Semantic Retrieval Approach for Web Documents

Hany M. Harb
Computers and Systems Engineering Dept., Faculty of Eng., Al-Azhar Univ., Egypt.

Khaled M. Fouad
Computer Science Dep., Community College , Taif Univ., Kingdom of Saudi Arabia (KSA).

Nagdy M. Nagdy
Engineering Applications and Computer Systems,
Al-Baha Private College of Science,
Kingdom of Saudi Arabia (KSA).

***Abstract*— Because of explosive growth of resources in the internet, the information retrieval technology has become particularly important. However the current retrieval methods are essentially based on the full text matching of keywords approach lacking of semantic information and can't understand the user's query intent very well. These methods return a large number of irrelevant information, and are unable to meet the user's request. Systems have been established so far failed to overcome fully the limitations of search based on keywords. Such systems are built from variations of classic models that represent information by keywords. Using Semantic Web is a way to increase the precision of information retrieval systems.**

In this paper, we propose the semantic information retrieval approach to extract the information from the web documents in certain domain (jaundice diseases) by collecting the domain relevant documents using focused crawler based on domain ontology, and using similar semantic content that is matched with a given user's query. Semantic retrieval approach aims to discover semantically similar terms in documents and query terms using WordNet.

Keywords- Semantic Web; information retrieval; Semantic Retrieval; Semantic Similarity; WordNet.

I. INTRODUCTION

In the current information retrieval models and systems, only when query words appear in document, the document may be retrievable. As a result, it probably appears that the related documents may be omitted because of expression difference. This kind of problem is one of the important reasons that influences the accuracy of information retrieval. Traditional text-based information retrieval systems and search engines are mostly based on keywords matching and statistic techniques [57]. Although they are widely used nowadays, users usually suffer from the so called "too many or nothing" problem for various reasons. One common reason is that the users may not have complete domain knowledge and often can not specify appropriate and exact keywords for a valid query. The other reason is that the target documents are expressed in terms of plain-text format that is hard for the search engine to parse, thus it is difficult to understand the semantic of the documents during the retrieval process.

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 information retrieval and the Semantic Web communities. However, these two fields have had a different understanding of the problem. The Semantic Web vision [54] was brought about the aim of helping automate tasks that require a certain level of conceptual understanding of the objects involved or the task itself, and enabling software programs to automatically find and combine information and resources in consistent ways.

At the core of these new technologies, ontologies were envisioned as key elements to represent knowledge that could be understood, used and shared among distributed applications and agents. Their potential to overcome the limitations of keyword-based search in the information retrieval context was soon envisaged, and was explored by several researchers in the Semantic Web area. A potential source of documents that is useful for information retrieval comes from the World Wide Web, it is important to develop document discovery mechanisms based on intelligent techniques such as focused crawling [55] to make this process easier for a certain domain.

In our work, we have used focused crawling to collect documents and information in a healthcare domain (jaundice diseases). 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. Our focus will be to identify Web pages for our domain in WWW. We present a critical semantic similarity approach for computing the semantic similarity between the terms 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.

II. RELATED WORKS

In [1], authors introduced the technique of focused crawling of country based financial data. The focused crawlers yield good recall as well as good precision by restricting themselves to a limited and selected domain. The focused crawlers try to predict whether or not a target URL is pointing to a relevant and high-quality web page before actually fetching the page. Their efficient focused crawler is made for collecting the financial data for a specific country.

The approach [2] was proposed to calculate the link score. First authors calculated the unvisited URL score based on its Anchor text relevancy, its description in Google search engine and calculated the similarity score of description with topic keywords, cohesive text similarity with topic keywords and Relevancy score of its parent pages. Relevancy score is calculated based on vector space model.

In the paper [3], authors explored four kinds of semantic models and semantic information to improve focused crawling, including thesauruses, categories, ontologies, and folksonomies. Main contributions of this work are : First, A statistical semantic association model to integrate different semantic models and support semantic interoperability. Second , Include added semantic information to improve focused crawling, especially semantic markups in the Semantic Web and social annotations in Web 2.0. Third, the Semantic Association Model(SAM) that is based focused crawler which adopts heterogeneous semantic information to make predictions and decisions about relevant URLs and web pages.

In the study [4], authors analyzed four focused crawling for retrieving chemical information. These focused crawlers were formed by combining two feature representations (Latent Semantic Indexing (LSI) and Mutual Information (MI)) and two classification algorithms (Support Vector Machines (SVM) and Naive Bayes (NB)) from machine learning. The study shows that the four focused crawling can keep a high precision to collect chemistry relevant pages. It was also found that the combination of SVM and LSI provided the best performance in gathering web pages on the topic of chemistry.

Authors in [5] proposed an algorithm for crawling Web pages with limited resources. First, existing Web pages are divided into 100 clusters using both static and dynamic features. Static features are extracted from the content of a Web page, the Web page's URL and hyperlinks. Dynamic features are extracted from changes to content, hyperlinks, page rank, and so on. The crawler fetches a sample of Web pages from a cluster to check if the pages have changed since their last download. If a significant number of pages in a cluster have changed, the other Web pages in the cluster are also downloaded. Clusters of Web pages have different change frequencies. Based on their change histories, different clusters is crawled at different frequencies. They demonstrated the superiority of their algorithm over various existing sampling-based Web page update detection algorithms.

