

SemApp: A Semantic Approach to Enhance Information Retrieval

Sameh Neji[1](✉), Tarek Chenaina[2], Abdullah M. Shoeb[3], and Leila Ben Ayed[2]

[1] Faculty of Economics and Management, Sfax University, Sfax, Tunisia
neji.sameh@yahoo.fr
[2] National School of Computer Science, University of Manouba, Manouba, Tunisia
Tarek.chenaina@ensi.edu.tn, leila.benayed@ensi-uma.tn
[3] Faculty of Computers and Information, Fayoum University, Fayoum, Egypt
ams02@fayoum.edu.eg

Abstract. The present work proposed a semantic retrieval approach to treat the issues of semantic ambiguity of indexed terms, the uncertainty, and imprecision that is inherent in the information retrieval process. The proposed approach constitutes of three different phases. The query meaning was discovered in the first phase by formulating a set of candidate queries from possible contexts. A score for each alternative was calculated based on its semantic tree and inherent dispersion between its concepts. This score assesses the overall meaning of the alternative query. This phase was finished by selecting the candidate query that attains the highest score to be the best representative to the original query. A semantic index was built in the second phase exploiting the classic and semantic characteristics of the document concepts to finally assign a weight for each concept to estimate its relative importance. The third phase proposed a ranking model that utilizes the semantic similarities and relations between concepts to calculate the query-document relevance. This ranking model is based on a query likelihood language model and a conceptual weighting model. The validity of the proposed approach was evaluated through performance comparisons with the related benchmarks measured in terms of the standard IR performance metrics. The proposed approach outperformed the compared baselines and improved the measured metrics. A statistical significance test was conducted to guarantee that the obtained improvements are true enhancements and are not a cause of random variation of the compared systems. The statistical test supported the hypothesis that the obtained improvements were significant.

Keywords: Information retrieval · Semantic information retrieval · Query reformulation · Semantic indexing · Conceptual weighting · Semantic relations · Ranking model · Query language model

1 Introduction

The aim of IR is to respond effectively to users' queries by retrieving information that better meets their expectations. Information needs are dependent on the user defined

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12951, pp. 62–78, 2021.
https://doi.org/10.1007/978-3-030-86970-0_6

context. Primarily, the level of domain expertise of the user directly influences the query richness. However, the information retrieval system (IRS) must be intelligent enough to identify information needs and respond effectively to meet the expected needs, regardless the level of user expertise level. This process is difficult and remains a major and open challenge in the domain of IR.

The present work fits into the general problem of searching for textual information. The huge increase in the amount of information accessible on the internet has resulted in an excessive demand for tools and techniques capable of semantically processing web data. The current practice of information retrieval is based mainly on keywords which are modeled with a bag of words. However, such a model lacks the actual semantic information in the text. To deal with this problem, ontologies are proposed for the representation of knowledge, which is the backbone of semantic web applications. Information retrieval processes can benefit from this metadata, which gives semantics to the text.

There is a semantic information search gap between the data, the application, and the used algorithms. Semantic Web technology has an intelligent role in bridging the semantic gap and retrieving relevant information through data mining algorithms for systematic knowledge incorporation. The model [1] was acquired by analyzing user behavior to record user information based on his/her interests.

Ontologies are the main tools for the representation of knowledge. They facilitate the classification as well as the cartography of concepts and their relations in a hierarchical structure. The literature shows that the use of conceptual knowledge (ontologies) in the information retrieval process has helped to overcome the main limitations of information retrieval. An ontology-based IRS is presented in [2]. In this model, the extraction of the ontology vocabulary is performed on the query terms. The terms of the documents are extracted by projection onto the ontology and they are presented by a vector space. The relevance between the concepts of the document and query is calculated using a correlation matrix between these concepts.

The main limitations concern the relations between the different concepts are the difficulty of inference of semantic information from concepts, the need to unify the semantic representations, and the mapping of knowledge from heterogeneous ontologies.

In [3], an OWL-based (Web Ontology Language) search engine was proposed. The proposed indexing algorithm offers better ontology maintenance and the retrieval algorithm facilitates the processing of user queries. The limitation here is that the authors manually tested the indexing algorithm considering a sample of ten OWL ontologies. In [4], the authors presented an IRS based on domain ontology. The meanings of user query concepts are inferred from the domain ontology and they are used for query expansion. The expansion of queries has positive effects on the search for relevant information. The goal of query expansion can be either to expand the set of returned documents by using terms similar to the query words, or to increase search accuracy by adding completely new terms to the query. Semantic similarity between inferred concepts and domain ontology concepts is obtained using structure-based measures [5].

After obtained the semantic knowledge and represented it via ontologies, the next step is to inquiring the semantic data to extract the relevant information. There are query languages designed for semantic enquiries, but these formal query languages are not intended to be used by end users because it requires knowledge of the ontology

domain as well as the syntax of the language. Combining the ease of using key-word-based interfaces with the power of semantic technologies is considered among the most difficult areas of semantic search.

In our view, all efforts to increase retrieval performance while maintaining usability will ultimately improve semantic search with keyword-based interfaces. This is a difficult task because it requires answering complex queries with only a few natural language keywords. The present work introduces a novel approach that utilizes three successive phases to improve IR: query meaning extraction, conceptual weighting documents representation, and a hybrid ranking model based on a query likelihood language model.

