Personalized Semantic Retrieval and Summarization of Web Based Documents

Salah T. Babekr
Computer Science Dept., College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia (KSA)

Khaled M. Fouad
Computer Science Dept., College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia (KSA)

Naveed Arshad
Computer Science Dept., College of Computers and Information Technology, Taif University,Kingdom of Saudi Arabia (KSA)

***Abstract* —The current retrieval methods are essentially based on the string-matching approach lacking of semantic information and can't understand the user's query intent and interest very well. These methods do regard as the personalization of the users. Semantic retrieval techniques are performed by interpreting the semantic of keywords. Using the text summarization allows a user to get a sense of the content of a full-text, or to know its information content, without reading all sentences within the full-text.**

In this paper, a semantic personalized information retrieval (IR) system is proposed, oriented to the exploitation of Semantic Web technology and WordNet ontology to support semantic IR capabilities in Web documents. In a proposed system, the Web documents are represented in concept vector model using WordNet. Personalization is used in a proposed system by building user model (UM). Text summarization in a proposed system is based on extracting the most relevant sentences from the original document to form a summary using WordNet.

The examination of the proposed system is performed by using three experiments that are based on relevance based evaluation. The results of the experiment shows that the proposed system, which is based on Semantic Web technology, can improve the accuracy and effectiveness for retrieving relevant Web documents.

Keywords-Semnatic Web; WordNet; Personalization; User Model; Information Retrieval; Summerization.

I. INTRODUCTION

Internet access, such as World Wide Web (WWW), has made document retrieval increasingly demanding as collection and searching of documents has become an integral part of many people's lives. Accuracy and speed are two key measurements of effective retrieval methodologies. Existing document retrieval systems use statistical methods [1] and natural language processing (NLP) [2] approaches combined with different document representation and query structures. Document retrieval [3] has created many interests in the information retrieval (IR) community.

Document retrieval refers to finding similar documents for a given user's query. A user's query can be ranged from a full description of a document to a few keywords. Most of the extensively used retrieval approaches are keywords based searching methods, e.g., www.google.com, in which untrained users provide a few keywords to the search engine finding the relevant documents in a returned list [4]. Another type of document retrieval is to use a query context by using language modeling, to integrate several contextual factors so that document ranking will be adapted to the specific query contexts [5]. Using an entire document as a query performs well in improving retrieval accuracy, but it is more computationally demanding compared with the keywords based method [6].

The effectiveness of processes models based on keywords is limited by the phenomenon known as "keywords barrier", i.e., the internal representation of an information item by a set of words extracted from texts through statistical and / or syntactic techniques does not allow a considerable improvement of the effectiveness of IR systems and, in particular, the precision of their results. These limitations have stimulated the development of several techniques trying to extract meaning from texts, such as semantic analysis [7] to obtain more accurate internal representations of information items. However, there is a lack of semantic retrieval process models providing appropriate abstraction representations of the activities, products and techniques involved in such retrieval processes [8].

Several IR process models, such the Boolean [9], the vector space [10] and the probabilistic models [11] have been proposed to cover the activities and technical user queries as well as storage and retrieval of information items from unstructured sources. Classic models represent the documents with a set of keywords extracted from text and propose different approaches to retrieval and presentation of retrieved information items sorted according to their relevance.

Some of the reasons that the classical IR approaches tend to be less effective as the web evolves can be identified as follows:

Content of the current web is created using natural language and HTML is a formatting language which is used to render presentation to human. The content of the web pages are not understandable with agents.

- Classical IR models are based on the computation of words or word occurrence which is a semantically imprecise calculus.

- The metadata is not available with the current web resources and there is no such a standard for creating the metadata.
- Interoperability and reusability of the web content is difficult due to heterogeneity of the web contents.

Personalized Semantic retrieval and summarization architecture aims at improving the conventional IR which is based on semantic Web technology. The personalized semantic enhanced retrieval and summarization framework is proposed that meets our objectives. The work begins with an overview of the research and then provides a comprehensive literature review on the related research topics. In particular, we conducted a selected study on the existing semantic IR systems and provide a detailed survey. More importantly, we suggest some improvements after the study of the existing systems. The idea also outlines our methodology towards designing a personalized semantic IR system.

II. BASIC CONCEPTS

A. Semantic Web

The Semantic Web [12] is a Web-based technology that extends XML by providing the means to define ontologies; the definition of objects and relationships between them. This allows machines to make intelligent inferences about objects across the Web. This allows intelligent agents [13] embodying knowledge about certain aspects of software development (much of it may be organization-specific) to make intelligent inferences that can be used as the basis for improved decision-making on software development processes, and usability issues. In addition, the semantic web is an approach to facilitate communication by making the web suitable for machine-to-machine communication [14]. It can be used to encode meaning and complex relationships in web pages. A major challenge for the emerging semantic-web field is to capture the knowledge required and structure it in a format that can be processed automatically (e.g., by agents).

Informally, ontology [15] of a certain domain is about terminology (domain vocabulary), all essential concepts in the domain, their classification, their taxonomy, their relations (including all important hierarchies and constraints), and domain axioms. More formally, to someone who wants to discuss about topics in a domain using a language, ontology provides a catalogue of the types of things assumed to exist in a domain; the types in the ontology are represented in terms of the concepts, relations, and predicates of language.

WordNet

WordNet [16, 17] has been used in several capacities to improve the performance of IR systems. WordNet can be used to solve the research problems in IR.

To overcome the weaknesses of term-based representation that is found in the conventional IR approaches, an ontology-based representation has been recently proposed [18], which exploits the hierarchical is-a relation among concepts, i.e., the meanings of words. For example, to describe with a term-based representation documents containing the three words: “animal”, “dog”, and “cat” a vector of three elements is needed; with an ontology-based representation, since “animal” subsumes both “dog” and “cat”, it is possible to use a vector with only two elements, related to the “dog” and “cat” concepts, that can also implicitly contain the information given by the presence of the “animal” concept. 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.

