ISSN 2278-3091

Volume 9 No.2, March - April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse213922020.pdf

https://doi.org/10.30534/ijatcse/2020/213922020

Semantic Similarity Measures for Medical Information Retrieval

Karim Gasmi[1], **Mouna Torjmen**[2]

[1]College of Sciences and Arts of Tabarjal
Jouf University , Saudi Arabia
gasmikarim@yahoo.fr
[2]ReDCAD Laboratory, National School of Engineering
Sfax, Tunisia
mouna.torjmen@redcad.org

ABSTRACT

The conceptual representation is one of the most commonly used approaches as a solution for semantic information retrieval. Most approaches apply NLP tools to map terms from queries and documents to concepts and then compute the relevance scores based on the concept representation. However, the mapping results are not perfect due to the erroneous concepts that are generated out of the document context. To overcome this problem, we propose to add a concept selection step in the indexing. Furthermore, we propose in this paper to study the use of semantic similarity distances in the matching step. Then, we propose a method based on adaptive genetic algorithm to combine two SSD.

Key words : Information retrieval, semantic similarity concept, medical information, UML**.**

1. INTRODUCTION

The traditional indexing methods are based on single words as an entity to represent the information in textual corpus. This representation is based on the co-occurrence of words in a text and does not take into account the semantic relationships that may exist between them. The problem of these models is that the meaning of a word can be expressed in different words, and one word can express different meanings in different contexts. This is due to the richness of the mechanisms of reflection and linguistic expression.

Some studies [3, 33] have highlighted the inadequacy of document representation based on simple words. The authors in [12] showed that only 20% of Internet users use a 100% accurate application depending on their needs. Indeed, this wealth can be a source of ambiguity in a natural language. Although the solutions based on relevance feedback allow partially overcoming the problem of synonymy and helping to improve the recall, yet the problem of polysemy still persists.

In recent years, much work has highlighted the inadequacy of words representation based on simple words[2, 33]. Yet, it has been difficult to go beyond the performance achieved so far. Some works have suggested exploring the semantic textual representation of information. Then appeared several studies attempting to incorporate the semantic information in the Information Retrieval process.

Among such works, we can distinguish two main approaches: the conceptual indexing and the semantic indexing[34]. The latter is based on the meaning of words. These representations are based either on word-based correlations or dictionaries[24] for synonyms extraction. Lately, in [25] and [28] suggested expanding query terms by medical terms. The first 6 study converts automatically user terms to medical terms using UMLS. Then, it adds the 7 medical terms to the initial query. However, there has not been sufficient evaluation for 8 this method. On the hand, authors in [13], employed the MetaMap online tool to recognize 9 medical terms in queries through selecting the highest score mapping of each phrase.

Using some statistics related to the query and the collection, corresponding medical terms were extracted, filtered (by a stop word list) and weighted to be added to the original query finally. Other studies suggested to select expansion terms by taking of irrelevant terms by employing different means like Quantum Mechanic, [54], the document frequency chisquare [35], and the Rank Score Method [21].

On the contrary, conceptual indexing is based on concepts extracted from semantic resources and taxonomies to index documents [52]. As part of information retrieval research, authors believe that the conceptual indexing can be seen as a generalization of the semantic indexing as the concepts convey the meanings of words or terms. The objective of the conceptual approaches is to identify all the terms of the document and could represent them as concepts using an external resource. The concepts extracted are based on the external resource which focuses on the keywords generated from the text[5, 27, 39].

In the medical field, most of the conceptual indexing techniques are based on the UMLS Metathesaurus and use the MetaMap tool 3 to translate a text to UMLS concepts[14]. Several researchers [50, 19]have proved that the main disadvantage when using such tools concerns performance and inaccurate concept as mapping results.

The first drawback is related to the fact that the selection of candidate concepts is based on the concepts of simple words, which causes the problem of over-generation. For example, given the noun phrase ”ocular complications,” we obtain three concepts ”Ocular”, ”Complications” and ”Complications Specific to Antepartum or Postpartum” because they share at least one word.

Recently, a new solution for the over-generation problem was suggested by [45] using an Integer Linear Programming model. This new suggested solution allows the selection of the most relevant concepts by converting the highest number of important terms in the text, which actually was not evaluated. Other works in [55] suggested to use a semantic recourse to map the text of the n top returned into MeSH concepts and compute a similarity score between each of the concepts, and the query so the top-k concepts can be added to the original query. The second drawback is related to the strict comparison between 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,” the ”gyrB” can not be identified as a protein.

To solve the problem of non-significant concept selection, authors in [23] adapt the graph based term weighting method proposed by [6] and apply it to concepts. In most of the occurrence of the concepts in the graph or in the external semantic resource.

To select the most significant concept, we proposed to construct a graph where the document is represented by a graph of concepts and the weights of the arcs are calculated relying on the semantic measure between each pair of concepts. We studied in a previous work [16] some distances between concepts .Results show that the semantic similarity distance have a great impact on the conceptual retrieval process. To better study the similarity degree between concepts and better identify its impact on the information retrieval model, we propose in this paper to study the use of semantic distances and their performance using semantic resources. In fact, we use all semantic distances adopted to the UMLS ontology. Then, we propose combining two different distances with a new combination method based on adaptive version of the genetic algorithm. Our proposed conceptual-based method for medical information retrieval consists of 3 phases that will be detailed later in this paper.

