KWSim: Concepts Similarity Measure

Youssef MATAR — Elöd EGYED-ZSIGMOND — Sonia LAJMI

20, Avenue Albert Einstein
69621 Villeurbanne Cedex - France
LIRIS Laboratory, INSA - Lyon

youssef.matar@insa-lyon.fr
elod.egyed-zsigmond@insa-lyon.fr
Sonia.lajmi@insa-lyon.fr

ABSTRACT. *The comparison of manually annotated medical images can be done using the comparison of keywords in a lexical way or using the existing medical thesauri to calculate semantic similarity. In this paper, first we introduce the $_{KW}$Sim measure, a fully automated technique of measuring semantic similarity by mapping concepts(keywords) to different medical thesauri and examining the "is-a" relation type. A keyword vector similarity is also presented, based on the $_{KW}$Sim measure. Our approach is implemented using MeSH, ICD-10 and SNOMED CT thesauri and compared with two other existing approaches. We illustrate our method with a real time online annotation assistant.*

RÉSUMÉ. *La comparaison des images médicales annotées manuellement peut être réalisée grâce à une comparaison lexicale entre des mots-clés ou en utilisant des thésaurus médicaux existants pour calculer une similarité sémantique entre ces mots. Dans cet article, nous présentons tout d'abord la mesure $_{KW}$Sim, une technique entièrement automatisée pour le calcul de la similarité sémantique en mappant des concepts (mots-clés) aux différents thésaurus médicaux et en examinant le type de relation « is-a ». Une similarité entre les vecteurs de mots-clés est également présentée, basée sur la mesure $_{KW}$Sim. Notre approche est implémentée en utilisant MeSH, ICD-10 et SNOMED CT thésaurus et comparée avec deux autres approches existantes. Nous illustrons notre méthode avec un assistant d'annotation en ligne et en temps réel.*

KEY WORDS: *keyword vector similarity, medical thesauri, semantic similarity.*

MOTS-CLÉS: *similarité entre vecteurs de mots-clés, thésaurus médicaux, similarité sémantique.*

1. Introduction

This work aims at providing a user assistant for the description and comparison of manually annotated medical images. A medical image is annotated by a vector of keywords that are concepts stemming from medical thesauri (MeSH[1], ICD-10[2] and SNOMED CT[3]). Thus, the comparison of two medical images consists of comparing the semantic similarity degree between the keywords.

Semantic similarity relates to computing the similarity between concepts which are not lexically similar. This is an important problem in Natural Language Processing (NLP) and Information Retrieval (IR) research and has received considerable attention in the literature. Several algorithmic approaches for computing semantic similarity have been proposed. Detection of similarity between concepts is possible if they share common attributes or if they are linked with other semantic concepts in an ontology or medical thesaurus (Li et al., 2003) (Resnik 1999). To relate concepts in different ontologies, semantic similarity works by discovering linguistic relationships between ontological terms across different ontologies (Rodriguez et al., 2003).

We present the $_{KW}$Sim measure, a fully automated technique of measuring semantic similarity between concepts stemming from the same medical thesaurus. A keyword vector similarity is also presented, based on the $_{KW}$Sim measure. We illustrate our method with experimental evaluation and a prototype application enriched with keyword recommendation based on this measure.

The rest of this paper is organized as follows: Section2 discusses the background and related work. Section3 reports the $_{KW}$Sim measure. In section4, we propose a cross thesaurus similarity algorithm. Section5 discusses the keyword vector similarity. Section6 demonstrates the experimental evaluation. Finally, Section7 concludes our work and point outs some future research directions.

2. Background and Related Work

The semantic similarity relates to computing the similarity between concepts (keywords) which are not necessarily lexically similar. Three main categories of algorithms for computing the semantic similarity between terms organized in a hierarchical structure (e.g. MeSH) have been proposed in the literature:

– **Distance-Based algorithms:** The general idea behind the distance-based algorithms (Li et al., 2003) (Leacock et al., 1998) (Wu et al., 1994) (Rada et al., 1989) is to find the shortest path between two keywords in terms of number of edges

1. http://www.nlm.nih.gov/mesh
2. http://www.who.int/classifications/apps/icd/icd10online
3. http://snomed.vetmed.vt.edu/sct/menu.cfm

(nodes) to pass in a given thesaurus in order to get from one to the other. This distance is then translated into a semantic distance.

– **Information Content-Based algorithms:** These algorithms (Seco et al., 2004) (Lord et al., 2003) (Resnik 1999) are inspired by the perception that pairs of words which share many common contexts are semantically related. Thus, the idea of these methods is to quantify the frequency of the co-occurrences of words within various contexts.

