19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Graph-based methods for Significant Concept Selection

Gasmi Karim[a,*], Torjmen-Khemakhem Mouna[a], Tamine Lynda[b], Ben Jemaa Maher[a]

[a]*ReDCAD Laboratory, ENIS Sokra km 3.5 , University of Sfax, TUNISIA*
[b]*IRIT Laboratory, University of Paul Sabatier, TOULOUSE*

Abstract

It is well known in information retrieval area that one important issue is the gap between the query and document vocabularies. Concept-based representation of both the document and the query is one of the most effective approaches that lowers the effect of text mismatch and allows the selection of relevant documents that deal with the shared semantics hidden behind both. However, identifying the best representative concepts from texts is still challenging. In this paper, we propose a graph-based method to select the most significant concepts to be integrated into a conceptual indexing system. More specifically, we build the graph whose nodes represented concepts and weighted edges represent semantic distances. The importance of concepts are computed using centrality algorithms that levrage between structural and contextual importance. We experimentally evaluated our method of concept selection using the standard ImageClef2009 medical data set. Results showed that our approach significantly improves the retrieval effectiveness in comparison to state-of-the-art retrieval models.

Keywords: information retrieval, Natural Language Processing (NLP), semantic similarity, concept selection;

1. Introduction

In information retrieval (IR)[1] area, traditional document indexing methods are based on the concept of bag of words. The latter lead to flat-based text representation relying on the distribution of word occurrence both in the query and in the document collection. While this indexing model served as a building block to very effective retrieval models such as the vectorial[2] and the language modelling[3] model, it faces several limitations. One of the limitations addressed in this paper concerns the semantic representation of document and query contents. For instance, basic natural language properties such as synonymy and polysemy are not considered in the retrieval model design and then lead to the selection of irrelevant results according to a query. To tackle this limitation, several research studies explored the use of semantic representations of both query and documents contents[4]. These representations are based either on word-based correlations or dictionaries[5] or concepts issues from reference terminologies or ontologies[6]. The latter rely heavily on concept extraction methods of terminological concept-entries that are generally based on the assumptions that: 1) the mapping between texts and terminologies is accurate, and 2) the semantic relations embedded

* Corresponding author. Tel.: +216 22 492 152 ; fax: +216 74 275 595.
E-mail address: gasmikarim@yahoo.fr

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.170

within the resource are valid. However, a previous work[7] shows that these assumptions are not systematically true, particularly, in the case of UMLS terminology. After a thorough study on this meta-thesaurus and METAMAP as a concept extraction tool, authors in[8] observed that the embedded concept extraction method suffers from several drawbacks. The authors showed that the inaccurate concepts generated by METAMAP significantly affect the results of the retrieval model and the annotation quality of some queries. Moreover, authors in[7] found that some ontological relationships are wrong or undefined. Among the several cases that can produce erroneous relations, we cite:

- Cases where the semantic category of the child is very broad whereas the parent's semantic type is too specific;
- Situations where the parent-child relationship is erroneous;
- Cases where a parent-child relationship is lacking and has to be added to the UMLS semantic network;
- Conditions where the parent or the child is missing in a semantic category;

To address these limitations, we propose in this paper a method for selecting the best significant document concepts. In fact, the indexing step will be based on those selected concepts. We use a traditional NLP method as a starting point for selecting a candidate concept and then build the corresponding document-based graph concept where the semantic relations are leveraged from document context and terminology structure. Moreover, we apply a centrality algorithm in order to weight the concepts according to their importance in the document. Experimental evaluation over a ImageClef2009[1] dataset shows the effectiveness of our approach in the medical domain.
The remainder of this paper is organized as follows: section 2 reviews the related work. Section 3 details our approach for the concepts selection. Section 4 presents and discusses the experimental evaluation of our method, based on the standard ImageClef2009 medical collection. Finally, we draw our conclusion in Section 5.

2. Related work

We review below two lines of related work: 1) concept extraction from medical documents and 2) concept-based techniques involved in both document indexing and ranking within an IR setting.

