navigation dans les hyperdocuments relève encore trop généralement d'une désorientation de l'utilisateur provoquée en partie par l'immensité et le désordre anarchique des liens. Afin d'y remédier, certains chercheurs ont imaginé de nouvelles manières de situer spatialement l'information en s'appuyant sur de nombreux procédés cartographiques, topologiques. Nous verrons ici comment ces initiatives ont favorisé l'apparition de nouvelles méthodes de représentation de l'information et également comment elles ont contribué à la création de nouveaux modèles perceptifs en milieu virtuel.

Abstract

Browsing within hypertext documents is still, and too often, related to the confusion of the user caused partly by the immenseness and the anarchistic disorder of links. In order to solve this problem, certain researchers imagined new ways of spatially locating information while basing themselves on many cartographic, topological processes. We shall see here how these initiatives favoured the coming of new methods for the representation of information and how they contributed to the creation of new perceptive models in a virtual context.

1 Introduction

Les premiers environnements non-linéaires, conçus au milieu des années soixante, ont très tôt tenté d'adapter certains procédés de repérage en vue de réguler la charge mentale mobilisée durant les opérations de traitement de l'information. Plus tard, d'autres techniques ont permis notamment d'identifier l'état des liens ou encore de positionner des marqueurs à plusieurs endroits stratégiques. Cependant, cela ne permit pas de résoudre efficacement certains problèmes persistants liés à ce qu'il convient d'appeler une désorientation cognitive. C'est pourquoi on voit apparaître de plus en plus massivement certains travaux de recherche s'intéressant aux représentations graphiques ou encore linguistiques qui calquent le réseau des liens d'informations, sous forme de hiérarchie, d'arbre ou encore d'association. Ces dispositifs d'orientation, de synthèse de l'information offrent de nouveaux modes d'exploration. Ils structurent l'information suivant un principe de positionnement grapho-spatial devant fournir une visualisation respectant les correspondances des données entre elles. Le terme « *mapping* », initié par G. Roman en 1993, définit alors « *une*

transformation d'un programme en représentation graphique ». Plus tard Jerding et Stasko démontreront l'intérêt de l'animation dans la description visuelle des fonctions opératrices.

Ces expérimentations d'ordre formel et également fonctionnel laissent entrevoir à plus long terme d'importantes modifications du point de vue des processus de diffusion, de traitement et enfin de production de l'information spatialisée. Elles sollicitent activement l'ensemble des mécanismes régissant notre compréhension spatio-temporelle en faisant apparaître de nouveaux comportements navigationnels au travers de situations d'interactivité, de parcours totalement originales. Il est indéniable qu'elles constituent une alternative efficace notamment aux interfaces de manipulation trop complexes comme celles qui utilisent des techniques de représentations multi-vues basées sur la superposition dynamique d'images. Ces hypermédias qu'il convient d'appeler ici hypermédias cartographiques ont comme principal objectif de proposer des représentations interactives qui soient entièrement destinées à favoriser l'appropriation des savoirs par l'utilisateur. Ils interrogent les propriétés d'une éventuelle iconicité dynamique cognitive supposant une plus grande contribution de la mémoire visuelle iconique, spatiale et également celle de la mémoire du geste en vue d'inaugurer une nouvelle phénoménologie de l'image.

2 L'exploration cartographique interactive

Les hypermédias cartographiques offrent la possibilité de localiser des objets visibles dans l'espace. Ils favorisent ainsi l'analyse des couches d'information et des données et permettent de prédire des résultats, de planifier des stratégies en vue de prendre des décisions éclairées. Ces environnements s'attachent donc à définir une véritable communication cartographique qui soit adaptée aux fonctions cognitives et d'appui à la prise de décision. Pour cela, ils déterminent le niveau de pertinence des données qui seront localisées en vue d'éviter toute surcharge cognitive et également de favoriser une certaine adaptabilité de l'information aux besoins de l'utilisateur.

Ce type d'hypermédia envisage donc les potentialités offertes par le réseau à la visualisation cartographique en vue d'améliorer le traitement et l'analyse de l'information. Par ailleurs, il démontre l'intérêt de visualiser un ensemble très large d'items pour accéder à l'un ou l'autre le plus rapidement possible et également de proposer, dans certains cas, des représentations adaptables aux besoins des différents utilisateurs dans l'objectif d'améliorer la compréhension du fonctionnement des processus spatiaux. L'enjeu majeur est donc de définir un concept de représentation graphique performant capable de traduire les nombreuses relations d'appartenance de diverses informations à priori hétérogènes en multipliant la puissance de la communication visuelle de données.

2.1 Les métamoteurs

De nombreux métamoteurs à l'étude utilisent aujourd'hui la technologie « *Flash* » afin de bénéficier d'une présentation dynamique. L'un des exemples les plus célèbres est sans aucun doute « *Kartoo* » [Kartoo, 2001]. Il convoque une organisation perceptuellement riche situant le contexte informationnel au travers d'un réseau de relations complexes. On y retrouve des boules de couleur dont le diamètre varie suivant la pertinence de l'hôte représenté. L'utilisateur peut alors choisir de visualiser les informations les plus pertinentes qui ont été retenues ou bien encore d'affiner plus en profondeur sa requête. Ce système exploite la recherche avancée, ce qui le rend très performant. Il possède également plusieurs atouts comme celui de permettre une vue globale des pages et de leurs relations grâce à des termes de recherche. Par ailleurs, il se base, à l'instar de « google », sur certains travaux en scientométrie de Price et Garfield visant à considérer qu'une citation est une indication suffisante en vue de garantir la qualité d'un article. Ainsi, une page "cible" devrait donc refermer une information nécessairement utile.

Certains s'accordent à penser que ce type de modélisation formelle de l'information et de la correspondance empruntant certaines associations métaphoriques ou encore métonymiques vise à encourager la projection et l'imaginaire individuel. Il ne fait que complexifier la charge de travail de l'utilisateur et n'offre à aucun moment de véritables techniques de sélection de l'information pertinente pouvant prétendre à améliorer la prise de décision. A ceux-ci, il convient de répondre que le concept d'hypermédia cartographique appliqué ici à la fouille de données facilite tout d'abord le traitement d'une relativement grande quantité d'informations au travers d'une visualisation et d'une interaction conviviale et performante. Il sollicite, par conséquent, activement les capacités psycho-perceptives de l'utilisateur en induisant une perception, une compréhension ou encore une interprétation spatiale du contexte graphique de recherche. La notion d'un déplacement, d'une déambulation au cœur du contexte informationnel constitue alors les fondements d'une entreprise plus vaste visant à interroger notre propre représentation mentale des connaissances. Ces environnements entretiennent donc une analogie significative à l'organisation mentale de l'utilisateur en vue d'améliorer ses capacités d'analyse et de traitement. Analogie et spatialisation deviennent dès lors étroitement liées, laissant apparaître de véritables représentations métaphoriques donnant raison aux théories de Lakoff et Johnson sur la détermination de notre expérience perceptivo-motrice.

Fig1. « *Kartoo* » est un métamoteur représentant les liens sémantiques entre les sites récoltés.

D'autre métachercheurs, comme « *Mapstan serach* » [Mapstan Search, 2001], développent des dispositifs de synthèse visuels au sein desquels chaque page référencée est regroupée par site. Ils parviennent ainsi admirablement à articuler les résultats ainsi que les liens de proximité envisagés et également à éviter d'innombrables multiplications des requêtes en vue d'accéder plus rapidement à l'information recherchée. Dans le cas de « *Mapstan* search», une technologie de cartographie personnalisée a été mise au point dénommée « *Web Positioning System™ (WPS)* ». Elle permet notamment d'associer la sélection à d'autres recherches similaires menées par l'ensemble de la communauté de « *mayeticVillage* »*. Il en résulte alors une nouvelle forme de travail collaboratif en matière de recherche d'information.

* Le « *mayeticVillage* » inaugure un nouveau type de société, entièrement axée sur les nouvelles technologies et les organisations d'entreprise. Il s'organise au travers de la création d'espaces de travail collaboratif sur le web où sont représenté plus de 25.000 membres, 950 sociétés référencées et 83 pays utilisateurs.

Ce système, très similaire à celui de « *Kartoo* », obéit également à certains critères ergonomiques assurant la cohérence des moyens de locomotion ainsi qu'une certaine souplesse de l'interaction. Le mode de visualisation satisfait ainsi le même souci d'observabilité de l'espace de recherche. Il convient donc d'assister l'utilisateur face à une importante quantité d'informations en améliorant, en premier lieu, l'utilisabilité du système. La notion de Non-préemption chère à Joëlle Coutaz et Laurence Nigay, au sens d'une absence de *« contraintes dans la trajectoire interactionnelle* », s'impose ici comme un avantage majeur. L'utilisateur peut alors obtenir plus efficacement l'information recherchée sans risquer de se perdre au travers d'une métrique trop longue.

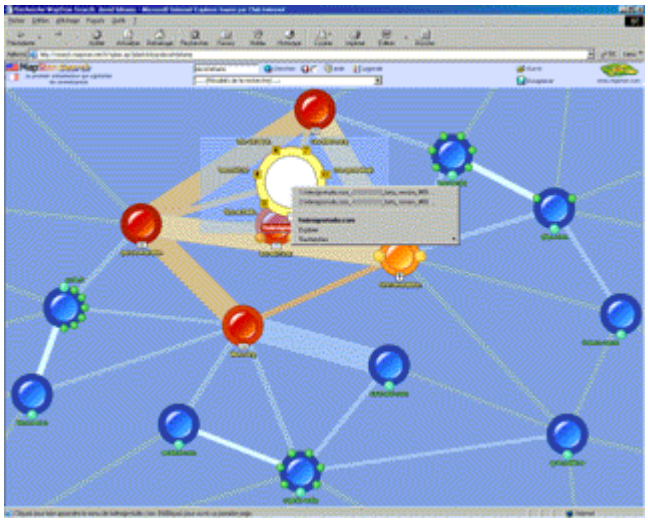


Fig2. « *Mapstan Search* » est un métamoteur offrant une recherche d'information enrichie par capitalisation.

Ces différentes expérimentations souhaitent démontrer la prééminence de la représentation d'une structure générale des données dans le cas prioritairement de la recherche d'information car cette tâche implique nécessairement une vue d'ensemble ainsi qu'une analyse sémantique des associations ou encore des relations entre les différentes entités représentées. Pour cela, ellles se fondent sur plusieurs théories rapportées notamment par Kulhavy, Schawarz et Sahha démontrant que la configuration réticulaire d'une carte facilite considérablement la compréhension. Le repérage visuel y est plus efficace et la mémorisation devient tributaire de la proéminence des objets. D'autre part, Elles s'appuient également sur les théories de Rossano et de Morrisson en développant un cheminement d'apprentissage stratégique guidé par la structure de la carte.

Sur le plan de la recherche documentaire, ce type de méta-moteurs offre des solutions tout aussi satisfaisantes, pour le moment, que les cartes auto-organisantes qui classent automatiquement des documents liés entre eux ou que toutes autres méthodes de "clustering" uniquement basées sur un algorithme. Il améliore considérablement le taux de rappel au travers de la reformulation de la requête de recherche par l'emploi de traitements lexicaux, d'un dictionnaire de synonymes, voire de traitements linguistiques. A ce propos, il faut rappeler l'excellente initiative d'Altavista en implémentant une fonction « Refine » couplée à une interface graphique, appelée Graph, permettant de trier rapidement les réponses obtenues lors d'une requête en proposant d'autres mots-clés en rapport avec l'interrogation de départ.

Cependant, il semble néanmoins que les réseaux classificateurs neuro-mimétiques non supervisés représentent l'avenir de la cartographie informationnelle pour la recherche et

l'exploration. Ils offrent non seulement la possibilité de faire émerger des classes sans avoir à formuler de critères mais en plus ils détectent les similitudes entre classes. Parmi les modèles les plus répandus, on retrouve ART, Néo-cognitron, SOM — SOM ("Self Organising maps"), souvent appelés « *réseaux de Kohonen* » permet de rapprocher les objets les plus semblables sur la carte offrant ainsi une visualisation claire des regroupements. L'utilisateur peut alors infléchir sur la géométrie du réseau en manipulant, comme pour une structure élastique, les données représentées.

2.2 Les sites d'information

Dans le cas du projet « *Webmap* » [Webmap, 2001], L'hypermédia devient un espace d'information organisé au travers d'une représentation topologique 2D multi-niveaux à forte teneur interactive. L'utilisateur peut alors "zoomer" sur le secteur d'activité de son choix et solliciter davantage de précisions sur l'information recueillie. L'ensemble des données est ici représenté sous forme symbolique. Il s'agit alors de petites montagnes aux sommets enneigés autour desquelles gravitent plusieurs autres icônes venant apporter des renseignements complémentaires sur la nature des données récoltées.

La particularité de ce type de représentation réside dans sa capacité à saisir, grâce à des techniques originales d'analyse et de placement des données dans l'espace, plusieurs niveaux de détails. A ce propos, il est intéressant de noter que la variabilité de l'échelle de perception revêt une importance capitale dans le processus de traitement de l'information spatialisée en milieu virtuel. Il devient alors possible de décrire les relations et structures de base de l'hypermédia mais également de définir la composition, les modifications et transformations éventuelles de l'information s'y trouvant.

Les lois d'une communication graphique se substituent dès lors à une communication d'ordre cartographique laissant apparaître une triade temps, espace, dynamique au sein de laquelle l'utilisateur explore, compare des multiples données accessibles interactivement. L'espace alloué est ainsi réglé par des transformations, des perturbations d'ordre logique ou encore "opto-logique" s'ajustant aux différents niveaux de détails ainsi qu'au degré d'intérêt exprimé par l'utilisateur.

Par conséquent, les hypermédias cartographiques s'imposent comme une des plates-formes interprétatives les plus probantes. Ces environnements perceptifs dynamiques fournissent alors une définition explicite et précise de la notion de multimodalité de traitement tout en réaffirmant la prise en compte d'une contextualisation à la fois de l'action et de l'information.

Malgré tout, on est en droit d'être dubitatif quant à la capacité de ces environnements à pouvoir garantir l'ensemble des fonctions de représentation dont fait état, de manière tout à fait exhaustive, Michel Denis et qui apparaissent indispensables à tout système formel visant à articuler une structure organisatrice et phénoménale de l'information. Elles doivent donc permettre notamment « *de rendre accessibles des informations qui ne le sont pas dans les conditions normales, "naturelles", de perception* [c'est le cas ici au travers d'une répartition schématique des groupements sémantiques] ; *d'expliciter de l'information* [Comme le souligne Daniel Peraya : "[…]*elles sont susceptibles de remplacer les objets originaux pour effectuer certaines fonctions, principalement de nature cognitive"*] ; de *guider*, d'*orienter* et de *réguler* mais encore de *systématiser*, de *transmettre* et de *communiquer* une information. » [Denis, 1989].

Cette difficulté semble en partie dû au fait qu'il s'élabore ici de véritables environnements scripto-visuels, régis par certains processus complexes d'échanges d'informations, interindividuels ou sociaux, se risquant à une articulation difficile entre une fonctionnalité symbolique de l'image et une volonté d'exploitation rationnelle des signes

iconiques en vue précisément de « *guider*, d'*orienter* et de *réguler* » les actions de l'utilisateur.

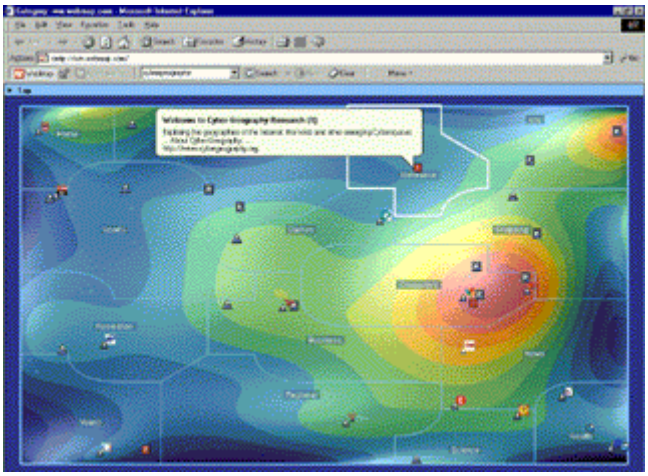


Fig3. « *Webmap* » est un site d'information réunissant plus de 2 millions de liens URL.

« *Map.net* » [Map.net, 2000] est un site d'information très complet ayant choisi de développer une représentation schématique hypertextuelle 2D multi-niveaux d'un large réseau de relation informationnelle. On y retrouve plusieurs rectangles, de tailles variables suivant l'importance du domaine représenté, accolés entre eux de manière à signifier les potentialités d'échange entre les disciplines. L'utilisateur est donc invité à les parcourir au travers d'une succession de "clics" l'amenant à préciser sa requête.

Ce type d'expression cartographique introduit indubitablement une dimension conceptuelle nouvelle faisant référence à une véritable problématisation de l'espace. Elle permet une meilleure compréhension des glissements sémantiques présupposés de l'information en insistant sur la construction analytique d'une représentation de l'espace. Le mode cartographique se révèle donc être le lieu d'appréciation idéal des interconnexions entre différentes sources d'information distantes. Il s'élabore ainsi une tactique d'appropriation de l'espace permettant d'organiser un véritable territoire relationnel, une sorte d'espace de médiation thématique. Désormais, le développement d'une répartition locale appelle une réévaluation de la structure d'organisation globale de l'information. Cela nécessite donc une procédure d'articulation des échanges interdisciplinaires définissant la territorialité comme un environnement structurant.

Par ailleurs, ce dispositif tend précisément à démontrer que les hypermédias sont avant tout des lieux sociaux d'interaction et de coopération procédant d'une intention discursive. Ce sont des systèmes de communication médiatisée où s'entremêlent plusieurs formes d'énonciation veillant à manifester un point de vue délibérément critique sur le monde qui nous entoure. Cette composante fondamentale de toute manifestation du langage n'a bien sûr pas échappé à J.P. Desgoutes qui nous rappelle que le discours est composé de deux champs complémentaires : *« le champ référentiel, informatif ou constatif, et le champ relationnel, que l'on peut qualifier de performatif ou d'énonciatif.* » [Desgoutes, 1999]. Plus loin, il explique que ces deux champs relèvent de deux fonctions essentielles dissociant l'appareil communicant de l'expression subjective.

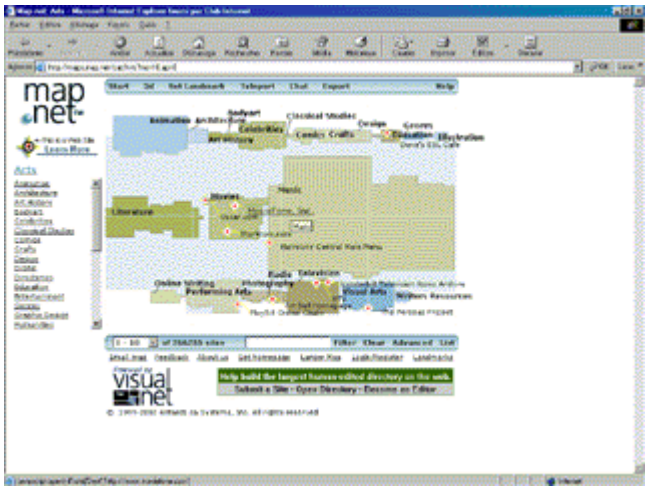


Fig4. « *Map.net* » est un site d'information répertoriant des adresses URL issus de l'Open Directory.

Du même ordre, il nous faut citer également l'excellent projet « *HistoryWired* » [HistoryWired, 2001] proposant une carte précise de la collection d'objets du musée national d'histoire américain, élaborée à partir des données du marché des technologies. Cette application utilise une technique de visualisation bien connue appelée « *Treemap* ». Elle consiste en un découpage morpho-spatial laissant apparaître alors une construction hiérarchique arborescente au sein de laquelle s'opère, comme le souligne Moutaz Hascoët et Michel Beaudouin-Lafon, « *un traitement surfacique* ».

Le mode de représentation choisi laisse ici une plus large part à l'abstraction au profit d'un repérage plus efficace des zones de correspondance de l'information. Les différents niveaux de profondeurs sont ainsi obtenus dynamiquement au travers d'un véritable "zoom" sémantique permettant d'accéder à une répartition plus détaillée. Il en résulte une diminution des effets de décontextualisation dûe en partie au principe de visualisation progressive de l'information pertinente.

Ce type d'espace multi-échelle, multi-dimensionnel reste, à mon sens, le paradigme de présentation et d'interaction le plus efficace. Il repose, comme le précise Mountaz Hascoët, « *sur le couplage, entre une transformation graphique élémentaire et une modélisation des données* ». La fonction de "zoom" devient alors « *l'élément graphique tangible »* facilitant non seulement la localisation de la position courante et des informations recherchées mais également l'exploration "déambulatoire" destinée à favoriser des découvertes "accidentelles".

Le système propose ici une sorte de schéma topographique relationnel très efficace mettant en évidence la complexité des liens. Il fait également émerger le sens inhérent de la masse d'information en proposant une synthèse graphique inter-objets. L'utilisateur opère alors sur un plan segmenté ouvrant vers une distribution spatiale de renseignements aussi bien qualitatif et quantitatif. Ce système s'intéresse particulièrement à favoriser la localisation de données ordinales grâce à une analyse hiérarchique exploratoire.

Le type d'interfaçage spécifique de gestion cartographique est conçu sur la base du matériau langagier. Les métadonnées d'ordre textuel deviennent dès lors les acteurs de l'espace d'interaction et de navigation. Cet environnement se fonde sur l'hypothèse d'une économie des modalités de représentation graphique en vue d'obtenir un meilleur

découpage fonctionnel de l'espace. Par conséquent, il choisit de se focaliser sur la fonction et l'utilité d'un développement minimal de manière à accéder à la spécificité du langage cartographique comme moyen de visualisation et de communication. Le message cartographique, véhiculé au travers d'opérateurs de base, se réalise ainsi via une carte mentale permettant de structurer la pensée et également gérer de manière stratégique les connaissances critiques.

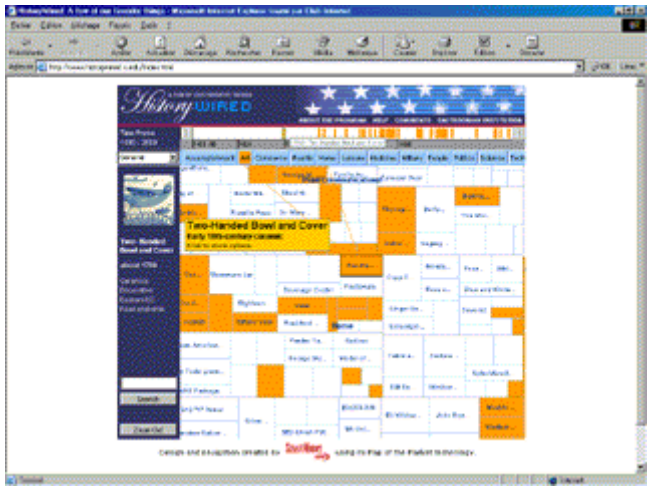


Fig5. « *HistoryWired* » est site d'information présentant la collection d'objets du musée national d'histoire américain.

Un des projets les plus réussis de Martin Wattenberg, « *Map of the Market* » [Map of the Market, 2001] propose une cartographie des performances boursières américaines au travers d'une représentation poly-composante ordonnée également par zone. Plusieurs centaines d'entreprises y sont regroupées sous forme de rectangles dont le format varie cette fois suivant la capitalisation. Le code couleur, quant à lui, renseigne l'utilisateur sur les modifications du prix de l'action.

Cet hypermédia ne déroge pas à la règle élémentaire qui veut que toute représentation cartographique soit déterminée en priorité par la nature de la mesure. En effet, la répartition chorochromatique, dû au caractère nominal de la variable, définit les spécifications de la carte et également sa mise en forme en langage graphique. On voit apparaître alors des objets zonaux stimulant notre sensibilité chromatique différentielle afin de faciliter le décryptage des variables visuelles et ainsi de favoriser la lecture de certaines informations. On remarque par là même que la programmatique colorée revêt une importance majeure car elle détermine très fortement les modes d'appréhension du système de repérage visuel.

Ainsi, le mode de perception graphique est investi par l'étude du traitement des signaux sensoriels. Le langage cartographique est alors une expression visible où s'entremêlent diverses sensations physiques obéissant à certaines conditions morpho-dispositionnelles de lecture.

Toutefois, il est légitime de s'interroger sur la pertinence de telles représentations d'origine perceptive dans le cadre du processus d'extraction des connaissances et d'approfondissement des calculs, des simulations, des inférences, des comparaisons en phase de post-traitement. En effet, la dénotation des similitudes et des différences entre les diverses informations visuelles, leur reconnaissance, association ou bien catégorisation pourrait, à priori, entraîner une trop forte augmentation de la difficulté de la tâche conduisant ainsi à une détérioration des performances. Or, de nombreuses expériences nous

confirment que cette ambivalence de traitement constitue effectivement deux types de représentation mentale pourvus de propriétés fortement différenciées, mais dont la coopération est attestée dans de nombreuses formes du fonctionnement cognitif. Ainsi, l'étude de leurs éventuelles interactions devient capitale en vue de garantir une parfaite exploration, focalisation mentale sur des objets visualisés.

Fig6. « *Map of the market* » est site d'information représentant les échanges boursiers américains.

3 Conclusion

Plusieurs études démontrent que l'activité spatiale qu'implique la navigation cartographique tend à améliorer considérablement la capacité de sélection et de traitement de l'information grâce à une meilleure compréhension du fonctionnement des processus spatiaux. En effet, il est indéniable que leur habileté à faire la saisie et la gestion des données spatiales, des requêtes concourent à une meilleure orientation de l'utilisateur au sein de l'environnement favorisant ainsi la fouille de données.

Ce type d'hypermédia graphique explorateur relève, comme le précise Claire Bélisle : « *d'une activité de méta-lecture facilitant une maîtrise réflexive de l'information*. […] *Circuler dans des informations ayant des contextes et environnements très divers oblige à repérer ces différences et à se situer par rapport à elles*. » [Bélisle, 1999]. Ils incitent, en augmentant la puissance de la communication visuelle de données, à une réelle représentation dynamique de l'information. Cette méta-lecture dont parle Claire Bélisle, est d'ordre métacognitive. Elle inaugure une nouvelle vision expériencielle de l'information hypermédiatique, sorte de phénoménologie d'une interactivité environnementale, capable de renouveler en profondeur les comportements stratégiques en matière de recherche d'information.

Cependant, certains chercheurs persistent à penser que les hypermédias cartographiques peuvent susciter certains problèmes de compréhension en monopolisant l'attention de l'utilisateur sur la macrostructure des échanges informationnels. À ce propos, il faut rappeler que la représentation acquise du système par l'utilisateur est une condition indispensable de toute organisation de l'information structurée et que par conséquent il est donc primordial de posséder une bonne connaissance de la configuration de l'espace en vue d'accéder précisément à l'information recherchée. Autrement dit, cette démarche conceptuelle ne se détourne en aucune façon des objectifs premiers d'un hypermédia explorateur qui est de faciliter l'acquisition de connaissances. Ces systèmes pêchent peut-être aujourd'hui encore en matière d'évaluation sémio-cognitve des modes de visualisation

et de navigation, mais il est certain qu'ils ne tarderont pas à devenir l'un des terrains d'étude privilégié en matière de recherche documentaire et peut-être même, comme le souligne Thierry Chanier, « *d'apprentissage collaboratif, voire social* ».

Références

[Bélisle, 1999] Bélisle C. (1999). *La navigation hypermédia : un défi pour la formation à distance*. Revue de l'enseignement à distance: 14,1.
Adresse URL (Iuicode) : <http://www.icaap.org/iuicode?151.14.1.3>

[Denis, 1989] Denis M. (1989). *Image et cognition*. PUF, Paris.

[Desgoutes, 1999] Desgoutes J.P. (1999). *La mise en scène du discours audiovisuel*. L'Harmattan, Paris.

[HistoryWired, 2001] Site d'information, « *HistoryWired* », lancement en 2001 par la société « *SmartMoney Inc.* » (Chef de projet : Martin Wattenberg). [en ligne], (page consultée le 13/01/03),
Adresse URL : <http://www.historywired.si.edu>

[Kartoo, 2001] Métamoteur de recherche, « *Kartoo* », lancement en avril 2001 par une société française de Clermont-Ferrand. [en ligne], (page consultée le 13/01/03), Adresse URL : <http://www.kartoo.com>

[Map.net, 2000] Site d'information, « *Map.net* », lancement en 2000 par la société « *Antarctica Systems Inc.* ». [en ligne], (page consultée le 13/01/03),
Adresse URL : <http://map.net/cat/>

[Map of the Market, 2001] Site d'information, « *Map of the market* », lancement en 2001 par la société « SmartMoney Inc ». [en ligne], (page consultée le 13/01/03),
Adresse URL : <http://www.smartmoney.com/marketmap/>

[Mapstan Search, 2001] Métamoteur de recherche, « *Mapstan Search* », lancement en décembre 2001 par une société française du même nom. La technologie utilisée par « *Mapstan* » a été appelée le « *WPS* » ("Web Positioning System"). [en ligne], (page consultée le 13/01/03),
Adresse URL : <http://www.mapstan.net>

[Peraya, 1996] Peraya D. (1996). *L'image: une troublante analogie*. Chronique d'images, Journal de l'enseignement primaire, n° 54, janvier/février, 34-35.

[Webmap, 2001] Site d'information, « *Webmap* », lancement en 2001 par une société du même nom. [en ligne], (page consultée le 13/01/03),
Adresse URL : <http://www.webmap.com>

Annotations sur le Web : types et architectures

E. Desmontils, C. Jacquin, L. Simon

IRIN,
2 rue de la Houssinière, BP 92208
44322 Nantes, Cedex 3, FRANCE.
Email : {desmontils,jacquin,simon}@irin.univ-nantes.fr
Tél : +33 2 51 12 58 33 Fax : +33 2 51 12 58 12

Résumé

De nombreux systèmes de partage d'information existent de nos jours mais les spécificités du Web en font des outils extrêmement difficiles à exploiter. Les outils d'annotation visent à améliorer échange, communication et interopérabilité sur le Web. L'objectif de cet article est d'une part, de faire une synthèse des caractéristiques des annotations et des architectures des systèmes d'annotation et d'autre part, de proposer une nouvelle architecture pour un système d'annotation simple d'utilisation, léger, efficace, non-intrusif, évolutif, partagé et indépendant d'une plate-forme.

Abstract

Many systems of information exchange exist nowadays but specificities of the Web make these tools extremely difficult to exploit. Annotation tools aim at improving exchange, communication and interoperability on the Web. The aim of this paper is on the one hand, to make a synthesis of the characteristics of the annotations and architectures of annotation systems, and in addition, to propose a new architecture for an annotation system which is simple of use, light, effective, not-intrusive, evolutionary, shared and platform independent.

1 Introduction

De nombreux systèmes de partage d'information existent : cela va du Web aux outils très évolués de travail collaboratif (Lotus Notes...). Ces derniers visent un groupe réduit de personnes travaillant souvent ensemble, avec un vocabulaire commun, sur des thèmes proches et donc avec des habitudes spécifiques. Cependant, ce qui est possible avec les outils de travail collaboratif ne l'est plus avec le Web. Par contre, le Web concerne potentiellement des millions de personnes non seulement ne se connaissant pas mais ayant en plus des centres d'intérêt différents, des habitudes différentes, des cultures différentes... De plus, sur le Web, l'information est fortement distribuée, extrêmement volumineuse, évolutive, volatile, très "bruitée", très hétérogène et souvent très peu structurée. Dans ce contexte, il est nécessaire de proposer des méthodes et des outils pour comprendre, manipuler et partager des documents, pour mettre en place des services pertinents et performants. Dans ce but sont nés les outils d'annotation qui visent à améliorer l'appréhension des documents HTML ainsi que la communication et l'interopérabilité sur le Web. Deux types de systèmes d'annotation préexistent, l'un repose sur des annotations sémantiques et l'autre sur des annotations libres. Les annotations sémantiques sont des méta-données basées sur des ontologies et ajoutées au document. Elles sont le plus souvent exploitées par des systèmes automatiques à des fins de construction de réponses à des questions. Les annotations libres quant à elles, permettent d'associer des notes de lectures aux documents, de partager de l'information, d'effectuer des tâches rédactionnelles en groupe... Grâce aux systèmes d'annotation, le lecteur devient aussi rédacteur. Plus généralement, dans un contexte comme le Web, le système passe du "one-to-many" (un rédacteur et des millions de lecteurs) au "many-to-many" (tout utilisateur du Web est Lecteur/Rédacteur) [Zohar, 1999].

Dans cet article, notre objectif est premièrement de présenter ce que sont les annotations et quels sont leurs rôles dans la communication lecteur/rédacteur (section 2). Ensuite, nous présentons et discutons les différents types de systèmes d'annotations libres sur le Web qui permettent d'assurer cette communication entre des acteurs (section 3). Finalement, en prenant en compte l'analyse précédente, nous proposons un nouveau type d'architecture pour les systèmes d'annotation qui pallient aux problèmes rencontrés avec les systèmes existants (sections 4 et 5).

2 Les annotations

Une annotation est une information graphique ou textuelle attachée à un document et le plus souvent placée dans ce document. Cette place est donnée par une ancre. Les annotations font référence à des entités diverses : un ensemble de documents, un document, un passage, une phrase, un terme, un mot, une image...

C. C. Marshall [Marshall, 1998] propose de caractériser les annotations selon différentes dimensions. Ces dimensions sont des espaces mono-dimensionnels continus. Elles décrivent les propriétés de l'annotation au niveau de sa structuration, de sa fonction et de son rôle dans la communication rédacteur/lecteur. Nous allons décrire ces dimensions en spécifiant les cas extrêmes, mais il est souvent possible de trouver des exemples d'annotation "intermédiaires".

2.1 Dimensions liées à la structuration de l'annotation

Une première classe de dimensions concerne la structuration de l'annotation. La première dimension proposée décrit le niveau de formalisation de l'annotation : c'est la dimension For-

melle - Informelle. Les annotations sont représentées de manière plus ou moins structurée. Les notes peuvent aller du renseignement de champs spécifiques (par exemple par réponse à un questionnaire) au texte en langage naturel (mais aussi être des sigles intuitifs...). Les annotations sémantiques, quant à elles, sont par nature formelles (reposent sur une connaissance définie à priori - une ontologie - qui est représentée à l'aide d'un langage spécifique)

Une autre dimension concerne la signification intrinsèque de la note, c'est-à-dire entre l'explicite et l'implicite. Une annotation explicite se suffit à elle-même (destinée à une autre personne que le rédacteur implicite demande une connaissance complémentaire (table de lecture...) et est destinée à un lecteur instruit des conventions adoptées (souvent le rédacteur lui-même). Les annotations sémantiques, quant à elles, étant donné qu'elles font référence à une ontologie sont implicites.

Il faut noter que certaines annotations peuvent "se déplacer" sur les axes des dimensions au cours du temps. Par exemple, les smiley ("frimousse" en français !) étaient au départ des annotations informelles et sont devenues (à travers tous les livres sur Internet) des annotations formelles (des règles de construction "normalisées" sont proposées...).