In the text representation, the terms are replaced by their associated concepts in WordNet [19]. In the pretreatment phase, it firstly convert uppercase characters into lowercase characters and then eliminate from text punctuation marks and stop words such as: are, that, what, do. This representation requires two more stages: a) the “mapping” of terms into concepts and the choice of the “merging” strategy, and b) the application of a disambiguation strategy. The first stage is shown in example, as found in figure 1, is about mapping the two terms government and politics into the concept government (the frequencies of these two terms are thus cumulated). Then, among the three “merging” strategies offered by the conceptual approach (“To add concept”, “To replace terms by concepts” and “concept only”), the strategy “concept only“ can be chosen, where the vector of terms is replaced by the corresponding vector of concepts (excluding the terms which do not appear in WordNet).

Voorhees [20] suggested that WordNet can be used in IR for query expansion. Query expansion is considered to be one of the techniques that can be used to improve the retrieval performance of short queries. Most of the indexing and retrieval methods are based on statistical methods; short queries posed challenges to this model due to the limited amount of information that can be gathered during its processing. In expanding the query, Voorhees suggested using of synonyms, hypernyms, hyponyms, and their combinations. The results showed that using of synonyms, hypernyms, and hyponyms are significant in the retrieval performance for short queries, but little improvement when they are applied to the long query.

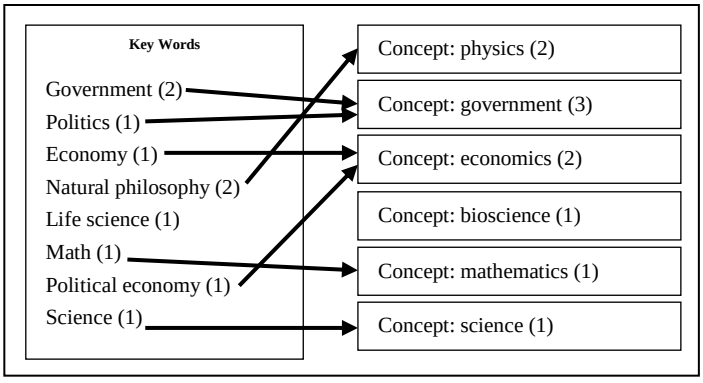

Figure 1. Example of mapping words in concepts

B. Personalization

The goal of personalization [21] is to endow software systems with the capability to change (adapt) aspects of their functionality, appearance or both at runtime to the particularities of users to better suit their needs. The recent rapid advances in storage and communication technologies

stress the need for personalization. This need is more evident in consumer oriented fields, like news content personalization systems, recommendation systems, user interfaces, and applications like home audiovisual material collection and organization, search engines in multimedia browsing and retrieval systems, providing services for personalized presentation of interactive video. The core idea of personalization is to customize the presentation of information specifically to the user to make user interfaces more intuitive and easier to understand, and to reduce information overload.

User modeling [22] describes the process of creating a set of system assumptions about all aspects of the user, which are relevant to the adaptation of the current user interactions. This can include user goals, interests, level of expertise, abilities and preferences. The most reliable method of user modeling is by explicit entry of information by the user. In most practical systems, this is too time-consuming and complex for the user. Hence implicit user modeling, based on analysis of past and current user interactions, is critical. The user profile is a machine-processable description of the user model [23, 24].

C. Text summarization

Text summarization [25] is a data reduction process. The use of text summarization allows a user to get a sense of the content of a full-text, or to know its information content, without reading all sentences within the full-text. Data reduction increases scale by (1) allowing users to find relevant full-text sources more quickly, and (2) assimilating only essential information from many texts with reduced effort. Text summarization is particularly useful in certain domain, where oncologists must continuously find trial study information related to their specialty, evaluate the study for its strength, and then possibly incorporate the new study information.

Text Summarization [26] methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization [26] attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language.

Extractive summaries [26, 27]are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted.

Extractive text summarization process can be divided into two steps [28]:

1) Preprocessing step, in this step Sentences boundary identification, Stop-Word Elimination and Stemming are performed and,

2) Processing step, in this step features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

III. Related Work

Personalized search [29] is addressed by a number of systems. Persona [30] used explicit relevant feedback to update user profiles that are represented by means of weighted open directory project taxonomy [31]. These profiles are used to filter search results. Personalized variants of PageRank, is as found in PersonalizedGoogle or the Outride Personalized Search System [32]. Persival [33] re-ranked the search results of queries for medical articles profiles keywords, associated concepts, and weights generated from an electronic patient record.

In [34] it was filtered search results on the grounds of user profiles obtained from earlier queries. These profiles consist of a set of categories, and weighted terms associated with each category. In their work on personalizing search results, [35] distinguish between long-term and short-term interests. While aiming at personalization in a broader sense, [36] use click-through data to increase the performance of search results.

Nowadays [37], personalization systems are developed by considering ontology to reduce the limitation of traditional IR such as information overload or cold start problem. So considering ontology to build an accurate profile brings some extra benefit in user modeling. A user profile can be presented as a weighted concept hierarchy for searching and browsing in the web. User profile can be created by user with his/her personal information and interest or it can be a reference one. However, profile can be created by manually entering the user's information or automatically by watching the use's activities.

Jin and others [38], proposed a novel approach which enables intelligent semantic web search for best satisfying users search intensions. The proposed approach combines the user's subjective weighting importance over multiple search properties together with fuzziness to represent search requirements. A special ranking mechanism based on the above weighed fuzzy query is also presented. The ranking method considers not only fuzzy predicates in the query, but also the user's personalized interests or preferences.

MedSearch is a complete retrieval system for medical literature [39]. It supports retrieval by SSRM (Semantic Similarity Retrieval Model), a novel IR method which is capable for associating documents containing semantically similar (but not necessarily lexically similar) terms. SSRM suggests discovering semantically similar terms in documents and queries using term taxonomies (ontologies) and by associating such terms using semantic similarity methods. SSRM demonstrated very promising performance achieving significantly better precision and recall than Vector Space Model (VSM) for retrievals on Medline.

