eHealth Beyond the Horizon – Get IT There
S.K. Andersen et al. (Eds.)
IOS Press, 2008

Using Knowledge for Indexing Health Web Resources in a Quality-Controlled Gateway

Michel JOUBERT[a], Stefan J DARMONI[b], Paul AVILLACH[a,c], Badisse DAHAMNA[b], Marius FIESCHI[a]

[a] *LERTIM, Faculte de Medecine, Universite de la Mediterranee, Marseille, France*
[b] *CISMeF, Centre Hospitalo-Universitaire de Rouen, France*
[c] *LESIM, INSERM U593, ISPED, Université Victor Segalen, Bordeaux, France*

Abstract. Objectives: The aim of this study is to provide to indexers MeSH terms to be considered as major ones in a list of terms automatically extracted from a document. Material and methods: We propose a method combining symbolic knowledge - the UMLS Metathesaurus and Semantic Network - and statistical knowledge drawn from co-occurrences of terms in the CISMeF database (a French-language quality-controlled health gateway) using data mining measures. The method was tested on CISMeF corpus of 293 resources. Results: There was a proportion of 0.37 ± 0.26 major terms in the processed records. The method produced lists of terms with a proportion of terms initially pointed out as major of 0.54 ± 0.31. Discussion: The method we propose reduces the number of terms, which seem not useful for content description of resources, such as "check tags", but retains the most descriptive ones. Discarding these terms is accounted for by: 1) the removal by using semantic knowledge of associations of concepts bearing no real medical significance, 2) the removal by using statistical knowledge of non-statistically significant associations of terms. Conclusion: This method can assist effectively indexers in their daily work and will be soon applied in the CISMeF system.

Keywords. Knowledge-based systems; Terminology-vocabulary; Indexing.

Introduction

The dissemination of medical scientific knowledge to the general public, students and health professionals is today well and truly based on the Internet. In France and countries in Europe, North America, Africa and Asia, where French is spoken, there exist numerous web sites providing health-related information. Some sites deliver accredited information, the information delivered by the others is not. The former received their accreditation from recognized bodies such as the CISMeF gateway (acronym for Catalog and Index of French-language Health Resources on the Internet, the most important and quality-controlled source of institutional health information in French) [1], or the Health on the Net Foundation throughout the rest of the world [2].

In a system as CISMeF, for instance, the problem is to index a large corpus of health-related documents. Today, indexing scientific documents is conducted using generally the MeSH thesaurus. Researchers in this field have envisaged doing

automated indexing systems for French-language documents, which should achieve the same capabilities as those developed for English [3]. However, in the long term, other terminologies in addition to MeSH should be used in order to index documents other than of a pedagogical or scientific nature, such as clinical reports, laboratory examinations, etc. The VUMeF (Unified Medical French-language Vocabulary) project [4], sponsored by the French National Research Agency (ANR), has been working along these lines. Its ambitious objective was to integrate the various French-language terminologies into a single homogenous package. To this end, it has drawn on the results of another project, the UMLF (Unified Medical Lexicon for French) [5], also sponsored by the ANR, the aim of which was to produce an as-complete-as-possible French-language Specialist Medical Lexicon.

The aim of this study is to propose a method related to the development of tools designed to facilitate indexing a large documentary database in the health domain. The objective is to develop a complete automated indexing process involving two components: 1) an indexer that extracts MeSH terms from documents, and 2) a process that filters extracted terms and retains only the most informative ones. We describe in this paper the method we propose to implement this latter component. We use symbolic knowledge: the UMLS Metathesaurus and Semantic Network [6]. We also use statistical knowledge derived from co-occurrences of MeSH terms in the CISMeF gateway documentary database. The method we propose is inspired by a model of documents coding/indexing we previously applied on coding standardized discharge summaries in a French hospital [7]. We illustrate here the application of the proposed method to a set of records provided by the CISMeF gateway each containing a list of MeSH terms. This work is part of a wider project of semi-automated indexing of some documents in the CISMeF database: a first process extracts automatically MeSH terms from documents, after what our method selects some of them according to their computed relevance, providing indexers with an effective help system. This study follows previous research on the automated indexing of French-language health-related web resources [8].