– **Feature-Based algorithms:** These algorithms (Petrakis et al., 2006) (Tversky 1997) measure the similarity between two terms as a function of their properties or based on their relationships to other similar terms in the thesaurus where this information is present.

Semantic similarity algorithms can also be distinguished between:

– ***Single Thesaurus*** similarity algorithms that assume that the concepts (keywords), which are compared, are from the same thesaurus (e.g., MeSH).

– ***Cross Thesaurus*** similarity algorithms, which compare concepts from different thesauri (e.g., MeSH and SNOMED CT).

Distance-based and information content-based algorithms are best suited for comparing concepts from the same thesaurus. Cross thesaurus algorithms usually call for feature-based approaches.

3. The $_{KW}$Sim Semantic Similarity Measure

We define our semantic similarity measure $_{KW}Sim$ as a measure based on the *path distance*. The path distance measures the relatedness of two concepts (keywords) by counting the minimal path of nodes between the two concepts through the structural relation of a thesaurus (*is-a* relations). The path distance is based on four factors: d_{K1} is the number of nodes from concept1 (K_1) to the closer common parent in the hierarchical structure and d_{K2} is the number of nodes from concept2 (K_2) to the closer common parent; D is the maximum depth of the thesaurus hierarchical structure; w_1 and w_2 are the weight values for K_1 and K_2 respectively.

The weight values w_1 and w_2 are computed as a function of the depth of the concepts (keywords) K_1 and K_2 in their hierarchical structure:

$$w_1 = \frac{\text{depth}(K_1)}{\text{depth}(K_1) + \text{depth}(K_2)} \quad [1]$$

$$w_1 + w_2 = 1 \mid w_1, w_2 > 0 \implies w_2 = 1 - w_1$$

Now, the path distance is formulated as:

$$pathDist(K_1, K_2) = \begin{cases} \frac{w_1 d_{K1} + w_2 d_{K2}}{(2 * \mathrm{D} * w_1 w_2) + 1} & if\ K_1 \neq K_2 \\ 0 & if\ K_1 = K_2 \end{cases} \quad [2]$$

The $_{KW}Sim$ semantic similarity measure then becomes:

$$_{KW}Sim(K_1, K_2) = 1 - pathDist(K_1, K_2) \quad [3]$$

$$_{KW}Sim(K_1, K_2) \in [0, 1]$$

The $_{KW}Sim$ measure from [3] always returns a value between 0 and 1, where 1 denotes a perfect semantic match between two concepts and zero indicates the absence of match. In addition, the $_{KW}Sim$ semantic similarity measure is symmetric.

Since the existing medical thesauri enclose different design structures, we propose a simple relational database structure that involves a set of common fields between several existing thesauri as follows:

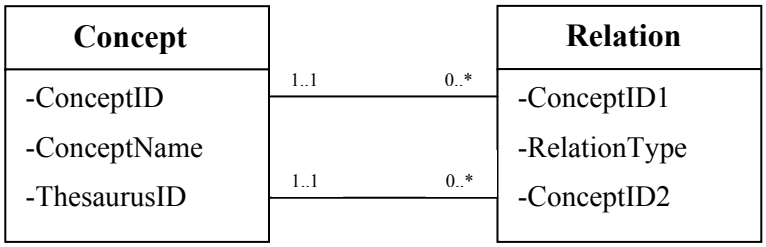

Figure 1. *Database Structure*

In this work, we have only dealt with the *"is-a"* relationship type between concepts. This way the database presents acyclic graphs, where a concept can have several parents and several children.

4. Cross Thesaurus Similarity Algorithm

The following algorithm generates a similarity measure between two concepts from different thesauri. The main idea is to transform the similarity measure computation task from cross thesaurus to a single thesaurus, since our approach is best suited for comparing concepts from the same thesaurus.