2 Related Works

In natural language processing, the similarity between words is often distinct from the relations between them. Similar words are almost synonyms while related words are connected to each other through semantic relations. When interpreting the query, the system can integrate other information to better meet the information user's needs. The content of the user's query is not always optimal for effective search. To improve the search performance which mainly depends on this query, various methods seek to improve its content by coordinating between the query and an effective index. While other methods, are interested in studying the relevance between the query and the document. Other systems study the behaviors and interests of the user to extract relevant information. Several methods use expansion techniques to extend the query by enriching the content of this query with other information [6, 7]. The extension of the query can be done by referring to ontologies and thesauri to discover syntactic, lexical or semantic relationships between words [8]. Other methods use additional information to identify meanings of query concepts in a disambiguation task [9]. Although statistical methods are used, the principle techniques are based on term co-occurrence to generate correlations between pairs of terms [10]. In [11], the author evaluates the co-occurrence by considering the frequency of terms found only in the first retrieved document. Other relevance feedback methods use user feedback to assess the relevance of documents that will be used to extend the query. This technology is based on the semantics of words in order to recognize their meaning in a specific context.

Discovering the correct meaning of the query is insufficient on its own to cope with IR challenges. Optimizing the representations of documents and queries is very important to extract the semantic information they contain. Term weighting technique is used in indexing phase rather than in the search phase. Several indexing approaches are based on external semantic resources. The basic principle of these approaches is to extract the representative elements from the document and the query. To assign degrees of importance to indexed terms, weighting algorithms are used to reduce the feature space to more specific terms in the search domain using knowledge bases such as ontologies [12]. In particular, the authors in [13] have proposed a term weighting scheme which combines the semantic similarity between each term for a category of the text and its frequency in the document to obtain the characteristic vector of each document. This approach outperforms TF-IDF in case of training data shortage and in case of the content of documents is focused on well-defined categories.

In the weighting model described in [14], the importance of a concept is assessed by a measure of its semantic similarity with the other concepts of the document which is combined with the frequency of their occurrences in this document.

Another semantic indexing approach [15] was extracted concepts by a construction of a semantic network by projection on WordNet. A semantic weight, which is adapted from the TF-IDF model, is calculated for each concept. In [16], the authors propose an indexing approach which is based on the frequency of the concept in the document with estimation of the semantic distances between each concept and the other most frequent concepts in a document. The results show that the use of TF-IDF weighting in semantic indexing increases precision by more than 50%. However, using the proposed semantic weighting, the precision is still low.

Recent approaches have been proposed for semantic indexing. This work [17], tackled the problem of indexing documents in the Arabic language. The author uses the Universal WordNet ontology to extract the index's concepts. This method improves the capacity of the semantic vector space model VSM.

To customize the search according to the user's interests, the work [18], presents a model to disambiguate queries. This model highlights the behavior of users via the history of conducted searches which is limited and insufficient to detect new information needs. To solve this problem, the user profile detection method uses semantic relations of ontologies to deduce its interests according to a hierarchy of concepts [19, 20].

Indexing is therefore an effective solution which consists in extracting descriptive elements from a document. In this case, the document is converted to a smaller structure than the original full text. This new representation contains the same amount of information. To successfully perform IR, index descriptors must be well exploited in the search phase. Ranking retrieved documents is an important task that measures query-document relevance. Some search models have been proposed to obtain semantic information between formal concepts [21]. However, assigning equal importance to all terms can significantly reduce the quality of the IRS.

Several approaches using fuzzy concept networks based on fuzzy formal concept analysis (FFCA) [22–24] have been proposed to address this challenge. They adopt the partial order relation of the concepts to calculate the relation between them and return a list of documents associated with the given query. However, these methods neglect explicit semantic information between concepts.

In [25], the authors conceived the idea of an intelligent IRS based on semantic web. In this model, the semantic relations between web pages are used to estimate the similarity between queries and documents. In [26], the authors proposed a model of knowledge search process using semantic metadata and artificial intelligence tech-niques. The works [27, 28] were introduced an automatic derivation of concept hyponymies and meronymies relations from web search results. Its method was based on generating derivative features from web search data and applying the machine learning techniques.

Language models have been applied in various applications such as ad hoc research, expert research and social research. These models estimate the relevance of a document to a query using the probability that this query can be generated by this document [29]. This model is based on the maximum likelihood estimate. Using of a bag of words, which

assumes the independence of words, is common in the approaches based on language models. However, natural language terms are often dependent.

Returning relevant results is the primary goal of IR. To achieve this goal, the proposed approach considers two main factors before displaying results. The first factor is the extraction of the information from the semantic processing which discovers the meaning of the query. The second factor, that directly influences the precision of retrieval system, is the estimation of query-document relevance exploiting semantic indexing of concepts the related to terms of documents and queries.

3 The Proposed Approach

The present work is based on general domain ontology for extraction and retrieval of semantic information. Previous works [30–33] were interested in developing individual steps to improve some parts of the IRS separately. The first phase of the proposed framework is to disambiguate the user's query to resolve the ambiguity caused by the multiple senses for each term in the query. The second phase is to disambiguate the queries and documents terms. They are replaced and indexed with a new format. A semantic weight is calculated for each concept that reflects the degree of importance of each unit in the index. This step improves the representation of documents and queries. The third phase measures the query-document relevance by calculating a semantic score for each document. This score exploits the conceptual relations between the documents, query terms, and the meanings of these terms in the context of the document. The search results will be ranked according to this relevance score.