The remainder of this paper was organized as follows: Some related work about the combination of two scores is described in Section 2. Following this, in section 3 we give the details of our approach about our approach for the selection of meaningful concepts. Section 4 was devoted to the evaluation of our approach. After that, we draw our conclusions with some future work in Section 5.

2. SEMANTIC SIMILARITY DISTANCES (SSD)

Several semantic similarity distances have been proposed [37] [4], which use the hierarchical structure in a single ontology or many ontologies. Some SSD are based on hierarchical structure of the semantic resources to compute the similarity between two concepts (hierarchical distances). Other SSD are based on the information content of both concepts to compute similarity (contextual distances).There are also some SSD based on both hierarchical structure and information content (hybrid distances). Finally, we found some SSD based on common features between concepts as common terms in information content or common ancestor in the hierarchical structure. SSD could be so classified into four types:

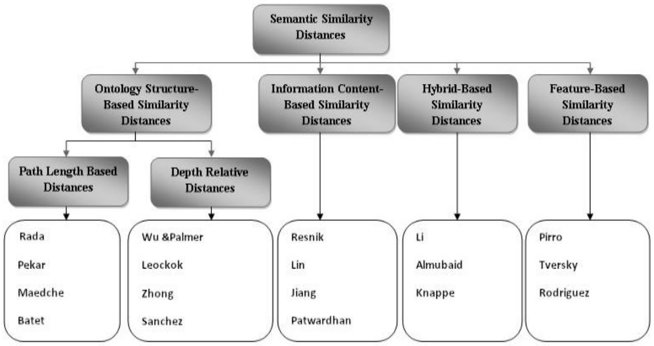
Semantic Similarity Distances
Ontology Structure-Based Similarity Distances
Information Content-Based Similarity Distances
Hybrid-Based Similarity Distances
Feature-Based Similarity Distances
Path Length Based Distances
Depth Relative Distances
Rada
Pekar
Maedche
Batet
Wu &Palmer
Leockok
Zhong
Sanchez
Resnik
Lin
Jiang
Patwardhan
Li
Almubaid
Knappe
Pirro
Tversky
Rodriguez

Figure.1: Semantic Similarity Distances classification

– Hierarchical Distances [38, 49, 32, 4, 46, 56, 53, 29]: they are based on the arcs (edge counting) in the hierarchical structure of the semantic resource. This latter is considered as a tree, and then, the number of arcs between two concepts is computed, or/and the depth of the concepts in the overall structure is computed to quantify similarity.

Hierarchical SSD are direct, natural and simple to implement.

– Contextual Distances [40, 31, 20, 36]: they are based on the notion of information content (i.e. the terms of the definition of the concept) of the paired concepts, or/and the information content of lowest common subsumer (lcs).

– Hybrid Distances [30,1, 22,58] :they are based on the combination of hierarchical and contextual distance factors together.

– Feature-Based Distances[37, 41, 47] : they are based on the the amount of common and non-common knowledge features to estimate similarity between two concepts.

Features could be based on the concept properties (e.g., concept description or ”glosses” in WordNet or ”scope notes” in MeSH) or based on concept relationships to other similar concepts in the taxonomy.

3. MEDICAL INFORMATION RETRIEVAL MODEL

Most concept-based retrieval approaches apply NLP tools to map terms from queries and documents to concepts and then compute the relevance scores based on the concept representation. These approaches usually use similarity distances between concepts to calculate the similarity between queries and documents.

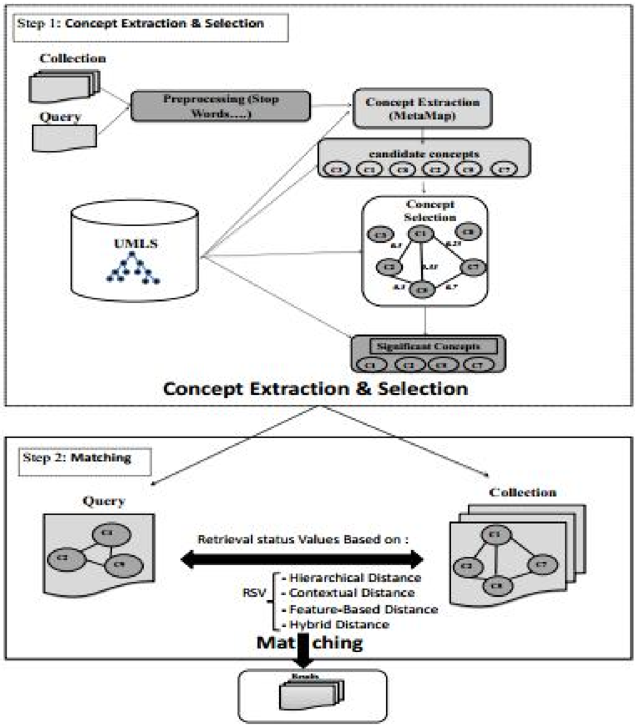

Figure 2: Overview of the proposed medical image retrieval model using

We recall that those conceptual approaches suffer from two majors problems; the non-significant concept selection and the missing relationships as shown in James [11]: 20% of relationships between concepts are erroneous.

In this paper, we propose a conceptual retrieval model which is composed of three parts: first, we extract concepts by using the UMLS ontology and MetaMap tool. Second, we select the most significant concepts through a graph-based method and SSD. Third, we adapt the matching function between queries and document by studying the impact of SSD. The overview of our model is presented in Figure 2.