In [40], the authors proposed a new approach to User Model Acquisition (UMA) which has two important features. It doesn't assume that users always have a well-defined idea of what they are looking for, and it is ontology-based, i.e., it was dealt with concepts instead of keywords to formulate queries.

The first problem is that most approaches assume users to have a well-defined idea of what they are looking for, which is not always the case. They solved this problem by letting fuzzy user models evolve on the basis of a rating induced by user behavior. The second problem concerns the use of keywords, not concepts, to formulate queries. Considering words and not the concepts behind them often leads to a loss in terms of the quantity and quality of information retrieved. They solved this problem by adopting an ontology-based approach.

In [41], authors introduced a method for learning and updating a user profile automatically. The proposed method belongs to implicit techniques. It processes and analyzes behavioral patterns of user activities on the web, and modifies a user profile based on extracted information from user's web-logs. The method relies on analysis of web-logs for discovering concepts and items representing user's current and new interests. The mechanism used for identifying relevant items is built based on a newly introduced concept of ontology-based semantic similarity.

Dıaz and Gervas [42] have proposed the personalized summarization as a process of summarization that preserves the specific information that is relevant for a given user profile, rather than information that truly summarizes the content of the news item. The potential of summary personalization is high, because a summary that would be useless to decide the relevance of a document if summarized in a generic manner, may be useful if the right sentences are selected that match the user interest. Authors defend the use of a personalized summarization facility to maximize the density of relevance of selections sent by a personalized information system to a given user.

Lv, Zheng and Zhang [43] have developed the method of IR based on semantics. In addition, they took the "wine" ontology instances provided by Stanford University as a reference, and develop a Chinese "wine" model by using protégé tools. Finally, the retrieval results show that the proposed method has higher recall and precision.

Rinaldi [44] have given the solution for the problem of IR on the Web using an approach based on a measure of semantic relatedness applied to evaluate the relevance of a document with respect to a query in a given context: the concepts of lexical chains, ontologies, and semantic networks. The proposed methods, metrics, and techniques are implemented in a system called DySE (Dynamic Semantic Engine). DySE implements a context-driven approach in which the keywords are processed in the context of the information in which they are retrieved, in order to solve semantic ambiguity and to give a more accurate retrieval based on the real of the user interests.

Huang and Zhang [45] proposed the approach to expand the set of query keywords based on associational semantics. Firstly, they constructed a group of semantic trees for original keywords one by one based on WordNet, an online lexical system. The original keywords perch on the roots of the trees. Secondly, they removed noise nodes in the trees by computing the similarity between words, and assemble the trees into a big integrated tree, i.e. Tree of Associational Semantics Model, by expanding the roots of the trees upward until finding the common origin of the trees. They assigned a weight to each word on the trees, and selected candidates from the trees by referring to thresholds. Finally, they executed the document retrieval by importing the weights and distribution density of keywords into calculation of similarities between query and documents.

Gauch, Speretta and Pretschner [46] explored the use of ontology-based user profiles to provide personalized search results. In this work, authors used the ontology that consists of hierarchies of concepts in which each concept is defined by a set of documents, and hierarchy is induced by an informal specialization relationship. They reviewed a variety of sources of information from which the ontology-based profiles can be created, and described improvements in accuracy achieved when the user profiles are used to select search results.

IV. Architecture of Personalized Semantic Retrieval and Summarization

The semantic retrieval approach embeds background knowledge with explicitly defined semantics can help to build intelligent IR applications. Based on some of the weaknesses of conventional IR techniques, the motivations towards a semantic IR framework have been identified.

Using of text summarization allows a user to get a sense of the content of a full-text, or to know its information content, without reading all sentences within the full-text. Data reduction increases scale by (1) allowing users to find relevant full-text sources more quickly, and (2) assimilating only essential information from many texts with reduced effort. Text summarization is particularly useful in certain domain, where oncologists must continuously find trial study information related to their specialty, evaluate the study for its strength, and then possibly incorporate the new study information.

The improved and practical approach is presented to automatically summarizing Web documents by extracting the most relevant sentences from the original document to create a summarization. The idea of proposed approach is to find out key sentences from the Keyword extraction based on statistics and synsets extraction using WordNet. These two properties can be combined and tuned for ranking and extracting sentences to generate a list of candidates of key sentences. Then semantic similarity analysis is conducted between candidates of key sentences to reduce the redundancy. The entire architecture of the proposed approach is shown in the figure 2.

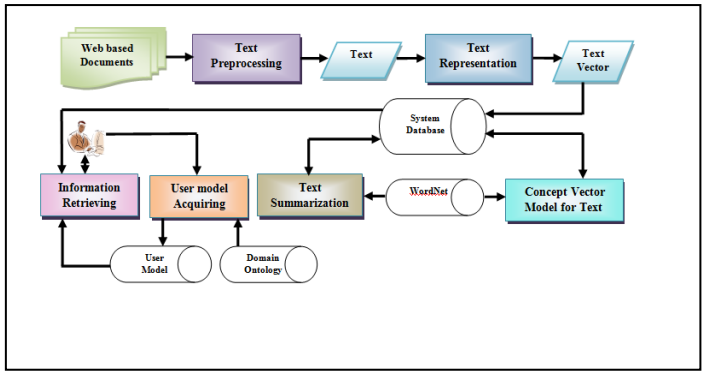

Figure 1. Architecture of the proposed approach

The detail of each process is found in the next sections.

A. Text Preprocessing

The most widely accepted document representation model in text classification is probably Vector Space Model (VSM) [10]. VSM is adapted in the proposed system to achieve effective representations of documents. The documents must be preprocessed before the text representation. The main procedures of preprocessing are shown in figure 3.

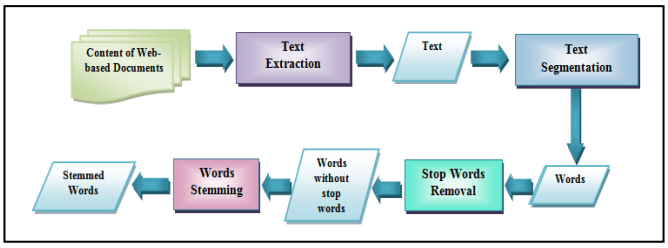

Figure 2. Main steps for text preprocessing

Text Extraction

The first step of the text representation process is extracting textual data from the web pages. Then convert each page into individual text document to apply text preprocessing techniques on it. This step is applied on input Web documents dataset by scanning the web pages and categorizing the HTML tags in each page.