Most IRSs use a keyword-based search mechanism. However, the proposed framework defines information in terms of concepts. The concept can be considered as a sequence of terms which identify a clear meaning of the context concerned. The present work introduces a complete semantic processing of information that encompasses the three major phases of information retrieval in a single approach. The originality of this work compared to previous works is that it presents a general integrated semantic approach for processing information and retrieving relevant documents with respect to a query introduced in natural language in the form of generally ambiguous keywords. The limitations that mentioned in the previous section show the need to a unified system that conducts all the necessary steps for an effective IR. The present work contribution is proposing an IR system which processes information in three consecutive phases, Fig. 1. This system uses ontologies and semantic relations to represent documents and define the meanings of terms. It is intelligent enough to retrieve documents that do not contain the terms of the query.

3.1 Discovering Query Meaning

The query disambiguation process leads the search of documents that are truly related to the user's information needs. The ambiguity can be lexical (polysemy, disambiguation), or structural (type of information expected). The ambiguity of the query can lead to heterogeneity of documents belonging to different contexts.

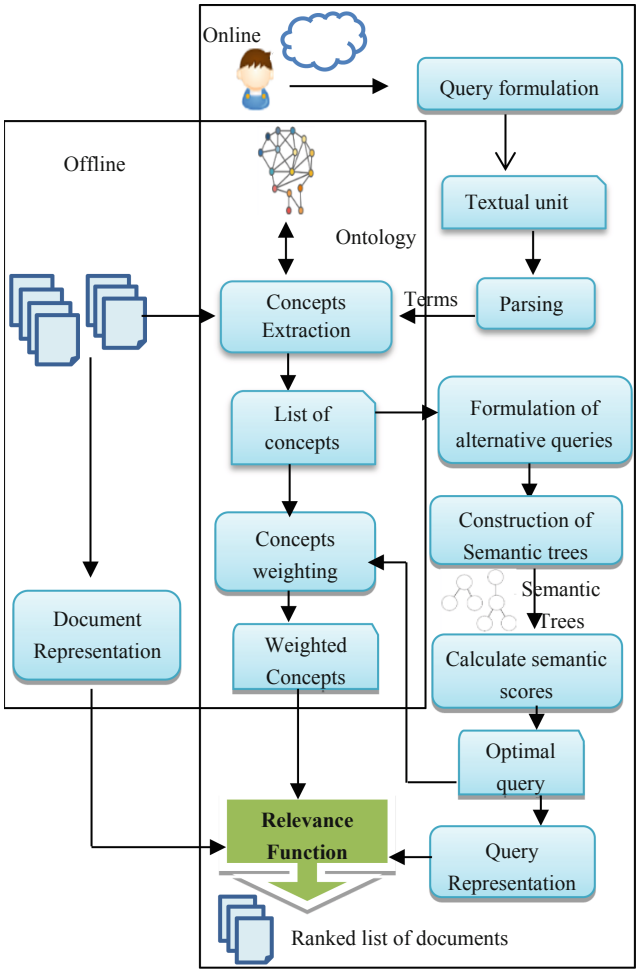

Fig. 1. The proposed approach outlines

Usually, the query introduced by the user is insufficient to return the most relevant documents. The query must be very clear to express the user's needs. The system is based on this query to retrieve the relevant documents. The search results will therefore be meaningless if the query is poorly expressed. Accordingly, the first step in the proposed technique for improving the performance of IRS is to discover the meaning expected by the user.

3.1.1 Extraction of Query Concepts

This task consists of identifying the concepts of the query. This is very important step to improve the performance of the system. Identifying concepts requires a word sense disambiguation step because of word ambiguity in natural language. A concept in the text can be represented by more than a term. The purpose of term disambiguation is

therefore to select the most appropriate meaning. In specific domain, it is easier to define the meaning of terms since using more specialized terms in a specific domain can help to define the precise context than using more ambiguous terms in a general domain.

Consider the example of the query 'let the cat out of the bag'. The single word concepts are: 'let', 'cat', 'out', 'bag'. Table 1 shows the vast variety of possible meanings of this query in different contexts. Discovering the meaning of the query or otherwise the meaning expected by the user is a great challenge especially in cases where the query has several different meanings.

Table 1. Number of possible senses with the POS of concept

Concept	Number of senses				
	Noun	Verb	Adjective	Adverb	Total
Let	2	9	–	–	11
Cat	8	2	–	–	10
Out	1	3	10	3	17
Bag	9	5	–	–	14
Number of possible senses					26180

The multi-word concept 'let the cat out of the bag' has only one meaning as a verb. Its definition in WordNet is: "spill the beans, let the cat out of the bag, talk, tattle, blab, peach, babble, sing, babble out, blab out - (divulge confidential information or secrets; "Be careful – his secretary talks")". If the different meanings of each term were considered, it yields 26180 different meanings for the query in the case of separate words. So, the discovery of query meaning presents a difficult problem.

A very precise word sense disambiguation algorithm is needed to increase performance of IRS. Extended Lesk algorithm was used to disambiguate words by calculating a semantic score for each sense of the concept. This score is based on the extended measure of gloss overlap of different relations between synsets in WordNet. This algorithm returns better results for nouns and adjectives [34].

The original query was transformed to alternatives by a combination of all the possible concepts it contains. The same grammatical structure of the original query was kept so as not to deviate from the global meaning of the query.

