Ubiquity: Technologies for Better Health in Aging Societies
A. Hasman et al. (Eds.)
IOS Press, 2006

Strategies for Health Information Retrieval

Lina F.SOUALMIA[a,b], Badisse DAHAMNA[a,b], Benoît THIRION[a,b] , Stéfan J.DARMONI[a,b1]
[a] *CISMeF Team, Rouen University Hospital, Rouen, France*
[b] *GCSIS, PSI Laboratory, INSA & Rouen University, Mt St Aignan, France*

Abstract. Background: The amount of health data accessible on the Web is increasing and Internet has become a major source of health information. Many tools and search engines are available but medical information retrieval remains difficult for both the health professional and the patients. Objective: In this paper we describe heuristics that aim at matching as much as possible queries with the content of the documents in the context of the CISMeF catalogue (Catalogue and Index of Health Resources in French) and its Doc'CISMeF search tool. The queries are represented by terms and the content of the documents is indexed by a terminology based on the MeSH thesaurus. Results: Several operations are performed to match the terms of the terminology: natural language processing techniques on multi-words queries, phonemisation, spelling correction, plain text search with adjacency etc… Each one is tested to evaluate its contribution in matching the terminology and the indexed documents. Conclusion: The implemented heuristics contribute significantly with good results in maximising as much as possible the recall of the Doc'CISMeF search tool.

Keywords: Information Retrieval, MeSH, Internet.

1. Introduction

Internet is a major source of health information. Many people including health professionals, patients and general public, now search health care information on the Web. The access to structured medical information remains difficult when using directories such as Yahoo or search engines such as Google. Therefore many tools and applications have been developed for the healthcare professionals and until recently bibliographic databases such as Medline were available only to experts [1]. In medical information retrieval, there is a need of support. In this context, the objective of CISMeF [2] (Catalogue and Index of Health Resources in French) is to assist the health professional during the search of electronic information available on the Internet. The CISMeF health gateway describes and indexes high quality-controlled information resources written in French. We present in this paper some strategies and heuristics to match as much as possible the users' queries with the French adaptation of the MeSH and thus to reduce the silence of the system.

[1] Corresponding Author : Stéfan Darmoni, CHU de Rouen, 1 rue de Germont, 76031 Rouen Cedex France, stefan.darmoni@chu-rouen.fr, http://www.chu-rouen.fr/cismef.

2. Methods

2.1. CISMeF Metadata and Terminology

Since February 1995, the CISMeF catalogue describes and indexes a large number of health information resources (*n=15,090; Dec. 2005*). Each catalogue resource is indexed by its container using *metadata* used to improve information retrieval [3] and by its contents using the *terms* of the CISMeF terminology. CISMeF metadata are described in [2, 4]. The CISMeF terminology 'encapsulates' the French version of the MeSH thesaurus [5]. However, the MeSH was originally intended to index scientific articles for the Medline database. In order to customise it to the broader field of health Internet resources we have developed several enhancements to the MeSH since 2000. In addition to MeSH keywords and subheadings, the concepts of metaterms (*n=105*) and resource types (*n=257*) were added. As defined by the Dublin Core Metadata Initiative [4], a resource type is used to categorize the nature of the content of the resource. MeSH (term/subheading) pairs describe the topic of the resource. A metaterm (in most cases MeSH terms) is a medical specialty or a biological science, which has semantic links with one or more MeSH terms, subheadings and resource types (e.g. *cardiology, bacteriology)*. The keywords, headings and resource types are organised hierarchically. Compared to the publication types of Medline, the CISMeF resource types are more diverse, with specific resource types dedicated to electronic health resources (e.g. *association*, *clinical guidelines*). Nonetheless, the MeSH thesaurus largely inspires this list as 187 resource types (76%) are deliberately ambiguous because they are also MeSH terms (e.g. *magnetic resonance imaging*). The objective of this ambiguity is to maximise the number of search results (the Doc'CISMeF search the answers for the MeSH term and for the resource type) when the user query contains this kind of ambiguous term. Furthermore, to be as close to a standard as possible, 28 resource types (11%) are also Medline publication types (e.g. *technical report*). Each metaterm has a semantic link with one or more keywords, headings and resource types. Each term can have a set of synonyms and can belong to several trees.