2.1. Concept Extraction Approaches

Over the recent years, several tools have been developed to map medical texts to concepts such METAMAP[9], MicroMeSH[10], CHARTLINE[11], CLARIT[12], and SAPHIRE[13]. Each of these systems employed one or more of the following features: lexical analysis (more often using a specialized lexicon), syntactic analysis or a mapping procedure accounting for partial matching. These tools are generally based on the UMLS Metathesaurus as the target knowledge source, rather than a smaller source such as MeSH.
Unlikely, authors in[14] propose the MaxMatcher tool based on a dictionary-based matching and relying on (MeSH, SNOMED, ICD_10) multiterminology. Given a document, MaxMatcher extracts a set of terms or phrases denoting domain concepts as well as their corresponding concept unique identifiers (CUIs). With Maxmatcher, the search for a string in a dictionary of concepts can be exact with MeSH or approximate with UMLS. Because the MeSH thesaurus is maintained by an organization (NLM), it contains no ambiguous terms denoting the concepts. For the UMLS, ambiguity arises from the fact that only one term may be presented by several concepts. However, MaxMatcher does not measure the importance of each concept for describing the document semantic. For that, authors in[14] proposed a strategy which is mainly based on ranking concepts extracted from documents using a combined score, it involves three steps: (1) computing a content-based matching score between a concept and a document, (2) computing a rank correlation to compute the word rank correlation between words in a document and a concept, (3) selecting the document semantic kernel by ranking the concepts according to their combined score.
Authors in[15] proposed a WSD component to select the most significant concepts issue from UMLS, which is an implementation of the WSD knowledge-based system. In this component, the appropriate semantic type of a target word is determined with the assumption that each of the possible concepts of the target word has a unique semantic

[1] Cross Language Evaluation Forum

type. So, for all of the concepts of the target word, the authors propose to create the first-order concept vector. The elements in the vector indicate whether or not the CUI of the term is assigned to that semantic type. In fact, journal Descriptor Indexing JDI is the basis for selecting the best meaning that is correlated to UMLS semantic types (STs) assigned to ambiguous concepts in the Metathesaurus. Then, the authors create the test vector whose elements indicate whether or not a feature is one of the semantic types of the words surrounding the target word. Next, the vectorial similarities between the test vector and each of the concept vectors are computed using the cosine Measure. The latter allows determining the concept from the concepts vector which is closest to the test vector. Those concepts will be assigned to the target word.

2.2. Concept-based document representation and ranking

The basic idea behind concept-based document representation is to use the concepts extracted from the document as information units, making the difference between the word-based information units and then apply a concept-document weigthing schema. In Baziz[4] authors proposed an indexing method based on a method of concept-extraction and then a CF.IDF weithing schema. The latter is a revised form of the traditional TF.IDF schema[16]. The core idea behind the CF.IDF weighting schema is that it allows to weight simple terms and compound terms associated with concepts. Indeed, in this approach, the weight of a compound term is based on the cumulative frequency of the term itself and its components.
In the medical domain, concepts are generally extracted using METAMAP from UMLS or MeSH terminologies and then the correpsonding preferred terms are used to index the documents[17]. There are two methods for the expansion of documents/query: (1) an expansion based on relevance feedback (called local context based expansion) and (2) an expansion based on adding semantic concepts from an external resource (called global context based expansion). In this paper, we focused on the second kind, which is the most suitable for our work. Authors in[18] proposed an indexing method based on a multi-terminology to index the hospital output summaries of patients. It is an indexing method based on a concept extraction method using biomedical symbolic knowledge as ontologies and statistical knowledge extracted from a field of application. The extracted terms are ordered to highlight their importance in the document. The importance of a term is determined by the number of relationships it shares with other terms. Relations between concepts can be exploited from the meta-thesaurus UMLS and co-occurrence relationships between concepts from one or several terminologies. In[19] authors assessed the impact of the document/query expansion in the medical domain by exploiting the concepts and their semantic relationships in UMLS. Their expansion method deals with the case of general- specific relatedness between query concepts and document concepts. They chose all concepts connected to query or document concepts by direct IS-A relation to expand because this type of relation is found between the concepts strongly connected by the notion of general-specific. More specific concepts are added to queries while more general concepts are added to documents, in the purpose of increasing the matching concepts subsets. They chose to apply DFR (Divergence From Randomness) proposed by[20] like weighting to documents and queries indexed by UMLS concepts.
Likewise, authors in[21] exploited several medical knowledge sources such as MeSH, SNOMED, and UMLS, for expanding the query with synonyms, abbreviations and hierarchically related terms identified by using the PubMeds automatic term mapping service. Furthermore, they also defined several rules for filtering the candidate terms according to each knowledge source. Some authors, like in[22], proposed to combine between the both expansion type. They proposed a knowledge-intensive conceptual retrieval by combining both the global context (i.e., concepts in several terminoontological resources such as MeSH, Entrez Gene, ADAM) and local context (top-ranked documents).