Then exclude the tags that contain no textual information like formatting tags and imaging tags (i.e. <HTML>, <BODY>, <IMG>, etc.). Also exclude all the scripts and codes that are found in the page like JavaScript and VBScript. Then extract the textual data from other tags (like paragraphs, hyperlinks, and metadata tags) and store it into individual text documents as input for next steps. To extract the text from Web documents, open source high-performance .NET C# module is used that was created to parse HTML [47] for links, indexing and other purposes.

1) Stop Words Removal

Stop words, i.e. words thought not to convey any meaning, are removed from the text. In this work, the proposed approach uses a static list of stop words about all tokens. This process removes all words that are not nouns, verbs or adjectives. For example, stop words removal process will remove all the words like: he, all, his, from, is, an, of, your, and so on. Removing these words will save spaces for storing document contents and reduce time taken during the search process.

2) Words Stemming

The stem is the common root-form of the words with the same meaning appear in various morphological forms (e.g. player, played, plays from stem play). In the proposed approach, the morphology function [48] based on WordNet [49] to perform stemming process. Stemming will find the stems of the output terms to enhance term frequency counting process because terms like “computers” and “engineering” come down from the same stem “computer” and " engineer". This process will output all the stems of extracted terms [50, 51].

B. Text Representation

VSM [10] is adapted in the proposed system to achieve effective representations of documents. Each document is identified by n-dimensional feature vector where each dimension corresponds to a distinct term. Each term in a given document vector has an associated weight.

The weight is a function of the term frequency, collection frequency and normalization factors. Different weighting approaches may be applied by varying this function. Hence, a document j is represented by the document vector d_j :

$$d_j = (w_{1j}, w_{2j}, w_{nj}) \qquad (1)$$

Where, w_{nj} is the weight of the kth term in the document j.

The term frequency reflects the importance of term k within a particular document j. The weighting factor may be global or local. The global weighting factor clarifies the importance of a term k within the entire collection of documents, whereas a local weighting factor considers the given document only. The document keywords were extracted by using a term-frequency and inverse-document-frequency (*tf-idf*) calculation [52], which is a well-established technique in IR. The weight of term k in document j is represented as:

$$w_{kj} = tf_{kj} \times (\log_2^n - \log_2^{df_k} + 1) \qquad (2)$$

Where: tf_{kj} = the term k frequency in document j, df_k = number of documents in which term k occurs, n = total number of documents in collection. The output of this step is the weight of terms in selected document.

C. Concept Vector Model of Text using WordNet

The purpose of this step is to identify WordNet concepts that correspond to document words. Concept identification [53] is based on the overlap of the local context of the analyzed word with every corresponding WordNet entry. The words mapping into concepts algorithm for the terms is given in figure 4.

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 = ∩ ($Cont_i$, 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 3. The algorithm of Words Mapping into Concepts

1) Weight of Concept Computation

The concepts in documents are identified as a set of terms that have identified or synonym relationships, i.e., synsets in the WordNet ontology. Then, the concept frequencies Cf_c are calculated based on term frequency tf_{tm} as follows [54]:

$$Cf_c = \sum_{t_m \in r(c)} tf_{tm} \qquad (3)$$

Where r(c) is the set of different terms that belongs to concept C .

D. Text Summarization

Text summarization [25] aims at compressing an original document into a shortened version by extracting the most important information out of the document.

Extractive summary [26, 27] is used in the proposed system by extracting key text segments from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. In the proposed system, the text summarization is performed by extracting the most relevant sentences, that are key sentences, from original document by calculating the weight of sentences [55] and then select the heigher weight. Semantic simalirity using WordNet is used to filter and refine the selected senetence to extract semantic dissimilar sentences. Figure 5 shows the algorithm of text summarization.

Input: document *D*.
Output: Set (Set_{Sum}) of Summary Sentences (Sum_S) in document *D*.
Procedure:
Split and get set of sentences (Sen) in document D
// Step1: Calculate the weight Sen_w of sentence (Sen)
Do While *i* <= C_{Sen} // Count of *Sen* in D
For Each term T **In** Sen
Get Term Weight (T_w) as found in equation 2.
Insert T_w to the set of term weight Set_{tw}
EndFor
Get Length of Sentence (Len(Sen))
Calculate $Sen_w = \frac{\sum T_w}{Len(Sen)}$
If *Sen* in title or subtitle **Then** // if the sentence is found in distinguished location
Calculate weight of the sentence location Loc_w //where $1 \le Loc_w \le 1.6$
Else
$Loc_w = 1$
EndIf
If *Sen* contains special phrases **Then** // such as "this paper propose; this article introduce;.."
Calculate weight of the sentence Sp_w
Else
$Sp_w = 1$
EndIf
Calculate $Sen_w = \frac{\sum T_w}{Len(Sen)} \times Loc_w \times Sp_w$ // The weight of sentence
Insert *Sen* and its weight Sen_w to list $list_{sen}$
Rank the list by the Sen_w
EndDo
// Step2: Filtering and refining the output sentence *Sen* using semantic similarity using WordNet
Do While *j* <= C_{list} // C_{list} is Count of *Sen* in $list_{sen}$
Get Semantic Similarity ($SemSim_{Sen(j),(j+1)}$) of *Sen* (*j*) *and Sen* (*j* + 1)
Get Sentences (Sum_S) that are Semantic Dissimilar
Insert (Sum_S) *in* (Set_{Sum})
EndWhile

Figure 4. The algorithm of Text Summarization

E. User Model Acquiring

This step aims at building the user model using user behavior in the system. There are roughly two kinds of automatic way to capture a user's interest implicitly: behavior-based and history-based. 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 [56] always constitute the most important part of the user profile in adaptive IR and filtering systems that dealt with large volumes of information.

