Extracting and Evaluating Knowledge from e-Health Documents: A Contribution to Information Retrieval and Indexing

Lina F. Soualmia
LIM&Bio EA 3969
University of Paris 13, Sorbonne Paris Cité, France

Badisse Dahamna, Stéfan J. Darmoni
CISMeF & TIBS EA 4108
Rouen University Hospital, France

1 Introduction

The Internet is a major source of biomedical information. This chapter presents a simple yet efficient approach for extraction of information from biomedical documents available on the Internet. The main objective here is to re-use the information extracted during document retrieval and document indexing. In this work, health-information seekers are categorized into three main profiles according to their needs for health information: health professionals, students in medicine and the general public. The information-extraction approach of information extraction is based upon a data-mining process, called association rules, which can infer relations between medical concepts. A data-mining system may generate several thousands and even several millions frequent association rules, and only some of these will be interesting. In this chapter we will show how only the most relevant association rules are mined using Formal Concept Analysis and Galois closure. We consider a relevant association rule as being non-redundant with a minimal antecedent and a maximal consequent. This chapter also presents methods to evaluate the extracted information automatically and manually. The automatic evaluation of association rules relies upon the domain background knowledge *i.e.* the existing relations between medical terms in terminologies or thesauri and medical concepts in ontologies. Background knowledge is combined with conventional statistical measures such as "support" and "confidence". The manual evaluation is realized by experts of the medical domain (physicians and medical librarians). For that, a set of criteria have been modeled to single out the best relation when several are produced. This chapter also shows how to re-use extracted information in the form of association rules as an input in information-retrieval processes using query expansion (or query refinement), and in document indexing and classification.

1.1 Information Retrieval on Internet

Today, a web search is clearly one of the foremost methods for finding information. The growth of the Internet and the increasing availability of online resources has made the task of searching a crucial one. However, searching the web is not always as successful as users expect it to be and Internet users have to make a great effort to formulate a search query that returns the required results. Information retrieval concentrates on

developing algorithms to locate and select documents from a corpus that are relevant to a given query. The development of online information retrieval tools, such as search engines or search robots many of which utilize hyperlink analysis (Hou & Zhang, 2003, Liu, 2007), has been greatly beneficial to Internet users. The most popular search engine is Google[1].

Many of the users find the current process of searching the web unsatisfactory and feel overwhelmed by the large amount of documents retrieved. This is not only attributed to the search engine program: the lack of homogeneity in the structure of documents and in their indexing by the search robots makes it difficult to find the relevant information. Most of the retrieved sets of documents in a web search meet the search criteria but do not satisfy the user needs. Indeed, since the start of information retrieval research, the user's query has only been treated as an approximation of the seeker's information needs. However, a major problem is the inability of the user to formulate the appropriate query. A user often has only a vague idea of what the relevant query terms may be and has to rely upon an iterative process in which the retrieved query results are used to formulate the next query. One cause of this lack of specification in the formulation of a query is that most of the time the user does not know the vocabulary of the topic and the query terms are not those of the domain. Therefore, the query does not perfectly reflect what the seeker is searching for and its meaning is misinterpreted.

Research by (Grefenstette, 1997, Spink *et al.*, 2001) has shown that most search-engine users typically formulate very short queries of two to three words. Such short queries lack many useful words and do not sufficiently describe the subject that the user wants to investigate. (Spink *et al.*, 2001) have suggested that web users more often tend to go from a broad to narrow formulation in queries because the most common query modification is to add terms. In this way, seekers are unconsciously performing a well-known solution for information retrieval, namely *query expansion*. After the query process is performed, new terms are added and removed from the query to discard uninteresting retrieved documents and to retrieve interesting documents that were not retrieved (Eftihimiadis, 1996).

Another method for searching information on the Internet is browsing catalogs or gateways developed for a particular field. (Koch, 2000) defines quality-controlled subject gateways as Internet services that apply a comprehensive set of quality measures to support systematic electronic document discovery. Considerable manual effort is used to process a selection of electronic documents which meet quality criteria and to display an extensive description and indexing of these resources with standards-based metadata. Regular checking and updating ensure optimal management of collection. The main goal is to provide a high quality of subject access through indexed electronic documents using controlled vocabularies and by offering a deep classification structure for advanced searching and browsing.

1.2 Health Information Retrieval on Internet

The Internet is increasing in preeminence in numerous fields as a source of information, including health. In this domain, as in the others, users are now experiencing huge difficulties in finding precisely what they are looking for among the documents available online, in spite of existing tools. In medicine and health-related information accessible on the Internet, general search engines, such as Google, or general catalogues, such as Yahoo[2], cannot solve this problem efficiently. This is because they usually offer a selection of documents that turns out to be either too large or ill-suited to the query. Free text word-based search engines typically return innumerable completely irrelevant hits, which require much manual weeding by the user, and also miss important information resources. In this context, several health gateways have been developed. Some of them are evaluated in (Abad Garcia, 2005).

The CISMeF project (Catalog and Index of French Medical Web Sites[3]) (Darmoni *et al.*, 2001) was initiated in February 1995. As opposed to Yahoo, CISMeF is cataloguing the most important and quality-

[1] http://www.google.com
[2] http://www.yahoo.com
[3] http://www.cismef.org

controlled sources of institutional health information in French. The CISMeF catalogue describes and indexes a large number of health-information resources (n=13,452 in October 2003; n=73,960 in August 2010). CISMeF references high quality information resources. A resource can be a web site, web pages, documents, reports and teaching material: any support that may contain health information.

CISMeF takes into account the diversity of the end-users and allow them to find good quality resources. These resources are selected according to strict criteria by a team of librarians and are indexed according to a methodology (Darmoni *et al.*, 2001) which involves a four-fold process: resource collection, filtering, description and indexing. CISMeF is a quality-controlled gateway such as defined by (Koch, 2000). The following elements that characterize a typical quality-controlled health gateway are fulfilled in CISMeF: selection and collection development, collection management, intellectual creation of metadata, resource description (a metadata set), resource indexing (with controlled vocabulary system).

To include only reliable resources, and to assess the quality of health information on the Internet, the main criteria (*e.g.* source, description, disclosure, last update) of CISMeF are NetScoring[4] and HONCode[5]. In the following sections we describe the set of metadata elements and the terminology ‘‘oriented’’ ontology (Desmontils & Jacquin, 2002, Soualmia *et al.*, 2003) used in the catalogue.

1.2.1 CISMeF Metadata

The notion of metadata was around before the Internet but its importance has grown with the increasing number of electronic publications and digital libraries. « The Semantic Web dream is of a Web where resources are machine understandable and where both automated agents and humans can exchange and process information[6].**»**. The World Wide Web Consortium (W3C) have proposed that metadata should be used to describe the data contained on the web and to add semantic markup to web resources, thus describing their content and functionalities, from the vocabulary defined in terminologies and ontologies. Metadata are data about data, and in the web context, these are data describing web resources. When properly implemented, metadata enhance information retrieval.

The CISMeF uses several sets of metadata. Among them there is the Dublin Core (DC) (Baker, 2000) metadata set, which is a 15-element set intended to aid discovery of electronic resources. The resources indexed in CISMeF are described by eleven of the Dublin Core elements: *author, date, description, format, identifier, language, editor, type of resource, rights, subject* and *title.* DC is not a complete solution; it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF: *institution, city, province, country, target public, access type, sponsorships,* and *cost.* The user type is also taken into account. The CISMeF have defined two additional fields for resources intended for health professionals: indication of the *evidence-based medicine,* and the *method* used to determine it. For teaching resources, eleven elements of the IEEE 1484 LOM (Learning Object Metadata) “Educational” category are added.