1. Material and methods

1.1. UMLS symbolic knowledge

The UMLS knowledge sources we used (2007AA release) are the Metathesaurus and the Semantic Network. The Metathesaurus provides an inventory of the concepts used in the health domain drawn from the nomenclatures included in the UMLS, more than a hundred today, including MeSH. Each MeSH term is thus connected individually to a concept in the Metathesaurus. Each concept is connected to at least one type of concept in the Semantic Network. Binary semantic relationships are defined between the different types of concepts conveying declarative medical knowledge such as “causes”, “diagnoses”, “prevents”, “complicates”, etc. These relationships thus make it possible to propose meaningful connections between MeSH terms.

In a previous study we shown how these relationships could be conveyed by qualifiers that are associated with the terms in order to build queries for a document server using MeSH for indexing purpose [9]. For instance, “a diagnostic procedure diagnoses a disease” associates the qualifier “diagnostic” with the term designating the disease. This knowledge thus enables us to propose associations of MeSH terms, some

of which are qualified, for comparison with the associations found in the list of terms in a record. Only pairs of terms having a semantic association issued from the Semantic Network are considered after this first step.

1.2. Statistical knowledge

Each resource is indexed in the CISMeF database with a set of MeSH terms. It is then possible to construct the entire set of term co-occurrences from the set of records and to associate a frequency to each pair of co-occurring terms. We thus obtain a set of distinct pairs of co-occurring terms and their frequency of occurrence.

The *confidence* (*Conf*) we have to find a term B when a term A is present is the ratio between the number of times that A and B, *Freq* (A,B), co-occurred and the number of times that A has occurred independently of the other terms, *Freq*(A):

$$Conf(A,B) = Freq\ (A,B) / Freq\ (A)$$

Confidence is the equivalent to the conditional probability of B if A in terms of probabilities [10]. It indicates the degree of interest of the association between A and B. However, confidence alone is not sufficient.

The *interest* (*Int*) we will have to consider the association of A and B is measured by the ratio between the confidence of the association of A and B and the frequency of B:

$$\begin{aligned} Int\ (A,B) &= Conf(A,B) / Freq\ (B) \\ &= Freq(A,B) / (Freq\ (A)*Freq\ (B)) \end{aligned}$$

Interest is the equivalent to *lift* as used in data mining and which is the coefficient by which we have to multiply the a priori probability of B in order to obtain the conditional probability of B, A being given. Interest measures the informational value of the association of A and B as compared with A and B taken separately. Thus, a high value of *Int*(A,B) means that A being given, the probability of B is high, and conversely.

Let us take the list of terms for a given record. Let us consider one of the terms that we will name A. The other terms in this same list will be named B_i. Knowing A, and the entire set of the ordered (A,B_i) pairs, we can obtain $Conf(A,B_i)$ and $Int(A,B_i)$ for each of the B_i. This allows us to assert, knowing A, which B_i are of interest. And conversely, knowing the B_i, we can suggest whether A is of interest or not. In both instances, it is necessary to determine the value of the interest above which the pairs of co-occurring terms are to be considered. This value is entirely heuristic and has an impact on the quality of expected results. It determines the terms to take into account in each record according of the relative interest they have to be associated with the other ones in the same record.

1.3. Material

Each CISMeF record contains MeSH terms, with or without qualifiers, and librarians according to indexing rules have designated certain terms as major terms. In order to test the above method, we built a set of co-occurrence pairs using CISMeF records and

applied it to the records returned by the CISMeF gateway in response to 100 requests formulated by five different teams. The system replied successfully to 89 requests. We limited to 10 the number of responses to each request as this generally matches the first page of responses returned by a search engine (between 66% and 85% of expected results are present in the first page displayed [11]). This gave us a total of 293 successful answers.