In the paper [6] authors introduced document vector compression, which significantly reduces the size of the vectors and increases the total F-measure (cluster quality). Document vector compression involves the use of the Discrete Cosine Transform(DCT) on the document vectors to obtain a spectral representation of the document vectors. Due to the energy compaction property of the DCT, the majority of the energy is concentrated at the low frequency subbands. The high frequency subbands can be deleted without significantly degrading the content of the original vector. Any standard clustering algorithm, such as K-means , was used to cluster the compressed document vectors.

In this approach [7], vector space is taken as an example to describe the construction process of the document representation model based on query and content information. The basic idea is as following: At the initial stage of information retrieval, the traditional vector space model is adopted to represent documents. Then, the information of the query space can be introduced into the document representation model gradually, thus the document-representing vector space becomes the integration of query space and document space. This model can improve the fitness, reliability and accuracy of the feature terms of documents.

The approach [8] is for representing text data. The method translated the text clustering problem into query processing. The intuition behind this approach is if a set of documents belongs to the same cluster, authors expected that they respond similarly to the same queries, which can be any combination of terms from the vocabulary. While in information retrieval, the target is to retrieve relevant document(s) to a query, in text clustering, the goal is finding relevant queries which generates high quality clusters (lowest inter-cluster and highest intra-cluster similarities). In this paper, authors proposed approach to generate relevant and non-redundant queries from the domain taxonomy which is extracted from document collection. Using this new model, the terms in BOW model are transformed to the similarity scores of Bag-Of-Queries (BOQ) model. The effectiveness of the proposed approach is evaluated by extensive numerical experiments using benchmark document data set.

It was proposed [9] to use 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. This is in contrast to previous WordNet-based work which focused on WSD to replace or supplement words with their senses. The method discovered important concepts, even if they are not explicitly mentioned in the document.

The goal in the work [10] was to study the use of the WordNet expansion technique over a collection with minimal textual information. The integration of knowledge through the use of ontologies has been very successful in many systems. Specifically, WordNet has been used with success in many works related to information retrieval, image retrieval, disambiguation and text categorization.

In the study [11], authors developed a new information retrieval system integrating Semantic Web with Multi-agent that handles the processing, recognition, extraction, extensions and matching of content semantics to achieve the following objectives: (1) Using Resource Description Framework (RDF) to analyze and determine the semantic features of users' queries, to present a new algorithm to extract semantics in the

content and build up semantic database; (2) to present a new matching algorithm using semantics extracted from content which can feedback useful and accurate information meeting users' requirements; (3) a new Information Retrieval based on Multi-agent is put forward, the Agents in this model can adapt users' own interests and hobbies, collect information based on users' behavior, dig up semantics in internet and feedback and share information between different users, so the search results will be more in line with users' needs and help users to complete complex tasks.

Authors in [12] introduced ontology into query expansion and makes good use of semantic relations of concepts in ontology to expand query keywords and to make the retrieval results more accuracy and comprehensive. Experimental results showed that this method can improve the precision and recall ratios of information retrieval.

The method proposed in [13] focused on semantic based expansion. There are three important improvements in the query expansion. First of all, this method categorizes the query terms based on their semantic similarities, and expands each category on words which show the relationship between words in the same group, as a result in this method selected words are not related to only an individual query term. Therefore it avoids outweighting problem in query expansion. Secondly, it avoids selecting vague and noise words to expand the query. Thus it avoids making the query noisy. Thirdly, it uses spreading activation algorithm to select candidate expansion words. Using spreading activation algorithm eases the selection of appropriate depth for hierarchical relations.

Authors in [14] proposed a new semantic similarity based model (SSBM) and they used this model in document text clustering. The model analyzed a document to get the semantic content. The SSBM assigns new weights to reflect the semantic similarities between terms. Higher weights are assigned to terms that are semantically close. In this model, each document was analyzed to extract terms considering stemming and pruning issues. They used the adapted Lesk algorithm to get the semantic relatedness for each pair of terms. SSBM solved the ambiguity and synonym problems that may lead to erroneous grouping and unnoticed similarities between text documents.

III. The Proposed Framework

Ontologies 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.

Due to the tremendous size of information on the Web, it is increasingly difficult to search for useful information for certain domain. For this reason, it is important to develop document discovery mechanisms based on intelligent techniques such as focused crawling. In a classical sense, crawling is one of the basic techniques for building data storages. Focused crawling goes a step further than the classical approach. It was proposed to selectively seek out pages relevant to a predefined set of topics called crawling topics.

In order to leave a lot of irrelevant noisy pages out, we propose an ontology-based focused crawling framework for Web. Crawling topics are based on our domain ontology. We focus on building an effective web-based documents discovery crawler that can autonomously discover and download pages from the web relevant to our domain, that is jaundice diseases. This is considered as semantic-based focused crawling, it makes use of an ontology to improve decision accuracy. The figure 1 shows the architecture of the proposed system.

