Correcting and Refining Users' Queries: the Contribution of Morphological Knowledge and Association Rules

Lina F. Soualmia
PSI Laboratory
INSA & Rouen University, CNRS FRE 2645
Place Emile Blondel, BP 68
76131 Mont Saint Aignan, France
Lina.Soualmia@chu-rouen.fr

Stéfan J. Darmoni
CISMeF team & L@STICS,
Rouen University Hospital & Medical School
1, rue de Germont
76031 Rouen, France
Stefan.Darmoni@chu-rouen.fr

Abstract

We propose in this paper to combine two knowledge-based methods to correct and to refine the user queries submitted over the CISMeF catalogue, in which the resources are indexed according to a structured terminology of the medical domain. There are two possible cases of queries: queries that don't match any term of the terminology and then give no answer, and queries that are too general and give too many answers. The first method consists of building and using morphological knowledge of the terms to correct the queries. The second method consists of extracting association rules between terms by applying a data mining technique over the indexed resources of the catalogue, the aim being to refine the queries. We show the preliminary results that we have obtained.

Keywords: Natural Language Processing, Data Mining, Information Retrieval, Medicine.

1 Introduction

The amount of information available on Internet is considerable and is growing quickly. It is the same case for health information. Information retrieval remains problematic: it is difficult to find exactly what ones is looking for, in spite of existing tools such as Web-catalogues (for example Yahoo) or search engines (for example Google). Free text word-based (or phrase-based) search engines typically return innumerable completely irrelevant hits requiring much manual weeding by the user and might miss important information resources. Free text search is not always efficient and effective: the sought page might be using a different term (synonym) that points to the same concept; spelling mistakes and variants are considered as different terms; search engines cannot process HTML *intelligently*, the most widespread language on the Web. In catalogues such as Yahoo, the resources are manually indexed and classified under topics and categories, which are too general to answer specific requests: there is an overlap between the categories as well as imprecision regarding the definition of their scope. This leads to confusion as to what to expect under a given category [12], and the user may not know in which category he can find what he is looking for.

We limit our context of information retrieval to the CISMeF[1] project [4] (acronym of Catalogue and Index of French-speaking Medical Sites) which has been developed since 1995 to help health professionals, as well as students and the general public, during their search for electronic health information. All the resources indexed in the CISMeF catalogue are described according to a structured terminology that is similar to an ontology of the medical domain, and a set of metadata elements. We propose here to use two knowledge-based components, natural language processing and knowledge discovery in databases, in our search engine KnowQuE (Knowledge-based Query Expansion) to correct and refine the users' queries. The first

[1] http://www.chu-rouen.fr/cismef/

component use lexical knowledge, in particular morphological knowledge. The second component use association rules between terms, extracted from the indexed resources using a data mining technique. We show how we have built these two components and give some preliminary results. (This work follows that done in [13]).

2 Towards a Medical Semantic Web

Nowadays the problematic is *intelligent information retrieval* on the Web. The Semantic Web [3] is an infrastructure that has to be built. It aims at creating a web where information semantics are represented in a form that can be understood by human as well as machines, better enabling computers and people to work in co-operation. One of its advantages is to bring sufficient information on the resources, by adding annotations in the form of *metadata* and to describe formally and significantly their content according to an ontology. This infrastructure must be formalized. The current Web is informal: HTML pages hand-written or generated automatically for only a human treatment mainly compose it. Ontologies and metadata are two major components for the construction of the Semantic Web. Ontologies are considered to be powerful tools to lift ambiguity: they provide a controlled vocabulary of terms and some specification of their meaning and are very useful for interoperability and for browsing and searching. Metadata describe Web information resources enhancing information retrieval and enabling accurate matches to be made while being totally transparent to the user.

The CISMeF catalogue describes and indexes a large number of health information resources. A resource can be a Web site, Web pages, documents, reports and teaching material: any support that may contain health information. The resources are selected according to strict criteria and a four-step methodology by the librarian team. The resources are described according to a structured terminology and several sets of metadata. This structure enables us to place the project at an overlap between the actual Web and the forthcoming Semantic Web.