Table 1 shows the number of different terms, the number of different major terms, and the number of co-occurrences between terms (major terms included) in the test sample. Co-occurrences were limited by adopting a lower threshold set at 0.5 and 1.0 for interest.

Table 1. Numbers of terms, major terms and co-occurrences for interest set at 0.5 and 1.0.

Number of terms	1,400
Number of major terms	740
Number of initial co-occurrences	9,242
Number of co-occurrences, interest 0.5	6,962
Number of co-occurrences, interest 1.0	4,460

2. Results

Table 2 shows the average numbers of terms, major terms and proportion of major terms per record in the test sample. The second column of the table presents these values for records indexed humanely. They will constitute our standard in what follows. Columns 3 and 4 shows the values obtained after applying the method with an interest value set at 0.5 and 1.0. The results show three things.

- the average numbers of terms decreases when the value of the interest increases
- the average number of major terms is not statistically different between filtered lists of terms and initial ones
- the average proportion of major terms in processed records increases with the value set to the interest.

Even if there are no statistically significant differences, we nevertheless try to interpret the results. A value of interest set to 1 is too restrictive because it decreases too much the average number of major terms. So, considering a value of interest set to 0.5, the average number of major terms is more or less the same than the standard value, and the proportion of major terms in the lists of processed records is higher than the standard value.

Table 2. Average number of terms, major terms and proportions of major terms per record in the test sample, and after applying the method with interest value set at 0.5 and 1.0

Interest		0.5	1.0
Average number of terms	8.17 ± 3.36	4.32 ± 2.41	3.55 ± 2.05
Average number of major terms	2.52 ± 1.41	1.94 ± 1.24	1.74 ± 1.11
Average proportion of major terms	0.37 ± 0.26	0.50 ± 0.31	0.54 ± 0.31

3. Discussion

We applied the combination of symbolic and statistical knowledge successfully in a previous work [7]. Even if the intent was not the same (coding discharge summaries on one hand, indexing documents on the other one), and the implementation of the model was different, they both exploit the symbolic knowledge of the UMLS and co-occurrences (of diagnoses and medical acts on one hand, of terms on the other one).

The joint use of both symbolic and statistical knowledge enables our proposed method to achieve an appreciable level of efficacy. Symbolic knowledge uses the UMLS Semantic Network to provide associations of medically significant concepts. The statistical knowledge built in a field of application allows taking into account a context for the symbolic knowledge and makes it possible to attribute a numerical value to the medically relevant associations. This signifies that it eliminates a number of terms deemed to be of lesser importance for the description of the documents content. Facing the challenge of a semi-automated indexing process of health documents, it retains the most descriptive ones. Values of Table 2 show that the average number of retained terms decreases when the value set to interest increases. This signifies that a small value of the interest is in favour of a number of terms greater than those retrieved with a high value of interest, due to the retained number of associations (Table 1). But, a small value of the interest makes the average number of major terms quite equal to the standard value. Lastly, we can remark that the number of records that the method has difficulties to fully process increases with the threshold set for the interest. So, we have to tune the value set to the interest according to the relevance of the method. In the case of this experiment our choice would be to set the value of the interest to 0.5. That allows discarding a notable number of terms while retaining the most of major ones.

Some terms, such as “check tags” (concepts of potential interest, regardless of the general subject content of the resource including terms such as *human*, *animal*, and *case-report*), having neither semantic nor statistical relations with other terms in records are discarded. Nevertheless indexers for resources indexing/retrieval purpose must introduce them. The method sometimes supplies a term with no qualifier and the same term with qualifiers. Let us consider, for instance, a record containing the following terms where the symbol « * » indicates a major term: *hematuria*; **hematuria/diagnosis*; *hematuria/etiology*; *urology/teaching and education*; *signs and symptoms*. After processing the list of suggested terms is: *hematuria*; *hematuria/diagnosis*; *hematuria/etiology*. The document in question deals with diagnosis of hematuria and therefore there are good grounds for indicating that “hematuria” and “etiology of hematuria” could be considered as significant terms. The method discards the terms *signs and symptoms* and *urology/teaching and education* that are important for retrieving purpose, but not significant for describing precisely the content of the document, as it does with “check tags”.