2.2 Dimensions concernant les fonctions de l'annotation

Une classe de dimensions concerne les dimensions décrivant les fonctions des annotations. La première dimension concerne l'utilisation des annotations dans les processus de lecture et de rédaction. Pour aller plus loin que C. Marshall, nous pensons que les annotations ont cinq utilisations majeures :

1. *information, illustration, extension du document* : lorsque le lecteur rédige ses annotations, il devient alors rédacteur[1],
2. *forum* : en permettant à un ensemble de lecteurs de débattre sur le document,
3. *opérationnalisation de l'information* : les annotations sémantiques permettent une opérationnalisation de l'information contenue dans des documents. En effet, elles sont destinées à être traitées par des machines (par opposition aux annotations libres en langage naturel ou composées de symboles souvent implicites). Leur objectif majeur est de *désambiguïser* le document pour un traitement automatique.
4. *aide au processus rédactionnel* : en permettant d'indiquer des consignes de rédaction (corrections, mouvements d'informations...),
5. *support de lecture* : (la mise en évidence de passages importants...) permettant l'appropriation du texte par le lecteur, l'annotation est alors le reflet de l'engagement par rapport à un texte d'un lecteur qui le personnalise (trace de lecture) afin de faciliter un futur retour.

C. C. Marshall propose aussi une dimension concernant le rôle de l'annotation par rapport au niveau de lecture du document. Certaines notes (hyper-liens par exemple) permettent une lecture extrêmement superficielle du document (surf). À l'opposé, des annotations sont utilisées pour une lecture approfondie d'un texte[2]. Entre les deux, il y a les notes de lecture rapide comme l'utilisation de la typographie (mise en gras ou en italique par exemple) ou le surlignage.

C. C. Marshall propose aussi une dimension qui concerne la "durée de vie" de l'annotation, c'est la dimension temporelle.

[1] Ce type d'annotation est présente dans beaucoup de systèmes d'annotation sur le Web. C'est une plus-value importante sur le document.

[2] Cf. rôle des notes dans les œuvres de Tolkien par exemple.

2.3 Dimensions concernant le rôle de l'annotation dans la communication rédacteur/lecteur

Une dimension particulièrement importante concerne la relation lecteur/rédacteur, autrement dit sur le choix des destinataires potentiels de l'annotation. Une annotation peut être soit privée, c'est-à-dire que le rédacteur la destine à lui-même, soit publique. Dans ce dernier cas, plusieurs "degrés" sont possibles : le groupe de travail, l'institution... jusqu'au niveau le plus global. Le comportement du rédacteur et des lecteurs vis à vis du document et des annotations mène à des systèmes de type "one-to-many" (un rédacteur destine son document à un ou plusieurs lecteurs) ou "many-to-many" (un groupe de lecteurs/rédacteurs travaille en collaboration). Les systèmes d'annotation sur le Web visent à atteindre ce dernier cas.

Après avoir présenté les annotations, leurs fonctions et leur rôle dans la communication rédacteur/lecteur, nous nous attachons dans la section suivante à analyser les systèmes d'annotations libres existant sur le Web qui assurent la communication entre les divers acteurs.

3 Les systèmes d'annotations libres sur le Web

Les outils d'annotation libre doivent prendre en compte un certain nombre de contraintes et particularités du Web : les acteurs (lecteurs et serveurs Web) sont répartis, les communications se font par le réseau, le système est fondamentalement multi-utilisateurs (utilisateurs par ailleurs très nombreux), le langage de communication (HTTP), les données au format HTML ou XML... Globalement, tous les systèmes respectent le même schéma d'architecture (figure 1) [Vasudevan and Palmer, 1999] : un intermédiaire "observe" les transactions entre le client Web et les serveurs Web. Cet intermédiaire agit sur la requête, les pages obtenues et, éventuellement, sur les événements issus du navigateur. Cet élément est composé d'un intercepteur qui est chargé de récupérer requête et/ou pages HTML, d'un composeur qui se charge d'associer aux pages les annotations attachées (présentes dans une base de données). Cette combinaison peut dépendre du profil de l'utilisateur.

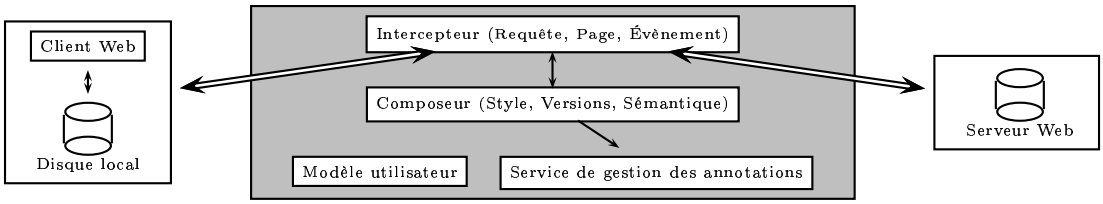


FIG. 1 – Schéma général d'un outil d'annotation sur le Web ([Vasudevan and Palmer, 1999])

En pratique, les systèmes d'annotations libres sur le Web se divisent en deux grandes catégories : ceux basés sur des serveurs mandataires ("proxy") et ceux utilisant un intermédiaire attaché au navigateur (intermédiaire client).

Les intermédiaires de type serveur mandataire [Palme, 1998, Ovsiannikov et al., 1999, Yee, 1997] sont des serveurs indépendants du client et des serveurs Web. Dans la configuration la plus standard, le serveur mandataire observe toutes les requêtes du client et gère la page ayant des annotations présentes dans sa base. Avec un tel système, toutes les transactions du client doivent passer par ce serveur posant ainsi un problème de confidentialité. De plus, ces systèmes posent aussi

des questions autour du droit d'auteur. En effet, certains sites refusent que leurs pages soient modifiées. Pour éviter ces deux problèmes, [Yee, 1997] (CritLink) propose un système basé sur un proxy facultatif (figure 2). L'utilisateur fait appel au système d'annotation uniquement quand il le désire (confidentialité) et les pages des sites peuvent être consultées sans les annotations (pointillés dans la figure 2) [3]. Ce type de serveur d'annotations pose d'autres problèmes :

- il n'est pas possible d'annoter des documents locaux,
- l'ajout d'annotations passe nécessairement par un dialogue spécifique (parfois assez lourd) et par un rechargement de la page,
- il ne sait pas gérer les pages dynamiques,
- il est lent (goulot d'étranglement) puisque toutes les requêtes doivent passer par lui,
- le document d'origine est forcément modifié (les notes sont insérées dans le document).

Les serveurs mandataires ont tout de même un certain nombre d'avantages non négligeables : le partage d'annotation est facile, les outils sont indépendants du système d'exploitation et du navigateur utilisés. De plus, ils sont faciles à installer (et il n'y a rien à désinstaller !).

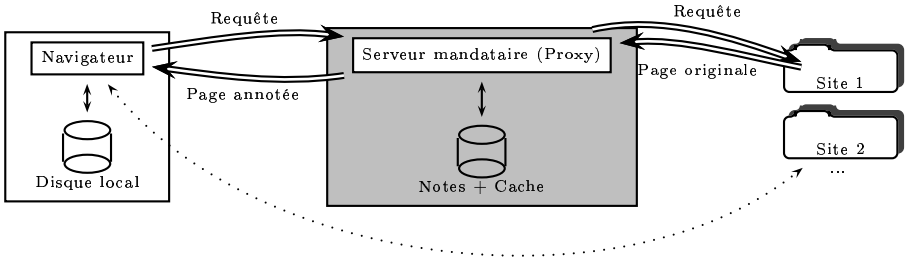


FIG. 2 – Architecture en proxy simple facultatif

Les systèmes à base d'intermédiaire client [Röscheisen et al., 1994, Koivunen et al., 2001, Imarkup Solutions, 2002, Denoue and Vignollet, 2000, ThirdVoice, 2000] sont des systèmes à installer sur le navigateur (figure 3). Par conséquent, ces systèmes sont totalement dépendants du système d'exploitation et du navigateur utilisés. Dans ces systèmes, le problème du partage des annotations n'est pas aussi simple à gérer. Certains systèmes proposent des protocoles pour les envoyer par courrier électronique ou d'utiliser un serveur d'annotation. Cette dernière solution réintroduit le problème de goulot d'étranglement normalement résolu du fait de la distribution du système d'annotation (en opposition à la centralisation des serveurs mandataires). Dans ces systèmes, la création et la visualisation des annotations sont beaucoup plus souples, quelque soit la structure du document et ce qu'il contient. Comme il est intégré au navigateur, il a accès à la structure DOM (Document Object Model) du document ainsi qu'aux différents évènements issus du navigateur.

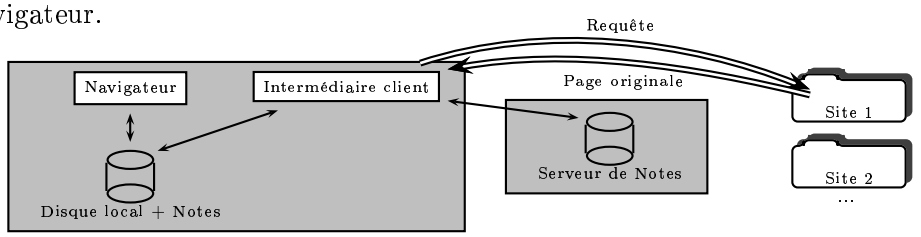


FIG. 3 – Architecture en intermédiaire client

[3] Pour accéder à la page "www.www.fr" d'origine, il suffit d'entrer sur son navigateur "http ://www.www.fr". Par contre, pour la même page avec les annotations, il suffit d'entrer "http ://crit.org/http ://www.www.fr".

Ces systèmes, et les architectures sous-jacentes, posent un certain nombre de problèmes en dehors des problèmes de confidentialité et de droit d'auteur déjà abordés. Le principal d'entre-eux est le passage à l'échelle. En effet, le problème est d'arriver à gérer des milliers d'utilisateurs et donc de manipuler des millions d'annotations. La solution la plus courante consiste soit à distribuer les serveurs soit à les dupliquer (serveurs miroirs) avec tous les problèmes habituels dans le domaine des bases de données, associés à ces solutions. Un autre problème est l'impossibilité de partager les annotations entre plusieurs systèmes, chacun utilisant son propre formalisme. Les normalisations avenir du W3C devraient résoudre en partie ces problèmes.

Finalement, dans tous les systèmes rencontrés, les annotations sont souvent simplistes et ne tiennent pas compte des objectifs visés, des types d'annotations manipulées (dimensions) et des utilisateurs concernés. Le grand nombre d'échec pour ces systèmes (très peu sont actuellement maintenus ou dépassent l'état de prototype de recherche) montre que ces problèmes sont loin d'être résolus et demandent à être travaillés[4].

4 Une nouvelle architecture : le système Dinosys

Les propriétés attendues pour un système d'annotation sur le Web sont : la légèreté, l'efficacité, la transparence, l'indépendance vis à vis de la plate-forme support, le passage à l'échelle [Vasudevan and Palmer, 1999]. Une autre propriété devient aussi de plus en plus primordiale (par exemple avec l'e-learning), c'est le support du travail collaboratif. En effet, dans le contexte de personnes travaillant ensemble et en même temps, le système permet la mise à jour en temps réel des annotations effectuées par chacune d'elles. [Vasudevan and Palmer, 1999] et l'analyse de la section précédente montrent que les divers systèmes existants n'ont pas tous les qualités énoncées précédemment. Notre objectif étant de définir une nouvelle architecture respectant les propriétés décrites, nous énonçons ici, nos motivations et nos choix :

1. *Une architecture distribuée* : l'intérêt porté par un utilisateur à une application est très souvent, sans parler de l'intérêt même du logiciel, lié à sa fiabilité et à la « fluidité » des fonctionnalités proposées. En ce qui concerne la fiabilité, elle représente une part importante dans la satisfaction et l'utilisation d'une application. Cette constatation nous a donc motivé dans la mise en place d'un système de proxy distribués géographiquement, spécialisés sur des domaines précis. Cette architecture distribuée apporte une plus grande disponibilité de services en cas d'interruption volontaire ou involontaire d'un proxy et assure une bonne répartition de la charge de travail en permettant la mise en place de nouveaux proxy au gré des sollicitations. Tout ceci garantit à l'utilisateur final une plus grande fluidité dans l'utilisation de l'application et, par conséquent, un plus grand intérêt dans son utilisation. La distribution contribue donc à ce que notre système ait les propriétés d'efficacité et de passage à l'échelle.
2. *Support du travail collaboratif* : plusieurs systèmes d'annotations actuels disposent de fonctionnalités de partage d'annotations sous la forme d'envois de mails aux participants. Ces fonctionnalités ne permettent pas des échanges «temps réel» des informations. Dans notre architecture, même si cette fonctionnalité reste présente (notamment pour prévenir le ou

[4]Pour des comparatifs techniques plus précis (techniques mises en œuvre, présentations des annotations choisies, types d'annotations...), il est possible de consulter [Garfunkel, 1999, Denoue, 2000, Bremer, 2002, Heck et al., 1999, Perry, 2001, Vasudevan and Palmer, 1999, Zohar, 1999].

les participants non connectés lors d'une session d'annotations collaborative), elle se voit couplée à un système de mise à jour automatique des annotations. Ce dernier sera utilisé lors de sessions d'annotations collaboratives en mode «forum» (i.e. un ensemble de personnes participe en même temps à un échange d'annotations sur un même document) ce qui rend ainsi les échanges plus directs.

3. *Indépendance vis à vis de la plate-forme, transparence et légèreté* : beaucoup de systèmes d'annotations existants sont basés sur des technologies propriétaires. Certains ne sont utilisables que sur une seule plateforme ou ne fonctionne qu'avec un seul type de navigateur. Pour dépasser ces limitations, la partie cliente de l'architecture se compose d'une applet Java. Ce choix technique permet à l'application de fonctionner sur toutes les plateformes dotées d'une machine virtuelle Java et avec tous les navigateurs prenant en charge le DOM et le JavaScript. L'utilisation d'une applet Java pour la partie cliente permet une mise à jour automatique de l'application sur le poste client sans aucune intervention particulière à l'ouverture d'une session d'annotation sur le portail d'accès (il est bien évident que si aucune mise à jour n'est nécessaire, c'est la version en cache sur le poste client qui est utilisée n'engendrant ainsi, aucun téléchargement superflu). Enfin, ce système de mise à jour automatique permet de garantir une utilisation de l'application dans sa dernière version de manière totalement transparente.

5 Description de l'architecture de Dinosys

Comme le constate [Vasudevan and Palmer, 1999], et comme nous l'avons vu à la section 3, les critères énoncés sont peu respectés dans les solutions existantes, et de toute manière jamais respectés tous ensembles pour un même système. Nous proposons donc une nouvelle architecture en exploitant les avantages des deux types d'architectures et en ajoutant une nouvelle dimension : la distribution.

Notre système, appelé *Dinosys* pour DIstributed NOtation SYStem, comprend trois composants principaux (figure 4) : un client (comme les systèmes à base d'intermédiaire client, figure 3), un portail et un ensemble de proxy (comme les systèmes à base d'intermédiaire par serveur mandataire, figure 2) tous identiques.

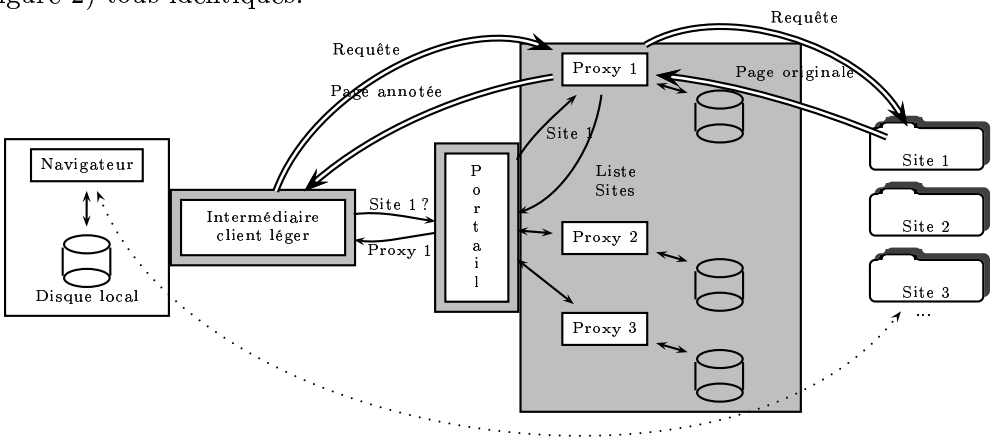


FIG. 4 – Architecture en proxy distribué facultatif

Le client[5] proposé se charge de la visualisation de la page annotée et de la gestion des annotations (création, modification, suppression, visualisation...). Il interagit avec les autres composants du système de manière transparente pour l'utilisateur. Ce client est téléchargé à partir du portail lorsque l'utilisateur commence à utiliser le système ou lorsqu'une nouvelle version est disponible. Dans ce dernier cas, la mise à jour est effectuée sans que l'utilisateur n'ait à intervenir (toujours dans un soucis de simplicité et de transparence).

Le portail joue plusieurs rôles dans notre système. Il est d'abord la vitrine du système sur le Web en présentant le projet. Il permet aussi aux nouveaux utilisateurs de s'inscrire et de charger le client[6]. Ensuite, lorsqu'un utilisateur désire manipuler une page, le client passe par l'intermédiaire du portail pour accéder au proxy chargé de la gestion de cette page. Il se charge aussi de fournir les éléments pour mettre en place de manière simple un proxy (quelque soit l'environnement du serveur). Pour finir, il permet aux différents intervenants (utilisateurs, administrateurs de proxy, administrateur principal) de gérer leurs profils et les informations les concernant.

Chaque proxy est chargé de gérer les annotations d'un ensemble de pages. Ces pages sont soit pré-sélectionnées par l'administrateur du proxy soit attribuées par le portail dans l'objectif de répartir les charges (processeur et disque) ou de regrouper des pages (pour éviter les accès au portail, un utilisateur travaillant sur un site pourra annoter ses pages sans changer de proxy). Un proxy a pour tâche, après contact par le client d'un utilisateur de récupérer la page demandée en y associant des annotations déjà présentes dans la base de données. Les nouvelles annotations saisies par l'utilisateur sont transmises au proxy par le client afin d'y être stockées. Lorsqu'un utilisateur désire changer de page, le client s'adresse d'abord au proxy puis, en cas d'échec, s'adresse au portail. Le proxy fournit dynamiquement au client les nouvelles annotations effectuées par les autres utilisateurs de la page. De plus, il informe le client des utilisateurs "connectés" sur la page courante.

Dans le cadre de Dinosys, deux grandes catégories d'intervenants se dégagent : les utilisateurs et les administrateurs. Les premiers consultent des pages annotées et, éventuellement, ajoutent et modifient des annotations. Les utilisateurs sont organisés en groupes pour prendre en compte les besoins de confidentialité. En effet, les annotations effectuées par un utilisateur d'un groupe peuvent être destinées à cet utilisateur uniquement, au groupe auquel il appartient ou au groupe le plus général (groupe publique[7]). L'autre catégorie d'intervenant est le groupe des administrateurs. Chaque proxy est géré par un administrateur qui décide des pages ou sites que son proxy doit (peut) gérer. Plus généralement, il joue le rôle de modérateur. Le système lui-même est géré par un administrateur qui gère utilisateurs (contrôles des utilisateurs et des utilisations, nomination des administrateurs de proxy...) et proxy (attributions des sites aux proxy, réaffectation de pages pour répartir les charges, lancement de nouveau proxy...). Du point de vue implémentation, le client est une applet Java (signée pour pallier aux problèmes de sécurité liés à la communication entre les différents cadres d'une page HTML). Les proxy sont des servlets associées à un moteur de servlet tel que Tomcat (donc portable dans n'importe quel environnement qui peut intégrer un moteur de servlets : IIS, Apache...). Les annotations, quant à elles, transitent dans le système sous forme de données XML, elles sont stockées sur les proxy dans des bases de données de type Mysql ou PostgresSql.

[5]Une version de démonstration du client est proposée à l'adresse suivante : http ://www.sciences.univ-nantes.fr/info/perso/permanents/desmontils/

[6]Une inscription allégée permet une évaluation du système.

[7]toutes les annotations sont visibles par tout le monde y compris les utilisateurs en mode évaluation

Pour utiliser Dinosys, il est donc nécessaire de se rendre à l'adresse du portail via un navigateur internet. Ce portail est chargé d'afficher dans le navigateur l'interface cliente du système (i.e. l'intermédiaire client) qui se présente sous la forme d'une barre d'outils. Notons que ce portail se charge également de vérifier la présence ou non de cet intermédiaire sur le poste client et le cas échéant, d'en effectuer un téléchargement automatique. Il est alors possible à l'utilisateur d'entrer dans la zone d'adresse de la barre d'outils l'URL du document à annoter et/ou consulter.

La requête est transmise au portail qui va établir la correspondance entre le site demandé et le proxy qui en a la charge. Une fois la correspondance établie, le portail communique en retour, l'adresse du proxy au client. A partir de cet instant, les échanges ne se feront plus qu'entre l'intermédiaire client et le proxy pour ce qui concerne le site consulté. La barre d'outils de l'application met à la disposition de l'utilisateur un certain nombre de fonctions pour gérer les annotations des documents consultés.

Pour un document, les nouvelles annotations sont transmises au proxy pour être enregistrées dans une base de données. Les clients connectés à ce document sont dynamiquement mis à jour (en respectant les permissions attribuées aux annotations : publiques, privées à un groupe...). Notons que l'utilisateur peut à tout moment repasser en mode "classique" de navigation en utilisant la barre d'adresse initiale de son navigateur.

6 Conclusion

Les spécificités du Web en font un environnement très difficile à exploiter. Les outils d'annotation sont une voie prometteuse pour l'échange et le partage d'informations. Ils permettent d'espérer atteindre un objectif primordial des concepteurs du Web à savoir un environnement collaboratif où chacun est aussi bien lecteur que rédacteur. Concrètement, les systèmes développés soit se sont soldés par des échecs commerciaux [ThirdVoice, 2000, Röscheisen et al., 1994] soit, projet de recherche, ont été abandonnés [Yee, 1997, Palme, 1998, Denoue and Vignollet, 2000]. Nous pensons que ces échecs sont particulièrement dus à deux causes : une mauvaise connaissance des annotations et des utilisations que l'on peut en faire ainsi que des architectures relativement inadaptées. De plus, les systèmes d'annotation sont des outils prometteurs pour améliorer la compréhension et la manipulation des documents dans le cadre du partage d'information et de l'interopérabilité sur le Web. Nous avons présenté une nouvelle architecture distribuée pour les systèmes d'annotation qui a comme propriété d'être légère, efficace, non-intrusive, indépendante de la plate-forme, supportant le passage à l'échelle et le travail collaboratif. Elle reprend à la fois la philosophie des systèmes basés sur des serveurs mandataires (mais les proxy sont distribués) et celle des systèmes à base d'intermédiaire client (l'intermédiaire client ici ayant pour rôle la communication avec le portail et avec le proxy qui référence les documents à annoter). Une première version du système a été développée. Nous l'utiliserons comme plate-forme expérimentale pour des applications de e-learning et surtout comme support à notre plate-forme d'indexation sémantique de document [Desmontils and Jacquin, 2002]. Les annotations apposées à l'aide de Dinosys seront exploitées lors de la phase de détermination des descripteurs des documents pour gérer le retour utilisateur. Dans ce cadre, nous nous intéressons à définir et exploiter la sémantique des annotations libres. Ceci nous permet aussi d'améliorer le système Dinosys en proposant une représentation visuelle des annotations en fonction de leur sémantique et du profil de l'utilisateur.

Références

[Bremer, 2002] Bremer (2002). Web annotations. http ://www.db.cs.ucdavis.edu/ bremer/annotations.html.

[Denoue, 2000] Denoue, L. (2000). *De la création à la capitalisation des annotations dans un espace personnel d'informations*. PhD thesis, Univ. De Savoie.

[Denoue and Vignollet, 2000] Denoue, L. and Vignollet, L. (2000). An annotation tool for web browsers and its applications to information retrieval. In *RIAO'00*, pages 180–195, Paris, France.

[Desmontils and Jacquin, 2002] Desmontils, E. and Jacquin, C. (2002). Indexing a web site with a terminology oriented ontology. In Cruz, I., Decker, S., Euzenat, J., and McGuinness, D. L., editors, *The Emerging Semantic Web*, pages 181–197. IOS Press.

[Garfunkel, 1999] Garfunkel, J. (1999). Web annotation technologies. http ://look.boston.ma.us/garf/webdev/annote/software.html.

[Heck et al., 1999] Heck, R. M., Luebke, S. M., and Obermark, C. H. (1999). A survey of web annotation systems. Technical report, Dep. Of Mathematics and Computer Science, Grinnell College, USA.

[Imarkup Solutions, 2002] Imarkup Solutions (2002). Imarkup. http ://www.imarkup.com/.

[Koivunen et al., 2001] Koivunen, M.-R., Kahan, J., Swick, R., and Prud'hommeaux, E. (2001). Annotea project. http ://www.w3.org/2001/Annotea/. W3C.

[Marshall, 1998] Marshall, C. C. (1998). Toward an ecology of hypertext annotation. In *ACM Hypertext*, pages 40–49. ACM Press.

[Ovsiannikov et al., 1999] Ovsiannikov, I., Arbib, M., and McNeill, T. (1999). Annotation technology. *Int. J. Human-Computer Studies*, 50 :329–362. http ://www-hbp.usc.edu/Projects/annotati.htm.

[Palme, 1998] Palme, J. (1998). Web4groups. http ://www.dsv.su.se/jpalme/w4g/web4groups-summary.html.

[Perry, 2001] Perry, P. (2001). Web annotations. http ://www.paulperry.net/notes/annotations.asp.

[Röscheisen et al., 1994] Röscheisen, M., Mogensen, C., and Wonograd, T. (1994). Commentor. http ://hci.stanford.edu/commentor/doc/.

[ThirdVoice, 2000] ThirdVoice, I. (2000). Thirdvoice 2000. http ://www.thirdvoice.com.

[Vasudevan and Palmer, 1999] Vasudevan, V. and Palmer, M. (1999). On web annotations : Promises and pitfalls of current web infrastructure. In *32nd Hawaii International Conference on System Sciences (HICSS-32)*. IEEE Computer Society Press.

[Yee, 1997] Yee, K.-P. (1997). Critlink. In *Fifth Foresight Conference on Molecular Nanotechnology*, http ://crit.org/critlink.html.

[Zohar, 1999] Zohar, R. (1999). Web annotation - an overview. Technical report, Dept. of Electrical Engineering, Israel Institute of Technology.

Propagation de métadonnées par l'analyse des liens

C. Prime-Claverie, M. Beigbeder

Laboratoire RIM - SIMMO
Ecole Nationale Supérieure des Mines de Saint-Etienne
158, cours Fauriel 42023 Saint-Etienne Cedex, FRANCE
Email : {prime,mbeig}@emse.fr
Tél : +33 4 77 42 66 12 Fax :

T. Lafouge

Laboratoire RECODOC
Université Claude Bernard Lyon 1
43, bd du 11 novembre 1918 69622 Villeurbanne Cedex, FRANCE
Email : lafouge@enssib.fr
Tél : +33 4 72 44 58 34 Fax :

Résumé

La Toile apparaît comme une véritable mine d'information et une des difficultés pour ses utilisateurs est de retrouver les documents répondant à leur besoins. Outre le problème de pertinence thématique, les documents rendus par les moteurs ne sont pas toujours en adéquation avec les attentes de l'utilisateur : document trop généraliste, ou contraire d'un niveau élevé, d'un genre différent de celui attendu par l'utilisateur, etc. Nous pensons que l'ajout de métadonnées aux pages pourrait considérablement améliorer la recherche d'information sur la Toile. Dans cet article nous proposons une méthode permettant d'ajouter ces métadonnées de manière semi-automatique. Elle se base sur la propagation des métadonnées dans le graphe de co-citation formé à partir du graphe web.

Abstract

The World Wide Web currently has a huge amount of pages and it is extremely difficult to retrieve documents corresponding to one's informational needs. In addition to the problem of thematic relevance, documents returned by the search engines do not correspond to the user expectations, documents are either too difficult or on the contrary too easy. We realize that the way to clearly improve information retrieval on the Web is to add metadata to web pages (thematic or non-thematic). In this paper, we present a method able to add metadata in a semi-automatic way. It is based on the propagation in the co-citation graph coming from the Web graph.

1 Introduction

Le Web apparaît comme une véritable mine d'information regroupant des ressources très différentes les unes des autres, aussi bien au niveau de leur contenu thématique, que de leur genre, leur langue, leur niveau, etc. Une des difficultés pour ses utilisateurs est de retrouver les ressources pertinentes à leurs besoins. Contrairement aux bases de données documentaires traditionnelles qui sont gérées et organisées par une même autorité, le Web est un espace d'expression libre qui ne connaît aucune organisation. Une des manières d'améliorer l'accès à son contenu serait d'ajouter de manière systématique des méta-informations aux pages web. Bien que prévu par les langages HTML et maintenant XML et malgré tous les efforts de normalisation (Dublin Core [Dublin, 2003]) l'utilisation de métadonnées est encore peu répandue. Ces métadonnées d'auteurs sont d'ailleurs assez mal utilisées, soit par un manque de pratique ou d'objectivité de la part des auteurs honnêtes, soit détournées de leur objectif initial pour permettre une meilleure visibilité par ceux qui les maîtrisent.

Pour que les pages web soient décrites de manière uniforme et systématique, nous pensons que ce sont les systèmes de recherche d'information eux-mêmes qui doivent affecter les méta-informations, de la même manière que ce sont les professionnels de la documentation qui effectuent le catalogage et l'indexation. Précisons que ces opérations documentaires sont effectuées manuellement et sont donc très coûteuses, et étant donné le nombre de pages disponibles sur le Web, leur volatilité, il n'est pas envisageable que les métadonnées soient affectées manuellement. Il faut donc s'orienter vers des méthodes automatiques ou semi-automatiques.

Cet article présente nos recherches en cours concernant la possibilité de propager des métadonnées aux documents web par l'analyse du graphe formé par les liens hypertextuels.

2 La représentation des documents Web

Les principaux outils de recherche d'information (moteurs) disponibles sur la Toile s'appuient sur les techniques des SRI (Système de Recherche d'Information) traditionnels notamment pour la représentation des documents et des requêtes, et pour le calcul des fonctions de correspondance. Rappelons toutefois que les SRI traditionnels travaillent sur des corpus de documents, que l'on appelle aussi collections. Une collection est un ensemble de documents sélectionnés, rassemblés par une même autorité et parfois classés. Les collections constituent donc des ensembles cohérents et homogènes où les documents partagent des propriétés communes (collections d'articles scientifiques, de brevets, etc.). Dans de telles collections la priorité est donc d'appréhender l'apport informationnel de chaque document et c'est pourquoi ceux-ci sont représentés sémantiquement au cœur du SRI par des mots-clés. Or sur le Web, espace hétérogène, il serait dommage de se limiter à une simple représentation thématique des documents comme le font les moteurs et les annuaires. C'est pour cela que nous envisageons d'ajouter aux pages en plus de leur représentation sémantique des métadonnées non thématiques.

3 L'analyse des liens

Actuellement, deux communautés scientifiques s'intéressent de près à l'analyse des liens du Web : les bibliométres et les informaticiens. Les premiers, dont l'un des objectifs est de structurer l'univers du savoir à partir de grands volumes d'information, étudient les équivalences entre les concepts établis en bibliométrie et le graphe du Web [Ingwersen, 1998], [Björneborn and Ingwersen, 2001], [Aguillo, 1999], [Egghe, 2000]. En effet, comme dans le réseau des publications scientifiques [Garfield, 1972] un lien hypertexte peut matérialiser une citation et indiquer une relation intéressante entre la page d'origine et la page vers laquelle il pointe. Les seconds utilisent les méthodes mathématiques de la théorie des graphes dans l'objectif d'améliorer la recherche d'information sur le Web. Parmi les applications les plus connues nous pouvons citer les algorithmes de classement de Google [Brin and Page, 1998], la découverte de communautés d'intérêts [Kumar et al., 1999].

Marchiori [Marchiori, 1998] propose dès 1998, une méthode permettant de propager des métadonnées de classification (thématique) le long des liens. Dans cette méthode, chaque page est décrite par des métadonnées thématiques (mots-clés) pondérées par un coefficient variant entre 0 et 1 (1 lorsque la métadonnée décrit parfaitement la page, 0 lorsqu'elle est inappropriée). Son hypothèse est la suivante : si une page P (décrite par une métadonnée A pondérée par le coefficient ν) est citée par une page P', alors on peut supposer que P sert à expliciter (à appuyer) des idées évoquées dans la page P'. Les métadonnées de P peuvent donc être propagées à P' avec un facteur d'affaiblissement f ($0 < f < 1$). La métadonnée A décrit à présent le document P' avec le coefficient $\nu \times f$.

Nous pensons comme lui que si un document P contient un lien hypertexte vers un document P', il existe (au moins pour le créateur de la page P) une association entre ces deux documents. Celle-ci se traduit par des valeurs identiques pour une ou plusieurs métadonnées (c'est-à-dire que deux pages reliées dans le Web partagent au moins un point commun, même origine géographique, même thème, même niveau...). Cependant nous pensons qu'une analyse plus poussée du graphe, utilisant des relations plus complexes que la simple relation "citant-cité" peut permettre d'extraire des ensembles de pages très homogènes partageant une majorité de métadonnées identiques.

4 Notre méthode

Nous envisageons comme Marchiori de propager des métadonnées en utilisant les relations entre les pages du Web. Notre méthode ne s'appuie pas directement sur le graphe Web, mais sur un graphe obtenu de manière indirecte, le graphe des co-citations (fig. 1). La méthode proposée comporte deux étapes :

- la structuration du corpus par la méthode des co-citations en vue d'obtenir une hiérarchie de sous-corpus que nous supposons homogènes,
- la propagation de métadonnées dans les sous-corpus.

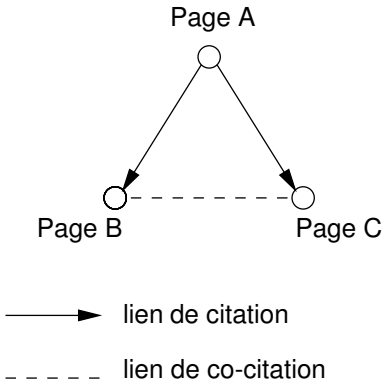


FIG. 1 – Liens de citation et de co-citation sur le Web

4.1 Structuration du corpus par la méthode des co-citations

4.1.1 La méthode des co-citations

La méthode des co-citations, utilisée en bibliométrie depuis 1973 [Marshakova, 1973] [Small, 1973], a pour objectif de créer à partir d'articles scientifiques d'un même domaine de recherche, et plus précisément de leurs références bibliographiques, des cartes relationnelles de documents ou d'auteurs qui reflètent à la fois les liens sociologiques et thématiques de ce domaine. Cette méthode repose sur l'hypothèse que deux références bibliographiques de date quelconque, fréquemment citées ensemble ont une parité thématique. Le lien hypertexte lui aussi peut matérialiser une citation, et plusieurs auteurs [Larson, 1996] [Pitkow and Pirolli, 1997] [Prime et al., 2002a] se sont intéressés à la transposition de la méthode des co-citations de documents pour caractériser les univers du Web. Ils mettent en évidence les limites théoriques et techniques de l'analogie, mais ont montré l'intérêt de la structuration pour rapprocher thématiquement les pages. Une des limites de cette analogie est de considérer tous les liens hypertextes comme des liens de citation ou de référence. En effet, il faut aussi prendre en compte les liens de publicité, mais surtout ceux qui servent à se déplacer dans un même site web : les liens de navigation interne. C'est pourquoi notre méthode ne tient compte que des liens inter-serveurs entre les pages citantes et citées espérant ainsi supprimer la majorité des liens de navigation.

