Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/100114

A Novel Conceptual Weighting Model for Semantic Information Retrieval

Sameh Neji [1,*], *Leila Jemni Ben Ayed* [2], *Tarek Chenaina* [2] *and Abdullah M. Shoeb* [3]

[1]Faculty of Economics and Management, Sfax University, Tunisia
[2]National School of Computer Science, University of Manouba, Tunisia
[3]Faculty of Computers and Information, Fayoum University, Egypt

Received: 4 Dec. 2020, Revised: 17 Dec. 2020, Accepted: 19 Dec. 2020.
Published online: 1 Jan 2021.

Abstract: Nowadays, the traditional information retrieval (IR) is inadequate for the user who requires precise results. Hence, the importance of the semantic IR arises. It is very important to move from the level of ambiguous terms to that of well-specified concepts in the indexing phase to enrich the search process. To handle the problems of semantic ambiguity of indexed terms as well as the uncertainty and imprecision inherent in the information retrieval process, a semantic indexing approach was proposed for a better document representation. It is based on indexing the associated synsets with document terms that are identified by mapping on the WordNet ontology. These synsets are defined following a term disambiguation process (WSD). The key to the proposed system is a weighting model which calculates the importance of each index item considering many factors that improve the performance of the information retrieval system. Proposed conceptual weight is based on local and global integrality, degree of re-homogenization, and degree of specificity of the concept. A corrector parameter to reduce the impact of errors in WSD process is included. The experimental evaluation of the introduced semantic IR model shows very satisfactory results compared to well-cited benchmarks.

Keywords: Degree of re-homogenization, Degree of specificity, Importance of the concept, Information retrieval, Semantic indexing, Semantic term weighting.

1 Introduction

Most information retrieval systems (IRS) are based on a keyword-based indexing system, in which document items and query items are represented by a set of weighted keywords. Indexing by keywords is imprecise [1] and does not consider the semantic relationship between the words. This imprecision occurs because of the semantic ambiguity of natural language words. Therefore, it is impossible to find relevant items even if they contain words that are synonymous with the query terms. In addition, irrelevant documents containing words that are lexically identical to the query terms will be retrieved without considering the meaning of these words.

Classic indexing also do not allow documents to be retrieved whatever they are semantically, and not lexically, close to the user specific needs. When the user wishes to search for documents explaining a subject, he/she will also be interested in documents describing subjects which are semantically related to it. A traditional indexing neither allows to define the meanings of the terms nor to find semantic links between them. To overcome these limitations, many works have focused on considering the semantic aspects of indexed terms. This process is denoted by semantic or conceptual indexing. These works move from the class of simple word processing to that of concept processing.

For semantic indexing, the index elements can be in the form of concepts identified from a semantic resource. The concept is referenced by one or more synonymous terms (synsets) in the terminology resources [2]. Semantic indexing is based on the use of semantic resources for better representation of information in the document. Depending on type of the used resource, we can classify the work in two classes: specialized [3] or general resources [4] and [5]. A state of the art on different semantic indexing and term weighting approaches is presented in the next section.

The present paper aims to improve the search

*Corresponding author e-mail: neji.sameh@yahoo.fr

effectiveness. To achieve the study objective, a semantic information retrieval approach (SIRA) based on conceptual weighting using WordNet ontology is proposed. What makes it different from other works is consideration of non-empty terms which have no entry in the ontology in addition to the concepts that are covered there and proposal of a conceptual weight to assess the importance of the role of the indexed concepts. The proposed formula calculates the importance of the concept in the document according to four main factors: the degree of integrality, the degree of re-homogenization, the degree of specificity and a corrective parameter to reduce the impact of uncertainty in the process of WSD. A comparative study on different term weighting schemes is conducted.

Section 2 covers the related works in the context of semantic indexing and weighting of index terms. The proposed semantic model is introduced in sections 3 and 4. Section 5 shows the experimental work that validates the approach. The last section involves conclusion and further research.

2 Related Works

The IRS tends to retrieve all relevant documents for a user query. An indexing phase is first performed to represent documents and queries. This phase requires measuring the importance of the concepts associated with the terms, and so the birth of the notion of "concept weighting". This weighting directly affects the quality of the obtained documents and so the overall accuracy. Semantic indexing in IR deals with the problem of ambiguity in natural language words. To improve search results, documents and queries are represented by word meanings, which help resolving ambiguity. To find the correct sense of the words to be indexed, word sense disambiguation techniques are used.

The identification of concepts is done by techniques of projection of the text on the semantic resources used [3] and/or by a process of disambiguation of the meaning of words (WSD) [6]. During indexing phase, concepts will be weighted to reflect their importance in the document (or query). These weights are integrated, in the search phase, in the document classification formula (document / query correspondence) which calculates the relevance score of each document compared to the query.

In the medical domain, an approach [7] is proposed based on the contextual and structural similarity which was expressed by the relationships between the concepts of thesaurus associated with the terms of the documents and the expressions represented by the successive terms of these words. In [3], the authors used the specialized MeSH thesaurus to perform semantic indexing of electronic patient records. Semantic indexing is carried out in two stages: semantic annotation (extraction and determination of the meaning of concepts) and generation of semantic index.