3. A Graph-based method for concept selecting

3.1. Motivation and contribution

The METAMAP Transfer (MMTx) concept extraction method has several limits reported by several studies[8,15]. We present two main drawbacks that lead to irrelevant concept-based representations of documents and consequently to inaccurate document-query matching. The first drawback is related to the fact that the core concept extraction method triggers the problem of over-generation. For example, given the noun phrase ”ocular complications,” the METAMAP

combines three concepts ”Ocular”, ”Complications” and ”Complications Specific to Antepartum or Postpartum” because they share at least one word. The second drawback is related to the strict comparison between the terms and nominal group entries in the UMLS. This strict comparison causes the problem of under-generation of relevant variants. For example, for the phrase ”gyrB and p53 protein,” METAMAP can not identify ”gyrB” as a protein because it is recorded as ”gyrB protein” in MeSH or UMLS.
To solve this problem of non-significant concept selection, authors in[23] adapted the graph-based term weighting method proposed by[24] and apply it to concepts. In most of works that use the graph-based representation[23,25], only one semantic relation between concepts is assumed to be accurate and used to compute the importance of their relatedness. The latter is generally based on the co-occurrence of the concept-entries in the text or on relations issued from external semantic resources.
To overcome this limitation, we propose in this paper a method that selects the most significant concepts associated with a document. In fact, those significant concepts will be the basis for document indexing. Our method is based on different sources of evidence to compute their relatedness in order to alleviate the topic drift that could be induced by using only one of them. The candidate concepts issued from a METAMAP are first organized into a graph. Second, an importance of concept relatedness is computed using both terminological and document content as sources of evidence. Finally, a graph-based voting algorithm allows identifying the best significant concepts to be retained for representing the document under consideration.

3.2. The general framework

The general IR process presented in Figure 1 highlights the following steps: (1) preprocessing and concept extraction using METAMAP, (2) building a semantic graph, (3) selecting the significant concept.

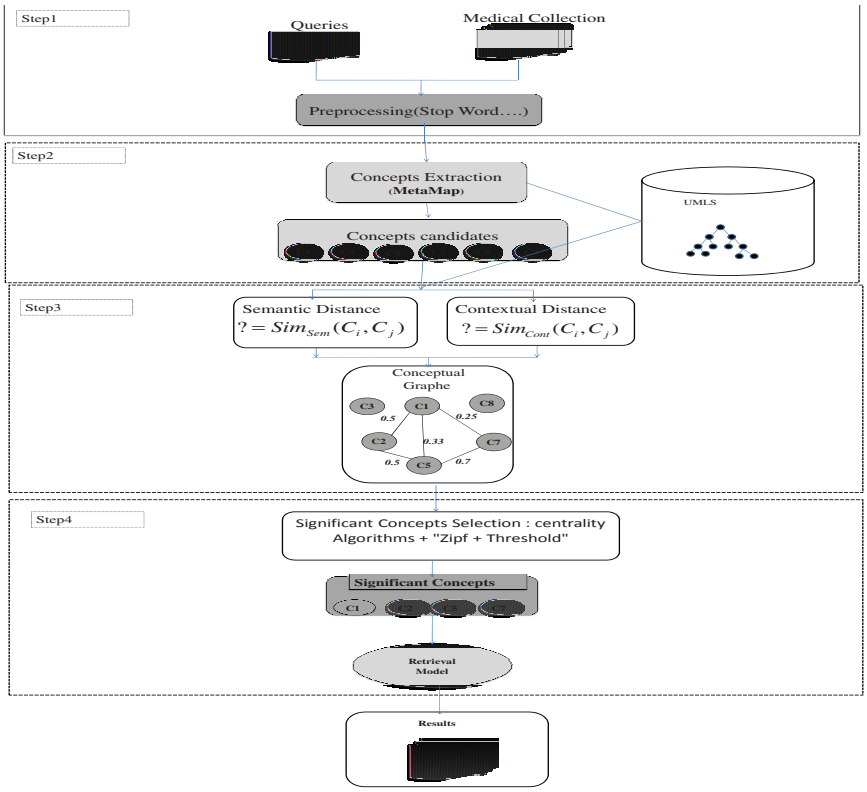

Fig. 1. Overview of the proposed method process

We detail below the main steps of our approach related to the semantic graph building and significant concepts selection.