La première phase de la méthode consiste à déterminer la proximité des pages entre elles. Pour cela on définit un indice de similarité qui doit traduire mathématiquement l'idée suivante : deux pages P_i et P_j sont proches, si par rapport à leurs fréquences de citation respectives (C_i et C_j), leur fréquence de co-citation (C_{ij}) est importante. Il existe plusieurs indices possibles qui par convention varient de 0 à 1 : 1 lorsque les pages sont toujours citées ensemble, et 0 lorsque celles-ci ne le sont jamais. L'indice que nous utilisons est l'équivalence :

$$E_{ij} = \frac{{C_{ij}}^2}{C_i \times C_j}. \quad (1)$$

Dans la suite de l'article nous utiliserons la distance d_{ij} entre les pages, plutôt que leur proximité avec $d_{ij} = 1 - E_{ij}$. Toutes les valeurs d_{ij} sont inscrites dans une matrice de co-citations à partir de laquelle on peut construire le graphe de co-citations, graphe valué où les nœuds sont les pages et les arcs les liens de co-citations entre les pages valués.

La seconde phase, le regroupement des pages les plus proches, utilise des méthodes de classification automatique issues de l'analyse de données. Nous utilisons une classification hiérarchique ascendante. Plusieurs choix sont possibles : le simple lien (voisin le plus proche), le lien complet (voisin le plus éloigné), le chaînage moyen. Cette classification permet de créer une hiérarchie de classes (agrégats de pages). Les documents les plus similaires sont regroupés dans des classes au plus bas niveau, tandis qu'au plus haut niveau les documents sont tous regroupés ensemble. La hiérarchie obtenue peut être visualisé graphiquement par un dendrogramme (fig. 2).

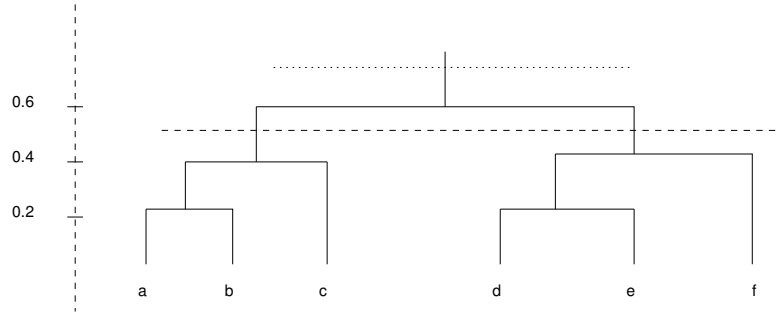


FIG. 2 – Exemple d'un dendrogramme

Une des difficultés de la méthode consiste à déterminer le niveau de coupure du dendrogramme qui donnera à la fois des classes de taille importante et les plus homogènes possible.

4.1.2 Expérience et résultats

Nous avons mené en 2001 une expérience sur un corpus contenant des pages relatives au thème de l'astronomie [Prime et al., 2002b]. Nous avons classé 198 pages par la méthode des co-citations que nous avons indexées à la main pour des métadonnées liées au genre (type) de document. Nous nous sommes intéressés à l'homogénéité des classes obtenues par la méthode du lien complet. Les résultats observés ont été très encourageants.

4.2 Propagation par l'analyse des liens

La méthode de propagation que nous proposons repose sur l'hypothèse que deux pages proches par l'indice de co-citation partagent des métadonnées communes. Elle permet de propager la (ou les) valeur(s) d'une (ou de plusieurs) métadonnée(s). Contrairement à Marchiori, il n'est pas nécessaire d'utiliser des métadonnées pondérées car notre méthode n'influe pas sur les pondérations. Elle s'appuie à la fois sur le dendrogramme obtenu par la classification et le graphe des co-citations. En effet, l'expérience menée montre que les classes ne se forment pas toutes "à la même vitesse". Certaines sont déjà de taille importante et bien homogènes à un seuil relativement bas dans le dendrogramme, alors qu'à ce même seuil persistent encore beaucoup de singletons. A un seuil plus élevé, certains singletons ont pu se regrouper ou rejoindre d'autres classes pour former des ensembles homogènes, tandis que d'autres classes qui étaient homogènes se sont "bruitées". C'est pourquoi nous trouvons dommage de ne travailler qu'à un seul niveau de coupure du dendrogramme et de ne pas utiliser toute la richesse de la hiérarchie.

Nous définissons un seuil S à partir duquel nous supposons que les classes sont déjà de taille importante et que celles-ci ne sont pas encore trop bruitées. Ce seuil dépend de la méthode d'agrégation choisie, plus la distance inter-classe est exigeante, plus le seuil S pourra être élevé. Au niveau de coupure S, nous obtenons une partition du corpus de départ. Chaque classe induit un sous-graphe du graphe de co-citations (figure 3). Pour chacune d'elles nous identifions le couple d'éléments les plus éloignés par la distance dans un graphe valué[1].

La figure 3 montre le graphe qui a permis de générer le dendrogramme de la figure 2 avec la méthode du simple lien. Pour un seuil supérieur à 0,6, nous obtenons une classe à 6 éléments. Les deux éléments les plus éloignés sont b et f (distant de $1,6$). Au seuil 0,6 cette classe sera divisée en deux créant ainsi deux sous-graphes à trois éléments. Les éléments b et f sont indexés pour les trois métadonnées "type de site", "type d'autorité" et "thème" avec les valeurs respectives *site de ressources*, *entreprise*, *astronomie* et *site de ressources*, *association* et *astronomie*.

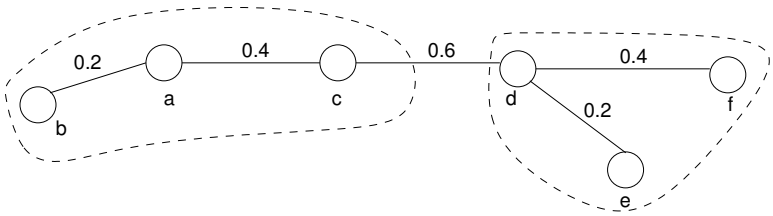


FIG. 3 – Visualisation du graphe de la classe de la figure 2 pour un seuil supérieur à 0,6

Pour chaque classe, nous examinons les valeurs de métadonnées du couple d'éléments les plus distants. Lorsqu'un de ces éléments n'est pas indexé, ce qui est le cas au départ, nous le faisons manuellement. Ces valeurs sont comparées. Le principe de notre méthode est de propager les valeurs de métadonnées du couple lorsqu'elles sont identiques, aux autres éléments de la classe. En effet ceux-ci ont une forte probabilité de partager les mêmes valeurs, puisqu'ils sont plus proches les uns des autres. Lorsque l'on travaille avec plusieurs métadonnées le couple peut partager les mêmes valeurs pour certaines métadonnées et avoir des valeurs différentes pour les autres. Dans l'exemple ci-dessus, les éléments b et f partagent les deux valeurs *site de ressources* et *astronomie*, alors que les valeurs de la métadonnée "type d'autorité" diffèrent. Deux choix sont possibles :

- propager les valeurs de métadonnée de manière individuelle. Dès que le couple partage une valeur commune de métadonnée, celle-ci est propagée aux autres éléments. Pour ce choix et dans l'exemple ci-dessus, les valeurs *astronomie* et *site de ressources* sont propagées aux éléments a, c, d, et e.
- propager les valeurs de métadonnées en groupe, c'est-à-dire, exiger qu'une partie ou la totalité des valeurs de métadonnées du couple soient identiques pour les propager aux autres éléments de la classe. Dans l'exemple ci-dessus, les valeurs de la métadonnée "type d'autorité" sont différentes. Si l'on exige que toutes les valeurs de métadonnées de couple soient identiques 2 à 2, alors aucune valeur n'est propagée, et la classe est divisée en deux au seuil 0,6.

[1]La distance $d(x, y)$ entre 2 sommets x et y est la longueur du plus court chemin entre x et y. Dans un graphe valué, c'est la somme des valuations des arêtes de ce chemin.

Tant que les valeurs de métadonnées sont différentes, nous "descendons" dans le dendrogramme au seuil qui partionnera cette classe et recommençons l'opération espérant ainsi intervenir le moins possible manuellement et propager le plus de valeur de métadonnées automatiquement.

5 Limites et perspectives

La méthode présentée ci-dessus est en cours de test sur le corpus utilisé dans l'expérience antérieure pour les 3 métadonnées "type de site", "type d'autorité", "type d'information". Nos premiers résultats avec une méthode qui propage les valeurs de métadonnées en groupe, nous donnent une bonne qualité de propagation (peu d'erreurs) mais une très faible rentabilité : l'indexation manuelle est trop importante par rapport à l'indexation automatique par propagation.
D'autre part, nous savons qu'une des limites de cette méthode est le faible taux de pages indexées. En effet, sur le Web de nombreuses pages ne sont pas citées par des pages hébergées sur d'autres sites, si bien qu'elle ne peuvent être *a fortiori* co-citées et classées. Il faudra donc envisager une méthode d'indexation et de propagation de métadonnées au sein des sites Web. Notons aussi que notre indice de similarité (l'équivalence) ne dépend pas du nombre de liens émis par les pages citantes. Or sur le Web le nombre de liens émis par chaque page est extrêmement variable, et il serait judicieux d'en tenir compte pour calculer la proximité entre les pages. Actuellement, nous commençons une expérience de plus grande envergure sur un corpus contenant 5 millions de pages qui correspond au Web francophone de décembre 2000 (collecté par M. Géry et D. Vaufreydaz du laboratoire CLIPS `http://www-clips.imag.fr`), pour identifier clairement le pourcentage de pages co-citées et le nombre de pages pouvant être ainsi classées.

Références

[Aguillo, 1999] Aguillo, I. (1999). Statistical indicators on the internet : The european science technology industry system in the world-wide web. *at http://diotima.math.upatras.gr/weborg/aguillo2.*

[Björneborn and Ingwersen, 2001] Björneborn, L. and Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1) :65–82.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International WWW Conference*. IW3C2.

[Dublin, 2003] Dublin (2003). Dublin core metadata initiative (dcmi). *at http://dublincore.org consulté en février 2003.*

[Egghe, 2000] Egghe, L. (2000). New informetric aspects of the internet : some reflections, many problems. *Journal of information science*, 26(5) :329–335.

[Garfield, 1972] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, (178) :471–479.

[Ingwersen, 1998] Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation*, 54(2) :236–243.

[Kumar et al., 1999] Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. In *Proceedings of the Eighth World Wide Web Conference.*

[Larson, 1996] Larson, R. (1996). Bibliometrics of the world wide web : An exploratory analysis of the intellectual structure of the cyberspace. In *Proceedings of the Annual Meeting of the American Society of Information Science*, Baltimore.

[Marchiori, 1998] Marchiori, M. (1998). The limits of web metadata and beyond. In *Proceedings of the Seventh International WWW Conference*. IW3C2.

[Marshakova, 1973] Marshakova, I. V. (1973). Document coupling system based on references taken from science citation index. *Russia, Nauchno - Teknicheskaya Informatsiya*, 2(6,3).

[Pitkow and Pirolli, 1997] Pitkow, J. and Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97*, pages 118–125.

[Prime et al., 2002a] Prime, C., Bassecoulard, E., and Zitt, M. (2002a). Co-citations and co-sitations : a cautionary view on an analogy. *Scientometrics*, 54(2) :291–308.

[Prime et al., 2002b] Prime, C., Beigbeder, M., and Lafouge, T. (2002b). Clusterisation du web en vue d'extraction de corpus homogènes. In *Actes du 20ème congrès INFORSID*, pages 229–242, Nantes.

[Small, 1973] Small, H. (1973). Co-citation in the scientific literature. *Journal of the American Society for Information Science*, 24 :265–269.

Deuxième partie

Communications Affichées

Art sur Internet
Analyse d'usages de sites web d'artistes

G. VIDAL
LabSIC - Université Paris 13,
99, avenue J.B. Clément,
93430 Villetaneuse, FRANCE
Mail : gvidal@sic.univ-paris13.fr
Tél. : +33 1 49 40 32 76 Fax : +33 1 49 40 44 80

Résumé

L'analyse qualitative d'usages de sites Web d'artistes confirme le fait que les internautes se réfèrent à un cadre d'usages Internet et culturels préalables. Ils cherchent à maîtriser la situation de communication médiatisée par ordinateur. De fait, ils établissent des relations avec les œuvres en ligne et vivent des expériences inédites. Ainsi, les usagers de sites d'artistes sur Internet adoptent une posture interactive les conduisant à éprouver des émotions.

Abstract

How do the Internet users use artists' Websites ? For our inquiry on the uses of a corpus of twelve artists' Websites, we adopted a qualitative method. In order to use these Websites, the users refer to a framework of previous uses. They need to control the computer-mediated communication. In fact, few users live new experiences during the relations with pieces and artists on line. Finally, we observe new emotions towards on-line works and interactive posture.

A partir des observations et des entretiens auprès de trente-deux internautes[1], qui consultaient trois sites d'artistes[2], durant au moins une demi-heure, à deux reprises[3], nous avons mené une analyse qualitative des usages [Jouët, 2000 ; Chambat, 1994] d'œuvres interactives sur Internet.

Internautes rencontrés	Sites d'artistes
4 femmes 4 hommes – 5 jeunes 3 adultes temps moyen de consultation/site : 9, 13, 13 minutes	**Jodi** http://www.jodi.org œuvre **Lecielestbleu** http://www.lecielestbleu.com œuvres de plusieurs artistes **Turux** http://www.turux.org œuvres d'un artiste
4 femmes 4 hommes - 4 jeunes 4 adultes temps moyen de consultation/site : 13, 12, 15 minutes	**Mongrel** http://www.tate.org.uk/netart/mongrel/home/default.htm Oeuvre dans le site d'un musée **Adaweb** http://adaweb.walkerart.org oeuvres de plusieurs artistes **Metaorigine** http://metaorigine.free.fr/ œuvres d'un artiste
4 femmes 4 hommes - 4 jeunes 4 adultes temps moyen de consultation/site : 13, 12, 14 minutes	**Cityparadigms** http://www.cityparadigms.timsoft.com œuvre **Panoplie** http://www.panoplie.org œuvres de plusieurs artistes **Présentation** http://PRESENTAT.IO-N.NET/ oeuvres d'un artiste
3 femmes 5 hommes - 4 jeunes 4 adultes temps moyen de consultation/site : 12, 12.5, 14 minutes	**Postal** http://postal.free.fr/ œuvre **Rhizome** http://www.rhizome.org site d'informations artistiques et accès aux oeuvres **Mouchette** http://www.mouchette.org oeuvres d'un artiste

Nous avons vérifié l'hypothèse considérant le fait que pour utiliser des sites d'artistes, les internautes se réfèrent à un cadre d'usages Internet et culturels préexistants. Mais cette étude exploratoire, étant donnée la faible connaissance des usages réels de cette production artistique en ligne, offre en outre une visibilité sur les relations médiatisées par ordinateur entre les internautes et les œuvres réalisées et mises en ligne par les artistes.

[1] Nous avons rencontré ces 32 internautes dans différents lieux : bibliothèque, cybercafé, université, lieu de travail de l'observé, domicile de l'observé. Dès lors, les usages réels ont été étudiés dans le contexte des usagers [Proulx 1994].

[2] Douze sites d'artistes, classés en quatre groupes de trois, ont été retenus selon une typologie de sites Web d'artistes : Fermés, les sites présentent une œuvre : Jodi, Postal, Cityparadigms, Tate-Mongrel, Ouverts, les sites présentent plusieurs oeuvres de plusieurs artistes : LeCielestBleu, Panoplie, Adaweb, Rhizome, Variables, les sites présentent plusieurs oeuvres d'un même artiste : Turux, Presentation, Mouchette, Metaorigin.

[3] 32 personnes x 2 consultations = 64 consultations x 3 sites = 192 observations. 15 femmes et 17 hommes - 17 jeunes et 15 adultes.

Pratiques informatiques et culturelles préalables

L'ensemble des usagers s'appuient sur leurs culture technique et pratiques culturelles, pour utiliser les interfaces qui leur sont ou non familières[4]. Lorsque ces dernières offrent des accès inédits aux contenus, les internautes éprouvent le besoin de se référer à des habitudes pour se repérer. Des tâtonnements et des tactiques [de Certeau, 1990] sont alors mis en oeuvre pour connaître petit à petit le site, découvrir les œuvres en ligne et donner un sens à leur exploration.

Avec les sites proposant des listes/menus notamment, plusieurs usagers choisissent une démarche méthodique, fondée sur une lecture linéaire, en commençant de gauche à droite, de haut en bas, pour ouvrir les menus, les liens, les images. Les sites sont dès lors pensés comme des « espaces » où sont entreposés des contenus à ouvrir et parcourir. Nous pouvons faire l'hypothèse que cette démarche systématique relève d'une volonté de se rassurer, en se référant à des usages préalables, au cours de la découverte d'un site sur Internet pensé potentiellement inépuisable. Ces usagers consultent les sites de façon superficielle, sans volonté d'approfondir leur consultation, mais de « faire le tour » des sites [5].

D'autres usagers, employant souvent le terme « chemin » -emblématique de la posture de la consultation qui consiste à partir d'une page pour aller vers une autre page-, définissent des objectifs, liés à un thème évocateur ou à un titre qui les attire. Ces usagers approfondissent davantage leur consultation par rapport à ceux visant un tour complet des sites.

Les internautes désorientés, ne parvenant pas à donner un sens à leur navigation, éprouvent un sentiment d'échec[6]. Celui-ci peut provoquer un rejet ; certains souhaitent en effet ne pas revenir sur les sites lors de la seconde rencontre ou abandonnent leur consultation en cours. Dans le cas du site Jodi, le sentiment d'échec est exacerbé par l'affichage non conventionnel des données :

> « Je ne trouve pas ça ludique. Ça me fait penser à un programme informatique » … « j'en ai assez vu », (une enseignante 1). « Il n'y a rien de plus illisible, on ne peut même pas reconnaître la page de départ » (un retraité).

Ce sentiment d'échec donne lieu à deux interprétations : certains usagers estiment être responsables de la situation d'échec : « Je n'ai toujours pas compris » (le retraité) et d'autres n'estiment pas en être responsables : « Il y a une faute, une erreur (…) il a un problème le site » (une hôtesse d'accueil).

[4] Sommaire, lien mentionné par un souligné, la couleur bleue ou un gras, curseur se transformant en main, retour dans la barre de navigation, ouverture de liens dans une nouvelle fenêtre, ascenseur pour descendre dans les pages téléchargées, page d'accueil considérée comme point de repère pour circuler dans les sites.

[5] Nous faisons l'hypothèse que le sentiment de pouvoir visiter un site dans sa globalité, « d'en faire le tour », sentiment pouvant provenir d'une appréhension de l'hypertexte qui fournit la possibilité de tout voir, a pour conséquence un temps de consultation plus long que les sites dont l'hypertextualisation des données est peu conventionnelle, provoquant un manque de repères. Toutefois, cette hypothèse reste à vérifier statistiquement, compte tenu de la tentative de faire le tour du site Postal, à l'hypertexte peu conventionnel, de la part d'une étudiante/analyste.

[6] Nous définissons l'échec comme une situation au cours de laquelle l'usager cesse de naviguer sur le site parce qu'il dit ne pas comprendre de quoi il s'agit, parce qu'il se dit perdu. N'est pas considéré comme un échec le manque d'intérêt pour le site.

Nouveaux repères et posture interactive

Mais, le manque de repères dans de nouveaux sites peut également être à l'origine de motivations et d'inventions d'usages au fil de la navigation des internautes. Etant donné que les internautes veulent savoir où ils se trouvent dans le site, ils cherchent alors à maîtriser la situation de communication médiatisée par ordinateur. Pour ce faire, ils se réfèrent aux normes d'usages du Web récentes certes, mais déjà prégnantes et associées aux critères d'efficacité et de performance[7].

Concrètement, les internautes, acceptant d'être surpris, tentent de se créer de nouveaux repères et sont fiers d'adopter une posture interactive. Ils deviennent dès lors les acteurs, voire se revendiquent co-auteurs de l'œuvre en ligne, en dépassant l'usage de l'interactivité comme simple possibilité de sélection. Pour plusieurs usagers, il s'agit de nouvelles expériences d'interactivité :

> « C'est moi qui fais la page … c'est moi qui fais bouger » (une assistante d'édition dans le site Panoplie). « C'est moi qui fais le tableau» (une enseignante 1 dans les sites Turux et Lecielestbleu). « Mes mouvements font bouger les croix » (un adolescent dans le site Turux).

Nous avons par ailleurs constaté une forte et fréquente volonté de comprendre le sens de l'œuvre, de l'interpréter :

- soit en observant : cette posture passe par une appréhension visuelle indépendamment de l'interface interactive ou par un mode de consultation descriptif,

-soit en interagissant avec les œuvres : cette seconde démarche tournée vers l'action conduit les usagers, main toujours sur la souris, à manipuler et faire évoluer les œuvres. Ils souhaitent comprendre ce qu'ils modifient et comment cela se produit.

Il semble que les internautes revendiquent une responsabilité dans l'acte de création, dans la mesure où il y a une volonté auctoriale de manipuler les œuvres à leurs convenances, de contrôler la situation en ligne de façon autonome et de partager cette responsabilité avec l'auteur du site, et ce, même si le site ne permet pas, de façon fonctionnelle, à l'usager de participer. Deux exemples sont intéressants à relever :

- un étudiant 3, dans le site Panoplie, souhaite revenir sur une œuvre vidéo en ligne pour vérifier son contenu : « je veux voir si on peut aller saisir l'image de cette fille ». Pour cela, il utilise la fonction zoom dans le navigateur, rallongeant ainsi le temps initialement prévu de visionnage, et modifiant l'image, grossie.

- Une étudiante 4 intervient dans l'ordre de défilement des pages dans le site Postale, en changeant les numéros de pages dans l'adresse du site (url). Ainsi, elle change à sa guise les étapes, initialement prévues par l'auteur.

Les représentations d'Internet (notamment le pouvoir par l'interactivité et la transparence pour tout voir) les conduisent à adopter une attitude de prise de pouvoir. Et si ce n'est pas le cas, ils n'apprécient pas d'être dans l'impossibilité d'agir, sont mal à l'aise, agacés et vexés.
En s'appuyant sur cette volonté d'autonomie, nous pouvons faire l'hypothèse que l'usager s'émancipe, en s'appropriant l'œuvre, l'interactivité programmée et en tentant de contourner les possibilités prescrites, les contraintes imposées par l'auteur, pour créer son propre parcours.

[7] Aller vite, atteindre une cible le plus rapidement possible, bénéficier de services personnalisés, notamment.

L'envie de s'engager dans une relation avec les œuvres se manifeste également par le fait de relier des repères du monde avec de nouveaux repères du « monde virtuel » :

> « Tu vas te noyer, (…) on ne va pas le laisser tenir trop longtemps sous l'eau sinon il n'aura plus de souffle » (une hôtesse d'accueil dans le site Lecielestbleu, s'adressant à un personnage en ligne).

En opérant des interprétations avec des éléments issus du réel, les internautes abordent les œuvres en accordant une place aussi importante à cette démarche qu'à la manipulation de l'œuvre.

Au delà de la relation qui implique l'usager dans l'œuvre, se manifeste la figure de l'auteur. Les internautes font référence à l'artiste, au concepteur sans forcément savoir à qui ils s'adressent. Ils sont conscients de son contrôle sur l'œuvre, de sa présence. L'interactivité fournit le temps et le lieu d'une action qui permet d'instaurer un contact entre l'œuvre, l'internaute et l'artiste. Ce contact peut prendre la forme d'un dialogue certes via l'interactivité, mais aussi d'échanges indirects (remarques, questions sur les intentions, critiques) de l'usager qui s'adresse, à voix haute, à un auteur absent.

Le recueil de ces propos nous conduit à deux remarques. D'une part, l'auteur ne revêt pas la même identité, il peut être « la personne, un artiste, ils, il ». D'autre part, une disparité apparaît dans la représentation que les usagers ont de l'auteur : tantôt il est créateur d'un projet artistique, tantôt concepteur d'un dispositif technique. Certains usagers s'adressent en priorité à l'artiste, tandis que d'autres au concepteur.

Le dialogue avec l'artiste et le rapport concret [Couchot, 1998] qui s'établit avec les œuvres constituent une scène des émotions. Le plaisir est présent tout au long de la navigation et ne relève pas seulement d'émotions liées aux qualités esthétiques des oeuvres, mais aussi de la qualité humoristique de l'œuvre et encore peut se déclarer par des appréciations négatives : « un peu d'air ! » (un sculpteur dans le site Mongrel. L'expression semble indiquer un sentiment d'étouffement éprouvé lors de la consultation).

Les usagers éprouvent du plaisir quand ils se sentent acteurs des multimédias en ligne, quand ils commandent du bout des doigts : « Je veux voir là » (une collégienne dans le site Cityparadigms. Le " là " est tout à fait évocateur de ce rapport physique à un univers immatériel et à une posture de commande).

Conclusion

Cette étude exploratoire permet de saisir la façon dont les usagers s'approprient des sites d'artistes sur Internet, selon leurs identités sociales (compétences, centres d'intérêt, habitudes, imaginaires). Ces appropriations contribuent à l'invention d'usages des œuvres en ligne et de postures interactives dans les sites. L'analyse qualitative des usages réels a mis en lumière des besoins de repères ancrés dans des pratiques informatiques et culturelles préexistantes. Celle-ci a également permis d'explorer des expériences interactives et des émotions en ligne stimulant les navigations-consultations. Nous faisons l'hypothèse que ces relations avec les œuvres et les artistes participent d'une esthétique de l'interactivité.

Enfin, nous relevons une quasi-absence de solennité[8], face aux œuvres interactives sur Internet. Et lorsque nous comparons l'analyse d'usages de ces dernières avec l'analyse de

[8] Nous avons observé un regard « solennel » sur des œuvres (Tate, Adaweb, Metaorigin) de la part d'internautes se référant à leur cadre d'usages des musées (étudiante 3), ou à des cours d'histoire de l'art (femme au foyer). En effet, sans intervenir dans l'œuvre interactive, ils les décrivent, interprètent, analysent et expriment ce qu'ils ressentent. Sans se borner à la

visites de musées d'art [9], nous nous demandons si cette quasi-absence provient de ce rapport ludique à Internet, aux hypermédias et aux écrans ou de la technicisation du procès de réception de l'art ?

Références

[Chambat, 1994] Chambat, P. (1994). « Usages des technologies de l'information et de la communication (TIC) : évolution des problématiques », *Technologies de l'Information et Société,* vol. 6, n°3 : 249-270.

[Couchot, 1998] Couchot E. (1998). « Autre corps, autre image. Autre image, autre corps », dans « Ce corps incertain de l'image », Art/Technologies, *Champs Visuels* n° 10, juin 1998, Paris, L'Harmattan : 12-16.

[de Certeau, 1990] De Certeau M. (1990). *L'invention du quotidien. Tome 1 : arts de faire,* Paris**,** éditions Gallimard, collection Folio/Essais.

[Jouët, 2000] JOUËT Josiane (2000) « Retour critique sur la sociologie des usages », *Réseaux,* n°100 : 487-521

[Proulx, 1994] Proulx, S., (1994). « Les différentes problématiques de l'usage et de l'usager », *Médias et nouvelles technologies. Pour une socio-politique des usages*, Vitalis A. (sous la direction de), Rennes : Apogée, coll. Médias et nouvelles technologies : 149-159

description (ils abordent la « nudité », « le désir »), il semble que les usagers partent « à la rencontre des œuvres » par un regard solennel, avec un registre de vocabulaire signifiant cette posture : « faire penser, refléter, inspirer, dégager ». L'étudiante 3 présente les œuvres de Tate comme « des œuvres exposées ».

[9] Vidal Geneviève, 1999, « L'appropriation sociale du multimédia de musée. Les interactions entre pratiques de musée et de multimédia de musée », Thèse de doctorat, Université Paris 8, Saint-Denis, 1999.

Traduire des schémas RDF avec TransRDF(S)

Yolaine bourda, biêch-lien doan

Supélec,
3 avenue Joliot Curie,
91192 gif-sur-Yvette Cedex
Email : {Yolaine.Bourda,Biech-lien.Doan}@supelec.fr
Tél : +33 1 69 85 14 88 Fax : +33 1 69 85 14 99

Résumé

L'existence de schémas RDF différents sur un même domaine entraîne une perte d'interopérabilité. Traduire des descriptions RDF devient donc obligatoire dès qu'un utilisateur veut interroger des ensembles de descriptions RDF basés sur des schémas différents (même s'il ne sait pas que des schémas différents existent). Cette transformation des descriptions RDF doit être transparente à l'utilisateur et automatique. Elle repose donc sur une transformation préalable des schémas RDF. Mais, transformer un schéma en un autre doit se faire en préservant la sémantique. Cet article décrit TransRDF(S) un outil permettant d'une part, une transformation semi-automatique des schémas RDF tout en préservant leur sémantique et, d'autre part, une transformation automatique des descriptions.

Abstract

Multiple RDF schemas can be found for the same domain (education for example) resulting in the lack of interoperability. So, translating RDF descriptions is mandatory if a user wants to request descriptions using multiple schemas even if one doesn't know that different schemas exist. But, transforming one schema into another must be done while preserving the meaning of the schemas. This paper describes TransRDF(S) a tool allowing semi-automatic translation of RDF schemas and automatic translation of RDF descriptions, preserving the semantics.

1 Introduction

Les métadonnées peuvent être considérées comme une première étape vers le Web sémantique et RDF un moyen de les implémenter [Bourda and Hélier, 2000].

Une description RDF [Lassila and Swick, 1999] associe à une ressource une propriété ainsi que la valeur de celle-ci. Les schémas RDF [Brickley and Guha, 2002] permettent de définir les vocabulaires, classes et propriétés partagés par une communauté.

Il paraît peu probable que différentes communautés travaillant dans le même domaine partagent le même schéma RDF. Ainsi, comme il existe actuellement plusieurs ensembles de métadonnées décrivant des ressources du même type (comme par exemple les métadonnées pédagogiques), il existe aussi des schémas RDF différents.

Il faut donc pouvoir transformer, de façon automatique, des descriptions RDF en d'autres descriptions RDF basées sur des schémas différents. Ceci signifie qu'il faut parvenir à exprimer la transformation d'un schéma RDF en un autre schéma RDF. Celle-ci ne peut se faire que de façon semi-automatique. L'expression de la transformation de schémas doit permettre ensuite l'automatisation complète des transformations des descriptions RDF. Enfin, la transformation de schémas doit être conforme (elle doit respecter la sémantique et donc, entre autres, les hiérarchies de classes et de propriétés). le modèle d'un schéma RDF est un graphe [Hayes, 2003], donc des schémas RDF différents auront comme modèles des graphes différents.

Des travaux existent sur la transformation de schémas RDF, mais leur approche est différente de la nôtre. Triple [Sintek and Decker, 2002] est basé sur la logique de Horn et permet requêtes, inférences et transformations. RDF-T [Omelayenko, 2002] permet une spécification déclarative de la correspondance.

Ces travaux, bien que plus avancés que les nôtres, n'intègrent pas d'outil de validation complexe vérifiant :

- la conservation des hiérarchies des classes et des propriétés.
- l'analyse détaillée des contraintes `rdfs:range` et `rdfs:domain`

Dans la suite, nous notons les description RDF sous la forme de triplets.

2 Traduire des schémas RDF

Une des premières questions qui se pose est de savoir jusqu'où il est possible d'aller dans la transformation en prenant en compte quelques-unes des difficultés suivantes :

- N'ayant aucune connaissance de la sémantique, le programme est incapable de décider automatiquement les correspondances correctes (entre deux propriétés ou deux classes). En conséquence, celles-ci ne peuvent être trouvées que par un utilisateur.
- Des communautés différentes peuvent avoir (et même ont très souvent) des visions différentes non seulement sur la construction des schémas mais aussi sur le domaine. Ainsi, les contraintes sur les propriétés telles que `rdfs:domain` et `rdfs:range` peuvent être définies de façons différentes ou ne pas être définies du tout.

- La modularité et l'extensibilité des schémas RDF peuvent aussi être cause de problèmes. En effet, une communauté peut très bien étendre un schéma RDF existant alors qu'une autre, pour les mêmes besoins de modélisation, construira un schéma différent. On peut aussi faire une extension d'une extension d'une extension ...
- Les hiérarchies de classes et de propriétés doivent être respectées.

Afin de mieux comprendre ces problèmes, nous donnons ci-dessous, un petit exemple. Soit un schéma RDF `Schéma1` d'espace de nom `ex1` :

```
(ex1:Animal, rdfs:subClassOf, rdf:resource)
(ex1:Human,rdfs:subClassOf,Animal)
(ex1:Occupation, rdfs:subClassOf, rdf:resource)
(ex1:hasAnOccupation, rdf:type, rdf:Property)
(ex1:hasAnOccupation,rdfs:domain, ex1:Human)
(ex1:hasAnOccupation,rdfs:range, ex1:Occupation)
(ex1:instructs, rdf:type, rdf:Property)
(ex1:instructs,rdfs:domain, ex1:Human)
(ex1:instructs,rdfs:range, ex1:Human)
(ex1:Teacher, rdf:type, ex1:Occupation)
(ex1:Student, rdf:type, ex1:Occupation)
```

Soit un autre schéma RDF `Schéma2` d'espace de nom `ex2` :

```
(ex2:Person, rdfs:subClassOf, rdf:Resource)
(ex2:Animal, rdfs:subClassOf, rdf:Resource)
(ex2:Teacher, rdfs:subClassOf, ex2:Person)
(ex2:Student, rdfs:subClassOf, ex2:Person)
(ex2:teaches, rdf:type, rdf:Property)
(ex2:teaches,rdfs:domain, ex2:Teacher)
(ex2:teaches,rdfs:range, ex2:Student)
```

À première vue, les deux exemples ci-dessus décrivent la même chose : « des professeurs enseignent à des étudiants » et les propriétés `Instructs` et `Teaches` ont le même sens. Mais si nous essayons de transformer une description basée sur le premier schéma en une description basé sur le second, tout en préservant le sens, quelques problèmes arrivent :

- `Schéma1` utilise deux instances de la classe `Occupation` : `Teacher` and `Student`, tandis que `Schéma2` utilise deux classes (c'est-à-dire deux ensembles de ressources) avec le même nom.
- Le domaine et l'espace de nom de la propriété `Instructs` de `Schéma1` ne sont pas compatibles avec ceux de la propriété `Teaches` de `Schéma2`.
  Pourquoi `rdfs:range` et `rdfs:domain` sont-ils importants au point d'empêcher les deux propriétés de pouvoir être équivalentes ? Une instance de `ex1:Person` peut être n'importe quelle personne, et la propriété `Instructs` peut lui être appliquée sans aucune restriction (ainsi, un boucher peut instruire un apprenti).
  Ce qui est possible, par contre, c'est de définir `Instructs` comme une généralisation de `Teaches`. Ainsi, nous pouvons transformer une description basée sur `Schéma2`

en une description plus générale (donc avec une sémantique plus faible) basée sur `Schéma1`. L'inverse n'est bien évidemment pas possible.
– Si nous voulons préserver la hiérarchie des classes, nous ne pouvons pas dire, en même temps, que les classes `ex1:Animal` et `ex2:Animal` sont équivalentes ainsi que les classes `ex1:Human` et `ex2:Person`. En effet, supposons une description basée sur `Schéma1`, si nous la transformons, toute personne devient un animal. La sémantique de `Schéma2` n'est pas préservée (animaux et personnes sont des classes distinctes sans aucun lien d'héritage).
Ce qui est possible, par contre, c'est de dire que les classes `ex1:Animal` et `ex2:Animal` sont équivalentes (dans ce cas, `ex1:Human` et `ex2:Person` doivent être distinctes) ou que `ex1:Human` et `ex2:Person` sont équivalentes (dans ce cas, `ex1:Animal` et `ex2:Animal` doivent être distinctes).
La conservation de la hiérarchie des propriétés est aussi nécessaire que la hiérarchie des classes pour préserver la sémantique.

