Web-based Semantic and Personalized Information Retrieval

Khaled M. Fouad 1, Ahmed R. Khalifa 2, Nagdy M. Nagdy 3 and Hany M. Harb 4

**1 Computer Science Dept., Community College,
Taif Univ., Kingdom of Saudi Arabia (KSA)**

**2, 3, 4 Computers and Systems Engineering Dept.,
Faculty of Eng., AlAzhar Univ., Egypt.**

Abstract

Effective retrieval of the most relevant documents on the topic of interest from the Web is difficult due to the large amount of information in all types of formats. Studies have been conducted on ways to improve the efficiency of information retrieval (IR) systems. To arrive to suitable solutions in IR systems, machines need additional semantic information that helps in understanding Web documents.

In this paper, the semantic IR model is investigated, oriented to the exploitation of domain ontology and WordNet to support semantic IR capabilities in Web documents, stressing on the use of ontologies in the semantic-based perspective. The system; called SPIRS, that uses Semantic Web and agent to support more expressive queries and more accurate results is proposed. The examination of the proposed system is performed by an experiment that is based on relevance based evaluation and user satisfaction based evaluation. The results of the experiment shows that the proposed system, which is based on Semantic Web and agent, can improve the accuracy and effectiveness for retrieving relevant Web documents in specific domain.

Keywords: Semantic Web; Ontology; Semantic Information Retrieval; Personalized Retrieval.

1. Introduction

Information retrieval (IR) [1] is the technology for providing the required content based on the request from the user. It involves the searching of the content based on the keywords, with assistance from the metadata. To facilitate the retrieval, the documents are clustered based on some commonalities. Identification of these commonalities is quite involved.

Current information retrieval [2] techniques are unable to exploit the semantic knowledge within documents and hence cannot give precise answers to precise questions.

Artificial intelligence technologies have been widely applied in retrieval systems. Exploiting knowledge more efficiently is a major research field. In addition, user oriented value added systems require intelligent processing and machine learning in many forms [3]. Using Semantic Web [4] aims at enhancing the ability of both people and software agents to find documents, information and answers to queries on the Web. This new Web paradigm is to insert some level of knowledge into Web resources so that software agents can be able to intelligently process Web contents [5].

The objective of this paper is to collect domain relevant documents from Web by using search and crawler agent based on domain ontology. Our model aims at representing extracted text in terms of the synsets in the WordNet [6]. Because of the clustering can increase the efficiency and the effectiveness of the retrieval, we use the documents clustering algorithm [7] to group the similar documents to form a coherent cluster. The semantic query expansion technique using WordNet [6] and ConceptNet [8] for the documents searching and retrieving will be proposed and implemented. This technique should be able to convert a user demand into set of discrete concepts.

Finally, document similarity is computed by associating semantically similar terms in the documents and in the queries respectively by calculating the semantic similarity [9] between the clusters labels and the expanded concepts of the query terms. The semantic similarity approach is based on WordNet [9]. In proposed system, the user model [10] is acquired by analyzing the user behavior in the system to record user profile that is based on user interests [11]. Then, the acquired user model is used to re-rank the retrieved documents that match the user interests.

2. Related Work

Information retrieval [1, 12] accesses the information as well as its representation, storage and organization. The fundamental issues regarding the effectiveness of information gathering from the Web are the mismatch and it is discussed in [13]. The traditional term weighting methods measure the importance of the text in the document [14]. The keyword-based searches suffer from several inadequacies such as it can miss many highly related pages. Authors in [15] argued that the clustering quality depends on the similarity measure and it has ability to discover the hidden patterns.

Shamsfard, Nematzadeh, and Motiee in [16] have used semantics to improve search results. The relation based search engine “Ontolook” makes use of core ontologies for

Semantic Web [17]. The term reweighting approaches based on ontology are used in information retrieval applications [18]. To improve the recognition of important indexing terms, it is possible to weight the concepts of a document in different ways [19]. Kothari and Russomanno in [20] developed the OWL enhanced prototype using the Web Ontology Language and include more semantic relations.

Thomas, Redmond, and Yoon in [21] developed an expert system implementation using the ontology language OWL to express the semantics of the representations and the rule language SWRL to define the rule base for contextual reasoning. The system can be used to guide users in an e-commerce environment. Ontology based approach has been effectively used in information retrieval process in the work of [22, 23]. Iosif and Potamianos in [24] presented Web-based metrics for semantic similarity computation between words or terms. The performance of context-based term similarity metrics are evaluated with various feature weighting schemes. Zhang and Wang in [25] showed that the ontology-based clustering algorithm with feature weights do a better job in getting domain knowledge and a more accurate result.

It was proposed in [26] using WordNet for document expansion, proposing a new method: given a full document, a random walk algorithm over the WordNet graph ranks concepts closely related to the words in the document.

The work in [27] aimed at studying the use of the WordNet expansion technique over a collection with minimal textual information.

The method proposed in [28, 29] focused on semantic based expansion. There are three important improvements in the query expansion.

Ontology-based similarity measure [30] has some advantages over other measures. First, ontology is manually created by human beings for a domain and thus more precise. Second, compared to other methods such as latent semantic indexing, it is much more computational efficient. Third, it helps integrate domain knowledge into the data mining process. Comparing two terms in a document using ontology information usually exploits the fact that their corresponding concepts within ontology usually have properties in the form of attributes, level of generality or specificity, and their relationships with other concepts [31].

3. The Proposed Framework

Ontologies [32] play an important role in providing a controlled vocabulary of concepts, each with an explicitly defined and machine understandable semantics. They are largely used in the next generation of the Semantic Web which focuses on supporting a better cooperation between humans and machines.

In this work, the system SPIRS is proposed. We have used ontology based focused crawling agent to collect Web pages (documents) in a medical domain from Web. Due to the huge number of retrieved documents, we require an automatic mechanism rather than domain experts in order to separate out the documents that are truly relevant to our domain of interest.

The focused crawler in a domain specific search engine must crawl through the domain specific Web pages in the World Wide Web. For a crawler, it is not an easy task to download the domain specific Web pages. Ontology can play a vital role in this context.

In the proposed system, the most widely accepted document representation model in text classification is probably vector space model.

This representation of the documents mainly resulted to inaccuracies of the user query results due to the ambiguity, expressionless of the single words. Ontology-based information retrieval approaches promise to increase the quality of responses since they aim at capturing some part of the semantics of documents. In document representation, known as semantic indexing, the key issue is to identify appropriate concepts that describe and characterize the document content using WordNet [6].

Within the information retrieval, clustering of documents has several promising applications, all concerned with improving efficiency and effectiveness of the retrieval process. Text document clustering groups similar documents to form a coherent cluster, while documents that are different have separated apart into different clusters. The fact that the users query is not matched against each document separately, but against each cluster can lead to an increase in the effectiveness, as well as the efficiency, by returning more relevant and less non relevant documents.

