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

Searching Related Resources in a Quality Controlled Health Gateway: a Feasibility Study

Tayeb MERABTI[a, b], Suzanne PEREIRA[a, b, c], Catherine LETORD[a], Thierry LECROQ[a], Michel JOUBERT[b], Badisse DAHAMNA[a], Stéfan J. DARMONI[a]
[a] *CISMeF, Rouen University Hospital, France & GCSIS, LITIS EA 4108, Institute of Biomedical Research, University of Rouen, France*
[b] *LERTIM, Faculte de Medecine, Universite de la Mediterranee, Marseille, France*
[c] *VIDAL Company, Issy les Moulineaux, France*

Abstract. Objective: The neighbors of a document are those documents in a corpus that are most similar to it. The objective of this paper is to develop and evaluate the related resources algorithm (CISMeF-RRA) in the context of a quality-controlled health gateway on the Internet CISMeF. **Method**: CISMeF-RRA is inspired by the PubMed Related Citations Articles. CISMeF-RRA combines statistical distances with a semantic distance using MeSH terms/qualifiers. **Material:** In this feasibility study an evaluation was performed using 50 CISMeF resources randomly chosen. **Results:** Overall, 49% of the related documents were ranked as relevant. **Conclusion**: if this feasibility study is confirmed by another evaluation of more resources, CISMeF-RRA will be implemented in the CISMeF catalog

Keywords. Algorithms; Automatic Data Processing; Catalog; Medical Subject Headings

Introduction

The Internet and in particular the Web has become an extensive health information repository. In this context, several quality-controlled health gateways have been developed [1]. Quality-controlled subject gateways were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery.

Among several quality-controlled health gateways, CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the Internet) [3] was designed to catalog and index the most important and quality-controlled sources of institutional health information in French in order to allow end-users to search them quickly and precisely (N= 36,851). CISMeF is used by Netizens and health professionals mainly from the French-speaking countries (N≈ 50,000 users per working day).

CISMeF is manually indexed by a team of four indexers, who are medical librarians. Its URLs are http://www.chu-rouen.fr/cismef or http://www.cismef.org The Doc'CISMeF search engine has four searches types: Simple, Advanced, Boolean, and Step by Step. In the Simple search, the end-user enters a query in a natural language in French or in English. This query is then automatically transformed by natural language processing tools [4] (e.g. phonemization, stemming) to map this query to the terminology used by CISMeF based on the MeSH thesaurus developed by the US National Library of Medicine [5]. The display of resources answering the end-user query is common to the four search types of the Doc'CISMeF search engine. Then, the end-user must choose the most interesting resources function of his/her context, which is most of the time much more complex than the one expressed in the query.

From one resource, it is important to obtain the nearest neighbors of a resource (or most related resources). The neighbors of a resource are those documents in the database that are the most similar to it [6].

The objective of this paper is to develop and to evaluate the algorithm "most closely related resources" in the CISMeF database. The CISMeF Related Resources Algorithm (CISMeF-RRA) is derived from the original idea of the work performed by Kim *et al.* [7]. The Related Citations Articles (NLM-RCA) feature is available from PubMed, which is a service of the US National Library of Medicine that mostly includes over 17 million citations from the MEDLINE bibliographic database. CISMeF-RRA was clearly inspired by NLM-RCA, but the algorithm was modified to adapt it to the more heterogeneous scope of Internet resources from the CISMeF gateway, when compared to scientific articles from the MEDLINE bibliographic database. The main difference of our approach consists in combining the statistical distance between documents as established by Kim *et al.* [7] with a semantic distance using the MeSH terms/qualifiers and the CISMeF resources type (RT).

1. Methods

1.1. CISMeF Terminology

The CISMeF terminology is exploited for several tasks: manually performed resource indexing, automatically performed resource categorization, visualization and navigation through the concept hierarchies in a CISMeF Terminology Server (URL http://www.chu-rouen.fr/terminologiecismef/) and information retrieval using the Doc'CISMeF search engine. CISMeF uses two standard tools for organizing information: the MeSH thesaurus and several metadata element sets, in particular the Dublin Core metadata set (URL:http://www.dublincore.org)[10]. The MeSH terms (24,357 in 2007) 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. The "is-a" and the "part-of" relations between concepts are extracted from the MeSH files to define the subsumption relationships in the CISMeF terms hierarchy.

