Ranking Documents Based on the Semantic Relations Using Analytical Hierarchy Process

Ali I. El-Dsouky
Computers and Systems Department,
Faculty of Engineering,
Mansoura University,
Egypt

Hesham A. Ali
Computers and Systems Department,
Faculty of Engineering,
Mansoura University,
Egypt

Rabab S. Rashed
Electrical Engineering Department,
Faculty of Engineering,
Kafr elsheikh University,
Egypt

***Abstract*—With the rapid growth of the World Wide Web comes the need for a fast and accurate way to reach the information required. Search engines play an important role in retrieving the required information for users. Ranking algorithms are an important step in search engines so that the user could retrieve the pages most relevant to his query.**

In this work, we present a method for utilizing genealogical information from ontology to find the suitable hierarchical concepts for query extension, and ranking web pages based on semantic relations of the hierarchical concepts related to query terms, taking into consideration the hierarchical relations of domain searched (sibling, synonyms and hyponyms) by different weighting based on AHP method. So, it provides an accurate solution for ranking documents when compared to the three common methods.

***Keywords*—Semantic rank; ranking web; ontology; search engine; information retrieval**

I. INTRODUCTION

Web based information retrieval systems; especially search engines are the basic tools to assist users to find information on the World Wide Web. Despite the vital role in reaching information, many of the returned results are irrelevant to the user's needs as they are ranked based on the string matching of the user's query. This has created a semantic gap between the meanings of the keywords in the retrieved documents and the meanings of the terms used in users' queries.

Search is the most popular applications on the Web. The bulk of traditional retrieval systems usually make use of metadata keywords matching with the query. However, these systems don't take into account the semantic relationships between query terms and other concepts that might be significant to users. Thus, the addition of explicit semantics can improve the search process. Semantic search is an application of the Semantic Web to search. It tries to improve traditional search results (based on Information Retrieval technology) using data from the Semantic Web [1]. This approach offers an enhancement to traditional search as it allows retrieval to incorporate the underlying terms semantics [2]. It improves the traditional search that focuses on word frequency by trying to understand hidden meanings in the retrieved documents and users' queries [3, 4]. The problem of poor retrieval information system exists when users cannot clearly express their information needs or poor ranking methods to evaluate pages if they are related to query or not.

In order to overcome the irrelevant documents that result from search process, there are many solutions such as: using query expansion (QE), taking into account the semantic meaning; or by improving the ranking of documents, taking into account not only the occurrence of query terms, but also the semantic relation between the user search and the document context.

QE is considered a viable solution, expanding process by expanding query keywords with related terms. With an expanded query, the retrieved documents are not only based on the query terms, but also on the related terms to that query which can improve the search process. This is suitably broadened and more accurate results may be obtained by retrieving more relevant documents .Web search ranking algorithms play an important role in ranking web pages so that the user could get good results more relevant to the user's query.

This paper presents two methods to solve these problems. The first is an expansion query method taking into consideration the relations between expanded query terms in the ranking process of documents, by organizing all terms of an expanded query as a tree model of multi-levels, regarding their hierarchical relationships defined in a specific ontology. The second method is a ranking process for documents based on the semantic relation between document contents and the query terms.

II. RELATED WORK

Search engines accuracy is improved based on how they will search for the meaning of query terms, and how they will present the results to users by evaluating the documents containing the query terms. There are many solutions for improving the search engine: by expanding query taking into account the semantic meaning related to user's query terms; or by improving the evaluation of documents not only by the occurrence of terms, but also by how it semantically relates to the topic search.

Query expansion (QE) is a technique used to aid users to express their requirements. There are many works in QE techniques, such as the mechanisms of relevance feedback [5] and statistical term co-occurrence [6]. The drawback of relevance feedback and statistical term co-occurrence methods is the analysis of pervious results documents which may provide a relationship between extracted terms and the

original query. But this cannot be ensured if there are no sufficient documents used for analysis before a search process.

The semantic meaning is a method based on ontology to disambiguate the query meaning [7]. This method is used to expand query terms by their synonyms using WordNet ontology, or by adding synonyms and terms related to them based on ontology domain. But adding these terms to query without taking into consideration their hierarchical relationships may affect the relevance of documents to the main query terms [8].

Ranking methods are applied to arrange the documents in order of their relevance, importance and content score using web mining techniques to do this [9]. Web mining techniques are applied in order to extract only relevant documents from the database and provide the intended information to users. They classify the web pages and internet users by taking into consideration the contents of the page (WCM), behavior of internet user in the past (WUM), and web structure mining based on links in pages (WSM) [9-13].

There are many ranking algorithms that can be classified based on the parameters used to describe them and the parameters used to calculate the ranking score. We will discuss this in the following section.

Page rank algorithm is an algorithm used by Google to rank pages. It is based on a web graph, where web pages are represented as nodes; and links as edges between pages. The page rank depends on the number of links it has. The page linked to many pages with high PageRank receives a high rank itself [14-16].

Weighted links rank (WLRank) is the modification of the standard page rank algorithm [17]. This algorithm provides weight value to the link based on three parameters; the length of the anchor text, tag in which the link is contained, and relative position.

Time Rank Algorithm is based on the visit time of a webpage [18] to overcome the keywords query match without taking into account the context of user meaning. User's preferences in content and in a link are used to rank pages [19-20]. Also, user behavior can be used to indicate the importance of webpages and websites, by analyzing the individual user sessions to rank the web pages [21].