1.2.2. CISMeF Controlled Vocabulary

Thesauri are a proven key technology for effective access to information as they provide a controlled vocabulary for indexing information. They therefore help to overcome some of the problems of free-text search by relating and grouping relevant terms in a specific domain. The main thesaurus used for medical information is the Medical Subject Headings[7] thesaurus used by the U.S. National Library of Medicine.

[4] http://www.chu-rouen.fr/netscoring
[5] http://www.healthonnet.org/HONcode
[6] Ian Horrocks, IEEE Intelligent systems, March/April 2002, 74-76.
[7] http://www.nlm.nih.gov/mesh/meshhome.html

The core of MeSH is a hierarchical structure that consists of sets of descriptors. At the top level we find general headings (*e.g.* diseases), and at deeper levels we find more specific headings (*e.g.* asthma). The 2010 version contains over 24,357 main headings (*e.g.* hepatitis, abdomen) and 83 subheadings (*e.g.* diagnosis, complications). Together with a main heading, a subheading allows to specify which particular aspect of the main heading is being addressed. For example, the pair (hepatitis/diagnosis) specifies the diagnosis aspect of hepatitis. For each main heading, MeSH defines a subset of allowable qualifiers so that only certain pairs can be used as indexing terms (*e.g.* aphasia/metabolism and hand/surgery are allowable, but hand/metabolism is not).

MeSH main headings and subheadings are organized hierarchically. However, these hierarchies do not allow a complete view concerning a specialty. The main headings and subheadings in the CISMeF controlled vocabulary are brought together under metaterms (*e.g.* cardiology). Metaterms (n=73) concern medical specialties and it is possible by browsing to know sets of MeSH main headings and subheadings qualifiers which are semantically related to the same specialty but dispersed in several trees.

MeSH was originally used to index biomedical scientific articles for the MEDLINE database. In addition to the set of metaterms, the CISMeF team has modeled a hierarchy of resource types (n=127), to customize MeSH to the field of e-health resources. These resource types describe the nature of the resource (*e.g.* teaching material, clinical guidelines, patient forums), and are a generalization or extension of the MEDLINE publication types.

Each resource in CISMeF is described with a set of MeSH main headings, subheadings and CISMeF resource types. Each main heading, (main heading/subheading) pair, and resource type is allotted a ‘minor’ or ‘major’ weight, according to the importance of the concept it refers to in the resource. Major terms are marked by a star (*).

Metaterms have been created to optimize information retrieval in CISMeF and to overcome the relatively restrictive nature of MeSH headings. For example a search on “guidelines in cardiology” or “databases in virology”, where cardiology and virology are descriptors and guidelines and databases are resource types, will yield few or no answers. Introducing cardiology and virology as metaterms is an efficient strategy to obtain more results because instead of exploding one single MeSH tree, the use of metaterms will result in an automatic expansion of the queries by exploding other related MeSH or CISMeF trees besides the current tree (Soualmia & Darmoni, 2005). The structure of the CISMeF controlled vocabulary is exploited for several tasks including resource indexing (manually and automatically), visualization and navigation through the concept hierarchies and document retrieval.

The importance of query expansion to improve retrieval effectiveness of the PubMed engine has been highlighted by (Aronson & Rindflesch 1997, Srinivasan, 1996). For the indexing process they both use an MeSH-indexed representation (in MEDLINE, a set of relevant MeSH terms for every record is manually associated as a representation of the content of the document). For the query, (Srinivasan, 1996) exploited a statistical thesaurus containing correlations between MeSH concepts and text, and (Aronson & Rindflesch, 1997) used the MetaMap system to associate the UMLS Metathesaurus concepts (Unified Medical Language System) with the original query.

Among the many sources that support users, including morphological bases, phonemic correction, and dynamic and contextual search tools (Soualmia & Darmoni, 2004), the MeSH structure is fully exploited in CISMeF. To complete these sources, we will now discuss how mining e-documents can be used to discover new associations between medical concepts through data-mining.

2 Data-Mining

Knowledge extraction from databases or data-mining in computer science is the process of identifying patterns in large sets of data. The aim is to “discover” previously unknown associations and thus consists of discovering

additional information from large structured sets of data. This knowledge could be used to perform predictions about new data as well as explain existing data.

2.1 Association Rules

The discovery of *association rules* is a widely used technique in data-mining. The general problem was described by (Agrawal *et al.*, 1993), who discovered relations among pieces of data (called *items*). The prototypical application of this task was the analysis of customers' basket data. For example, an association rule extracted from a French market database is expressed as follows: bread, cheese→wine. The meaning of this rule is quite intuitive, it says that whenever customers buy bread and cheese, they probably also buy wine in this supermarket.

A data-mining system may generate several thousands and even several millions frequent association rules, but only some of them are of interest. An association rule is interesting if it is easily understood by the users, valid for new data, useful, or confirms a hypothesis. The task of association rule mining can be applied to various types of data: any data set consisting of "baskets" containing multiple "items".

2.1.1 Definitions

Let I be a set of items, called *itemset*, and D a database of transactions where each transaction $T \in D$ is an itemset. An association rule is an implication rule expressed in the form of:

$$I_1 \Rightarrow I_2$$

where I_1 and I_2 are two itemsets $I_1, I_2 \subseteq I$ so that $I_1 \cap I_2 = \varnothing$. The rule expresses that whenever a transaction T contains I_1 then T probably also contains I_2. In other words, the implication rule means that the apparition of the itemset I_1 in a transaction, T, implies the apparition of the itemset I_2 in the same transaction. However, the reciprocal implication does not have to happen necessarily. I_1 is called antecedent and I_2 is called consequent.

2.1.2 Support

The support of an association rule represents its utility. This measure corresponds to the proportion of objects which contains at the same time the rule antecedent and consequent. In our example, the support measures the proportion of customers who bought bread, cheese and wine. It is possible to calculate the support of an association rule from the support of an itemset. $Supp(I_k)$ the support of the itemset I_k is defined as the probability of finding I_k in a transaction of T:

$$Supp(I_k) = \frac{\left|\{t \in T / I_k \subseteq t\}\right|}{\left|T\right|}$$

The support of the rule $I_1 \Rightarrow I_2$ written as $Supp(I_1 \Rightarrow I_2)$ is calculated as follows:

$$Supp(I_1 \Rightarrow I_2) = Supp(I_1 \cup I_2)$$

2.1.3 Confidence

The confidence of an association rule represents its precision. This measure corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent. In our example, the confidence measures the proportion of customers who bought wine among those whom bought bread and cheese. The confidence of the rule $I_1 \Rightarrow I_2$, written as $Conf(I_1 \Rightarrow I_2)$ is calculated as follows:

$$Conf\ (I_1 \Rightarrow I_2) = \frac{Supp\ (I_1 \cup I_2)}{Supp\ (I_1)}$$

Two types of rules are distinguished: *exact* association rules that have a confidence equal to 100%, *i.e.* verified in all the objects of the database and *approximative* association rules that confidence<100%. For example, if bread, cheese→wine (sup=20%; conf=70%), this rule says that 20% of customers buy bread cheese and wine together, and those who buy bread and cheese also buy wine 70% of the time. It is an approximative association rule: 70%, but not all customers, buy wine when buying bread and cheese.