1. Let K_1 and K_2 be two concepts from T_1 and T_2 respectively where (T_1 denotes the first thesaurus and T_2 the second).
2. Let R be the set of all ascendants, descendants and direct siblings of K_2 where:
   i. ***a*** is an ascendant of ***b*** if ∃ set of nodes $n_1, n_2, n_3, \dots, n_k \mid n_i$ is the direct parent of n_{i+1} and $n_1 = \boldsymbol{a}$ and $n_k = \boldsymbol{b}$.
   ii. ***a*** is a descendant of ***b*** if ∃ set of nodes $n_1, n_2, n_3, \dots, n_k \mid n_i$ is the direct child of n_{i+1} and $n_1 = \boldsymbol{a}$ and $n_k = \boldsymbol{b}$.
   iii. ***a*** is a direct sibling of ***b*** if ***a*** and ***b*** have the same direct parent.
3. Compute the distance *d* between K_2 and x where:
   a. S is a set of x ; x is a concept | x ∈ R and x ∈ T_1.
   b. *d* is the number of nodes separating the determined concepts.
4. Retrieve x from step 3, having the shortest distance *d*.
5. Compute the similarity measure between K_1 and each x from S via $_{KW}$Sim since the two concepts K_1 and x belong to the same thesaurus T_1.
6. Define an error measure ε based on the distance *d* from step 4.
7. Max_{Sim} is the maximal similarity measure between K_1 and x from step 5.
8. Let Sim_{Sem} be the semantic similarity measure between K_1 and K_2 :

$$Sim_{Sem}(K_1, K_2) = Max_{Sim}(K_1, x) - \varepsilon$$

In the previous algorithm, an error measure ε was proposed due to the indirect comparison between the initial concepts. This error will be defined according to empirical experiments.

5. Keyword Vector Similarity

Since a medical image is annotated by a vector of keywords, thus the comparison of two images requires computing the similarity between its keyword vectors. We have already presented a semantic similarity measure that computes the similarity degree between two keywords; therefore the similarity between two vectors could be formulated as follows based on this measure:

$$R = \sum_{i=1}^{|V1|} max_{j=1}^{|V2|} [\ {}_{KW}Sim(K_{1i}, K_{2j})] \qquad [4]$$

$$S = \sum_{j=1}^{|V2|} max_{i=1}^{|V1|} [\ {}_{KW}Sim(K_{2j}, K_{1i})]$$

$$Sim(V_1, V_2) = \frac{R + S}{|V1| + |V2|}$$

$$Sim(V_1, V_2) \in [0, 1]$$

where V_1 and V_2 denote the keyword vectors. $|V_1|$, $|V_2|$ indicate the number of keywords in V_1 and V_2 respectively. K_{1i} represents the i^{th} keyword in vector V_1 and K_{2j} the j^{th} keyword in vector V_2. Equation [4] always returns a value between 0 and 1, where 1 stands for perfect match and zero indicates absence of match between two vectors of keywords.

6. Experimental Evaluation

For the experimentations, we implemented and integrated our method into an image management system ***PhotoMot*** (Egyed-Zsigmond et al., 2006) (Iszlai et al., 2006) enriched with keyword recommendation based on the $_{KW}$Sim measure (Figure 3). We added our tables to the MySQL database of the system and filled it with the three medical thesauri cited above. This system enables collaborative manual online image annotation with user tracing and assistance.

We also implemented two distance-based algorithms: Leacock & Chodorow algorithm and Wu & Palmer algorithm in order to check the effectiveness of our approach by comparing the three methods.