Semantic ranking is based on the domain ontology by similarity between ontology concepts and document page terms using the term frequency of terms [22]. Semantic ranking is based on the user logs and IS A and Part of hierarchy relations, the extension similarity is based on the user browsing patterns and their hyperlinks, the content similarity between two nouns are constructed based on the IS A and Part of hierarchy using user's web log to find the semantic ranking web page [23]. Modifying graph base sentence ranking by summarizing a text to nodes and edges as relations, to hypergraph to overcome the group of relationships between sentence, where a sentence represents as nodes and edges may be group relationship or pairwise relationship Text hypergraph for summarization and hypergraph based semi-supervised learning algorithm for sentence ranking [24].

III. Analytic Hierarchy Process (AHP)

The Analytic Hierarchy Process (AHP) is an effective tool for dealing with complex decision making; it aids the decision maker to determine the priorities of used criteria. It based on a series of pairwise comparisons and then synthesizing the results, it also incorporates a useful technique for checking the consistency of the decision maker's evaluations.

The AHP generates a weight for each evaluation criterion according to the decision maker's pairwise comparisons of the criteria. The higher the weight, the more important the corresponding criterion is. Next, for a fixed criterion, the AHP assigns a score to each option according to the decision maker's pairwise comparisons of the options based on that criterion. The higher the score is, the better the performance of the option with respect to the considered criterion. Finally, the AHP combines the criteria weights and the options scores, thus determining a global score for each option and a consequent ranking. The global score for a given option is a weighted sum of the scores it obtained with respect to all the criteria [25].

In AHP (Analytic Hierarchy Process Matrix) a matrix is constructed where the Rows and Columns have the same parameters. The first row and the first column have the same parameter and the so on for other rows and columns. once the matrix is arranged ,the comparison between each row with all columns are done to determine the score, where a maximum score implies that the row is more important than the column. The diagonal of the matrix is allocated a score of 1. The score value of cell below the main diagonal is just inverse of the scores in the corresponding row. Likewise calculate all the columns. Add the columns. Calculate the new table to normalizing the scores; divide each value of a cell of a column by the column total. Likewise do for all columns. Add the rows of this new table. This will be the Normalized score for each parameter. Convert into percentage by dividing the normalized score for a parameter with the column total of the Normalized Score Column and multiplying by 100. This will be the Percent Ratio Scale Of Priority (PRSP) for each parameter and will also be the priority of your customer.

TABLE I. The Score Matrix

	X	Y
X	1	3
Y	1/3	1
Sum	1.3	4

TABLE II. Normalized and Priority Table

	X	Y	Sum	Priority
X	1/1.3	3/4	1/1.3+3/4	(1/1.3+3/4) /S1*100
Y	0.33/ 1.3	1/4	0.33/1.3 +1/4	(.33/1.3+1/ 4)/S1*100
Sum	1/1.3 +.33/ 1.3	3/4+ 1/4	S1	

IV. SEMANTIC SIMILARITY

The semantic similarity techniques are used to determine how two concepts or terms are similar, they are used in many applications such as intelligent information retrieval, knowledge integration systems, sense disambiguation, classification and ranking, detection of redundancy, and detection and correction of malapropisms [26,27]. Semantic similarity between words is measured by using semantic web (ontology) which define words with their define meaning, and describes the relationships between terms or concepts and their properties.

There are many techniques used to semantic similarity using domain ontology, wordnet, and corpus. Also semantic similarity can be measured based on the information content based approaches that use ontology structure and corpus-based features such as Resnik [28], Jiang & Conrath [29], Lin [30], and structure based approaches such as path length [31], Leacock.& Chodorow [32], Wu & Palmer [33].

Semantic similarity is important approach in information retrieval, semantic similarity can evaluated using page count, and text snippets retrieved from search engine for two terms. Using page count to count the result of searching of each term alone, and pages contain two terms to evaluate how they depend or independent terms [34]. Google used to evaluate semantic relatedness to calculates the similarity between two words, and distance between them [35].

In text snippets retrieved , searching about two terms and extract a snippet from results such as Wikipedia pages for two terms and processing the result to extract only the main terms in original form, then using a five similarity measure of association that is simple similarity. Jaccard similarity comparing the similarity and diversity of given sample set. Dice similarity also related to the jaccard measure. Over Lap method is used to find the overlapping between the two sets. The cosine similarity is a measure of similarity between two vectors of n dimensions by finding the angle between them [34].

V. THE PROPOSED SEARCH ENGINE TECHNIQUE

The proposed engine enhances a search engine through two methods. The first is the disambiguation of query terms by expansion process using general purpose ontology and domain ontologies selected by searching in the domain it is dealing with. The domain ontology is selected by searching in the domain dealing with and taking into consideration the relation of expanded terms through ontology domain description. The second method improves the ranking process taking into account the semantic relation between terms found on the page. This engine retrieves a high amount of the available semantic documents and enhances current search technology on the web. It performs the basic functionalities of the traditional search engine including: crawling web documents, indexing, ontology selection, query manipulation and expansion, and thus ranking documents.

As Fig.1 depicts, the architecture of the proposed engine indicates the two suggested methods, each of them composed of some modules. The Search engine has a main module that is a user interface module, and an additional module that is semantic search for ontology domain search.

- User Interface Module: is an easy interface for user to enter their queries and show required results.
- Semantic Search Module: In this module, the process of searching for the semantic documents is related to the domain search using the user queries to provide a suitable ontology.

A. The Query Expansion Method

This method is an expansion query process to disambiguate the query terms and to explain the meaning of query terms using their synonyms from WordNet ontology (general purpose ontology), and their related terms from domain ontology taking into account their relationships. It consists of three modules fig.2: query manipulation (expansion), semantic query and weighting module (building tree model, using AHP algorithm).