2.2 Data-Mining Algorithms

Several methods are used to extract all of the association rules from a database. The simplest method consists of enumerating all the itemsets from which all the possible association rules could be generated. The total number of itemsets for a database that contains *n* Boolean attributes is 2^n. This naïve method is inapplicable to real-life databases. A more efficient method involves computing itemsets that have a support higher than a given threshold. They are called *frequent* itemsets. The association rules extraction time depends on the frequent itemsets extraction time.

Several accesses to the database are necessary to compute the number of database objects in which each frequent itemset candidate is contained. The association rules algorithms by level consider in each iteration a set of itemsets of a particular size, *i.e.* a set of itemsets in a level of the itemsets lattice. The following properties are used by these algorithms to limit the number of the itemsets candidates: all of the super-sets of an infrequent itemset are infrequent, and all the subsets of a frequent itemset are frequents (Agrawal & Srikant, 1994, Mannila *et al.*, 1994). These algorithms are founded on the following two-stepped model that finds all of the rules that satisfy user-specified minimum support and confidence:

1. Generate all large itemsets that satisfy minimum support
2. From large itemests generate all association rules that satisfy minimum confidence

The algorithms Apriori (Agrawal & Srikant, 1994) and OCD (Mannila *et al.*, 1994) realize a number of database accesses equal to the size of the larger frequent itemsets. Many researchers (Brin *et al.*, 1996, Park *et al.*, 1995, Thoivonen, 1996) have tried to improve various aspects of Apriori, such as the number of passes and accesses to the data-bases or the time efficiency of those passes.

The *frequent closed itemsets* (Pasquier *et al.*, 1998, Pasquier *et al.*, 2005) are defined by using the closure operator of the Galois connection of a finite binary relation (Ganter & Wille, 1999). In the A-Close algorithm, the binary relation *R* is between objects and items. Frequent itemsets are closed by the closure operator γ of the Galois connection which is the composition of the application φ which associates to a set $O \subseteq$ O the common items to all the objects $o \in O$, and of the application ψ which associates to an itemset $l \subseteq$ I the objects in relation with all the items $i \in l$.

$$\varphi(O) : \mathrm{O} \rightarrow \mathrm{I} \qquad \psi(O) : \mathrm{I} \rightarrow \mathrm{O}$$
$$\varphi(\mathrm{O}) = \{i \mid \forall o, o \in O \rightarrow (o,i) \in R\} \qquad \psi(O) = \{o \mid \forall i, i \in l \rightarrow (o,i) \in R\}$$

The operator $\gamma = \varphi\ o\ \psi$ associates to an itemset *l* the maximal set of items in common to all the objects that contain *l*, *i.e.* the intersection of these objects. All frequent itemsets and their support, and therefore all association rules, are deduced efficiently from the frequent closed itemsets without accessing the database. The frequent closed itemsets form a lattice whose size is limited by the size of the frequent itemsets lattice. The closed itemsets generators are also derived by the A-Close algorithm: the generators of a closed itemset *f* are the minimal itemsets *g* which closure $\gamma(g)=f$.

In real-life applications and especially for data other than that used for classic market basket analysis, some adaptations should be done on the algorithms. The problem of the relevance and the usefulness of extracted association rules is of primary importance because real-life databases lead to several thousands and even millions of association rules whose confidence measures are high, and among which there are many redundancies, *i.e.* rules conveying the same information among them. New bases (Duquenne & Guigues, 1986) for association rules are deduced from the closed frequent itemsets and their generators in A-Close. These bases consist of non-redundant association rules of minimal antecedents and maximal consequents, *i.e.* the most relevant association rules.

2.3 Knowledge Extraction Process

Mining association rules is generally included in a knowledge extraction process which is realized in several steps:

1. data and context preparation (*i.e.* objects and items selection) from the database(s),
2. extraction of the frequent itemsets (compared with a minimum support threshold),
3. generation of the most informative rules using a data-mining algorithm (compared with a minimum confidence threshold),
4. and finally, interpretation of the results and deduction of new knowledge (Fayyad *et al.*, 1996).

3 Extracting Knowledge from e-Health Documents

Our experiments are carried out on a database of electronic health documents, the CISMeF database. An extraction context is a triplet $C= (O, I, R)$ where O is the set of objects, I is the set of all the items and R is a binary relation between O and I. Applying this model to our database, the objects are the indexed electronic health documents. Each document has a unique identifier and a set of associated descriptors. These descriptors may be main headings and associations between main headings and subheadings. The relation R represents the indexing relation between an object and an item, *i.e.* descriptor that belongs to I.

We studied different extraction contexts by applying and adapting the A-Close algorithm. We distinguished two cases of data-mining: the technique based upon the conceptual indexing of the e-health documents and the technique based upon the plain-text indexing, in the latter a document being represented by the terms it contains and not only by the descriptors. Using conceptual indexing, we also studied the context of categorized documents, according to the user type and to the medical specialties.

There is an average of 6.5 descriptors by document in CISMeF with a minimum of 1 and a maximum of 300. This constraint on the number of descriptors — *i.e.* the size of the set of items — has been considered in the implementation phase of the A-Close algorithm. Indeed, A-Close works on databases with a maximum of 12 items. We have added another requirement to the implementation to avoid long time generation: maximal size of the closed itemsets is fixed to 300 items as it corresponds to the maximum number of descriptors for the documents. As an output, the association rules may be visualized in a file or automatically added to the database to be used in the information retrieval process (Figure 1.)

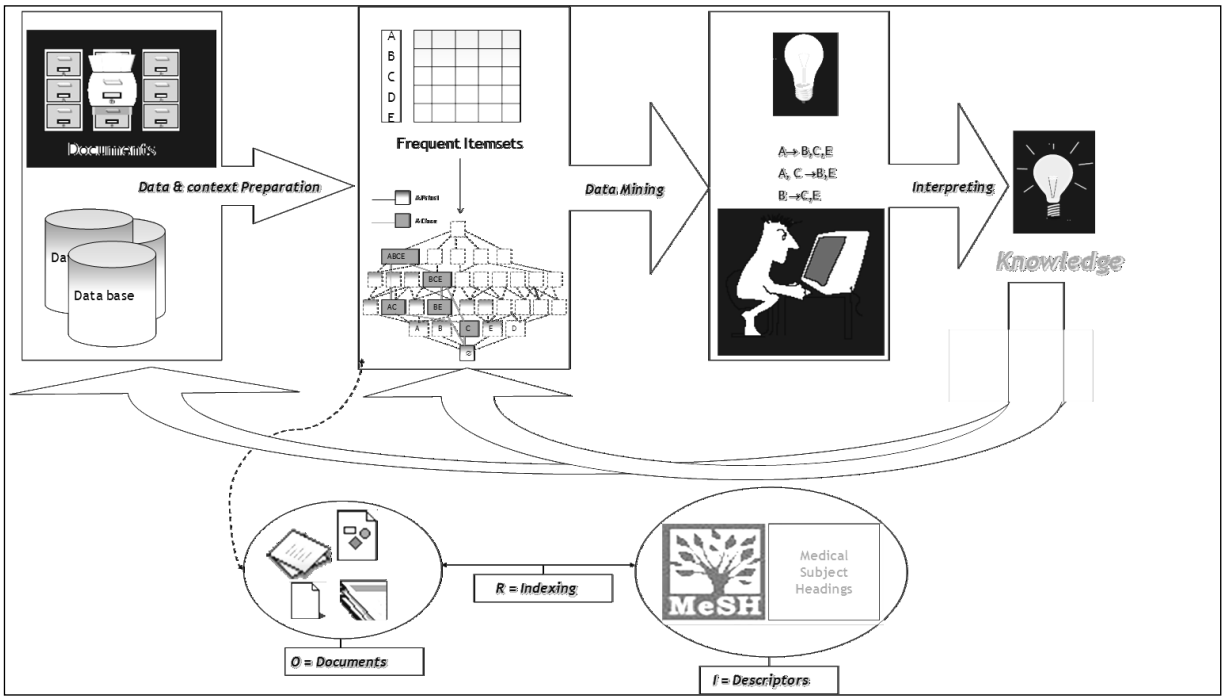

Figure 1: This figure describes the general process of knowledge extraction from the CISMeF database: data and context preparation, association rules mining using A-Close, more efficient than Apriori, knowledge interpreting by experts and knowledge re-injection in the database.