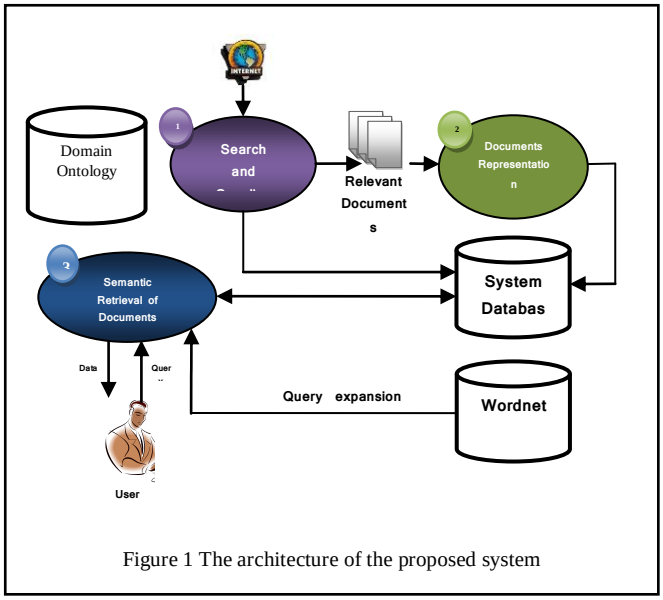

Figure 1 The architecture of the proposed system

A. The Domain Ontology

The term "Ontology" [21] is becoming frequently used in many contexts of database and artificial intelligence researches. However, there is not a unique definition of what an ontology is [22,23]. An initial definition was given by Tom Gruber: "an ontology is an explicit specification of a conceptualization" [22]. However, this definition is general and remains still unsatisfied for many researchers. In [24] Nicola Guarino argues that the notion of "conceptualization" is badly used in the definition. We note that many real-world ontologies already combine data instances and concepts [25]. The definition in [21] differs from this point of view as we show later . Informally, an ontology is defined as an intentional description of what is known about the essence of the entities in a particular domain of interest using abstractions, also called concepts and the relationships among them.

Ontologies [26] 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. Because of the intrinsic complexity of the concepts involved, the medical domain is one of the most active ones in defining and using ontologies.

The ontology in our system is focused in the medical domain that is "Jaundice diseases". Jaundice [27], is a yellowing of the skin, conjunctiva (clear covering over the sclera, or whites of the eyes) and mucous membranes caused by increased levels of bilirubin in the human body. When red

blood cells die, the heme in their hemoglobin is converted to bilirubin in the spleen and in the hepatocytes in the liver. The bilirubin is processed by the liver, enters bile and is eventually excreted through feces. Consequently, there are three different classes of causes for jaundice. Pre-hepatic or hemolytic causes, where too many red blood cells are broken down, hepatic causes where the processing of bilirubin in the liver does not function correctly, and post-hepatic or extrahepatic causes, where the removal of bile is disturbed. Figure 2 shows part of our ontology for "Jaundice diseases". Jaundice diseases [28] are divided into three types as shown, as it follows: Pre-hepatic jaundice is caused by anything which causes an increased rate of hemolysis (breakdown of red blood cells). Hepatic (in hepatocellular jaundice there is invariably cholestasis) jaundice causes include acute hepatitis, hepatotoxicity, Gilbert's syndrome. Post-hepatic jaundice, also called obstructive jaundice, is caused by an interruption to the drainage of bile in the biliary system. The most common causes are gallstones in the common bile duct, and pancreatic cancer in the head of the pancreas.

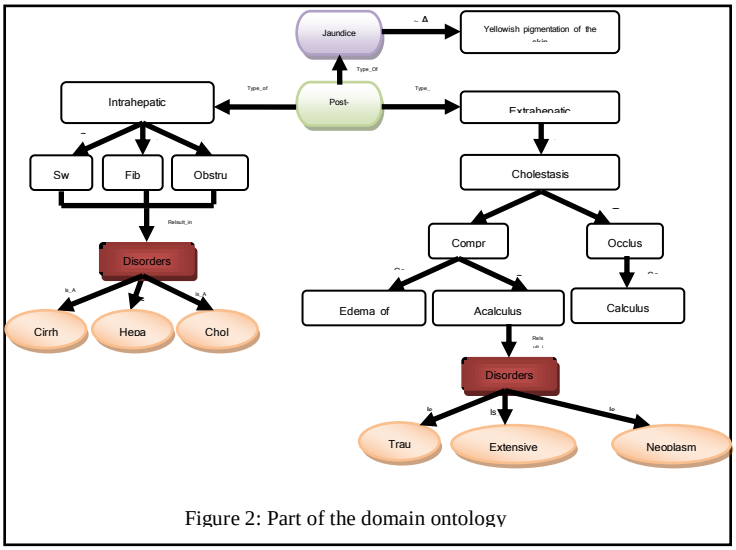

Figure 2: Part of the domain ontology

Our domain ontology is represented in Web Ontology Language (OWL) [29]. The Web Ontology Language (OWL) [30] describes classes, properties, and relations among these conceptual objects in a way that facilitates machine interpretability of Web content. OWL is the result of the Web Ontology Working Group (now closed) and descends from DAML+Oil, which is in turn an amalgamation of DAML and OIL.

B. Search and Crawling Web

A web crawler [15] is a program that collects web content from the World Wide Web automatically and stores this content into the storage. It starts with a list of URLs (seeds) to be visited. When it visits these pages, it parses all links from these pages. After collecting all these links, the web crawler inserts them in the URL queue. Web crawler continuously visits the unseen links and also scans them for discovering more links and put only the unseen links in the URL queue. The basic steps in the crawler [15] are:

1. Fetch a page
2. Parse it to extract all linked URLs
3. For all the URLs not seen before, repeat (1)–(3).

One of the features that characterizes a focused crawler [16] is the way it exploits hyper-textual information. Traditional crawlers convert a web page into plain text extracting the contained links, which will be used to crawl other pages. Focused crawlers exploit additional information from Web pages, such as anchors or text surrounding the links. Semantic focused crawlers [17, 18, 19] are considered as a subset of focused crawlers enhanced by various semantic web technologies. Table 1 shows different category of semantic focused crawlers and its definitions.