- Query manipulation: In this module, query is interpreted by performing preprocessing, stemming and disambiguating the query. Disambiguating the query is done by adding semantic meaning to terms with their synonyms using general purpose ontology (WordNet).
- Semantic query Module: In this module after connect to WordNet to extract the synonyms for each query terms and based on domain ontology extract hyponyms for query terms and their sibling, we construct all semantic meaning to query terms as a vector of terms.
- Weighing Modules: consist of two parts.

The first step is building a hierarchy tree based on domain ontology and the synonym terms in two-level trees. A tree model is a technique used to build a tree with multi-levels. All terms of an expanded query are organized as a tree with multiple levels regarding their hierarchical relationships defined in a selected ontology. In this model, the synonyms are located at the same level as the query terms and the hyponyms are distributed at a lower level. The relevance scores generated by those expansion terms and documents are evaluated upon the degree of relation between terms, original query and documents.

The second step is to evaluate the weight values based on AHP algorithms. AHP is a multi-criteria decision support methodology used in management science. We estimate the mutual importance values between relevance generated by original query terms and synonyms and hyponyms estimated based on the AHP score [36]. Where the original query terms and their synonyms are in the same degree of importance, but their hyponyms terms have different degree.

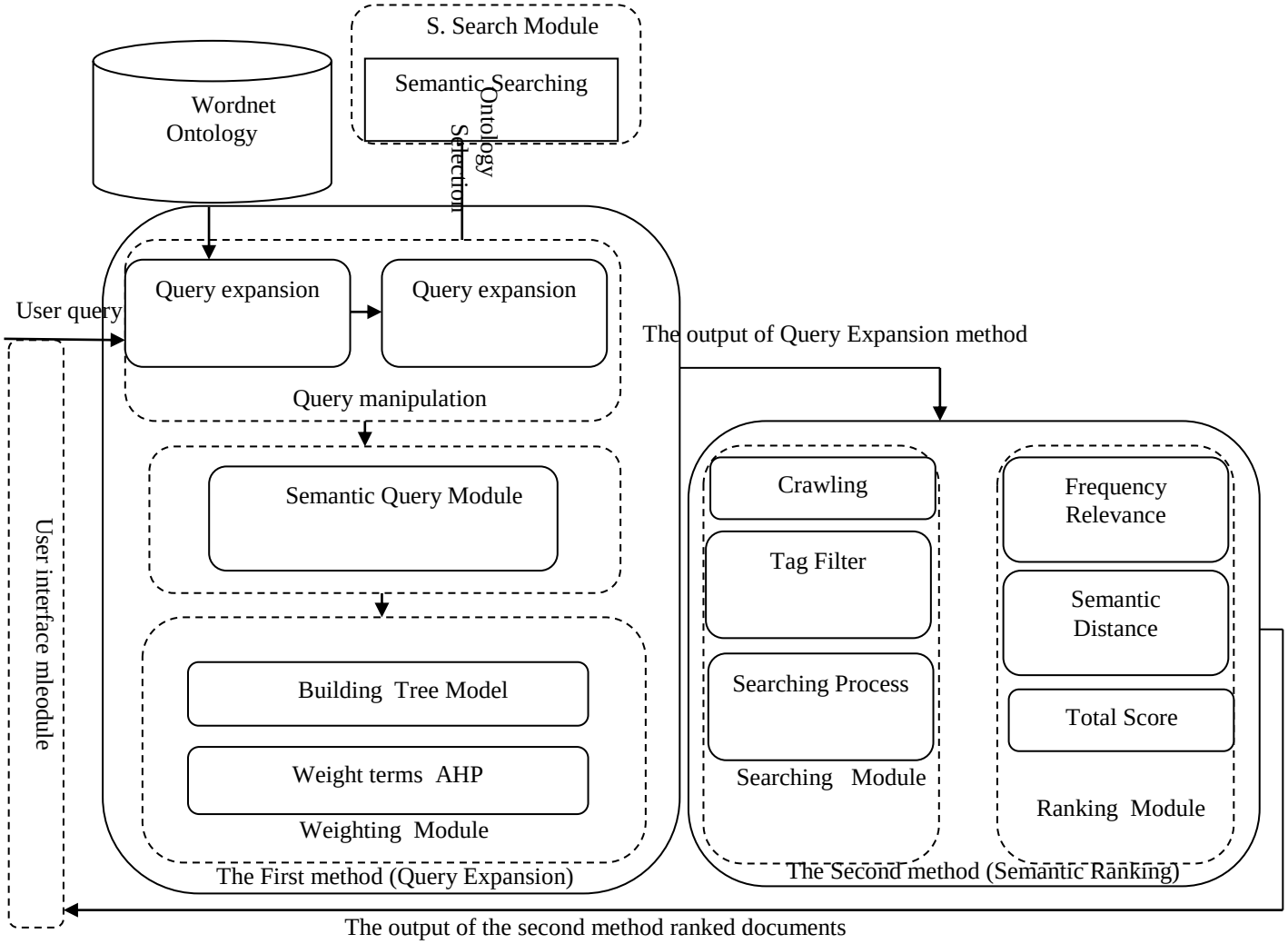


Fig. 1. Proposed System Architecture

B. The Semantic Ranking Method

Ranking process is considered an important step in any search engine. A good search engine is evaluated by whether the user's requirement exists in relevant documents which are returned, and evaluated by ranking techniques. This method consists of two modules a Searching Module, and Ranking Module (Semantic distance in content, term frequency) as shown in fig.3.