The author in [8] addressed the issue of semantic indexing as an extended vector model where the principal components can be analyzed using latent semantic indexing. In the indexing method [9], the synsets represent the different possible meanings of the word. These concepts are related to the WordNet ontology. To assign the correct meaning to the target word, each synset of that word is classified according to the number of overlaps between the local context (the sentence in which the word appears) and its neighborhood.

In [10], the authors indexed the words by defining the correct sense of each one in the local context of these words (ordered list of words starting with the useful word closest to the left or right neighborhood up to the word target). Based on a similar principle, the approaches proposed by Baziz et al. [11, 12] allowed to project the contents of the documents and the queries on WordNet to extract terms associated with concepts. The representations of documents and queries are done through these concepts and the relationships between them.

The authors of [13] proposed a conceptual indexing approach. First, indexed items are identified by following the steps of the classic index. Then, extraction of the resource entries containing the indexed words is performed by following the mapping technique on WordNet. Frequency of the term in a document and its semantic distance from other more common concepts in the document are used to disambiguate indexed terms [14]. The importance of a concept in a document is assessed by measuring its semantic similarity with other concepts in that document. This measure is combined with the frequency of occurrences of a concept in a document [15]. A similar approach has been proposed for indexing multimedia XML documents [16].

Another approach [17] was based on the notion of the centrality of a concept which is defined through the number of semantic relations of WordNet that it shares with the other concepts of the document. The semantic indexing approach proposed in [18] allows to discover the various associations between concepts and improves IR in massive text. This approach allows building a network of ontologies based on unsupervised learning. Simple index terms are extracted by classic indexing and compound words are identified by statistical methods based on the frequency of words that appear mutually in the text of the document (or query) [19].

The author of [4] defined a model of semantic representation of documents and queries through a set of concepts which are collected between them in the form of a semantic network using WordNet. This approach shows that the results improved when the index representation is performed by concepts identified from ontology (semantic index) combined with the keywords of documents (classic index). The identification of concepts is done by a text mapping technique on WordNet. The disambiguation of the

sense of words is based on a combination of WordNet and its extension WordNetDomains [20].

The weighting of terms is an important issue in the field of IR. It defines degree of representativeness of index descriptors. Many approaches to weight concepts in IRS have been proposed. In [21], the authors proposed a conceptual approach based on specialized ontologies.

Several works define efficient weighting models, such as TF-IDF [22], Okapi-BM25 [23] and the rotated normalization [24]. Although they are different models, they are essentially based on the same basic principle: The obvious importance of the term is quantized mainly according to the frequency of its occurrence and the frequency of term throughout the collection. It is an efficient term weighting system in information retrieval [25] and many text mining tasks [26]. This method based on the number of occurrences of words does not allow the expression of the potential importance of the concepts related to the semantic contribution of their concepts to the content of the document.

A semantic indexing approach has been introduced [27]. It utilizes the logical structures. The author of [11] has proposed a weighting scheme called CF-IDF, which presents an extension of the weight TFIDF to increase the weight of compound terms. In [28], the authors presented a supervised model to estimate the weights depending on training dataset. The same approach was revisited again by many works [29, 30, 31, 32, 33, 34, 35].

The weight of a term is enriched in [17, 36] with two factors: centrality and specificity. The centrality of a concept reflects its relationship to concepts in the same document and the specificity reflects the specialty of that concept in the field of research addressed by the document in which it appears. In [37], the calculation of term weight is induced by its context and the semantic similarity between the concept associated with this term and the other concepts found in the context of the same tag.

3 Overview of SIRA

SIRA consists of three main steps:

(i) Definition and disambiguation of concepts: Extraction of concepts describing the document and the query and transform them into meaningful concepts through a contextual process of disambiguation.

(ii) Construction of the index: Construction of a semantic index with concept weights. This step solves the weak document representation in the classical index.

(iii) Assess of document-query relatedness to validate the proposed weighting model compared to other weighting approaches.

Figure (1) shows the algorithm of SIRA.

3.1 WSD Module

In addition to semantic information (meaning of terms), this module also adds a descriptive part to each concept. Each document is processed as input to this module and its terms transforms to the following structure:

Term /POS/Sense/Spec/Tau

Where: Term is the word to be processed, POS is the term part of speech, *Sense* is the meaning of the concept associated with this term in WordNet, *Spec* is evaluated by the sense "depth" in the ontology (WordNet) induced by the "is-a" relation, and *Tau* is the synset cardinality per document.

If no meaning assigned by the WSD algorithm exists, the approach assigns the common sense to the concept (# 1).

3.2 Indexing Module

This is the indexing phase of documents after their processing by the WSD module. Retrieving semantic information is also necessary to retrieve the relevant information. The generated field POS, sense, specificity, tautology fields are added to the index. This step is crowned with success by finding the weight of each identified sense $w(s_t)$ associated with term t.