3.1.2 Extraction of the Most Relevant Candidate Query

The retrieved documents may not meet the needs of the users because of wrong expression of the query. In this case, the level of user expertise and its direct effect on the accuracy of the results present a challenge to be tackled. Ambiguous words lead to a dispersion of the query context. To solve this problem, our approach extracts the closest context that maximizes the meaning of the query to help the user find relevant information. A semantic score is calculated for each candidate query. This score (*Can_Score*) reflects the semantic rapprochement with the user's expectations. The most relevant

(optimal) query is selected to pass to the next phase.

$$Can_Score = \frac{1}{D^{\alpha} \sum_{i=1}^{n} \frac{1}{(depth(c_i))^{\beta}} \sum_{\substack{c = c_i \\ c \in path(c_i, LCS)}} log(1 + (c, hyper(c)))} \quad (1)$$

Where n is the number of query concepts, c_i is the concept of rank i, LCS is least common subsumer of c_i, $hyperc$ is the hypernym of the concept c, δ $(c, hyper(c))$ is the similarity between the current concept c and its subsume, D is the coefficient of dispersion that calculate the number of branches containing at least one concept of the query, $depth(c_i)$ is the depth of the concept c_i with respect to LCS and α, β are two constants. Many experiments are conducted to set the values of α and β to finally assign to 2 and 1 respectively.

3.2 Indexing Phase

The use of semantic notion allows solving the problems of ambiguity of the indexed terms. Thus, exploitation of different semantic relations makes it possible to limit the disparity between the representation of documents and user's query. A semantic indexing technique based on semantic relations (is-a) and the synsets associated with the terms of the documents and the query was utilized. This technique makes it possible to represent the terms of documents and queries by their associated senses in the context in which they appear. An index based on the identified synsets was built. Representation of documents and queries was done by these synsets rather than the terms themselves.

3.2.1 Semantic Representation of the Document

The main problem with traditional IRSs is that they generally return too many results or irrelevant results. As keywords have multiple lexical characteristics or have multiple senses, in this case, relevant information may be lost because of not specifying the exact search's context. The resolution of the ambiguity problem by integration of the conceptual representation becomes necessary to represent the documents by a set of descriptors in the form of concepts belonging to a predefined semantic resource (ontology). Two issues that can be reported; the used external resource could be limited to a certain domain or it may miss certain terms. This representation brings new challenges because of the presence of synonymies and polysemy's in natural language, concerning the association of a term with the best concept.

A combined index based on terms and their senses was designed as the representation of data is a very important task in the IR process. The output of this phase is a new representation of the document and the query. Each term is replaced by the following format:

Term /POS/ Sens/ Spec/Tauto/ Weight

Where *Term* is the common term, *POS* is the part of speech of this term in the document (Part Of Speech), *Sens* is the concept sense associated with this term in WordNet,

Spec is estimated by the "depth" of the current concept in the ontology, *Tauto* is the cardinality of the concept's synsets in the document, *Weight* is the concept weighting. The system also indexes the terms that are not defined in the ontology.

3.2.2 Concepts Weighting

Term weighting is also an essential task in defining the importance of query and document terms. This weight is generally measured by a statistical method which is based on the presence of the term and its frequency in a text. Traditionally, the weighting score does not consider the senses of the terms. There is a problem with this method which neglects the notion of term's senses. This measure only considers the factor of existence of the concept in a document and is reflected by its frequency. This metric does not consider the semantic importance of the concept in the document. To remedy these problems, we utilize concept weighting model that calculates the importance of concepts according to different criteria that allow better presentation of relevant information.

Estimating the importance of each index item considers many factors that have improved the performance of the system. The index is built by weighted synsets. To give importance to each concept, we use the weighting model, Eq. (2).

$$w(s_t, d) = \frac{F + |Syn(s_t)|_d + \#Rel(hyper)(s_t, root) * \frac{1}{|C|} * \sum_{\forall t, POS(t)=NOUN} freq(t, C)}{\sum_{\forall si \in synset(s_t)} \sum freq(s_i, C) + \overline{POS(N, C)}} \tag{2}$$

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

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.

3.3 Ranking Phase

This process involves calculating a relevance measure for each document against a query. This measure reflects the semantic similarity between the query and the document. Unlike classical systems based on keywords, the semantic relevance formula uses the relations between the concepts. These relations are expressed in the used semantic resource.

An effective representation of documents and queries has been used to ensure effective retrieval of information expected by the user. The proposed approach benefits from the advantages of classical information retrieval. It combines between weighting formula and the language model which estimate the semantic similarities between the concepts. The document ranking score is defined by:

$$SemRank(D) = \left(\prod_{i=1}^{|Q|} P(q_i|D) \right)^{\gamma} * (Score_W(D))^{\delta}$$

Where γ and δ are two parameters to distinguish importance of each ingredients of the SemRank. The default values for these two constants are 1 and 3, respectively.

$$P(q_i|D) = \frac{\sum_{j=1}^{m} Sim(q_i, d_j)}{|D| * \underset{\substack{\forall d_k \in D' \\ D' \in C}}{\max} Sim(q_i, d_k)}$$

$$Score_W(D) = \prod_{i=1}^{n} \left(\frac{\left(w(q_i, D) - \min_{\forall D' \in C} w(q_i, D')\right) * \left(\underset{\forall D' \in C}{Max} P(q_i|D') - \min_{\forall D' \in C} P(q_i|D')\right)}{\left(\max_{\forall D' \in C} w(q_i, D') - \min_{\forall D' \in C} w(q_i, D')\right)} + \min_{\forall D' \in C} P(q_i|D') \right)$$

$$w(q_i, D) = \frac{\beta * F + Tau(q_i, D) + Spec(q_i) * \overline{POS(N, C)}}{freq(q_i, C) + \overline{POS(N, C)}}$$

$$F = \alpha freq(q_i, D) + (1 - \alpha) * (freq(t_i, D) - freq(q_i, D))$$

This ranking model combines between classical and semantic search. The semantic similarity measure "Resnik" is used to find $Sim(q_i, d_j)$.