The main purpose of this step is acquiring the interested concepts of the user in the web page (document), and then gets concept frequency that reflects the importance of concept, and finally gets 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.

User model in the proposed system is built in ontological representation by using domain ontology. User model is built by mapping of user's interest information and the concept in domain ontology; convert the contents of the user's interest into the form of ontology concept, and using these ontology concepts to construct user interest ontology.

Figure 6 shows the algorithm of user model acquiring.

Input: (B_C) Bag of concepts (C_i) in represented document *D* that was browsed by the user during using the system as found in section IV(B); Concepts C_{Ont} in domain ontology *DO*.
Output: User Model (UM) in ontological representation**.**
Procedure:
// Step 1: Acquire User Interest to build UM .
Do While *i* <= C_c // (C_c) *is the count of Concepts* C_i *in the bag,*
Get concept weight W_{Ci} for C_i by using equation 3
Save concept C_i and its weight W_{Ci} as user interest and its weight in UM
EndDo
// Step 2: Build the UM as ontological representation (user ontology).
For Each C_i **In** UM
If C_i is similar to concept C_{Ont} in *DO* then
Get Concept relations Rel_c for C_i from DO
Get W_{Ci} for C_i from UM
Insert C_{Ont} and its W_{Ci} to user ontology node.
Insert Rel_c of C_{Ont} to all related concepts
Else
Insert C_i and its W_{Ci} to user ontology node.
EndIf
EndFor

Figure 5. The algorithm of User Model acquiring

F. Information Retrieving

The document retrieval [57] is based on semantic similarity of the query term vector and document vector using equation 5.

$$sim(q, d) = \frac{\sum_i \sum_j q_j w_i \, sim(i,j)}{\sum_i \sum_j q_j w_i} \qquad (4)$$

where w_i is term weight of concept 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 [58] of the term i and term j.

Finally, the system should arrange the retrieved documents by using the semantic similarity score of the query term vector and document vector. After building the user model; that is based on the user interest, the system uses the user model to rerank the retrieved documents. Reranking the retrieved documents user model makes the documents appears in the order as the user interest is matched. 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.

V. IMPLEMENTATION AND EXPERIMENTATION

In this section, the results of the experiments carried out to evaluate the performance of proposed system will be discussed from a quantitative point of view by running some experiments to evaluate the precision of the results. A test set collection is used to evaluate the proposed system. The test collection is a set of documents, queries and a list of relevance documents. These are used to compare the results of proposed system using the ranking strategies described.

The proposed system is implemented in ASP.Net as Web-based system using Visual Studio 2010, .NET Framework 4, and SQL Server 2008. The number of stored documents is about 3000 documents. These Web documents are about computer science domain.

Figure 7 shows the samples of the extracted texts from the collected documents.

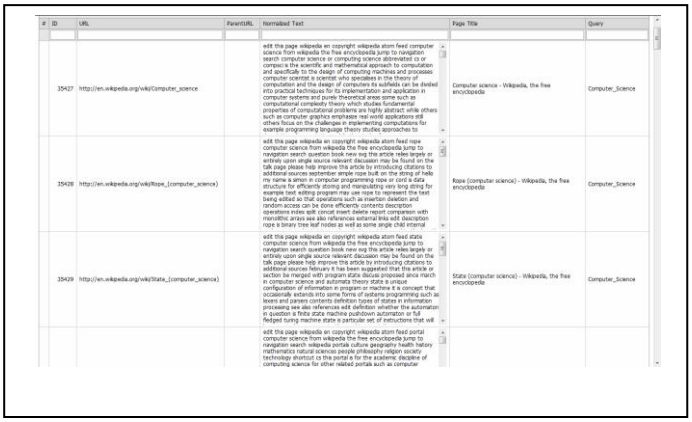

Figure 7. Samples of the extracted texts from the collected documents

Figure 8 shows the results of the retrieving subsystem. This subsystem retrieves the documents based on semantic similarity between the query and the collected documents.

The improvement is measured by performing three experiments. In the experiments, relevance based evaluation method [59] is used. It uses the metrics; precision, recall, f-measure, average precession (AP) and mean average precision (MAP), to measure the performance of proposed system. In this method, proposed system is judged according to the search results' ability to satisfy an easily pleased user or hard to please user.

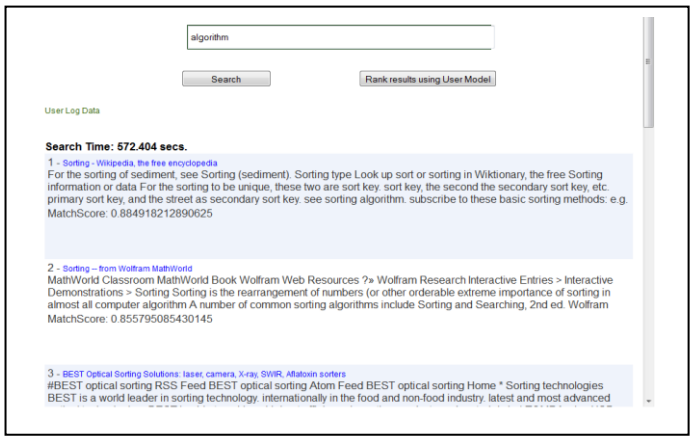

Figure 6. The results of the retrieving subsystem for "algorithm" term

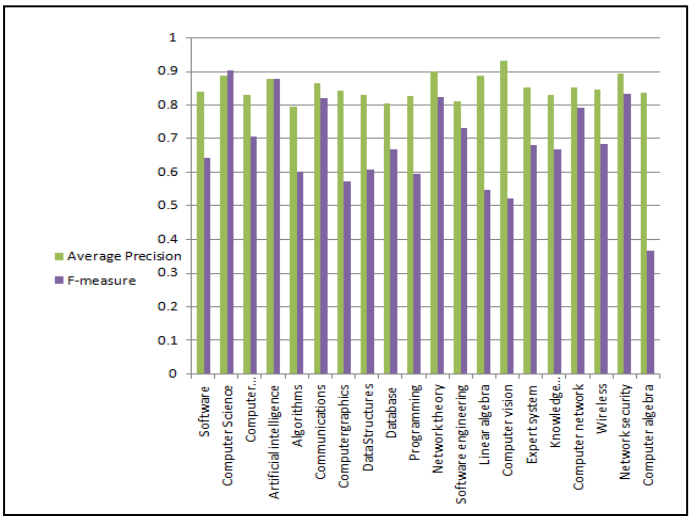

Figure 8. The average precession and f-measure for different terms

In the first experiment, figure 9 shows the average precession and f-measure that is based on different retrieving results for different user query.