3.3 Search Module

This phase is dedicated to retrieve the documents that are relevant to a query. The approach uses a combined indexing based on keywords and synsets, so it retrieves documents containing single keywords (in case of the keywords that do not belong to WordNet), keywords with a sense assigned to each, or synonyms of the keywords. The final scoring of a document is:

$$Score(d) = \sum_{\forall\, t \in d} w(s_t, d)$$

4 Indexing Process

Indexing of documents based only on concepts may be inadequate as disambiguation techniques are not completely reliable and may result in loss of information. The approach also indexes terms that are unrelated to ontology. This is valuable if new documents have been added to the collection and the system has not yet linked their contents to the ontology, but the system can still retrieve them.

The indexing phase describes the process of representing the content of the documents in the collection as well as the query through representative elements that serve to facilitate processing of information during the search. These elements are called descriptors or indexing terms and come from documents or external semantic resources. The approach used controlled language to index all documents the same way. The problems of polysemy,

```
Algorithm: Proposed Weighting Model
Input : Collection C, List of topics
Output : Score of a document d, Score(d) ∀ d ∈ C
Begin algorithm
Extract Collection statistics, |C|, POS(N,C)
Set α value
Set MinTopicID, MaxTopicID
   For QueryID from MinTopicID to MaxTopicID do
     Read(Query_Terms[i])  ∀i, 1 ≤ i ≤ Topic(QueryID).length
     Ex_Query[] ← Extending(Query_Terms[])
     Evaluate Spec(Ex_Query[j]), ∀j, 1 ≤ j ≤ Ex_QueryID.length
     Evaluate freq(Ex_Query[j],C), ∀j, 1 ≤ j ≤ Ex_QueryID.length
     Split C into n partitions    //for concurrency
     For partID from 1 to n
         Fork a thread, partID_Scoring, to handle partID
     END For                  //End of all partition documents score
     Join all threads
   END For                    //End of all topics
End Algorithm

Thread partID_Scoring(partID) {
    For all document d ∈ partID
        For all concept St associated to term t∈d, St in Ex_Query
            Evaluate freq(St, d)
            Evaluate tautology(St, d)
            Evaluate freq(t, d)
            Evaluate w(St)
            SIRA_Score(d,QueryID) += w(St)
        END For      //End of weighting all senses in d
    END For          //End of document score
}                    //End of all partition documents score
```

Figure (1). Proposed weighting model algorithm

synonymy, homonymy … etc. are avoided using this language because it refers to an external semantic resource. WordNet is made up of a structured list of concepts that are linked by semantic relationships. Thus, it has been used during the indexing process to ensure consistency in the representations of documents and queries. The first indexing step is to define which elements will be used to represent documents and queries to build the index space. Disambiguated concepts (synsets) have been chosen as descriptors.

This phase consists of three sub steps: The first sub step is to extract the concepts of the ontology that are attached to the documents, the second one is to disambiguate these concepts, and the last one is to weight them according to the mentioned factors.

4.1 Concepts Identification

In this part, it is required to define from a text a list of concepts belonging to WordNet that represent its content. For a term t_i of a document d, we have attached a unique WordNet concept c_i of ontology O by projection of t_i on O.

In general, to clarify the content meaning of a document, the first step is to extract its different concepts. This step defines the concepts (c_1, …, c_n) of the ontology O associated with the terms (t_1, …, t_n) identified in the document.

4.2 Disambiguation of Terms

Polysemy and synonymy are two fundamental problems that affect the representation of text and the classification of documents. Removing ambiguity of polysemous words, synonyms gives better document scoring [38]. WSD is a process of replacing the original terms of a document with the most appropriate meaning dictated by the context of the document.

There are many algorithms to perform this step. Extended Lesk has been used in the current work as a WSD algorithm. It is based on the extended gloss overlap measure of different relationships between synsets in WordNet [39]. This algorithm returns best results for nouns and adjectives [40].

$$SenseScore_k = \sum_{i=-n}^{n} \sum_{j=1}^{|w_i|} Rel(S_{0,k}, S_{i,j}), \quad i \neq 0$$

where, $SenseScore_k$ is the score of the k^{th} sense of a word w_0, $|w_i|$ designates the number of candidate senses of the word w_i, n is the context window around the target word w_0, $S_{0,k}$ is the k^{th} sense of w_0, $S_{i,j}$ denotes the sense j of w_i, and Rel is the relatedness measure based on WordNet.

4.3 Concept Weighting

The purpose of weighting is to give the indexed terms weights that tend to reflect their importance in the documents in which they appear. In classic weighting methods, a Boolean aspect is based on the presence or absence of query elements in a document. In this case, the weight of a relevant document which does not lexically contain the query vocabulary, but is semantically linked to it, is null. Moreover, documents which are lexically identical but semantically different will have high weights. To remedy these limitations, we have proposed a weighting formula at the conceptual level to give importance to the elements which are semantically related by moving from the level of terms to the level of concepts. Indexed items are concepts related to words contained in documents through semantic relationships.