3 L'algorithme de transformation

Les principales étapes de l'algorithme de transformation de schémas RDF sont les suivantes :

1. Construire les hiérarchies de classes et de propriétés pour chaque schéma.
2. Définir des équivalences valides entre deux classes ou deux propriétés.
3. Finaliser l'ensemble des associations par déduction.
4. Générer une sortie qui permettra la transformation automatique des descriptions.

L'algorithme de transformation est basé sur la sémantique des schémas RDF qui peut être exprimée par quelques règles [Champin, 2001].

3.1 Construire les hiérarchies

La hiérarchie de classes est triviale à implémenter en utilisant la règle de transitivité et le fait qu'il existe une racine `rdf:resource`. Par contre, cette racine n'existe pas pour les propriétés. Il suffit de la créer et le problème se ramène alors au précédent.

3.2 Faire des associations valides

Cette étape est réalisée par l'utilisateur. C'est lui qui exprime que telle classe (respectivement telle propriété) du premier schéma est équivalente à telle classe (respectivement telle propriété) du deuxième schéma. Mais les équivalences proposées par l'utilisateur peuvent ne pas être valides, elles sont donc vérifiées par le programme.

Quand l'utilisateur associe une classe `C1` du schéma `S1` avec une classe `C2` du schéma `S2`, le programme vérifie que :

1. si une sous-classe de `C1` a un équivalent, celui-ci doit être une sous-classe de `C2`
2. si une super-classe de `C1` a a un équivalent ; celui-ci doit être une super-classe de `C2`.

Les équivalences entre propriétés doivent respecter les mêmes règles. Mais celles-ci ne sont pas suffisantes et il faut aussi prendre en compte les contraintes venant de `rdfs:domain` et `rdfs:range`. Si la propriété `P1` a comme `rdfs:domain D1` et la propriété `P2` a comme `rdfs:domain D2`, nous imposons que `D1` et `D2` soient équivalentes (si elles ont déjà été associées). Si `D1` a été associée à une sous-classe (super-classe) de `D2` alors `P1` est une spécialisation (généralisation) de `P2`. Si `D1` et `D2` n'ont pas été associées, nous en déduisons qu'elles sont équivalentes.

3.3 Finaliser l'ensemble des associations

Quand l'utilisateur a fini de faire toutes les équivalences qui lui paraissent nécessaires, le programme peut alors en déduire des associations comme : si une classe `C1` n'a pas d'équivalent mais que l'une de ses super-classes (`C2`) en a, `C1` devient une spécialisation de l'équivalent de `C2`.

Les règles pour les associations entre propriétés sont du même ordre en prenant, en plus, en compte le fait que deux propriétés peuvent déjà être dans une relation de généralisation/spécialisation.

3.4 Génération de la sortie

Quand toutes les correspondances possibles ont été faites entre les deux schémas, il est nécessaire de générer une sortie traitable par un ordinateur afin que les transformations des descriptions soient automatisées. Cette sortie est elle même une description RDF. Nous avons donc préalablement construit un schéma RDF qui définit les relations entre classes et propriétés dont voici un extrait :

```
(ese:equivalence,rdf:type, rdf:Property)
(ese:sameClass, rdfs:subPropertyOf, ese:equivalentTo)
(ese:sameProperty, rdfs:subPropertyOf, ese:equivalentTo)
(ese:specialization,rdf:type, rdf:Property)
(ese:specializationOfClass,rdf:type, rdf:Property)
(ese:specializationOfProperty,rdf:type, rdf:Property)
(ese:generalization,rdf:type, rdf:Property)
(ese:generalizationOfClass,rdf:type, rdf:Property)
(ese:generalizationOfProperty,rdf:type, rdf:Property)
```

La sortie générée par l'application sera donc une description RDF respectant ce schéma.

3.5 Transformer les descriptions

Il est alors simple d'écrire un programme (ce que nous avons fait) réalisant, à partir de la sortie générée, la transformation d'une description respectant un schéma en une

description respectant un autre schéma. Bien évidemment, toutes les classes et propriétés ne peuvent pas être transformées.

4 Conclusion

Un des services principaux offerts par notre application, écrite en Java, est la traduction automatique d'un ensemble de descriptions RDF en un autre. Un autre interêt est la création d'un langage de transformation, un schéma RDF encodant les relations existant entre les classes et les propriétés des deux schémas.

Nous avons construit un mécanisme simple, mais efficace pour extraire non seulement, les parties communes (ayant le même sens) des deux schémas mais aussi les relations de généralisation/spécialisation. L'application préserve les hiérarchies de classes et de propriétés, prend en compte les contraintes sur les propriétés et infère des relations à partir de celles données par l'utilisateur.

Cette transformation semi-automatique de schémas RDF pourrait néanmoins être améliorée en proposant des équivalences à l'utilisateur basées sur les noms des classes et propriétés.

Références

[Hayes, 2003] Hayes, P. (2003). Rdf semantics. `http://www.w3.org/TR/rdf-mt/`.

[Champin, 2001] Champin, P.-A. (2001). Rdf tutorial. `http://www710.univ-lyon1.fr/~champin/rdf-tutorial/rdf-tutorial.html`.

[Bourda and Hélier, 2000] Bourda, Y. and Hélier, M. (2000). Métadonnées, rdf et documents pédagogiques. In *Cahiers GUTenberg 35-36, Toulouse, May 2000,.*

[Omelayenko, 2002] Omelayenko, B. (2002). Rdft : A mapping meta-ontology for business integration. In *Proceedings of the Workshop on Knowledge Transformation for the Semantic Web (KTSW 2002) at the 15-th European Conference on Artificial Intelligence.*

[Brickley and Guha, 2002] Brickley, D. and Guha, R. (2002). Rdf vocabulary description language 1.0 : Rdf schema. `http://www.w3.org/TR/rdf-schema`.

[Lassila and Swick, 1999] Lassila, O. and Swick, R. R. (1999). Resource description framework (rdf), model and syntax specification. `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222`.

[Sintek and Decker, 2002] Sintek, M. and Decker, S. (2002). Triple—a query, inference, and transformation language for the semantic web. In *International Semantic Web Conference (ISWC).*

Extraction de la Terminologie du Domaine : Étude de Mesures sur un Corpus Spécialisé Issu du Web

M. Roche, O. Matte-Tailliez, J. Azé, Y. Kodratoff

Équipe Inférence et Apprentissage
Laboratoire de Recherche en Informatique
Bât 490, Université Paris-Sud
91405 Orsay Cedex, FRANCE.
Email : {roche,oriane,aze,yk}@lri.fr
Tél : 01 69 15 64 09 Fax : 01 69 15 65 86

Résumé

Le Web recèle de nombreuses données textuelles spécialisées qui peuvent être exploitées. Une des tâches intéressantes à effectuer est la construction d'ontologies spécialisées à partir des textes disponibles sur le Web. La première étape d'un tel travail consiste à déterminer la terminologie du domaine. Il existe de nombreuses mesures qui peuvent être utilisées afin d'extraire les termes pertinents d'un corpus. Les travaux que nous présentons dans cet article consistent à étudier la qualité des termes que l'on peut extraire en utilisant différentes mesures.

Abstract

Very large amounts of specialized textual data are presently available on the web, but are not yet really usable. One of the tasks necessary to start exploiting this wealth of knowledge is building specialized ontologies from texts. In order to carry out this task, gathering the terminology of the specialized domain is a first unavoidable step. This paper presents a study the quality of the different terminologies obtained by using several quality measures rating the terms.

1 Introduction

L'objectif, dans les travaux que nous présentons dans cet article, est de construire des ontologies [Gruber, 1995] à partir de corpus spécialisés issus du Web. L'utilisation des ontologies permet une recherche d'informations dans les textes plus efficace. En effet, dans les tâches d'extraction de données dans les textes, les ontologies permettent de construire des patrons d'extraction plus généraux [Freitag, 1998, Faure et Poibeau, 2000]. Les ontologies ont également un rôle essentiel dans la recherche de règles d'association dans les textes [Azé et Roche, 2003].

En général, la constitution de corpus spécialisés est une tâche difficile à effectuer. Avec l'émergence du Web, de nombreuses données textuelles sont plus facilement disponibles et la constitution de corpus spécialisés se révèle plus aisée. Ainsi, la constitution de corpus peut s'effectuer, en interrogeant des bases de données spécialisées ou en chargeant directement des pages Web (au format HTML et/ou XML). L'exploitation des corpus spécialisés au format HTML et/ou XML, demande une phase de nettoyage importante afin d'enlever les informations non pertinentes du corpus représentées, par exemple, par les balises. A contrario, l'exploitation des informations sémantiques contenues dans les balises XML aide à la construction d'ontologies spécialisées [Giraldo et Reynaud, 2002]. Dans nos travaux, nous nous intéresserons uniquement à l'exploitation des bases de données textuelles en ligne. En effet, le traitement des corpus constitués de pages hypertextes reste similaire au traitement d'un corpus textuel brut en incluant un pré-traitement afin de supprimer les balises et les en-têtes des fichiers hypertextes.
La première étape nécessaire pour la construction des ontologies à partir de corpus consiste à déterminer les termes pertinents des textes [Fontaine et Kodratoff, 2003]. Ces termes représenteront les instances des concepts des ontologies.

Dans cet article, nous allons nous intéresser à la recherche terminologique d'un corpus traitant de la Biologie Moléculaire. Après avoir décrit la collecte et le nettoyage du corpus (section 2), nous allons plus particulièrement nous attacher à déterminer la mesure la plus efficace afin d'extraire des termes pertinents pour un domaine (section 3). En effet, de nombreuses mesures existent dans la littérature afin d'ordonnancer statistiquement les termes extraits d'un corpus. Dans cet article, nous allons tester des mesures traditionnellement utilisées dans le domaine de l'extraction de la terminologie du domaine [Jacquemin, 1997, Daille et al., 1998] ainsi que des mesures du domaine de l'extraction de règles d'association [Lallich et Teytaud, 2003].

2 Collecte et nettoyage du corpus

2.1 Description du corpus issu du Web

Afin de constituer un corpus spécifique sur les protéines de la levure, nous avons effectué la requête "DNA-binding protein yeast" sur la base bibliographique Medline[1].

1. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

Ainsi, nous avons pu télécharger un corpus de plus de 6000 résumés d'articles. Le corpus que nous avons constitué a une taille relativement importante (10 Mo).

2.2 Nettoyage

Le nettoyage consiste à éliminer les bruits et à uniformiser le corpus. Le premier nettoyage a consisté à mettre au format notre corpus : enlever les noms des auteurs, supprimer les noms des laboratoires, éliminer les codes propres à la base de données, etc. Dans un second temps, nous avons mis en place des règles pour uniformiser le corpus pour notamment généraliser certains termes propres au domaine. Par exemple, nous avons remplacé les termes "carboxyl terminal", "carboxyl termini", "COOH-terminal", "CO2H-terminal", etc. par "C-term". L'établissement de telles règles de nettoyage est effectué manuellement par un expert du domaine. Un autre exemple de règles consiste à remplacer les noms de gènes spécifiques par leur nom générique admis dans la communauté biologique[2]. Avec ces informations propres au domaine, de telles règles peuvent être établies automatiquement.

3 Recherche terminologique

Comme dans de nombreux travaux sur la terminologie, nous nous sommes intéressés aux termes nominaux [Aussenac-Gilles et Bourigault, 2000, Bourigault et Jacquemin, 1999]. Une étude comparative des termes nominaux que nous avons extraits par rapport à une ontologie de référence dans le domaine[3] a montré que notre méthode permet d'extraire de très nombreux termes pertinents pour le domaine [Matte-Taillez et al., 2002].
Afin d'extraire les termes, une étape préliminaire consistant à étiqueter notre corpus est nécessaire (section 3.1).

3.1 Étiquetage

L'étiqueteur de Brill [Brill, 1994] appose une étiquette grammaticale à chacun des mots du corpus. Pour le corpus de biologie moléculaire, 70% des mots n'ont pas été reconnus en utilisant le lexique standard du logiciel. Avec l'intervention d'un expert du domaine de la biologie moléculaire, nous avons ajouté 30 règles lexicales adaptées à la biologie moléculaire. Avec de telles règles, il ne reste plus que 15% de mots inconnus dans le lexique de l'étiqueteur de Brill. Ces mots inconnus sont des mots très spécifiques au domaine qui n'ont pas été reconnus par nos règles lexicales. Une autre contribution a consisté à introduire une nouvelle étiquette appelée "Formule" qui est spécifique à la biologie. La règle permettant de repérer de telles formules (par exemple, `ABF1, Rad51, G3-Pa`), a été construite en collaboration avec l'expert du domaine. L'ensemble de ces adaptations de l'étiqueteur de Brill à notre domaine, nous a permis de construire un étiqueteur spécialisé que nous appelons *GenoBrill*.

2. La liste des noms génériques est disponible à l'adresse :
ftp://genome-ftp.stanford.edu/yeast/data_download/gene_registry/registry.genenames.tab
3. Gene Ontology : http://www.geneontology.org/

3.2 Extraction des termes et évaluation de leur pertinence

L'extraction des termes consiste à exhiber les mots voisins ayant une étiquette spécifique. Dans cette étude, nous allons nous intéresser aux candidats-termes binaires et ternaires de type *nom-nom (N-N)*, *adjectif-nom (Adj-N)*, *nom-préposition-nom (N-Prep-N)*, *nom-verbe_gérondif (N-VG)* et *formule-nom (For-N)*. Cependant, dans l'extraction des candidats-termes, nous n'avons pas retenu les candidats composés d'un mot non significatif tels que les adjectifs, "such", "more", "same", etc.

Le but de notre travail est d'exhiber parmi ces listes de candidats-termes, les termes les plus pertinents pour le domaine. Pour évaluer la pertinence de chacun des candidats-termes, nous utilisons la mesure de **Précision**, et son complément la courbe d'élévation ("lift chart"), qui sont les seules mesures dont nous puissions disposer dans un cadre d'apprentissage non-supervisé . Il est à noter qu'il est impossible de connaître le nombre exhaustif de termes pertinents du corpus, c'est la raison pour laquelle nous ne calculerons pas le Rappel, ni la courbe ROC.

La définition de la précision est indiquée ci-dessous où $\mathcal{E}$ représente une liste de termes trouvée par le système.

$$Précision = \frac{\text{nombre de termes pertinents présents dans } \mathcal{E}}{\text{Nombre de termes dans } \mathcal{E}}$$

La précision donne donc la proportion de termes corrects parmi les termes extraits.

Cependant, comme le précisent [Aussenac-Gilles et Bourigault, 2000], selon l'utilisation que l'on compte effectuer de la terminologie (indexation, ontologie, etc.), l'évaluation de la pertinence des termes peut différer. Pour notre part, nous jugerons qu'un terme est pertinent s'il représente une instance de concept de l'ontologie que nous construisons.

Après différents traitements, l'expert en biologie moléculaire a validé 7026 termes : termes jugés comme corrects s'ils représentent une instance de concepts et non corrects sinon. Nous précisons que, dans la liste $\mathcal{E}$ de termes obtenus en utilisant différentes mesures (section 3.4), les termes non expertisés ne seront pas pris en compte dans le calcul de la précision. De plus, dans notre évaluation de la qualité des termes extraits, nous utiliserons les courbes d'élévation consistant à donner la variation de la précision en fonction de la proportion de termes extraits par le système. Ainsi, selon la proportion de termes que nous fournissons à l'expert, nous pourrons évaluer la précision associée.

3.3 Candidats-termes extraits

La première étape de notre travail consiste à extraire l'ensemble des candidats-termes de notre corpus. Selon les types de relations, leur nombre varie mais celui-ci est globalement important (voir TAB. 1). Dans la suite de notre étude, nous avons effectué un élagage consistant à ne conserver que les termes ayant un nombre d'occurences supérieur à un certain seuil fixé par l'expert et ainsi de ne pas prendre en compte les candidats-termes apparaissant trop rarement dans le corpus. La taille du corpus est un critère essentiel afin

de fixer le seuil d'élagage. En effet, avec un corpus de taille conséquente, comme c'est le cas dans cette étude, il semble essentiel d'avoir un seuil d'élagage important afin de privilégier les termes les plus représentatifs du domaine.

	Adj-N	**N-N**	**N-prep-N**	**N-VG**	**For-N**
nb candidats-termes	23284	22241	4362	1943	6539
nb candidats-termes après élagage	2547	3332	151	171	319
% élagage	89%	85%	96%	91%	95%

TAB. 1 – *Candidats-termes avant et après élagage à* 4.

3.4 Comparaison de différentes mesures

Après élagage, le nombre de candidats-termes reste important, particulièrement pour certains types de relations. Il semble donc délicat de fournir à l'expert du domaine l'ensemble des candidats-termes. C'est pourquoi, il est nécessaire de fournir automatiquement à l'expert les termes les plus pertinents du domaine. Afin d'ordonnancer les termes selon leur pertinence, nous pouvons utiliser différentes mesures : mesures propres à la recherche terminologique [Jacquemin, 1997, Daille et al., 1998] mais également des mesures propres au domaine de l'extraction de règles d'association décrites dans [Lallich et Teytaud, 2003].

3.4.1 Description des mesures

Les premières mesures que nous allons expérimenter sont des mesures traditionnelles dans le domaine de l'extraction de la terminologie du domaine. Une des mesures couramment utilisée est l'information mutuelle : $log_2 \frac{P(A,B)}{P(A)P(B)}$. Cette mesure calcule l'indépendance de chacun des mots composant le candidat-terme. Une telle mesure a tendance à extraire des termes rares et peu fréquents. L'information mutuelle au cube [Daille et al., 1998] et la mesure d'association [Jacquemin, 1997] sont des mesures qui s'appuient sur l'information mutuelle mais en privilégiant davantage les candidats-termes fréquents. Nous allons, de plus, tester une mesure décrite dans [Jacquemin, 1997] également moins sensible aux candidats-termes de faible fréquence. Cette mesure est le coefficient de Dice qui est donné par la formule $\frac{2P(A,B)}{P(A)+P(B)}$.

Enfin, la dernière mesure du domaine de l'extraction de la terminologie que nous allons tester est le rapport de vraisemblance [Dunning, 1993]. Pour définir le rapport de vraisemblance, définissons, dans un premier temps, une table de contingence associée à chaque couple de mots (L_i, L_j) comme ci-dessous :

	L_j	$L_{j'}$ avec $j' \neq j$
L_i	a	b
$L_{i'}$ avec $i' \neq i$	c	d

Les valeurs a, b, c et d définissent les occurences des couples et $a+b+c+d=N$ est le nombre total d'occurences des couples trouvés.

On définit alors le rapport de vraisemblance ainsi :

$$RV(L_i, L_j) = a log(a) + b log(b) + c log(c) + d log(d) - (a+b) log(a+b) - (a+c) log(a+c) - (b+d) log(b+d) - (c+d) log(c+d) + N log(N)$$

Contrairement aux mesures précédentes, le rapport de vraisemblance ne prend pas uniquement en compte l'indépendance des deux mots du candidat-terme. Cette mesure prend également en compte la présence des candidats-termes formés avec un seul des mots ainsi que les candidats-termes formés avec aucun de ces mots.

La seconde série de mesures que nous allons décrire sont des mesures propres au domaine de l'extraction des règles d'association. La J-mesure, la mesure de conviction, la mesure de Sebag-Schoenauer et la moindre contradiction (ces mesures sont décrites dans [Lallich et Teytaud, 2003]), sont des mesures qui prennent en compte le nombre de contre-exemples $(A, \overline{B})$ d'un candidat-terme (A, B). Enfin, l'intensité d'implication [Gras, 1979] et l'intensité d'implication normalisée [Lerman et Azé, 2003] d'un candidat-terme (A, B) sont fondées sur l'étonnement statistique d'observer très peu de contre-exemples $(A, \overline{B})$ par rapport au nombre attendu sous l'hypothèse d'indépendance des mots A et B.

Nous précisons que dans le domaine de la fouille de textes, nous n'utilisons pas les mesures de la manière exacte dont elles sont décrites dans la littérature. Dans les mesures décrites précédemment, $P(A, B)$ correspond à la probabilité de rencontrer A et B. Une telle probabilité est symétrique c'est-à-dire que $P(A, B) = P(B, A)$. Cependant, dans les mesures que nous utilisons, $P(A, B)$ signifie la probabilité de rencontrer le mot A avant le mot B. Cet ordre a une influence dans le domaine de la fouille de textes c'est la raison pour laquelle, nous utiliserons ces mesures pour lesquelles l'effectif de (A, B) est, dans la majeure partie des cas, différent de l'effectif de (B, A). Nous précisons qu'il existe, en général, peu d'occurences d'une des deux formes.

3.4.2 Expérimentations

Selon les mesures utilisées et la proportion des termes fournis à l'expert, la précision varie. Dans TAB. 2, nous noterons les précisions obtenues pour les mesures testées si l'on fournit 100, 200, 500, 1000 termes à l'expert avec la relation ayant le plus grand nombre de candidats-termes avant élagage, en l'occurence la relation *adjectif-nom*. Parmi les termes que l'on extait, une proportion appartient aux 7026 termes analysés par l'expert. Nous noterons dans TAB. 2 cette proportion de termes entre parenthèses. Plus cette proportion est élevée et plus nous pouvons avoir confiance dans les mesures de précision données.

TAB. 2 montre que les mesures qui ont tendance à particulièrement bien se comporter sont le rapport de vraisemblance [Dunning, 1993], la conviction [Brin et al., 1997], l'intensité d'implication [Gras, 1979] et l'intensité d'implication normalisée [Lerman et Azé, 2003]. En effet, ces quatre mesures donnent la précision la plus élevée si l'on fournit moins de 500 termes aux experts. [Daille et al., 1998] montrent également que le rapport de vraisemblance est efficace dans les tâches d'extraction de la terminologie, ce que nous confirmons dans nos travaux. De plus, nous pouvons avoir une confiance significative dans ce résultat car le rapport de vraisemblance possède la proportion de termes expertisés la plus élevée.

De plus, cette étude confirme l'assertion de [Daille et al., 1998] qui précise que les

Nombre de termes proposés à l'expert		100	200	500	1000
1	Information mutuelle	89.0 (100%)	90.8 (76%)	92.2 (51%)	91.9 (43%)
2	Information mutuelle au cube	96.0 (100%)	**97.5** (100%)	94.0 (87%)	94.1 (61%)
3	Mesure d'association	90.0 (100%)	91.2 (80%)	93.0 (55%)	92.5 (46%)
4	Coefficient de Dice	92.0 (100%)	92.9 (92%)	92.6 (73%)	93.0 (53%)
5	Rapport de vraisemblance	**98.0** (100%)	**97.5** (100%)	**95.4** (92%)	94.1 (62%)
6	J-mesure	89.0 (27%)	89.1 (23%)	89.4 (26%)	**95.2** (42%)
7	Conviction	**96.9** (97%)	**97.4** (79%)	**97.2** (57%)	**95.2** (42%)
8	Sebag-Schoenauer	93.1 (58%)	94.9 (60%)	94.7 (53%)	**94.7** (43%)
9	Moindre contradiction	96.0 (99%)	96.1 (77%)	95.3 (43%)	**95.9** (32%)
10	Intensité d'implication	**99.0** (100%)	**96.6** (89%)	**95.5** (67%)	93.0 (50%)
11	Intensité d'implication normalisée	**99.0** (100%)	**96.6** (89%)	**95.5** (67%)	92.8 (50%)

TAB. 2 – *Précision (en %) selon le nombre de termes extraits pour la relation adjectif-nom. Le pourcentage entre parenthèses représente le pourcentage de termes analysés par l'expert parmi les termes extraits. Pour chaque nombre de termes proposés à l'expert, nous notons "en gras" les 4 mesures ayant la précision la plus élevée.*

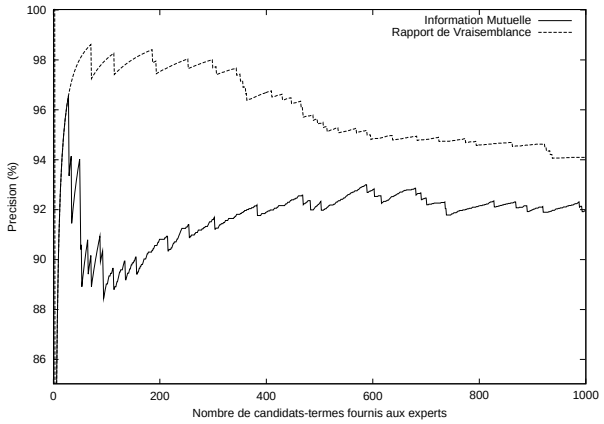


FIG. 1 – *Courbes d'élévation avec la relation adjectif-nom et un élagage à 4 pour l'information mutuelle et le rapport de vraisemblance.*

termes extraits avec l'information mutuelle [Church et Hanks, 1990] ne sont pas toujours de bonne qualité (voir FIG. 1). En effet, l'information mutuelle extrait des termes rares et spécifiques et pas nécessairement très représentatifs pour le domaine. A contrario, l'information mutuelle au cube [Daille et al., 1998] et la mesure d'association [Jacquemin, 1997] se comportent mieux que l'information mutuelle dans notre tâche d'extraction de la terminologie pour notre classification conceptuelle. Ceci s'explique par le fait que ces mesures, fondées sur l'information mutuelle, favorisent davantage les termes fréquents.

Nous remarquons de plus que l'ensemble des mesures utilisées dans d'autres domaines telles que l'extraction des règles d'association se comportent globalement bien pour l'extraction de la terminologie du domaine. Toutes ces mesures donnent une précision élevée,

en particulier, l'intensité d'implication et l'intensité d'implication normalisée (voir TAB. 2). De plus, avec ces mesures, la proportion de termes expertisés permettant de calculer la précision est assez élevée. Ceci donne une confiance intéressante dans la précision calculée avec ces mesures. Comme le précisent [Lerman et Azé, 2003], nous pourrions expliquer le peu de différence entre l'intensité d'implication et l'intensité d'implication normalisée par le nombre, peut-être trop restreint, de candidats-termes pour ces deux mesures.

4 Méthodologie globale de l'extraction de la terminologie du domaine

Le but fixé dans nos travaux est d'extraire de manière automatique les termes les plus pertinents pour le domaine. Cette étape est essentielle afin de faciliter et de guider le travail de l'expert dans l'objectif de construire une ontologie du domaine.

Dans cette étude, nous avons évalué la mesure la plus adaptée pour extraire les termes binaires. Cependant, dans de nombreux domaines et particulièrement dans le domaine de la biologie, il est nécessaire d'extraire des termes composés de plus de deux mots, c'est-à-dire des termes spécifiques. Notre algorithme de recherche terminologique consiste à introduire les termes binaires trouvés dans le corpus nettoyé en ajoutant un trait d'union entre chacun des mots constituant le terme binaire. Ainsi, ce terme sera reconnu comme un mot à part entière par l'étiqueteur de Brill. Dans ce corpus, avec prise en compte de la terminolgie trouvée lors d'une première passe, nous pouvons effectuer une nouvelle itération de recherche terminologique suivant le même principe que celui développé dans cet article. Ceci permet alors de former des termes composés d'un nombre de mots plus important et particulièrement intéressant en biologie[4].

De plus, afin de privilégier le vocabulaire du domaine, nous avons ajouté différents paramètres développés dans [Roche, 2003]. Un des paramètres essentiel consiste, par exemple, à privilégier les termes composés de mots inclus dans les termes extraits aux itérations précédentes de notre algorithme de recherche terminologique.

5 Conclusion et perspectives

L'étude des différentes mesures permettant d'extraire des termes de qualité est primordiale pour les tâches de classification conceptuelle. Dans les expérimentations que nous avons effectuées, nous avons remarqué que les mesures utilisées dans le domaine de la recherche des règles d'association se comportaient bien pour l'extraction de la terminologie. Par exemple, avec la conviction [Brin et al., 1997], l'intensité d'implication [Gras, 1979] et l'intensité d'implication normalisée [Lerman et Azé, 2003], mesures typiques du domaine de la recherche des règles d'association, nous obtenons une précision élevée. Cependant, nous confirmons l'assertion de [Daille et al., 1998] précisant que le rapport de vraisemblance [Dunning, 1993], typiquement utilisé pour l'extraction de la terminologie, est une

4. Des centaines d'exemples de termes que nous avons extraits à partir de notre corpus sont consultables à l'adresse `http://www.lri.fr/ia/Genomics/`.

mesure permettant d'extraire de nombreux termes pertinents. Afin d'augmenter la qualité des termes proposés à l'expert, une solution pourrait consister à selectionner seulement les termes communs trouvés en utilisant différentes mesures.

Une caractérisation précise des types de termes que l'on extrait pourrait être intéressante à mener. Par exemple, si l'on cherche à extraire des termes rares et spécifiques, nous pouvons utiliser l'information mutuelle. A contrario, si nous cherchons des termes particulièrement représentatifs d'un domaine, par exemple pour la construction d'ontologies, le rapport de vraisemblance, la conviction, l'intensité d'implication ou l'intensité d'implication normalisée semblent plus appropriés.

La méthodologie et les mesures choisies afin d'extraire des termes issus d'un corpus spécifique semblent avoir des résultats prometteurs. Nous avons obtenu des résultats particulièrement intéressants sur d'autres corpus [Roche, 2003] qui ne sont pas issus du Web. Il serait intéressant de valider notre méthode sur de nouveaux corpus spécifiques de domaines différents. Avec les nombreuses données du Web, la collecte de corpus spécialisés est facilitée. Comme, nous l'avons effectué ici, une telle étude nécessiterait, non seulement un travail rigoureux dans le pré-traitement des données (nettoyage, étiquetage) mais également dans l'expertise des termes extraits.

Dans notre étude, nous nous sommes attachés à l'étude des termes nominaux. Une prochaine étape de notre travail consistera en l'étude des relations verbales.

Références

[Aussenac-Gilles et Bourigault, 2000] Aussenac-Gilles, N. et Bourigault, D. (2000). The Th(IC)2 initiative: Corpus-based thesaurus construction for indexing www documents. Dans *Proceedings of the EKAW'2000 Workshop on Ontologies and Texts, Vol-51.*

[Azé et Roche, 2003] Azé, J. et Roche, M. (2003). Une application de la fouille de textes : l'extraction des règles d'association à partir d'un corpus spécialisé. *Revue RIA-ECA numéro spécial EGC03*, 17:283–294.

[Bourigault et Jacquemin, 1999] Bourigault, D. et Jacquemin, C. (1999). Term extraction + term clustering: An integrated platform for computer-aided terminology. Dans *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL '99), Bergen.*, pages 15–22.

[Brill, 1994] Brill, E. (1994). Some advances in transformation-based part of speech tagging. Dans *AAAI, Vol. 1*, pages 722–727.

[Brin et al., 1997] Brin, S., Motwani, R., et Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. Dans *Proceedings of ACM SIGMOD'97*, pages 265–276.

[Church et Hanks, 1990] Church, K. W. et Hanks, P. (1990). Word association norms, mutual information, and lexicography. Dans *Computational Linguistics*, volume 16, pages 22–29.

[Daille et al., 1998] Daille, B., Gaussier, E., et Langé, J. (1998). An evaluation of statistical scores for word association. Dans *J.Ginzburg, Z. Khasidashvili, C. Vogel, J.-J. Levy, and E. Vallduvi (eds) The Tbilisi Symposium on Logic, Language and Computation: Selected Papers, CSLI Publications*, pages 177–188.

[Dunning, 1993] Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

[Faure et Poibeau, 2000] Faure, D. et Poibeau, T. (2000). First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX. Dans *Actes du workshop Ontology Learning, 14 th European Conference on Artificial Intelligence.*

[Fontaine et Kodratoff, 2003] Fontaine, L. et Kodratoff, Y. (2003). Comparaison du rôle de la progression thématique et de la texture conceptuelle chez les scientifiques anglophones et francophones s'exprimant en anglais. Dans *Journée de Rédactologie scientifique : L´écriture de la recherche. Nantes. Publication à paraître.*

[Freitag, 1998] Freitag, D. (1998). Toward general-purpose learning for information extraction. Dans Boitet, C. et Whitelock, P., editors, *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pages 404–408. Morgan Kaufmann Publishers.

[Giraldo et Reynaud, 2002] Giraldo, G. et Reynaud, C. (2002). Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine. Dans *13èmes Journées Francophones d'Ingénierie des Connaissances.*

[Gras, 1979] Gras, R. (1979). Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques. Master's thesis, Université de Rennes 1.

[Gruber, 1995] Gruber, T. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer*, pages 907–928.

[Jacquemin, 1997] Jacquemin, C. (1997). Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Dans *Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes.*

[Lallich et Teytaud, 2003] Lallich, S. et Teytaud, O. (2003). Evaluation et validation de l'intérêt des règles d'association. *Contribution au rapport d'activité du Groupe Qualité de l'action Gafo-Données, à paraitre.*

[Lerman et Azé, 2003] Lerman, I. C. et Azé, J. (2003). Une mesure probabiliste contextuelle discriminante de qualité des règles d'association. *RSTI série RIA-ECA*, 17(1-2-3):247–262.

[Matte-Taillez et al., 2002] Matte-Taillez, O., Roche, M., et Kodratoff, Y. (2002). A precise automatic extraction of terminology in genomics. Research Report 1344, LRI.

[Roche, 2003] Roche, M. (2003). Extraction paramétrée de la terminologie du domaine. *Revue RIA-ECA numéro spécial EGC03*, 17:295–306.