Many ways of navigation and information retrieval are possible in the catalogue [6]. The most used is the *simple search* (free text interface). It is based on the subsumption relationships. If the query can be matched with an existing term of the terminology, thus the result is the union of the resources that are indexed by the term, and the resources that are indexed by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. If the query cannot be matched, the search is done over the other fields of the metadata and in a worse case a full-text search is carried out. Contrary to Medline, the resource types and the metaterms were voluntary made ambiguous to maximize the recall (e.g. in the query *guidelines in virology*, *virology* will be recognized as a metaterm (instead of a term) and *guidelines* will be recognized as both the term and the resource type because we assume most of end users confuse content and container). We propose in the following, some enhancements for query matching.

2.2. Basic Natural Language Processing (NLP) techniques

The basic natural language processing steps developed in [7] are founded on the following operations.

***Query segmentation*:** the query is segmented in words thanks to a list of characters and *string tokenizers*, composed by all non-alphanumerical characters (e.g.: * *$,!§;*|*@*).

***Character normalisations*:** we apply two types of characters normalisation at this step. The MeSH terms are in the form of non-accented upper case characters. Nevertheless, the terms used in the CISMeF terminology are in mixed-case and accented [8]. (1) *Lowercase conversion*: all the uppercased characters are replaced by their lowercase version; "*A*" is replaced by "*a*". This step is necessary because the controlled vocabulary is in lowercase. (2) *Deaccenting*: all accented characters (*"éèêë"*) are replaced by non-accented (*"e"*) ones. Words in the French MeSH are not accented, and words in queries can be accented or not, or wrongly accented (*h**è**patite"* instead *"h**é**patite"*).

***Stop words*:** we eliminate all the stop words (such as *the, and, when*) in the query. Our stop words list is composed by 1,422 elements [9].

***Exact expression*:** we use regular expressions to match the 'exact' [10] expression of each word of the query with the terminology. This step allows taking into consideration the complex terms of the vocabulary and avoiding some inherent noise generated by the truncations. The query *'**sida**'* is matched with the terms *'lymphome lié **sida**'* and *'**sida** atteinte neurologique'* but not with the terms *'gluco**sida**ses'*, *'agra**sida**e'*..

Phonemisation: the study [9] of the users' queries have shown that a great percent of no answer result from spelling mistakes. We have developed a Word phonemisation module that converts a word to its French phonemic transcription: e.g. the query *alzaymer* is replaced by the reserved term *alzheimer*, which is, has the good orthography. If the phonemic transcription of the query couldn't be matched, a spelling correction is proposed.

Spelling correction (optional)*:* the module of spelling correction propose to the user the reserved term that has a similar phonemic (according to a score and taking into account the possible characters' inversions) with a reserved term. The query is not replaced and a correction suggestion is proposed to the user.

Bag of word: this algorithm [7] searches in the user's query the greatest set of words that corresponds to a reserved term. The reserved terms bags are formed iteratively until no possible combinations. The query *'therapy of the breast cancer'* gives two reserved words: '*therapeutics'* and '*breast cancer'* (*therapy* is a synonym of the reserved term *therapeutics).*

2.3. Heuristics to return documents from the database

The complex terms matching is more requiring than simple terms matching. The CISMeF team editorial policy concerning the queries' rewriting consists in maximising as much as possible the Doc'CISMeF recall. This approach is mainly due to the size of the CISMeF's corpus (n=15,090 vs. several million in the MEDLINE database). When all the terms of the query couldn't be recognized as reserved terms, we have implemented 5 main heuristics for information retrieval that was largely inspired by the PubMed heuristics developed to access the MEDLINE bibliographic database.