3.1 Extracting Knowledge from all the Database

3.1.1 Case 1: Itemset is the Set of Descriptors

In the first case, let I be the set of main headings (MH), which, via R, are used to index the set O of 11,373 documents. I={MH}.The 11,373 documents were selected at random. We have fixed the support threshold as minsup=20 and the confidence threshold as minconf=70%. A total of 11,819 rules were mined (2,438 exact with confidence=100%; 9,381 approximative with confidence≥70%). The number of rules is too high to be manually analyzed by our experts (physicians and medical librarians). Indeed, as indicated by (Gras *et al.*, 2001) the number of generated association rules may be high and the interpretation tasks of the results may become complex and inextricable.

3.1.2 Case 2: Itemset is the Set of Main Headings and Subheadings

In the second case, let I be the set of main headings (MH) and subheadings (SH) associated with the set of documents O. I={MH}∪{SH}. The 11,373 documents were selected at random. The support threshold was 20 documents and the confidence threshold was 70%. We obtained 16,976 rules (5,241 exact; 11,738 approximative). The same conclusions are drowned from the case 1.

3.1.3 Case 3: Itemset is the Set of Associations between Main Headings and Subheadings

In the third case, I is the set of the associations of main headings and subheadings (MH/SH) related to the documents. I={[MH/SH]}. Association rules between couples of (MH/SH) are more precise than association rules between main headings, and between main headings and subheadings since a subheading specifies a

particular aspect of a main heading (section 1.2.2.). With the same thresholds as in case 1 and case 2, the number of rules is 2,565 (648 exact rules; 1,917 approximative rules).

3.2 Categorizing Documents

The extracted association rules in the precedent cases are related to the medical domain. To obtain more precise rules we performed experiments on categorized documents according to groups of users (students in medicine, health professionals, and general public) and to medical specialties (*e.g.* cardiology, pharmacology) to evaluate the influence of categorization on the generation of association rules.

3.2.1 Categorization according to Users

In CISMeF, mainly three types of end-users are categorized: professionals, students in medicine, patients and lay people. We consider three "major" (section 1.2.2) resource types: guidelines*, education* and patients*. We also consider two kinds of itemsets: the set of major main headings I={MH*} and the set of major (main heading/subheading) pairs I={[MH/SH]*)}. The collection is detailed in Table 1.

Resources	Documents	Items	Min	Max	Mean
Guidelines*	2,727	MH*	1	64	5.21
		MH/SH*	1	70	6.12
Patients*	3,272	MH*	0	25	1.63
		MH/SH*	0	30	1.82
Education*	3,610	MH*	0	25	2.22
		MH/SH*	0	34	2.73

Table 1. Description of the collections of documents.

For all contexts, the minimum support threshold was fixed to minsup=20 and the minimum confidence threshold was fixed to minconf=70% (Table 2). We obtained association rules between major main headings MH* in the first context where I={MH*} and between (MH/SH)* pairs for I={[MH/SH]*}. For the major resource types patients* and education* all association rules (100%) are between two MHs* and between (MH/SH]* *i.e.* one descriptor in the antecedent and one descriptor in the consequent. For the major resource type guidelines*, 24% of the rules are between more than two descriptors. The characteristics of documents may explain these results: average descriptors were from 1.63 to 2.22 for patients* and education* whereas they were from 5.21 to 6.12 for guidelines*.

	Context: item=MH*				Context: item= [MH/SH]*			
Resources	*Nb rules*	*ER Conf=1*	*AR Conf≥0.7*	*Nb pairs*	*Nb rules*	*ER Conf=1*	*AR Conf≥0.7*	*Nb pairs*
Guidelines*	50	12 (24%)	38 (76%)	38 (76%)	39	8 (20.51%)	31 (79.49%)	35 (76%)
Patients*	20	9 (45%)	11 (55%)	20 (100%)	19	8 (42.1%)	11 *(57.9%)*	19 (100%)
Education*	23	6 (26.09%)	17 (73.91%)	23 (100%)	25	13 (52%)	12 (48%)	25 (100%)

Table 2. Number of rules, number of exact rules (ER), approximative rules (AR), and number of pairs.

Another experiment was performed in the context of documents with the resource type guidelines* (2,727 documents) to obtain more complete association rules: we considered the descriptors MH and MH/SH pairs without allotted minor or major weight. We obtained a high number of association rules with a minimum support threshold, minsup=20 and a minimum confidence threshold, minconf=70% (Table 3). However, only 0.95% are between two MH and only 1.92% are between two (MH/SH) pairs. By reducing the confidence from 1 to 0.7 the number of rules between MH growths with a factor of 5, and it growths with a factor of 4.42 between MH/SH.

Items	Nb Rules	ER Conf=1	AR Conf≥0.7	Nb pairs
MH	35,454	6,990 (19.71%)	28,464 (80.29%)	338 (0.95%)
MH/SH	27,011	6,102 (22.6%)	20,909 (77.4%)	520 (1.92%)

Table 3. Association rules between MH and MH/SH in the context Guidelines*.

3.2.2 Categorization according to Specialties

Each catalogue resource was indexed according to the vocabulary (the terms being main headings, subheadings, and resource types as detailed in section 1.2.2). Using heuristics and the rule-based classification algorithm that we have previously developed (Darmoni *et al.*, 2006), the related specialties to a resource are deduced from the existing semantic links between metaterm and main heading, metaterm and subheading, and metaterm and resource type. These were then ranked according to their level of importance. The categorization algorithm was processed on the initial set of 11,373 documents. A document may belong to several specialties. Different contexts are prepared, depending on the itemsets.

		Item=MH			Item=[MH/SH]		
Speciality	**Documents**	**ER**	**AR**	**Total**	**ER**	**AR**	**Total**
Allergy	509	101	231	332	93	206	299
Cardiology	558	251	542	793	151	332	483
Oncology	644	154	329	483	119	358	477
Psychiatrics	515	76	337	413	57	155	212
Gastroenterology	501	85	300	385	96	248	344
Neurology	1 137	169	520	689	83	285	368
Environment	1 254	257	924	1 181	148	584	732
Diagnosis	883	465	1 218	1 683	112	312	424
Therapeutic	782	555	2 010	2 565	206	562	768
Paediatrics	906	1 116	5 629	6 745	205	634	839
	Total	**3 229**	**12 040**	**15 269**	**1 270**	**3 676**	**4 946**

Table 4. Number of rules *minsup*=20, *minconf*=70%, ER Exact rules, AR Approximate rules

The number of rules (Table 4) is nearly the same before categorization (section 3.1.1) but the rules are not the same as they don't have the same support and confidence measures. The documents are categorized, the number of documents in the collections are different, and therefore the measures of support and confidence for the same antecedent and consequent are different.