Chaînes de Markov Combinées pour la Prédiction de Parcours WWW

Y. HAFRI

Institut National de l'Audiovisuel,
4, avenue de l'Europe
94366 Bry-sur-Marne Cedex, FRANCE.
Email : yhafri@ina.fr
Tél : +33 1 49 83 33 07 Fax : +33 1 49 83 28 82

Résumé

Le réseau Internet donne accès à plusieurs centaines de millions de sites et s'enrichit d'un million de pages par jour. Toutefois, en raison de son développement rapide et anarchique, Internet reste un réseau d'informations sans organisation ni structure, de telle sorte qu'une recherche efficace y est difficile. Nous nous intéressons dans cet article à l'adaptation structurelle automatique des sites Web. Nous proposons un outil offrant un accès à une information hypermédia contextualisée et personnalisée selon le profil de l'utilisateur. A cet effet, nous utilisons une forme particulière de Modèles de Markov que nous avons appelé *Chaînes de Markov Combinées*. Ces modèles spécialisés ont été développés pour la prédiction des pages Web susceptibles d'intéresser les internautes au cours de leur navigation. Nous présentons également quelques applications annexes de notre approche à d'autres problématiques telles que la prédiction de requêtes HTTP et la génération de parcours virtuels.

Abstract

Internet network gives access to several hundreds of million sites, but because of its lack of organization, the need for reliable navigation tools becomes a major issue for effective information retrieval. Our work describes a stochastic tool for accessing hypermedia information according to the user profile. The personalisation of the navigation is based on combined Markov's models, dynamically predicting next visited pages from former ones. We will also apply our model to the prediction of HTTP requests according to the profile and the generation of virtual tours.

1 Introduction

L'explosion du nombre de sites Web connectés sur Internet et la croissance du nombre d'internautes confirment de plus en plus la position du Web comme un média de masse. Par conséquent, la recherche d'informations dans ce genre de média est un problème difficile. Dans ce contexte, de nombreux travaux se sont intéressés à étudier la problématique de l'auto-adaptation des sites Web. Un site Web est dit *adaptatif*, s'il peut évoluer automatiquement en fonction de l'intérêt des utlisateurs. L'objectif d'un processus d'adaptation est d'améliorer l'utilité d'un site en tirant des enseignements de l'usage du site. Par exemple, du point de vue de l'utilisateur, l'objectif peut être de faciliter l'accès aux informations fournies par le site (ex. raccourcis hypertextes) ou de filtrer le contenu du site pour mieux cibler l'utilisateur. Dans cet article nous proposons une approche originale pour l'auto-adaptation structurelle de site Web basée sur les Modèles de Markov.

La partie 2 présente les principes de base des chaînes de Markov (CM) et décrit leurs utilités dans un contexte d'aide à la navigation. La partie 3 présente une nouvelle extension des CM appelée *Chaînes de Markov Combinées*. Cette extension riche en avantages nous permettra de traiter la problématique de manière plus précise que les modèles classiques. En partie 4, nous détaillons les résultats d'expérimentations de notre approche et son application à deux problèmes annexes à savoir la prédiction de requêtes HTTP et la génération automatique de parcours virtuels. Nous comparons ensuite notre modèle à plusieurs approches existantes en 5 et finirons en partie 6 par une conclusion et des perspectives.

2 Chaînes de Markov combinées pour la prédiction de liens

Avant de décrire notre système, nous allons présenter le modèle probabiliste de base que nous avons choisi pour la représentation des traces (sessions de navigation) des utilisateurs sur le Web : les *modèles de Markov observables*.

D'une manière générale, un processus ou modèle stochastique observable est un processus aléatoire qui peut changer d'état s_i, $i = 1,...,n$ au hasard, aux instants $t = 1,2,...,T$. Dans notre modèle, chaque état représente *une ou plusieurs pages Web*. Le résultat observé est la suite d'états dans laquelle le processus est passé. Ce dernier émet des séquences $S = s_1,s_2,...,s_T$ avec une probabilité $P(S) = P(s_1,s_2,...,s_T)$. Afin de calculer $P(S)$, il faut se donner la probabilité initiale $P(s_1)$ et les probabilités d'être dans l'état s_t, connaissant l'évolution antérieure entre les états. Un processus stochastique est *markovien*[1] (ou *de Markov*) si son évolution est entièrement déterminée par une probabilité initiale et des probabilités de transitions entre états. Autrement dit, en notant ($q_i = s_i$) le fait que l'état observé à l'instant t est s_i, alors :

$$\forall t, \qquad P(q_t = s_i | q_{t-1} = s_j, q_{t-2} = s_k \ ...) = P(q_t = s_i | q_{t-1} = s_j)$$

$$\text{d'où :} \qquad P(q_0...q_T) = P(q_0) \times \prod_{t=1}^{T} P(q_t | q_{t-1})$$

1. Au sens strict : markovien d'ordre k=1.

Ceci nous autorise à définir une matrice de probabilité de transitions $A = [a_{ij}]$ entre états telle que :

$$a_{ij} = P(q_t = s_j | q_{t-1} = s_i) \quad 1 \leq i,j \leq n \quad \text{avec} \quad \forall i,j \quad a_{ij} \geq 0 \quad \text{et} \quad \forall i \quad \sum_{j=1}^{j=n} a_{ij} = 1$$

En ce qui nous concerne, à savoir la prédiction des prochaines actions entreprises par un internaute, a_{ij} pour un modèle d'ordre 1 représente la probabilité de passage d'une page i à une autre page j à un instant t.

Plusieurs travaux [16, 1, 9, 8] ont démontré que les chaînes de Markov d'ordre $k = 1$ n'assuraient pas une bonne prédiction des actions effectuées par les utilisateurs. La principale raison est la prise en compte d'un seul état d'historique, ce qui rend difficile la discrimination entre différents *types de comportements* et qui restreint de façon non négligeable la couverture des données (taille de l'historique considéré). Une solution originale introduite par [16] pour ajuster les prédictions est d'entraîner plusieurs chaînes de différents ordres $k \geq 1$ et de combiner leurs prédictions respectives. Nous appellerons ces chaînes : *Chaînes de Markov Combinées* (CCM).

Cependant, les CCM souffrent de deux inconvénients majeurs : (i) une complexité de l'espace des états et (ii) une prédiction non pertinente dans certains cas (taille des sessions trop petite <= 3 états). Néanmoins, la principale raison reste le nombre d'états dans ces modèles qui augmente de façon exponentielle et limite de ce fait les performances de calcul pour les systèmes embarqués ou pour des applications ayant des contraintes temporelles.

3 Construction et optimisation des chaînes de Markov combinées

Le modèle que nous avons adopté tente de combiner de façon intelligente différentes CM afin d'avoir un espace réduit d'états, une plus grande précision lors des prédictions et une couverture totale des parcours. L'idée sous-jacente est l'élimination des états/transitions estimés comme générant des prédictions de faible précision par l'application des phases suivantes :

- ☞ Nous considérons dans cette étape qu'un état intervenant peu ou pas du tout lors de l'apprentissage n'est pas un bon prédicteur. Tous les états entrant en jeu dans moins de φ exemples d'apprentissage seront éliminés. φ est le seuil minimum à franchir pour garder un état.
- ☞ Les CM combinées sont testées sur des exemples de validation et pour chaque état on retient le nombre de *bonnes prédictions* qu'il a effectuées. Si un état d'ordre 4 correspondant à l'action $\{s_{12},s_1,s_4,s_7\}$ a prédit plus de mauvaises actions que son sous état d'ordre 3 $\{s_1,s_4,s_7\}$, il sera immédiatement supprimé. Cette procédure ne s'applique pas aux états d'ordre 1 pour garder une bonne couverture des données.

Une fois les CCM construites et optimisées, les prédictions sont effectuées en utlisant la chaîne d'ordre k, ensuite la chaîne d'ordre $k-1$ et ainsi de suite jusqu'à la chaîne d'ordre 1.

$sw_1 : \{s_3,s_2,s_1\}$
$sw_2 : \{s_3,s_5,s_2,s_1,s_4\}$
$sw_3 : \{s_4,s_5,s_2,s_1,s_5,s_4\}$

2*ème* ordre	s_1	s_2	s_3	s_4	s_5
$s_1 = \{s_1,s_4\}$	0	0	0	0	0
$s_2 = \{s_1,s_5\}$	0	0	0	1	0
$s_3 = \{s_2,s_1\}$	0	0	0	1	1
$s_4 = \{s_3,s_2\}$	1	0	0	0	0
$s_5 = \{s_1,s_5\}$	0	0	0	1	0
$s_6 = \{s_4,s_5\}$	0	1	0	0	0
$s_7 = \{s_5,s_2\}$	2	0	0	0	0
$s_8 = \{s_3,s_5\}$	0	1	0	0	0

TAB. 1 – *Exemple de sessions de navigation et la matrice des fréquences correspondante*

Considérons un exemple simple de 3 sessions de navigation sw_1,sw_2,sw_3 du tableau 1. Prédire la prochaine page d'un utilisateur ayant s_1,s_5,s_2 comme parcours par une chaîne d'ordre d'ordre 2 se fait comme suit : on identifie d'abord l'état s_7 associé à s_5,s_2, ensuite, on recherche la page s_i de plus haute probabilité correspondant à s_7 dans la matrice de transition. Dans notre cas, le système suggèrera la page s_1. Si cette prédiction est différente de 0, on s'arrête, sinon on essaye de prédire avec une chaîne d'ordre 1.

L'objectif final étant la mise au point d'un système de prédiction *en ligne*, nous avons défini deux paramètres empiriques pour la création et la mise à jour des différentes matrices de transitions a_{ij}.

- Le paramètre $\alpha \in [0,1]$ pour le renforcement des pages les plus *récemment* visitées. A chaque accès à une page, ce paramètre détermine celle dont la probabilité de visite sera augmentée ou diminuée en fonction de son âge. Plus une page est ancienne et moins sa probabilité sera élevée. Respectivement, plus une page est récente et plus sa probabilité se verra augmentée.
  Considérons le schéma suivant où un individu effectue le parcours s_i,s_j. Au moment où la page s_j est chargée, toutes les probabilités de transition de toutes les chaînes en ligne i seront mises à jour comme suit : $a_{i,k} = a_{i,k} \times \alpha$ avec $k \notin j$. Du fait que la page s_i est la plus récemment visitée (s_i précède s_j), sa probabilité sera augmentée par la formule suivante $a_{i,j} = a_{i,j} \times \alpha + (1 - \alpha)$.
  Notons que pour $\alpha = 0$, l'algorithme correspond à un apprentissage qui favorise uniquement la prédiction des pages visitées à l'étape q_{t-1} (étape précédente) et que pour $\alpha = 1$, il correspond à un algorithme qui ne change jamais ses probabilités (distribution uniforme).
- Le paramètre $\beta \in [0,1]$ pour le renforcement des pages *nouvellement* visitées. A chaque accès à une nouvelle page s_i, sa probabilité initiale est calculée comme suit :

$$P(q_0 = s_i) = \frac{1}{N} \times \beta + (1 - \beta)\frac{\mu_i}{\sum_i \mu_i}$$

  où N représente le nombre d'états du système et μ_i le nombre de fois où la page s_i a été accédée.
  Pour $\beta = 0$ l'algorithme fournit une distribution initiale selon le maximum de vraisemblance et pour $\beta = 1$ il correspond à une distribution initiale équiprobable.

4 Expérimentations

Nous avons évalué les performances des CCM sur un grand jeu de données (776986 sessions). Les logs proviennent d'un site de e-commerce *www.gazelle.com* fourni par KDDCup [2]. Ces logs ont tout d'abord été nettoyés et transformés en sessions de navigation de longueur minimale $minLength \geq 3$ états. Notons que ces dernières contiennent uniquement les accès aux pages car ceux aux images ont été ignorés.

Afin d'effectuer une validation croisée, nous avons divisé les données initiales en deux parties égales (388493 sessions chacune) : une d'apprentissage est une autre de test. 300 échantillons aléatoires ont été tirés pour la phase de test à raison de 40% par échantillon. Nous avons apporté quelques modifications à notre modèle pour obtenir celui de Jacobs et Blockeel[10] et l'avons comparé à deux versions : celle des CCM classiques et celle utilisant les deux paramètres alpha et beta décrits dans la partie 3.

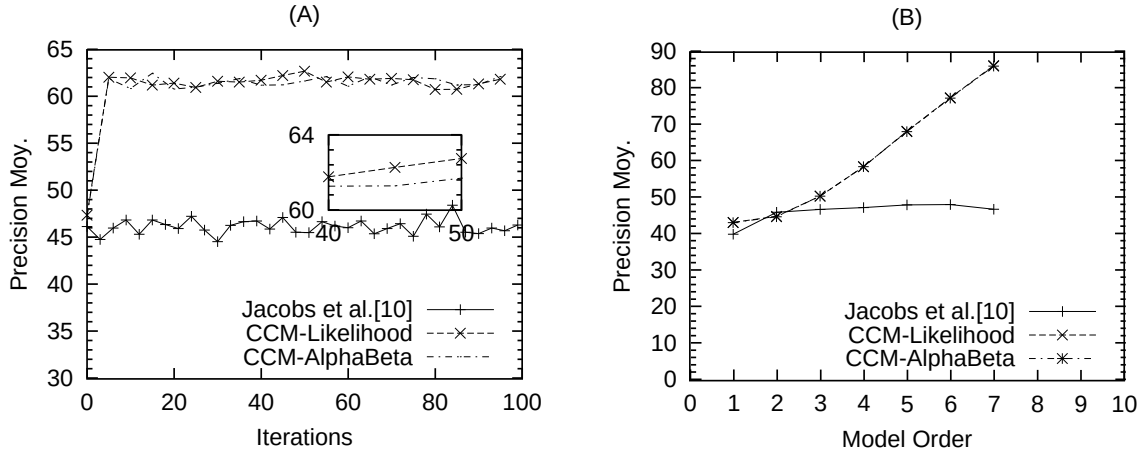


FIG. 1 – *(A) Prédiction moyenne des CCM (B) et Prédiction moyenne par modèle*

La figure 1.A montre que si les résultats de Jacobs et Blockeel atteignent au maximum 47.81% de bonnes prédictions, les CCM dépassent le taux de 62% et généralisent très bien (sans overfitting) face à des données inconnues. Notons cependant que la courbe CCM-AplhaBeta fluctue moins que la courbe des CCM classiques (voir zoom figure 1.A).

La figure 1.B valide bien notre hypothèse de départ ; utiliser plusieurs chaînes d'ordres différents au-lieu d'une seule chaîne d'ordre 1 vient renforcer la prédiction pour les 3 modèles testés. Néanmoins, les CCM bénéficient d'une optimisation au niveau de leurs états et transitions ce qui leur confèrent le moyen d'éviter des distributions de probabilités spécifiques pouvant engendrer des sur-apprentissages.

Le tableau 2 (partie gauche) compare les variations du taux de bonne prédiction (Pred.) à l'ordre du modèle. La taille de chaque modèle est indiquée par la colonne (S.States) et la taille cumulée par (C.States). De la même façon, le nombre de transitions pour chaque modèle est indiqué par la colonne (S.Edges) et celui cumulé par (C.Edges). On passe alors d'une précision

2. http://www.ecn.purdue.edu/KDDCUP/

Ordre	Pred.	S. States	C.States	S.Edges	C.Edges	Pred.	S.States	C.States	S.Edges	C.Edges
1	40.72%	1543		4000		47.29%	60		489	
2	48.20%	1631	3174	4265	8265	56.84%	528	588	1607	2096
3	48.66%	1723	4897	4543	12808	59.96%	606	1194	3067	5163
4	49.48%	1826	6725	4835	17643	60.73%	711	1905	4555	9718
5	47.07%	1939	8664	5146	22789	60.80%	942	2847	5859	15577
6	46.65%	2055	10719	5477	28266	61.95%	1081	3928	6770	22347
7	47.38%	2183	12902	5828	34094	62.07%	1199	5127	7229	29576

TAB. 2 – *Résultats de l'optimisation des CCM*

de 40.72% avec une seule chaîne d'ordre 1 à une précision de 47.38% en combinant 7 chaînes. Le prix à payer est une explosion du nombre d'états passant de 1543 états (resp. 4000 transitions) à 12902 états (resp. 34094 transitions). La partie droite du tableau 2 montre les résultats des CCM-AlphaBeta avec les paramètres $\alpha = 0.75$, $\beta = 0.80$ et $\varphi = 23$. Le modèle optimisé arrive à prédire les actions des utilisateurs avec un taux de 62.07% en passant de 12902 états (resp. 34094 transitions) à 5127 états (resp. 29576 transitions).

4.1 Applications annexes

Dans cette partie, nous allons présenter deux applications annexes de notre modèle.

4.1.1 Prédiction de requêtes HTTP

Si l'accès à l'information dans le Web est un problème crucial, la capacité des serveurs à fournir cette information d'une manière rapide est tout aussi importante. La prédiction des liens peut être utilisée pour *précharger* des documents lourds (images, vidéos, etc.) pendant que l'utilisateur parcourt le document actuel. Ce qui permet au serveur de réduire le temps de latences des requêtes.

Beaucoup de travaux ont porté sur l'analyse des requêtes HTTP afin d'améliorer les performances des serveurs Web. La majeure partie s'est concentrée sur l'étude de la taille des fichiers, les types des requêtes et les mécanismes de cache [3]. A notre connaissance, peu de travaux utilisant les modèles de Markov ont été appliqués pour prédire les requêtes HTTP [2, 17]. L'adjonction d'une CCM à un serveur ou un proxy est très aisée. Le client envoie une requête au serveur, ce dernier utilise le module de prédiction de liens et *anticipe* ses prochaines requêtes en se basant sur son historique.

4.1.2 Génération de parcours virtuels

La génération de parcours virtuels a été étudiée dans plusieurs travaux comme ceux de [11, 17]. Nous allons présenter notre algorithme (algo. 1) simple de génération de parcours virtuels évoluant au cours du temps (dynamique) et basé sur le modèle CCM. L'algorithme reçoit en entrée une simple URL de départ et génère ensuite une séquence d'URLs en utilisant les CCM. La séquence est affichée au client telle une visite guidée. L'avantage de l'algorithme,

1. /openperl/index.html	9. /openperl/syntaxe/fonction.html
2. /openperl/intro/exemple/hello_world.pl	10. /openperl/in_out/index.html
3. /openperl/structure/data.html	11. /openperl/regex/index.html
4. /openperl/structure/tableau.html	12. /openperl/regex/metacaractere.html
5. /openperl/syntaxe/commande.html	13. /openperl/regex/recherche.html
6. /openperl/structure/variable.html	14. /openperl/regex/substitution.html
7. /openperl/structure/hash.html	15. /openperl/regex/translation.html
8. /openperl/syntaxe/operateur.html	

TAB. 3 – *Exemple de génération d'un parcours prédéfini*

par rapport aux travaux cités, est sa capacité à générer des parcours différents au cours du temps même si l'URL de départ est la même. L'algorithme fait face à quelques obstcales pour générer des parcours intéressants. Les états produits par l'algorithme peuvent apparaître plus d'une fois (formation de cycle), mais en marquant ceux déjà rencontrés, on évite de les revoir. Un autre problème surgit quand toutes les probabilités à partir d'un état sont équiprobables ou en deça d'un certain seuil ϵ. Dans ce cas, on redémarre l'algorithme à partir de l'état précédent. Enfin, on contourne les nœuds à choix multiples (i.e. accès à plusieurs états avec la même probabilité) en sélectionnant le lien ayant le plus grand préfixe URL en commun avec l'état courant ou en le choisissant au hasard.

Algorithm 1 Algorithme de génération de parcours

Require: $CM\ d'ordre\ k$;

soit s_0 l'état initial, V une liste des états visités, Q une pile et s' := s_0;

while longueur du parcours n'est pas atteinte **or** condition de sortie non vérifiée **do**

for $\forall$ les états non visités s' **do**

$add(V,s')$; $push(Q,s')$; ## Calculer $P(s' \rightarrow s")$ appartenant à la CM de plus grand ordre. Mettre les états correspondants dans S ##

$S \leftarrow \max\{P(s' \rightarrow s") > \epsilon\}$;

if $|S| > 1$ **then**

choisir s" de S tel que URL(s") et URL(s') possèdent le plus grand préfixe **or** s" := $rand(S)$;

end if

if $|S| = 0$ **then**

$del(V,s')$; $pop(Q,s')$; **next**;

end if

s' := s";

end for

end while

L'exemple du tableau 3 met en évidence un échantillon limité à 15 URLs produites par l'algorithme sur un cours d'introduction au langage Perl. Il est cependant difficile d'évaluer qualitativement le parcours généré pour un site aussi petit (40 pages seulement). Une interprétation générale de ce parcours peut être la suivante : commencer par la page d'introduction puis tester l'exemple hello_world.pl. Aller ensuite à la partie structure de données et syntaxe et voir les entrées-sorties puis terminer par les expressions régulières.

5 Travaux antérieurs

Dans Perkowitz [14], le problème a également été traité d'une manière un peu différente. Pour prédire les actions futures d'un internaute, Perkowitz a utilisé une heuristique de prédiction très simple : pour chaque page S du site, il calcule le nombre de fois où les pages P sont accédées après avoir visité S. Quand S est re-visitée une nouvelle fois, m liens sont affichés par ordre décroissant de fréquence des pages P. Effectivement, cette heuristique est tout à fait applicable. Notre système considère aussi ces probabilités mais fournit en plus un modèle comportemental qui cible chaque utilisateur particulièrement, alors que l'approche de Perkowitz dégage plutôt un comportement globale.

Letizia [13] est un agent côté client qui parcours le web en parallèle avec l'internaute. En se basant sur les actions effectuées par ce dernier (ex. liens suivis, pages mises en bookmark, etc.), Letizia estime son *intérêt*, traite différentes pages et les lui propose avant même qeu celui-ci n'y ait accédé. A la différence de notre approche qui se place du côté serveur, Letizia est contraint par les ressources du client (ce qui n'est pas intéressant pour des systèmes embarqués PDA, Mobile, etc). De plus, Letizia ne prend pas en compte les expériences d'autres utilisateurs face au même site et se contente de l'expérience d'un seul internaute à la fois.

SurfLen [6] et PageGather [15] proposent une solution à ce problème en se basant sur les co-occurences des requêtes effectuées par le passé. Ces algorithmes suggèrent les m pages les plus probables d'apparaître dans la session de l'utilisateur sous forme d'une liste de liens à suivre (SurfLen) ou en construisant une page d'index (PageGather). L'incovénient est que ces deux systèmes proposent parfois de longues listes d'URLs pouvant dérouter la recherche. Dans notre approche, le système se contente d'afficher les 7 meilleurs liens que le système a trouvé.

MINPATH [1] vient en aide aux Wireless Devices et n'a qu'un seul but à satisfaire : *minimiser le nombre de clicks pour atteindre l'information voulue*. Le but du système est de maximiser une fonction de gain *expected savings* en parcourant N niveaux dans la structure du site après chaque accès à une page. A partir de la page courante, MINPATH propose les meilleurs *raccourcis* pour accélérer la navigation. Néanmoins, plusieurs paramètres sont à fournir en entrée comme le nombre de clusters à construire et la profondeur de recherche dans le site.

ILASH [4, 5] est une extension du shell UNIX tcsh. Il permet à l'utlisateur d'éviter les fautes de frappes lors d'exécution de commandes et offre un mode de complétion automatique très agréable. Jacobs et Blockeel [10] se sont fortement inspirés des travaux de Davison [4, 5] et ont apporté une amélioration à ce système. Au lieu d'apprendre avec des modèles de Markov à une seule mémoire, l'appentissage se fait sur les n dernières commandes exécutées. Ce choix est plus judicieux que le modèle proposé par Davison pour les raisons cités en section 2. De plus, l'ordre maximum n est estimé à la volée. Dans un premier temps, le système procède à des prédictions avec des chaînes d'ordre 1. Si les prédictions sont correctes, il passe aux chaînes d'ordre 2 et ainsi de suite. A chaque échec, le système utilise une chaîne d'ordre inférieur. Cependant, plusieurs contraintes assez fortes ont été relaxées pour la construction de ce modèle. La somme des probabilités de transition d'une commande à une autre n'est pas égale à 1 mais est supposée tendre vers 1 à l'infini. Ceci dit, ce système reste le seul à pouvoir être comparé à notre modèle vue son apprentissage en ligne et sa capacité à prédire avec plus d'une mémoire

d'historique.

Notre modèle peut tout à fait être appliqué à cette problématique de prédiction de commandes UNIX sans changement aucun au niveau de la conception. Son avantage par rapport à tous ceux exposés est l'aptitude à réduire de façon drastique le nombre de ses états à intervalles réguliers ce qui lui confère une très grande vitesse lors des prédictions, une bonne fiabilité et un faible coût de stockage.

6 Conclusion

La recherche exposée s'inscrit dans la perspective du Dépôt Légal du WWW Français. Elle présente des modèles de Markov combinés pour l'aide à la navigation obtenus par l'élimination des états et transitions non pertinentes. Les résultats des tests sur 388493 sessions sont pertinents à 62.07%. Des tests complémentaires sont effectués actuellement sur des données plus larges du serveur Web de l'INA [3] (depuis 1999 à aujourd'hui). Dans un autre temps, nous avons présenté deux applications annexes du système à savoir la prédiction des requêtes HTTP et la génération de parcours virtuels.

Notre modèle (CCM) n'a aucunement besoin de stocker toutes les sessions de navigation (apprentissage en ligne), possède un espace réduit d'états et une une bonne pertinence lors des prédictions. Sa complexité est de l'odre de $O(2N_{t,k}-1)$ où $k \in [1,kmax]$ qui dépend du temps, de l'ordre des chaînes utilisées $kmax$ et de la taille de l'espace des états N.

Références

[1] C. R. Anderson, P. Domingos, and D. S. Weld. Adaptive web navigation for wireless devices. *Seventeenth International Joint Conference on Artificial Intelligence*, IJCAI, 2001.

[2] Yonatan Aumann, Oren Etzioni, Ronen Feldman, Mike Perkowitz, and Tomer Shmiel. Predicting event sequences : Data minign for prefetching web-pages. *The Fourth International Conference on Knowledge Discovery and Data Mining*, New York City, 1998.

[3] P. Barford, A. Bestavros, A. Bradley, and M. Corvella. Changes in web client access patterns: Characteristics and caching implications. *World Wide Web, Special Issue on Characterization and Performance*, 1998.

[4] B.D. Davison and H. Hirsh. Experiments in unix command prediction. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 1997.

[5] B.D. Davison and H. Hirsh. Predicting sequences of user actions. *Predicting the Future: AI Approaches to Time-Series Problems*, 1998.

[6] X. Fu, J. Budzik, and K. J. Hammond. Mining navigation history for recommendation. *In Proc. 2000 Conf. on Intelligent User Interfaces*, 2000.

[7] P. Gorniak and D. Poole. Predicting future user actions by observing unmodified applications. *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence*, USA, 2000.

[8] Y. Hafri and C. Djeraba. Détection automatique de profils pour la prédiction de parcours www par chaînes de markov combinées. 5ème *Congrès de la Société Française de Recherche Opérationnelle et d'Aide à la Décision*, Avigon, 2003.

[9] Y. Hafri, C. Djeraba, P. Stanchev, and B. Bachimont. A markovian approach for web user profiling and clustering. *In The Seventh Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD03*, Seoul, 2003.

[10] N. Jacobs and H. Blockeel. Sequence prediction with mixed order markov chains. *Proceedings of BNAIC'02*, 2002.

3. http://www.ina.fr/

[11] T. Joachims, D. Freitag, and T. Mitchell. Webwatcher: a tour guide for the world wide web. *IJCAI*, 1997.

[12] J. Kleinberg. Authoritative sources in a hyperlinked environment. *In Proc.* 9th *ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[13] H. Lieberman. Letizia: An agent that assists web browsing. *In Proc. 14th Intl. Joint Conf. on AI*, 1995.

[14] M. Perkowitz. Adaptive web sites: Cluster mining and conceptual clustering for index page synthesis. *PhD thesis, Dept. of Comp. Sci. and Eng.*, University of Washington, 2001.

[15] M. Perkowitz and O. Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Art. Int. J.*, 118(1-2), 2000.

[16] J. Pitkow and P. Pirolli. Mining longuest repeating subsequence to predict world wide web surfing. *In second USENIX Symposium on Internet Technologies and Systems*, Boulder C0, 1999.

[17] R. Sarukkai. Link prediction and path analysis using markov chains. *Computer Networks*, pages 33(1–6):377–386, 2000.

La notion de distance comme référence pour gérer la pertinence de documents sur le Web

J.Révault

Laboratoire VALORIA,
rue Mainguy, Tohannic
56000 Vanness, FRANCE.
Email : Joel.Revault@univ-ubs.fr
Tél : +33 2 97 01 72 08 Fax : +33 2 97 01 70 71

Résumé

Le but de cet article est de montrer que la notion de distance peut jouer un rôle important pour calculer des indices de pertinence de documents d'un réseau. Nous considérons que les documents possèdent des indices locaux de pertinence vis-à-vis de sujets et que ces indices sont interprétables comme des duals de distances normalisées.

Nous proposons la notion de distance comme référence pour résoudre deux problèmes différents :
- la construction d'un indice synthétique (à partir d'indices élémentaires) pour répondre à des demandes portant sur plusieurs sujets.
- la prise en compte des difficultés d'accès aux documents pour propager et évaluer un indice de pertinence distant.

En raison de la stabilité des opérateurs choisis sur les distances toutes les combinaisons peuvent être enchainées sans sortir du cadre formel de référence.

Mots clés : distance, indice de pertinence, méta-information, multicritère, propagation.

Abstract

The aim of this paper is to prove that the distance model is very important to evaluate pertinence indicators of documents on the net in relation to topics. We assume, first, theses documents have local indicators, second, each indicator is the dual of a normalized distance.

We use the formal distance frame in two case :
- we built a global multicriterion indicator to resolve questions using several topics.
- we introduce a fadding factor to evaluate a propagated remote indicator.

We show that all these combinations are stable in the distance frame. So, we can repeat as we like, the main properties of the distance model are preserved.

Keywords : distance, pertinence indicator, metadata, multicriterion, propagation.

1 Introduction

Le volume des informations et le nombre de documents disponibles sur le Web rendent l'identification de documents pertinents vis à vis d'une question tout à fait difficile.

De nombreux documents contiennent localement des méta-informations comme par exemple des indices de pertinence vis à vis de différents sujets. Ces méta-informations pouvant selon les besoins des utilisateurs être utilisées aussi bien pour l'accès aux documents que pour l'interdiction d'accès à ceux-ci.

Ces informations ne sont cependant pas toujours suffisantes pour identifier les documents d'intérêt et ceci pour plusieurs raisons. Citons, par exemple, le fait qu'elles n'ont aucune raison d'être directement en relation avec les demandes des utilisateurs (questions multicritères), le fait qu'elles restent enfermées dans les documents et qu'on ne dispose pas "à distance" d'une forme dérivée de ces méta-informations.

Dans ce contexte il est nécessaire de définir différentes manipulations et combinaisons des méta-informations dans le but d'approcher au mieux les notions sous-jacentes dans les demandes des utilisateurs. Ces manipulations et combinaisons se traduisent par des opérations sur des objets qui peuvent être disparates. Or l'utilisation empirique d'opérateurs sur des objets très divers risque fort de fournir des résultats difficilement interprétables, autrement dit il devient difficile d'affirmer, même si on en avait la volonté, que l'on approche au mieux la demande de l'utilisateur.

Nous nous proposons de choisir un concept central pour évaluer l'intérêt des documents relatifs à une demande. Le choix d'un unique concept se fonde sur deux raisons :

- sa présence permet de définir des opérations stables qui peuvent donc *a priori* être répétées autant de fois qu'il convient sans sortir du cadre de ce concept,
- ce concept étant préservé au cours des manipulations, nous conservons *de facto* ses propriétés dans tous les processus d'adaptation et de combinaison des méta-informations.

On peut considérer que le premier soucis est d'abord d'ordre formel, le second ayant plutôt trait à la signification, mais nous pensons que ces deux aspects ne sont pas totalement dissociables.

Il reste à choisir le concept central et à définir des opérateurs auquels on peut attribuer un sens en rapport avec un type de demande.

Le but de cet article est de montrer le rôle central qui peut être joué par la notion de distance et différentes opérations sur cette notion pour évaluer la pertinence de documents vis à vis d'un ou plusieurs sujets, et ceci, aussi bien localement (indice natif) que depuis un site distant (indice propagé).

La section suivante indique les travaux proches qui ont influencé cette étude. Les troisième et quatrième sections traitent des indices locaux et du rôle des distances dans ce contexte en particulier dans les évaluations multicritères. La section suivante aborde la question des indices propagés, du rôle et de la capacité des distances à exprimer des caractéristiques qui ne paraissent pas être de leur ressort. Elle aborde enfin les problèmes issus de la multiplicité des sources d'information. La dernière section fait le bilan de notre actuelle réflexion sur le sujet traité.

2 Travaux proches

De nombreux travaux concernent la gestion des métadonnées [1], [12], [3] et plus spécialement l'évaluation de l'indice de pertinence d'un document vis à vis d'un sujet. Ces évaluations concernent un indice local, qui ne dépend que du document ([13], [6]), ou une estimation de la difficulté d'accès au document comme en [11] et [5]. L'objectif étant de calculer un indice propagé, qui prend en compte à la fois la pertinence intrinsèque du document et la difficulté d'y accéder. Dans ce cas il s'agit de faire une fusion entre deux évaluations de nature différente, cette question de fusion est abordée dans [13], [4], [8], [9], [11]).

Les différentes formes d'évaluation utilisent des distances [5], des probabilités [2] ou des logiques floues [4], c'est à dire des formalismes différents. Leur fusion (comme en [4] ou [11]) pose donc des problèmes de cohérence globale dans la mesure où le résultat composite n'est plus a priori interprétable dans aucun des formalismes dont il est issu. Les exigences énoncées dans l'introduction, c'est-à-dire la référence à un cadre formel unique pour uniformiser l'interprétation du résultat et permettre son exploitation ultérieure, nous ont conduit à utiliser un concept central : la notion de distance.

Nous réutilisons et étendons plus particulièrement les travaux d'E.Spertus ([13]), pour les indices locaux, ainsi que ceux de M.Marchiori ([4]) et J.Révault ([8], [9]), pour les calculs d'indices propagés fusionnant les deux composants évoqués ci-dessus.

3 Hypothèses sur les indices de pertinence natifs

Sur ce point, les hypothèses de cet article concernent la présence d'indices de pertinence dans certains documents (indices locaux simples ou natifs), elles peuvent s'énoncer ainsi :

- pour ces documents du réseau on a une évaluation de leur pertinence vis à vis d'un sujet S, cet indice simple (ou monocritère) d'un objet O vis à vis de S sera noté $v(O, S)$.
- les valeurs de v sont interprétables comme le dual de distances normalisées dans l'espace des documents et des sujets, c'est à dire que $v(O, S) \in [0, 1]$ et $d(O, S) = 1 - v(O, S)$ a les propriétés des distances.