Query expansion (QE) is the process of adding more terms to an original query in an attempt to refine the information search and improve retrieval effectiveness. We use the query expansion in the proposed system to improve results by including terms that would lead to retrieving more relevant documents. Our proposed technique expands the user query lexically as well as semantically. Lexically the query was expanded by using WordNet [6], while the semantic based query expansion is done by using ConceptNet [33].

We present a critical semantic similarity [9] approach for computing the semantic similarity between the terms in the query and expanded words and the labels of each clusters using WordNet. We also propose the semantic retrieval approach to discover semantically similar terms in documents and query terms using WordNet by associating such terms using semantic similarity methods. In the proposed system, we re-rank the search results based on user model to get personalized search results.

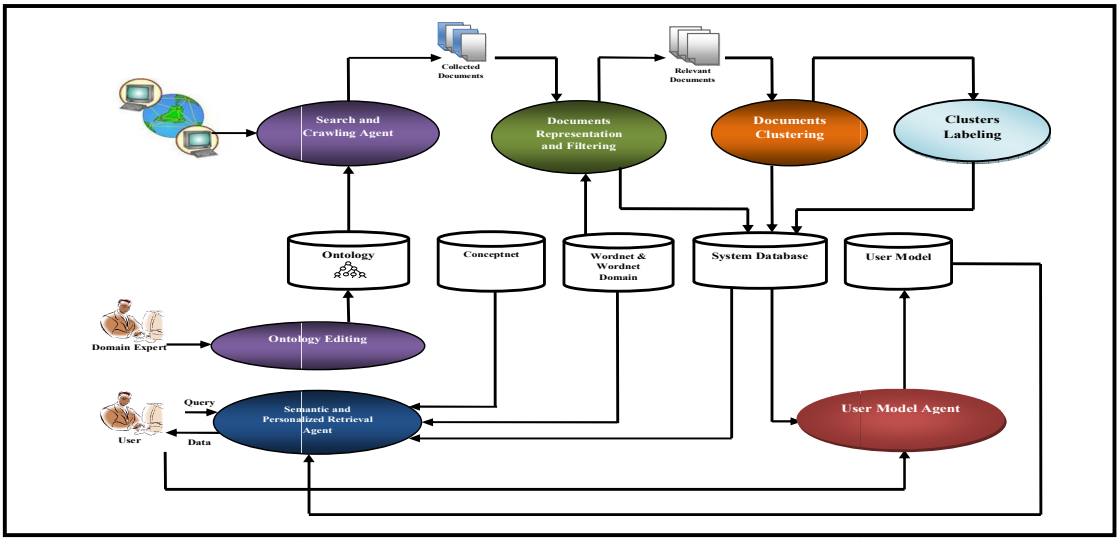

Figure 1: Architecture of the proposed System

The user model is built and updated by analyzing the user behavior during the user browsing the Web documents and inserting the query in the SPIRS system. Our user model is based on the user interests. Figure 1 shows the architecture of the proposed system.

3.1 The Domain Ontology

Ontologies [34] are designed for being used in applications that need to process the content of information, as well as, to reason about it, instead of just presenting information to humans. They permit greater machine interpretability of content than that supported by XML, and OWL, by providing additional vocabulary along with a formal semantics. Figure 2 shows main steps of the ontology development process [35]. Identifying the purpose and the requirement specification concerns to clearly identify the ontology purpose, scope and its intended use, which is the competence of the ontology. Ontology acquisition is to capture the domain concepts based on the ontology competence. The relevant domain entities (e.g. concepts, relations, slots, and role) should be identified and organized into hierarchy structure. This phase involves three steps as follows: first, enumerate important concepts and terms in this domain; second, define concepts, properties and relations of concepts, and organize them into hierarchy structure; third, consider reusing existing ontology. Ontology implementation aims at explicitly representing the conceptualization captured in a formal language. Evaluation/Check means that the ontology must be evaluated to check whether it satisfies the specification requirements. Documentation means that all the ontology development must be documented, including purposes, requirements, textual descriptions of the conceptualization, and the formal ontology [35]. The ontology in the proposed system is focused in the medical domain that is "Jaundice diseases" as found in [36].

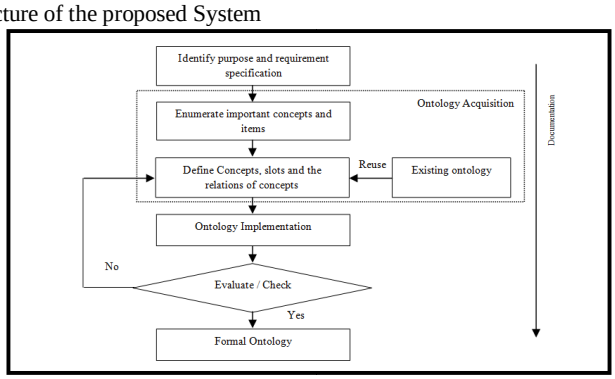

Figure 2: Main steps of the ontology development

Figure 3 shows part of our medical ontology for "Jaundice diseases".

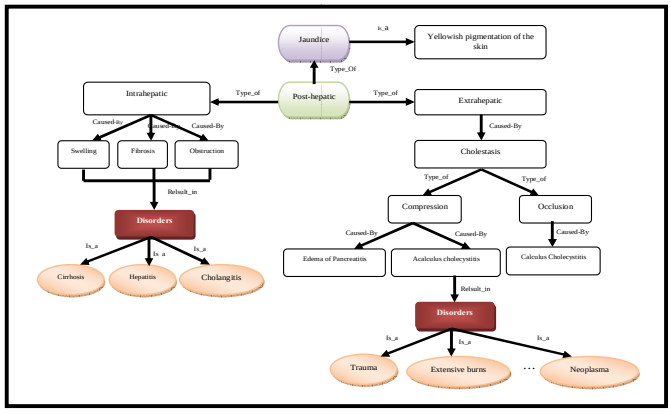

Figure 3: Part of the domain ontology

3.2 Search and Crawling Agent

This agent is a Web crawler (spider) which searches the Web automatically. The agent takes the concepts from the domain ontology and gets the seed URL from the search results URL. The spider then identifies all the hyperlinks in the first page which necessitates other URLs to be crawled again. The spider updates its URL list when

identifying a new URL to crawl it and so on. It passes the page title, address, and ontology concepts to the system database.

Although documents are retrieved selectively through restricted queries and by focused crawling, we still need a mechanism to evaluate and verify the relevance of these documents to the predefined domain of Jaundice domain. To remove unexpected documents, first the agent automatically removes those that are blank, too short, duplicated documents, or those that are in a format that is not suitable for text processing. It then performs the relevance calculation to extract the relevant documents and discard the irrelevant document to our domain. In relevance calculation [32], the relevancy of a Web document on a specific domain is calculated. Relevance calculation algorithm calculates the relevance score of a Web page as shown in figure 4.