Step 1. The reserved terms: The process consists in recognizing the user query expression. If it matches a reserved term of the terminology, the process stops, and the answer of the query is the union of the resources that are indexed by the term, and the resources that are indexed by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. If it doesn't match a reserved term, the query is segmented to seek if it contains one ore more reserved terms. The query '*enfant asthme'* is replaced by (*enfant.mr* AND *asthme.mr*), where *enfant* and *asthme* are reserved terms (*mr*). The reserved terms are matched thanks to the *bag of words* algorithm independently of the words query order.
Step 2. The documents' title: The search is performed over the other fields of the metadata. The field *title* of the documents is considered in priority. The stop words are eliminated and the search is realised over the union of the words of the query with a truncation (*) at the right in the field title (*ti*), as the following : $word_1$*.*ti* AND $word_2$*.*ti* for a 2-words query.
Step 3. Mixing the reserved terms and the titles: The system seeks if some words are reserved terms. A new Boolean query is generated with the fields reserved term (*mr*), if the word is a reserved term, and title (*ti*) if not. The query '*allergie infantile'* is replaced by the Boolean query : (*allergie.mr* AND *infantile.ti*).
Step 4. Mixing the reserved terms, all fields and adjacency in the titles : The search is processed over all the fields (*tc*) of the documents' metadata for the words that couldn't be recognized as reserved terms UNION the initial query processed over all the fields with adjacency (*at)* at *n* words with *n*=5×(nb words of the query–1). The query '*les problèmes respiratoires des enfants'* is replaced by the Boolean query [*(enfant.mr* AND *problemes.tc* AND *respiratoires.tc)* OR *(problemes respiratoires enfant.at)*]. In this query, the word *enfant* is recognized as a reserved term because it has the same sonority as the reserved term *enfants*. The words *problèmes* and *respiratoires* are searched over all the fields and the initial query *problèmes respiratoires enfants* is searched over all the fields with adjacency of 10 which means that these 3 words shouldn't be distant at more than 10 words.
Step 5. Mixing the reserved terms, all fields and adjacency in the plain texts : A plain text search over the documents with adjacency (*ap*) of *n* words with *n* = 10×(nb words of the query – 1) is realised. The query '*bronchite asthmatiforme'* is replaced by the Boolean query (*bronchite asthmatiforme.ap*) where the words *bronchite* and *asthmatiforme* shouldn't be distant at more than 10 words in the plain texts of the documents. The plain text search is possible with the Intermedia Text tool of Oracle® 9.i. which required a pre-treatment of the CISMeF corpus (~72 hours).
An intuitive scale of interpretation (from Step 1 to Step 5) is available to inform the users about their queries operations and rewritings.

2.4. Evaluation methodology

To evaluate the strategies that we have implemented, we have extracted from the Doc'CISMeF http log server of the first version of the search engine a set of 250 queries that gave no answer in the month of September 2002. The contribution of each treatment is measured. The difference with the evaluation method we have developed [9] lives in the matching of the terms with the terminology whereas here we want to match the queries with the documents of the catalogue.

3. Results

Among the 250 "difficult" queries (with no answer in 2002), the results show that a total of 176 (65%) queries give now answer(s) and 74 (35%) still not. The study of the set of the queries that give no answer shows that different reasons are possible: (a) 8 queries (10%) are matched but there is no CISMeF resources that corresponds to; (b) 27 (36%) are non corrected spelling errors returned as suggestions to the users; (c) 18 (24%) have no relationships with the medical domain and (d) 21 are unknown words. However, thanks to the heuristics (a) 27% of the queries are matched with a reserved term (Step 1); (b) 7% are matched with the title of the documents (Step 2); (c) 4% are matched with a mix of the reserved terms and the titles (Step 3); (d) 10% are matched with a mix of reserved terms, all fields and adjacency in the title of the documents (Step 4) and (e) 17% are matched with a mix of reserved terms, all fields and adjacency in the plain text of the documents. The response time is acceptable. Steps 1 and 2 response time is less than one second. Step 5 average response time ranges from 2 to 3 seconds.