TABLE I. DIFFERENT CATEGORY OF SEMANTIC FOCUSED CRAWLERS AND ITS DEFINITION

Crawler category	Definition
Ontology-based focused crawlers	The focused crawlers that utilize ontologies to link a crawled web document with the ontological concepts (topics), with the purpose of organizing and categorizing web documents, or filtering irrelevant web pages with regards to the topics.
Metadata abstraction focused crawlers	The focused crawlers that can abstract and annotate metadata from the fetched web documents, in addition to fetching relevant documents.
Other semantic focused crawlers	The focused crawlers that employ other semantic web technologies than ontology-based filtering and metadata abstraction.

1) Search and Crawling Web approach

In our approach, we crawl through the Web and add Web pages contents to the database, which are related to a specific domain and discard Web pages which are not related to the domain [20]. The block diagram of the search and web crewing is shown in figure 3.

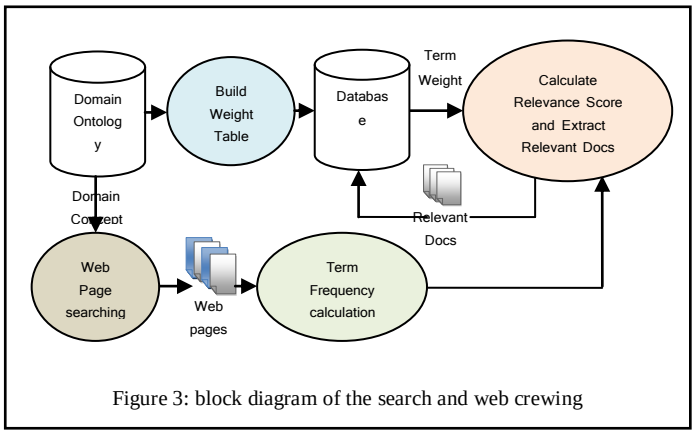

Figure 3: block diagram of the search and web crewing

2) Building the weight table

In order to build the weight table, we must determine some weights to each term in our ontology The strategy of assigning weights is that, the more specific term will have more weight

on it. The terms which are common to more than one domain have less weight. The sample weight table for some terms of a given ontology is shown in table 2. The weight table is built with the assistance of the knowledge experts.

3) Relevance Score Calculation

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 we automatically remove those that are blank, too short, duplicated documents, or those that are in a format that is not suitable for text processing. We then perform the relevance calculation to extract the relevant documents and discard the irrelevant document to our domain.

In the relevance calculation, the relevancy of a Web page to a specific domain is calculated. Relevance calculation algorithm [20], which calculates the relevance score of a Web page, is shown in figure 4.

TABLE II. WEIGHT TABLE FOR THE PART OF ONTOLOGY

Concepts	Weight
Jaundice	1
Biliruin	1
Post-hepatic	0.9
Prehepatic	0.9
Hepatic	0.9
Hepatitis	0.8
Cholestasis	0.8
Thalassemia	0.7
Damaged Hepatocytes	0.6
Tumours	0.4
Trauma	0.2
Compression	0.1
Disorder	0.1

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

In our approach we go along the links that are found in domain specific pages to crawl the result web page. The figure 5 shows the crawling algorithm that is based on the relevance of domain ontology.

Input: Crawling Level (C_L), Start URL list (U_{List}) from search engine, Relevance threshold (R_T).
Output: The Relevant Web pages (P).
Procedure:
Do While $i <= CU_{List}$ // CU_{List} : the count of URL in the U_{List}.
Get URL.
Do While $j <= C_L$ //Crawling until the Crawling Level
Do Crawling Function.
Get Web Page.
Calculate relevance score of the Web page RS_P.
If ($RS_P > RT$) // if the page is relevant to the domain ontology and take high relevance score.
Get the Page content.
Save the content in the database.
Else // if the page is not relevant to the domain ontology and take high relevance score.
Discard the Page
Endif
EndDo
EndDo

Figure 5 : Algorithm of crawler based on domain relevance

C. Documents Representation

In information retrieval, the most widely accepted document representation model in text classification is probably vector space model [31]. The Vector Space Model is adapted in our 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:

$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$ Where, w_{kj} is the weight of the k_{th} term in the document j. Figure 5 shows the approach for representation and classification of the output documents of web pages.

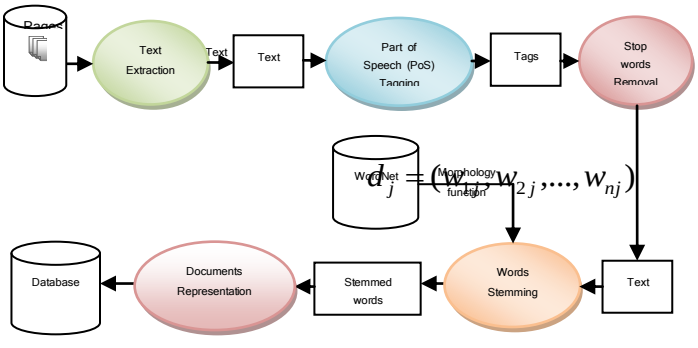

Figure 5 : Approach of documents

The first step of the documents classification process is to extract 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, we used the open source high-performance .NET C# module that was created to parse HTML [32] for links, indexing and other purposes.