3.3 Documents Representation

In proposed system, we used the vector space model (VSM) [32, 37] to represent the documents. In VSM, a document j is represented by the document vector d_j:

$d_j = (w_{1j}, w_{2j}, \ldots . w_{kj}, w_{nj})$ where, w_{kj} is the weight of the k_{th} term in the document j.

Input: A Web page (P), a weight table.
Output: The relevance score of the Web page (P).
Procedure:
Get a Web page (P).
Read Ontology (O) (C: Concept in the ontology, and C_c: Concept count in ontology).
Do While $j <= C_c$ //Calculate and save weight table
Open Weight Table (WT) of the domain concepts that contains concepts and
its weights.
Calculate Concept Weight (CW_j).
Save CW_j in DB table.
EndDo
Read WT.
Get (T_c : Concept Count in weight table).
Do While $i <= T_C$ //Calculate Relevance Score for the page
Reset Relevance Score (RS_P) of the Web page (P) (RS_P=0).
Get (T_i) and its (W_i).
Calculate (TF_P :Term Frequency in P).
Calculate $TW_P = TF_P \times W_i$. (TW_P: Term Weight in P).
Calculate $RS_P = RS_P + TW_P$.
Output RS_P (for the Web page).
Save RS_P and its Content in DB table.
EndDo

Figure 4: Algorithm of calculation of relevance score for the Web pages

This has several limitations, including:

1. Different vector positions may be allocated to the synonyms of the same term; this way, there is an information loss because the importance of a determinate concept is distributed among different vector components,
2. The size of a document vector must be at least equal to the total number of words of the language used to write the document, and
3. Every time a new set of terms is introduced (which is a high-probability event), all document vectors must be reconstructed; the size of a repository thus grows not only as a function of the number of documents that it contains, but also of the size of the representation vectors.

To overcome these weaknesses of term-based representations, an ontology-based representation using WordNet [6] is performed. Moreover, by defining an ontology base, which is a set of independent concepts that covers the whole ontology, an ontology-based representation allows the system to use fixed-size document vectors, consisting of one component per base concept. We used the method, ontology-based representation that is based on WordNet, which improves traditional vector space model (VSM). We used WordNet to identify WordNet concepts that correspond to document words. This representation requires two more stages:

a) The "mapping" of terms into concepts and the choice of the "merging" strategy, and

b) The Word Sense Disambiguation (WSD) strategy.

Concept identification [38] is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry. The entry which maximizes the overlap is selected as a possible sense of the analyzed word. The concept identification algorithm [39] is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry.

Document representation method in our proposed system requires two steps:

(1) Mapping terms into synsets, and

(2) Capturing relationships between synsets.

Before we start to perform the text representation, we must prepare the text by performing Part of Speech (PoS) [40], Stop words Removal, and Words Stemming. Figure 5 shows the approach for representation of the retrieved documents of web pages.

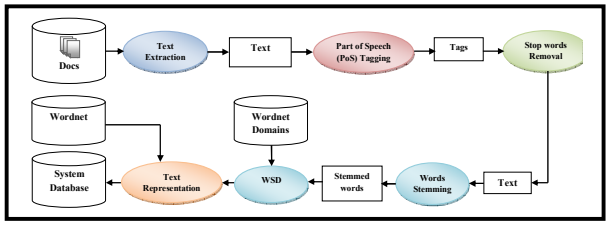

Figure 5: Document Representation

- *WSD using WordNet Domains*

Lee and Mit in [41] have discussed using the knowledge of domain provided by WordNet to Word Sense Disambiguation (WSD). Each synset in WordNet would be annotated by one or more domain labels. Therefore instead of only assigning the score to every occurrence of the domain, Cliozzo, Magnini and Strapparava in [42] suggested to assign the domain relevance (DR) to every domain that annotated to the synsets. The DR is a measure to weigh the score of the domain according to the number of the domain types

annotated to the word. The WSD algorithm described in our work has a basic idea underlying that work is that the disambiguation of a word in its context is mainly a process of comparison between the domain of the context and the domains of the word's senses. The WSD algorithm that is based on WordNet domains [43] is shown in figure 6.

Input: Terms in the text sentences (S)
Output: The sense belonging to obtained domain is the correct sense.
Procedure:
Get sentence (S).
Perform PoS tagging to extract the tagged sentence (S_t).
Separate word in (S) according to (S_t).
Insert words into bag (B_1) that has count (C_w).
Do While $i <= C_w$
  Insert set of domains corresponding to its PoS tag into bag (B_2)
EndDo
Get target word (W).
Insert target word (W) domains corresponding to its PoS tag into bag (B_3).
Compare each domain in (B_3) with set of domains of remaining content words.
If domain in (B_3) > domains of other content **Then**
  Set content words are the domain of the text.
  Set the sense belonging to domain obtained is the correct sense.
Endif

Figure 6: WSD algorithm using WordNet Domains

- *Mapping Words into Synsets*

The purpose of this step is to identify WordNet concepts that correspond to document words. Concept identification [38] is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry. The concept identification algorithm for the terms is given in figure 7.

Input: (B_w) Bag of words (W_i) in document D that was gotten from Words Stemming phase.
Output: Set of all WordNet concepts belonging to terms (words) in document D.
Procedure:
// (C_w) is the count of words in the bag, and ($Cont_i$) the context of the word in the document, it is the sentence in document D that contains the word occurrence being analyzed.
Do While $i <= C_w$
  Get WordNet entries C_i set ($CSet_i$) that is containing the word W_i , where $C_i \in CSet_i$.
  Save W_i and its C_i in database table.
EndDo
Rank concepts C_i in $CSet_i$ where $|C_1| > |C_2| > |C_3| \ldots > |C_n|$ *// | | denotes the concept length, in terms of the number of words in the corresponding terms. $CSet_{ri}$ is the ranked concepts set.*
FOR each C_i in $CSet_{ri}$
  Get common words between $Cont_i$ and representative term of C_i , which is the intersection $Cint = \cap (Cont_i, C_i)$.
  If $|Cint| < |C_i|$ then
    The concept-sense C_i is not within the context $Cont_i$.
  EndIf
  If $|Cint| = |C_i|$ then
    The concept-sense C_i is within the context $Cont_i$.
    Add C_i to the set of possible senses associated with the document.
  EndIf
EndFor

Figure 7: The concept identification of words algorithm

- *The Weight of Concept Computation*

The concepts in documents are identified as a set of terms that have been identified or synonym relationships [44], i.e., synsets in the WordNet ontology. Then, the concept frequencies Cf_c are calculated based on term frequency tf_t as follows:

$$Cf_c = \sum_{t_m \in r(c)} tf_{tm} \quad (1)$$

where r(c) is the set of different terms that belongs to concept C. Note that WordNet returns an ordered list of synsets based on a term. The ordering is supposed to reflect how common it is that the term is related to the concept in Standard English language. More common term meanings are listed before less common ones. Hypernyms of concepts can represent such concepts up to a certain level of generality. For example, the concept corresponding to 'hepatitis_C' can represent the concept corresponding to 'viral_hepatitis'. The concept frequencies are updated as follows:

$$hf_c = \sum_{b \in H(c,r)} Cf_b \quad (2)$$

where $H(c, r)$ is the set of concepts C_H, which are all the concepts within *r* levels of hypernym concepts of *c*.