3.3 Weighted Association Rules

We explored the classical item-object relationships with the degrees 0 and 1 which traduce the presence or absence of the item in one object of the context. In our e-health documents, the descriptors are "major" (marked with a star "*") or "minor" terms. The weight of "major" has a factor of 2 compared to "minor". We generated association rules between "major" descriptors as in section 3.3.1. We considered I={MH*} in the first case, and I={[MH/SH]*]} in the second case.

The association rules that are extracted (Table 5) are related to the concerned speciality which is itself deduced from the major pairs (MH/SH)* and (MH)*.

		Item=MH*			Item=[MH/SH]*		
Speciality	**Documents**	**ER**	**AR**	**Total**	**ER**	**AR**	**Total**
Allergy	509	4	12	16	2	12	14
Cardiology	558	7	37	44	5	31	36
Oncology	644	2	13	15	0	20	20
Psychiatrics	515	1	8	9	0	3	3
Gastroenterology	501	4	34	38	2	12	14
Neurology	1 137	4	34	38	0	25	25
Environment	1 254	6	85	91	5	53	58
Diagnosis	883	7	36	43	4	36	40
Therapeutic	782	2	32	34	2	18	20
Paediatrics	906	6	90	96	4	61	65
	Total	**43**	**381**	**424**	**24**	**271**	**295**

Table 5. Number of rules *minsup*=20, *minconf*=70%, ER exact rules, AR approximative rules.

3.4 Text-Mining

Text-mining shares many characteristics with classical data-mining but it differs in the nature of the studied data. We performed an automatic indexing of the plain text of the documents using the InterMediaText tool of Oracle® which indexes textual contents to improve information retrieval into Web applications without storing the Web pages. The number of extracted association rules was very high for expert evaluation. However, these rules could be used in the information-retrieval process by automatic query expansion as they associate medical as well as non-medical terms used in the documents.

Speciality	**Documents**	**Exact Rules**	**Approximative Rules**	**Total**
Neurology	1 137	1 354	105 202	106 556
Environment	1 254	2 815	397 073	399 888

Table 6. Number of rules *minsup*=20, *minconf*=70%, I={term}

4 Evaluating Extracted Knowledge

Not all of the association rules extracted are evaluated: according to the context extraction and the itemset I there are more or less association rules. The more the collection is specialized, and the itemset size is reduced, the less we have association rules to evaluate. In the following we consider only the collections that are categorized according to the users and according to the specialties.

4.1 Evaluation for the Categorized Documents according to Users

As defined, an interesting association rule confirms or states a new hypothesis (Fayyad *et al.*, 1996). Here, we proposed to combine background domain knowledge with simple statistical measures used traditionally in association rules mining for evaluation. We considered several cases of interesting association rules according to relations between MeSH headings. As these relations are defined between two main headings and between two subheadings, we considered only the association rules between two elements. Hence, an interesting existing association rule could associate: a (in)direct son and its father (relation FS); two descriptors that belong to the same hierarchy (same (in)direct father) (relation B); two descriptors with See Also relation (relation SA). These rules are automatically classified thanks to MeSH structure. The other rules that satisfy the minsup and minconf are then considered as «new» interesting association rules.

Exact association rules, except for collection patients*, are mostly new interesting rules: from 62.5% to 99.86%. Therefore, existing rules are mainly from the patients* collection: 77.77% for MH* and 75% for MH/SH*. Approximative rules, except for the guidelines* collection with items MH and MH/SH pairs, are mostly existing interesting rules: from 58.07% to 78.73%. New interesting rules are between MH and MH/SH from the collection guidelines*: 99.73% for MH and 99.52% for MH/SH (Table 7).

Subjective interest measures are based on expert knowledge about the data, *i.e.* that of physicians and medical librarians in this context. New interesting rules for the contexts MH* and (MH/SH)* pairs are evaluated manually. 93.75% (resp. 84.78%) of the interesting new rules with conf=1 (resp. conf≥0.7) between major descriptors are validated.

		Exact rules: Confidence=1				***Approximative rules: Confidence≥0.7***			
		Existing knowledge			*New*	*Existing knowledge*			*New*
		FS	*B*	*SA*		*FS*	*B*	*SA*	
Patients*	MH*	0 0%	5 55.55%	2 22.22%	2 22.22%	2 18.18%	2 18.18%	4 36.36%	3 27.27%
	MH/SH*	0 0%	5 62.5%	1 12.5%	2 25%	2 18.18%	2 18.18%	3 27.27%	7 36.36%
Education*	MH*	1 16.66%	1 16.66%	0 0%	4 66.67%	2 11.76%	6 35.29%	3 17.64%	6 35.29%
	MH/SH*	1 7.69%	0 0%	1 7.69%	11 87.62%	2 16.76%	3 25%	2 16.76%	5 41.66%
Guidelines*	MH*	0 0%	0 0%	4 33.33%	8 66.67%	2 5.26%	7 18.42%	10 26.31%	12 31.57%
	MH/SH*	1 12.5%	1 12.5%	1 12.5%	5 62.5%	3 9.67%	3 9.67%	9 29.03%	13 41.93%
	MH	0 0%	2 0.03%	8 0.14%	6,980 99.86%	12 0.04%	37 0.13%	30 0.10%	28,382 99.73%
	MH/SH	6 0.1%	4 0.06%	7 0.11%	6,085 99.73%	25 0.12%	50 0.23%	27 0.13%	20,807 99.52%

Table 7. Association rules evaluation according to MeSH structure.

aging → aged
breast cancer/diagnosis → mammography
aids/prevention and control → condom
influenza vaccines→influenza/prevention and control
Turner syndrome ∧ child → human growth hormone ∧ growth disorders
obstetric delivery→ pregnancy
depression → depressive disorder
prostate cancer/surgery→ biopsy ∧ prostatectomy
amniocentesis → prenatal diagnosis ∧ chorionic villi sampling
opioids analgesics/administration and dosage → pain/drug therapy

Figure 2. Examples of association rules evaluated as new interesting ones.

4.2 Evaluation of the Categorized Documents according to Specialties

All of the extracted rules (n=294) were evaluated as in 4.1 according to MeSH relations. The rules that describe the existing relationships are automatically classified thanks to the MeSH thesaurus (n=52). This left 242 rules to be analyzed manually. We obtained the following results among the 178 interesting association rules:

New rules (NW): 70.78 % (42.85% of the total)
See Also (SA): 13.48 % (08.16% of the total)
Same hierarchy (B): 09.55 % (05.78% of the total)
Father-Son (FS) 06.18% (03.74% of the total)

Speciality	NW	SA	B	FS	Other
Allergy	12	1	0	1	0
Cardiology	26	3	2	1	4
Oncology	17	1	1	0	1
Psychiatrics	0	1	0	0	2
Gastroenterology	5	1	0	0	8
Neurology	8	4	1	0	12
Environment	18	4	7	5	24
Diagnosis	32	3	0	1	3
Therapeutic	3	3	3	1	10
Paediatrics	5	3	3	2	52
Total	**126**	**24**	**17**	**11**	**116**

Table 8. Association rules evaluation.