2) Part of Speech (PoS) Tagging

The PoS tagger [33] relies on the text structure and morphological differences to determine the appropriate part-of-speech. This requires the words to be in their original order. This process is to be done before any other modifications on the corpora. For this reason, if it is required, PoS tagging is the first step to be carried out. After this, stop word removal is performed, followed by stemming. This order is chosen to reduce the amount of words to be stemmed. We used Stanford POS Tagger to tag the tokens [34].

The output of this step is the tags of each tokens. For example: "*Jaundice is yellowish pigmentation of the skin*", when this sentence is passed to the POS Tagger. The output is the tagged text as follows: "*Jaundice/NNP is/VBZ yellowish/JJ pigmentation/NN of/IN the/DT skin/NN*". The table 3 shows the different symbols that is found in PoS tagging.

TABLE III. SYMBOLS OF POS TAGGING

CC	Coordinating conjunction	**IN**	Preposition or subordinating conjunction	**MD**	Modal
CD	Cardinal number	**JJ**	Adjective	**NN**	Noun, singular or mass
DT	Determiner	**JJR**	Adjective, comparative	**NP**	Proper noun singular
EX	Existential there	**JJS**	Adjective, superlative	**NPS**	Proper noun plural
FW	Foreign word	**LS**	List item marker	**PDT**	Predeterminer
POS	Possessive ending	**RBR**	Adverb, comparative	**TO**	To
PP	Personal pronoun	**RBS**	Adverb, superlative	**UH**	Interjection
PP$	Possessive pronoun	**RP**	Particle	**VB**	Verb, base form
RB	Adverb	**SYM**	Symbol	**VBD**	Verb, past tense
VBG	Verb, gerund or present participle	**VBP**	Verb, noun-3rd person singular present	**WDT**	Wh-determiner
VBN	Verb, past participle	**VBZ**	Verb, 3rd person singular present	**WP**	Wh-pronoun
WP$	Possessive wh-pronoun	**WRB**	Wh-adverb		

3) 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 with PoS information 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.

4) 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, we used the morphology function [35] provided with WordNet [36, 37] that is used for stemming process. Stemming will find the stems of the output terms to enhance term frequency counting process because terms like “diseases” and “diagnosing” come down from the same stem “disease” and "diagnose". This process will output all the stems of extracted terms [38, 39].

5) Documents Representation

As part of the key vocabulary extraction process from documents, tf × idf (term frequency times inverted document frequency), takes place. Terms (T_k) in the documents is represented as the document-term frequency matrix ($D_j \times TF_{jk}$) as shown in figure 6.

D_j/T_k	T_1	T_2	....	T_k
D_1	TF_{11}	TF_{12}	....	TF_{1m}
D_2	TF_{21}	TF_{22}	....	TF_{2m}
D_3	TF_{31}	TF_{32}	....	TF_{3m}

Figure 6 : The document-term frequency matrix

D_j is referring to each document that exists in the system database where j=1,….,n. Term frequency TF_{jk} is the number of how many times the distinct term T_k occurs in document D_j where k=1,…,m.

The calculation of the terms weight W_{jk} of each term T_k is done by [40, 41, 42]:

$$W_{jk} = TF_{jk} \times idf_k \quad \text{.....(1)}$$

where the document frequency df_k is the total number of documents in the database that contains the term T_k. The inverse document frequency is :

$$idf_k = \log_2 n - \log_2 df_k + 1 \quad \text{.....(2)}$$

where (n) is the total number of documents in the database.

$tf \times idf$ is a mathematical algorithm [40, 41, 42], which is used to efficiently find key vocabulary that best represents the

texts by applying the term frequency and the inverted document frequency together. $tf(T_k, D_j)$ is the term frequency of term T_k that appears in Document D_j, and (n) is the total number of documents of the corpus. $df(T_k)$ is the number of the documents in which the Term T_k appears at least once and represents how often Term T_k appears in other documents. $tf \times idf$ for Term T_k is defined as:

$$tf \times idf(T_k, D_j) = tf(T_k, D_j) \times \log\left(\frac{n}{df(T_k)}\right) \quad \text{.....(3)}$$

For vocabulary with a low or rare appearance frequency, the value of $tf \times idf$ is low, compared to that with a high appearance frequency, thus resulting words successfully classifying the documents. In the term selection process, a list of all terms contained in one document from the text collection is made. Then, the document selection process chooses term T_k that maximizes *W(k),* which is expressed as a vector for document D_j as follows.

Document D_j includes $tf \times idf(T_k, D_j)$, which is $tf \times idf$ for the most appropriate term.

$$W(k) = \sum_{j=1}^{n} tf \times idf(T_k, D_j) \quad \text{.....(4)}$$

D. Semantic Retrieval of Documents

Semantic retrieval [43] 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 will get a result set which is sorted according to semantic similarity.

Nowadays semantic retrieval mainly implements concept retrieval [44, 45] by interaction terms, which does not take the concept's attributes and other information in to consideration. This semantic retrieval method based on concepts often cannot meet practical requirements.

So, we organize concepts with ontology, calculating the semantic similarity between concepts, whose basis is that there are some semantic correlations between two concepts. There are several semantic similarity methods were used which have certain limitations despite the advantages. No one method replaces all the semantic similarity methods. When a new information retrieval system is going to be build, several questions arises related to the semantic similarity matching function to be used. In [46] authors discussed the survey of different similarity measuring methods used to compare and find very similar concepts of an ontology.