In WordNet, $H(c, r)$ is obtained by gathering all the synsets that are hypernym concepts of synset *c* within *r* levels. In particular, $H(c, \infty)$ returns all the hypernym concepts of *c* and $H(c, 0)$ returns just *c*. The weight of each concept *c* in document *d* is computed as follows:

$$wh_c = hf_c \times idf_c \quad (3)$$

Where idf_c is the inverted document frequency of concept *c* by counting how many documents in which concept *c* appears as the weight of each term *t* in the document *d*.

3.4 Clustering of Documents

The fuzzy c-means (FCM) seems to be the most popular algorithm in the field of fuzzy clustering [45]. FCM is an iterative algorithm. The aim of FCM is to find cluster centers (centroids) that minimize a dissimilarity function. The algorithm minimizes a dissimilarity (or distance) function

The algorithm of FCM is shown in figure 8. By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the "right" location within a data set.

Input:
  x : Data Set for n where X={$x_1, x_2, \ldots x_n$}.
  c : Number of clusters.
  t : Convergence threshold (termination criterion).
  m : Exponential weight.
Output: U : Membership Matrix
Procedure:
  Randomly initialize matrix U with c clusters that fulfils.
Repeat
  Calculate C_i .
  Compute dissimilarity (distance) between centroids and data points.
  Compute a new U.
Until the improvement over previous iteration is below *t*.

Figure 8: The FCM algorithm

3.5 Labeling of Clusters

The proposed algorithm learns the weights for each feature in order to pick out the good labels from the rest of the data. It is not merely a case of choosing the most common feature or the most frequently occurring feature

word. Our solution is to assign variable weights to features, which reflect their relative importance with respect to the likelihood of containing appropriate cluster labels. We refer to these weights as feature scores [46].

3.6 User Model Agent

This agent aims at building the user model using user behavior in the system and updating the user model using user query. There are roughly two kinds of automatic way to capture a user's interest implicitly: behavior-based and history-based. The behavior-based research proves that the time spent on a page, the amount of scrolling on a page and the combination of them has a strong positive relationship with user interests. Browsing histories capture the relationship between user's interests and his click history in which sufficient contextual information is already hidden in the web log. User interests [47, 48] always constitute the most important part of the user profile in adaptive information retrieval and filtering systems that dealt with large volumes of information.

In the proposed system, user interest model's knowledge expression uses the thought, which is based on the space vector model's expression method. This method for acquiring user's interest was shown in [47, 48]. Figure 9 shows certain steps to acquire user interest.

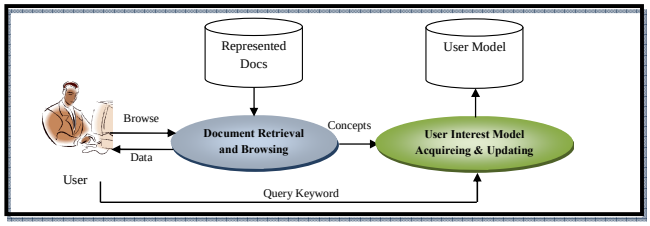

Figure 9: The steps to acquire user interest

The main purpose of this step is to acquire the interested concepts of the user in the web page (document), then get concept frequency that reflects the importance of concept, and finally get the weight of concepts in the selected page. The output of this step is the weight of concepts in the selected page that can be used to build user interest model. During the user is working through proposed system, user interests often change quite, and users are reluctant to specify all adjustments and modifications of their intents and interests. Therefore, techniques that leverage implicit approaches for gathering information about users are highly desired to update the user interests that are often not fixed.

In order to update user interest [47], we should analyze user's history query keywords. For certain keyword, we extract the words which have the semantic relationships with the keyword and add them into the user interest model as nodes according to semantic relationships in WordNet.

With new words added constantly, user is always interested in the kind of the words with higher score which stands for some type of knowledge. We must constantly update the user interest model after the users enter new specific keywords. User interest model is updated by the new keywords. The incremental updating strategy is used here, and gives the related words the different score according to the relations which reflect their importance of different words in order to render the interestingness of the words. As a result, the words that are more frequent have higher score. This means that if the system wants to update the user interest, it can add the initial score value of the interest to the semantic similarity score of the query term (T_q), the synonym (T_{Syn}), hyponym (T_{Hyp}), and meronym (T_{Mer}) that can be extracted from WordNet.

Because the keywords are added constantly and the scale of the user interest model becomes bigger, some old nodes must be removed in order to reduce user model. The update of user ontology method is shown in the flowchart of figure 10.

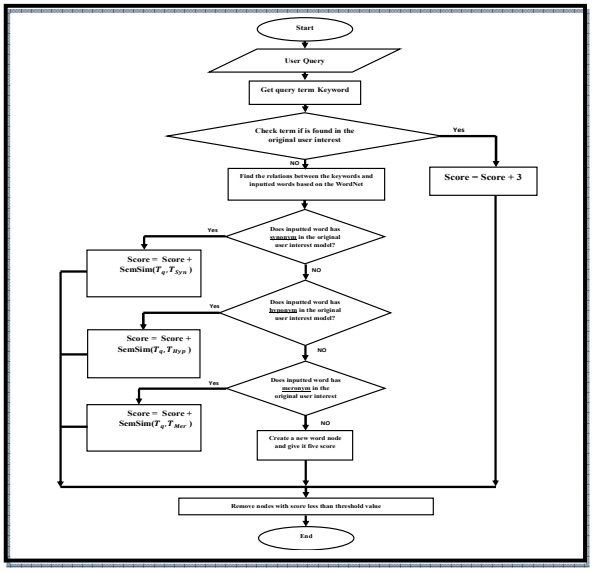

Figure 10: The steps to acquire user interest

3.7 Semantic and Personalized Retrieval Agent

Semantic retrieval [36] plays an increasingly important role in information retrieval. It overthrew the shackles of traditional idea of information retrieval.