3.3. Semantic graph building

We note $G = (C, R)$ an undirected semantic graph made up of a concept set C and a semantic relation set R. An arc connects two concepts C_i and C_j having a semantic similarity weighted according to a score computed using a combined similarity distance detailed below. Formally, we consider:

- A set $C = \{C_1, C_2, ..., C_n\}$ whose elements are called nodes,
- A set $R = \{R_1, R_2, ..., R_m\}$ whose elements are called nodes arcs.

the adjacency matrix $m_{i,j}$ of a graph G is defined by $m_{i,j}$:

$$m_{i,j} = \begin{cases} Rel(C_i, C_j)\ if\ Rel(C_i, C_j) \neq 0 \\ 0\ else \end{cases}$$

Relations between the two concepts C_i and C_j are classified as follows:

$$Rel(C_i, C_j) = Dist(C_i, C_j) \tag{1}$$

In the literature, semantic distances are classified into three types:

- The hierarchical distance[26]: it is based on the semantic resource hierarchy and computed using the number of edges between concepts.
- The contextual distance[27]: it relies on the information content, and computed using the comparison between the definition of the two concepts.
- The hybrid distance : it is the combination of hierarchical and contextual distances.

For the contextual distance, we propose to use the ontology-independent Context Vector measure proposed by Pederson in[28]. The authors in[28] have shown the efficiency of this measure compared to other distances. This measure is computed as follows:

$$Dist(C_i, C_j) = \frac{E(C_i), E(C_j)}{||E(C_i)||.||E(C_j)||} \tag{2}$$

Where $E(C_i)$ and $E(C_j)$ are the vectors associated of concept terminological definitions. Each concept definition $E(C_i)$ is represented using a vector of terms: $E(C_i) = (w_{1,i}, w_{2,i}, w_{3,i}, ..., w_{k,i})$.

Where $||E(C_i)|| = \sqrt{w_{1,i}^2 + ... + w_{k,i}^2}$; And $w_{k,i}$ is the number of occurrences of k^{th} term in concept definition $E(C_i)$;

For the hierarchical distance, we propose to use the rada distance[26] which is computed as follows:

$$Dist(C_i, C_j) = \frac{1}{Path(C_i, C_j)} \tag{3}$$

Where $Path(C_i, C_j)$ is the shortest path between concepts C_i and C_j
The main strengh of this method is its low computation complexity. Taking into account the concept relation problems, we suggest to use the two types of distance jointly: hierarchical and contextual. This hybridization is done by a combination after a normalization step of the two distances taking into account all concepts weight in each document.

$$Dist(C_i, C_j) = Dist_{Sem}(C_i, C_j) + Dist_{Con}(C_i, C_j) \tag{4}$$

3.4. Significant concept selection

3.4.1. Concept weighting

Our objective here, is to select the best significant concept of the document. To achieve this goal, we compute the importance of the concepts based on the semantic graph-based representation detailed above. We apply a centrality

algorithm in order to weight each concept according to their importance in the document. The basic idea held through a centrality graph algorithm is that the importance of a node in a graph can be determined by taking into account the relationship between the node under consideration and the whole related nodes in the graph. In our experiments, we used three centrality algorithms: Closeness, Betweenness and PageRank.

- Closeness centrality focuses on how close a node is to all the other nodes in a graph. This algorithm describes the extent of influence of a node on the graph[29]. Closeness centrality of a node C_i is the reciprocal of the sum of the shortest path distances from C_i to all $N-1$ other nodes. Since the sum of the distances depends on the number of nodes in the graph, closeness is normalized by the sum of the $(N-1)$ minimum possible distances[30].

$$Weight(C_i) = \frac{(N-1)}{\sum_{j=1}^{N-1} Dist(C_i, C_j)} \tag{5}$$

  With $C_i \neq C_j$, and N is the number of nodes in the graph. In the weighted graphs built in our experiments, we use a weighted version of the closeness measure, which takes into account the weights on the edges while computing the shortest path.