Ainsi, rechercher des documents traitant du sujet S consiste à repérer les documents dont l'indice de pertinence est supérieur à un seuil σ, ou encore en termes de distance à rechercher les documents O qui se trouvent dans la sphère de centre S et de rayon $1 - \sigma$, *i.e.* $d(O, S) < 1 - \sigma$. En posant $\alpha = 1 - \sigma$, ce que nous appliquerons dans toute la suite, cette contrainte devient $d(O, S) < \alpha$.

La connaissance des indices de pertinence locaux simples n'est cependant pas suffisante lorsqu'on s'intéresse aux documents conjointement relatifs à plusieurs sujets. Dans ce cas il s'agit de construire un indice de pertinence multicritère.

4 Indice de pertinence multicritère

Ici le but consiste à fabriquer un indice local pour identifier des documents aussi proches que possible de chacun des sujets cibles. Nous considérons d'abord le cas d'un indice bicritère, par la suite nous étendons les résultats à un nombre quelconque de sujets.

4.1 Indice bicritère

Pour trouver des documents relatifs à deux sujets S_1 et S_2 une première idée consiste à sélectionner, d'abord, les documents O_1 tels que $d(O_1, S_1) < \alpha_1$, puis, parmi ceux-ci les documents O_2 tels que $d(O_2, S_2) < \alpha_2$, ce qui revient à ne retenir que les documents communs.

Conceptuellement il s'agit de construire l'intersection de deux sphères conformément au shéma (a) de la figure 1. Or le choix de l'intersection pose au moins deux problèmes :

- si les niveaux de pertinence demandés sont assez élevés (α_1 et α_2 proches de 1) l'ensemble des documents risque fort d'être vide,
- au delà, si cet ensemble n'est pas vide, un document très pertinent vis à vis de S_1 mais assez peu vis à vis de S_2 (ou l'inverse) sera très probablement écarté.

Ainsi on ne récupère que des documents médiocrement satisfaisants pour les deux sujets, et intuitivement on fige préalablement le dosage du compromis qui n'est pourtant pas nécessairement bien clair à ce stade. Dans ces conditions le choix de l'intersection n'est pas totalement satisfaisant.

4.2 Une ellipse pour un compromis moins figé

Afin d'éviter ce dernier écueil E.Spertus [13] propose de conserver l'ellipse de l'espace des documents dont la somme des distances est inférieure à un seuil α (voir figure 1 schéma (b)).

D'un point de vue conceptuel ce choix correspond à une extension de la notion de distance antérieure en $d(O, S_1 * S_2)$ où $S_1 * S_2$ représente la conjonction des critères S_1 et S_2, et où $d(O, S_1 * S_2)$ représente la distance de l'objet O vis-à-vis de la conjonction des sujets S_1 et S_2.

Pour conserver les propriétés de la normalisation on prendra

$$d(O, S_1 * S_2) = \frac{d(O, S_1) + d(O, S2)}{2}$$

ce qui correspond au choix suivant pour l'indice de pertinence
$v(O, S_1 * S_2) = 1 - d(O, S_1 * S_2) = 1 - (d(O, S_1) + d(O, S2))/2$ c'est à dire $v(O, S_1 * S_2) = (v(O, S_1) + v(O, S2))/2$.

Ainsi l'ellipse des documents retenus correspond à $d(O, S_1 * S_2) < \alpha$, ou en termes d'indice de pertinence $v(O, S_1 * S_2) \geq 1 - \alpha$. Cette configuration traduit l'assouplissement de la condition d'appartenance à une zone de pertinence en substituant une condition globale à deux conditions séparées.

4.3 La prise en compte de l'importance de chaque sujet

La proposition d'E.Spertus [13] traite les deux sujets d'intérêt S_1 et S_2 de façon similaire. Or bien souvent on peut établir un ordre de priorité ou de préférence pour des documents plutôt plus pertinents vis-à-vis de l'un des deux sujets.

Dans ce but on peut introduire des pondérations k_1 et k_2 indiquant l'importance relative accordée au traitement de chacun des sujets. En conservant le même principe on obtient

$$v(O, (S_1, k_1) * (S_2, k_2)) = \frac{k_1 \times v(O, S_1) + k_2 \times v(O, S2)}{k_1 + k_2}$$

et v est encore le dual d'une distance normalisée à savoir

$$d(O, (S_1, k_1) * (S_2, k_2)) = \frac{k_1 \times d(O, S_1) + k_2 \times d(O, S2)}{k_1 + k_2}$$

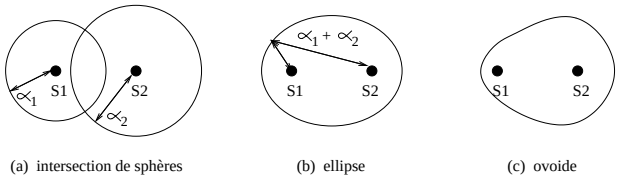


FIG. 1 – les ensembles des documents retenus

L'espace des documents retenus pour un seuil fixé à α sera constitué de l'ensemble vérifiant $d(O,(S_1,k_1)*(S_2,k_2)) < \alpha$

Ce qui peut se représenter par l'appartenance à une zone ovoïde de l'espace des documents comme l'illustre la figure 1 schéma (c).

Cette représentation traduit à la fois le déséquilibre dans l'intérêt porté aux deux sujets et l'assouplissement de la condition d'appartenance à une zone de pertinence.

4.4 Extension au dela de deux critères

Soit $S_1, S_2, ..., S_n$ un ensemble de n sujets ; la détermination de l'indice de pertinence d'un document O vis à vis de cet ensemble de sujets sera exprimable de manière similaire (sous forme normalisée). Nous noterons $v(O,(S_1,k_1)*...*(S_n,k_n))$ sous la forme $v(O, *_{i=1}^{i=n}(k_i,S_i))$, ainsi nous obtenons :

$$v(O, *_{i=1}^{i=n}(k_i,S_i)) = \frac{\sum_{i=1}^{i=n} k_i v(O,S_i)}{\sum_{i=1}^{i=n} k_i}$$

$$v(O, *_{i=1}^{i=n}(k_i,S_i)) = \frac{\sum_{i=1}^{i=n-1} k_i v(O,S_i) + k_n v(O,S_n)}{\sum_{i=1}^{i=n-1} k_i + k_n}$$

Ce qui signifie que le calcul est cumulatif ; concrêtement tout se passe comme si $S_1 * ... * S_{n-1}$ représentait un seul sujet dont la pondération serait $\sum_{i=1}^{i=n-1} k_i$.

Autrement dit l'ajout d'un nouveau critère n'engendre pas de recalcul, au contraire on réutilise complètement et directement les calculs antérieurs. Il est facile d'en déduire que l'indice de pertinence de O vis à vis de $S*T$, pour $S = S_1 * ... * S_n$ et $T = T_1 * ... * T_m$, se calcule encore de la même manière, à savoir

$$v(O,(S,k_s)*(T,k_t)) = (k_s v(O,S) + k_t v(O,T))/(k_s + k_t)$$

où $k_s = \sum_{i=1}^{i=n} k_{Si}$, $k_t = \sum_{j=1}^{j=m} k_{Tj}$, et où, $v(O,S)$ et $v(O,T)$ sont les indices de pertinence préalablement calculés pour les deux familles de sujets S et T. C'est à dire que tout calcul partiel est récupérable dans une évaluation vis à vis d'un surensemble de sujets.

5 La construction d'un indice de pertinence distant

Dans cette section et les suivantes le web est identifié à un ensemble de documents reliés par des hyperliens, cette structure est celle d'un graphe dont on ne retient, *a priori*, que l'organisation logique, ce qui n'empêche pas, *a posteriori*, de prendre en compte des éléments d'organisation physique (par des choix de paramètres adaptés).

La propagation d'un indice de pertinence dans un sous ensemble limité du réseau a été proposée dans [4]. La méthode permet d'accéder à distance (depuis un document O') à l'indice de pertinence d'un document O (accessible depuis O') vis à vis d'un sujet S. L'indice résultant est calculé en faisant subir un affaiblissement à l'indice natif, cet affaiblissement représente en quelque sorte la difficulté d'accès à O depuis O'.

La technique de calcul proposée comporte tout de même deux lacunes : on ne connait pas le cadre formel sous-jacent (on peut cependant l'interpréter en termes de distance comme le montre [9]), et surtout, l'interprétation de l'indice affaibli conduit dans certains cas à des évaluations contraires à l'intuition.

Dans le prolongement, la proposition d'un cadre formel stable est fournie par [9] : il s'agit du cadre des distances orientées et normalisées muni de la h-opération, à savoir $h(d_1, d_2) = d_1 + d_2 - d_1 d_2$. On y fournit la preuve que cette opération correspond à la multiplication pour les indices duals de d_1 et d_2. Appliqué au calcul d'un indice de pertinence distant on peut affirmer qu'aussitôt que $v(O, S)$ (indice natif) et $f(O', O)$ (facteur d'affaiblissement) sont des duals de distances alors $c(O', S) = f(O', O) \times v(O, S)$ est aussi le dual d'une distance (orientée et normalisée).

Le problème du cadre formel étant résolu de façon convenable il reste à trouver un facteur d'affaiblissement f qui traduise correctement la difficulté d'accès à O' dans ses différents aspects.

5.1 L'élaboration d'un facteur d'affaiblissement

Tout en conservant le contrôle des limites de la propagation défini par [4], un premier pas est franchi dans [9] en proposant un facteur d'affaiblissement $f = r^{\delta}$, dans lequel δ représente l'évaluation du plus court chemin de O à O' et r une constante à déterminer dans l'intervalle $[0, 1]$.

Ce choix traduit la distance physique dans le réseau, en nombre de "clics" si on l'applique naïvement. Il permet aussi éventuellement de prendre en compte des coût différents pour les "clics" par exemple selon qu'ils provoquent ou non un changement de site (ces catégories de coût des liens sont aussi retenues par [13]). Une telle évaluation de f donne donc une représentation d'un des aspects de la difficulté d'accès à l'information dans le réseau. Elle ne fournit plus des résultats contre-intuitifs, cependant elle n'exprime pas tous les aspects de la difficulté d'accès. En effet elle ne prend pas en compte les multiples possibilités de navigation en présence de plusieurs liens et donc les risques d'égarement dans le réseau en raison de cette liberté.

Un deuxième pas est proposé par [8], il porte sur la prise en compte du risque d'égarement pendant la navigation. Celle-ci est facilement exprimable à l'aide de probabilités : dans la suite $q(O', O)$ représente la probabilité d'atteindre O lorsque l'on a déjà atteint O'. Pour être plus précis notons que l'on peut distinguer deux cas suivant la qualité des méta-informations disponibles ou la volonté de les utiliser plus ou moins complètement : le cas de l'équiprobabilité si on ne prend pas en compte la signification des étiquettes des liens et de leur environnement, le cas de la non équiprobabilité si on prend en compte la signification des étiquetttes des liens, de leur environnement et plus généralement de leur relation au sujet d'intérêt. La figure 2 donne un exemple de valuation pour q.

La difficulté provient ici du fait qu'*a priori* on passe du cadre formel des distances à celui des probabilités, ce qui limite la fiabilité d'une combinaison entre le premier facteur issu de la distance physique dans le réseau et le deuxième facteur probabiliste. Au dela d'un défaut, qui serait de pure forme, nous pensons que ce glissement fait peser une lourde hypothèque sur la

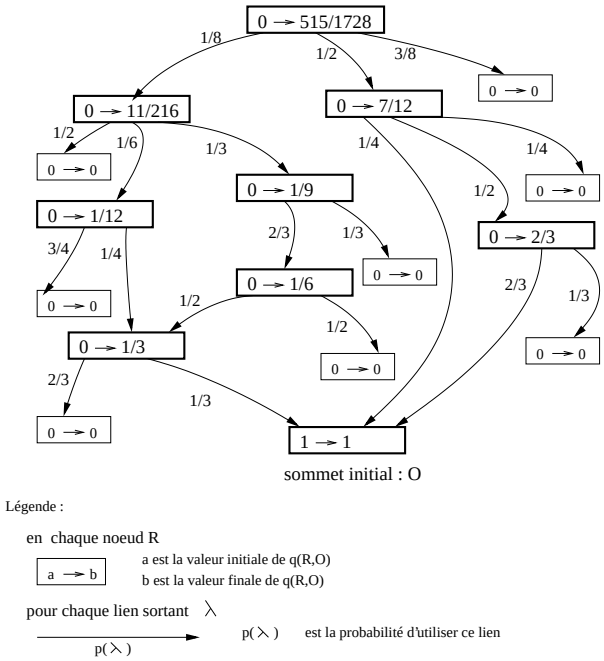


FIG. 2 – Exemple de valeurs pour $q(R, O)$

signification des résultats. Une solution figure dans [8], elle définit, en adaptant [14] à ce contexte, une méthode de fusion des indices d'origine métrique et probabiliste dans le cadre des distances. Techniquement on prend $f = r^{\delta} \times s^{log_a(q)}$ où r, s et a peuvent être fixés arbitrairement dans $[0, 1]$. Notons enfin qu'accessoirement ceci permet de choisir, quand on le souhaite, des formes particulièrement simples à travers des choix particuliers de r, s et a.

La figure 3 donne deux exemples simples de valeurs de f dans le cas particulier où il y a équiprobabilité et où $a = s$.

Finalement l'ensemble des résultats de [4], [9] et [8] permet d'évaluer un indice distant, en O', dérivé de la pertinence de O sur un ensemble de sujets S, prenant en compte divers aspects de la difficulté d'accès, et ceci sans sortir du cadre formel des distances.

A ce stade il reste une limitation importante, il faut que O soit la seule source d'information sur S ayant propagé son indice jusqu'à O'. Il s'agit comme indiqué en début de section 5 de

$$c(O', S) = f(O', O) \times v(O, S)$$

Mais l'hypothèse d'unicité de cette source n'est pas acceptable puisque les informations sont à priori réparties dans l'ensemble du réseau. Nous abordons donc le problème de la répartition des sources dans la sous-section suivante.

5.2 Un indice composite pour des informations réparties

L'incidence de la répartition des documents en divers noeuds du réseau conduit à l'examen de deux problèmes dans le calcul des indices distants.

Le premier est celui des pluri-valuations de cet indice pour un même sujet ou un ensemble de sujets S, cette distinction n'ayant plus lieu d'être ici. Il s'agit du cas d'un document O' recevant

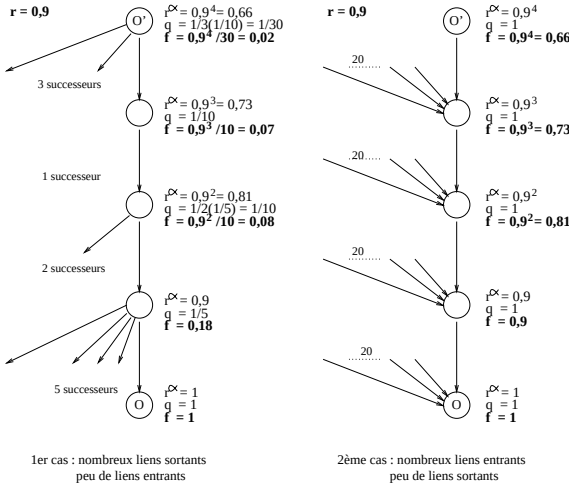


FIG. 3 – Exemple simple de facteur d'affaiblissement

par propagation deux (ou plusieurs) valeurs de pertinence pour un même sujet S, chaque valeur étant issue d'un document différent. La détermination d'un indice résultat a été étudiée dans [4] et [10], dans les deux cas le choix s'est porté sur le maximum des indices reçus (*i.e.* le minimum des distances) : pour [4] il est conforme au cadre des logiques floues, pour [10] il est stable dans le cadre des distances. Le choix du maximum ne s'impose pas de façon impérative, en particulier parce qu'il n'exprime pas tous les aspects que l'on pourrait prendre en compte dans les cas de pluri-valuation, il a cependant au moins deux qualités : il préserve le cadre formel des distances et on peut l'interpréter comme celui du compromis le plus efficace.

Le second problème est celui de la construction d'un indice multicritère distant en O' relatif à $S = S_1 * ... * S_n$ quand ces sujets sont traités soit conjointement soit exclusivement dans plusieurs documents d'une famille $O_1, ..., O_p$.

Chaque document O_j aura propagé un indice de pertinence pour la partie S_{Oj} de S qu'il traite et on sait que $S = \cup_{j=1}^{j=p} S_{Oj}$. Deux cas peuvent alors se présenter : dans le cas le plus simple les S_{Oj} sont deux à deux disjoints, dans le cas contraire ils se recouvrent plus ou moins et nous montrons qu'il est possible de revenir au cas précédent.

Lorsque les sous-ensembles S_{Oj} sont disjoints on peut réutiliser la technique de combinaison vue en section 4 pour les indices multicritères locaux, elle s'adapte directement au cas des indices distants.

Lorsque les sous-ensembles se recouvrent partiellement on peut construire une partition maximale d'ensembles disjoints à partir de la famillle $S_{O1}, ..., S_{Op}$, puis raisonner en deux temps : appliquer la méthode retenue pour les multi-valuations aux sous-ensembles qui étaient l'objet d'un recouvrement, appliquer ensuite la technique de combinaison retenue dans le cas précédent à tous les sous-ensembles de la partition maximale puisqu'ils sont disjoints et que chacun n'est plus soumis maintenant qu'à une seule valuation de l'indice.

6 Conclusion

Dans cet article nous avons illustré le principe qui consiste à s'appuyer sur un concept de référence unique. Nous avons montré que ce choix pouvait être judicieux pour maitriser les aspects formels et signification de méta-informations inférées lors de différentes manipulations qui s'avèrent nécessaires pour fournir des réponses acceptables aux utilisateurs des grands réseaux de documents comme le web.

Nous nous sommes plus particulièrement intéressé à l'évaluation des indices de pertinence des documents vis à vis d'un ensemble de sujets. Dans ce contexte nous avons choisi la distance comme notion pivot. A ce propos nous avons montré qu'elle permet :

- des évaluations locales souples pour des requêtes multi-critères,
- des évaluations distantes intégrant un facteur d'affaiblissement qui ne semblait pas initialement être de son ressort.
- des fusions d'indices lorsque l'information recherchée se trouve répartie dans plusieurs documents.

Cette étude appelle immédiatement plusieurs précisions complémentaires :

- la notion de distance ne fournit pas le seul cadre formel envisageable,
- la création de passerelles entre des cadres formels différents enrichi les possibilités d'expression de chacun,
- l'acceptation d'un cadre formel n'interdit pas la diversité des opérations sur les méta-données et il permet d'en définir plus clairement la signification.

Ajoutons encore que choisir une notion pivot n'interdit pas d'en sortir à condition de le savoir. A ce sujet il est clair que, d'une part, tant qu'on reste à l'intérieur on préserve les propriétés inhérentes à cette notion et on bénéficie de toutes les libertés permises par les opérateurs (selon les cas : ordonnancement, regroupement, factorisation, ...) ; d'autre part, une fois sorti il est prudent d'en tenir compte aussi bien dans la signification des résultats (éventuellement ambigüs) que dans leurs futures possibilités de manipulations qui peuvent être réduites voire interdites.

Une difficulté supplémentaire concerne le maintient et la cohérence de l'ensemble des méta-informations inférées, nous n'avons pas encore étudié la faisabilité des solutions et leur complexité. Il nous semble cependant que si l'étude d'une solution complète peut apporter des idées intéressantes, elle sera probablement trop lourde à mettre en œuvre, c'est pourquoi sur ce point nous orientons nos recherches vers des solutions approchées.

Références

[1] Patrice Bellot and Marc El-Beze Classification Automatique et Recherche d'information (expérience dans le cadre de TREC) *Technical Report in LIA - 1998*

[2] L.Denoyer, H.Zaragoza and P.Gallinari HMM-based Passage Models for Document Classification and Ranking *23rd European Colloquium on information Retrieval Research - mars 2001*

[3] Tim Krauskopf, Jim Miller, Paul Resnick and Win Treese PICS Label Distribution Label Syntax and Communication Protocols Version 1.1, http ://www.w3.org/PICS/labels.html, W3C - 1997

[4] M.Marchiori The Limits of Web metadata, and beyond *http ://dec-web.ethz.ch/WWW7/1896/com1896.htm, 2001*

[5] Y.Matsuo, Y.Ohsawa and M.Ishizuka Average-clicks : A New Measure of Distance on the World Wide Web *Journal of intelligent information systems, vol. 20 n°1, p. 51 - janvier 2003*

[6] Jim Miller, Paul Resnick and David Singer Rating Service and Rating Systems (and their Machine Readable Descriptions) *Version 1.1, http ://www.w3.org/PICS/services.html, W3C - 1997*

[7] Paul Resnick and Jim Miller PICS : Internet Access Controls without Censorship *Communication of the ACM , http ://www.w3.org/WWW/PICS/iacwcv2.htm - 1996*

[8] Joël Révault Fusion de distances et de probabilités pour évaluer un facteur d'affaiblisserment pendant la propagation d'un indice de pertinence *Journées Web Semantique - octobre 2002*

[9] Joël Révault Propagation of Pertinence Indicator using Distance Models *IPMU2002 - juillet 2002*

[10] Joël Révault Quelques réflexions sur la propagation d'indices à travers un réseau *Internal report in VALORIA - décembre 2001*

[11] Chris Ridings and Mike Shishiging Pagerank Uncovered *http ://www.website-promotion-ranking-services.com/PageRank.pdf - septembre 2002*

[12] François Role Panorama des travaux en cours dans le domaine des metadonnées *INRIA report N° 3268 - février 1999*

[13] Ellen Spertus ParaSite : Mining Structural Information on the Web *http ://www.scope.gmd.de/info/www6/technical/paper206/paper206.htm*

[14] M.Szijártò, D.Gröger and G.Kallós A qualitative Model for Conditions in Safety-Critical systems - *1998*

Génération d'un portail multilingue sur la thématique cinéma

N. STIENNE ET N. LUCAS

GREYC - UMR 6072
campus II - BP 5186
Université de Caen
14000 Caen, FRANCE
Mail :Nadine.Lucas@info.unicaen.fr
Tél : 02 31 56 73 36 Fax : 02 31 56 73 30

Résumé

Nous présentons un système de constitution de portail sur le cinéma et les sorties de films à partir de sites spécialisés multilingues d'internet. L'expérience consiste à extraire, analyser et ré-exploiter des données et du vocabulaire acquis sur des sites spécialisés. Le système n'a aucune ressource au départ et doit donc les construire. La fouille de sites et l'analyse des pages nécessite une technique robuste. Les données extraites alimentent une base de données sur le cinéma. Le robot devineur alimente une base de données en recoupant les informations issues de plusieurs sites. L'interface permet l'interrogation dans plusieurs langues européennes, et la consultation des critiques dans la langue de l'utilisateur. L'ajout d'une langue se fait par ajout d'une collection de liens sur des sites spécialisés.

Mots-clés : recherche d'information ciblée, documents multilingues, fouille robuste, thématique spécialisée, dépouillement de sites internet, acquisition de ressources, fouille de textes, cinéma.

Abstract

The paper introduces an experiment based on the robust multi-view approach in web mining. The task is to create a portal on movies via active learning from European websites specialised in cinema. Is is performed through robots that get information from websites and web pages, then extract and analyse the data with a guesser and feed a database, last cross information obtained from various sites. The portal allows to consult pages on films in the user's language, or to query by menu or in natural language. The system learns to adapt to a new language by visiting specialised sites from a new set of urls.

Key-words : robust learning, cross language information retrieval, topic oriented information retrieval, focused crawling, web mining, semi-supervised information analysis.

1 Introduction

Le présent travail a débuté lors d'un projet de DESS à l'Université de Caen. Il se situe à mi-chemin entre apprentissage et déduction contextuelle, empruntant tout à la fois au paradigme de la fouille ciblée (*focused crawling)*, aux techniques de fouille de site [Kushmerick 1999] et aux techniques robustes de TAL, appliquée en recherche d'information multilingue (*cross language information retrieval*) [Grefenstette 1998 ; Collins et Singer, 1999].
Mais l'approche adoptée n'utilise pas de lexique même en amorce, elle est semi-supervisée, elle correspond à l'apprentissage robuste préconisé par Muslea [Muslea *et al*., 2000]. Le système permet la création et la maintenance semi-automatique d'un portail appelé site cinéphile européen. Il assure l'extraction de données à partir de sites sur le cinéma et permet de suivre les sorties de films en Europe, de consulter les critiques (multilingues) ainsi que d'interroger le système sur des films à l'affiche ou récents dans la langue de l'utilisateur. Cela suppose quatre étapes distinctes : la fouille des sites, l'analyse des données par recoupements, l'alimentation de la base de données, la ré-utilisation des informations contextuelles pour répondre aux questions des utilisateurs en langage naturel. Soulignons d'emblée que le cinéma est un prétexte. Notre but n'est pas de constituer une base de données exhaustive, ni une ontologie parfaite pour le cinéma, mais plutôt d'étudier à quelles conditions un système de création de portail peut donner des informations *suffisantes* sur un thème donné en gérant l'aspect multilingue et surtout quels sont les modules ré-utilisables pour une autre thématique, suivant en cela [Muslea *et al*., 2000, 2002 a , b]. En effet, il existe un besoin de collecte et synthèse d'information par des moyens robustes et assez légers, pour des centres d'intérêt qui peuvent être temporaires ou très diversifiés. Il serait ainsi possible de développer des outils de fouille très sélective et multilingue.

2 Le système

2. 1 La fouille et le dépouillement des sites

Le système utilise un robot explorant régulièrement les sites internet spécialisés dans le cinéma pour constituer une base de données thématique sur les films, les réalisateurs, les acteurs, dates et pays de sortie, et donner accès aux critiques. Pour la première étape, le robot cueilleur de données utilise une liste de sites spécialisée fournie en amorce. La visite de sites est hebdomadaire.
La seconde étape est le dépouillement des sites, pour localiser correctement l'information recherchée parmi les publicités, interviews et autres informations [Kushmerick, 1999]. Les sites sont de facture très variable et il faut donc faire une analyse automatique de la structure HTML du site. Il est clair que ce problème fait partie des étapes obligées, et que le dépouillement des sites est un des modules ré-utilisables ou remplaçables dès qu'il s'en trouve un meilleur [Cohen & Fan, 1999 ; Kushmerick, 2000 ; Sakamoto, 2002].
Le robot cueilleur localise l'information dans des pages contenant des critiques ou synopsis de films. Nous ne nous attarderons pas sur cette étape. La stratégie utilisée est l'exploitation de la structure du site et de la structure de la page-fille considérée comme pertinente [Hersovici *et al*., 1998]. La structure de la page est étudiée à travers la comparaison des tableaux. L'heuristique adoptée ne permet pas d'analyser tous les sites[1], mais suffisamment de sites par pays pour la poursuite du processus. Les informations fournies par la page de une, appelée page d'index ou index sont conservées pour l'étape suivante. Celle-ci correspond à l'indexation des données, à l'aide d'un robot devineur (*guesser*).

[1] Environ 36% des sites ne sont pas exploités.

2. 2 Le robot devineur

Les informations pertinentes sont analysées automatiquement, mais sans dictionnaires, de façon à rapatrier des données multilingues et les distribuer dans les champs de la base de données. Contrairement à la démarche commune [Oard 1997 ; Grefenstette 1998 ; Gey *et al.* 2001], nous n'avons pas établi de ressources externes en filtrant très tôt les termes à l'aide d'une liste, même restreinte, car nous avons choisi de ne pas limiter a priori la couverture linguistique. En revanche, nous sommes en contexte spécialisé. Nous pouvons donc renverser la problématique. Il est nécessaire de stipuler ce que le logiciel doit deviner, donc de prévoir l'affectation des champs de la base de données.
L'algorithme général s'appuie sur la récupération du titre, à partir des textes de liens. On constitue alors une entité film. Les autres informations sont ajoutées le cas échéant et recoupées.

2.2.1 La fiche signalétique

Nous avons choisi de retenir comme informations pertinentes le titre du film, le nom du réalisateur ou des réalisateurs, le nom des acteurs, le pays d'origine du site, la date de récupération de l'information. Quelques champs sont prévus dans la base de données, mais ils ne sont pas exploités actuellement : le genre du film, la date de sortie du film et le nom des personnages (rôles) du film. Chaque champ de la base contient donc une entité. Les règles du devineur exploitent la redondance de l'information relayée par les sites, en filtrant les informations stables comme les noms propres et les séquences stables, comme la série des noms d'acteurs. Elles permettent l'identification par recoupement des entités recherchées auxquelles sont affectées les valeurs "titre de film", "réalisateur", "acteur", etc., ce qui constitue la fiche signalétique du film considéré. Le titre du film est l'entité la mieux renseignée, par l'utilisation de la redondance des liens entre la page d'index et les pages filles et par une exploitation du contexte immédiat des liens [Hersovici *et al.* 1998].
Notons que le titre est déduit comme tel, en raison du rôle particulier qu'il occupe, comme texte de lien et instance de texte (nous savons *où* le trouver) mais que sa forme est très variable, puisque nos sources sont multilingues. C'est même l'entité la moins stable lexicalement de la fiche signalétique. Nous ne cherchons donc pas une constance de forme, mais une constance dans la *structure des liens* et la répétition de la chaîne de caractères constituant le titre à divers endroits du site. Ceci nous permet d'établir une classe d'équivalence pour les diverses formes prises par l'entité titre. Ainsi *La chute du faucon noir* est une instance de titre dans un site français et *Black hawk down* dans un site anglais. L'inconvénient est que si le texte de lien ne contient pas le titre, on crée une entité film avec un faux titre (souvent un titre de rubrique comme "Fiche Film" ou "This week notice").

2.2.2 Détection des motifs

Comme indiqué ci-dessus, nous avons besoin d'associer une valeur aux chaînes de caractères extraites. Les règles du devineur se basent sur la reconnaissance de motifs et structures, essentiellement typographiques et dispositionnels, détectés par des expressions régulières. Il y a deux motifs pouvant fournir des noms d'acteurs et de réalisateurs. Le motif A définit une liste avec une mise en forme éventuellement contrastée pour la tête (la dénomination de la liste), un séparateur deux points et un corps de liste (les items).

Type A:
Réalisateur: Steven Spielberg
Acteurs: Leonardo DiCaprio, Tom Hanks, Christopher Walken

Un autre motif, de type B permet d'établir des séquences fréquentes dans le texte.

Type B:

Arrête-moi si tu peux, réalisé par Steven Spielberg, avec Leonardo DiCaprio, Tom Hanks, Christopher Walken

La détection du motif B s'appuie sur la co-occurrence d'un contexte immédiat du titre (motif T, avec mise en forme contrastée) contenant une chaîne de caractères courte de type P devant une séquence de type C. Celle-ci est une sous-structure "Prénom Nom", séquence capitalisée double. La structure T suivi de P suivi de C permet l'apprentissage des prépositions ou introducteurs introduisant des noms propres.

2.2.3 Affectation de valeurs

Les données doivent recevoir une valeur pour servir à l'affectation des champs. Pour cela, et dans la mesure où nous ne souhaitions pas exploiter le lexique, nous avons recours à une heuristique. Examinons les règles du devineur : s'il existe des listes A, (instance du motif A détecté par expression régulière, soit une liste) alors la liste la plus longue est la liste des acteurs et la plus courte celle du réalisateur. Par défaut, la première liste est celle du réalisateur. Elle peut être la seule. Ainsi, le film *le peuple migrateur* n'a pas d'acteurs.
S'il n'existe pas de liste (motif A), alors on cherche une instance du motif B dans le texte.

2.2.4 Indexation

Une fois la fiche signalétique construite, le système acquiert d'autres informations. Le système "apprend" par induction un vocabulaire technique spécialisé.
Pour établir une fiche signalétique par film, nous avons aussi besoin d'établir une classe d'équivalence entre *La chute du faucon noir* et *Black hawk down*, par exemple, pour pouvoir considérer que ces deux titres sont en fait des variantes de titre pour une même entité film. C'est le couple d'entités titre de film et réalisateur qui permet d'établir cette relation. Le couple d'entités titre de film et réalisateur permet ensuite de consolider les informations permettant d'enrichir la fiche signalétique. Tout nouveau film est comparé aux informations déjà présentes dans la base, et il n'y a qu'une fiche par film, en dépit de la grande variété de forme des titres de film en environnement multilingue, comme le montre la figure 1.
Nous procédons à la reconnaissance de diverses macro-structures. Techniquement, la variable reliée à l'entité "réalisateur", dans l'exemple suivant, est la chaîne de caractères Steven Spielberg. Elle est co-occurrente avec une chaîne de caractères en tête de liste, qui correspond donc au contexte gauche du séparateur deux points. Une position intéressante est détectée. En parcourant divers sites on établit par déduction une classe d'équivalence entre Réalisateur, Director, etc. fondée sur l'idée que la partie gauche de la liste est variable, et la partie droite est constante.

Réalisateur: Steven Spielberg comparé à Director: Steven Spielberg

Ensuite la comparaison est faite sur le couple entité titre de film et entité réalisateur

Arrête-moi si tu peux / Steven Spielberg comparé à Catch me if you can / Steven Spielberg

L'établissement d'un classement des expressions récupérées pour l'ensemble des films d'un même site permet d'écarter du bruit, dû à une détection incomplète des acteurs par exemple. En effet, le schéma retenu pour leur détection étant la sous-structure "Prénom Nom", des noms de personnes tels qu'Anémone ou Lio induisent des erreurs.
Les erreurs rencontrées sont imputables aux présupposés du devineur. Il peut y avoir confusion entre le nom du réalisateur et des éléments du titre, lorsqu'ils sont groupés dans l'url de l'index, ou encore entre le nom du réalisateur et le nom du compositeur de la musique du film, ou encore le nom de l'auteur d'une œuvre littéraire adaptée. Par exemple, nous avons trouvé The Truth en tant que réalisateur. En effet, le contexte étant celui d'un titre reconnu, la position attendue pour le nom du réalisateur (après le titre) est examinée, or le motif

capitalisation est détecté (cette chaîne ressemble à une sous-structure "prénom nom"), l'affectation est faite. Nous avons aussi trouvé Dumas en tant que réalisateur des *Trois mousquetaires*.

Figure 1 Synthèse de l'entité film

Comme l'apprentissage est incrémental, il est clair qu'une fiche devient rapidement complète et correcte pour un film commenté dans un grand nombre d'articles, et de plus dans plusieurs sites. Le recoupement des contextes des mots permettra par exemple d'éliminer The Truth comme réalisateur d'un film, puisqu'on trouvera *The Truth by …* et non comme dans le premier contexte exploré, une simple co-occurrence.