Semantic matched on information considerably improves the information recall and precision ratio. Given a query, if we can get enough semantic knowledge, acquire semantic similarity of the known query and optional data, then we get a result set which is sorted according to semantic similarity. Nowadays semantic retrieval mainly implements concept retrieval by interaction terms, which does not take the concept's attributes and other information into consideration. The proposed system applies the query

expansion approach based on WordNet and ConceptNet [33]. Query expansion [49] has been a well-known and popular technique to improve performance of typical information retrieval systems. The effectiveness of query expansion comes from the fact that users' queries (especially short queries) usually cannot describe their information needs clearly, and on the other hand, sometimes the vocabulary in a query is inconsistent with that found in relevant documents.

Figure 11 shows the block diagram of semantic and personalized information retrieval agent of Web documents using semantic similarity between query and documents data and query expansion of the query.

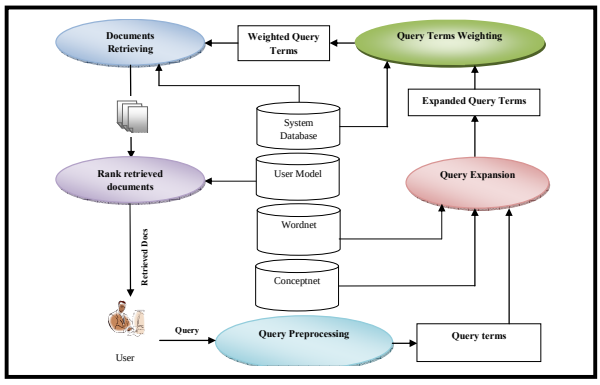

Figure 11: Semantic and personalized retrieval agent

- *Query Preprocessing*

Queries are in the natural language form. The first step of the proposed query expansion approach is concerned with detecting meaningful keywords in the query. To this purpose, in this step query terms that has stop words are also must be removed from query, then stemming the query terms, Parts-of-Speech (PoS) Tagging, and each query term is disambiguated through assigning appropriate WordNet domains.

- *Query Expansion (QE)*

QE refers to the process of adding new necessary terms to a user's initial query. The purpose of QE aims at improving retrieval performance. QE reformulates the original query that enables users' desired information to be retrieved. The major process of query expansion is the modification of the original query with new relevant and meaningful terms. The main aim of query expansion (also known as query augmentation) is to add new meaningful terms to the initial query.

To optimize the performance of proposed system, we propose the novel approach for our information retrieval system where the query is expanded lexically and commonsensical by using knowledge bases. The query expansion algorithm is shown in figure 12.

Input: Original query terms set (Q_{or}) where $Q_{or} = \{t_1, t_2, \ldots t_n\}$
Output: Query terms set (Q_{set}) where $Q_{set} = Q_{or} \cup Q_{ex}$, Q_{ex} is the expanded query terms
Procedure:

1. Get Q_{or} and add it to Q_{set}.
2. Do While $i <= n$ *//n is the number of terms t_i in Q_{or}*
   a. Search t_i in the WordNet
   b. If t_i is the root
      i. Get the Hyponym, and Get the Synonym.
      ii. Add the two concepts to the query set Q_{ex}, Then GOTO Step 4
   c. ElseIf t_i is not the root and is not the Leaf
      i. Get the Hyponyms, Hypernyms, and Synonym.
      ii. Add the three concepts to the query set Q_{ex}, Then GOTO Step 4
   d. ElseIf t_i is not the root and is the Leaf
      i. Get Hypernyms, and Synonym.
      ii. Add the two concepts to the query set Q_{ex}, Then GOTO Step 4
   e. Endif
3. EndDO

//Compute the Semantic Similarity (SemSim) between concepts t_j in Q_{ex} and concept t_i in Q_{or}, put pre-expansion concepts that has high relativity as expansion words.

4. Do While $i <= n$ *//n is the number of terms t_i in Q_{or}*
   a. DO While j <= m *//m is the number of terms t_j in Q_{ex}*
      If SemSim (t_i , t_j) < Sim_{th} *//Sim_{th} is the semantic threshold*
      Remove t_j from Q_{ex}
      Endif
   b. EndDO
5. EndDO
6. $Q_{set} = Q_{or} \cup Q_{ex}$
7. Do While i <= n *//n is the number of terms t_i in Q_{or}*
   Search the term in the ConceptNet
   If the term is found
   Get the related concepts form ConceptNet.
   Add the concept to the query set Q_{set}
   Endif
8. EndDO

Figure 12: QE using WordNet and ConceptNet

QE is still the existing issues to effective retrieve results from the large information corpus. However, the trends now move to the semantic expansion of the user queries. The systems those are heavily relied on lexical analysis, flunks in the complex queries. It does not discover the semantic relatedness or have no possibility for common sense reasoning. Despite the facts of these, that lexical analysis plays a vital role in the extracting the meaning from the user request. The common sense reasoning also plays a main role in the user query. Common sense knowledge; that is ConceptNet, includes knowledge about the social, physical, spatial, temporal and psychological aspects of daily life. WordNet has been used ordinarily for the query expansion. It has made some modification, but it was limited. Several studies expose the importance of common sense reasoning in information retrieval, data filtering, and data mining.

- *Query Term Weighting*

The query term weighting technique is responsible for weighting each term in the query submitted to the information retrieval system, indicating the significance of each query term. This is essential so that the ranking models can use this weighting information to calculate the rank scores for the documents.

Inverse Document Frequency (IDF) [50] is a statistical scheme that determines term specificity according to the number of documents a term appears in relative to the number of documents in the collection as found in equation 4. IDF and its extensions that depend on the document collection, has become the most popular and important term significance indicator for information retrieval models.

$$IDF = \log \frac{n_i}{N} \quad (4)$$

where N is the total number of the documents in the system database, and n_i is the number of documents where query term q occurs.

- *Document Retrieval*

The document retrieval is based on semantic similarity of the query term vector and document vector using equation 5.

$$sim(q,d) = \frac{\sum_i \sum_j q_j w_i \, sim(i,j)}{\sum_i \sum_j q_j w_i} \quad (5)$$

where w_i is term weight of term *i* in the documents vector, q_j is the term weight of term *j* in the query vector, and $sim(i,j)$ is semantic similarity of the term *i* and term *j*.

Finally, we arrange the retrieved documents by using the semantic similarity score of the query term vector and document vector.

- *Ranking Retrieved Documents based on the User Model*

After building the user model; that is based on the user interest, the system uses the user model to rearrange the retrieved and ranked documents. Ranking the retrieved documents user model makes the documents appears in the order as the user interest is matched. Figure 13 shows the flowchart of ranking algorithm for the retrieved documents based on the user model. Matching between the user interest term and the documents terms is based on semantic similarity to determine the documents that the user is interested to be ranked first by semantic similarity score.