3.1. Concept Extraction

The aim of this step is to map a text into concepts. For this purpose, we start by pre-processing the collection and removing the stop word in order to keep only significant words.

After the preprocessing step, we extract the concepts from the text by the MetaMap tool, as follows:

1. Parse the text into noun phrases;
2. Extract all the lexical variation from noun phrases found in step 1, with all its variant spellings, abbreviations, acronyms, synonyms, inflectional and derivational variants, and meaningful combinations of these;
3. Look for different concepts (called candidate concepts) from all metathesaurus containing one of the variants found in step 2;
4. By using an evaluation function, MetaMap calculates the mapping strength of candi-dates. This step has been performed for each candidate concept, found in step 3;
5.Combining candidate concepts involved with disjoint parts of the noun phrase, re-computing the matching strength based on the combined concepts, and selecting those having the highest score to form a set of best metathesaurus mappings for the original noun phrase.

The measure used to calculate the mapping strength is based on four components: centrality, variation, coverage and cohesiveness. A normalized value between 0 (the weakest match) and 1 (the strongest match) is computed for each of these components. Candidates mapping concepts are subsequently ordered according to this score.

3.2. Concept Selection

We recall that concept-based approaches suffer from the non-significant concept selection.

To select the most relevant concepts, we have proposed in a previous work [15] a method where the document is represented by a graph of concepts. For all extracted concepts, a semantic similarity distance is calculated between each pair of concepts.

In addition, we propose to improve the quality of the graph by weighting concepts using a voting algorithm as betweeness [7], pageRank [9] and closeness[10] and then use a threshold to eliminate concepts that are not enough semantically similar to other concepts.

Semantic graph building : We note SG = (C, R) a semantic graph composed of a concept set C and a semantic relation set R. A semantic relation connects two concepts having a semantic similarity.

- $C = \{C_1, C_2, ..., C_n\}$ is a set of concepts, and
- $R = \{Rel(C_1, C_2), Rel(C_2, C_3), ..., Rel(C_m - 1, C_m)\}$ is a set of relations.

In some SSD computing tools, if two concepts are non-similar, the distance equals -1.

For this reason, we propose to construct the graph relations as follows:

$$Rel(C_i, C_j) = \begin{cases} dist(C_i, C_j)\; if\; dist(C_i, C_j) \neq 0 \\ 0\; if\; dist(C_i, C_j) \neq -1 \end{cases} \quad (1)$$

With dist is the distance between Ci and Cj calculated by a SSD.

In a previous work [16], we have studied the impact of some SSD on the concept selection step using equation 1, and we found that combining Rada [38] and Patwardhan [36]SSD gives the best results. Consequently, in this paper, the relation weights in the graph are computed using the combination between Rada and Patwardhan SSD.

Weighting graph concepts using a centrality algorithm : At this level, we have already, for each document, a graph composed of concepts related one each other according to a similarity distance. We aim to improve the quality of this graph by eliminating some concepts that are not well semantically related to the most of other concepts. For this reason, we propose to use a centrality algorithm to weight concepts according to their importance in the graph and then use a threshold to decide which concepts eliminate. Although the are several centrality algorithms, we propose to use three algorithms: Closeness [10], Betweenness [7] and PageRank [9].

– 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 [10]

– Betweenness centrality, according to Borgatti [7], 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 [8].

– PageRank [9,57] is based on the idea of ”vote” or ”recommendation”. When a node is connected to another, it is essentially voting for the other node. The more votes are for a node, the greater is the importance of this node. Although PageRank was originally defined for directed graphs, it can also be applied to undirected graphs.

We apply the centrality algorithm on the weighted graph to compute weights for concepts.

Concept elimination by threshold : In order to keep only the most significant concepts, we propose to use the dyadic technique for calculating the eliminatory threshold(2) and then eliminate noisy concepts which have a score less the threshold. The threshold is calculated as follows:

With W eightmin=max is the min and max values of the concept weights in the document graph.

$$X_{i,N} = Weight_{min} + \frac{i * (Weight_{max} - Weight_{min})}{2^N} \qquad (2)$$

We notice here that the values of these thresholds have been determined empirically by examining the regularity of the relationships frequency between concepts extracted from the constructed graph. We represent graphically the number of concepts according to their weights by a Zipfian distribution in figure 5 [26]. This distribution shows a decreasing curve, which can be divided into three zones depending on the concepts significance:

(I) a first zone containing few concepts, describe the trivial, marginal and noise information; (ii) a second zone containing interesting information represented by the concepts constructing the graph; and (iii) a third zone containing significant concepts that acts as a connection between nodes of the graph. From this representation, it’s clear that the second and the third zones are the targets for extracting the most relevant concepts from the texts.

Thus, we use a threshold that allows the elimination of the area of insignificant concepts.

The remaining areas contain the significant concepts representing the documents.

3.3 Matching Based on semantic similarity between concepts

Semantic similarity based matching : We have used the vector space model [44] , [43] as basic retrieval model in our work. Documents and queries are represented by two vectors with N dimensions, where N is the number of the ontology concepts[48].

The value of the system relevance is calculated using the similarity function RSV (Q; d) where Q is a query and Dj is a document as follows:

$$RSV\,(Q, D_j) = \frac{\sum_{i=1}^{N} w_{ij}.w_{q_i}}{\left(\sum_{i=1}^{N} w_{q_i}^2\right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^{N} w_{ij}^2\right)^{\frac{1}{2}}} \qquad (3)$$

$$w_{ij} = cf_{ij} * idf_{ij} \quad and \quad wq_i = cf_{qi} * idf_{qi} \qquad (4)$$

Where:

cfij(respectively cfqi) is the concept frequency in the document Dj (respectively the query q);

idfij is the inverse document frequency of the concept.