2.1 The CISMeF Metadata

Metadata is data about data and specifically in the context of the Web, it is data that describe Web resources. When properly implemented, metadata can enhance information retrieval. In CISMeF several sets of metadata elements are used. The resource indexed are described the Dublin Core (DC) elements set [2] (e.g. *author, date*). 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 [4] (e.g. *institution, target public*). Two additional fields are in the resources intended for the health professionals: indication of the *evidence-based medicine* and the *method* used to determine it. In the teaching resources eleven elements of the IEEE 1484 LOM (Learning Object Metadata) “Educational” category are added. The metadata format was the HTML language in 1995. Since December 2002, the format used is RDF [7], a Semantic Web language, within the ongoing MedCIRCLE project [8], developed to qualify health information quality.

2.2 The CISMeF Terminology

The CISMeF terminology is founded on the MeSH [9] concepts and its French translation. The MeSH thesaurus, in its 2003 version, is composed by approximately 22,000 keywords (e.g.: *abdomen, hepatitis*) and 84 qualifiers (e.g.: *diagnosis, complications*). These concepts are organized into hierarchies going from the most general, at the top of the hierarchy, to the most specific at the bottom of the hierarchy. For example, the keyword *chromosomal aberration* is more general than the keyword *trisomy.* The qualifiers, also organized into hierarchies, allow specifying which particular aspect of a keyword is addressed. For example the association of the keyword *trisomy* and the qualifier *diagnosis* (noted *trisomy/diagnosis*) restrict the *trisomy* to its *diagnosis* aspect. The MeSH was selected because it responds to the waiting of the librarians and it is well known by the health professionals. MeSH keywords and qualifiers are organized into hierarchies that do not allow a complete view concerning a specialty. The

keywords and qualifiers in CISMeF are gathered according to *metaterms.* Metaterms (*n*=66) concern medical specialties. They are similar to super-concepts and allow knowing the sets of the MeSH keywords and qualifiers that are dispersed in several trees and semantically related to the same specialty. In addition to the set of metaterms, a hierarchy of *resource types* (*n*=127) has been modeled by the CISMeF team. The resource types describe the nature of the resource (e.g.: *teaching material, clinical guidelines*). The metaterms and resource types enhance information retrieval into the catalogue when searching "*guidelines in cardiology*" or *"databases in virology"*, where *cardiology* and *virology* are metaterms and *guidelines* and *databases* are resource types. The *"is-a"* and "*part-of"* relations between concepts are extracted from the MeSH text files to define the subsomption relationships in the CISMeF keywords hierarchy.

A CISMeF resource can be indexed by keywords, couples of (keyword/qualifier) and resource types.

The CISMeF terminology (Figure 1) has the same structure as a terminological ontology [14]. The vocabulary describes major terms of the medical domain and is well known by the librarians and the health professional. Each concept has a *preferred term* to express it in natural language, a set of properties, a natural language definition that allows differentiating it from the concepts it subsumes and those that subsume it, a set of synonyms and a set of rules and constraints. All these relations are exploited in information retrieval. The resources and all the information concerning the terminology are stored and managed by the Relational DBMS Oracle 8.i.

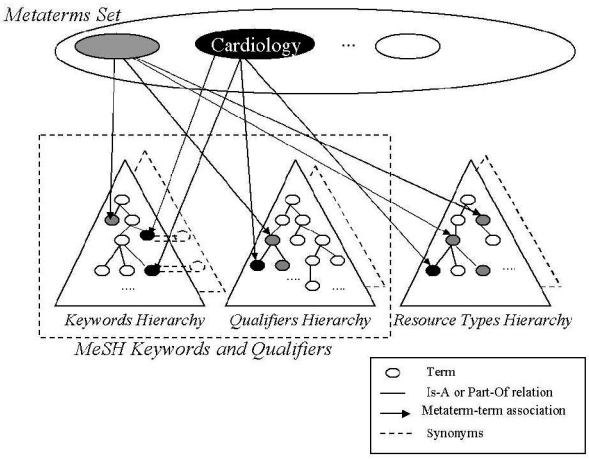

Figure 1: The structure of the terminology.

2.3 Information Retrieval

Many kinds of search are available in the catalogue [4]. We have extracted and analyzed the kind of queries from the http server log and their associated number of answers between the 15th August 2002 and the 6th February 2003. 1,552,776 queries were extracted. Among them 892,591 (58.62%) were submitted via the *simple search* interface and 365,688 (40.97% of the simple queries) had no answer. A refined analysis over the simple queries showed that 12.01% of the null queries can be matched with the terminology: they do not correspond to a bad query they simply do not have any related resource indexed in the catalogue.