4. Implementation and Experimentation

We have implemented a framework for the SPIRS (Semantic and Personalized IR system) to test its performance, addressing mainly the web-based semantic and personalized information retrieval based on Semantic Web and agent. The proposed system depends on two phases, first is collecting the domain relevant documents, representing the documents, clustering, and labeling. Second phase aims at fetching the relevant documents from database that are semantically matched with the user query, user query expansion, acquiring and updating user model, and re-ranking the documents based on user model. The second phase of the system has the components of the user search agent and user model agent.

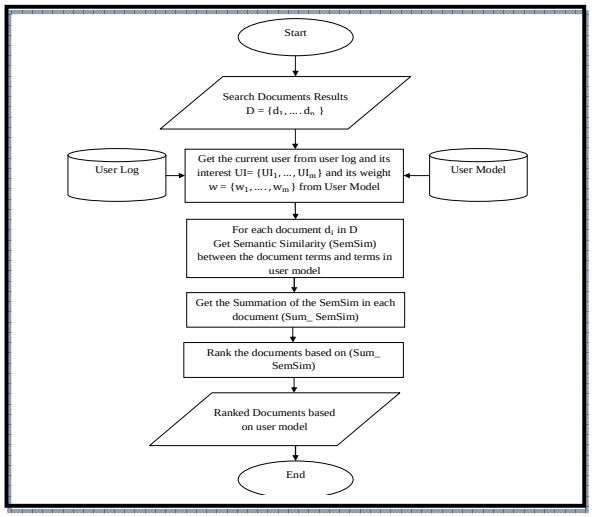

Figure 13: The documents ranking based on user model

Figure 14 shows the main user interface of user search in the system. The result of the semantic information retrieval is shown in figure 15.

Figure 14: The main user interface for user search

Figure 15: The results of user search for semantic information retrieval

In this work, we intend to improve IR performance by using semantic web and agents. In order to quantify the improvement of our approach, we need to define the experimental strategy that allows us, without any ambiguity, to evaluate our hypotheses. The improvement is measured by performing the experiment. In our experiment, we used two methods from eight methods in the evaluation studies that were discussed in [51]. The used evaluation methods are relevance based evaluation, and user satisfaction based evaluation. In relevance based evaluation method, we used precision, recall, f-measure, and mean average precision to measure the performance of proposed system. The method of user satisfaction based evaluation

aims at measuring the capabilities with considerations of the potential differences in the background of end users, such as domain experts. In this method, proposed system is judged according to the search results' ability to satisfy an easily pleased user or hard to please user. The experiment is based on evaluating number of domain experts in our medical domain by using proposed system. They gave the required measures of system performance, subjective measures of seeming satisfaction and the relevance degree of the search results to the query.

The implementation is tested on a set of 3780 documents that are collected and extracted during running search and crawling agent in the proposed system. The collected documents from Web own the same medical domain; that is jaundice disease. The documents filtration in proposed system defined that 2869 documents are relevant to our domain and 911 are irrelevant documents. These relevant documents are used during representation, clustering, labeling and calculating semantic similarity with user query. The proposed system is implemented in ASP.Net as Web-based system using Visual Studio 2010, .NET Framework 4, and SQL Server 2008.

The experiment aims at proving enhancements in the performance of retrieving the Web documents based on certain domain. In this experiment, seven domain experts are motivated to test and evaluate the system. The domain experts entered number of user query then they checked each retrieved document and defined whether each document is relevant or no. After each expert had finished him test, he filled an evaluation form. This evaluation form shows the query terms for each expert, the number of total retrieved documents, total relevant documents, and number of relevant retrieved documents during each test session.

After the domain expert has finished the results, the calculation of the recall, precision, f-measure, and average precision are performed for each query then calculation of the mean average precision MAP for each domain expert is performed.

The performance of proposed system can be examined by check the differences between the mean average precision (MAP) of each domain expert and then calculating the average of mean average precision. The average of MAP for all experts is accurate measure of performance enhancement for proposed system, because each domain experts can examine the retrieved document and consider this document as relevant but another domain expert can consider the same document as irrelevant. This difference of the domain experts because we select number of domain experts with different expertise.

Figure 16 shows the differences between MAP value for each domain expert and the average of MAP.

In this experiment, also we consider additional factors. The factors are system speed to measure the response of the request for each user, and using simplicity to measure the degree of familiarity and user effort of using the system.

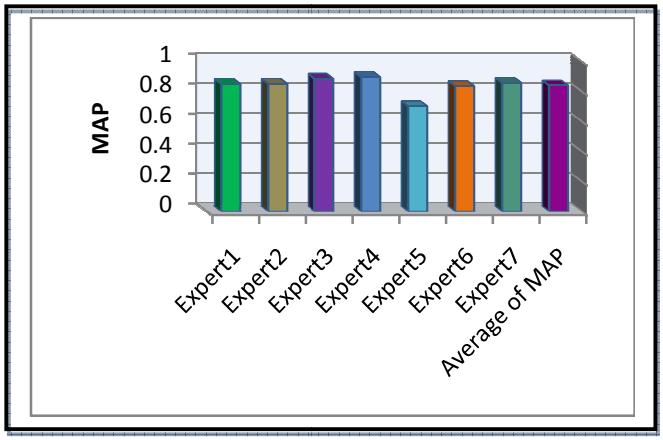

Figure 16: Differences between MAP value for each domain expert and the average of MAP

Figure 17 shows the comparison of these factors for each expert.

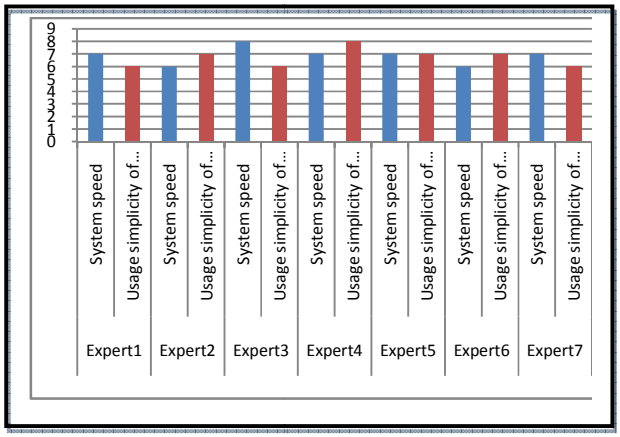

Figure 17: Comparison of extra factors for each expert

5. Conclusions

Semantic Web provides a very flexible framework for content based retrieval. Semantic web would serve as a good integration platform for content based retrieval.

The proposed semantic IR prototype system called SPIRS has been implemented. The system is designed using a highly modular approach that hides most of the complex processing tasks from users. We have conducted the evaluation based on the SPIRS search engine using Web documents that were collected by the system from Web. The proposed system experimentation showed that the SPIRS system can improve the accuracy and effectiveness for retrieving the Web documents. The reported experimental results demonstrate the proof of concept and show that proposed approaches perform as good as syntactic analogues while allowing for an improvement whenever semantics is available and can be

exploited. The proposed system is Web-based and is now online. It aims at providing the relevant Web-documents in certain domain that are matched to user request. The proposed system can be used in other domains by editing the domain ontology through the ontology editor and building the domain concepts weight table. A user model is proposed to improve the ranking of the relevant documents retrieved to user based on its interests.