However, the MeSH thesaurus was originally intended to index scientific articles for the Index Medicus and for the MEDLINE database. In order to customize it for the broader field of health Internet resources, we developed several enhancements [3] to the MeSH thesaurus, with the introduction of two new concepts, metaterms (MT) and resource types respectively.

A metaterm is a medical specialty or a biological science (e.g. cardiology, bacteriology), which has semantic links with one or more MeSH terms, subheadings and RTs. CISMeF resource types are an extension of the publication types of MEDLINE. As defined by the Dublin Core Metadata Initiative (URL: http://www.dublincore.org/documents/dcmi-terms/) [10], a CISMeF RT (N=278) is used to categorize the kind of the content of a resource. MeSH <term/subheading> pairs describe the topic of the resource. For example, in the case of a clinical guideline about carbon monoxide intoxication, 'carbon monoxide poisoning' is the MeSH term and 'clinical guidelines' is the resource type. The RT controlled list is available at the following URL: http://www.chu-rouen.fr/documed/typeeng.html. The RT list has been manually built and maintained by the CISMeF team since 1997.

Major Topics exist in the MEDLINE database and the CISMeF catalogue for terms and qualifiers. A term is said to be "major" if the concept it represents is discussed throughout the whole document, or on the contrary "minor" if it is referred to only in a few paragraphs. Major terms are marked in MEDLINE and CISMeF by a star. In CISMeF, Major Topics are extended to resource types and metaterms. This task is manually performed by the CISMeF medical librarians for resource types. It is automatically performed for metaterms: a metaterm is "major" for a CISMeF resource if and only if at least one term, qualifier or resource type semantically linked to this metaterm is major for the same CISMeF resource (otherwise, the metaterm is minor).

1.2. Similarity calculation between documents

As mentioned by Kim *et al.* [7], the similarity between documents is measured by the words they have in common, with some adjustment for document lengths. In our work, the criteria allowing similarity calculation between documents are based on the description and indexing by the CISMeF medical librarians. There are four criteria as follows: Title of the document, Abstract, MeSH terms (or pairs MeSH term/subheading) and CISMeF resource types. These four criteria belong to the Dublin Core metadata set [11] and comprise the overall representation of a document.

The concept of document and its representation play a fundamental part in the step of an effective computation of inter-document similarity as well as on the treatment level or on the relevance level. However, the most used representation is the vectorial representation in which a document is represented by a t-dimensional vector, where t is the total number of terms in the document database. The inter-document comparison can then be performed by a cosine measure of these two documents vectors [12]. Two steps may be used to reduce the space dimension: elimination of stopwords [1] and stemming [2] to reduce the grammatical variations of words to a possible root word.

Having obtained the set of terms that represents every documents, the next step is to assign a numerical weight for every stemmed word. Thus, each word will be balanced with a TF-IDF weight [12], which is computed as being the multiple of the frequency of a term in a document (TF) by the inverse weight of the frequency of the document in the collection (IDF). In this way a frequent term occurring in a small number of documents will have a greatest weight. In addition to the vectorial distance, three heuristic weightings were defined by the CISMeF team: 1) in order to give an additional weight to the words in the title vs. the words in the abstract (7 and 1

[1] Words with very low discrimination values in the retrieval process
[2] We use a stemming strategy developed in CISMeF

respectively), 2) to give additional weights to major MeSH terms and major CISMeF RT vs. minor MeSH terms and minor CISMeF RT (7 and 3 respectively), 3) to give respective weightings to the four MeSH relations: Hierarchy, See Also, Pharmacological Action, Do Not Confuse (1, 0.1, 0.1, -0.1 respectively) reflecting their respective importance in computing the overall semantic distance.

1.3. Semantic distance

In this work, the similarity between documents also has a semantic dimension in addition to the syntactic dimension previously defined. A word-by-word distance can be defined between the MeSH terms and the MeSH subheadings. The MeSH hierarchical relation is defined as the traditional relation that exists between the concepts in a tree structure. The distance in this relation will be computed in particular by being based on the taxonomic links "is-a", and "part-of": the more distant in the hierarchy the two terms are, the larger the distance. There is no computation of distance for the three other relations, because for each relation there is a list of word pairs (in the relation) and they will be given a score reflecting the weight of the section 1.2. For example for the relation "Do not confuse" the two MeSH terms "sunstroke" and "heat stroke" are in connection and a score of "-0.1" will be given according to this relation.