As described in Figure 1, the association rules are re-injected in the information retrieval process. Other scenarios of exploitation of the association rules are detailed in the following section.

5 Exploiting Extracted Knowledge

5.1 Query Expansion: a Solution for Information Retrieval

Our objective was to use the numerous association rules that we extracted from the CISMeF database into the information-retrieval process by query expansion. We use Interactive Query Expansion with the seeker. For example, the association rule *breast cancer→mammography* is extracted from the corpus because the keywords *breast cancer* and *mammography* are frequently used together to index the documents. This association rule is as a "new" one because it doesn't exist in the domain knowledge which is, in our case, a structured terminology of the medical domain. When applying the association rule *breast cancer→mammography* on a query containing the term "breast cancer", an IQE proposes to the user electronic documents related to "mammography" to complete the search. In medicine and health-related information, (Kahng *et al.*, 1997) have already investigated an efficient algorithm for association-rule mining using the MeSH thesaurus. They adopted a MeSH-indexed representation of MEDLINE records, but the evaluation of the interest of the mined associations with respect to the task of PubMed retrieval improvement was not considered by the authors. In (Prince & Roche, 2009) many other works on information retrieval and query expansion in the biomedical domain are presented.

In the literature, a number of methods for performing query expansion have been developed. The solutions given are based mainly on two approaches. The first is the augmentation of query terms to improve the retrieval process without user intervention. The second is the suggestion of new terms to the user which can to be added to the original query to guide the search towards a more specific document space. The first case is called automatic query expansion (Buckley *et al*., 1994, Gauch & Smith 1993) whereas the second case is called semi-automatic query-expansion (Peat &Willet, 1991, Vélez *et al.*, 1997). (Magennis & van Rijsbergen, 1997, Ruthven 2003), as well as others, tried to evaluate and compare the efficiency of the two methods. Despite the fact that their experiments were based on simulations and not on real human users in most of the cases, the results of the experiments showed that the interactive query expansion method gave more control to the searcher who knows her utility better than any automated system.

Early methods involved extracting terms from thesauri (Gauch &Smith, 1991, Voorhees, 1993). However, as these proved to be labor-intensive, researchers turned to methods such as lexical co-occurrence (Vechtomova, 2003) and clustering (Jones, 1971, de Loupy *et al.*, 1998, Leuski, 2001, Shamim Khan & Khor, 2004). Lexical co-occurrence is the process of developing relationships between words based upon their co-occurrence in documents. In clustering, documents that share a significant number of terms are grouped together and representative words from each cluster are used for expansion of the original query. However, most systems that used clustering for query expansion reported rather pessimistic conclusions on their performance (Eftihimiadis, 1996). The similarity of the method proposed here with these methods with lexical co-occurrence and clustering is that the source, which provides the candidate terms for expansion, is the set of the retrieved documents as opposed to some knowledge structure as in thesaurus-based approaches. As a consequence, if the user chooses terms that do not yield results from the expected domain, the terms suggested by the query-expansion algorithm are unlikely to be helpful to the user. In (Grabar *et al.*, 2003, Soualmia & Darmoni 2003) we proposed the use of lexical variants for query expansion in CISMeF and obtained good results.

Relevance feedback (Harman, 1988, Robertson *et al.*, 1986, Mitra *et al.*, 1998) is another way of performing query expansion. In this method, users submit a query which yields an initial set of results. From this set, they select a number of documents that are thought to be relevant. The system expands the query based upon the terms in the selected documents. Based on the set from which the terms are selected, different cases can be distinguished. If a document collection is considered as a whole from which the terms are extracted to be added to the query, the technique is called global analysis (Xu & Croft, 1996). However, if query expansion is performed based on the documents retrieved from the first query, the technique is denominated local analysis, and the set of documents is called local set (Croft & Thompson, 1987). Local analysis can also be classified into two types: local feedback and global feedback. Local feedback adds common words from the top-ranked documents of the local set (Mitra *et al.*, 1998). These words are identified sometimes by clustering the document collection. We can include the relevance-feedback process in this group, because the user has to

evaluate the top ranked documents from which the terms to be added to the query are selected. On the other hand, local context analysis combines global analysis and context local feedback to add words based on relationships of the top-ranked documents. The calculus of co-occurrences of terms is based on passages (text windows of fixed size), as in global analysis, instead of complete documents. In general, the authors show that local analysis performs better than global analysis.

Despite the significant improvement in the quality of results the method produces, research carried out by (Spink *et al.*, 2001) shows low use of the relevance feedback facilities in search engines. The low use should not necessarily be attributed to the interactive nature of this method. Sometimes when a user has already found a set of relevant documents they may not wish to expand the query. Furthermore, relevance feedback algorithms are only useful when relevant documents are returned within the top ranked documents of the results.

Recent methods to perform query expansion with promising results involve mining user logs (Cui *et al.*, 2003, Liu, 2007) and constructing user profiles (Nikravesh *et al.*, 2002). Another study on logs in PubMed for searching biomedical and life-science literature online has been performed by (Lu & Wilbur, 2009). There has also been work that utilizes fuzzy association rules (Marìn-Bautista *et al.*, 2004).

5.2 Evaluating Query Expansion based on Association Rules

Many ways of navigation and information retrieval are possible in the catalogue. The most used is the *simple search* (free text interface). It is based on the subsumption relationships. A query (a word or an expression) can be matched with an existing concept. In this case, the result of the query is the union of the resources that are indexed by the concept, and the resources that are indexed by the concepts it subsumes, directly or indirectly, in all of the hierarchies it belongs to. The co-occurrence tools developed for information retrieval bring the terms which frequently appear in the same documents closer together. These terms thus have a semantic proximity. This technique was used very early to allow query expansion. By analogy, association rules may be exploited in a search engine by carrying out an interactive query expansion. This helps the user to formulate his query by using the result of a query to reformulate, filter and re-orientate the query by exploiting the terms related to his query terms. In fact, the user can select suggested terms sets to add them to his initial query. It is useful in the case of non-precise information needs. IQE requires user implication. We developed a web-based evaluation tool of the IQE used by a set of 500 users[8] which are subscribers of the weekly letter "What's new" of CISMeF. 20 queries, and for each one a set of medical terms derived from the extracted association rules were proposed. The evaluation was performed thanks to a Likert scale. The results (76% of the users were satisfied by the propositions) demonstrate the usefulness of this approach. An expanded query by association rules contains more related terms. By using the vectorial model, for example, more documents will be located and this treatment increases recall (Guarino *et al.*, 1999). In addition, association rules are indication on the possible definition of a term or its context environment.

5.3 Correcting Indexing

Our initial corpus was manually indexed. According to the indexing policy, the more precise descriptor should be used to index a document, *i.e.* the descriptor at the lower level in the hierarchy. However, 1,466 documents contained descriptors that had father-son relationships. For example, a document was indexed by the main headings *trisomy* and *chromosome aberrations*, but *trisomy* is a *chromosome aberration*. 478 documents were indexed by related subheadings and were associated with the same keyword. For example, *diabetes/therapy* and *diabetes/pathology* describe the same document whereas *pathology* is an *anatomy-histology*. This may explain the proportion of existing interesting associations between FS and a correction will be proposed to the indexers. CISMeF has been criticized by (Abad Garcia, 2005) because "failure on precision may be due to exhaustive

[8] http://www.chu-rouen.fr/enquete/

indexing". However, in our case of knowledge acquisition, this exhaustive indexing is very useful. This may explain the proportion of existing associations. Correction should be proposed to the indexers.