- Crawling: crawling the documents and indexing them [37,38].In crawling we based on crawler built using java code enter a start url and extract a list of urls from pages ,indexing process by parsing url document using jsoup java tools that deal with html pages ,it parsing html based on tags ,which allow us extract each text tag separately, split them based on (. dot) for each statement or (" " space) for terms , removing stopwords and stemming them ,calculate the frequency of each term and storing them in database.
- Tag Filter: Most information are represented in internet pages in HTML documents, which it contains a set of markup tags that represent the content. These tags have different priorities in documents. Many retrieval information works deal with tf (term frequency),VSM(vector space model) and many other techniques deal with all document as a whole.

But HTML have many parts (tags) which mean different priorities, such as the document that have query term in title tag mean related to the query more than the document have a query term in other tag, the query term in <a href> is related to another page that explain it in detail ,and so on, then it becomes difficult to weight all document as the same in final ranking [39]. Due to the above mention some works deal with document as classify document based on tags, but it deal with single query term [40], also dealing with document tag by adding extra weight to term found in special tag [41].In this paper we deal with main document tags (title, head, body) construct a weight to each tag based on AHP dealing with semantic distance between query terms in each tag.

We implement our system using jsoup as a tool in java working with real world HTML(jsoup: Java HTML Parser),it provides a way to extract html tags ,extract the text for each

tag select("title"), select("body"), select("head")), then split text by " ,_-" any special characters, remove stopwords ,and stemming each term to restore in original form, connecting with WordNet ontology to return the synonyms for each term. All these data are stored in database relate the terms to original text contains and in which tag, to measure the semantic distance.

- Searching Process: Searching for documents that have query terms and their expansion and taking into account their frequency of each term found.
- Ranking Module: Ranking plays an important role in searching. In this paper the documents are evaluated based on the semantic relation between terms in statements (semantic distance) and term frequency. The related terms found are weighted based on the result of the tree module .These two values are calculated according to the following subsections.