4 Evaluation

The aim of the experimental work is to evaluate the quality of the proposed approach as a function of the quality of its retrieved ranked list by comparing this result with other benchmarks. The plan is started by selecting a widely used dataset collection. Then, we choose the relative benchmarks to compare with. The effectiveness of each system is assessed through the standard IR measures.

TREC collection, one of the standard test collections, and its relevance judgements were chosen to conduct the evaluations. The proposed approach exploits the semantic and classic characteristics of the documents; therefore, we built a semantic and a classic index to the dataset. SemApp is compared to TFIDF, BM25, and QL (query likelihood language model). Each benchmark is implemented on both classic and semantic index and it will be indicated by a prefix C or S respectively before its name to indicate the underlying index. Hence, there are six models for comparison: C_TFIDF, S_TFIDF, C_BM25, S_BM25, C_QL, and S_QL. A set of forty TREC queries, and their qrels, was selected to conduct the experiments. Each model was run on every topic and the effectiveness measures are assessed and compared to each other. SemApp ranking part sets the following parameters: $\alpha = 0.9$, $\beta = 2$, $\gamma = 1$, and $\delta = 3$.

4.1 Evaluation of MAP, GMAP, Rprec, and MRR Metrics

In this experiment, four IR measures were assessed; mean average precision (MAP), geometric mean average precision (GMAP), precision after R (the number of relevant documents for a query) documents retrieved (Rprec), and mean reciprocal rank (MRR).

Figure 2 demonstrates a comparison among the seven models in the four measures. Also, Table 2 shows the percentage of improvements that achieved by SemApp.

The MAP, among of other measures, plays an important role in evaluating the IR systems because MAP, by its definition, assesses the gross precision of the examined system. SemApp attains the highest MAP among the other competitors. The improvement percentages are ranged from 15.5%, over C_QL, to 130.9%, over S_TFIDF. The descending order of systems performance is as follows: SemApp, C_QL, C_BM25, S_QL, S_BM25, C_TFIDF, and then S_TFIDF respectively. The benchmark model that built on classic index shows significant improvement over the corresponding one that built on semantic index. This result may happen because of change in frequency of words that appears in the classical index into several frequencies corresponding to the different meanings of these words in the semantic index. Also, word sense disambiguation algorithm may drift the meaning away from the correct sense.

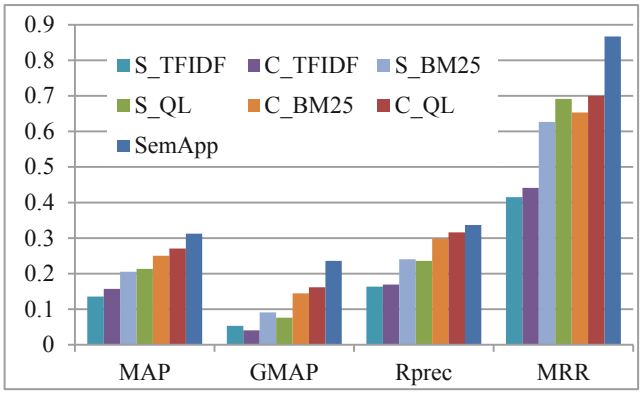

Fig. 2. Comparison of MAP, GMAP, Rprec, and MRR

Table 2. SemApp improvements percentage of MAP, GMAP, Rprec, and MRR

Improvement %	C_QL	S_QL	C_TFIDF	S_TFIDF	C_BM25	S_BM25
MAP	15.5%	46.6%	99.1%	130.9%	24.8%	52.2%
GMAP	46.0%	210.2%	483.3%	344.3%	63.5%	160.0%
R_{prec}	6.6%	42.9%	99.1%	106.5%	12.9%	40.2%
MRR	23.9%	25.4%	96.5%	109.0%	32.8%	38.3%

MAP is defined as the arithmetic mean of average precisions, so it assigns equal weight to each topic score change regardless of the relative size of the change. Alternatively, GMAP concentrates on the low performed topics; the topics that have low average precision. GMAP measures the impact of improvement in these special topics on the overall system performance. The experiment shows that the greatest GMAP appears in SemApp by a considerable gap with the other benchmarks. While C_QL and C_BM25

come in second and third places respectively, the worst behavior appears in C_TFIDF. This result shows that SemApp is capable of improving the low performed topics by 46% over C_QL and 63.5% over C_BM25.