- Betweenness centrality, according to Borgatti[31], is defined as "the share of times that a node i needs a node k (whose centrality is being measured) in order to reach j via the shortest path". The more times a node lies on the shortest path between two other nodes, the more control that the node has over the interaction between these two non-adjacent nodes[32]. This is achieved for all the nodes of the graph in the eq 6. Here, it computes the relatedness degree between a concept and the other extracted concepts in the document.

$$Weight(C_i) = \sum_{C_j, C_k \in C} \frac{\sigma_{C_j,C_k}(C_i)}{\sigma_{C_j,C_k}} \tag{6}$$

  where C is the set of concepts; σ_{C_j,C_k} represents the total number of shortest geodesic paths between C_j and C_k; $\sigma_{C_j,C_k}(C_i)$ is the number of those paths passing through a node C_i other than C_j and C_k. We apply the betweenness centrality of a weighted graph concept, where a weight of an arc corresponds to a similarity between two concepts.

- PageRank is the third measure we propose to use[33]. The main idea implemented by PageRank is that of "vote" or "recommendation". When a node is connected to another, it is essentially voting for the other node. The more votes for a node, the greater the importance of the node increases. Although PageRank was originally defined for directed graphs, it can also be applied to undirected graphs. The PageRank score associated with a node is defined using a recursive function. The weight of C_i within a document is initially set to 1 and the following PageRank function is run for several iterations. Here, the high pageRank measure for a candidate concept highlights the importance of the considered concept, considering the fact that it is related to many others concepts in the documents.

$$Weight(C_i) = \frac{1-d}{N} + d * \sum_{C_j \in C} \frac{Weight(C_j)}{\mid Degree(C_j) \mid} \tag{7}$$

  where :

$$Degree(C_j) = \sum_{C_i \in C} Dist(C_j, C_i) \tag{8}$$

– d is a parameter that lies between 0 and 1. A typical value for d is 0.85, and this is the value we use in our implementation; N is the total number of concepts; C is the set of concepts which are connected to C_j.

3.4.2. Concept selecting

At the stage of the concept selection, a cutoff score is used allowing us to eliminate the non-significant concepts. We use the dyadic subdivision (9) to calculate each threshold. Then the concept that has a weight above the threshold is selected.

$$X(i,n) = Weight_{min} + \frac{i * (Weight_{max} - Weight_{min})}{2^n} \quad (9)$$

Where $Weight_{min}$ and $Weight_{max}$ are respectively the min and max values of the concepts-weights in a document. This weight was generated by the centrality algorithms; n is the numbre of landmark partition; i is the partition N°.
It should be noted that the values of these thresholds have been determined empirically by studying the regularity of the relationships frequency between a concept and others-concepts after the graph building. Indeed, the graphical representation of the distribution Zipfian[34] relative to the number of concepts according to their weight shows a decreasing curve, which is traditionally divided into three zones: (i) a first area describing the trivial information represented by the general concepts, the marginal information and noise, illustrated by few concepts; (ii) a second zone containing interesting information represented by the concepts used to construct the graph related with different document topics; and, (iii) a third zone representing a very significant information, illustrated by the significant concepts connecting the majority of the graph nodes. The three zones are noted to be contiguous and that the second and the third zones are the target of extracting the most relevant concepts from the texts. The threshold allows the elimination of the area of insignificant concepts. The remaining areas are those with significant concepts used for indexing the collection of documents.

4. Evaluation

The objectives of the evaluation are (1) to study the impact of centrality measures as concept-document weighting measures and(2) to measure the effectiveness of our method for selecting the significant concepts.

4.1. Data sets and Evaluation metrics

To evaluate our approach, we used the ImageClef 2009 [2] test collection including: 1) 74,902 medical images and annotations associated with them. This collection contains images and captions from two Radiological Society of North America (RSNA)[3] journals; 2) a set of 25 queries selected by medical experts; for each query is assigned a list of relevant documents assessed by human assessors involved in the CLEF evaluation campaign.
To evaluate our indexing approach, we use the vector model as a retrieval model[35]. For measuring the IR effectiveness, we used i) P@5, P@10 representing respectively the mean precision values at the top 5, 10 returned documents and ii) MAP representing the Mean Average Precision calculated over all queries.

4.2. Impact of the centrality algorithm and the threshold

Table 1 presents the IR performance using three centrality algorithm. For each centrality algorithm, we have applied four thresholds according to the concepts distribution. Hence, each threshold X (i, N) belongs to a part of the Zipfian graph and a threshold X(i=0, n=0) presents the classical indexing, denoted *Baseline*. In fact, each threshold was determined by equation 9.