In our model, we adapt Equation 15 by adding the similarity score between document concepts and query concepts as follows:

$$RSV\,(Q, D_j) = \frac{\sum_{i=1}^{N} w_{ij}.w_{q_i}.Dist(C_{q_i}, C_{ij})}{\left(\sum_{i=1}^{N} w_{q_i}^2\right)^{\frac{1}{2}} \cdot \left(\sum_{i=1}^{N} w_{ij}^2\right)^{\frac{1}{2}}} \qquad (5)$$

Where : Dist(Cqi ; Cij) is the distance between the concepts Cqi and Cij.

GA-based hierarchical and contextual distances combination : Authors in [11] reveal that a total of 17 022 (24.3%) relations (parent- child) between UMLS concepts can not be justified according to the semantic categories of concepts. Taking into account the concept relation missing, the semantic weight between two concepts is not always exact and doesn’t always reflect the real semantic degree between two concepts. For example, if a child-parent relation between two concepts is missing, a hierarchical distance score will be affected and not correct. Moreover, when using contextual distance, the score can be affected due to the term ambiguity problem of concept definitions. To improve the semantic distance results between concepts, we propose to combine both distances using the genetic algorithm.

The idea id to find the most optimal weight for each distance as follows:

$$Dist_i = \alpha Dist_{hier_i} + \beta Dist_{cont_i} \qquad (6)$$

Disthieri : is the hierarchical distance between two concepts
Distconti : is the contextual distance between two concepts
In genetic algorithms [18], the basic idea is to simulate the population evolution process. We start from a population of N solutions to the problem represented by individuals.

The randomly selected population is called a relative population. The individual adaptation degree to the environment is expressed by the value of the cost function f (x), where x is the solution that the individual represents. It is declared that more an individual is better adapted to its environment, more the cost of the solution is lower.

Figure 3 shows the procedure of the proposed GA weighting combining-annotation method as follows:

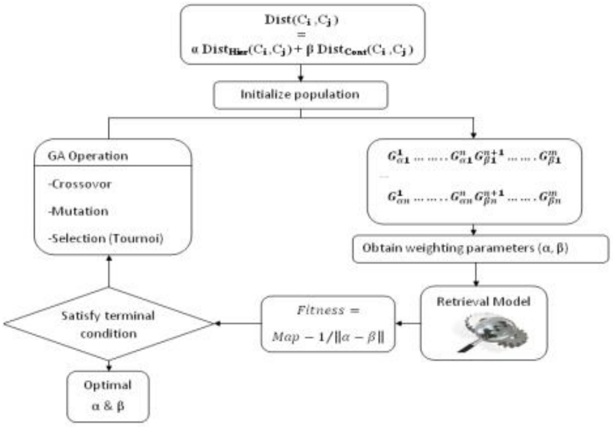
Dist(Ci ,Cj) = α DistHier(Ci ,Cj)+ β DistCont(Ci ,Cj)
Initialize population
GA Operation
-Crossovor
-Mutation
-Selection (Tournoi)
$G^1_{\alpha 1} \dots\dots G^n_{\alpha 1} G^{n+1}_{\beta 1} \dots\dots G^m_{\beta 1}$
$G^1_{\alpha n} \dots\dots G^n_{\alpha n} G^{n+1}_{\beta n} \dots\dots G^m_{\beta n}$
Obtain weighting parameters (α, β)
Retrieval Model
Fitness = Map − 1/‖α − β‖
Satisfy terminal condition
Optimal α & β

Figure 3: Genetic algorithm process

Most Genetic Algorithm operations require a fitness function to calculate the individual adaptation. In the proposed approach, the fitness function depends on tow factors: (1) the results returned by the retrieval model according to mean average precision (MAP):

MAP is considered as a user who selected in each iteration the relevant documents in order to compute the Fitness for each chromosome; and (2) the absolute value of subtraction between the two weights. The aim of the second factor is to minimize the cost of any annotation influence compared to the other on the Fitness function (Equation 7).

$$Fitness = MAP - \frac{1}{\|\alpha - \beta\|} \qquad (7)$$

The genetic algorithm process is as follows:

The genetic algorithm process is as follows:

The genetic algorithm process is as follows:

Step 1 : Encodes the chromosomes and the parameters representing the weighting indexing method as a binary string. Figure 4 shows an example of a chromosome creation from two weights.

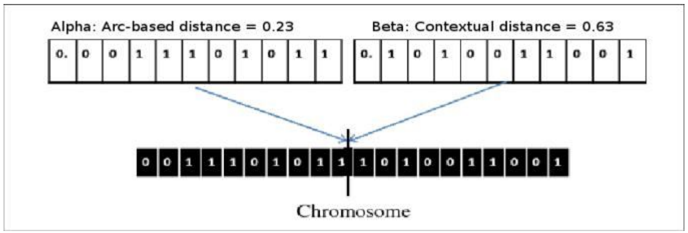

Figure.4: Coding of a Chromosome

– Step2 : Initializes the population and produces the initial population of chromosomes arbitrarily.
– Step3 : The fitness for each chromosome must be computed, this is related to the calculated results obtained by the fitness equation.
– Step 5 : The main feature is that the fitness value decreasing during the last M generation or N is reached as the maximum generation number.
– Step 6 : The iteration process stops only when the two criteria are achieved. Otherwise you have to move to step 5.
– Step 7 : To generate a offspring generation genetic operations (crossover, mutation and tournoi reproduction) should be performed.