The *simple search* is based on the subsumption relationships. If the query (a word or an expression) can be matched with an existing concept, then the result of the query is the union of the resources that are instances of the concept, and the resources that are instances of the concept it subsumes, directly or indirectly, in all the hierarchies it belongs to. For example a query on *"hepatitis"* will return as answer all the resources related to *hepatitis* but also those related to *hepatitis A, hepatitis B…*etc. If the query cannot be matched with an existing concept, the search is done over the other fields of the metadata. In the worse case, a full-text search is carried out. But as said before it is not an optimized solution and this kind of search requires a good knowledge of the medical domain, which is not obvious for any user.

Therefore, to enhance this kind of information retrieval, which is largely used over the catalogue, we propose to use morphological knowledge and association rules.

3 Morphological Knowledge

The submitted queries over the search engine are seldom matched to the terms of the terminology. We apply a morphological analysis of the queries. The results of a preliminary work [15] showed that using a morphological knowledge base enhance information retrieval tasks. The proposed algorithm consists in correcting the user query (only in case of null answer) by eliminating the stop words (such as *the, and, when*) and replacing each word of the query by a disjunction of all the terms in its morphological family. We consider that any term of the morphological family is equivalent to the others. A morphological family of a term is composed by its *inflexions* (for example {*accident, accidents*}) and *derivations* (for example {*probability, probabilistic*}). If the user query is "*interaction between the drugs"* it will be replaced by the term *"drug interactions"*. The problem is that this kind of knowledge base doesn't exist for the French medical language. It is the UMLF[2] ongoing project [16].

3.1 Extracting Derived Words

To build a morphological knowledge-base according to the CISMeF terminology, we have used a terminological resource Lexique [10], which contains the lexicon of the contemporary French deduced from a corpus of texts written between 1950 and 2000. Lexique is not specific to the medical domain but it allowed us to obtain 31,016 derived terms that match exactly 1,484 terms of CISMeF. In order to precise the coverage we have analyzed the structure of the CISMeF terminology. We have only considered here the terms that are already used in the indexed resources: 3,953 keywords; 78 qualifiers; 127 resource types; so a total of 4,158 terms.

[2] Unified Medical French Lexicon

By merging the first morphological knowledge base [15] with that built from Lexique we have obtained the following results (Table 1).

Table 1: Matching the Terminology;

K: keywords, Q: Qualifiers, R: Resource Types.

Matched Terms	K	Q	R
Number	1405	54	25
Terms of 1 word	97.77%	98.18%	89.28%
Terms of 2 and +	83.58%	78.48%	41.73%
Total	35.54%	68.35%	19.68%

By analyzing the other terms composed by 2 or more words, we have found that 1,935 terms (1,899 keywords; 8 qualifiers; 22 resource types) are *semi-matched.* We consider that a term is *semi-matched* when at least one of the words that compose it is matched. For example the keyword *"accidents"* has as family: *{accident, accidents, accidenté, accidentées, accidentel, accidentels, accidentelle, accidentelles, accidentellement, accidenter}.* Therefore, the keyword "*accident circulation"* is semi-matched because *accident* is matched and not *circulation*.

3.2 Correcting the Queries

We have implemented the algorithm in Java with an ODBC connection to the CISMeF database. The different functions of the algorithm (Segmentation, Normalization, Accents, Stop Words, and Derived Words) were expressed using SQL queries and Regular Expressions.

For preliminary results, we have tested the algorithm on a set of 6,954 *null* queries, which were segmented into 17,827 terms to be matched with the terminology (Segmentation step using String Tokenizers (e.g.: ** $,!§;|@*).

We have obtained the following results with the other functions (Table 2). A total of 13,181 terms (73.94%) were matched and 4,646 terms are unknown (26.06%). Many of the unknown words are misspelling errors but, in addition to morphological knowledge, semantic knowledge is necessary, for example *heart* and *cardiac* are

semantically related and a syntactic analysis is not adapted. The set of the CISMeF synonyms were created with the collaboration of patients associations and the French National League Against Cancer. We are currently analyzing the log queries to complete this set.