The results of the different queries show that the system gives high precision during retrieving documents.

The second experiment aims at determining the importance of semantic similarity during determining the documents that are relevant to the user query. This experiment measures the performance degree when the system uses this function. It compares the recall, precision, average precision and MAP with and without using this function.

Figure 10 shows the charts of the MAP of comparison for using the semantic similarity (SemSim) between documents vector and query vector or not. This experiment shows that using semantic similarity, when the system determines the documents that are matched user query, increases the accuracy of document retrieving.

The third experiment aims at determining the importance of personalization by using generated user model (UM) during using the system. The user model is used to re-rank the retrieved documents to match the user interest.

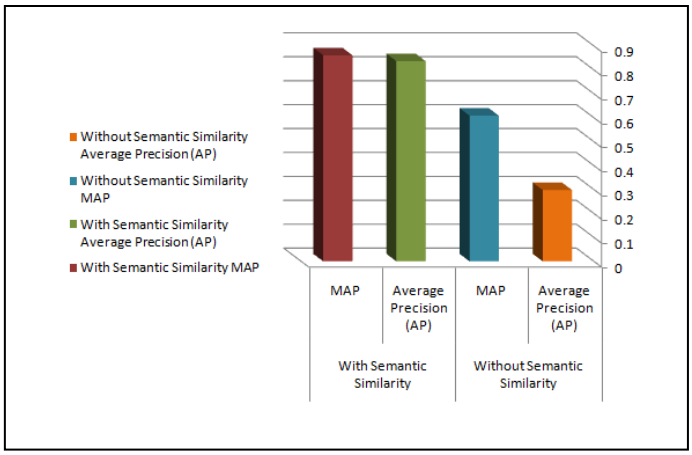

Figure 9. Coparison of using Semantic Similarity during determining the documents that are relevant to the user query

This experiment measures the performance degree when the system uses this function. It compares the recall, precision, average precision and MAP with and without using this function. Figure 11 shows the charts of the MAP of comparison for using the UM to re-rank the retrieved documents. This experiment emphasizes that using UM to realize the personalization aims at improving documents retrieving results by re-ranking the retrieved results based on user interests.

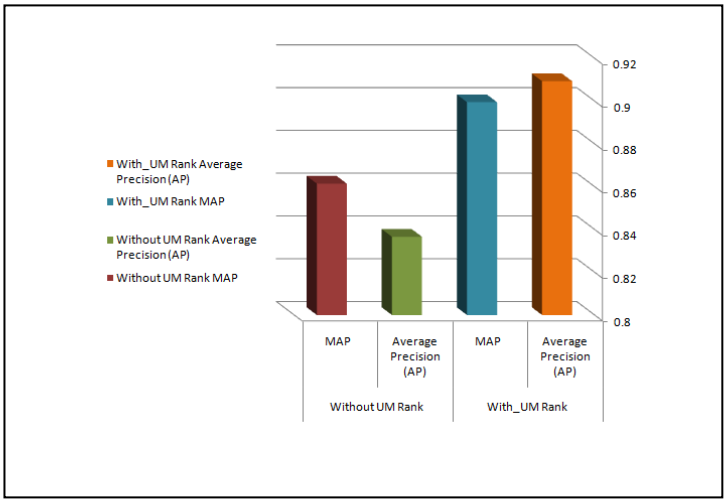

Figure 10. Coparison of using User Model during re-ranking the retrieved documents

VI. Conclusion

Aiming to solve the limitations of keyword-based models, the idea of semantic search, understood as searching by meanings rather than literal strings, has been the focus of a wide body of research in the IR and the Semantic Web communities.

A system for personalized semantic IR and summarization has been presented. The Semantic Web is a new approach for organizing information and it represents a great interest area for the international research community, but it is still far from a large-scale implementation. In this work, we have proposed and implemented the system for IR based on Semantic Web, defining a strategy for scoring and ranking results by means of a novel metric to measure semantic relatedness between terms. In the proposed system, user model, which is user interests, is used to realize the personalization. It is acquired by using concept vector model and WordNet ontology to be represented in semantic representation. In the proposed approach, summarization is based on extractive summary. Summarization is implemented by extracting the most relevant sentences, that are key sentences, from original document by calculating the weight of sentences and then select the heigher weight. Semantic simalirity using WordNet is used to filter and refine the selected .

In the system evaluation, three experiments; that are based on relevance evaluation method, show that the system can improve the accuracy of the IR because it depends on the Semantic Web tecknolgy. The system performs the summarization to allow users to find relevant full-text sources more quickly.

References

[1] Ramachandra, M. (2010). Information Retrieval. In: Web-Based Supply Chain Management and Digital Signal Processing: Methods for Effective Information Administration and Transmission. PP: 182-194. DOI: 10.4018/978-1-60566-888-8.ch014. IGI Global.

[2] Yue, X., Di, G. Yu, Y. Wang, W. & Shi, H. (2012). Analysis of the Combination of Natural Language Processing and Search Engine Technology. 2012 International Workshop on Information and Electronics Engineering (IWIEE). Procedia Engineering 29 (2012) 1636 – 1639. Elsevier Ltd.

[3] Liddy, M. (2006). Document Retrieval, Automatic. In: Encyclopedia of Language & Linguistics (Second Edition) 2006, Pages 748–755. Elsevier Ltd.

[4] MITRA, M. & CHAUDHURI, B. (2000). Information Retrieval from Documents: A Survey. Information Retrieval 2, 141–163 (2000). Kluwer Academic Publishers.