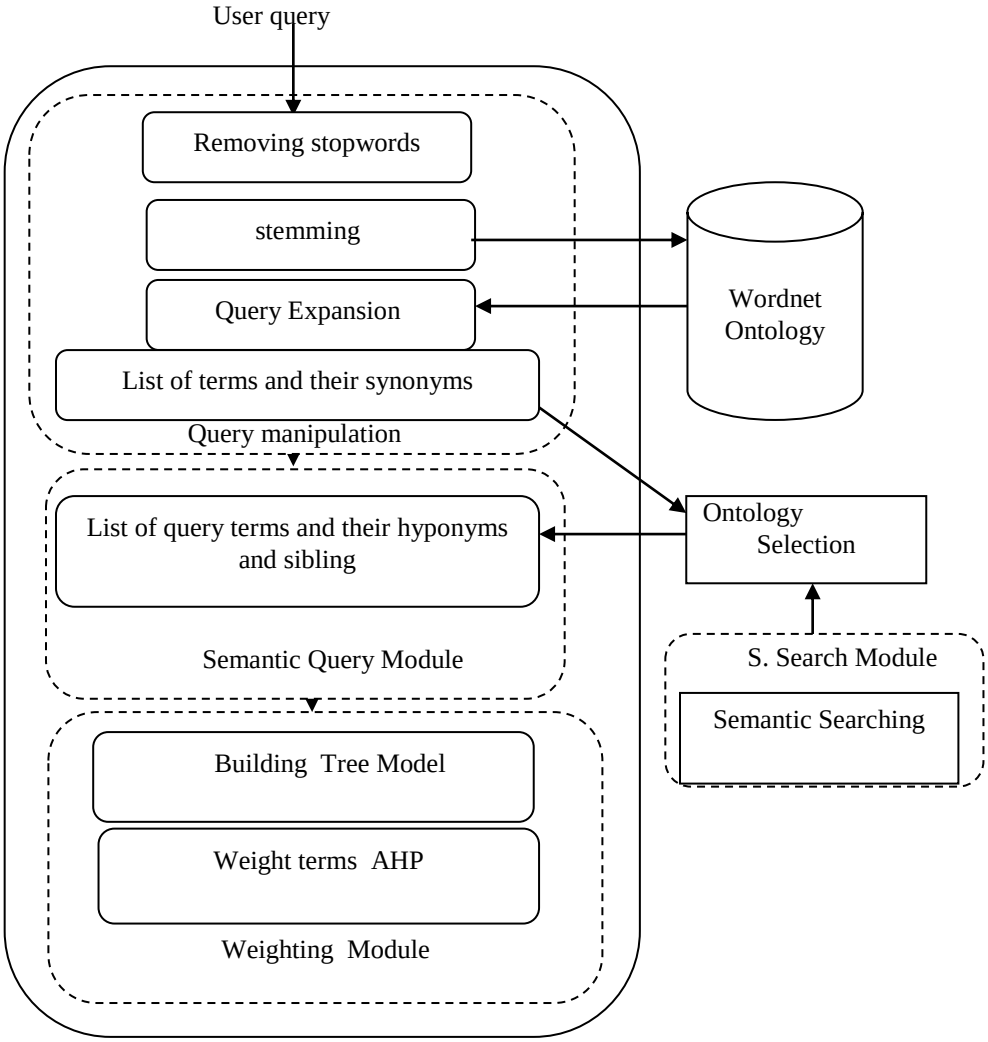


Fig. 2. The first Part of query expansion and weighting

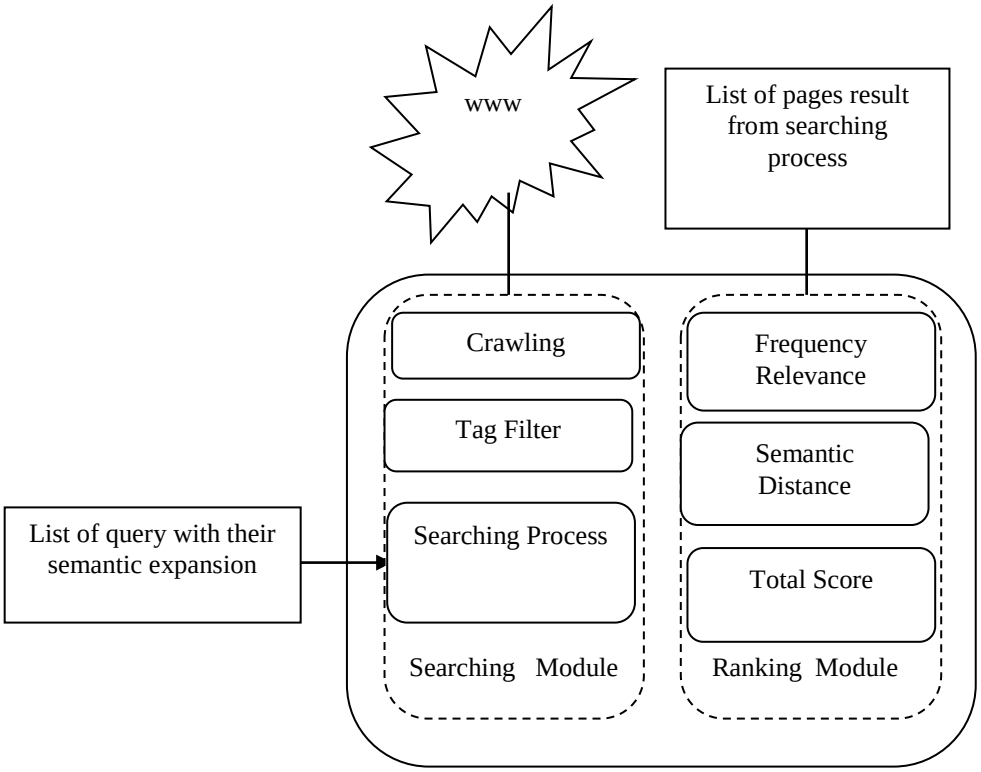


Fig. 3. The Second part Ranking Process

1) Frequency Relevance

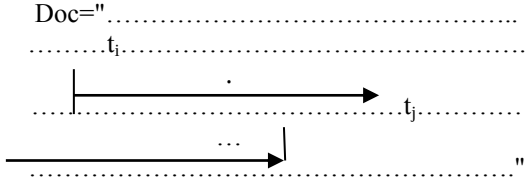
Frequency is used to evaluate how documents are related to the user query, by searching in the document for the number of occurrences of the required terms. In the previous works, they took into account the summation of the frequency of all terms found. But in our proposed method, due to the expansion of query; we take into account the semantic relationships (synonyms, hyponyms) to ensure that the page is related to the domain selected. We weight each frequency term to indicate their priority on the page. fi is the frequency of term i, so the total frequency relevance is the summation of all query terms.

$$F(D)=\frac{\sum_{i=1}^{k} w_i \, f_i}{|D|} \quad (1)$$

w_i : weight value estimated using AHP algorithm, f_i : frequency of term i, K: number of query terms, |D| :is the length of document to make normalized for total frequencies of terms.

2) Semantic Distance Function

Using the frequency relevance for query terms may introduce multiple topics and irrelevant information within relevant documents. In order to provide the semantic distance between two terms, the weights of their hierarchical structure in documents are taken into account. We measure the distance between query terms and their hyponyms found in documents; the terms that have higher distance between them become less related terms. The distance function is a weighting function to measure the semantic distance between the terms of queries and their hyponym found in the document. Where the frequency based of terms dealing with terms in any position within the documents, whatever these terms are related to each other or not. So, it is important for assessing if a term is close enough to query terms and their expansions, which indicates if the document related to specific topic or not, the position distance is adapted from one proposed[42]. Based on the relevance model, the main idea of the positional relevance model (PRM) is to further distinguish different positions of a term and discount the occurrences of a term at positions that are far away from a query term in a document. We modify this work to be suitable on document ranking and semantic distance as follows in (2).



We calculate the position between two terms, based on the semantic distance between two concepts in ontology which is calculated by measuring the distance (length of path between two concepts). We estimate the distance between two terms by the length between terms in statement.

$$P_{pos}(t_1|D) = \sum_{j=1}^{m} SD(t_1, t_j) = \\ = \sum_{j=1}^{m} \left(\frac{1}{len\left(t_1 \xrightarrow[j=1:m]{} t_j\right)}\right)/|ST| \quad (2)$$

Where, $SD(t_1, t_j)$ is the semantic distance between two terms t_1,t_j, where j terms from 1 to m (m: number of all query terms and their expansions); |ST| length of statement.

For each term t_1, we measure the distance between it and all the other terms, their synonymous, other query terms and their expansions (t_1 and t_j). For each sentence or statement or paragraph separated by (.,\n) are splitted remove the stopwords and return each term to their original form and measure the length between terms. $\text{len}(t1 \xrightarrow[j=1:m]{} tj)$ is the length between ti and tj. Because we deal with different length statements, normalize this result by divide by the length of statement |ST|.

In our proposed method, we calculate the semantic relation between query terms in each part in web page (title, head, and body). The semantic relation of each part is calculated as shown in (3):

$$SR(D) = \sum_{h=1}^{3} w_h * \sum_{i=1}^{k} w_i * \sum_{j=i}^{k} SD(t_i, t_j) \quad (3)$$

Where, SR(D): the semantic relations, k: number of query terms with expansion, $SD(t_i,t_j)$: semantic distance between terms shown in (2), w_i : the weight of t_i, w_h is the weight for each tag in html document contain query terms.