2.4. Interrogation

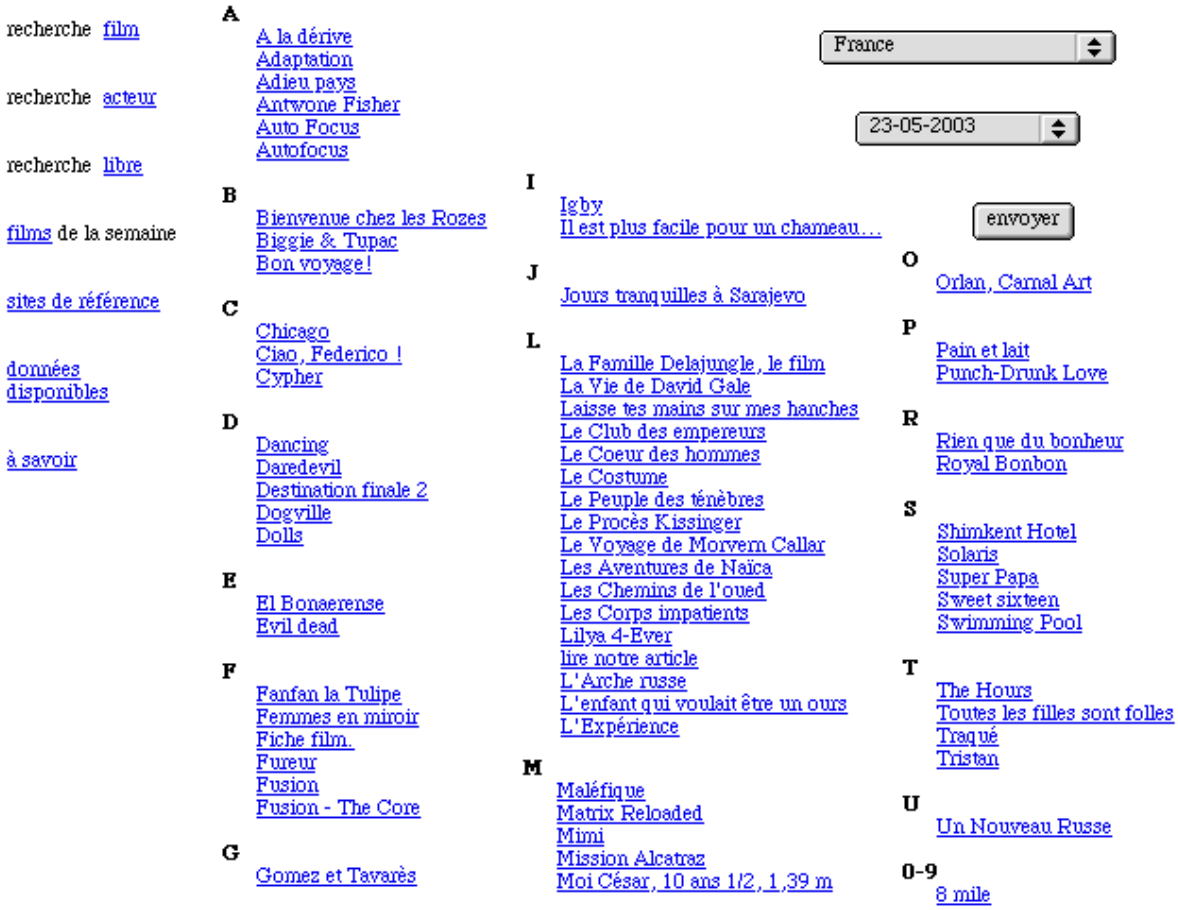
Le résultat est l'affichage des films de la semaine, avec possibilité de consulter les fiches signalétiques et les critiques parues dans chacune des langues rencontrées. Dans l'interface, un drapeau symbolise la langue de communication entre le système et l'utilisateur. Le système indique si le film est également à l'affiche dans un autre pays et affiche les traductions du titre dans une ou plusieurs autres langues, le cas échéant comme indiqué par la figure 1. D'autres modes d'interrogation sont possibles, par exemple sur les films de la semaine (Figure 2).

Le système permet également l'interrogation rétroactive sur les films, les réalisateurs et les acteurs. Cette recherche n'est naturellement possible que sur la période de temps couverte par le moteur du portail.

La constitution du catalogue de films référencés permet l'interrogation à partir de plusieurs points d'entrée par menu, et prévoit aussi une interrogation en langage naturel. Celle-ci renvoie des résultats corrects. L'utilisateur posant une question en espagnol verra s'afficher les

critiques en espagnol, celui qui pose les questions en néerlandais verra les réponses en néerlandais.
L'ajout de nouveaux sites pour un nouveau pays entraîne la possibilité d'interrogation dans une nouvelle langue, car le système réinjecte dans le module d'interrogation les termes co-occurrents avec les entités déduites. Ainsi, si on ajoute des sites norvégiens, le système "apprend" que Steven Spielberg est introduit par "regi" par observation de la co-occurrence dans les listes; de même dans les textes, cette variable est co-occurrente avec "med". Il est ainsi possible de faire des requêtes en langage naturel, car le système va associer par exemple "de" ou "by" ou "von" à l'entité nom du réalisateur et "avec", "with" ou "mit" aux entités reconnues comme noms d'acteurs.

Figure 2. Interface française interrogation sur la date



2.5. Résultats et évaluation

Le système mis en place donne satisfaction pour la récupération et la synthèse des informations. Le catalogue des titres de films a été correctement incrémenté sur la période d'activité du système (février-mars 2002 puis depuis novembre 2002). Il y a peu de bruit mais on relève du silence. Pour évaluer la pertinence des résultats nous avons procédé à une vérification manuelle sur les enregistrements du 23 mai. Les calculs sont effectués de la manière suivante : l'ensemble de référence P est l'ensemble des films sortis, recensés manuellement, et l'ensemble B les enregistrements de la base. On calcule pour chaque rubrique les entrées manquantes, (soit p-b dans chaque colonne), les entrées sur-générées (doublons non éliminés, soit b-p) et les erreurs. Le rappel est établi par le rapport entre les

films retrouvés corrects et le nombre de films effectif. La précision est établie par le rapport entre nombre d'entrées correctes sur le nombre d'entrées enregistrées. Les taux habituels de silence (1- rappel) et de bruit (1- précision) sont donnés à titre indicatif pour cette semaine.
La totalité des films sortis, recensés manuellement, n'a pas été enregistrée dans la base, mais la couverture est satisfaisante (Tableau 1). Le rapport entre l'investissement en ressources et en règles d'une part et le nombre d'informations effectivement collectées et analysées d'autre part est très intéressant. Cela montre l'intérêt de l'approche active.
Concernant la qualité ou la précision des résultats, elle nécessite une supervision humaine, car on trouve des erreurs. Pour les titres de film, il s'agit de silence, 189 entrées dans la base pour 207 films effectivement sortis dans la semaine. Il y a aussi du bruit : des entités films sont créées en surnombre. Cependant, nous avons constaté des fluctuations sur la durée. Dans les périodes où des films à très large distribution sortent, il y a très peu d'erreurs ; en revanche dans le cas de rétrospectives ou de films à audience nationale, des erreurs se glissent dans le catalogue construit, faute de recoupements suffisants. On pourra noter un silence plus important sur les noms d'acteurs mais aussi une meilleure précision. Les erreurs concernent surtout l'attribution des statuts réalisateur ou acteurs. Sachant qu'il est possible qu'un acteur soit aussi réalisateur, l'efficacité du système est jugée bonne. Par exemple, le système répond correctement aux requêtes sur Clint Eastwood sous ces deux casquettes.
Les disfonctionnements locaux à un site se produisent à tous les niveaux : dans le rapatriement des données quand les textes de critique ne sont pas ou plus téléchargeables, dans le choix des liens pour détecter le titre, dans l'attribution des champs. Généralement ils résultent en entités sur-générées mais aussi en silences. Le cas le pire se produit lorsqu'un un faux titre comme *Fiche Film* est détecté plusieurs fois, il est rentré une seule fois dans la base (les doublons sont fusionnés), cela crée du silence localement, on perd les titres des films signalés sous cette rubrique.

Tableau 1 Evaluation de la pertinence de la base

	Titres de film	Réalisateur	Acteurs
Ensemble de référence P	207	130	403
Total enregistré B	189	154	306
Correctement détecté	165	110	289
Sur-généré	24	35	17
Non détecté	66	20	140
Autres erreurs (inversions…)		9	9
Rappel	0,8	0,85	0,72
Précision	0,87	0,71	0,94
Silence	0,20	0,15	0,28
Bruit	0,12	0,29	0,06

Il semble que les objectifs de départ soient à dissocier. L'objectif de synthèse est bien atteint, avec une excellente couverture, mais celui de thésaurisation de l'information pour une consultation ultérieure souffre d'une trop grande approximation dans l'interprétation des informations recueillies, notamment pour les films à distribution restreinte. On envisage de faire un test sur le premier enregistrement d'une entité film dans la base pour n'autoriser les incrémentations d'information que si les rubriques sont renseignées et restent cohérentes. Une autre stratégie consiste à écarter les titres de rubrique par comparaison des sites d'une semaine sur l'autre.
La fonction multilingue est particulièrement agréable, car elle est totalement transparente à l'utilisateur. L'ajout de sites a été testé pour le norvégien, et a permis de vérifier qu'il suffit effectivement d'ajouter de nouvelles urls.

3. Discussion

Ce travail illustre une voie d'approche jusqu'ici utilisée avec succès dans la fouille de documents, mais qui n'a à notre connaissance pas été étendue à la fouille multilingue en dehors du cas particulier de la détection de citations. Les techniques de fouille ciblée de la toile permettent en effet d'ajouter de nouvelles pages pertinentes pour une thématique particulière à partir d'une collection de liens donnée en amorce [De Bra *et al*. 1994 ; Cho *et al.* 1998 ; Chakrabarti *et al.*1999 ; McCallum 1999 ; Diligenti *et al.*, 2000]. De nombreuses techniques permettent de raffiner et d'étendre le principe de classes d'équivalence pour des url utilisées comme mots-clés avec ou sans contexte, avec ou sans pondération des pages. L'exploration d'un site obéit à des règles similaires [Mukerjea 1998]. Mais dans la mesure où la similarité des pages est définie par co-occurrence des mots de la langue, à partir du document ou des pages servant d'étalon de comparaison, la fouille thématique est *de facto* ordonnée et limitée par la langue. Même si le travail de constitution de lexique n'est pas très lourd dans l'application visée, il est fastidieux à la longue.

Dans les approches orientées par le genre ou la situation, des données du code graphique quasi-indépendantes de la langue (parenthèses, capitalisation, dates et chiffres) ont été utilisées avec le succès que l'on sait pour la détection des citations du genre académique et la constitution automatique de bases de données bibliographiques [Giles *et al.* 1998], la plus connue et la plus élaborée à ce jour étant CiteSeer (http://citeseer.nj.nec.com).

Nous nous inscrivons à la croisée de ces travaux [Muslea *et al.* 2000]. Au lieu de penser la diversité des langues comme un handicap, nous construisons le jeu d'inductions et de déductions sur une constante, qui est l'uniformité du thème. Un raffinement possible pour ce thème particulier des sorties de film serait de s'appuyer la périodicité régulière du renouvellement de l'information recherchée. Le devineur pourrait également être amélioré, par une exploration du document entier. Mais alors on perd en simplicité.

Sachant que les sites de cinéma offrent des informations qui fournissent elles-mêmes la matière des interrogations possibles, nous produisons automatiquement un lexique de noms propres associé à un lexique de noms de métiers du cinéma, au lieu de l'injecter à partir d'un travail de recensement manuel. Dans le cas où un robot explorateur est lancé, il faut éviter que les nouvelles données recueillies sur de nouveaux sites ne puissent pas être exploitées à cause de la variété des lexiques. Dans notre système, le robot interpréteur travaille immédiatement derrière le robot cueilleur et tout site dépouillé est exploité.

La part de supervision du système consiste à vérifier et mettre à jour la base de liens. La vérification et le toilettage de la base de données formeraient le post-traitement manuel. La généricité de l'approche tient à l'absence de lexique. Pour adapter le système à une autre problématique, par exemple le dépouillement de pages sportives pour alimenter une base de données sur le football, il faut établir les classes d'équivalence pour créer les entités "équipes", "sportifs", "matchs", "scores" etc. et trouver les expressions régulières pour alimenter ces champs.

Cette expérience permet de situer une zone de recherche médiane, entre exploration ciblée de la toile, dépouillement de site et interprétation de données multilingues. Elle montre la possibilité de faire une synthèse multilingue et de rétro-activer l'acquisition du vocabulaire. Le travail préliminaire, qui consiste à cerner automatiquement la zone à analyser dans des sites formatés de façon très variable, est une clé pour la fiabilité du résultat. Si peu de sites fournissent des données recoupables, la base de données est alimentée avec un taux d'erreur imputable au devineur. On peut objecter bien sûr que pour arriver à un résultat identique, la liste des films de la semaine, il est inutile d'élaborer une méthode alternativement inductive et déductive. Il serait plus simple et certainement plus fiable de donner des mots-clés, correspondant au domaine. Cependant, nous cherchons surtout à cerner les heuristiques

s'appliquant à un vaste ensemble de problèmes d'extraction d'information, de façon à pouvoir adapter rapidement un système de fouille et indexation à une nouvelle thématique.

Références bibliographiques

[Chakrabarti *et al*. 1999] "Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery", S. Chakrabarti, M. van den Berg and B. Dom. In *Proceedings of the 8th International WWW Conference*, Toronto, 1999. http://www8.org/w8-papers/5a-search-query/crawling/index.html

[Cho *et al.* 1998] "Efficient Crawling Through URL Ordering", J. Cho, H. Garcia-Molina, L. Page. In *Proceedings of the 7th International WWW Conference*, Brisbane, 1998. http://www7.scu.edu.au/programme/fullpapers/1919/com1919.htm

[Cohen et Fan 1999] "Learning Page-Independent Heuristics for Extracting Data from Web-pages" W. Cohen and W. Fan. *Proceedings of 8th International Conference on World Wide Web* 1999.

[Collins et Singer 1999]. "Unsupervised models for named entity classification" M. Collins and Y. Singer In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999.

[De Bra *et al*. 1994] "Information Retrieval in Distributed Hypertexts", P. De Bra, G. Houben, Y. Kornatzky and R. Post. In *Proceedings of the 4th RIAO Conference*, 481 - 491, New York, 1994. http://citeseer.nj.nec.com/debra94information.html

[Diligenti *et al.* 2000] "Focused Crawling Using Context Graphs", M. Diligenti, F. Coetzee, S. Lawrence, C. Giles and M. Gori. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000*), Cairo, 2000. http://citeseer.nj.nec.com/diligenti00focused.html

[Gey *et al.* 2001] "Entry Vocabulary - A Technology to Enhance Digital Object Search" F. Gey, M. Buckland, A. Chen, and R. Larson. In *Proceedings of the First International Conference on Human Language Technology*, 2001.

[Giles *et al*. 1998] "CiteSeer: an automatic citation indexing system" L. Giles, K. Bollacker et S. Lawrence. In Witten, Ackscyn, Shipman (eds) *Digital libraries 98*, ACM Press, 1998.

[Grefenstette 1998] *Cross-Language Information Retrieval* G. Grefenstette (ed) Kluwer, 1998.

[Hersovici *et al*. 1998] "The Shark-Search Algorithm - An Application: Tailored Web Site Mapping", M. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalhaim and S. Ur. In *Proceedings of the Seventh International World Wide Web Conference*, Brisbane, 1998. http://www7.scu.edu.au/programme/fullpapers/1849/com1849.htm

[Kushmerick 1999]. "Learning to remove internet advertisements", N. Kusmerick. *Proceedings of Autonomous Agents* 1999.

[Kushmerick 2000] "Wrapper induction: Efficiency and Expressiveness", *Artificial Intelligence* 118, 2000.

[McCallum *et al.* 1999] "Building Domain-Specific Search Engines with Machine Learning Techniques", A. McCallum, K. Nigam, J. Rennie, and K. Seymore. In *1999 AAAI Spring*

Symposium on Intelligent Agents in Cyberspace, Stanford University, 1999. http://www.ri.cmu.edu/pubs/pub_2716.html

[Mukherjea 2000] "WTMS: A System for Collecting and Analysing Topic-Specific Web Information", S. Mukherjea. In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, 2000. http://www9.org/w9cdrom/293/293.html

[Muslea *et al*. 2000] "Selective sampling with redundant views" In *Proceedings of national conference on Artificial Intelligence*, 2000.

[Muslea *et al.* 2002 a] "Adaptive view validation : a case study on wrapper induction", I. Muslea, S. Minton, C. Knoblock. In *Proceedings19th ICML*, 2002.

[Muslea *et al.* 2002 b] "Active+ Semi-Supervised Learning = Robust Multi-View Learning" I. Muslea, S. Minton, C. Knoblock *Proceedings19th ICML*, 2002.

[Oard 1997] Douglas W. Oard, "Alternative Approaches for Cross-Language Text Retrieval," in *Cross-Language Text and Speech Retrieval*, AAAI Technical Report SS-97-05. http://www.clis.umd.edu/dlrg/filter/sss/papers

[Sakamoto 2002] "Knowledge Discovery from Semistructured Texts", in Arikawa & Shinohara (eds) *Progress in Discovery Science* (LNAI 2281) Springer, 2002.

Troisième partie

Démonstrations

Web-R, la mémoire exhaustive de ma Toile

A. LIFCHITZ
CNRS - LIP6 (Laboratoire d'Informatique - Université Paris 6),
8, rue du Capitaine Scott,
75015 Paris, FRANCE
Mail : alain.lifchitz@lip6.fr
Tél : +33 1 44 27 43 32 Fax : +33 1 44 27 70 00

J.D. KANT
LIP6 (Laboratoire d'Informatique - Université Paris 6),
8, rue du Capitaine Scott,
75015 Paris, FRANCE
Mail : jean-daniel.kant@lip6.fr
Tél : +33 1 44 27 88 05 Fax : +33 1 44 27 70 00

Résumé

Dans cette démonstration, nous présentons le système *Web-R,* un outil non-intrusif et rapide d'enregistrement qui réalise un stockage complet et systématique des pages web de l'utilisateur et de sa navigation. Il sauvegarde tous les composants qui seront nécessaires et suffisants à restituer hors-ligne la page de la même façon qu'elle l'a été vue en-ligne. Nous montrons que cet enregistrement systématique du Web personnel est non seulement techniquement réalisable, mais aussi réaliste car ne demandant qu'une faible fraction du volume de stockage des disques durs actuels. De surcroît, *Web-R* fournit un mécanisme de gestion de cet espace de stockage local incluant la comparaison du contenu de pages, évitant une inutile redondance. Enfin, puisque l'ensemble du contenu des pages visitées est stocké, *Web-R* permet à l'utilisateur d'avoir une vue globale de sa navigation personnelle par des statistiques, des outils de tri et de filtrage, pour sa réutilisation et plus tard sa structuration semi-automatique.

Abstract

In this demo, we present the *Web-R* system, a non-intrusive and fast recording tool that performs a systematic and complete storage of user's web pages and navigation. It saves all the components that are necessary and sufficient to visualize offline the page the same way it was displayed online. We show that a systematic storage of the personal web is non only technically feasible but also realistic since it requires only a small fraction of the storage space available in current disks. Moreover, *Web-R* also provides a way to manage this storage space and integrates a page comparison mechanism to avoid unnecessary redundancy. Finally, since the full content of visited pages is stored, *Web-R* allows the user to have a global view on his/her personal navigation by incorporating statistics, sorting and filtering tools and later on its semi-automated structuring.

Description de la démonstration

Nous effectuerons la démonstration du logiciel *Web-R*, un outil non-intrusif et rapide d'enregistrement qui tend à réaliser un stockage complet et systématique des pages web de l'utilisateur et de sa navigation. Il sauvegarde tous les composants qui seront nécessaires et suffisants à visualiser hors-ligne la page de la même façon qu'elle l'a été en-ligne. Afin d'éviter la saturation du disque local, *Web-R* fournit aussi un mécanisme de gestion de l'espace de stockage ainsi qu'un autre pour la comparaison de pages évitant une inutile redondance. De plus, puisque l'ensemble du contenu des pages visitées est stocké, *Web-R* permet à l'utilisateur d'avoir une vue globale de sa navigation personnelle par des statistiques, des outils de tri et de filtrage pour sa réutilisation.

Toutes ces fonctionnalités seront illustrées lors de cette démonstration, qui procèdera en 4 étapes :

1. *Prise en main du logiciel* : login, interface principale, premier enregistrement de page

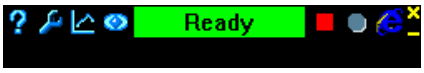


Figure 1. L'interface racine de *Web-R* (échelle réelle).

2. *Enregistrement des pages dynamiques* : il s'agit de montrer que *Web-R* parvient à enregistrer des pages délicates (car dynamiques, *e.g.* scripts, *frames*, animations, *Flash Macromedia*®, etc.) que la plupart des autres systèmes (*i.e.* ceux qui sont fondés sur le stockage d'URL, les fonctions « Enregistrer sous » des navigateurs, les navigateurs hors-ligne, etc.) ne parviennent pas à enregistrer correctement. Pour vérifier que ces pages sont complètement stockées, nous réaliserons un visionnage hors-ligne de leur contenu grâce à la visionneuse intégrée dans *Web-R*. Nous montrerons également l'enregistrement de pages sécurisées (`https//…`).

3. *Gestion de l'espace de stockage* : nous montrerons l'efficacité du mécanisme de comparaison de pages qui évite de stocker plusieurs fois des pages identiques par leur contenu, ainsi que les outils (limitations de taille, purge automatique) fournis à l'utilisateur pour gérer l'espace disque dévolu à l'enregistrement .

4. *Visionnage hors-ligne et statistiques* : nous terminerons cette démonstration par la présentation de la visionneuse hors-ligne des pages intégrée à *Web-R*, qui permet notamment de consulter un sous-ensemble de la base des pages stockées grâce des filtres appliquant différents critères croisés. Nous présenterons également les différentes statistiques affichées par *Web-R* qui permettent d'avoir une vue plus précise sur ce qui est enregistré sur le disque.

GeniMiner, un outil pour la veille stratégique

F. Picarougne, N. Monmarché, M. Slimane, G. Venturini

Laboratoire d'Informatique,
École Polytechnique de l'Université de Tours (Dépt. Informatique),
64 avenue Jean Portalis,
37200 Tours, FRANCE.
Email : fabien.picarougne@etu.univ-tours.fr;{monmarche,slimane,venturini}@univ-tours.fr
Tél : +33 2 47 36 14 14 Fax : +33 2 47 36 14 22

Résumé

Nous proposons dans cette démonstration un moteur de recherche complémentaire aux moteurs de recherche standard. Cet outil s'inscrit dans le cadre de la veille stratégique. Nous supposons que l'utilisateur peut attendre plusieurs heures avant d'obtenir ses résultats. Ainsi, nous utilisons ce temps pour effectuer une recherche plus approfondie des pages grâce à une requête utilisateur plus complexe que celle présentée généralement dans les moteurs standards. Nous nous intéressons spécialement au problème suivant : quelle stratégie peut efficacement maximiser le gain de l'utilisateur (i.e. quelles pages doivent être explorées dans le but de trouver des documents intéressants qui vont minimiser le temps dévolu à une analyse manuelle). Nous utilisons le fait que les algorithmes génétiques résolvent de manière optimale le problème d'exploration vs. exploitation (i.e. utiliser les résultats produits par les moteurs de recherche standards vs. explorer les liens/pages des documents connus).

Abstract

We propose in this demonstration a search engine that is complementary to standard ones. This tool lies within the scope of the strategic watch. We suppose that the user can wait for his results during several hours. So, we use this time for performinge a more deeper search on the pages thanks to a more complex user request than those used by standard search engines. We are especially interested in the following problem: what strategy can hopefully maximize the user's gain (i.e. which pages should be explored in order to find interesting documents that will minimize the time devoted to manual analysis). We use the fact that the genetic algorithms optimally solve the problem of exploration vs. exploitation (i.e. use the results given by standard search engines vs. explore links/pages of known documents).

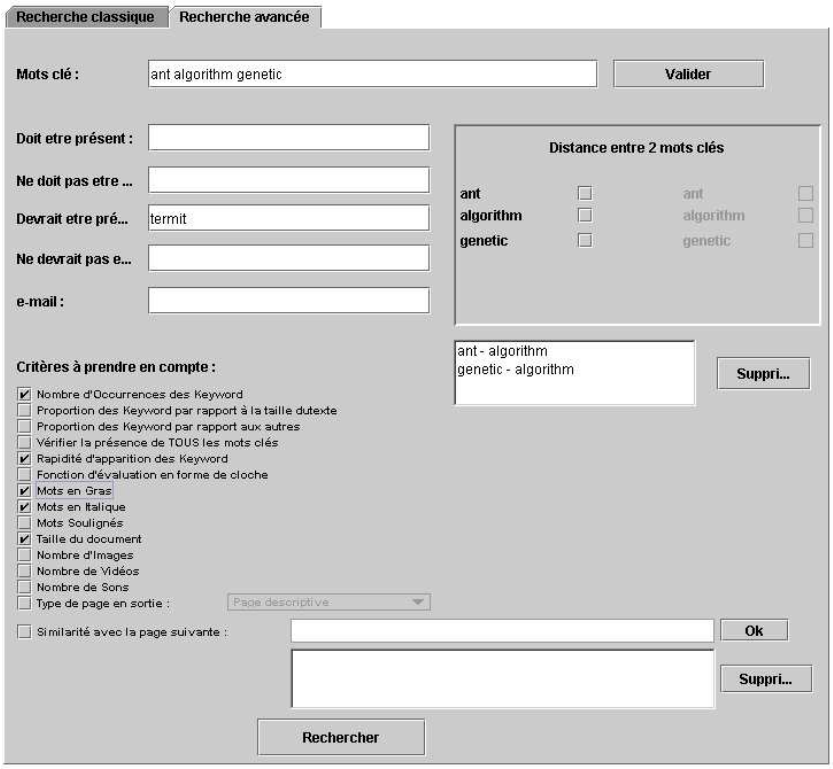


FIG. 1 – *Interface d'interrogation de GeniMiner.*

Nous abordons dans cette démonstration le problème de la recherche d'information sur le web. Ce problème consiste à trouver une ou plusieurs pages web qui répondent de manière pertinente à la requête définie par un utilisateur. Nous modélisons ce problème comme un problème d'optimisation auquel peuvent être appliqués de nombreux concepts et méthodologies issus des techniques d'optimisation. Pour cela, nous considérons Internet comme un graphe dans lequel les noeuds sont représentés par les documents présents sur le réseau et les arcs par les liens hypertextes contenus dans ces documents. La requête de l'utilisateur définit la fonction d'évaluation qui va mesurer l'intérêt d'une page web donnée, où le voisinage local présent dans cet espace de recherche est défini par les liens entre les pages. Cette requête est basée sur un grand nombre de critères qui vont être choisis par l'utilisateur pour sélectionner les documents correspondant le mieux à ses souhaits (proximité des mots clés, prise en compte de la typographie, ...). Nous avons alors définit un algorithme génétique à partir de cette modélisation pour résoudre le problème d'optimisation obtenu.

Nous avons réalisé plusieurs tests sur le web dans son ensemble avec des requêtes réelles en comparant la qualité des pages obtenues (au sens de notre fonction d'évaluation) par une interrogation simple des moteurs de recherche standard (Google, etc) et par le complément d'une recherche locale effectuée par notre algorithme génétique. Les résultats obtenus confirment l'intérêt de notre approche et sa compétitivité et complémentarité par rapport aux moteurs de recherche existants.

Construction de sites portails par des fourmis artificielles

H. Azzag, N. Monmarché, M. Slimane, G. Venturini

Laboratoire d'Informatique,
École Polytechnique de l'Université de Tours (Dépt. Informatique),
64 avenue Jean Portalis,
37200 Tours, FRANCE.
Email : hanene.azzag@etu.univ-tours.fr;{monmarche,slimane,venturini}@univ-tours.fr
Tél : +33 2 47 36 14 14 Fax : +33 2 47 36 14 22

C. Guinot

CE.R.I.E.S,
20, rue Victor Noir,
92521 Neuilly-sur-Seine Cédex, France.
Email : christiane.guinot@ceries-lab.com
Tél : +33 1 46 43 43 59 Fax : +33 1 46 43 46 00

Résumé

Dans cette démonstration nous présentons un nouvel algorithme de classification non supervisée pour la construction automatique d'une hiérarchie. Il utilise le principe d'auto-assemblage observé chez une colonie de fourmis. Chaque fourmi représente une donnée (ex : une page web). Les déplacements et les assemblages des fourmis sur cet arbre dépendent de la similarité entre les données. Il en résulte une organisation arborescente qui permet de nombreuses applications : détermination automatique d'une classification «plane», exploration visuelle de l'arbre, génération d'un dendogramme, construction automatique de sites portails.

Abstract

In this demonstration, we present a new algorithm for unsupervised learning. It is inspired from the self-assembling behavior observed in real ants. The artificial ants that we have defined will similarly build a tree. Each ant represents one data (a text for exemple). The way ants move and build this tree depends on the similarity between the data. When all ants are fixed in the tree, this hierarchical structure can be used in several ways: it can be seen as a partitioning of the data, it can be used for data visualization purposes, it can transformed into a dendogram as in hierarchical clustering and it is used for the automatic generation of portal sites.

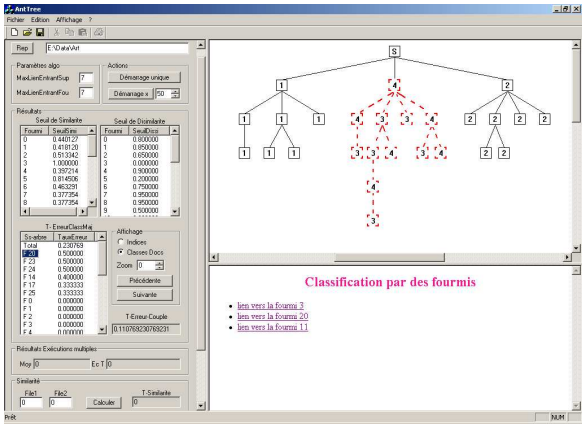


FIG. 1 – *Visualisation et exploration interactive de l'arbre.*

Nous proposons dans cette démonstration une modélisation nouvelle inspirée des fourmis réelles. Il s'agit de modéliser le phénomène d'auto-assemblage observé chez les fourmis pour construire des structures vivantes (grappe de fourmis, chaînes de fourmis) et d'utiliser ce comportement pour organiser les données à regrouper selon un arbre qui se construit de manière hiérarchique et distribuée.

À partir d'un support fixe sur lequel sont situées initialement les fourmis (les données), ces dernières vont s'accrocher successivement à ce point fixe, puis aux fourmis connectées au support, et ainsi de suite jusqu'à ce que, par exemple, une chaîne de passage soit construite entre deux points. Les fourmis se déplacent sur la structure vivante et s'accrochent sur celle-ci aux endroits les plus opportuns en fonction de la similarité entre les données.

Nous avons testé notre algorithme de classification sur des données numériques, des données du CE.R.I.E.S et sur des pages web issues d'Internet. L'interface de l'application permet de visualiser les résultats de deux manières différentes. La première propose de générer les résultats de manière similaire à un site portail et dont la construction est entièrement automatique. La première page représente le support et contient des liens vers d'autres pages que représentent les fourmis connectées à ce point fixe. Ces dernières contiennent elles-mêmes des liens vers d'autres pages représentant les fourmis connectées à elles et ainsi de suite pour toutes les autres fourmis. La seconde propose une vue sous forme d'arbre dont la racine est le support, les sous-arbres placés directement sous ce point constituent la classification «plane» trouvée par AntTree, chaque sous-arbre correspondant à une classe constituée de toutes les données présentes dans ce sous-arbre. L'interface permet aussi à l'utilisateur de naviguer au sein de l'arbre et de l'explorer visuellement et interactivement afin de mieux comprendre l'organisation arborescente des données, il peut évaluer directement la taille d'une classe et la répartition des données en sous-arbres. Une mesure d'erreur de classification est également indiquée pour chaque noeud de l'arbre avec des couleurs pour les sous-arbres commettant l'erreur. Il est possible aussi de transformer l'arbre construit en dendogramme. Enfin il reste beaucoup de problèmes à résoudre pour ce type d'approche et les résultats obtenus jusqu'ici sont plus qu'encourageants.

Survol de données géographiques en 3D temps réel sur Internet

C. Guéret, M. Slimane, C. Proust, G. Venturini

Laboratoire d'informatique,
École Polytechnique de l'Université de Tours (Dépt. Informatique),
64 avenue Jean Portalis,
37200 Tours, FRANCE.
Email : christophe.gueret@etu.univ-tours.fr,{slimane,proust,venturini}@univ-tours.fr
Tél : +33 2 47 36 14 14 Fax : +33 2 47 36 14 22

Résumé

La présentation consiste en une démonstration du système de survol de terrain en 3D dimensions. Cette application, FlyMap, permet à un internaute d'effectuer un survol totalement libre de la zone géographique modélisée.

Les deux principales caractéristiques de FlyMap sont l'utilisation d'une architecture réseau pour la gestion des données et l'utilisation de Java et OpenGL pour la réalisation du programme.

La démonstration est faite sur un jeu de données modélisant le département de l'Indre (36).

Abstract

The presentation consists in a demonstration of three dimensions terrain fly-over system. This application, named FlyMap, allows an Internet user to do a totally free flight above a geographical zone.

The two main FlyMap's characteristics are first the use of a network architecture to manage data and second the use of Java and OpenGL to realize the application.

The demonstration is done using a data set representing the French "Indre" county.

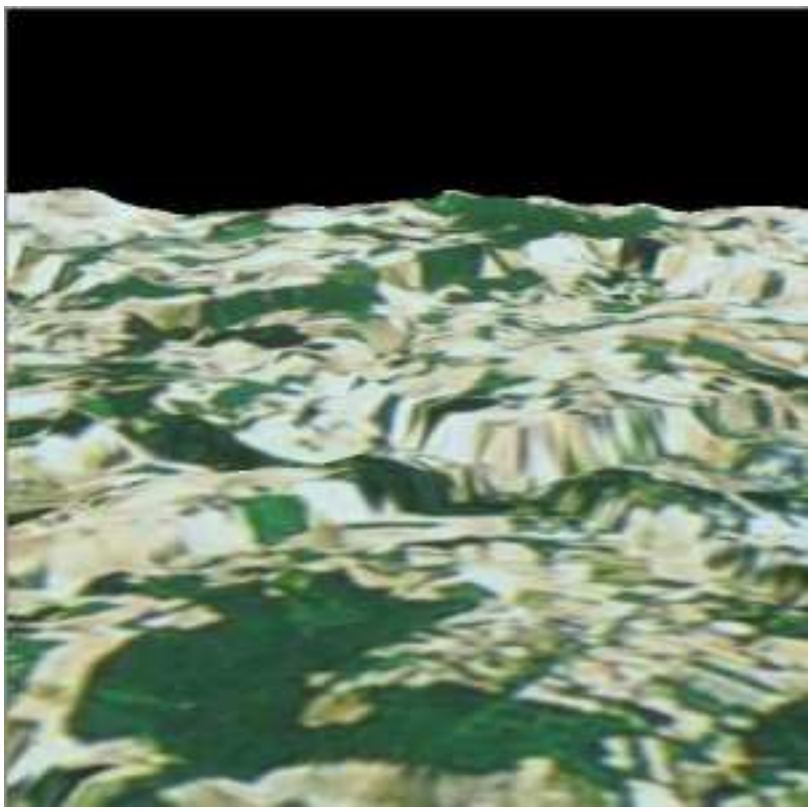


FIG. 1 – *FlyMap - Survol de l'Indre*

Les données sont gérées sur un modèle client/serveur. Un serveur est chargé de fournir à toutes les applications clients les données nécessaires au calcul de la scène 3D à un moment précis. Ces données sont transférées selon une méthode de transfert continu (ou *streaming*) afin que l'utilisateur de l'application cliente soit le moins possible dérangé par le temps nécessaire à l'envoi des données via le réseau.