4. EXPERIMENTS AND RESULTS

The aim of this section is to evaluate our proposed model and more precisely is to study empirically the impact of SSD on the retrieval effectiveness.

4.1. Data sets and evaluation metrics

To evaluate our approach, we used the 2009 ImageClef6 test collection composed of 74,902 medical images with their associated. This collection contains images and captions from two Radiological Society of North America (RSNA)[7] journals. The Medical Image Clef 2009 contains 25 queries selected by medical experts. For each query is assigned a list of relevant documents assessed by human assessors involved in the CLEFevaluation campaign.

For measuring the IR effectiveness, we used P5, P10 representing respectively the mean precision values at the top 5, 10 returned documents and the MAP measure representing the Mean Average Precision calculated over all topics.

4.2. Results of the retrieval step

The purpose of this section is to study the semantic similarity degree effectiveness for medical information retrieval. Then, we carried out two series of experiments: the first is based on the classical conceptual retrieval using the well known weighting scheme BM25, as the baseline, denoted MetaMap. The second concerns our retrieval approach and consists of four methods:

– the first concerns the hierarchical distance without taking into account the contextual relationship between concepts, denoted Arc-based, we have exploited 8 different distances

– The second concerns the contextual distance, denoted IC-based.

– The third concerns the hybrid distance, denoted Hybrid-based.

– The fourth concerns the feature-based distance, denoted feature-based

We computed the paired-sample wilcoxon[51] in order to test the significance of the results. We assume that the difference between the results of two retrieval approaches is significant if p <0.1 (noted *) and very significant if p <0.05 (noted **). Table 2 presents the MAP, P5 and P10 results over the baseline and the proposed retrieval model with different SSD.

Table 1: Impact of SSD using our retrieval model

Type		Distance	P5	P10	Map
MetaMap		–	0.3920	0.3440	0.2163
Arc-Based	PATH	Rada	0.424**	0.3520*	0.1993
		Pekar	0.3120	0.2680	0.1512
		Maedche	0.2480	0.1960	0.1342
		Batet	0.1280	0.1480	0.0953
	DEPTH	WP	0.2250	0.2400	0.1329
		leaChod	0.2800	0.2320	0.1406
		Zhong	0.4240**	0.4040**	0.2244*
		Sanchez	0.2480	0.2240	0.1364
IC-Based		Resnik	0.3238	0.2667	0.1435
		Lin	0.3200	0.2875	0.1700
		JiaCon	0.3520	0.3000	0.1780
Hybrid-Based		Li	0.1320	0.1452	0.1120
		Almu	0.1680	0.1680	0.1075
Feature-Based		Pirro	0.3520	0.2920	0.1770

As seen, the paired-sample wilcoxon-test shows that the Zhong run is statistically significant compared to the baseline. According to Table 1, we note that the Map value of hierarchical distances is the best value compared to other map values obtained by other distances. As for the hybrid distances, values of P5 and P10 are the lowest values. Theoretically these results can be justified by the following reasons: Firstly, to use the Rada distance, a semantic resource must be coherent and rich of concepts. The major drawback of this measure is that it is not valid for all reporting relationships. Furthermore, it does not support in the multiple hierarchies. For example, the UMLS-concepts are not connected only by purely hierarchical relationships like ”is a” relation, establishing a hierarchy of types. The UMLS-network also has 5 major categories of non-hierarchical (or associative) relationships, which constitute the remaining 53 relationship types. We cite such relationship examples: ”physically related to”, ”spatially related to”, ”temporally related to”, ”functionally related to” and ”conceptually related to”. Despite both semantic distances proposed by Rada [53] and Wu & Palmer [53] are easy to implement and have the advantage of implementing the semantic relationships between the concepts. However, these distances do not take into account the information content of the concept which devalues the contribution of the concept in terms of information.

Then, the semantic measure proposed by Resnik, depending on the corpus, actually returns the LCS or the root of both concepts. Note that two pairs of different concepts can have the same LCS.

A disadvantage of Resnik measure is that the probability of a concept is derived from the words found in the corpus without checking that each occurrence refers to the concept, the problem occurred when the term is ambiguous and designates different concepts.

Another disadvantage is that Resnik does not use the information content of each concept separately.

Also, using the Lin measure requires that the concepts are well organized in the hierarchy and they are interconnected. Then, we can’t use this measure to calculate the similarity between two concepts from different categories. In addition, Lin measure does not take into account the lexical fields.

4.3. Results of the selection step

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 MetaMap.
– The second one concerns the concept selection method proposed by Humphrey in [19], denoted WSD MM.

– The third one concerns our concept selection approach and consists of four scenarios
For each centrality algorithm evaluation, we have chosen thresholds X (i, N), each was determined by the equation 4. In fact, each threshold was belongs to a part of the Zip graph or on other words according to the concepts distribution.

Table 1 presents the results using PageRank, Betweneess and Closeness algorithms according to P5, P10 and MAP . By comparing the results against those obtained by traditional indexing, PageRank gives a significant improvement. This improvement is observable especially with the threshold X (1,4). According to the Table 1, we observe that Map value changes by variation of the threshold. This variation is presented in figure 5 using PageRank algorithm.