For each term, calculate the semantic distance between this term and all other searched terms and their synonyms and hyponyms for each part in documents. If the term occurred many times in the paragraph or sentence, we deal with each statement separately, if no terms we deal with the paragraph as a whole.

The total semantic relation for each page is calculated as the summation of semantic relation for three weighted parts using AHP algorithms that indicate title tag with higher priority than body tag, and body tag with higher priority than the head tag (0.607002, 0.303344, and 0.089654).

3) Total Score

Based on the pervious notice ranking document based on the term frequency or cosine similarity between query terms and document contents, they does not take into account the semantic relation between terms found in documents, so to aggregate the advantage of the occurrence of query terms and how they are close related to each other, we add two values of frequency and semantic distance between query terms as shown in(4)

$$R(D) = w1 * \sum_{i=1}^{3} SR(D) + w2 * F(D) \quad (4)$$

Where, $\sum_{i=1}^{3} SR(D)$: Is the Semantic relation calculated for document D for each part (title, head, body) in HTML documents, F(D) is the total frequency of terms found in documents.

VI. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned into a system. It can be considered the most critical stage in achieving a successful new system to give the user confidence that the new system will work and will be effective. The proposed system was implemented using Java and JENA software as a simulator.

As previously explained, the implemented system consists of some steps; it starts with the crawling process. Multi-threaded, multi crawlers were implemented to crawl both traditional and semantic web. The implemented crawlers make use of seed URLs to extract semantic web (pages with extended owl type). Another option for crawler is to enter the domain or keywords to search in swoogle API or Google API to search for all files by (file type: owl). After collecting web documents, their content is then parsed using Jena software to extract all semantic web details (concepts/classes, relations, instances /individuals). These data are stored in the index.

When the user enters his query through an interface, the implemented system comes with capabilities that enable the user to identify his intent by disambiguating his query using the WordNet database to extract the synonyms of his query. The important part of the interface is a list of semantic web document results with a summary such as their content description class number, property number and instances number. As well, it allows an immediate preview for related data found in the ontology, if the user selects one. The semantic web document with a high ranking is selected and is used as a domain ontology description to expand a user's query. After expansion, a tree of expanded terms is built and the weights are evaluated for expanded terms using AHP process.

A. Data Sets

Implementing our proposed system with the three real-world data sets below:

- Academic Staff & University (Staff): academic staff's full names (from 20 different universities) and their universities have been collected. (D1)
- Drug & Disease (Drug): This data set contains 200 drug names and the names of 183 different diseases they can cure. It was extracted from a drug list. (D2)
- Invention & Inventor (Invention): This data set contains 512 inventions' names with their chief inventors' full names (311 different people) from an inventive list on Wikipedia. (D3)

For D1, database is collected by crawling documents based on Google search engine for university (seed url).D2 is collected by crawling documents form Wikipedia site for pharmaceutical products, and Google search for drugs and diseases. The ontology which is used to expand query and build a tree to determine their priority is selected by user, when he/she enters his/her query search engine for semantic documents are working to present a ranked semantic documents and allowing immediate preview for ontology descriptions related to this query.

B. Results

For the experiment, two parameters are used to evaluate the information retrieval system. These two values are precision P and recall R, where Nc is the number of correct web pages returned, Nr is the number of related returned web pages, but they were not necessarily the correct web page, and Nt is the number of total returned web pages.

The documents are crawled and are stored in database. These documents are classified as 'relevant' and 'non-relevant'. The judgment for relevant and ranking the documents collect for D1 based on ranking result of (http://www.arwu.org/), for D2 the relevant and non-relevant ordering of documents are based on Google search.

The two values precision and recall are calculated using the following equations:

$$P = \frac{Nc}{Nt}$$

$$R = \frac{Nc}{Nc + Nr}$$

$$f - mesure = \frac{2*p*R}{P+R}.$$

To measure the performance of the suggested ranking method, there are four different documents search engines, named SR1, SR2, SR3 and SR4; respectively. They are implemented using Java, where SR1 represents a traditional keyword-matching search engine based on the user query terms only, which does not employ any QE techniques. SR2 is a search engine based on expanding user query taking into account their synonyms, SR3 is a search engine that does the search process by expanding queries based on the pervious retrieval pages for that domain based on the relative terms and their frequency[43]. SR4 uses the proposed ranking method based on expanding queries by disambiguating their meaning with synonyms using WordNet and their subclasses from a domain ontology taking into account their relative weights to that expanded term.

In our system, we evaluate the ranking by using the relative weights for expanded terms to measure how documents are related to query terms based on the priority of terms founds evaluated using AHP algorithm.

In D1 for example we search about academy staff by query terms "academy staff & university" are expanded based on the selected ontology to (staff, university, academia, faculty, research, clerical staff ,system staff, professor, research assistant, administrative staff, chair, dean, teacher ,organization, affiliated organization, course, lecture) with a relative weight to indicate the importance and their priority in documents(0.094811868,0.090237076,0.073245559,0.084874439,0.023292748, 0.008797915,0.099963995,0.05668714,0.053551544,0.028051472,0.037997043, 0.062339604, 0.067050112,0.021785616,0.02354974,0.037918655, 0.07472).

SR3 method based on the related terms from the previous query result, for the same search query "academy staff & university" which is expanded to (university, professor, school, faculty, technology, department, Dr., institute, lecture, PhD, edu). In the first method, our search is based on only the query terms; in SR2, the ranking method is based on the query terms and their synonyms.