As a result of evaluation of domain experts, the proposed system can improve the accuracy for retrieving the relevant Web documents. Using documents filtration, SPIRS improves the mean average precision MAP by 12.57964224%. Using semantic similarity between the documents vector and query vector increases the accuracy of documents retrieval, which is represented by MAP, by 25.25837347%. The query expansion in the proposed system improves the MAP by 13.69156%. The used user model to re-rank the retrieved documents that match the user requirements and interests increases the MAP by 3.809859%.

Acknowledgments

The authors would like to express their deepest sense of gratitude to each and every person who reviewed, helped and supported them to complete this research. The authors also would like to thank the domain experts who provided them by the domain knowledge and evaluated the system, specially, Dr. Osama Mohammed, Dr. Mohammed Elshora, Dr. Ayman Elmohamady, Dr. Thorya Abd Allah, and Dr. Mai.

References

[1] Ramachandra, M. (2010). Information Retrieval. In: Web-Based Supply Chain Management and Digital Signal Processing: Methods for Effective Information Administration and Transmission . 182-194 pp. IGI Global.

[2] Shah, U., Finin, T., Joshi, A., Mayfield, J. , & Cost, R. (2002). Information retrieval on the semantic web. The ACM Conference on Information and Knowledge Management, November 24..

[3] Mandl, T. (2009). Artificial Intelligence for Information Retrieval. In: Encyclopedia of Artificial Intelligence. 151-156 pp. IGI Global.

[4] McCuaig, J. (2011). The Semantic Web. In: Essential Software Architecture. DOI 10.1007/978-3-642-19176-3_12, Springer-Verlag Berlin Heidelberg.

[5] LUO, J. & XUE, X. (2010). Research on Information Retrieval System Based on Semantic Web and Multi-Agent. 2010 International Conference on Intelligent Computing and Cognitive Informatics. 978-0-7695-4014-6/10, IEEE.

[6] Fellbaum, C. (2010). WordNet. Theory and Applications of Ontology: Computer Applications, 231, PP: 231-243, Springer Science+Business Media B.V.

[7] Win, T. & Won, L. (2010). Document Clustering by Fuzzy C-Mean Algorithm. 978-1-4244-5848-6110, IEEE.

[8] Blanco, E., Cankaya, H. & Moldovan, D. (2011). Commonsense Knowledge Extraction Using Concepts Properties. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference.

[9] Li, H., Tian, Y., Ye, B. & Cai, Q. (2010). Comparison of Current Semantic Similarity Methods in WordNet. 20IO International Conference on Computer Application and System Modeling (ICCASM 2010). 978-1 -4244-7237-6/10, IEEE.

[10] Nidelkou, E., Papastathis, V., Papadogiorgaki, M., Kompatsiaris, I., Bratu, B., Ribiere, M. & Waddington, S. (2009). User Profile Modeling and Learning. In Encyclopedia of Information Science and Technology, Second Edition. DOI: 10.4018/978-1-60566-026-4.ch627. 3934-3939. IGI Global.

[11] Harb, H., & Fouad, K. (2010). Semantic web based Approach to learn and update Learner Profile in Adaptive E-Learning. Al-Azhar Engineering Eleventh International Conference, December 23-26.

[12] Sridevi, U. K. & Nagaveni, N. (2011). An Ontology Based Model for Document Clustering, International Journal of Intelligent Information Technologies (IJIIT) , 7 (3), 54-69, DOI: 10.4018/jiit.2011070105.

[13] Li, Y., & Zhong, N. (2008). Mining ontology for automatically acquiring web user information needs. IEEE Transactions on Knowledge and Data Engineering, 18(4), 554–568.

[14] Lan, M., Ta, C. L., Su, J., & Lu, Y. (2009). Supervised and traditional term weighting methods for automatic text categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(4), 721–735. doi:10.1109/TPAMI.2008.110.

[15] Dash, R., Mishra, D., Rath, A. K., & Acharya, M. (2010). A hybridized K-means clustering approach for high dimensional dataset. International Journal of Engineering . Science and Technology, 2(2), 59–66.

[16] Shamsfard, M., Nematzadeh, A., & Motiee, S. (2006). ORank: An ontology based system for ranking document. International Journal of Computer Science, 1, 225–231.

[17] Li, Y., Wang, Y., & Huang, X. (2007). A relation- based search engine in semantic web. IEEE Transactions on Knowledge and Data Engineering, 19(2), 273–282. doi:10.1109/TKDE.2007.18.

[18] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., & Milios, E. E. (2005). Semantic similarity methods in wordnet and their application to information retrieval on the web. In Proceedings of the 7th Annual ACM International Workshop on Web Information and Data Management (pp. 10-16).

[19] Valkeapää, O., Alm, O., & Hyvönen, E. (2007). An adaptable framework for ontology-based content creation on the semantic web. Journal of Universal Computer Science, 13(12), 1825–1853.

[20] Kothari, C. R., & Russomanno, D. J. (2008). Enhancing OWL ontologies with relation semantic. International Journal of Software Engineering and Knowledge Engineering, 18(3), 327–356. doi:10.1142/S0218194008003660.

[21] Thomas, M. A., Redmond, T. R., & Yoon, V. Y. (2009). Using ontological reasoning for an adaptive e-commerce experience. International Journal of Intelligent Information Technologies, 5(4), 41–52. doi:10.4018/jiit.2009080703.

[22] Sridevi, U. K., & Nagaveni, N. (2009a). Ontology based semantic measures in document similarity ranking. In Proceedings of the International Conference on Advances in Recent Technologies in Communication and Computing (pp. 482-486).

[23] Sridevi, U. K., & Nagaveni, N. (2009b). Ontology based correlation analysis in information retrieval. International Journal of Recent Trends in Engineering, 2(1), 134–137.

[24] Iosif, E., & Potamianos, A. (2010). Unsupervised semantic similarity computation between terms using web documents. IEEE Transactions on Knowledge and Data Engineering, 22(11), 1637–1647. doi:10.1109/TKDE.2009.193.

[25] Zhang, L., & Wang, Z. (2010). Ontology-based clustering algorithm with feature weights. Journal of Computer Information Systems, 6(9), 2959–2966.

[26] Eneko, A., Xabier, A. & Arantxa, O. (2010). Document Expansion Based on WordNet for Robust IR. COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Volume, pages 9–17, Beijing. ACM.