[5] Bai, J. & Nie, J. (2008). Adapting information retrieval to query contexts. Information Processing and Management 44 (2008) 1901–1922. Elsevier Ltd.

[6] Chow, T., Zhang, H. & Rahman, M.. (2009). A new document representation using term frequency and vectorized graph connectionists with application to document retrieval. Expert Systems with Applications 36 (2009) 12023–12035. Elsevier Ltd.

[7] Chen, M., Chu, H. & Chen, Y. (2010). Developing a semantic-enable information retrieval mechanism. Expert Systems with Applications 37 (2010) 322–340. Elsevier Ltd.

[8] Silva, F., Girardi, R. & Drumond, L. (2009). An Information Retrieval Model for the Semantic Web. Sixth International Conference on Information Technology: New Generations. 978-0-7695-3596-8/09, IEEE.

[9] Vester, K. & Martiny M. (2005). Information retrieval in document spaces using clustering. IMM-Thesis . Informatics and Mathematical Modelling, Technical University of Denmark, DTU.

[10] Liu. Y. (2009). On Document Representation and Term Weights in Text Classification. In: Handbook of Research on Text and Web Mining Technologies. DOI: 10.4018/978-1-59904-990-8.ch001, 1-22. IGI Global.

[11] Grossman, D. A. & Frieder, O. (2004). Information Retrieval: Algorithms And Heuristics. The Springer International Series in Engineering and Computer Science. Springer.

[12] Berners-Lee, T. (1998). Semantic Web Roadmap, W3C Semantic Web Vision Statement. http://www.w3.org/DesignIssues/Semantic.html, Last accessed on 12/09/2011.

[13] Hendler, J. (2001). Agents and the Semantic Web. IEEE Intelligent Systems, 16 (2). 30-37.

[14] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web, Scientific American 284(5):35-43, http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html.

[15] Devedžić, V. (2005). Introduction to the Semantic Web. In: Integrated Series in Information Systems, Volume 12, 29-69, DOI: 10.1007/978-0-387-35417-0_2 . Springer-Verlag Berlin Heidelberg.

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

[17] Maria, I. & Loke, S. (2010). The Impact of Ontology on the Performance of Information Retrieval: A Case of WordNet, In G. I. Alkhatib, D. C. Rine, Web Engineering Advancements and Trends: Building New Dimensions of Information Technology, DOI: 10.4018/978-1-60566-719-5.ch002, 24-37.

[18] Pereira, d. C., Tettamanzi, C. (2006). A.G.B.: An ontology-based method for user model acquisition. In: Ma, Z. (ed.) Soft computing in ontologies and semantic Web. Studies in fuzziness and soft computing, pp. 211–227. Springer, Heidelberg.

[19] Amine, A., Elberrichi, Z. & Simonet, M. (2010). Evaluation of Text Clustering Methods Using WordNet, The International Arab Journal of Information Technology, (7) 4.

[20] Voorhees, E. (1994). Query Expansion Using Lexical-Semantic Relations. The 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (Dublin Ireland, 1994), 61. ACM.

[21] Anke, J. & Sundaram, D. (2009). Personalization Techniques and Their Application. In: Ang , S. & Zaphiris, P. Human Computer Interaction: Concepts, Methodologies, Tools, and Applications. DOI: 10.4018/978-1-87828-991-9.ch013. 168-176. IGI Global.

[22] 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.

[23] Baishuang, Q., & Wei, Z. (2009). Student Model in Adaptive Learning System based on Semantic Web, In First International Workshop on Education Technology and Computer Science, 978-0-7695-3557-9/09, IEEE, DOI 10.1109/ETCS.2009.466.

[24] Li, F., Li, Y., Wu, Y., Zhou, K., Li, Z., Wang, X. (2008). Discovery of a User Interests on the Internet, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, DOI 10.1109/WIIAT.2008.18, 978-0-7695-3496-1/08, IEEE.

[25] Reeve, L., Han, H. & Brooks, A. (2007). The use of domain-specific concepts in biomedical text summarization. Information Processing and Management 43 (2007) 1765–1776. Elsevier Ltd.

[26] Gupta, V. & Lehal, G. (2010). A Survey of Text Summarization Extractive Techniques. JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010. ACADEMY PUBLISHER.

[27] Kyoomarsi, F., Khosravi, H. Eslami, E., Dehkordy, P. & Tajoddin, A. (2008). Optimizing Text Summarization Based on Fuzzy Logic. Seventh IEEE/ACIS International Conference on Computer and Information Science. 978-0-7695-3131-1/08, IEEE.

[28] Gupta, V. & Lehal, G. (2009). A Survey of Text Mining Techniques and Applications. JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST, ACADEMY PUBLISHER.

[29] Gauch, S., Speretta, M. & Pretschner, M. (2007), Ontology-Based User Profiles for Personalized Search, DOI10.1007/978-0-387-37022-4, Springer US.

[30] Tanudjaja, F. & Mui, L. (2002) Persona: A Contextualized and Personalized Web Search. Proc 35 th Hawaii Intl. Conf. on System Sciences.

[31] The Open Directory Project (ODP). http://dmoz.org.

[32] Pitkow, J., Schütze, H. & Cass, T.. (2002), Personalized search. CACM 2002; 45(9):50-55.

[33] McKeown, K., Elhadad, N. & Hatzivassiloglou, V..(2003), Leveraging a common representation for personalized search and summarization in a medical digital library. In Proceedings of the 3 rd ACM/IEEE-CS joint conference on Digital libraries 2003; 159-170.

[34] Liu, F., Yu, C. & Meng, W. (2002) Personalized web search by mapping user queries to categories. In Proceedings CIKM'02 2002; 558-565.

[35] Sugiyama, K., Hatano, K. & Yoshikawa, M. (2004), Adaptive web search based on user profile constructed without any effort from users. In Proceedings 13 th Intl. Conf. on World Wide Web 2004; 675-684.

[36] Xiangwei, M., Yan, C. & Nan, L. (2009), Modeling of Personalized Recommendation System Based on Ontology, 978-1-4244-4639-1/09, IEEE.