Let R be the number of relevant documents, as judged in the qrels, of a query. The third assessment measure is Rprec which measures the precision after retrieve R documents whatever the relevancy of the retrieved documents. So, if an IR system returns x relevant documents within the first R retrieved documents; Rprec remains the same and independent of the location of the x documents within the top R-sized returned list. While SemApp comes first in Rprec results, the differences between SemApp and the nearest two competitors, C_QL and C_BM25 respectively, are not considerable. The improvement percentages over all benchmarks are ranged from 6.6% to 106.5%.

Reciprocal rank is defined as the multiplicative inverse of rank of the first retrieved relevant document. MRR is the arithmetic mean of reciprocal rank over all topics. By this definition, MRR is arranged the studied systems according to their first correct answer. MRR of SemApp exceeds those of C_QL, S_QL, C_BM25, S_BM25, C_TFIDF, and S_TFIDF by 23.9%, 25.4%, 32.8%, 38.3%, 96.5%, and 109% respectively. So, SemApp is the fastest system to find its first correct answer with considerable improvements with the other benchmarks.

4.2 Evaluation of P@x Metric

This experiment calculates the accuracy of each studied system at standard predefined steps of retrieved documents. Systems accuracy is measured after retrieve 5, 10, 15, 20, 30, and 100 documents. Figure 3 shows a comparison between the seven systems at these steps. Table 3 demonstrates the improvements that SemApp achieve over its competitors. The results show that SemApp rejects more irrelevant documents in each step compared to the other benchmarks. SemApp succeeds to enhance P@5 by 36.7% to 132.1%, P@10 by 24.4% to 137.5%, P@15 by 20.6% to 134.2%, and P@20 by 19.1% to 139% over the other competitors. In fact, the first two steps are very important because the user is usually interested to browse the top ten retrieved documents. So, the systems try to enhance their accuracy especially at this step. Commonly, the precision is inversely

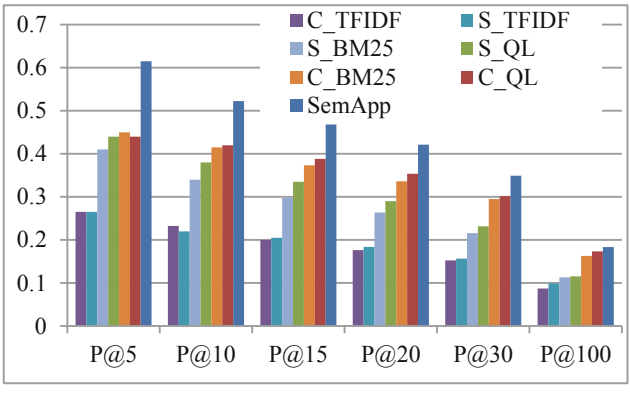

Fig. 3. Precision @ 5, 10, 15, 20, 30, and 100

proportional to the step because the system accepts more irrelevant documents in the higher steps. But, SemApp is still outperformed the other benchmarks in the rest of steps.

4.3 Evaluation of Recall-Precision Graphs

Recall-precision curves are means to compare the performance of IR systems. In this experiment, the recall-precision curve is plotted for each system. These curves are plotted by joining the average accuracy points at eleven recall steps. Figure 4 compares the seven graphs of the seven models. SemApp curve lies above the curves of the competitors which proves that SemApp is more accurate than the other benchmarks.

Table 3. SemApp improvements percentage of P@x

Improve%	P@5	P@10	P@15	P@20	P@30	P@100
C_QL	39.8%	24.4%	20.6%	19.1%	15.7%	5.6%
S_QL	39.8%	37.5%	39.8%	45.3%	50.7%	58.7%
C_TFIDF	132.1%	124.7%	134.2%	139.0%	129.0%	110.0%
S_TFIDF	132.1%	137.5%	128.4%	129.3%	122.9%	85.1%
C_BM25	36.7%	25.9%	25.4%	25.3%	18.4%	12.4%
S_BM25	50.0%	53.7%	57.0%	59.7%	61.8%	61.8%

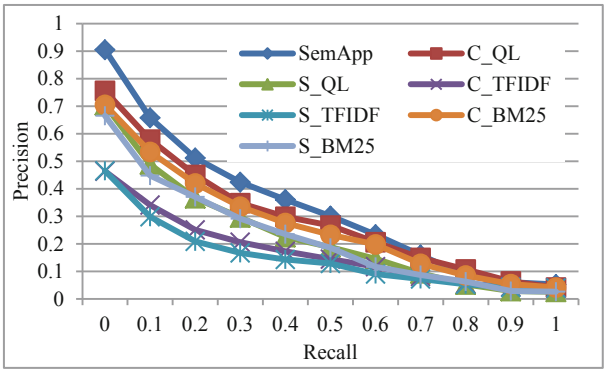

Fig. 4. Average recall-precision graph

4.4 Statistical Significance Testing

SemApp is capable of improving the most important IR metrics compared to baselines. However, these performance improvements may not be a true enhancement unless it goes through a statistical test to ensure that the difference between SemApp and the other benchmarks is not caused by random variation of the compared systems. The assumed

characteristics to conduct the statistical test, like distributions normality and absence of outliers, were validated before performing the test. The null hypothesis assumes that there is no difference between SemApp and the benchmark to be tested. If the probability of the null hypothesis is less than the specified significance level, the alternative hypothesis is accepted which supports the superiority of the proposed approach.