5.4 Modeling Expert Rules

The main return on extraction and evaluation was modeling and formalization of rules between (MH/SH) pairs based on observations. The pattern of the rule *hepatitis/prevention→hepatitis vaccines* is used to model *dysentery bacillary/prevention→shigella vaccines*. 463 rules were modeled. Formalization concerns different cases and contexts for retrieval and indexing.

We defined a pruning rule as taking the following form:

$$MH_1/SH_1 \xrightarrow{--} MH_2$$

It states that MH_1/SH_1 should be replaced by the main heading MH_2. For example *abdomen/radiography* $\xrightarrow{--}$ *radiography abdominal* states that when a user is searching for abdomen/radiography, the query should be replaced by radiography abdominal. It also states that when indexing, radiography abdominal should be preferred to the association (abdomen/radiography).

An adding rule was defined in the following form:

$$MH_1/SH_1 \xrightarrow{++} MH_2/SH_2$$

It states that the pair MH_2/SH_2 should be added to the pair MH_1/SH_1. For example, *appendectomy* $\xrightarrow{++}$ *appendicitis/surgery* states that the pair appendicitis/surgery should be added to queries (or to document description when indexing) containing the main heading appendectomy.

Different categories of main headings have been defined for the pruning and adding rules according to the hierarchies of the main headings. A complete list of these rules can be found in (Soualmia, 2004).

6 Conclusions

In the genomic model, text-mining activity is increasing (An *et al.*, 2009, Ananiadou & McNaught, 2005). In (Bodenreider *et al.*, 2005) co-occurrences between Gene Ontology (GO) terms are analyzed on annotation databases and association rules are mined to identify only pairs of related GO terms. Our A-Close algorithm adaptation generates all the valid non-redundant association rules composed by minimal antecedents and maximal consequents, *i.e.* the most relevant association rules which are particularly interesting in retrieval and indexing. In (Burgun & Bodenreider, 2001) the authors compared symbolic knowledge provided by the UMLS Semantic Network and the inter-concepts relationships in the MeSH with the UMLS source of co-occurrences between major MeSH terms in MEDLINE.

There has been an interesting study that uses fuzzy association rules (Marìn-Bautista *et al.*, 2004) to perform query expansion. The authors start from an initial set of documents retrieved from the Web and extract association rules from the texts between terms as we have performed in this chapter. The difference is that they model the problem using a fuzzy extension, the presence of the term in a text is determined with a value between 0 and 1. In this chapter, we used only the presence (1) or absence (0) of a concept as in the Boolean model. However, the authors do not give any evaluation about the returned fuzzy association rules, and the meaning of these fuzzy rules is not quite intuitive. Nevertheless, this approach is similar to the "weighted association rules" using major and/or minor concepts that are present in the indexed documents.

Association rules are more complete than co-occurrence measures between pairs of concepts. One of the challenging issues is the overabundance of associations that may be discovered, as described by (Berardi *et al.*, 2004). A-Close generates all of the valid non-redundant association rules composed by minimal antecedents and maximal consequents. Evaluation is processed in two steps. First, is the selection of all the most

informative rules. This is followed by the classification of the rules according to the taxonomy structure of MeSH to filter existing associations. Only the most frequent rules that are not classified are presented to the expert for a final evaluation. This method combines statistical measures and background domain knowledge.

With respect to the application of association-rule mining to the biomedical domain, especially biomedical literature, (Hristovski *et al.*, 2001) discovered new associations involving a concept of interest where the novelty of the relation was evaluated by matching transitive associations. They first find all the concepts Y related to the concept of interest X, then all the concepts Z related to Y. Finally they check whether X and Z appear together in the biomedical literature. If they do not appear together, they conclude that the system has discovered a potentially new relation that will be evaluated by the user. However, the search of associations is constrained to associations involving only two terms. In contrast this chapter has shown that more complex association rules involving an unknown number of medical concepts are quite useful in CISMeF.

Association rules are used in retrieval by query expansion (automatic and interactive) and enriching queries with new knowledge. As exact rules state that the antecedent and the consequent are at the same time in all documents, these kinds of rules should be used in automatic query expansion. By analogy, as approximative rules state that the antecedent and the consequent are at the same time in only some documents, these kinds of rules should be used in interactive query expansion. However, these expansions work only in the case of queries that return documents. Association and expert rules (pruning and adding rules) can be translated in the form of automata for processing automatic indexing of raw text documents. Formalized association rules could improve the power of reasoning based on MeSH-OWL (Soualmia *et al.*, 2004). Finally, as the CISMeF controlled vocabulary is moving from MeSH to multi-terminologies (Darmoni *et al.*, 2010), mining association rules over several terminologies and classifications should help to induce several mapping relations (Ghasvinian *et al.*, 2009) between these sources of biomedical knowledge.

References

Abad Garcia, F. (2005). A comparative study of six European databases of medically oriented Web resources. *Journal of the Medical Library Association*, 93(4), 467–479.

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the VLDB Conference* (pp. 478−499).

Agrawal, R., Imielinski, T. & Swami, A.N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 207–216).

An, L., Obradovic, Z., Smith, D. & Bodenreider, O. (2009). Mining association rules among gene functions in clusters of similar gene expression maps. In *Proceedings of workshop on Data Mining in Functional Genomics* (pp.254–259).

Ananiadou, S. & Mc Naught, J. (2005). *Textmining for biology and biomedicine.* Artech House.

Aronson, A.R. & Rindflesch, T.C. (1997). Query expansion using the UMLS Metathesaurus. In *Proceedings of AMIA Annual American Medical Informatics Association Conference* (pp. 485–489).

A grammar of Dublin Core. *D-Lib Magazine*, 6(10).

Berardi, M., Lapi, M., Leo, P., Malerba, D., Marinelli, C. & Scioscia, G. (2004). A data mining approach to PubMed query refinement. In *Proceedings of the 15th International Workshop on Database and Expert Systems Applications* (pp.401–405).

Bodenreider, O., Aubry, M. & Burgun, A. (2005). Non-lexical approaches to identifying associative relations in the Gene Ontology. In *Pacific Symposium on Biocomputing* (pp. 91–102).

Brin, S., Motwani, R., Ullman, J.D. & Tsur, S. (1996). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 255–264).

Buckley, C., Salton, G., Allan, J. & Singhal A. (1994). Automatic query expansion using SMART: TREC3. In *Proceedings of the Third Text REtrieval Conference* (pp. 69–80).

Burgun, A. & Bodenreider, O. (2001) Methods for exploring the semantics of the relationships between co-occuring UMLS concepts. In *Proceedings of Medinfo Conference* (pp.171–175).

Croft, W.B. & Thompson, R.H. (1987). I^3R: a new approach to the design of document retrieval systems. *Journal of the American Society Information Science*, 38 (6), 389–404.

Cui, H., Wen, J.R. & Ma, W.Y. (2003). Query expansion and classification by mining user logs. *Knowledge and Data Engineering*, 15 (4), 829–839.

Darmoni, S.J., Thirion, B., Leroy, J.P., Douyère, M., Lacoste, B., Godard, C., Rigolle, I., Brisou, M., Videau, S., Goupy, E., Piott, J., Quéré, M., Ouazir, S. & 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.

Darmoni, S.J.; Névéol, A.; Renard, J.M.; Gehanno, J.F.; Soualmia, L.F.; Dahamna, B. & Thirion, B. (2006). A MEDLINE categorization algorithm. *BMC Medical Informatics and Decision Making*, 6, 7.