By using a threshold in the interval] 0,X (1,4)], our model based on the PageRank algorithm generates better results compared to the MetaMap ones. The MAP was increased by augmenting the threshold until having a maximum value of the MAP equals to 0,26 with a threshold of X (1,4). Those results can be explained by the fact of eliminating the weak concepts when we increase the threshold value. Indeed, the result will be negatively or positively influenced according to the threshold rate. Thus, we found a negative effect of non significant concepts and a positive effect of our concept-selection method results compared to the ones obtained when using classical indexing. However, for a threshold in the interval [X (1,4), X(n,n)], the MAP has been decreased when increasing the threshold value. Thus, the significant concepts in this part of the curve will be eliminated and consequently the result will be negatively impacted. For the remaining of our experiments we chose X(1,4) as the best threshold.

Table 2: Comparison between centrality algorithms according to different thresholds

Category	P5	P10	MAP	Category	P5	P10	MAP	Category	P5	P10	MAP
MetaMap	0.3920	0.3440	0.2163	MetaMap	0.3920	0.3440	0.2163	MetaMap	0.3920	0.3440	0.2163
WSDMP	0.4400	0.428	0.2160	WSDMP	0.4400	0.4280	0.2160	WSDMP	0.4400	0.4280	0.2160
PageRank				Betweneess				Closeness			
X(7,3)	0.19	0.16	0.04	X(1,2)	0.260	0.195	0.0690	X(7,3)	0.1889	0.2111	0.0775
X(5,3)	0.2910	0.2580	0.1700	X(1,4)	0.3580	0.3250	0.1550	X(1,2)	0.3130	0.3217	0.1651
X(1,2)	0.4250	0.3750	0.1850	X(2,4)	0.4160	0.4250	0.2290	X(1,3)	0.4333	0.4333	0.2392
X(1,4)	0.4660	0.4580	0.2600	X(1,6)	0.3910	0.4370	0.2300	X(1,4)	0.4583	0.4458	0.2529

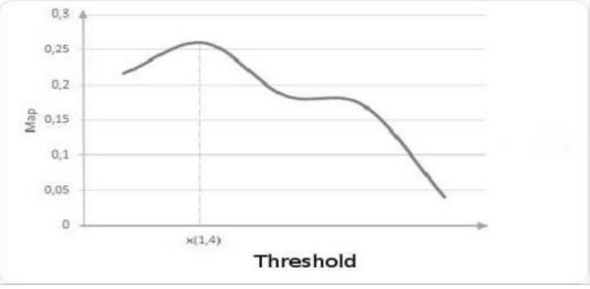

Figure 5: MAP variation relative to the cutoff threshold

4.4. Results of SSD's combination

In this section, we present results of combining hierarchical and contextual distances.
We chose distances according to the best results of Table 2. More precisely, we combine Zhong and JiaCon distances.
The combination is done after a normalization step according to Equation 6. Literature hybrid distances use several information sources to compute the similarity degree between concepts as LCS and path length, but does not give good results.

Table 3 contains the genetic algorithm parameters.

According to results, we observe an improvement of only 2.3% in the MAP value using the combination.
This improvement is achieved with a combination between the Zhong and the JiaCon distances. We notice that a score combination between two types of distances gives better results than literature hybrid distance as Li or Almu distances. Although the improvement is not very significant, we believe that the combination avoid the disadvantages of each type of distance. For example, using the contextual distances

Table 3: GA parameters

GA property	Value/Method
Size of generation	100
Initial population size	30
Selection method	Tournoi
Number of crossover points	1
Crossover Probability	0.9
Mutation method	Uniform mutation
Mutation Probability	0.05

allows to avoid the problem of missing relations at the hierarchical distance,because it is based on the concept information content instead of the path length.

4.5. Further Results and Discussion

Concept-based approaches allow to minimize the vocabulary problem as the use of concept faces and exceeds (1)keyword variation (gender, plural, verbal conjugation, etc.), (2) semantic association (specialization, generalization), and (3) semantic information (”to be” or ”not to be”). However, these methods generally could not outperform text-based approaches.

In order to takes advantages from concept-based approaches, combining these methods with textual approaches could be a good solution and it showed its interest. So, in our work, we propose to combine our method with textual one using the genetic algorithm (the run is denoted AG-TexCon).

In this section, we propose also to compare our method with official submissions of Medical Image Clef 2009.

Table 4 presents the results. Thanks to these experiments, our method should have been ranked second in the official Medical Image Clef 2009 (best MAP = 0.4300). Experimentations showed the efficiency of our method. According to the official runs, best result are obtained by the IRIS team method which uses a relevance feedback extension of queries with the n first returned documents (best results are obtained with n = 100)

Table 4: Comparison of our method with official runs of IMAGE CLEF 2009

	Rang	MAP	P5	P10
LIRIS maxMPTT extMPTT	1	0.4300	0.7000	0.6600
AG-TexCon		0.3870	0.6500	0.6300
sinai CTM t	2	0.3800	0.6500	0.6200
york.In expB2c1.0	3	0.3700	0.6100	0.6000
ISSR text 1	4	0.3500	0.5800	0.5600
ceb-essie2-automatic	5	0.3500	0.6500	0.6200
deu run1 pivoted	6	0.3400	0.5800	0.5200
clef2009	7	0.3400	0.6700	0.6000
BiTeM EN	8	0.3200	0.5200	0.5000
UNTtextrf1	9	0.2600	0.5300	0.4400
OHSU SR1	10	0.1800	0.5900	0.5400
MirEN	11	0.1700	0.6200	0.5500
uwmTextOnly	12	0.1300	0.4400	0.4000
Alicante-Run3	13	0.1300	0.3400	0.3600