Table 4 shows the metrics that have been subjected to paired two sample t-test: average precision (AP), Rprec, reciprocal rank (RR), P@5, P@10, P@15, P@20, and P@30 under 39 degrees of freedom. The listed 48 computed student t-statistics are compared to the t-critical one/two tail cases (2.429/2.712 in case of significance level 0.01 and 1.685/2.023 in case of significance level 0.05).

Table 4. Significance testing of SemApp metrics improvements

	C_QL	S_QL	C_TFIDF	S_TFIDF	C_BM25	S_BM25
t Stat of AP	3.038	3.646	5.528	5.354	3.873	3.968
t Stat of Rprec	1.533	3.945	5.500	5.277	2.537	3.407
t Stat of RR	2.916	2.333	5.650	6.684	3.793	3.181
t Stat of P@5	4.376	3.484	5.982	5.730	4.444	3.776
t Stat of P@10	3.295	3.439	5.378	5.381	5.047	4.149
t Stat of P@15	3.057	3.628	5.733	5.218	4.509	4.505
t Stat of P@20	3.004	3.787	5.683	5.333	4.260	4.646
t Stat of P@30	3.109	3.466	5.291	5.145	3.367	4.115

The first row of the table shows the computed t-statistic based on values of SemApp AP and values of each benchmark individually. All these values exceeds the threshold 2.712 which means that the AP improvements that achieved by SemApp over the other benchmarks are highly significant at significance level 0.01. The entries in the second row show some variations. First, the difference between SemApp and C_QL in Rprec metric is not statistically significant which means that C_QL and SemApp almost have the same performance in this metric. Secondly, the difference between SemApp and C_BM25 is statistically significant at significance level 0.05. The same result was happened between SemApp and S_QL in RR metric. To sum up: there are 45 out of 48 computed t-statistics show high statistically significant differences between SemApp and its competitors and 2 out of the 48 test show significant differences and only one test shows not statistically difference. These results lead to the superiority of the proposed approach over the other baselines.

5 Conclusions

The work presented in this paper is based on three integrated phases to enhance IR. First phase is devoted to query meaning extraction. Discovering the global meaning of the query is necessary to explore the true user's needs. A set of semantic trees corresponding

to the possible query alternatives was extracted. After scoring each alternative (semantic tree), the optimal alternative was selected as the best query representative. An effective document and query representations were conducted in the second phase. This indexing phase utilize a conceptual weighting scheme to distribute importance effectively on the document concepts. In phase three, a hybrid ranking model was used to rank the documents. This model is based on a query likelihood language model and a conceptual weighting model.

The validity of the proposed approach has been tested through empirical experiments. The conducted evaluations assessed the standard IRS metrics to the proposed integrated approach and also compared them to the metrics of related benchmarks. The results show that the proposed integrated system is outperformed the other baselines. The proposed approach improves MAP by 15.5%, GMAP by 46%, Rprec by 6.6%, MRR by 23.9%, P@5 by 36.7%, P@10 by 24.4, P@15 by 20.6%, and P@20 by 19.1% over the nearest competitor in each metric. Statistical tests were performed to ensure that the obtained metric improvements are statistically significant and they are not caused by either random variations or coincidence.

The main limitation of this model is that the terms in the ontologies are limited and that the conceptual relations that connect the different concepts are not complete. To improve the performance of the proposed system, it may supported by more external semantic resources. Also, the formulae of weighting and ranking may be enhanced.

References

1. Nidelkou, E., Papastathis, V., Papadogiorgaki, M.: User Profile Modeling and Learning. in Encyclopedia of Information Science and Technology, 2nd edn, pp. 3934–3939. IGI Global, Hershey (2009)
2. Yibing, S., Qinglong, M.: Research of literature information retrieval method based on ontology. In: 2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI), pp. 1–6 (2014)
3. Kumar, S., Singh, M., De, A.: OWL-based ontology indexing and retrieving algorithms for semantic search engine. In: 2012 7th International Conference on Computing and Convergence Technology (ICCCT), pp. 1135–1140 (2012)
4. Chauhan, R., Goudar, R., Sharma, R., Chauhan, A.: Domain ontology based semantic search for efficient information retrieval through automatic query expansion. In: 2013 International Conference on Intelligent Systems and Signal Processing (ISSP), pp. 397–402 (2013)
5. Ali, A., Bari, P., Ahmad, I.: Concept-based information retrieval approaches on the web: a brief survey. IJAIR **3**, 14–18 (2011)
6. Zhou, D., Lawless, S., Liu, J., Zhang, S., Xu, Y.: Query expansion for personalized cross-language information retrieval. In: 2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), pp. 1–5 (2015)
7. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**, 1–50 (2012)
8. Segura, A., Vidal-Castro, C., Ferreira-Satler, M., Salvador-Sánchez: Domain ontology-based query expansion: relationships types-centered analysis using gene ontology. In: Castillo, L., Cristancho, M., Isaza, G., Pinzón, A., Rodríguez, J. (eds.) Advances in Computational Biology. AISC, vol. 232, pp. 183–188. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-01568-2_27