[27] Manuel, D., Maria, M., Alfonso, U. L., & Jose, P. (2010). Using WordNet in Multimedia Information Retrieval. CLEF 2009 Workshop, Part II, LNCS 6242, pp. 185–188, Springer-Verlag Berlin Heidelberg.
[28] Hongsheng, W., Jiuying, Q. & Hong, S. (2009). Expansion Model of Semantic Query Based on Ontology. Web Mining and Web-based Application. WMWA '09. IEEE.
[29] Shabanzadeh, M., Nematbakhsh, M.A., & Nematbakhsh, N. (2010). International Conference on Intelligent Control and Information Processing August 13-15, 2010 - Dalian, China. 978-1-4244-7050-1/10,IEEE.
[30] Zhang, X., Jing, L., Hu, X., Ng, M., Jiangxi, J. & Zhou, X. (2010). Medical Document Clustering Using Ontology-Based Term Similarity Measures. In Taniar , D. & Irina, L. Strategic Advancements in Utilizing Data Mining and Warehousing Technologies: New Concepts and Developments. DOI: 10.4018/978-1-60566-717-1.ch007, 121-132. IGI Global.
[31] Pedersen, T., Pakhomov, S., Patwardhan, S., & Chute, C. (2007). Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, 40(3), 288-299.
[32] Harb, H., Fouad, K. & Nagdy, N. (2011). Semantic Retrieval Approach for Web Documents. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 9, 2011.
[33] Hsu, M., Tsai, M. F. & Chen, H. (2008). Combining WordNet and ConceptNet for Automatic Query Expansion: A Learning Approach. AIRS 2008, LNCS 4993, pp. 213–224, Springer-Verlag Berlin Heidelberg.
[34] Aida, V., Karina, G., David, S. & Montserrat, B., (2010), Using ontologies for structuring organizational knowledge in Home Care assistance, international journal of medical informatics 79 (2010) 370–387, Elsevier Ireland Ltd.
[35] Fouad, K., Nofal, M., Harb, H., Nagdy N. (2011). Using Semantic Web to support Advanced Web-Based Environment. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, P: 120 – 129.
[36] Harb, H., Fouad, K. & Nagdy, M. (2011) Semantic Retrieval Approach for Web Documents. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 9, P: 11 – 75.
[37] Ying, L. (2009). On Document Representation and Term Weights in Text Classification. IGI Global.
[38] B. Fatiha, B. Mohand, T. Lynda, D. Mariam. (2010). Using WordNet for Concept-Based Document Indexing in Information Retrieval, SEMAPRO: The Fourth International Conference on Advances in Semantic Processing, Pages: 151 to 157, IARIA.
[39] B. Rujiang, W. Xiaoyue, L. Junhua. (2010). Extract Semantic Information from WordNet to Improve Text Classification Performance. AST/UCMA/ISA/ACN 2010, LNCS 6059, pp. 409–420. Springer-Verlag Berlin Heidelberg.
[40] The Stanford Natural Language Processing Group. http://nlp.stanford.edu/software/tagger.shtml.
[41] Lee, W. & Mit, E. (2011). Word Sense Disambiguation By Using Domain Knowledge. International Conference on Semantic Technology and Information Retrieval. 978-1-61284-353-7/11, IEEE.
[42] Cliozzo, A., Magnini , B. & Strapparava, C. (2001). Unsupervised domain relevance estimation for word sense disambiguation. SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation System, 2001, Toulose, France, in press.
[43] Lee, W. & Mit, E. (2011). Word Sense Disambiguation By Using Domain Knowledge. International Conference on Semantic Technology and Information Retrieval. 978-1-61284-353-7/11, IEEE.
[44] Z. Hai-Tao, K. Bo-Yeong, K. Hong-Gee. (2009). Exploiting noun phrases and semantic relationships for text document clustering. Information Sciences 179, 2249–2262, doi:10.1016/j.ins,.02.019. Elsevier Inc.
[45] Oikonomakou, N. & Vazirgiannis, M. (2010). A Review of Web Document Clustering Approaches. In: Data Mining and Knowledge Discovery Handbook. Part 6, Pages 931-948. DOI 10.1007/978-0-387-09823-4_48, Springer Science+Business Media, LLC.
[46] T. Yuen-Hsien. (2010). Generic title labeling for clustered documents. Expert Systems with Applications 37 (2010) 2247–2254. Elsevier Ltd.
[47] Harb, H., & Fouad, K. (2010). Semantic web based Approach to learn and update Learner Profile in Adaptive E-Learning. Al-Azhar Engineering Eleventh International Conference, December 23-26.
[48] Fouad, K., Harb, H., & Nagdy, N. (2011). Semantic Web supporting Adaptive E-Learning to build and represent Learner Model, The Second International Conference of E-learning and Distance Education - eLi 2011.
[49] Z. Jiuling, S. Chuan, D. Beixing Deng, and L. Xing. (2009). Using WordNet in Conceptual Query Expansion. CCIS 30, pp. 210–218. Springer-Verlag Berlin Heidelberg.
[50] Jones, K. (2004). A Statistical Interpretation of Term Specificity and its Application to Retrieval. Journal of Documentation, 60 (5), p.493-502.
[51] Ali, R. & Beg, M. (2011). An overview of Web search evaluation methods. Computers and Electrical Engineering 37 (2011) 835–848. Elsevier Ltd.

Khaled M. Fouad received his Master degree of AI and expert systems. He is currently a PhD candidate in the faculty of engineering AlAzhar University in Egypt. He is working now as lecturer in Taif University in Kingdom of Saudi Arabia (KSA) and is assistant researcher in Central Laboratory of Agriculture Expert Systems (CLAES) in Egypt. His current research interests focus on Semantic Web and Expert Systems.

Ahmed R. Khalifa He received his PhD Degree in Computer Science from the City University of New York (CUNY) in 1993. He is currently an Associate Professor in Systems and Computer Engineering Department, AlAzhar University, Cairo, Egypt. His research interests include Information Security, Wireless Networks, Network Security, and Web Services Technologies and Security.

Nagdy M. Nagdy is professor of engineering applications and computer systems, Department of Systems Engineering and Computer Engineering - Faculty of Engineering AlAzhar University. He is working now in Al-Baha Private College of Science, Kingdom of Saudi Arabia (KSA) .He received his Ph.D in 1986. He has supervision of some master's and doctoral degrees in the department of Systems Engineering and Computer and Electrical Engineering.

Hany M. Harb is professor of Computers and Systems Engineering Department - Faculty of Engineering AlAzhar University. Doctor of philosophy (Ph.D.), Computer Science, Illinois Institute of Technology (IIT) , Chicago , Illinois, USA, 1986 He is Chairman of Computers and Systems Engineering department, Chairman of Systems and Networks Unit in AlAzhar university, and manager of WEB-Based Tansik program. He has supervision of many master's and doctoral degrees in the department of Systems and Computers Engineering.