In our approach, we depend on the semantic similarity based on Wordnet [47]. Five commonly used semantic similarity measures based on WordNet are discussed in [48, 49]. In [50] the authors conducted a comparative study on how different term semantic similarity measures including path-based, information content-based and feature-based similarity measure affect document clustering. WordNet is a controlled vocabulary and thesaurus offering a taxonomic hierarchy of natural language terms developed at Princeton University [56].

Figure 7 shows the block diagram of information retreieval of documents using semantic similarity between query and documents data.

Query expansion refers to the process of adding new necessary terms to a user's initial query. The purpose is to improve retrieval performance Query expansion 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. With query expansion [51], the user is guided to formulate queries which enable useful results to be obtained. The main aim of query expansion [52, 53] (also known as query augmentation) is to add new meaningful terms to the initial query. Our approach uses query expansion that computes good weights for the new terms introduced into the query by using semantic similarity based on wordnet. Queries is first syntactically analyzed and reduced into term vectors as performed in documents in section 3.3.5. Very infrequent or very frequent terms are eliminated. Each term in this vector is represented by its weight.

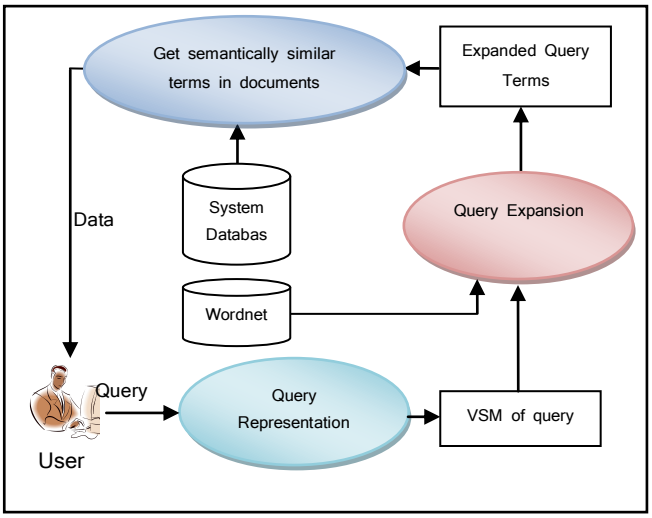

Figure 7 Semantic Retrieval of Documents

The query is augmented by synonym terms in wordnet, using the most common sense of each query term. Then, the query is augmented by terms higher or lower in the tree hierarchy in XML file which are semantically similar to terms already in the query.

The neighborhood of the term is examined and all terms with similarity greater than threshold t are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term.

Each query term i is assigned a weight as in formula 5 in which the summation is taken over all terms j that is introducing terms to the query. It is possible for a term to introduce terms that already existed in the query. It is also possible that the same term is introduced by more than one

other terms. After expansion and reweighting, the query vector is normalized by document length, like each document vector.

$$q_i' = \begin{cases} q_i + \sum_{sim(i,j)\geq t}^{j\neq i} \frac{1}{n} q_j \; sim(i,j), & i \; had \; weight \; q_i \\ \sum_{sim(i,j)\geq t}^{j\neq i} \frac{1}{n} q_j \; sim(i,j), & i \; is \; a \; new \; term \end{cases} \quad \text{.....(5)}$$

where t is a user defined threshold, n is the number of hyponyms of each expanded term j (for hypernyms n will be equal 1). The algorithm of the semantic retrieval of documents is shown in figure 8.

IV. CONCLUSION

We have proposed the approach of semantic retrieving for web documents in certain domain, extracting relevant information based on the semantic web. We have studied and implemented a focused crawler enabling us to retrieve web documents in the domain of jaundice diseases from the Web.

Semantic information retrieval method have exploited the advantages of the semantic web to retrieve the relevant data. It outperforms VSM, the classic information retrieval method and demonstrates promising performance improvements over other semantic information retrieval methods in retrieval.

Input: Query Vector $q = (q_1, q_2, ..., q_i)$, Document Vector $d = (d_1, d_2, ..., d_K)$, Thresholds T.
Output: Document similarity value $Sim(d, q)$.
Procedure:

// --- Get the semantic similarity of each term with another in the query of the same vector.
Do While $i <= q_{count}$ // q_{count} : the count of terms in the query vector q.
Get q_i.
Compute $q_i' = q_i + \sum_{sim(i,j)\geq t}^{j\neq i} q_j \; sim(i,j)$
EndDo

// --- Get Expanded terms based on wordnet.
Do While $i <= q_{count}$
Get q_i.
Open Wordnet.
If $sim(i,j) >= T$ // $sim(i,j)$: similarity between the term i in the query and term j in the wordnet.
Retrieve the term j from wordnet.
Add term j to the query vector q.
Endif
EndDo

// --- Re-weighting the terms in the query.
Do While $i <= q_{count}$

$$q_i' = \begin{cases} q_i + \sum_{sim(i,j)\geq t}^{j\neq i} \frac{1}{n} q_j \; sim(i,j), & i \; had \; weight \; q_i \\ \sum_{sim(i,j)\geq t}^{j\neq i} \frac{1}{n} q_j \; sim(i,j), & i \; is \; a \; new \; term \end{cases}$$

EndDo

// --- Compute the document similarity (The similarity between an expanded and re-weighted query q and a document d).
Do While $i <= q_{count}$

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

EndDo

Figure 8 : Algorithm of the semantic retrieval of documents

REFERENCES

[1] D. Manas, C. Hasan, S. Debakar, A. Khandakar. (2010). Focused Web Crawling: A Framework for Crawling of Country Based Financial Data. 978-1-4244-6928-4/10, IEEE.