5.CONCLUSION

In this work, we have proposed and evaluated a sense-based approach for retrieving medical information. Our main contribution consists of improving our conceptual retrieval model by adapting some semantic distances in the UMLS Meta-thesaurus. To improve our graph-based method for concept selection, we propose to enhance the constructed graph by centrality algorithm and dyadic threshold method. Then, we propose an automated combination method based on adaptive version of the genetic algorithm to combine two SSD. Our results on the test collection MedicalImageCLEF 2009 showed an improvement. As a future perspective, we plan to solve the problem of the relationship weight and suggest a semantic similarity measure. Specifically, we will improve our approach in a future work by the use of other relation categories as ”physically related to”, ”spatially related to”, ”temporally related to”, ”functionally related to” and ”conceptually related to”. Secondly, we aim at extending our proposed model by other termonology other than UMLS.

Finally, we aim to propose a new approach to automatically learn how to classify medical queries into a set of retrieval models.

.REFERENCES

1. Al-Mubaid, H., Nguyen, H.A.: A cluster-based approach for semantic similarity in the biomedical domain. In: In 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 2713–2717 (2006) https://doi.org/10.1109/IEMBS.2006.259235
2. Aronson, A.R.: Effective mapping of biomedical text to the umls metathesaurus: the metamap program. Annual Symposium AMIA pp. 17–21 (2001)
3. Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The nlm indexing initiative’s medical text indexer. In: World Congress on Medical Informatics Demner-Fushman and Lin Answering Clinical Questions. pp. 268–272 (2004)
4. Batet, M., Sanchez,´ D., Valls, A.: An ontology-based measure to comput semantic similarity in biomedicine. Journal of biomedical informatics 44(1), 118–125 (2011) https://doi.org/10.1016/j.jbi.2010.09.002
5. Baziz, M.: Indexation conceptuelle guidee par ontologie pour la recherche d’information. In: These` de doctorat, Universite Paul Sabatier, Toulouse, France (2005)
6. Blanco, R., Lioma, C.: Graph-based term weighting for information retrieval. Inf. Retr. 15(1), 54–92 (2012)
7. Borgatti, S.P.: Centrality and network flow. In: Social Networks. pp. 55–71 (2005) https://doi.org/10.1016/j.socnet.2004.11.008
8. Brandes, U.: A faster algorithm for betweenness centrality. In: Journal of Mathematical Soci-ology. vol. 25, pp. 163–177 (2001)
9. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Comput. Netw. ISDN Syst. vol. 30, pp. 107–117. Amsterdam, The Netherlands, The Netherlands (1998)
10. 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. pp. 1–13 (2011)
11. Cimino, J.J., Min, H., Perl, Y.: Consistency across the hierarchies of the umls semantic network and metathesaurus. Journal of Biomedical Informatics 36(6), 450–461 (2003)
12. Crestani, F.: Exploiting the similarity of non-matching terms at retrieval time. Journal of Infor-mation Retrieval 2, 25–45 (1999)

13. Diao, L., Yan, H., Li, F., Song, S., Lei, G., Wang, F.: The research of query expansion based on medical terms reweighting in medical information retrieval. EURASIP Journal on Wireless Communications and Networking 2018 (12 2018)
https://doi.org/10.1186/s13638-018-1124-3

14. Fiszman, M., Rindflesch, T.C., Kilicoglu, H.: Abstraction summarization for managing thebiomedical research literature. In: The HLT/NAACL 2004 workshop on computational lexical semantics. pp. 76–83 (2004)

15. Gasmi, K., Khemakhem, M.T., Jemaa, M.B.: Word indexing versus conceptual indexing in medical image retrieval. In: CLEF (Online Working Notes/Labs/Workshop) (2012)

16. Gasmi, K., Khemakhem, M.T., Jemaa, M.B.: A conceptual model for word sens disambiguationin medical image retrieval. In: AIRS. pp. 296–307 (2013)

17. Gasmi, K., Khemakhem, M.T., Tamine, L., Jemaa, M.B.: Graph-based methods for significant concept selection pp. 488–497 (2015)
https://doi.org/10.1016/j.procs.2015.08.170

18. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning (1989)

19. Hliaoutakis, A., Zervanou, K., Petrakis, E.G.: The fAMTExg approach in the medical document indexing and retrieval application. Data and Knowledge Engineering 68(3), 380 – 392 (2009)

20. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy(1997)
1.

21. Jothilakshmi, R., Moorthi, S.: Combining multiple term selection methods for automatic query expansion in pseudo relevance feedback using rank score method. Asian Journal of Research in Social Sciences and Humanities 7, 910 (2017)
2.

22. Knappe, R., Bulskov, H., Andreasen, T.: Similarity measures for content-based querying. Proceedings of the 10th International Fuzzy Systems Association World Congress (IFSA'03),(2003)
3.

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. pp. 80–87 (2012)
https://doi.org/10.1145/2407085.2407096
4.

24. 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. p. 245–252 (2000)

25. Kumar, K., Deepa, K.: Medical query expansion using umls. Indian Journal of Science and Technology 9 (04 2016)
5.

26. Lafouge, T., Boukacem, B.: Applications des lois informatiques en sciences de l'information : dualite,́ champ informatique d'usage et de production. ISDM 17, 1–25 (2004)
6.