Thus, the global semantic similarity takes into account not only the hierarchical relation ("is-a", "part-of") of both the MeSH thesaurus and the CISMeF resource types thesaurus but also the three other relations of the MeSH thesaurus. In this semantic distance computation, we are taking into account the subheading affiliation to a MeSH term, and the RT affiliation to a MeSH term (or a MeSH term/subheading pair) [8].

Contrary to NLM-RCA [7] the CISMeF-RRA takes into account Major/Minor indexing for MeSH terms, MeSH subheadings, and CISMeF resource types. For the hierarchical relation the score is computed according to the more information that two terms share in the MeSH tree structure. We have chosen the Lin's similarity [9] to compute this information, already used to compute semantic distance [13].

Given two terms m_i and m_j, the Lin similarity between them is defined as:

$$sim(m_i, m_j) = \frac{2 \times \max_{m \in S(m_i, m_j)} \left[\log\left(p(m) \right) \right]}{\log\left(p(m_i) \right) + \log\left(p(m_j) \right)} \qquad (1)$$

Where S (m_i ; m_j) is the set of the ancestor terms shared by both m_i and m_j, max represents the maximum operator and p (m) is the probability of finding m or any descendants in a reference corpus. It generates normalized similarity values between 0 and 1. Because Lin's similarity model relies on information content, when one term is the parent of another, their similarity is low when the parent term is placed high in the hierarchy. Conversely, it is high when the parent term is low in the hierarchy. Thus, the total similarity between the MeSH terms of two documents I and J will be measured by applying an average of the distances obtained between all their MT according to the four relations:

$$Sim(D_I, D_J) = \frac{\sum sim(a,b)}{card(D_I) \times card(D_J)} \qquad (2)$$

$\forall a \in D_I$ and $\forall b \in D_J$, where D_I and D_J are set of MeSH terms of documents

I, J respectively.

Finally, the total similarity between documents will be a combination of two measurements of similarity (syntactic and semantic).

1.4. Evaluation

In order to test our algorithm we extracted from the CISMeF corpus a randomly-chosen sample of 50 resources and we run two distance algorithms (CISMeF-RRA and NLM-RCA) on this sample as a feasibility study. A manual evaluation was carried out *a posteriori* by an expert medical librarian of the CISMeF team. (CL) She quantified the number of relevant results according to a qualitative Likert scale of 5 levels, her opinion being regarded as the reference (gold standard). The evaluation was performed in two steps: Step 1: For each of the 50 resources, all the resources classified by the algorithm as "related resources" were rated by the medical librarian. Step 2: for each of the 50 resources, only the top 3 resources were rated.

2. Results

The results of the two-step evaluation are presented in Table 1. For CISMeF-RRA, overall 49% of the related resources relevant were ranked as relevant (Good or Very Good) whereas 30% of them do not reach the average (Very Bad one or Bad). In the second step of the evaluation the resources considered as the nearest (first position) were ranked relevant (Very Good or Good) in 68% of the cases, while the resources in the third position were ranked relevant in 58% of these cases.

Table 1. Step 1 and Step 2 of the evaluation

	Step1				**Step2**											
Results by position	**All**				**1**				**2**				**3**			
	CISMeF RRA		NLM RCA		CISMeF RRA		NLM RCA		CISMeF RRA		NLM RCA		CISMeF RRA		NLM RCA	
	N	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Very Good	33	19	17	15	18	36	16	33	14	28	13	25	11	22	12	21
Good	52	30	20	18	16	32	11	22	12	24	09	17	18	36	12	21
Average	32	18	7	06	05	10	03	6	12	24	09	17	08	16	05	08
Bad	28	16	28	25	06	12	08	16	07	14	10	19	08	16	12	21
Very Bad	24	14	38	34	05	10	10	20	05	10	11	21	05	10	15	26

3. Discussion

In this feasibility study, the CISMeF-RRA gave satisfactory results as overall 49% of the related documents were rated as "very good" or "good" vs. 37% for the NLM RCA. This feasibility study is based on a relatively small sample (N= 50). It should be

followed by a more complete evaluation based on the whole manually indexed corpus (N=21,838).