[2] H. Debashis, S. Biswajit, K. Amritesh. (2010). Adaptive Focused Crawling Based on Link Analysis. 201O 2nd International Conference on Education Technology and Computer (ICETC). 978-1-4244-6370-1, IEEE.

[3] H. Rui, L. Fen, S. Zhongzhi. (2008). Focused Crawling with Heterogeneous Semantic Information. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-3496-1/08, IEEE.

[4] X. Zhaojie, G. Li, L. Chunyang, L. Xiaoxia, and Y. Zhangyuan. (2007). Focused Crawling for Retrieving Chemical Information. Innovations in Hybrid Intelligent Systems, ASC 44, pp. 433–438, Springer-Verlag Berlin Heidelberg.

[5] T. QINGZHAO and M. PRASENJIT. (2010). Clustering-Based Incremental Web Crawling. 2010 ACM 1046-8188/2010/11-ART17.

[6] T.W. Fox. (2005). Document Vector Compression and Its Application in Document Clustering. 0-7803-8886-0/05, IEEE.

[7] Z. ZHOU, H. JIANG, J. MA, X. YANG. (2010). Study on Application of Document Representation Model Based on Query and Content Information in Website Search Engine. 2010 International Conference on Web Information Systems and Mining, 978-0-7695-4224-9/10, IEEE.

[8] M. Masoud. (2010). Query-Relevant Document Representation for Text Clustering. 978-1-4244-7571-1/10, IEEE.

[9] A. Eneko, A. Xabier, O. Arantxa. (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.

[10] D. Manuel, M. Maria, U. L. Alfonso, and P. Jose. (2010). Using WordNet in Multimedia Information Retrieval. CLEF 2009 Workshop, Part II, LNCS 6242, pp. 185–188, Springer-Verlag Berlin Heidelberg.

[11] J. LUO, X. XUE. (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.

[12] W. Hongsheng, Q. Jiuying, S. Hong. (2009). Expansion Model of Semantic Query Based on Ontology. Web Mining and Web-based Application. WMWA '09. IEEE.

[13] M. Shabanzadeh, M.A. Nematbakhsh, and N. Nematbakhsh. (2010). International Conference on Intelligent Control and Information Processing August 13-15, 2010 - Dalian, China. 978-1-4244-7050-1/10,IEEE.

[14] G. Walaa and K. Mohamed. (2009). Enhancing Text Clustering Performance Using Semantic Similarity. LNBIP 24, pp. 325–335. Springer-Verlag Berlin Heidelberg.

[15] [D. Manas, C. Hasan, S. Debakar, A. Khandakar. (2010). Focused Web Crawling: A Framework for Crawling of Country Based Financial Data, 978-1-4244-6928-4/10, IEEE.

[16] M. Alessandro and G. Fabio. (2007.). Adaptive Focused Crawling. The Adaptive Web, LNCS 4321, pp. 231–262, 2007, Springer-Verlag Berlin Heidelberg .

[17] D. Hai, K. Farookh, and C. Elizabeth. (2009). State of the Art in Semantic Focused Crawlers. O. Gervasi et al. (Eds.): ICCSA 2009, Part II, LNCS 5593, pp. 910–924, Springer-Verlag Berlin Heidelberg.

[18] D. Hai, H. Farookh, C. Elizabeth. (2008). A Survey in Semantic Web Technologies-Inspired Focused Crawlers. 978-1-4244-2917-2/08, IEEE.

[19] D. Hai Dong, K. Farookh. (2010). Focused Crawling for Automatic Service Discovery. Annotation and Classification in Industrial Digital Ecosystems, Industrial Electronics, IEEE Transactions on Volume : PP , Issue:99 .

[20] M. Debajyoti, B. Arup, S. Sukanta. (2007). A New Approach to Design Domain Specific Ontology Based Web Crawler. 0-7695-3068-0/07, IEEE.

[21] S. Wei, L. Da-Xin. (2006), Using Ontologies for Semantic Query Optimization of XML Database, R. Nayak and M.J. Zaki (Eds.): KDXD

2006, LNCS 3915, pp. 64 – 73, 2006, Springer-Verlag Berlin Heidelberg.

[22] Gruber, T. (1993), A translation approach to portable ontology specifications. In: Knowledge Acquisition, 5(2) ,199-220.

[23] Chandrasekaran, B., Josephson, J., Benjamins, V. (1999), What are ontologies, and why do we need them? In: IEEE Intelligent Systems, 20-26.

[24] Guarino, N., Giaretta, P. (1995), Ontologies and knowledge bases: towards a terminological clarification. In: Knowledge Building Knowledge Sharing, ION Press. 25-32.

[25] Noy, N., Hafner, C.D. (1997), The state of the art in ontology design. AI Magazine 3, 53-74.

[26] V. Aida, G. Karina, S. David, B. Montserrat, (2010), Using ontologies for structuring organizational knowledge in Home Care assistance, international journal of medical informatics 79 (2010) 370–387, Elsevier Ireland Ltd.

[27] http://www.medic8.com/healthguide/articles/jaundice.html.

[28] http://en.wikipedia.org/wiki/Jaundice.

[29] http://www.w3.org/TR/owl-ref/

[30] K.K. Breitman, M.A. Casanova and W. Truszkowski. (2007), OWL, In: " Semantic Web: Concepts, Technologies and Applications ", Springer-Verlag London Limited.

[31] L. Ying. (2009). On Document Representation and Term Weights in Text Classification. IGI Global.