9. Laatar, R., Aloulou, C., Belguith, L.H.: Disambiguating Arabic words according to their historical appearance in the document based on recurrent neural networks. ACM Trans. Asian Low-Resource Lang. Inf. Process, **19**, 1–16 (2020)
10. Boughareb, D., Farah, N.: A query expansion approach using the context of the search. In: van Berlo, A., Hallenborg, K., Rodríguez, J., Tapia, D., Novais, P. (eds.) Ambient Intelligence - Software and Applications. AISC, vol. 219, pp. 57–63. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-319-00566-9_8
11. Kumar, N., Carterette, B.: Time based feedback and query expansion for twitter search. In: Serdyukov, P., et al. (eds.) ECIR 2013. LNCS, vol. 7814, pp. 734–737. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36973-5_72
12. Nagaraj, R., Thiagarasu, V., Vijayakumar, P.: A novel semantic level text classification by combining NLP and thesaurus concepts. IOSR J. Comput. Eng. **16**, 14–26 (2014)
13. Luo, Q., Chen, E., Xiong, H.: A semantic term weighting scheme for text categorization. Expert Syst. Appl. **38**, 12708–12716 (2011)
14. Boubekeur, F., Azzoug, W., Chiout, S., Boughanem, M.: Indexation sémantique de documents textuels. In: 14e Colloque International sur le Document Electronique (CIDE14), Rabat, Maroc (2011)
15. Boughanem, M., Mallak, I., Prade, H.: A new factor for computing the relevance of a document to a query. In: International Conference on Fuzzy Systems, pp. 1–6 (2010)
16. Boubekeur, F., Boughanem, M., Tamine, L., Daoud, M.: Using WordNet for concept-based document indexing in information retrieval. In: Fourth International Conference on Semantic Processing (SEMAPRO), Florence, Italy (2010)
17. Al-Zoghby, A.: A new semantic distance measure for the VSM-based information retrieval systems. In: Shaalan, K., Hassanien, A.E., Tolba, F. (eds.) Intelligent Natural Language Processing: Trends and Applications. SCI, vol. 740, pp. 229–250. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-67056-0_12
18. Luo, C., Liu, Y., Zhang, M., Ma, S.: Query ambiguity identification based on user behavior information. In: Jaafar, A., et al. (eds.) AIRS 2014. LNCS, vol. 8870, pp. 36–47. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-12844-3_4
19. Hawalah, A., Fasli, M.: Dynamic user profiles for web personalisation. Expert Syst. Appl. **42**, 2547–2569 (2015)
20. Safi, H., Jaoua, M., Belguith, L.H.: Intégration du profil utilisateur basé sur les ontologies dans la reformulation des requêtes Arabes. In: ACTES DU COLLOQUE, p. 40 (2015)
21. Codocedo, V., Lykourentzou, I., Napoli, A.: A semantic approach to concept lattice-based information retrieval. Ann. Math. Artif. Intell. **72**(1–2), 169–195 (2014). https://doi.org/10.1007/s10472-014-9403-0
22. Formica, A.: Semantic web search based on rough sets and fuzzy formal concept analysis. Knowl.-Based Syst. **26**, 40–47 (2012)
23. Poelmans, J., Ignatov, D.I., Kuznetsov, S.O., Dedene, G.: Fuzzy and rough formal concept analysis: a survey. Int. J. Gen. Syst. **43**, 105–134 (2014)
24. Kumar, C.A., Mouliswaran, S.C., Amriteya, P., Arun, S.R.: Fuzzy formal concept analysis approach for information retrieval. In: Ravi, V., Panigrahi, B., Das, S., Suganthan, P. (eds.) FANCCO - 2015. AISC, vol. 415, pp. 255–271. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-27212-2_20
25. Raj, T.F.M., Ravichandran, K.S.: A novel approach for intelligent information retrieval in semantic web using ontology. World Appl. Sci. J. **29**, 149–154 (2014)
26. Martín, A., León, C., López, A.: Enhancing semantic interoperability in digital library by applying intelligent techniques. In: 2015 SAI Intelligent Systems Conference (IntelliSys), pp. 904–911 (2015)

27. Damaševičius, R.: Automatic generation of concept taxonomies from web search data using support vector machine. In: Proceedings of 5th International Conference on Web Information Systems and Technologies, pp. 673–680 (2009)
28. Damaševičius, R.: Automatic generation of part-whole hierarchies for domain ontologies using web search data. In: 32nd International Convention Proceedings: Computers in Technical Systems and Intelligent Systems, vol. 3, pp. 215–220 (2009)
29. Tu, X., He, T., Chen, L., Luo, J., Zhang, M.: Wikipedia-based semantic smoothing for the language modeling approach to information retrieval. In: Gurrin, C., et al. (eds.) ECIR 2010. LNCS, vol. 5993, pp. 370–381. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12275-0_33
30. Chenaina, T., Neji, S., Shoeb, A.: Query sense discovery approach to realize the user search intent. Int. J. Inf. Retr. Res. **12** (2022)
31. Neji, S., Jemni Ben Ayed, L., Chenaina, T., Shoeb, A.: A novel conceptual weighting model for semantic information retrieval. 07-Inf. Sci. Lett. **10**, (2021)
32. Neji, S., Chenaina, T., Shoeb, A., Ben Ayed, L.: HyRa: an effective hybrid ranking model. In: 25th International Conference on Knowledge Based and Intelligent information and Engineering Systems, vol. 10 (2021)
33. Neji, S., Chenaina, T., Shoeb, A., Ben Ayed, L.: HIR: a hybrid IR ranking model. In: International Computer Software and Applications Conference (2021)
34. Billami, M.B., Gala, N.: Approches d'analyse distributionnelle pour améliorer la désambiguïsation sémantique. Journées internationales d'Analyse statistique des Données Textuelles (JADT) (2016)