According to the results, we observe that PageRank algorithm returned a significant result compared to the results obtained by the traditional indexing,ie, Baseline, for the threshold of X (1,4). While increasing the threshold value, we

[2] http://www.imageclef.org/
[3] http://www.rsna.org/

Table 1. Comparison between different centrality algorithms according to the thresholds

Category	P@5	P@10	MAP	Category	P@5	P@10	MAP	Category	P@5	P@10	MAP
PageRank				**Betweneess**				**Closeness**			
Baseline	0.3920	0.3440	0.2163	**Baseline**	0.3920	0.3440	0.2163	**Baseline**	0.3920	0.3440	0.2163
X(7,3)	0.19	0.16	0.04	**X(1,2)**	0.260	0.195	0.069	**X(7,3)**	0.1889	0.2111	0.0775
X(5,3)	0.291	0.258	0.170	**X(1,4)**	0.358	0.325	0.155	**X(1,2)**	0.3130	0.3217	0.1651
X(1,2)	0.425	0.375	0.185	**X(2,4)**	0.416	0.425	0.229	**X(1,3)**	0.4333	0.4333	0.2392
X(1,4)	**0.466**	**0.458**	**0.26**	**X(1,6)**	**0.391**	**0.437**	**0.230**	**X(1,4)**	**0.4583**	**0.4458**	**0.2529**

have observed that the MAP generated by the retrieval system changed. This variation of the MAP over the threshold is presented in the figure 2.

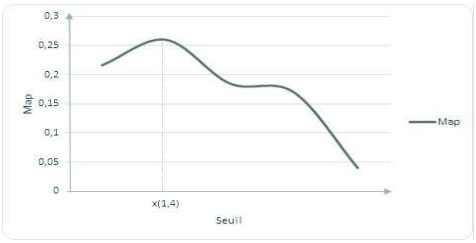

Fig. 2. MAP variation relative to the cutoff threshold

For the thresholds belonging to the interval of] 0,X (1,4)], the system based on the PageRank algorithm generates better results than the baseline. The MAP was increased simultaneously with the augmentation of the threshold until obtaining a MAP of 0.26 and a threshold of X (1,4). Those results are explained by the fact that when we increase the threshold value, we will remove the non-significant concepts. Indeed, the result will be impacted either negatively or positively depending on the threshold rate. So we found a negative effect of non significant concepts on RI system results. Thus, the positive effect of our concepts-selection method on the RI model result compared with the classical indexing. Whereas in the interval of [X (1,4), X(n,n)], the MAP value has decreased when increasing the threshold value. In this part of the curve 2, the significant concepts will be eliminated. Indeed, the result will be negatively impacted. For the remaining experiments, we have chosen X(1,4) as the best threshold.

4.3. Evaluation of retrieval effectiveness

The purpose of these experiments is to determine the effectiveness of the concept selection method. Hence, we carried out three series of experiments: the first one is based on the classical conceptual indexing of documents using the well known weighting scheme BM25, as the baseline, denoted *Baseline*. The second one concerns the concept selection method proposed by Humphrey in [15], denoted *WSD MM*. The third one concerns our indexing approach and consists of three scenarios:

- the first one concerns the graph-based method using the hierarchical distance without taking into account the contextual relationship between concepts, denoted *Hier-graph-based*,
- the second one concerns the documents indexing using concepts identified by the graph-based method using the contextual distance, denoted *cont-graph-based*.
- the third one concerns the documents indexing using concepts identified by the combination of the hierarchical and contextual distance, denoted *hyb-graph-based*.

We computed the paired-sample T-tests between means of each ranking obtained by each indexing method based in the concept selection method and the baseline, in order to test the significance of the results. We assume that the difference between two given rankings is significant if $p \prec 0.1$ (noted *) and very significant if $p \prec 0.05$ (noted

**). Table 2 presents the MAP results over the baselines and the different retrieval scenarios. As we can see, the paired-sample T-test shows that our best concept selection approach (*hyb-graph-based*) for indexing is statistically significant compared to the baseline.