In our method, the searching and ranking process does not only depend on the terms found or on their frequency, but it also takes into account the importance or priority of the expanded terms through domain ontology with the relationships, synonyms and hyponyms of a query term. This process is done by weighting values to indicate the important terms. SR3 is based on the related terms from the previous query result. For the same search query "academy staff & university" which is expanded to (university, professor, school, faculty, technology, department, Dr., institute, lecture, PhD, edu). In the first method, our search is based on only the query terms; while in SR2, it is based on the query terms and their synonyms.

In our method, the searching and ranking process not only based on the terms found or on their frequency, but it also takes into account the importance or priority of the expanded terms through domain ontology with the relationships synonyms and hyponyms of a query term, by weighting values to indicate the important terms.

TABLE III. THE PRECISION OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.53	0.67	0.73	0.9
D2	0.62	0.6	0.66	0.9
D3	0.55	0.53	0.35	0.9

TABLE IV. THE RECALL OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.588889	0.744444	1	1
D2	0.688889	0.666667	1	1
D3	0.611111	0.588889	1	1

TABLE V. THE F-MEASURE OF COMPARISON METHODS

	SR1	SR2	SR3	SR4
D1	0.557895	0.705263	0.843931	0.947368
D2	0.652632	0.631579	0.795181	0.947368
D3	0.578947	0.557895	0.518519	0.947368

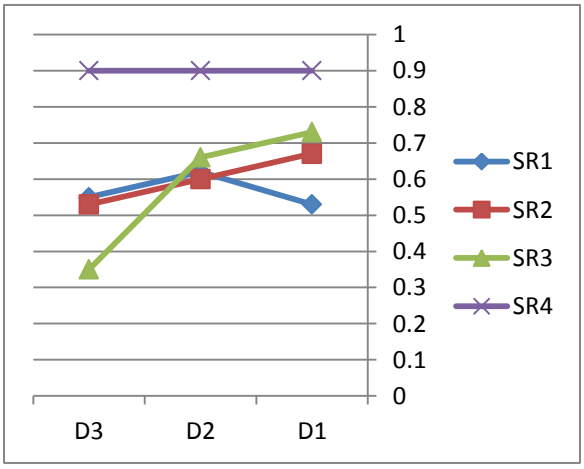


Fig. 4. The Precision of comparison between methods

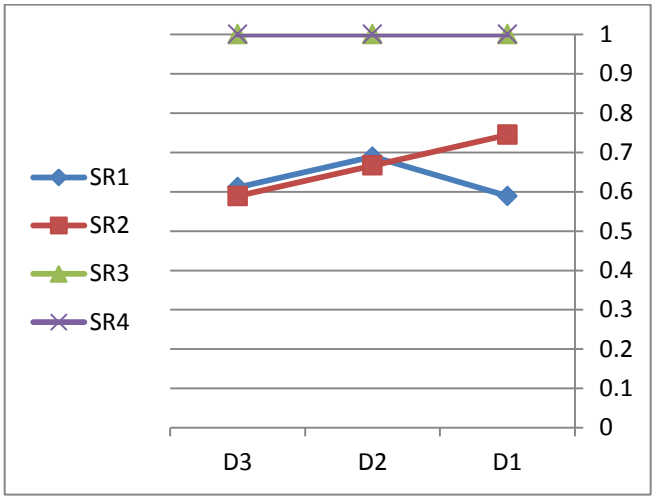


Fig. 5. The Recall of comparison between methods

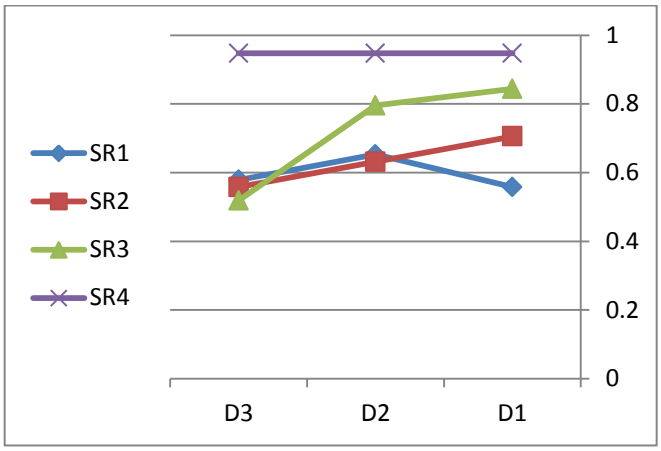


Fig. 6. The F-measure of comparison between methods

In SR1, the search results depend only on the query terms so it holds only the document; contain query terms unless it doesn't relate to the domain searched. In the SR2 and SR3, the relevant documents are increased based on the expansion, query terms by synonyms and related terms from pervious query results respectively. While in SR3, it depends on the good pervious results.

We notice that the expanded query in the SR3 method has the same recall as the proposed method, but it is still controlled using the related terms that expand from the previous query results.

SR4 depends on the expanded terms using domain ontologies that are searched for by our system, controlled by multiple parameters such as: properties of the concepts, properties and instances searched in ontologies (details of domain description).

For measuring the semantic similarity based on html tag, we take only three main tags (head, title, body)tags ,with the pervious weights .we measure the recall and precision for html documents based on tags, We notice the precision and recall of semantic similarity is increased based on numbers of query terms found in parts in html pages as shown.