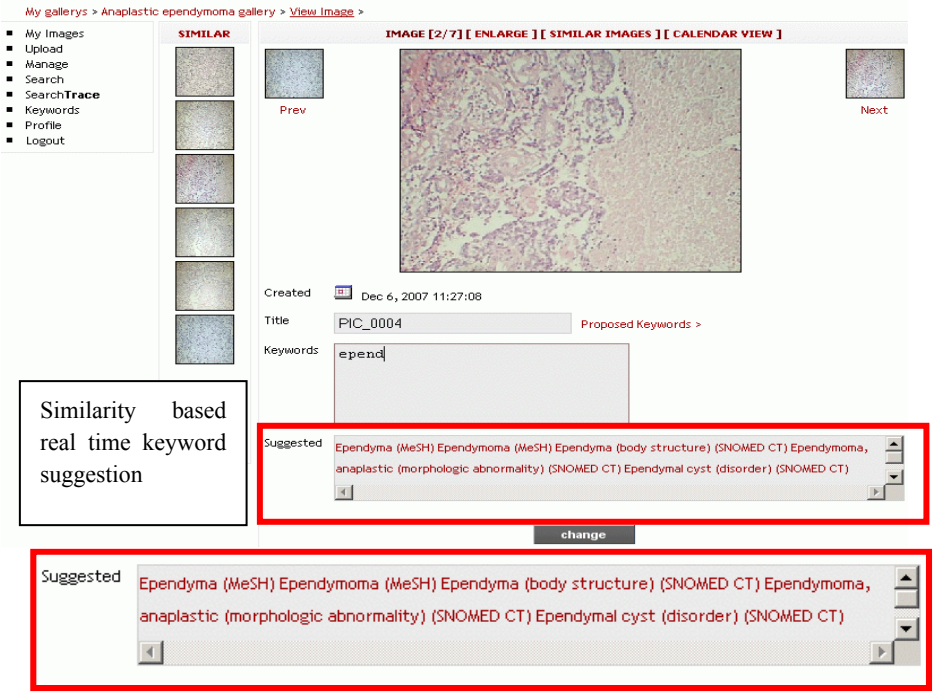

Figure 2. *PhotoMot Screenshot*

For the evaluation, we chose randomly a set of 40 pairs of concepts from MeSH thesaurus and we asked experts from the Tirgu Mures (Romania) University of Medicine to provide an estimate of similarity between 0 (not similar) and 10 (perfect similarity) for each pair. The similarity values obtained by each of the implemented methods are correlated with the average scores obtained by the humans. The correlation results are summarized in Table 3. These results show that our proposed method achieves 7% better correlation than Wu & Palmer method and 2% better correlation than Leacock & Chodorow method.

Method	**Correlation**
Wu & Palmer	0.84
Leacock & Chodorow	0.89
$_{KW}$Sim	0.91

Table 1. *Evaluation of the implemented methods on MeSH*

Indeed it was quite difficult for the experts to provide an estimate similarity value between two concepts. We are currently working on other experimentations that overpass this problem. We ask our medical partners to annotate several images with several keyword vectors in order to calculate the similarity degree between these vectors.

7. Conclusion and Perspectives

In this paper we presented several similarity measures and algorithms. These measures give semantic similarity between concepts present in one or several thesauri structured hierarchically as an acyclic graph by *"is-a"* relations. We have extended this measure in order to calculate the similarity between keyword vectors and keywords stemming from different thesauri.

We have implemented a test application and several other states of art similarity measure functions and compared the results with our method. We used human experts to evaluate the similarity results and found that our measure is at least as good as the other functions. We apply our measures to improve the annotation and search assistance in a collaborative online medical image management system.

We study the inclusion of contextual knowledge in the comparison of keyword vector annotated images. We also study the use of general purpose thesauri (WordNet) to be able to compare sentences.

Another important perspective is the evaluation, configuration and optimization of cross thesauri concept vector comparison.

8. References

Egyed-Zsigmond E., Lajmi S., Iszlai Z., "Concurrent use in an image management system." *13th ISPE INTERNATIONAL CONFERENCE ON CONCURRENT ENGINEERING: RESEARCH AND APPLICATIONS*, 2006, p. 403-417.

Iszlai Z., Egyed-Zsigmond E., "User centered image management system for digital libraries." *2nd IEEE International Conference on Document Image Analysis for Libraries*, 2006, p. 164-171.

Karanastasi A., Christodoulakis S., "The OntoNL Semantic Relatedness Measure for OWL Ontologies", *In Proceedings of the 2nd International Conference on Digital Information Management (ICDIM'07)*, 2007, p. 333-338.

Leacock C., Chodorow M., "Combining local context and WordNet similarity for word sense identification." *In Christiane Felbaum, editor, An Electronic Lexical Database,* 1998, p. 265-283.

Li Y., Bandar Z.A., McLean D., "An approach for Measuring Semantic Similarity between Words Using Multiple Information Sources", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15 no. 4, 2003, p. 871-882.

Lord P.W., Stevens R.D., Brass A., Goble C.A., "Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation", *Bioinformatics*, vol. 19 no. 10, 2003, p. 1275-1283.

Petrakis E., Varelas G., Hliaoutakis A., Raftopoulo P., "Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies", *4th Workshop on Multimedia Semantics (WMS'06)*, 2006, p. 44-52.

Rada R., Mili H., Bicknell E., Blettner M., "Development and Application of a Metric on Semantic Nets", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19 no. 1, 1989, p. 17-30.

Resnik P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language", *Journal of Artificial Intelligence Research,* vol. 11, 1999, p. 95 - 130.

Rodriguez M.A., Egenhofer M.J., "Determining Semantic Similarity among Entity Classes from Different Ontologies", *IEEE Transactions on Knowledge and Data Engineering*, vol. 15 no. 2, 2003, p. 442 - 456.

Seco N., Veale T., Hayes J., An intrinsic Information Content Metric for Semantic Similarity in WordNet, Technical report, 2004, Dublin City University, Ireland.

Tversky A., "Features of Similarity", *Psychological Review*, vol. 84 no. 4, 1997, p. 327 - 352.

Wu Z., Palmer M., "Verb semantics and lexical selection", *In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994, p. 133-138.