Table 2. Comparison between our significant concept selection method and other methods

		Improvement rate				**Improvement rate**				**Improvement rate**		
	P_5	MM	WSD MM	Hier-Graph -based	P_10	MM	WSD MM	Hier-Graph -based	MAP	MM	WSD MM	Hier-Graph -based
Baseline	0.3920	–	–	4.5%	0.3440	–	–		0.2163	–	+0.13%	
WSD MM	0.4400	+12% *	–	17.3% *	0.428	+24% **	–	14.3%	0.2160	–	–	
Hier-Graph -based	0.3750	–	–	–	0.3750	+9%	–	–	0.2271	%4.9	%5.13	–
Cont-Graph -based	0.448	14.2% *	1.8%	19.4% *	0.432	+25.5% **	0.9%	15.2% *	0.245	13.4% *	13.1% *	10%
Hyb-graph -based	**0.4667**	**+19%** **	**+6.6%** *	**+24.45%** **	**0.4583**	**+33.2%** **	**+7.07%** *	**+22.21%** **	**0.260**	**+20%** **	**+20.37%** **	**+14.48%** *

According to the results, we can see that the results obtained by the different concepts-selection methods are better than those obtained using the traditional METAMAP tool. This observation confirms that the use of our concepts selection method improves the indexing and therefore the retrieval effectiveness. These results are supported by the presence of concepts that are out of context. In an indexing method which does not use a selection method, the non-significant concepts negatively impacted the retrieval model.
As far as, if the measurement P @ 5 is concerned, our selection model outperforms the WSD METAMAP model by 6.6%. As for the measurement of MAP, our model outperforms the model WSD METAMAP by 19%. Therefore, we may conclude that the use of conceptual relationships to disambiguate concepts is a suitable solution.
According to the results, we note that the use of hybridization of both distances helps to correct the lack of a relationship in the UMLS. We can explain these results by the problems at the semantic similarity and relatedness between the concepts, either for those based on the arc or for those based on information content. The semantic weight with a hierarchical distance between two concepts is not always exact and does not always reflect the real semantic degree between two concepts. For example, if a child-parent relation between two concepts is missed, hierarchical distance score will be affected. Moreover, when using contextual distance, the score can be affected, due to the term ambiguity problem of concept definitions.

5. Conclusion

We have proposed in this paper a graph-based method to select the most significant concepts using an external semantic resource. We argued that the significant concepts-selection can be seen as a graph problem taking into account (i) the semantic similarity and relatedness between the identified concepts and (ii) the importance of a concept in a graphe calculated by a graph centrality algorithm. The selected concepts are used to index the collection in an attempt to close the semantic gap between the users query and documents in the collection. The results demonstrate that our graph-based concept selection approach provides a significant improvement over a state-of-the-art IR baseline approach.
Our future work aims at incorporating our concept selection method into a semantic information retrieval model for a global context based expansion with the preferred terms denoting significant concepts, which we believe to be able to overcome the limits of the bag-of-words based models. In addition, we also plan to use semantic distances other than those used in this work and to combine between them to avoid the semantic distance problems mentioned in this paper.