4.3.1 Concept frequency

Based on the principle that a concept is better representative of the content of a document if it is more frequent locally in the document and that on the other side and it is more discriminative if it is less frequent overall in the corpus, the following degrees were considered:

The locality degree: This factor is defined by the frequency of the concept in the document and all synonymous concepts are counted together.

$$freq(s,d) = \sum_{\forall\, s_i \in synset(s)} freq(s_i,d)$$

freq(*s*,*d*) the number of occurrences of a concept *s* in a document *d* which is equal to the sum of the frequencies for all $s_i \in synset(s)$ in the document *d*.

The integrality degree: This factor is defined by the sum of the occurrences of the concept and its synonyms in the entire corpus.

$$freq(s,C) = \sum_{\forall\, s_i \in synset(s)} freq(s_i,C)$$

freq(*s*,*C*) is the number of occurrences of a concept *s* globally in the corpus which is equal to the sum of the frequencies of all the $s_i \in synset(s)$ concepts in the corpus.

4.3.2 Specificity

The specificity *Spec* of a concept-sense *s* is estimated by its "depth" in the ontology (WordNet) induced by the "is-a" relation. As the quantity of web information is very huge these days, it is important to integrate specificity factor into the representation of documents. For very large collections made up of a variety of texts, the notion of specificity has a more discriminative aspect (high *Spec* value) than that of collections specialized in a precise field and which consists of limited number of documents. Thus, specificity is an important factor in measuring the degree of information specialization. Hence, the system retrieves the most general documents for novices by reference with the entered keywords. On the other hand, it retrieves documents that are more specialized in the search domain for an expert who introduces more precise keywords than a novice.

The measurement of specificity is formalized by:

$$Spec_S = \#Rel_{(hyper)}(s, root)$$

$Spec_S$ is estimated by the number of hypernyms relations (*hyper*) that must be traversed to reach the concept-meaning *s* from the root of the ontology.

4.3.3 Impact of Noun tagged terms

According to WordNet statistics, the unique strings of Noun POS constitute more than 75% of the total number of indexed words. Hence, the value of $\overline{POS(N,C)}$ has been used as a smoothing parameter of the final score.

$$\overline{POS(N,C)} = \frac{1}{|C|} * \sum_{\forall\, t, POS(t)=NOUN} freq(t,C)$$

Where, *freq*(*t*, *C*) is the number of terms tagged as nouns throughout the *C* collection, and |*C*| is the size of the corpus.

4.3.4 Tautology

This factor is defined here by the number of synonyms of a concept *s* in a document *d*. This factor is directly proportional to the importance of a sense *s* within the document *d* as it is used to enrich the text and to be able to express more clearly. The author [41] mentions the "re-homogenization" as a type of tautology with the function of re-homogenizing the sense to avoid the risk of context heterogenized by introducing a certain trait. In other words, using tautology enforces the homogeneity degree.

$$Tau(s,d) = |Syn(s)|_d$$

4.3.5 Final weight

The proposed weighting model $w(s_t)$ of a concept *s* associated with the term *t* is formalized, as follows:

$$w(s_t) = \frac{F + Tau(s_t,d) + Spec_{s_t} * \overline{POS(N,C)}}{freq(s_t,C) + \overline{POS(N,C)}},$$

$$F = \alpha\, freq(s_t,d) + (1-\alpha) * \big(\, freq(t,d) - freq(s_t,d)\big)$$

Where α is a parameter whose value is set empirically, *freq*(s_t,*d*) is the frequency of the sense identified by a term *t* within a document *d*, *freq*(*t*,*d*) is the frequency of the term *t* within a document *d*, and *freq*(s_t,*C*) is the frequency of the sense identified by a term *t* throughout the collection *C*.

The free parameter α is used to reduce the impact of error that may occur during the process of WSD. After many experiments, it is preferred to set α to 0.9. This

means that 10% of the frequency factor value comes from counting the term itself when the s_t is not present in the document, which treats the case in which WSD algorithm drifts the correct sense of the term away. The value of $\overline{POS(N,C)}$ is used to smooth the final scores and it is weighted by the specificity to highlight its significance.

5 Experimental Works

The retrieval model should generally be validated empirically rather than theoretically [42]. SIRA has been tested using the TREC test collection, which consists of documents in xml format. 50 queries were selected to carry out a comparative study between SIRA and two main Baselines: The first (denoted C_TFIDF) corresponds to a classic index based on keywords weighted by TFIDF and the second (denoted C_BM25) corresponds to a classic index based on key words weighted by Okapi-BM25. The proposed semantic index is then considered. Two more baselines based on semantic index are handled. They exploit TFIDF and Okapi-BM25 and denoted S_TFIDF and S_BM25 respectively.

5.1 α Setting

Many experiments have been carried out to set the optimized value of α parameter. Each experiment evaluates the standard measures for SIRA for a specific value of α. Twelve values of α are tested, starting from 0 to 1. The evaluated measures are MAP, GMAP, R_{prec}, bpref, MRR, and interpolated precision for the standard values of recall. Table (1) compares the performance of SIRA with respect to α values.