Table 2: Matching the Queries

FUNCTION	MATCHED TERMS	%
Accents (éèêàï replaced)	3865	21.68%
Stop Words (and, a, one)	3314	18.59%
Derivations + Inflexions	6002	33.68%

We have done a quantitative analysis to match the null queries of the users with the terminology enriched by a morphological knowledge base. It is not sufficient. A qualitative one must be realized using the (precision/ recall) metrics.

4 Data Mining : Refining the Queries

We want to discover "new" knowledge from the CISMeF database and in particular from the annotations and the terms. This knowledge will be exploited in the process of information retrieval in order to refine the queries, especially when the system returns too many answers.

We apply a Data Mining technique called *Association Rules* to extract interesting associations, previously unknown, from the database.

4.1 Association Rules Extraction

Association rules were initially used in data analysis and in data extraction from large relational databases [1]. We are interested in the discovery of Boolean association rules, which are expressed as:

$$AR : i_1 \wedge i_2 \wedge \ldots \wedge i_j \Rightarrow i_{j+1} \wedge \ldots \wedge i_n$$

The rule AR states that if an object has the items $\{i_1, i_2 \ldots, i_j\}$ it tends also to have the items $\{i_{j+1}, \ldots, i_n\}$. The AR *support* represents its utility. This measure corresponds to the proportion of objects which contains at the same time the rule antecedent and consequent. The AR *confidence* represents its precision. This measure corresponds to the proportion of objects that contains the consequent rule among those containing the antecedent.

The knowledge extraction process is realized by several steps: the data and context preparation (objects and items selection), the extraction of the frequent itemsets (compared with a minimum support threshold), the generation of the most informative rules using a Data Mining algorithm (compared with a minimum confidence threshold), and finally the interpretation of the results and deduction of new knowledge. Our extraction context is the triplet $C= (O, I, R)$ where O is the set of objects, I the set of all the items and R a binary relation between O and I. The objects are the annotations used to describe the indexed resources. The relation R represents the indexing relation between an object and an item. We consider the case I={(Keyword/Qualifier)}, the couples of (keyword/qualifier) that index the resources. First, the frequent itemsets are extracted. An itemset is frequent in the context C if its support is higher than the minimal threshold initially fixed. The extraction problem of frequent itemsets has an exponential complexity in size of n, the size of the potential frequent itemsets is 2^n. The itemsets form a lattice [5]. The most known algorithm used to extract frequent itemsets is Apriori [1]. In our case we use the A-Close algorithm [11], which calculates the *closed frequent itemsets* using the semantic based on the closure of the Galois connection [6], reducing by that itemsets space size studied. The algorithm calculates the generators of the frequent closed itemsets. The generators of a closed itemset I_{close} are the itemsets of maximal size which closure is equal to I_{close}. New bases for association rules are deduced from the closed frequent itemsets and their generators. These bases consist of non-redundant association rules of minimal antecedents and maximal, i.e. the most relevant association rules [11].

We have implemented the A-Close algorithm in Java and tested it on several sets of resources by fixing the support to 10 documents and the confidence to 100% (exact rules). The results are in Table 3.

Table 3: Number of resources and rules extracted for 10 specialties

SPECIALTY	RESOURCE	RULE
Environment	1254	53
Neurology	1137	25
Pediatrics	906	57
Diagnosis	883	33
Therapeutic	782	18
Oncology	644	20
Cardiology	558	2
Psychiatrics	515	3
Allergy	509	36
Gastro.	501	13

4.2 Evaluation

All the extracted rules (260) were evaluated by an expert (medical librarian) (Table 4). An interesting association rule is one that confirms a hypothesis or states a new hypothesis.

In our case, there are several cases of interesting association rules. An association rules that associate:

- a direct (or an indirect) son and its father in the hierarchy (Father-Son type)
- two terms (or more) that belong to the same hierarchy (have the same direct or indirect father) (Brother type)
- a See Also relationship that exists in the thesaurus (See-Also type)
- a new relationship judged interesting (New type).

Among the 260 rules, 142 (54.61%) were judged interesting by our expert. Examples of "New" rules:

breast cancer/diagnostic ⇒ *mammography* (support=25 documents, confidence=1)

aids/prevention and control ⇒ *condom* (support=10 documents; confidence=1)

Among the interesting rules we have obtained:

- 68.31% new rules
- 14.49% See Also relationships
- 10.56% Brother relationships
- 05.63% Father-Son relationships.

Table 4: links: FS: Father-Son; SA: See Also; B: Brother; NW: New ; OT: Other (not interesting).