References

1. Bayeza, R., Ribeiro-Neto, B.. *Modern information retrieval*. Addisson Wesley; 1999.
2. Salton, G., McGill, M.J.. *Introduction to modern information retrieval*. McGraw-Hill computer science series. New York: McGraw-Hill; 1983.
3. Ponte, J.M., Croft, W.B.. A language modeling approach to information retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA; 1998, p. 275–281.
4. Baziz, M.. Indexation conceptuelle guidee par ontologie pour la recherche dinformation. In: *Thse de doctorat, Universite Paul Sabatier, Toulouse, France*. 2005, .
5. Krauthammer, M., Rzhetsky, A., Morozov, P., Friedman, C.. Using blast, A DNA and protein sequence comparison tool, for finding gene and protein names in journal articles. In: *American Medical Informatics Association Annual Symposium*. 2000, p. 245252.
6. Xiaoyue, W., Rujianga, B.. Rdf ontologies to improve text classification. In: *CINC*. 2009, p. 118–121.
7. Cimino, J.J., Min, H., Perl, Y.. Consistency across the hierarchies of the umls semantic network and metathesaurus. *Journal of Biomedical Informatics* 2003;**36**(6):450–461.
8. Wang, Y., Liu, X., Fang, H.. A study of concept-based weighting regularization for medical records search. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL*. 2014, p. 603–612.
9. Aronson, A.R.. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Annual Symposium AMIA* 2001; :17–21.
10. Elkin, P., Cimino, J., Lowe, H., Aronow, D., Payne, T., Pincetl, P., et al. Mapping to mesh: The art of trapping mesh equivalence from within narrative text. 1988, p. 185–190.
11. Miller RA Gieszczykiewicz FM, V.J., GF, C.. Chartline: Providing bibliographic references relevant to patient charts using the umls metathesaurus knowledge sources. 1992, p. 86–90.
12. Evans, D., Webster, G.K., Hart, M., Lefferts, R., Monarch, I.. Automatic indexing using selective NLP and first-order thesauri. In: *Proceedings of RIAO-91*. 1991, p. 624–643.
13. Hersh WR Hickam DD, H.R., bon KA, M.. A performance and failure analysis of saphire with a medline test collection. 1994, p. 51–60.
14. Dinh, D., Tamine, L.. Voting techniques for a multi-terminology based biomedical information retrieval. In: *AIME*. 2011, p. 184–193.
15. Hliaoutakis, A., Zervanou, K., Petrakis, E.G.. The {AMTEx} approach in the medical document indexing and retrieval application. *Data and Knowledge Engineering* 2009;**68**(3):380 – 392.
16. Robertson, S.E., Walker, S.. On relevance weights with little relevance information. In: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA; 1997, p. 16–24.
17. Dinh, D., Tamine, L.. Biomedical concept extraction based on combining the content-based and word order similarities. In: *Proceedings of the ACM Symposium on Applied Computing*. New York, NY, USA; 2011, p. 1159–1163.
18. Avillach, P., Joubert, M., Fieschi, M.. A model for indexing medical documents combining statistical and symbolic knowledge. 2007, p. 31–5.
19. Le, D.T.H., Chevallet, J.P., Thuy, D.T.B.. Thesaurus-based query and document expansion in conceptual indexing with umls: Application in medical information retrieval. In: *In Research, Innovation and Vision for the Future*. IEEE; 2007, p. 242–246.
20. Amati, G., Van Rijsbergen, C.J.. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans Inf Syst* 2002;**20**:357–389.
21. Stokes, N., Li, Y., Cavedon, L., Zobel, J.. Exploring criteria for successful query expansion in the genomic domain. *Inf Retr* 2009; **12**(1):17–50.
22. Zhou, W., Yu, C., Smalheiser, N., Torvik, V., Hong, J.. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2007, p. 655–662.
23. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.. Graph-based concept weighting for medical information retrieval. In: *The Seventeenth Australasian Document Computing Symposium, ADCS'12, Dunedin, New Zealand*. 2012, p. 80–87.
24. Blanco, R., Lioma, C.. Graph-based term weighting for information retrieval. *Inf Retr* 2012;**15**(1):54–92.
25. Sajgalik, M., Barla, M., Bielikova, M.. From ambiguous words to key-concept extraction. In: *International Workshop on Database and Expert Systems Applications,Prague, Czech Republic*. 2013, p. 63–67.
26. Rada, R., Mili, H., Bicknell, E., Blettner, M.. Development and application of a metric on semantic nets. In: *IEEE Transaction Systems Man and Cybernetics*. 1989, p. 17–30.
27. Jay J. Jiang, D.W.C.. Semantic similarity based on corpus statistics and lexical taxonomy. In: *In International Conference on Research in Computational Linguistics (ROC)*. 1997, p. 19–33.
28. Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G.. Measures of Semantic Similarity and Relatedness in the Biomedical Domain. *Journal of Biomedical Informatics* 2007;**40**(3):288–299.
29. Chaoqun Ni, C.R.S., Jiang, J.. Degree, closeness, and betweenness: Application of group centrality measurements to explore macro-disciplinary evolution diachronically. In: *In Proceedings of ISSI*. 2011, p. 1–13.
30. Freeman, L.C.. Centrality in social networks conceptual clarification. *Social Networks* 1978;:215.
31. Borgatti, S.P.. Centrality and network flow. In: *Social Networks*. 2005, p. 55–71.
32. Brandes, U.. A faster algorithm for betweenness centrality. In: *Journal of Mathematical Sociology*; vol. 25. 2001, p. 163–177.
33. Brin, S., Page, L.. The anatomy of a large-scale hypertextual web search engine. In: *Comput. Netw. ISDN Syst.*; vol. 30. Amsterdam, The Netherlands, The Netherlands; 1998, p. 107–117.
34. Lafouge, T., Boukacem, B.. Applications des lois informatiques en sciences de linformation : dualit, champ informatique dusage et de production. *ISDM* 2004;**17**:1–25.
35. Ventresque, A., Cazalens, S., Lamarre, P., Valduriez:, P.. Improving interoperability using query interpretation in semantic vector spaces. In: *ESWC*. 2008, p. 539–553.