Table (1). Results of MAP, GMAP, R_{prec}, bpref, MRR for α values

α	MAP	GMAP	Rprec	bpref	MRR
1.0	0.2271	0.1283	0.2646	0.5896	0.6375
0.95	0.2479	0.1497	0.2871	0.6562	0.6566
0.9	0.2503	0.1538	0.2911	0.6711	0.6571
0.8	0.2511	0.1499	0.2849	0.6735	0.6477
0.7	0.2518	0.1407	0.2885	0.6651	0.6369
0.6	0.2488	0.1296	0.2828	0.6589	0.6146
0.5	0.2383	0.1131	0.2647	0.6477	0.5934
0.4	0.2158	0.0794	0.2445	0.5998	0.5421
0.3	0.2053	0.0615	0.2283	0.5793	0.4798
0.2	0.1916	0.0549	0.2180	0.5292	0.4484
0.1	0.1850	0.0475	0.2086	0.5127	0.4162
0.0	0.1464	0.0436	0.1699	0.4346	0.3133

The analysis of Table (1) results shows that the maximum value of MAP is 0.2518. The MAP hits the value of 0.25 by three values of α, i.e. 0.7, 0.8, and 0.9. MAP values are proportionally increasing as α increasing from 0.0 to 0.7 and it declined after that. It is noticeable that the absolute value of MAP difference regarding the three mentioned α values is just 0.0015 which drives more study to the rest measures in the three α values. The values of GMAP, R_{prec}, MRR regarding α = 0.9 are higher than the corresponding measures in the other two cases of α. Finally, bpref measure hits 0.67 twice, namely when α = 0.9 and α = 0.8. However the previous discussion shows that α = 0.9 is the most recommended value to be chosen. Other experiments are conducted to set α correctly.

Figure (2) shows the recall-precision graphs of SIRA with selected values of α. The performance of SIRA with α = 0.9 is higher than the other variants with other values of α. Hence, the overall performance of SIRA powered by α=0.9 outweighs the other variants.

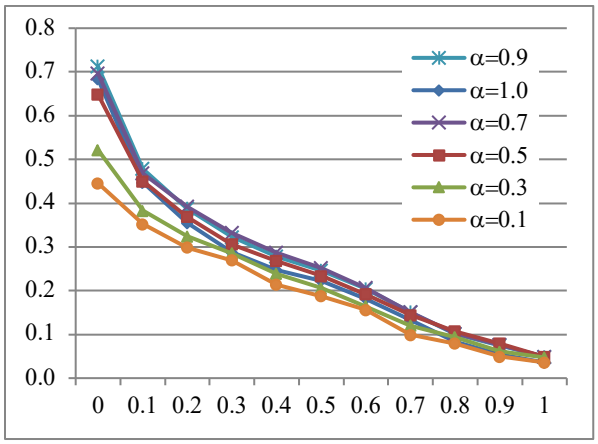

Figure (2). Recall-precision graphs for selected α's

5.2 Comparison of MAP, GMAP, and Rprec

In this experiment, three measures were evaluated to compare between the five models, namely MAP, GMAP, and R_{prec}. Table (2) shows the values of each measure in each model. Figure (3) visualizes these values and Table (3) shows the mutual models improvements. The experiment results show the great improvement of MAP by SIRA over the other models. The weak competitor was the classical TFIDF and the best one is classical BM25. Also, BM25 shows improvement in MAP ranging from 56% to 49% over the other three models. The S_BM25 outperforms the other models (C_TFIDF, S_TFIDF, S_BM25). The same trend appears in the other two measures (GMAP and R_{prec}). Indeed, SIRA considerably improves these two measures. Also, the GMAP and the Rprec of C_BM25 surpasses those of three models S_BM25, C_TFIDF and S_TFIDF.

The MAP is the most important and widely used measure in IR because it estimates the overall performance of the IR system. The power behind the MAP comes from the fact that it considers the average precision calculated for each query. The MAP results show that the overall performance of SIRA is better than the other competitors. GMAP by its definition focuses on improving low-performing queries. Results show that SIRA outperforms the rest models and classical BM25 comes at the second stage. The R_{prec} measure represents the calculated precision for the R^{th} relevant document returned. This measure de-

focuses the proper ranking of the retrieved relevant documents, which may be helpful when large count of relevant documents is present. SIRA outperforms classical TFIDF nearly twice and C_BM25 by 9.5%.