SPECIALTY	FS	SA	B	NW	OT
Environment.	5	4	7	13	24
Neurology	0	4	1	8	12
Pediatrics	1	2	2	0	52
Diagnosis	1	3	0	26	3
Therapeutics	0	3	2	3	10
Oncology	0	1	1	17	1
Cardiology	0	0	0	0	2
Psychiatrics	0	1	0	0	2
Allergy	1	3	2	26	4
Gastro.	0	1	0	4	8

The Father-Son relationships are already used in the information retrieval process. The other types of interesting association rules could be used to refine the users' queries.

5 Conclusion and future work

We have proposed different methods to enhance information retrieval into the CISMeF catalogue. The natural language processing is used to build morphological knowledge base. Data Mining enables association rules discovery between concepts. To our knowledge, no existing work has combined these methods in order to enhance information retrieval.

To evaluate the contribution of each method, first we apply an automatic expansion (or query enrichment, to enlarge the research) over the users' queries by using each resource (morphological knowledge base and association rules) separately and then conjointly. The users'

queries are those extracted from the http server log and those that have no answer. We can also take into account all the set of queries to evaluate the contribution before and after expansion. Secondly the expansion will be interactive: a representative set of users will have to evaluate for each submitted query the *utility* of the various expanded queries suggested by each method. This evaluation (automatic and interactive) on a larger scale will allow building a base of rules and a protocol to apply a method, or a combination of the two methods according to the type of query submitted to the search engine and the number of answers.

References

[1] R. Agrawal, R. Srikant (1994). Fast algorithms for mining association rules in large databases. *Proceedings VLDB Conference*, 478-499.

[2] T. Baker (2000). A Grammar of Dublin Core. *Digital-Library Magazine,* vol 6 n°10.

[3] T. Berners-Lee, J. Heudler, O. Lassila (2001). The Semantic Web. *Scientific American*, 284(5):34-43.

[4] SJ. Darmoni, B. Thirion, JP. Leroy, et al. (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.

[5] BA. Davey, HA. Priestley (1994). Introduction to lattices and order. *Cambridge University Press.*

[6] B. Ganter, R. Wille (1999). Formal concept analysis: mathematical foundations. *Springer-Verlag*.

[7] O. Lassila, R. Swick (1999). Resource description framework (RDF), model and syntax specification. *W3C candidate recommendation.*

[8] MA. Mayer, SJ. Darmoni, M. Fiene, et al. (2003). MedCIRCLE Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. In G. Surjan, R. Engelbrecht, P. Mc Nair (Eds): MIE 2003, IOS Press Publisher. *Stud Health Technol Inform.* 95:667-672.

[9] SJ. Nelson, WD. Johnson, BL. Humphreys, (2001). Relationships in Medical Subject Headings. Bean and Green (eds). *Kluwer Academic Publishers*, 171-184.

[10] B. New, C. Pallier, L. Ferrand, R. Matos (2001). Une base de données lexicales du français contemporain sur Internet: LEXIQUE, *L'Année Psychologique*, 447-462.

[11] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems*, 24(1):25-46.

[12] K. Risden (1999). Toward usable browse hierarchies for the Web. Bullinger and Zieder (Eds). *Human Computer Interaction: Ergonomics and User Interfaces*, 1:1098-1102.

[13] LF. Soualmia, C. Barry, SJ. Darmoni (2003). A knowledge-based query expansion system over a terminology oriented ontology. In M. Dojat, E. Keravnou, P. Barahona (Eds.): AIME 2003, *Lecture Notes in AI* # 2780, 209-213.

[14] JF.Sowa,(2000) Ontology, Metadata and Semiotics. B. Ganter, G. W. Mineau (Eds), Conceptual Structures: Logical, Linguistic, and Computational Issues, *Lecture Notes in AI* # 1867, 55-81.

[15] P. Zweigenbaum, SJ. Darmoni, N. Grabar (2001). The contribution of morphological knowledge to French MeSH mapping for information retrieval. Journal of the American Medical Informatics Association 8:796-800.

[16] P. Zweigenbaum, R. Baud, A. Burgun, et al. (2003). Towards a Unified Medical Lexicon for French. In G. Surjan, R. Engelbrecht, P. Mc Nair (Eds): MIE 2003, IOS Press Publisher *Stud Health Technol Inform.* 95:415-420.