Table 1. Repartition of the queries matched with documents in the CISMeF database

Operation	Number of queries with documents in return	Percentage
Step 1	57	27%
Step 2	14	7%
Step 3	8	4%
Step 4	22	10%
Step 5	36	17%
Total	176/250	65%

4. Discussion

In this paper we have presented strategies to support health information seeking using the CISMeF information gateway in the case of free queries that don't match the controlled vocabulary, i.e. that give no answer from the corpus. Simple but essential treatments such as spelling correction are processed online. McCray [11] has also presented strategies for supporting health information seeking. The major difference is in the treatment of the query and specifically in its expansion. We think that this type of query expansion (by relaxing the query) and suggestions to the users may led them too much tasks in first, choosing the expanded query and then, in navigating through the documents of the expanded queries to seek the wanted information. The second problem lives in the exponential growth of the query when it is composed by several words. Another treatment seems to us not necessary at all: the expansion of each word of the query by a set of its synonyms, derivations, and inflections. If the query contains a synonym of a reserved term, it should be replaced by the reserved term, which is more precise especially in the context of indexed resources with a controlled vocabulary. The last point concerns the search mode itself which is based on plain text search vs. indexing terms which is much more precise.

Our strategies are relatively powerful as 65% of the queries with 0 answer in 2002 give at least one answer in 2006 (see Table 1). Among the 35% of the queries with 0 answer, only 36% (n=27) are spelling errors which are not corrected. We will then focus on this problem. Furthermore, a recent Spanish study comparing 6 European health catalogues has shown that CISMeF was ranked second after OMNI in terms of precision and recall, mainly because "failure on precision may be due to exhaustive indexing" [12]. This external judgment is definitively true: we deliberately focused on maximizing recall in terms of terminology and in terms of heuristics. This approach may be explained by the relative small size of the CISMeF corpus. This Spanish study will lead to a rather serious modification of the CISMeF editorial policy. New adds-on on the CISMeF terminology or heuristics will now focus on maximizing precision (in the near future, the default query for the step 1 (reserved term) will answer CISMeF resources indexed with a MeSH major (or starred) term). The CISMeF heuristics for health information retrieval tried to improve the PubMed heuristics although the size of his respective corpus is not the same. We have introduced the step 2 (search on title) because, based on the know-how of the CISMeF Chief Librarian, this step provides very precise answers. In the step 4 (search on all metadata fields), we have introduced a search with adjacency once again to be more precise. Finally, we have also introduced the step 5 (search on plain text). This step very similar to a Google search (but more precise thanks to the adjacency) is feasible for the CISMeF Catalogue because it indexes full text resources, which is not the case for the Medline database. Nonetheless, the PubMed Website would be able to apply the step 5 to a subset of journals indexed in Medline, in particular those of PubMed Central.

References

[1] Eysenbach G, Jadad AR. Consumer Health Informatics in the Internet Age. JMIR, 2001; 3(2): e-19.

[2] Darmoni SJ, Leroy JP, Thirion B, et al. CISMeF a Structured Health Resource Guide. Methods Inf Med 2000; 39(1): 30-35

[3] Hudgins J, Agnew G, Brown E. Guetting Mileage Out of Metadata: Applications for the Library. ALA 1999.

[4] Dekkers M, Weibel S. State of the Dublin Core Metadata Initiative. D-Lib Mag. 2003; V9 n° 40.

[5] Nelson SJ, Johnson WD, Humphreys BL. Relationships in Medical Subject Headings. Kluwer Publishers, 2001; 171-84.

[6] Darmoni SJ, Thirion B, Leroy JP et al.. A Search Tool Based on Encapsulated MeSH Thesaurus to Retrieve Quality Resources on the Internet. Medl Inform Internet Med 2001; 26(3):165-178

[7] Soualmia LF. 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; December 2004.

[8] Zweigenbaum P, Grabar N. Restoring Accents in Unknown Biomedical Words: application to the French MeSH thesaurus. IJMI 2002; 113-26.

[9] Douyère M, Soualmia LF, Névéol A, et al. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004;21(4): 253-61.

[10] Riloff E. Little Words can make Big Difference for Text Classification. 18th ACM SIGIR 1995. 130-36

[11] McCray AT, Ide NC, Loane RR, Tse T. Strategies for Supporting Consumer Health Information Seeking. Medinfo 2004; 1152-56.

[12] Abad Garcia F, Gonzalez Teruel A, Bayo Calduch P, et al.. A comparative study of six European databases of medically oriented Web resources. J Med Libr Assoc. 2005 Oct; 93(4): 467-79.