Table (2). Results of MAP, GMAP, R_{prec}

Model	MAP	GMAP	R_{prec}
SIRA	0.250	0.154	0.291
C_TFIDF	0.135	0.036	0.150
S_TFIDF	0.135	0.042	0.158
C_BM25	0.210	0.129	0.266
S_BM25	0.141	0.053	0.177

Table (3). Models improvements of MAP, R_{prec}

Percentage of Improvement	MAP	Rprec
SIRA over C_TFIDF	85.6%	93.9%
SIRA over S_TFIDF	85.4%	83.7%
SIRA over S_BM25	77.4%	64.3%
SIRA over C_BM25	19.1%	9.5%
C_BM25 over C_TFIDF	55.9%	77.2%
C_BM25 over S_TFIDF	55.7%	67.8%
C_BM25 over S_BM25	48.9%	50.1%
S_BM25 over C_TFIDF	4.7%	18.0%
S_BM25 over S_TFIDF	4.5%	11.8%
S_TFIDF over C_TFIDF	0.1%	5.6%

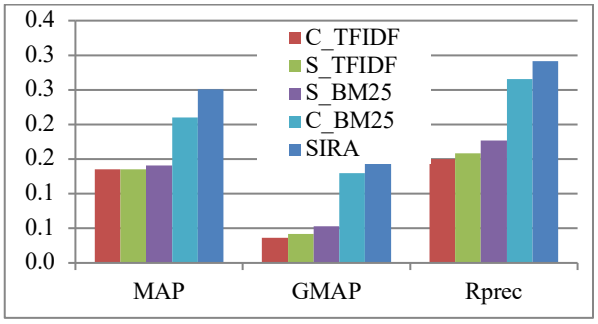

Figure (3). Comparison of MAP, GMAP, and R_{prec}

5.3 Comparison of MRR and Recall

The second experiment measures another two classic measures: Mean Reciprocal Rank (MRR) and Recall. Table (4) presents the measured values of both metrics. Figure (4) depicts these values and Table (5) shows SIRA percentage of enhancement. The results show that SIRA improves Recall by 65%, 61%, 44%, and 14% over S_TFIDF, C_TFIDF, S_BM25, and C_BM25, respectively. Recall measure by its definition reflects the power of a system to retrieve all relevant documents. Accordingly, SIRA can retrieve more relevant documents than the rest of competitors. Against to R_{prec}, MRR is based on the multiplicative inverse of the rank of the first retrieved relevant document which reflects the ranking quality. The value of MRR defines which model can identify the first correct hit in average on the total number of tested queries. SIRA comes at the first stage followed by C_BM25 by improvement equals 17.5%. Higher improvements are achieved with S_BM25, S_TFIDF, and C_TFIDF by 44%, 75%, and 59%, respectively.

Table (4). Results of MRR, Recall

Model	MRR	Recall
SIRA	0.657	0.671
C_TFIDF	0.414	0.459
S_TFIDF	0.375	0.499
C_BM25	0.559	0.611
S_BM25	0.456	0.463

Table (5). Improvements of MRR, Recall

Percentage of Improvement	MRR	Recall
SIRA over C_TFIDF	58.7%	60.6%
SIRA over S_TFIDF	75.2%	65.1%
SIRA over S_BM25	44.1%	44.3%
SIRA over C_BM25	17.5%	13.5%

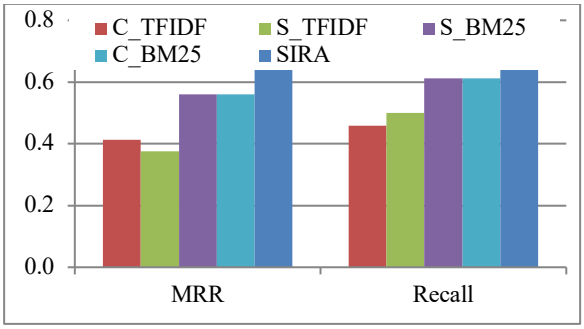

Figure (4). Comparison of models in MRR, Recall

5.4 Comparison of Recall-Precision Graphs

In this experiment, a comparison of recall-precision graphs of the five models is carried out. These graphs show the retrieved documents ranking at different standard values of recall. The recall-precision graph usually declined from left to right. This means that as we get more relevant documents (growing recall), we get more irrelevant documents (diminishing precision). Figure (5) presents graphical comparison among the five models. It shows that SIRA curve is on the top of the rest curves, which means that SIRA is the superior of the models. Indeed, it retrieves more relevant documents at all recall points. The curve closest to SIRA shows that C_BM25 is the most competitive model compared to the other three models.

5.5 Comparison of P@x

The fourth experiment compares precision@*x* (P@x) to the underlying models, where *x*=5, 10, 15, 20, 30, 100, 200, 500, 1000. P@x is evaluated to the precision in the x^{th} retrieved document. Figure (6) shows the accuracy of each model at different x points. It is noticeable that the precision of SIRA surpasses those of the other reference models for all the points of x. This means that SIRA rejects

more irrelevant documents in each standard step compared with the other models.