Darmoni, S.J., Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., Soualmia, L.F. & Thirion, B. (2010) Health Multi-Terminology Portal: semantics added-value for quality-controlled health gateway. Submitted to *Journal of Biomedical Semantics*.

de Loupy, C., Bellot, P., El-Beze, M. & Marteau, P.F. (1998). Query expansion and classification on retrieved documents. In *Proceedings of the Text REtrieval Conference* (pp. 382–389).

Desmontils, E. & Jacquin, C. (2002). Indexing a Web site with a terminology oriented ontology. In: *The Emerging Semantic Web*, 181–197.

Duquenne, V. & Guigues, J.L. (1986) Famille minimale d'implications informatives résultant d'un tableau de données binaires. *Mathématiques et Sciences Humaines,* 24(95), 5−18.

Eftihimiadis E. (1996). Query expansion. *Annual Review on Information Systems Technolgy*, 31, 121–187.

Fayyad, U.M., Piatetsky-Shapiro, G.P., Smyth, P. & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining.* American Association of Artificial Intelligence Press.

Ganter, B. & Wille, R. (1999). *Formal Concept Analysis: Mathematical foundations.* Springer.

Gauch, S. & Smith, JB. (1993). An expert system for automatic query reformulation. *Journal of the American Society Information Sciences,* 44 (3), 124–136.

Ghazvinian, A., Noy, N.F., Jonquet, C., Shah, N. & Musen M.A. (2009). What four million mappings can tell you about two hundred ontologies. In *Proceedings of the 8th International Semantic Web Conference* (pp.229–242).

Grabar, N., Zweigenbaum, P., Soualmia, L.F. & Darmoni, S.J.(2003). Matching controlled vocabulary words. In *Proceedings of the MIE Conference* (pp.445–450).

Gras, R., Kuntz, P. & Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines*, 155, 9–29.

Grefenstette, G. (1997). Short query linguistic expansion techniques : palliating one-word queries by providing intermediate structure to texts. *Lecture Notes in Computer Science*, 1299, 97–114.

Guarino, N., Masolo, C., & Vetere, G. (1999). OntoSeek: content-based access to the Web. *IEEE Intelligent Systems*, 14(3), 70–80.

Harman, D. (1998). Towards interactive query expansion. In *Proceedings of the 11th ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 321–331).

Hou, J., & Zhang, Y. (2003). Effectively finding relevant web pages from linkage information. *IEEE Transactional Knowledge Data Engineering*, 15(4), 940–951.

Hristovski, D., Stare, J., Peterlin, B. & Dzeroski, S. (2001). Supporting discovery in medicine by association rule mining in Medline and UMLS. In *Proceedings of Medinfo Conference* (pp.1344–1348).

Jones, K.S. (1971). *Automatic Keyword Classification for Information Retrieval*, Butterworth London.

Kahng, J., Liao, W.H.K. & McLeod, D. (1997). Mining generalized term associations: count propagation algorithm. In *Proceedings of the KDD workshop* (pp.203–206).

Koch, T. (2000). Quality controlled subject gateways: definitions typologies, empirical overview. *Online Information Review*, 24(1), 24–34.

Leuski, A. (2001). Evaluation document clustering for interactive information retrieval. In *Proceedings of the 10th International Conference on Information and Knowledge management* (pp. 33–40).

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents and usage data*. Springer.

Lu, Z. & Wilbur, W.J. (2009). Improving Accuracy for identifying related PubMed queries by an integrated approach. *Journal of Biomedical Informatics*, 42, 831–838.

Magennis, M. & van Rijsbergen, C.J. (1997). The potential and actual effectiveness of interactive query expansion. In *Proceedings of the 20th annual international ACM SIGIR Conference on Research and development in Information Retrieval* (pp. 324–332).

Mannila, H., Toivonen, H. & Verkamo, A.I. (1994). Efficient algorithms for discovering association rules. In *Proceedings of the KDD workshop* (pp.181−192).

Marìn-Bautista, M.J., Sànchez, D., Chamorro-Martìnez, J., Serrano, J.M. & Vila, M.A. (2004). Mining Web documents to find additional query terms using fuzzy association rules. *Fuzzy Sets and Systems*, 148(1), 85–104.

Mitra, M., Singhal, A & Buckley, C. (1998) Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp 206–214).

Nikravesh, M., Loia, V. & Azvine, B. (2002). Fuzzy Logic and the Internet (FLINT): Internet World Wide Web and search engines. *Soft Computing*, 287–299.

Park, J.S., Chen, M.S. & Yu, P.S. (1995). An Effective hash based algorithm for mining association rules. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp.175–186).

Pasquier, N., Bastide, Y. Taouil, R. & Lakhal, L. (1998). Pruning closed itemset lattices for association rules. In *Proceedings of Bases de Données Avancées* (pp.177−196).

Pasquier, N., Taouil, R., Bastide, Y., Stumme, G. & Lakhal, L. (2005) Generating a condensed representation of association rules. *Journal of Intelligent Information Systems*, 24(1), 29–60.

Peat, H.P., & Willet, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society Information Science,* 42(5), 378–383.

Prince, V. & Roche, M. (2009). *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration.* IGI Global.

Robertson, S.E., Thompson, C.L., Makaskill, M.J. & Dovey, J.D. (1986) Weighting ranking and relevance feedback in a front end system. *Information Science*, 12(1–2), 71–75.

Ruthven, I. (2003). Reexamining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 213–220).

Shamim Khan, M. & Khor, S.W. (2004). Web Document clustering using a hybrid neural network. *Applied Soft Computing*, 4(4), 423–432.

Soualmia, L.F. (2004). *Etude et Evaluation d'Approches Multiples d'Expansion de Requêtes pour une Recherche d'Information Intelligente : Application au Domaine de la Santé sur l'Internet*. PhD thesis, INSA de Rouen.

Soualmia, L.F. & Darmoni, S.J. (2004). Combining knowledge-based methods to refine and expand queries in medicine. In *Proceedings of the conference on Flexible Query-Answering Systems* (pp.243–255).

Soualmia, L.F. & Darmoni, S.J. (2005). Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. *International Journal of Medical Informatics*, 74(2-4), 41–150.

Soualmia, L.F., Barry, C. & Darmoni, S.J. (2003). Knowledge-based query expansion over a medical terminology oriented ontology. In *Proceedings of the 9th Conference on Artificial Intelligence in Medicine* (pp.209–213).

Spink, A., Wolfram, D. & Jansen, B.J. & Saracevic, T. (2001). Searching the web: the public and their queries. *Journal of the American Society Information Science Technology*, 52(3), 226–234.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing and Management*, 32(4), 431–443.

Thoivonen, H. (1996). Sampling large databases for finding association rules. In *Proceedings of the 22nd VLDB Conference* (pp. 134–145).

Vechtomova, O., Robertson, S. & Jones, S. (2003). Query expansion with long-span collocates. *Information Retrieval*, 6(2), 251–273.

Vélez, B., Weiss, R., Sheldon M.A. & Gifford, D.K. (1997). Fast and effective query refinement. In *Proceedings of the 20th ACM Conference on Research and Development in Information Retrieval.*

Voorhees, E.M. (1993). Using WordNet to disambiguate word senses for text retrieval. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 171–180).

Xu, J. & Croft, X.B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th ACM Conference on Research and Development in Information Retrieval* (pp. 4–11).