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

[33] S. Julian, K. Dimitar. (2004). WordNet-based Text Document Clustering. ROMAND '04 Proceedings of the 3rd Workshop on RObust Methods in Analysis of Natural Language Data. ACM.

[34] The Stanford Natural Language Processing Group. http://nlp.stanford.edu/software/tagger.shtml.

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

[36] 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.

[37] F. Christiane. (2010). WordNet. In: R. Poli et al. Theory and Applications of Ontology: Computer Applications, (pp. 231-243). 231-243, DOI: 10.1007/978-90-481-8847-5_10, Springer Science+Business Media B.V.

[38] G.Tarek, F. Mohammed, A. Mostafa. (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.

[39] G.Tarek, F. Mohammed, A. Mostafa. (2010). Fuzzy Document Clustering Approach using WordNet Lexical Categories. K. Elleithy (ed.), Advanced Techniques in Computing Sciences and Software Engineering, DOI 10.1007/978-90-481-3660-5_31, Springer Science+Business Media.

[40] S. Mu-Hee, L. Soo-Yeon, K. Dong-Jin, L. Sang-Jo. (2005). Automatic Classification of Web Pages based on the Concept of Domain Ontology. Proceedings of the 12th Asia-Pacific Software Engineering Conference (APSEC'05). 0-7695-2465-6/05, IEEE.

[41] S. Mu-Hee, L. Soo-Yeon, K. Dong-Jin, L. Sang-Jo. (2006). Ontology-Based Automatic Classification of Web Documents. D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNAI 4114, pp. 690 – 700, Springer-Verlag Berlin Heidelberg.

[42] S. Mu-Hee, L. Soo-Yeon, K. Dong-Jin, L. Sang-Jo. (2006). Ontology-Based Automatic Classification of Web Pages. Advances in Intelligent and Soft Computing, Volume 34/2006, 483-493, DOI: 10.1007/3-540-31662-0_37. Springer-Verlag Berlin Heidelberg.

[43] W. Hongsheng, H. Xiaoguang. (2010). Research on Similarity of Semantic Web. International Conference on Computer Application and System Modeling (ICCASM 2010). 978-1-4244-7237-6/10, IEEE.

[44] Q. Sheng, G. Ying. (2008). Measuring Semantic Similarity in Ontology and Its Application in Information Retrieval. Congress on Image and Signal Processing. 978-0-7695-3119-9/08, IEEE.

[45] L. Gang, Z. Cheng, Z. Li. (2009). Text Information Retrieval Based on Concept Semantic Similarity. Fifth International Conference on Semantics, Knowledge and Grid. 978-0-7695-3810-5/09, IEEE.

[46] K.Saruladha, G.Aghila, R. Sajina. (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.

[47] S. Michel, M. Radja, D. Gayo, S. Ana. (2010). Ontologies in the Health Field. In: Data Mining and Medical Knowledge Management: Cases and Applications. Pages: 37-56 pp. 10.4018/978-1-60566-218-3.ch002. IGI Global.

[48] L. Haisheng, T. Yun, Y. Ben, C. Qiang. (2010). Comparison of Current Semantic Similarity Methods in WordNet. International Conference on Computer Application and System Modeling (ICCASM 2010). 978-1 - 4244-7237-6/10, IEEE.

[49] H. Tran and S. Dan. (2008). Word Similarity In WordNet. Modeling, Simulation and Optimization of Complex Processes, 293-302, DOI: 10.1007/978-3-540-79409-7_19. Springer-Verlag Berlin Heidelberg.

[50] Z. Xiaodan, J. Liping, H. Xiaohua, N. Michael, X. Jiali, Z. Xiaohua. (2010). Medical Document Clustering Using Ontology-Based Term Similarity Measures. In: International Journal of Data Warehousing and Mining (IJDWM). Pages: 62-73 pp. IGI Global.

[51] J. Bhogal, A. Macfarlane, P. Smith. (2006). A review of ontology based query expansion. Information Processing and Management 43 (2007) 866–886. Elsevier Ltd.

[52] V. Giannis, V. Epimenidis, R. Paraskevi. (2005). Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. WIDM'05, November 5, Bremen, Germany. ACM.

[53] H. Angelos, V. Giannis, V. Epimenidis, P. Euripides, M. Evangelos. (2009). Information Retrieval by Semantic Similarity. In: Medical Informatics: Concepts, Methodologies, Tools, and Applications. Page: 647-665 pp. IGI Global.

[54] F. Miriam, C. Iván, L. Vanesa, V. David, C. Pablo, M. Enrico. (2010). Semantically enhanced Information Retrieval: An ontology-based approach. Web Semantics: Science, Services and Agents on the World Wide Web, doi:10.1016/j.websem.2010.11.003. Elsevier B.V.

[55] L. Hiep, G. Susan, W. Qiang. (2009). Ontology Learning Through Focused Crawling and Information Extraction. 2009 International Conference on Knowledge and Systems Engineering. 978-0-7695-3846-4, IEEE.

[56] F. Christiane. (2010) . WordNet. Theory and Applications of Ontology: Computer Applications, DOI 10.1007/978-90-481-8847-5_10, C_ Springer Science+Business Media B.V.

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

AUTHORS PROFILE

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 Al-Azhar 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.

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.

Nagdy M. Nagdy is professor of engineering applications and computer systems, Department of Systems Engineering and Computer Engineering - Faculty of Engineering Al-Azhar 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 Enginee