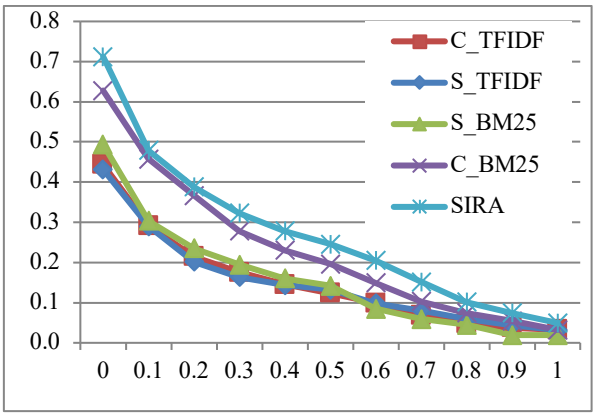

Figure (5). Average recall-precision graph

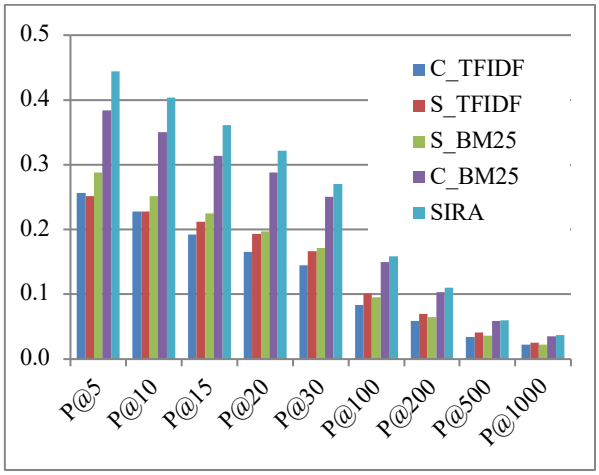

Figure (6). Precision @ different ranks

6 Conclusions

The present paper introduced a novel conceptual weighting model for semantic IR to bridge the gap in classic IR. The proposed weighting formula depended on various factors, such as locality and integrality degrees of the concept. Also, it exploited the concept tautology, degree of concept specificity, and a corrector parameter to alleviate the uncertainty resulting from the process of WSD. To validate the proposed model, it was integrated into a complete IR approach which comprised three main steps: Concept identification, indexing, and document scoring. Concurrency was used in every phase of our approach to overcome the complexity of the system. The proposed approach was validated by comparing it to some important well-cited weighting models. The results showed that the proposed approach outperformed the other benchmark models based on the standard field measures. This improvement reflects the importance of semantic IR to enhance the process of IR.

In future work, the technique of identifying concepts may be strengthened by combining several semantic resources to cover as many concepts as possible and to avoid nonempty terms that are unrelated to any concept in the ontology. Also, a WSD may be enhanced based on the different semantic relationships derived from combined resources.

References

[1] D. Genest and M. Chein, “A content-search information retrieval process based on conceptual graphs,” *Knowl. Inf. Syst.*, vol. 8, no. 3, pp. 292–309, 2005.

[2] J.-P. Chevallet, *Ressources endogènes et exogènes pour une indexation conceptuelle intermédia*. Université de Grenoble, 2009.

[3] D. Dinh and L. Tamine, “L.: Vers un modèle d’indexation sémantique adapté aux dossiers médicaux de patients (short paper,” in *In: Conférence francophone en Recherche d’Information et Applications (CORIA), Sousse, Tunisie, 18/03/2010-21/03/2010, Hermès (mars 2010) 325--336.*

[4] M. Baziz, M. Boughanem, Y. Loiseau, and H. Prade, “Fuzzy logic and ontology-based information retrieval,” in *Fuzzy Logic*, Springer, 2007, pp. 193–218.

[5] F. Boubekeur and W. Azzoug, “Concept-based indexing in text information retrieval,” *arXiv Prepr. arXiv1303.1703*, 2013.

[6] B. Scarlini, T. Pasini, and R. Navigli, “With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 3528–3539.

[7] H. Kammoun, I. Gabsi, and I. Amous, “MeSH-Based Semantic Indexing Approach to Enhance Biomedical Information Retrieval,” *Comput. J.*, 2020.

[8] D. C. A. Jaco, “Extended vectorial model ACP of latent semantic indexation in the natural language processing for the search and retrieval of information in electronic documents,” in *2018 IEEE 38th Central America and Panama Convention (CONCAPAN XXXVIII)*, 2018, pp. 1–6.

[9] E. M. Voorhees, “Using WordNet to disambiguate word senses for text retrieval,” in *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 1993, pp. 171–180.

[10] B. Katz, O. Uzuner, and D. Yuret, “Word sense disambiguation for information retrieval,” 1998.

[11] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, "Conceptual indexing based on document content representation," in *International Conference on Conceptions of Library and Information Sciences*, 2005, pp. 171–186.

[12] M. Baziz, M. Boughanem, and N. Aussenac-Gilles, "The use of ontology for semantic representation of documents," in *The 2nd Semantic Web and Information Retrieval Workshop (SWIR), SIGIR*, 2004, pp. 38–45.

[13] F. Boubekeur, "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets," 2008.

[14] F. Boubekeur, M. Boughanem, L. Tamine, and M. Daoud, "Using WordNet for Concept-based document indexing in information retrieval," in *Fourth International Conference on Semantic Processing (SEMAPRO), Florence, Italy*, 2010.

[15] F. Boubekeur, W. Azzoug, S. Chiout, and M. Boughanem, "Indexation sémantique de documents textuels," in *14e Colloque International sur le Document Electronique (CIDE14), Rabat, Maroc*, 2011.