4. Conclusion

The proposed method aims to provide two-pronged assistance to human indexers in their work. Firstly, by using statistics to check that the terms indicated by an indexer comply with the established statistics. Secondly, in the light of the established statistics, by interactively submitting to indexers appropriate keywords and qualifiers according

to the keywords that he/she has already introduced into the record of the document currently being indexed. The method can also be used to check indexing quality by verifying terms introduced by an indexer and can then issue warnings when unlikely associations of terms occur.

The CISMeF team developed automated indexing tools since 2002 [3, 12]. Then it was decided to use them in the daily practice for most of the Internet resources. Three levels of indexing were defined: 1) totally manually indexed resources (e.g. guidelines); 2) resources automatically indexed and then reviewed by medical librarians; 3) totally automatically indexed resources (e.g. teaching material). The method proposed in this work will be soon applied to resources falling within levels 2 and 3. The use of the method is particularly interesting in the latter case. An automated indexation of documents may produce a lot of terms that the method can filter to retain only the most descriptive ones according to statistical and symbolic knowledge bases.

Acknowledgements

The authors thank the whole consortium of the VUMeF project in the framework of which this work was conducted. They also thank the U.S. National Library of Medicine for free access to the UMLS knowledge sources. They thank Mr George Morgan for his translation.

References

[1] Darmoni SJ, Leroy JP, Baudic F, Douyère M, Piot J, Thirion B. CISMeF: a structured health resource guide. Methods Inf Med. 2000 Mar; 39(1): 30-5.
[2] Boyer C, Selby M, Scherrer JR, Appel RD. The Health On the Net Code of Conduct for medical and health Websites. Comput Biol Med. 1998; 28(5): 603-10.
[3] Neveol A, Mork JG, Aronson AR, Darmoni SJ. Evaluation of French and English MeSH indexing systems with a parallel corpus. AMIA Annu Symp Proc. 2005: 565-9.
[4] Darmoni SJ, Jarousse E, Zweigenbaum P et all. VUMeF: extending the French involvement in the UMLS Metathesaurus. AMIA Annu Symp Proc. 2003: 824.
[5] Zweigenbaum P, Baud R, Burgun A et all. Towards a unified medical lexicon for French. Stud Health Technol Inform. 2003; 95: 415-20.
[6] McCray AT, Nelson SJ. The Representation of Meaning in the UMLS. Meth Inform Med 1995; 34: 193-201.
[7] Avillach P, Joubert M, Fieschi M. A Model for Indexing Medical Documents Combining Statistical and Symbolic Knowledge. Proc. AMIA Symp. 2007. To appear.
[8] Joubert M, Peretti AL, Darmoni SJ, Dahamna B, Fieschi M. Contribution to an Automated Indexing of French-language Health Web Sites. AMIA Annu Symp Proc. 2006: 409-13.
[9] Aymard S, Falco L, Dufour JC, Joubert M, Fieschi M. Modeling and implementing a health information provider on the Internet. Proc. MIE2003. IOS Press; 2003: 89-94.
[10] Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. ACM Computing Surveys 2006; 38(3): art. #9.
[11] Spink A, Jansen, BJ. Web search: Public searching of the web. Kluwer Academic Publishers 2004.
[12] Névéol A, Pereira S, Kerdelhué G, Dahamna B, Joubert M , Darmoni SJ. Evaluation of a simple method for the automatic assignment of MeSH descriptors. to health resources in a French online catalogue. Stud Health Technol Inform. 2007; 129: 407-11.