Le programme est écrit sous la forme d'une "applet" Java et la librairie 3D choisie pour le calcul des images est OpenGL. Ces choix nous assurent une portabilité de l'application sur la majeure partie des systèmes d'exploitation du marché. FlyMap peut être utilisé depuis n'importe quel navigateur Internet de type Internet Explorer ou Mozilla supportant la technologie Java.

Dans le but de s'adapter aux contraintes de bande passante, plusieurs choix ont étés fait quant à la précision de la surface représentée. Ainsi un compromis entre qualité d'image et rapidité d'affichage nous permet d'avoir une application interactive et rapide tout en gardant une image suffisament précise.

Des algorithmes d'optimisation d'envoi du flux de données furent également mis au point afin de tirer partie de l'arrivée progressive des informations par le réseau.

Webxygen, un générateur de sites Web

A. OLIVER, F. PICAROUGNE, H. AZZAG,
G. VENTURINI, N. MONMARCHÉ, M. SLIMANE

Laboratoire d'Informatique, Ecole Polytechnique de l'Université de Tours
64 Avenue Jean Portalis,
37200 Tours, France

Mail :{oliver,venturini,monmarche,slimane}@univ-tours.fr
Fabien.picarougne@etu.univ-tours.fr

Résumé

Cette démonstration porte sur un générateur de sites Web baptisé Webxygen. Ce générateur concentre un certain nombre de fonctionnalités avancées dont vont pouvoir hériter tous les sites Web représentés dans ses bases de données. Parmi ces fonctionnalités, nous détaillerons : l'édition à distance de pages en multiples langues, la gestion de styles d'affichage et leur génération automatique par un algorithme générique interactif, l'analyse statistique de l'audience du site avec notamment la classification automatique des sessions utilisateurs, l'utilisation d'un système de suggestion à l'utilisateur, la gestion des droits d'accès permettant de définir des extranets.

Abstract

This demonstration deals with a web sites generator called Webxygen. This generator integrates many advanced functionalities which will inherited by all sites hosted in its databases. Among those functionalities, we will detail : the remote editing of pages with multilingual modes, changing and optimizing the sites styles with an interactive genetic algorithm, the statistical analysis of sites audience with, among other things, the clustering of users sessions, the use of a user suggestion system, the handling of users access rights in order to define extranets.

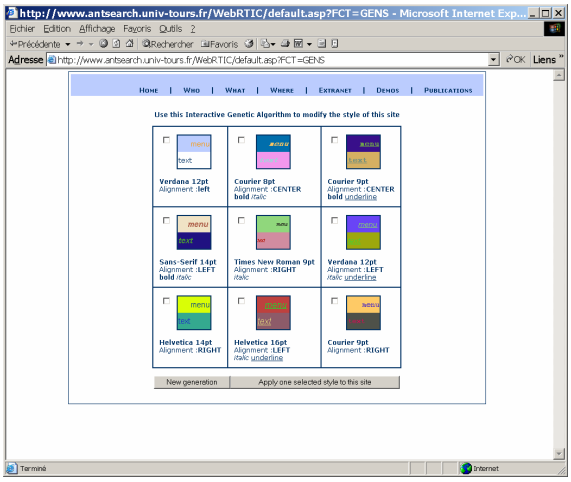


Fig. 1 : Génération interactive du style

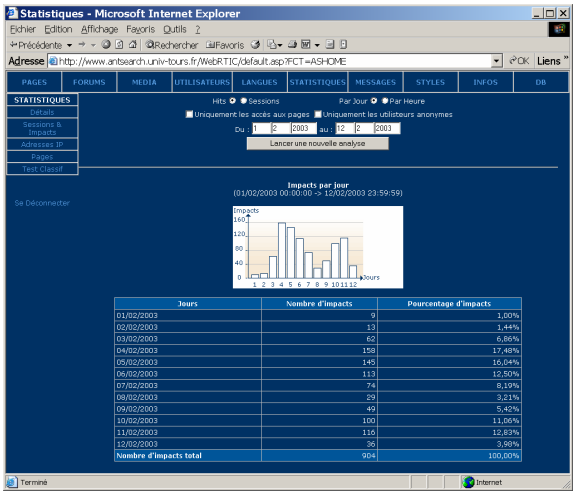


Fig. 2 : Analyse statistique

Webxygen (pour Web Sites Generator) représente un site Web de manière dynamique dans une base de données avec des champs pour chaque page, chaque média (image, son, etc), chaque utilisateur défini, chaque visite du site, etc. Cette représentation commune à tous les sites construits par Webxygen permet de centraliser un grand nombre de fonctionnalités :

- les sites sont entièrement administrables à distance à partir de n'importe quel navigateur (voir par exemple l'interface liée aux statistiques sur la figure 2),
- autant de langues que nécessaire peuvent être définies pour le site, l'administrateur pouvant modifier pour une même page les différentes versions,
- autant de styles que nécessaire peuvent être définis (voir figure 1). Nous nous attachons par exemple à définir un style pour les mal-voyants avec un contraste élevé. Chaque style peut être déterminé par un algorithme génétique interactif (voir la page démo de notre site),
- Webxygen enregistre toutes les visites ayant lieu sur le site, et place des cookies sur les machines des visiteurs. Il en découle tout un ensemble d'analyse statistique consultables en temps réel et à distance (voir figure 2),
- Les actions de l'utilisateur sont suivies en temps réel. Cela permet par exemple de lui suggérer en temps réel des pages proches de celles qu'il a déjà visitées,
- A chaque page et à chaque média est associé un niveau d'accréditation nécessaire pour sa visualisation. L'utilisateur anonyme a le niveau zéro et voit tous les éléments d'une page (et toutes les pages de niveau 0). Il est ensuite possible de définir des utilisateurs et leur associer un niveau d'accréditation. Webxygen filtre dynamiquement les informations qu'un utilisateur a le droit de voir ou non, ce qui permet de définir des extranets à plusieurs niveaux de confidentialité,
- Webxygen gère automatiquement les forums de discussion en enregistrant toutes les échanges qui y ont lieu. Les forums sont également administrables à distance.

Système de Questions-Réponses multilingue sur le Web

F. DUCLAYE, O.COLLIN, P.FILOCHE

France Télécom R&D,
2 avenue Pierre Marzin,
22307 Lannion Cedex,FRANCE.
Email : {florence.duclaye,olivier.collin,pascal.filoche}@rd.francetelecom.fr
Tél : +33 2 96 05 37 02, +33 2 96 05 26 10, +33 2 96 05 14 36 Fax : +33 2 96 05 32 86

Résumé

L'outil présenté est notre système de Questions-Réponses multilingue, permettant l'accès à des informations sur le Web en temps réel. L'utilisateur pose une question factuelle précise comme "Qui a acheté Netscape ?" et obtient la réponse précise : "AOL". Notre système repose sur des outils d'analyse linguistique et est adaptable à de nombreuses langues. La démonstration présente la version actuelle de notre système multilingue de Questions-Réponses sur le Web.

Abstract

The tool presented is our multilingual Question-Answering system, which makes it possible to access Web information in real-time. Users can ask precise factual questions such as "Who bought Netscape ?" and obtain the precise answer : "AOL". Our system is based on tools for linguistic analysis and is adaptable to many languages. The demonstration presents the current version of our multilingual Question-Answering system on the Web.

Les moteurs de recherche disponibles sur le Web permettent d'obtenir des informations rapidement. Cependant, dans ces modèles d'interaction par mots-clés, l'information recherchée n'apparaît pas toujours clairement. L'utilisateur désirant poser une question devra la traduire en mots-clés, envoyer la requête au moteur, puis chercher si la réponse se trouve parmi les documents retournés. Il devra alors lire les documents les uns après les autres pour rechercher la réponse à sa question. Dans de nombreux cas, le contenu des documents renvoyés ne correspond pas à l'attente de l'utilisateur. Dans d'autres cas, l'information est présente dans certains documents mais ces derniers se trouvent noyés dans la masse fournie par le moteur.

Les systèmes de questions-réponses automatiques dépassent cette solution d'interaction par mots-clés. Leur principe est que l'utilisateur recherche une information précise et qu'il existe le plus souvent une seule bonne réponse. La requête de l'utilisateur doit donc décrire non plus un thème, mais un problème précis. Les inconvénients liés aux moteurs de recherche sont évités. D'une part, l'utilisateur a la possibilité de définir précisément l'information recherchée (par une question en langage naturel). D'autre part, la tâche fastidieuse de lecture des documents retournés par le moteur de recherche est évitée. La qualité des informations est ainsi privilégiée à leur quantité.

Notre système de questions-réponses permet à l'utilisateur de formuler une question en langage naturel et d'obtenir en quelques secondes la réponse précise. Les questions traitées sont factuelles : elles visent des réponses localisées dans un seul document et dans une même phrase (noms de personnes, villes, lieux, dates, etc). La recherche de la réponse se fait par l'intermédiaire d'un moteur de recherche du Web. Elle peut également se faire dans des bases locales ou sur des services en ligne. Grâce à des outils de traitement automatique des langues, le système analyse la question de l'utilisateur, effectue la recherche de documents, puis analyse ces documents pour y trouver la réponse. Notre système fonctionne pour plusieurs langues comme le français, l'anglais, l'espagnol, l'allemand, le polonais, ou le portugais.

Étant donnée une question exprimée en langage naturel, les mots-clés pertinents sont extraits et envoyés à un moteur de recherche sur le Web. Les documents dont les URLs sont retournées sont téléchargés et analysés pour y trouver des phrases susceptibles de répondre à la question (phrases contenant tous les mots-clés). L'analyse linguistique de la question et des phrases extraites se fait à l'aide d'outils d'analyse morphologique, lexicale, et syntaxique de surface. À l'issue de l'analyse, les phrases sont lemmatisées et étiquetées en parties du discours et en catégories grammaticales. L'extraction de la réponse se fait en construisant des patrons syntaxiques d'extraction des réponses, générés par un ensemble de règles appliquées au moment de l'analyse linguistique de la question. Ces règles consistent à transformer la représentation syntaxique de la question en un ou plusieurs patrons possibles de réponse. Ces règles linguistiques génériques incluent des règles de reformulation sémantique de la réponse. De manière à présenter une vue synthétique des résultats, une fois que les réponses potentielles ont été extraites, le système les réordonne en fonction de leur fréquence et de leur position syntaxique au moment de l'application du patron d'extraction.

Réalisation accélérée de sites web dynamiques

http://www.softandem.com

OLIVIER FRINAULT
Softandem
La Ville Es Ray
35190 Québriac
Mail :olivier.frinault@softandem.com
Tél : +33 6 03 68 43 83

FRÉDÉRIC VISSAULT
Softandem
La Ville Es Ray
35190 Québriac
Mail :frederic.vissault@softandem.com
Tél : +33 6 72 55 87 55

Résumé

A l'aide d'exemples extrêmement concrets, nous vous présenterons les fonctionnalités et les avantages d'Omnigen et **PhpPager**, deux produits complémentaires pour la réalisation accélérée de sites web dynamiques.

Omnigen automatise toutes les tâches de production logicielle répétitives, et ceci dans n'importe quel langage informatique.

PhpPager, adapté aux technologies PHP, est une bibliothèque dédiée à la génération du code HTML des sites internet.

Abstract

Thanks to extremely practicle examples, we will demonstrate the functionalities and the advantages of Omnigen and **PhpPager**, two complementary products that speed-up the development of dynamical Internet websites.

Omnigen automates every repetitive task of the software development process, whatever be the coding language.

PhpPager, fitting to PHP technologies, is a software library dedicated to HTML code generation for Internet websites.

1. Produits de Softandem

1.1 Omnigen

Le produit Omnigen est un outil de génération de code ouvert à tout langage informatique, dont l'objectif est d'automatiser toutes les tâches de production logicielle répétitives. Le gain est double et immédiatement mesurable : d'une part, l'équipe de développement réduit substantiellement la charge consommée pour la production, et d'autre part, le code généré est d'une fiabilité incomparable à un code produit manuellement, ce qui rend inutiles certaines phases de mise au point (tests et corrections).

Le produit Omnigen est actuellement disponible dans une version adaptée à la génération de code pour les bases de données : à partir de la définition du modèle de données effectuée à l'aide de l'interface utilisateur du logiciel, il génère le code – personnalisable – relatif à :

- la définition de la structure de la base de données,
- la couche objet persistant, réalisant les accès à la base de données,
- la couche objet métier.

.

1.2 PHP Pager

PHP Pager est une brique logicielle prête à l'emploi écrite en PHP permettant d'accélérer notablement la réalisation des interfaces HTML des sites Internet.
PHP Pager utilise le principe de génération de code HTML à la volée et offre la possibilité de combiner toutes les technologies et fonctionnalités classiques des navigateurs Internet telles que Javascript, Applets, Flash, …

2. Démonstration proposée

Nous vous présenterons les avantages de nos outils de génération. La démonstration sera illustrée d'exemples concrets d'utilisation, mettant en œuvre les fonctionnalités des deux produits et leurs complémentarités.

---

1. Softandem products

1.1 Omnigen

Omnigen is a code generation tool for all type of programming language, that aims to automate every repetitive task of the software development process. The two main benefits are immediately visible : on one hand, the development team substantially reduces the production workload, and on the other hand, the generated code is incomparably more reliable than a hand-made program, which makes some debugging steps of testing and patching become useless.

The Omnigen product is now available with a version dedicated to code generation for databases : using the data model definition designed through the software graphic user interface, it generates – customizable – code relatively to :

- the database structure definition
- the persistent object layer that deals with database accesses
- the business object layer

1.2 PHP Pager

PHP Pager is a ready-to-use PHP software library that allows to speed up considerably the HTML interfaces development of Internet websites.
PHP Pager uses the on-the-flight HTML code generation principle, and offers the possibility to combine any classical browser-side technology and functionality as Javascript, Applets, Flash, …

2. Proposed show

We will demonstrate the advantages of our code generation tools. The show will be illustrated by practical use examples, showing the features of both products and their complementarities.

La toile comme un corpus dynamique

Cédrick Fairon – Anne Dister – Patrick Watrin
Cental – Centre de traitement automatique du langage
Université de Louvain
Collège Érasme - Place Blaise Pascal
1348 Louvain-la-Neuve
Belgique
{fairon, dister, watrin}@tedm.ucl.ac.be
Tél : +32 10 47 37 73

Le recours aux corpus pour la recherche d'exemples et d'attestations est aujourd'hui une pratique bien ancrée en linguistique : les chercheurs constituent des exempliers en collectant dans de vastes corpus les occurrences des phénomènes linguistiques qu'ils souhaitent étudier.

Des outils informatiques permettent d'automatiser les recherches avec plus ou moins de précision : il s'agit en général de logiciels entrant dans la catégorie des concordanciers. Lorsque le corpus est « épuisé », il doit être remplacé pour que la recherche puisse se poursuivre. GlossaNet est un service en ligne (http://glossa.fltr.ucl.ac.be) qui propose une solution à ce problème, en donnant accès à des corpus « ouverts », continuellement mis à jour grâce au prélèvement de nouveaux textes sur Internet.

Concrètement, le système est spécialisé dans la collecte des textes de presse : il télécharge quotidiennement sur Internet l'édition du jour de plus de 80 journaux dans 9 langues (français, anglais, italien, norvégien, portugais, espagnol, grec, néerlandais et allemand). Les textes récupérés sont ensuite analysés grâce aux programmes du logiciel Unitex[1], un analyseur de corpus qui permet l'application de ressources lexicales sur les textes. L'utilisateur de GlossaNet, dont la requête est mémorisée par le système, reçoit automatiquement, à chaque mise à jour du corpus, les résultats de la recherche par courriel sous la forme d'une concordance en HTML.

GlossaNet est principalement utilisé par des linguistes et chercheurs en TAL (pour la recherche et l'enseignement), mais est également utilisé pour ses possibilités de *veille* et comme moteur de recherche dans la presse en ligne.

Lors de cette démonstration, nous montrerons comment utiliser GlossaNet pour la *veille* (recherche d'information automatisée) et comment affiner les requêtes grâce aux outils linguistiques intégrés. Nous détaillerons également les quotidiens disponibles pour le français, en montrant que de nombreux pays de la francophonie sont représentés (France, Belgique, Suisse, Québec, Sénégal, etc.).

Matériel nécessaire : projecteur d'écran d'ordinateur et connexion internet.

[1] Unitex est un analyseur de corpus Open Source basé sur Unicode et développé par Sébastien Paumier à l'Université de Marne-la-Vallée (http://www-igm.univ-mlv.fr/~unitex/). Ce système permet d'appliquer des dictionnaires électroniques et des grammaires sur des textes. Les dictionnaires utilisés sont ceux du réseau RELEX: http://glossa.fltr.ucl.ac.be/dictionaries.html

ENGLISH VERSION

Using corpora for retrieving examples and linguistic attestations has become a common practice among linguists and NLP researchers. Various types of computer programs can help with that; they usually belong to the category of *concordancers*. Once the researcher has retrieved from a corpus all the examples he/she is interested in, the corpus is 'exhausted' and it is necessary to replace it, if more examples are needed. GlossaNet is an Web-based service (http://glossa.fltr.ucl.ac.be) which solves that problem: GlossaNet works on 'open corpora' that are continuously updated with new texts downloaded on the Internet. More precisely, the system generates a continuous flow of textual data by downloading online newspaper texts.

More than 80 online newspapers in 9 languages (French, English, Italian, Spanish, Portuguese, Norwegian, Greek, German and Dutch) are daily downloaded. Text corpora are analysed with Unitex programs and dictionaries.

GlossaNet users enter a query and select some newspapers through a Web Interface. Every time one of the selected newspaper is updated, the system re-apply the user query on the new corpus and if any results are found, they are sent to the user by email under the form of a concordance.

GlossaNet is mostly used by linguists and NLP researchers (for research and teaching) but is also used for IR in newspaper texts.

Fairon, Cédrick. 2001 : "Extension dynamique de ressources lexicales par consultation du Web". In *T.A.L. et Internet*. BULAG, N°26.

Fairon, Cédrick. 1999 : "Parsing a Web site as a corpus", In C. Fairon (ed.). 1998-1999. *Analyse lexicale et syntaxique: Le système INTEX*, Lingvisticae Investigationes Tome XXII (Volume spécial), Amsterdam/Philadelphia: John Benjamins Publishing, 450 p.

Démonstration d'un système de calcul des thèmes de l'actualité à partir des sites de presse de l'internet

JACQUES VERGNE
GREYC - UMR 6072
campus II - BP 5186
Université de Caen
14000 Caen, FRANCE
mail :Jacques.Vergne@info.unicaen.fr
tél. : 02 31 56 73 36 fax : 02 31 56 73 30

Résumé

Dans cette démonstration, nous présentons un système de constitution de revue de presse à partir des sites de presse présents sur l'internet[1]. Il s'agit de répondre à des questions telles que : "de qui, de quoi est-il question aujourd'hui dans la presse de tel espace géographique ou linguistique ?". L'utilisateur, qu'il soit un journaliste qui prépare sa revue de presse, ou simplement une personne intéressée par l'actualité, définit en entrée l'espace de recherche qui l'intéresse.

Ce système inverse la problématique des moteurs de recherche : au lieu de rechercher des documents à partir de mots-clés qui représentent des thèmes, il s'agit de produire en sortie les thèmes principaux de l'actualité, et de donner accès aux articles concernés par ces thèmes.

Les thèmes d'actualité sont capturés en relevant les termes récurrents dans les textes d'hyperliens des "Unes" des sites de presse. Le système calcule un graphe de termes dans lequel les nœuds sont les termes et les arcs sont les relations entre termes, relations définies par la co-occurrence de deux termes dans un texte de lien.

L'interface exploite ce graphe en permettant à l'utilisateur de naviguer parmi les termes et d'avoir accès aux articles contenant ces termes[2].

Mots-clés : hypertextes, web, internet, documents électroniques, web mining, recherche d'informations, veille stratégique, fouille de textes.

Abstract

In this demonstration, we present a system for building a news review, from news sites on the web. We want to be able to answer questions as : "who, what are papers speaking about today in the news of a given geographic or linguistic search space". The user, a journalist preparing his news review, or somebody interested in news, defines as input the search space he is interested in.

This system reverses the issues of search engines : in spite of searching documents from key-words which represents topics, we want to produce as output the main topics of the news, and to give access to related papers.

News topics are captured while computing recurrent terms in hyperlinks texts of front-pages of news sites. The system computes a graph in which nodes are terms and arcs are links between terms; a link is defined as a co-occurrence of two terms in a same link text.

[1] Une version préliminaire de la démonstration est accessible sur :
http://www.info.unicaen.fr/~jvergne/demoRevueDePresse/index.html

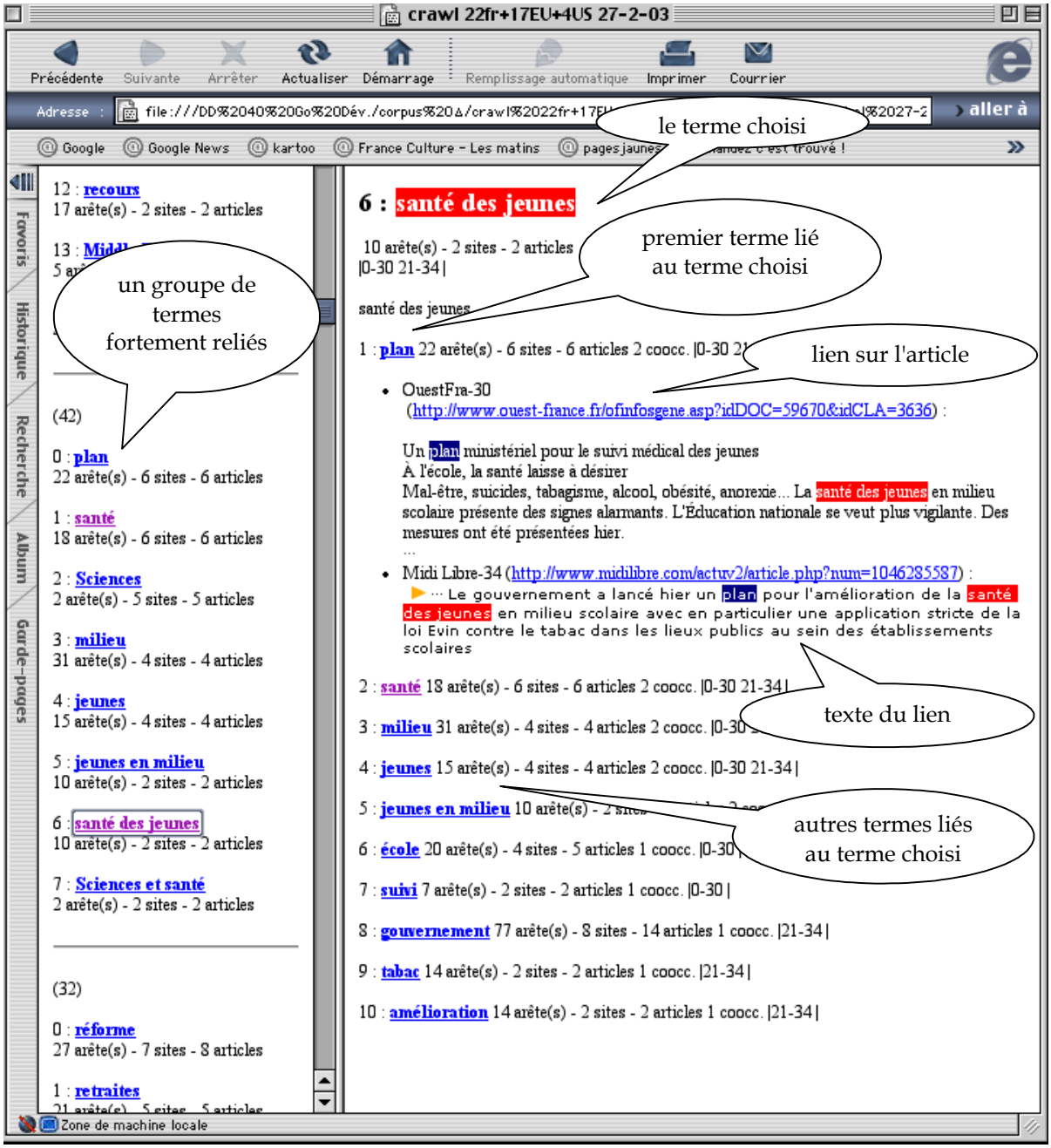
[2] Le système présenté a des analogies avec celui de Google News (http://news.google.fr), mais Google News n'a pas encore publié sur son processus de traitement.

The interface is based on this graph as the user can browse through the terms and have access to papers containing these terms.

Key-words : hypertexts, web, internet, electronic documents, web mining, information retrieval, strategic watching, text mining.

Cette démonstration correspond à l'article « Un système de calcul des thèmes de l'actualité à partir des sites de presse de l'internet ».

Interface



Le projet e-OCEA : vers une plateforme Internet dédiée aux problèmes d'ordonnancement

V. T'kindt, J.-C. Billaut, J.-L. Bouquard, C. Lenté,
P. Martineau, E. Néron, C. Proust, C. Tacquard
Laboratoire d'Informatique,
64 avenue Jean Portalis,
37200 Tours, FRANCE.
Email : tkindt@univ-tours.fr
Tél : +33 2 47 36 14 14 Fax : +33 2 47 36 14 22

Résumé

Cette démonstration présente un système d'aide à la décision dédié aux problèmes d'ordonnancement. Ce système, nommé e-OCEA (http://www.ocea.li.univ-tours.fr), est en cours de développement au Laboratoire d'Informatique de l'Université de Tours. Il fournit, notamment, aux utisateurs des outils pour créer des algorithmes de résolution, via l'Internet De la modélisation du problème à l'édition d'ordonnancement calculés, le système e-OCEA offre des outils qui peuvent être utilisés aussi bien par des chercheurs opérationnels que par des industriels. Dans cette démonstration nous présentons le projet dans son état actuel ainsi que ses développements futurs.

Abstract

This demonstration deals with a decision support system for scheduling problems. This system, called e-OCEA (http://www.ocea.li.univ-tours.fr), is being developed at the Laboratory of Computer Sciences of the University of Tours. It provides a user with tools to help creating an effective algorithm to solve his scheduling problem, via the Internet. From the modelisation of the problem to the visualization of a computed schedule, the e-OCEA system offers software that can be used either by operations researchers or industrial engineers. In this demonstration we present the current state of this system and provide future directions.

Le projet e-OCEA a pour objectifs de faciliter à la communauté scientifique et industrielle en ordonnancement l'étude et la résolution de tels problèmes, que ce soit dans un cadre théorique ou appliqué. Les problèmes d'ordonnancement ont fait l'objet de nombreuses études et leur applicabilité n'est plus à démontrer. Le projet e-OCEA a été initié par l'équipe Ordonnancement et Conduite du Laboratoire d'Informatique de l'Université de Tours et vise à offrir une plate-forme Internet à la communauté internationale. Cela comprend une base de données d'algorithmes, de problèmes et de références ainsi qu'une panoplie de modules pour aider un Décideur. L'équipe de Recherche Opérationnelle du DAI Politecnico di Torino (Italie) nous a récemment rejoint sur ce projet, ainsi que le groupe de recherche OCSD du LAAS à Toulouse. Notre objectif à 5 ans et de rendre disponible une première version complètement opérationnelle de cette plate-forme. Ce qui permettrait de structurer fortement la recherche en Ordonnancement et placerait les participants dans une position de leader dans ce domaine.

Lors de cette démonstration nous vous ferons voir l'ensemble des fonctionnalités de la plateforme e-OCEA. Celles-ci se décomposent en deux grandes parties : les fonctionnalités liées à la base de données et les fonctionnalités liées à l'identification et la résolution des problèmes d'ordonnancement. La base de données de données contient un ensemble de références bibliographiques, de jeux de données et de solutions ainsi qu'un ensemble de codes sources et exécutables correspondant aux algorithmes de recherche mis à disposition par la communauté. Les fonctionnalités liées à l'identification et la résolution des problèmes d'ordonnancement sont séparées en différents modules. La figure 1 présente l'écran du module d'identification. Ce module permet, à l'aide d'une représentation graphique simple, de "dessiner" un atelier et d'extraire à partir de cette représentation le problème d'ordonnancement sous-jacent. D'autres modules seront présentés lors de la démonstration.

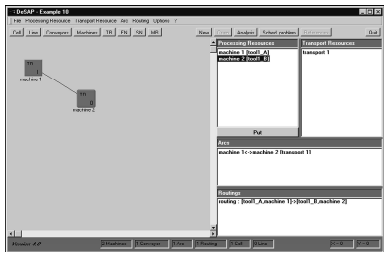


FIG. 1 – *Module d'identification de problèmes d'ordonnancement*

Textes, images, volumes : les bibliothèques numériques au Conservatoire National des Arts & Métiers

Pierre Cubaud, Jérôme Dupire, Alexandre Topol

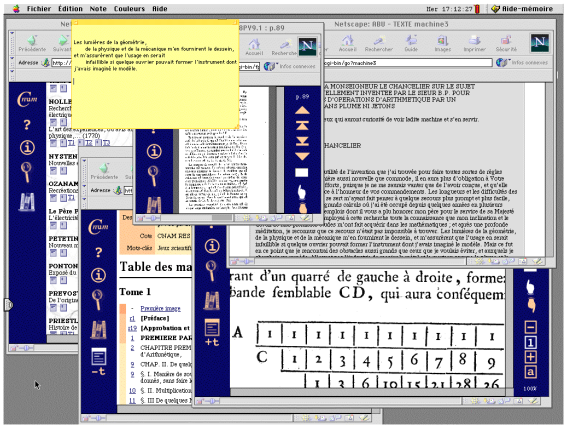
Centre d'études et de recherche en Informatique (CEDRIC)
CNAM, 292 rue St-Martin, F-75003 Paris
tél. : +33 (0)1 40 27 22 47
{cubaud, dupire_j, topol}@cnam.fr

Résumé

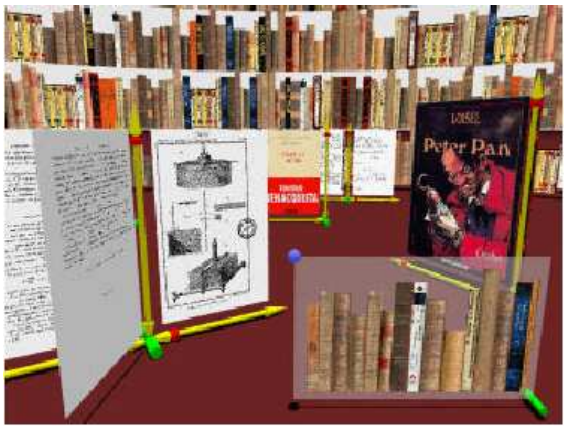
Nous présentons deux bibliothèques numériques utilisant une interface Web développées dans notre laboratoire, diffusant un corpus numérisé respectivement en mode texte et en mode image. L'étude des usages des ces sites (et d'autres similaires) met en évidence les limitations inhérentes aux interfaces « plates » (2D) du Web actuel en matière de navigation fluide dans un grand corpus documentaire ainsi qu'en matière de lecture à l'écran de fac-similés. Nous pensons que des métaphores d'interaction 3D pourraient répondre à ces difficultés et offrir aux lecteurs des outils plus proches de leur activité dans une bibliothèque réelle : un applicatif en cours de développement sera également présenté.

Abstract

We demonstrate two Web-based digital libraries developed in our laboratory. One offers a corpus digitalized in text mode, while the other relies on a image-based digitalization. Studying the effective usage of these sites (and those of other similar projects) reveals the inherent limitations of today's "flat" (2D) Web interfaces : the non fluidity of the navigation through a large documents corpus and the difficulty of on-screen reading for fac-similii. We believe that 3D interaction metaphors can provide a continuous navigation space for reading and browsing activities, closer to the library patrons' real life experience. A prototypal demonstrator will thus be presented.



(a) Session de travail avec ABU et le CNum



(b) Prototype de poste de lecture 3D

Les bibliothèques numériques ont bénéficié ces dernières années du progrès constant des technologies de captation, stockage et transmission numérique ainsi que de la chute de leur coût. Le développement du World-Wide-Web a également permis d'atteindre un public international considérable par le biais d' une interface standard et ergonomique. Lancée en octobre 1993 sur le premier site Web français, *l'Association des bibliophiles universels* (ABU, http://abu.cnam.fr) [1] dépasse aujourd'hui le millier d'ouvrages téléchargés quotidiennement, ce qui en fait un des sites les plus actifs de l'internet culturel francophone. Fruit d'un partenariat avec la bibliothèque centrale du CNAM et le Centre d' histoire des techniques (CDHT), le *Conservatoire numérique* (CNum, http://cnum.cnam.fr) a été mis en ligne en janvier 2000 (fig. a). Ce site diffuse 300 reproductions en fac-similés d'ouvrages scientifiques et techniques anciens [2]. Les courriers d'usagers de ces deux sites et les presque dix années de compte-rendus de leur activité (logs) constituent également pour notre équipe un observatoire des pratiques de lecture numérique sur le temps long. Il apparait que les interfaces Web de bibliothèques numériques actuellement en service n'offrent pas un confort d'utilisation suffisant pour dépasser le rôle bien restreint de diffusion de fac-similés pour l'impression à distance. Les évolutions constantes du Web, ainsi que la banalisation des cartes graphiques permettant la production d' images synthétiques en temps-réel ouvrent de nouvelles opportunités d'études en matière d'interaction humain-machine, pour : (1) une meilleure appréhension du contenu d'une bibliothèque numérique, (2) un meilleur support pour la lecture active des documents numérisés. Une première interface 3D de consultation du fond numérisé du CNum a d'abord été développée en VRML [3]. La difficulté de programmation de comportements interactifs dans ce langage nous a conduit à préférer le recours à des technologies propriétaires (Virtools dev, Renderware) et le développement d'un applicatif autonome. Après une phase de spécification des comportements des entités 3D du dispositif [4], nous souhaitons évaluer le nouveau prototype (fig. b) avec un pannel de lecteurs réguliers de l'ABU et du CNum.

Références

[1] P. CUBAUD, D. GIRARD, « ABU : une bibliothèque numérique et son public ». *Documents numériques*, vol. 2(3-4), 1998

[2] P. CUBAUD, G. DEBLOCK, B. ROZET « Le Conservatoire numérique des arts et métiers : une création patrimoniale ». *Bulletin des bibliothèques de France*, vol. 46(4), 2001.

[3] P. CUBAUD, A. TOPOL. « A VRML-based user interface for an online digitalized antiquarian collection ». *In* Proc. of ACM Web3D' 2001, Paderdorn, Germany, Feb. 2001.

[4] P. CUBAUD, A. TOPOL, P. STOKOWSKI. « Binding browsing and reading activities in a 3D digital library » *In* Proc. of the 2nd ACM/IEEE joint conf. on dig. libraries, Portland,USA, Jul. 2002.