[16] M. Torjmen, K. Pinel-Sauvagnat, and M. Boughanem, "Towards a structure-based multimedia retrieval model," in *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 2008, pp. 350–357.

[17] M. Boughanem, I. Mallak, and H. Prade, "A new factor for computing the relevance of a document to a query," in *International Conference on Fuzzy Systems*, 2010, pp. 1–6.

[18] F. Shi, L. Chen, J. Han, and P. Childs, "A data-driven text mining and semantic network analysis for design information retrieval," *J. Mech. Des.*, vol. 139, no. 11, 2017.

[19] F. Harrathi, C. Roussey, S. Calabretto, L. Maisonnasse, and M. M. Gammoudi, "Indexation sémantique des documents multilingues," *INFORSID, Ed. Atelier RISE Assoc. au 27ème Congrès INFORSID*, pp. 31–50, 2009.

[20] B. Magnini and G. Cavaglia, "Integrating Subject Field Codes into WordNet.," in *LREC*, 2000, pp. 1413–1418.

[21] L. Khan, D. McLeod, and E. Hovy, "Retrieval effectiveness of an ontology-based model for information selection," *VLDB J.*, vol. 13, no. 1, pp. 71–85, 2004.

[22] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *J. Am. Soc. Inf. Sci.*, vol. 27, no. 3, pp. 129–146, 1976.

[23] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, and others, "Okapi at TREC-3," *Nist Spec. Publ. Sp*, vol. 109, p. 109, 1995.

[24] A. Singhal, G. Salton, M. Mitra, and C. Buckley, "Document length normalization," *Inf. Process. Manag.*, vol. 32, no. 5, pp. 619–633, 1996.

[25] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Doc.*, 2004.

[26] F. Sebastiani, "Consiglio Nazionale Delle Ricerche," *Mach. Learn. Autom. text Categ. ACM Comput. Surv.*, vol. 34, pp. 1–47, 2002.

[27] S. Chagheri, S. C. CR, and C. Dumoulin, "Semantic indexing of technical documentation," *Lab. d'InfoRmatique en Image Systèmes d'information*, vol. 12, 2009.

[28] F. Debole and F. Sebastiani, "Supervised term weighting for automated text categorization," in *Text mining and its applications*, Springer, 2004, pp. 81–97.

[29] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 721–735, 2008.

[30] H. Alt\inçay and Z. Erenel, "Analytical evaluation of term weighting schemes for text categorization," *Pattern Recognit. Lett.*, vol. 31, no. 11, pp. 1310–1323, 2010.

[31] Y. Liu, H. T. Loh, and A. Sun, "Imbalanced text classification: A term weighting approach," *Expert Syst. Appl.*, vol. 36, no. 1, pp. 690–701, 2009.

[32] D. Wang and H. Zhang, "Inverse-category-frequency based supervised term weighting scheme for text categorization," *arXiv Prepr. arXiv1012.2609*, 2010.

[33] T. T. Nguyen, K. Chang, and S. C. Hui, "Supervised term weighting centroid-based classifiers for text categorization," *Knowl. Inf. Syst.*, vol. 35, no. 1, pp. 61–85, 2013.

[34] T. Peng, L. Liu, and W. Zuo, "PU text classification enhanced by term frequency--inverse document frequency-improved weighting," *Concurr. Comput. Pract. Exp.*, vol. 26, no. 3, pp. 728–741, 2014.

[35] Z.-H. Deng, K.-H. Luo, and H.-L. Yu, "A study of supervised term weighting scheme for sentiment analysis," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3506–3513, 2014.

[36] R. Blanco and C. Lioma, "Graph-based term weighting for information retrieval," *Inf. Retr. Boston.*, vol. 15, no. 1, pp. 54–92, 2012.

[37] H. Zargayouna, "Contexte et sémantique pour une indexation de documents semi-structurés.," *CORIA*,

vol. 4, pp. 161–177, 2004.

[38] S. Fodeh, B. Punch, and P. N. Tan, “On ontology-driven document clustering using core semantic features,” *Knowl. Inf. Syst.*, vol. 28, no. 2, pp. 395–421, 2011.

[39] S. Banerjee and T. Pedersen, “Extended gloss overlaps as a measure of semantic relatedness,” in *Ijcai*, 2003, vol. 3, pp. 805–810.

[40] M. B. Billami and N. Gala, “Approches d ’ analyse distributionnelle pour améliorer la désambiguïsation sémantique Approches d ’ analyse distributionnelle pour améliorer la désambiguïsation sémantique,” no. JUNE, pp. 2016–2020, 2016.

[41] H. Abé, “La tautologie et la notion subjective de ‘désirabilité,’” in *Current Issues in Unity and Diversity of Languages, Collection of the papers selected from the 18th International Congress of Linguists, Published by The Linguistic Society of Korea*, 2009, pp. 3266–3278.

[42] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*, vol. 520. Addison-Wesley Reading, 2010.