27. Latifur Khan, Dennis McLeod, E.H.H.: Retrieval effectiveness of an ontology-based model for information selection. In: VLDB journal. pp. 71–85 (2004)
https://doi.org/10.1007/s00778-003-0105-1
7.

28. 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. 2007 IEEE International Conference on Research, Innovation and Vision for the Future pp. 242–246 (2007)
8.

29. Leacock, C., Chodorow, M.: Combining local context and wordnet similarity for word sense identification. In: MIT Press. pp. 265–283 (1998)
9.

30. Li, Y., Bandar, Z.A., McLean, D.: An approach for measuring semantic similarity between words using multiple information sources. IEEE Trans. on Knowl. and Data Eng. 15(4), 871–882 (Jul 2003)
https://doi.org/10.1109/TKDE.2003.1209005
10.

31. Lin, D.: An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning. pp. 296–304 (1998)
11.

32. Maedche, A., Staab, S.: Comparing ontologies-similarity measures and a comparison study AIFB (2001)
12.

33. Marc Weeber, James G. Mork, A.R.A.: Developping a test collection for biomedical word sense disambiguation. Annual Symposium. AMIA Symposium pp. 746–750 (2001)
13.

34. Mihalcea, R., Moldovan, D.: Semantic indexing using wordnet senses. In: In Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval. pp. 35–45 (2000)
14.

35. Nabeel Asim, M., Wasim, M., Usman Ghani Khan, M., Mahmood, W.: Improved biomedical term selection in pseudo relevance feedback. Database 2018 (07 2018)
https://doi.org/10.1093/database/bay056

36. Patwardhan, S., Pedersen, T.: Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together. vol. 1501, pp. 1–8. Trento, Italy (April 2006)
15.

37. Pirro,́ G., Euzenat, J.: A feature and information theoretic framework for semantic similarity and relatedness pp. 615–630 (2010)
16.

38. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transaction Systems Man and Cybernetics. pp. 17–30 (1989)
17.

39. Rada Mihalcea, Paul Tarau, E.F.: Pagerank on semantic networks with application to word sense disambiguation. In: In International Conference on Computational Linguistics (COL-
18.
ING). pp. 1126–1132 (2004)
19.

40. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1. pp.448–453 (1995)
20.

41. Rodr´ıguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. IEEE Trans. Knowl. Data Eng. 15(2), 442–456 (2003)
21. https://doi.org/10.1109/TKDE.2003.1185844

42. Sajgalik, M., Barla, M., Bielikova, M.: From ambiguous words to key-concept extraction. In International Workshop on Database and Expert Systems Applications,Prague, Czech
22. Republic. pp. 63–67 (2013)

43. Salton, G. (ed.): The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice
23. Hall, Englewood, Cliffs, New Jersey (1971)

44. Salton, G., McGill, M.J.: Introduction to modern information retrieval. McGraw-Hill computer science series,
24. McGraw-Hill, New York (1983)

45. Samassi Adama, Brou Konan Marcellin, G.B.T., Kimou, P.: Efficient reduction of overgeneration errors for automatic controlled indexing with an application to the biomedical domain.International Journal of Advanced Computer Science
25. and Applications

46. Sanchez,´ D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications 39(9),
26. 7718–7728 (2012)

47. Tversky, A.: Features of similarity. Psychological Review 84(4), 327–352 (1977)
27. https://doi.org/10.1037/0033-295X.84.4.327

48. Ventresque, A., Cazalens, S., Lamarre, P., Valduriez:, P.: Improving interoperability using query interpretation in
28. semantic vector spaces. In: ESWC. pp. 539–553 (2008)

49. Viktor, P., Steffen, S.: Taxonomy learning: Factoring the structure of a taxonomy into a semantic classification decision
29. pp. 1–7 (2002)

50. 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
30. for Computational Linguistics, ACL. pp. 603–612 (2014)

51. Wilcoxon, F.: Individual Comparisons by Ranking
31. Methods. Biometrics Bulletin pp. 80–83 (1945)

52. Woods, W.A.: Conceptual indexing: A better way to organize knowledge. Tech. Rep. SMLITR-97-61, Sun
32. Microsystems, Inc. (1997)

53. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics. pp. 133–138
33. (1994)

54. Zhang, P., Song, D., Zhao, X., Hou, Y.: Investigating query-drift problem from a novel perspective of photon polarization. Proceedings of the Third International Conference on Advances in Information Retrieval Theory pp.
34. 332–336 (2011)

55. Zhou, W., Yu, C., Smalheiser, N., Torvik, V., Hong, J.: Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.pp.
35. 655–662. SIGIR '07 (2007)

56. Zhu, H., Zhong, J., Li, J., Yu, Y.: An approach for semantic search by matching rdf graphs. In:FLAIRS Conference. pp. 450–454 (2002)

57. L. Lakshmi , K. Pushpa Rani , M. Purushotham Reddy, A Comparative Study of Navigation Techniques and Information Retrieval Algorithms for Web Mining. International Journal of Advanced Trends in Computer Science and Engineering, Vol 8, 2019.

58. Omar, Nurul. Hyb-Tvx: A Hybrid Semantic Similarity Feature-Based Measurement for Multiple Ontologies. International Journal of Advanced Trends in Computer Science and Engineering. 8. 176-180. 10.30534/ijatcse/2019/3581.32019.