When compared to the NLM-RCA, the CISMeF-RRA has several differences. The CISMeF-RRA computes the inter-document similarity by using two distances. One is in common with NLM-RCA, which is based on a vectorial approach. Nevertheless CISMeF-RRA is based on a weighting of the terms by using the TF-IDF in opposition to the weighting derived from the Poisson model of term frequencies in NLM-RCA. These two weighting measures are based on a similar basic concept: most frequent terms in the documents will have small weights. The main innovation of the CISMeF-RRA relies on the use of semantic inter-document distance based on Lin's similarity metrics for the MeSH hierarchy relation, CISMeF resource types hierarchy relations, and the semantic links between MeSH terms according to the three other relations ("See Also", "Pharmacological Action", "Do Not Confuse"). Another difference between CISMeF-RRA and NLM-RCA relies on Major/Minor indexing processing, Major weighting differs from Minor one in CISMeF-RRA whereas the weights are similar in NLM-RCA.

In the near future, we will need to estimate in a more convincing way the various weightings that were manually assigned by the CISMeF medical librarians. We also envisage to make our semantic distance algorithm more complex by implementing several relations coming from other medical terminologies, in particular SNOMED CT semantic network. We will soon benchmark the CISMeF "Related Resources algorithm" vs. NLM "Related Articles algorithm" based on the overall manually indexed CISMeF corpus using a blind evaluation by a medical librarian.

References

[1] F. Abad Garcia, A. Gonzalez Teruel, P. Bayo Calduch, R. de Ramon Frias, L. Castillo Blasco, A comparative study of six European databases of medically oriented Web resources, J Med Libr Assoc 93(4) (2005), 467–79.

[2] T. Koch, Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways. Online Information Review 24(1) (2000), 24–34.

[3] M. Douyère, LF. Soualmia, A. Névéol, A. Rogozan, B. Dahamna, JP. Leroy, B. Thirion, SJ. Darmoni, Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 21(4) (2004), 253–61.

[4] LF. Soualmia, B. Dahamna, B. Thirion, SJ. Darmoni, Some Strategies for Health Information Searching. Stud Health Technol Inform 124 (2006), 595–600

[5] SJ. Nelson, WD. Johnson, BL. Humphreys, Relationships in Med. Subject Headings, In Relationships in the org. of knowledge, CA. Bean and R. Green, eds., Kluwer Academic Publishers, (2001), 171–84.

[6] Computation of Related Articles. [Web access] URL:http://www.ncbi.nlm.nih.gov/entrez/query/static/computation.html, Last rev.: Feb 6th, 2003.

[7] W. Kim, AR. Aronson, J.Wilbur, Automatic mesh term assignment and quality assessment. Proc. AMIA symposium (2001), 319–23.

[8] SJ. Darmoni, B. Thirion, F. Ionut-Florea, A. Rogozan, C. Letord, G. Kerdelhué, JN. Dacher, Affiliation of a resource type to a MeSH term in a quality-controlled health gateway Medinfo, 12th World Congress on Health and Medical Informatics (2007), 129:407-11.

[9] D. Lin, An information-theoretic definition of similarity. In Proc. Int. Conf. On Machine Learning (1998), 296–304.

[10] M. Dekkers, S. Weibel, State of the Dublin Core Metadata Initiative, D-Lib Magazine 9(4) (2003).

[11] S. Hoelzer, RK. Schweiger, H. Boettcher, J. Rieger, J. Dudeck, Indexing of Internet resources in order to improve the provision of problem-relevant med. inf. Stud Health Technol Inform 90 (2002),174–7.

[12] G. Salton, C. Buckley, Term-weighting approaches in autom. text retrieval. IPM 24(5) (1988), 513–23.

[13] A. Névéol, K. Zeng, O. Bodenreider, Besides Precision & Recall: Exploring Alternative Approaches to Evaluating an Automatic Indexing Tool for MEDLINE. Proc. AMIA Symposium (2006), 589–93.