[37] Trong, D., Mohammed, N, Delong, L., & Geun, J. (2009), A Collaborative Ontology-Based User Profiles System, N.T. Nguyen, R. Kowalczyk, and S.-M. Chen (Eds.): ICCCI 2009, LNAI 5796, pp. 540–552, Springer-Verlag Berlin Heidelberg.

[38] Jina, H. , Ninga, X., Jiab, W., Wua, H. & Luc, G.. (2008), Combining weights with fuzziness for intelligent semantic web search, Knowledge-Based Systems 21 (2008) 655–665, 0950-7051, Elsevier.

[39] Yang. Q., Sun, J., Li, Y. & Ca, K. (2010), Domain Ontology-based personalized recommendation research, 978-1-4244-5824-0, IEEE.

[40] Pereira, C. & Tettamanzi, A. (2006), An Evolutionary Approach to Ontology-Based User Model Acquisition, V. Di Ges´u, F. Masulli, and A. Petrosino (Eds.): WILF 2003, LNAI 2955, pp. 25–32, c_Springer-Verlag Berlin Heidelberg.

[41] Reformat, M. Koosha, S. (2009). Updating User Profile using Ontology-based Semantic Similarity, FUZZ_IEEE 2009, Korea, August 20-24, 978-1-4244-3597-5, IEEE.

[42] Dıaz, A. & Gervas, P. (2007). User-model based personalized summarization. Information Processing and Management 43 (2007) 1715–1734. Elsevier Ltd.

[43] Lv, G., Zheng, C. & Zhang, L. (2009). Text Information Retrieval Based on Concept Semantic Similarity. 2009 Fifth International Conference on Semantics, Knowledge and Grid. 978-0-7695-3810-5/09. IEEE.

[44] Rinaldi, A. M. (2009). An ontology-driven approach for semantic information retrieval on the Web. ACM Trans. Internet Technol, 9, 3, Article 10 (July 2009), 24 pages. DOI = 10.1145/1552291.1552293 http://doi.acm.org/10.1145/1552291.1552293

[45] Huang, G. & Zhang, X. (2010). Text Retrieval based on Semantic Relationship. 978-1-4244-7161-4/10, IEEE.

[46] Gauch, S., Speretta, M. & Pretschner, A. (2007). ONTOLOGY-BASED USER PROFILES FOR PERSONALIZED SEARCH. In: Ontologies A Handbook of Principles, Concepts and Applications in Information Systems, Volume 14, 2007, DOI: 10.1007/978-0-387-37022-4. Springerlink.

[47] Majestic-12: Projects : C# HTML parser (.NET). http://www.majestic12.co.uk/projects/html_parser.php.

[48] wordnetdotnet - Revision 262. http://wordnetdotnet.googlecode.com/svn/trunk/Projects/Thanh/.

[49] Bai, R., Wang, X. & Liao, J. (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.

[50] Gharib, T., Fouad, M. & Aref, M. (2010). Fuzzy Document Clustering Approach using WordNet Lexical Categories. In: Advanced Techniques in Computing Sciences and Software Engineering. DOI 10.1007/978-90-481-3660-5, Springer Science+Business Media.

[51] Tarek, G., Fouad, M. & Aref, M. (2008). Web Document Clustering Approach using WordNet Lexical Categories and Fuzzy Clustering. Proceedings of International Workshop on Data Mining and Artificial Intelligence (DMAI' 08), 24 December, 2008, Khulna, Bangladesh. 1-4244-2136-7/08, IEEE.

[52] Jones, K. (2004). A Statistical Interpretation of Term Specificity and its Application to Retrieval. Journal of Documentation, 60 (5), p.493-502.

[53] 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.

[54] Dragoni, M., Pereira, C. & Tettamanzi, A. (2010). An Ontological Representation of Documents and Queries for Information Retrieval Systems, IEA/AIE 2010, Part II, LNAI 6097, pp. 555–564, Springer-Verlag Berlin Heidelberg.

[55] Xu, X. (2009). Research on Automatic Summarization System based on topic partition. 2009 International Conference on Web Information Systems and Mining. 978-0-7695-3817-4/09, IEEE.

[56] 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.

[57] Lv, G., Zheng. C. & Zhang, L. (2009). Text Information Retrieval Based on Concept Semantic Similarity. Fifth International Conference on Semantics, Knowledge and Grid. 978-0-7695-3810-5/09, IEEE.

[58] Saruladha, K., Aghila.G. & Raj, S. (2010). A Survey of Semantic Similarity Methods for Ontology based Information Retrieval. Second International Conference on Machine Learning and Computing. 978-0-7695-3977-5/10, IEEE.

[59] Ali, R. & Beg, M. (2011). An overview of Web search evaluation methods. Computers and Electrical Engineering 37 (2011) 835–848. Elsevier Ltd.

AUTHORS PROFILE

Salah T. Babekr is an associate professor of Computer Engineering in College of Computers and Information Technology, Taif University. He is PhD holder for 16 years as a Computer Engineer bilingual Russian and English with extensive experience in administration and project control,analysis, design, consultancy, development and implementation of software, organization, establishment and improvement ofInternet band networks security, quality control on software products, implementation of best development practices, teamwork and support.

Khaled M. Fouad has received his PhD and Master degree of AI, and expert systems in of computer engineering from the faculty of engineering AlAzhar University in Egypt. He is working now as assistant professor in Taif University in Kingdom of Saudi Arabia (KSA) and is researcher in Central Laboratory of Agriculture Expert Systems (CLAES) in Egypt. His current research interests focus on semantic web, text mining, clustering and expert systems.

Naveed Arshad has completed his Ph.D. from University of Colorado at Boulder, USA. Before joining LUMS, Dr Naveed Arshad has worked with ABN AMRO Global IT Systems, Pakistan International Airline. He is part of the Software Engineering Research Group (SERG) at LUMS. This group is undertaking research in various areas of software engineering such as engineering of autonomic systems, conceptual modeling, large scale systems development, etc.