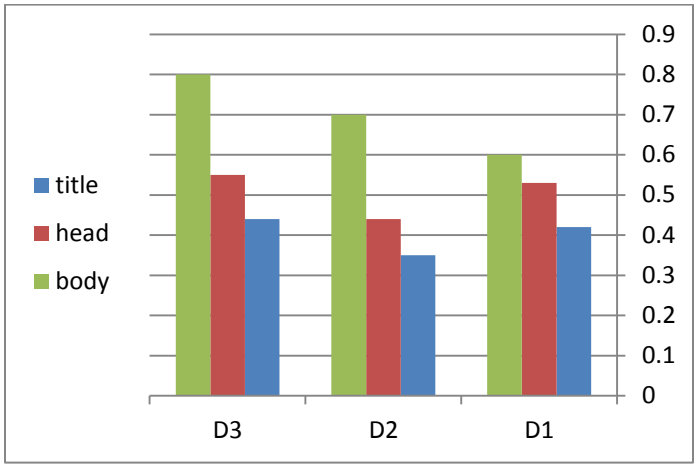


Fig. 7. The precision based on semantic distance using html tags.

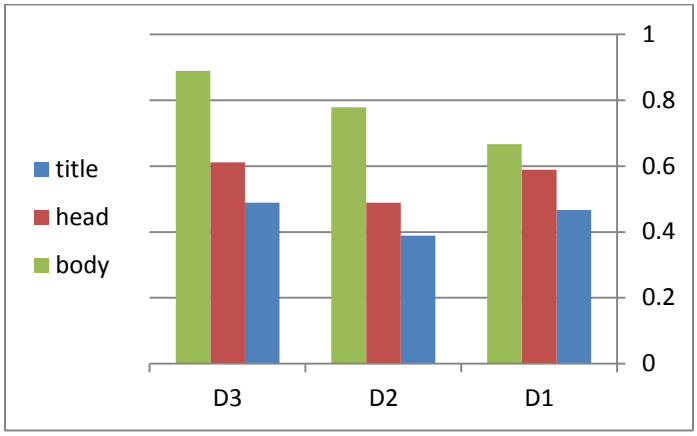


Fig. 8. The recall based on semantic distance using html tags

VII. Conclusion

In this paper, a system is proposed to improve the search process to overcome the traditional search problems by some methods, such as enhancing the expression of what the users actually mean and enhancing the evaluation process of the documents returned to users. The process of query expansion can be done using relevance feedback-based, statistical co-occurrence-based and domain ontology. But in the case of using domain ontology while dealing with all expanded terms from ontology that has different relationships with the same weighting which will affect in the evaluation to documents that contain them. The new proposed method used to search based on ontology and expanded query with domain ontology and ranking document taking into account the related weights in expanded terms as in the ontology domain in the hierarchical structure. These weights will affect the document accuracy related to the user main query terms.

VIII. Future Work

In this work we focus on single ontology for single domain. In future, we will focus on multiple ontologies which allow us to give an opportunity for employing the knowledge from different ontologies of single or different domains. Also, we will take into account the important html tags such as link (<a href> <.a>), bold tag <B>.

REFERENCE

[1] Preethi , Ms.N., and Devi , Dr.T., Case and Relation (CARE) based Page Rank Algorithm for Semantic Web Search Engines. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

[2] Lee, T. B., Hendler, J., and Lassila ,O.,"The semantic web". Scientific American, vol. 284(5), May 2001.

[3] Ch.-Qin Huan,Ru-Lin Duan, Y. Tang, Zhi-Ting Zhu, Y.-Jian Yan, and Yu-Qing Guo, ."EIIS: an educational information intelligent search engine supported by semantic services".International Journal of Distance Education Technologies ,January 1, 2011.

[4] Robin Sharma , Ankita Kandpa,and Priyanka Bhakuni, Rashmi Chauhan, R.H. Goudar and Asit Tyagi." Web Page Indexing through Page Ranking for Effective Semantic Search". Proceedings of7'h International Conference on Intelligent Systems and Control (ISCO 2013).

[5] Yuan LIN,Hongfei LIN, and Li HE." A Cluster-based Resource Correlative Query Expansion in Distributed Information Retrieval ".Journal of Computational Information Systems 8: 1 ,2012, 31–38.

[6] W. W. Chu, Z. Liu and W. Mao."Textual document indexing and retrieval via knowledge sources and data mining". Commun. Inst. Inf. Comput.Mach. (CIICM), Taiwan, 2002, 5, (2), pp. 135–160

[7] A. Vizcaíno, F. García, I. Caballero, J.C. Villar, M. Piattini."Towards an ontology for global software development". IET Softw., 2012, 6, (3), pp. 214–225

[8] N. Tyagi and S. Sharma."Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page". In International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.

[9] N. Duhan, A. K. Sharma and K. K. Bhatia."Page Ranking Algorithms: A Survey". In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

[10] Vishal Jain, Dr. Mayank Singh."Ontology Based Information Retrieval in Semantic Web: A Survey ", I.J. Information Technology and Computer Science, 2013, 10, 62-69

[11] M.Yadav,and Mr. P. Mittal." Web Mining: An Introduction ". International Journal of Advanced Research in Computer Science and Software Engineering, 2013,Volume 3, Issue 3, March , ISSN: 2277 128X

[12] Md. Z. Hasan, Kh. J. A. Chisty and Nur-E-Z. Ayshik . "Research Challenges in Web Data Mining". International Journal of Computer Science and Telecommunications , Volume 3, Issue 7, July 2012.

[13] S. Pal, V. Talwar, and P. Mitra ."Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions". In IEEE Trans. Neural Networks, 13(5), PP.1163–1177,2002.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web" . Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[15] P. Devi, A. Gupta and A. Dixit. "Comparative Study of HITS and PageRank Link based Ranking Algorithms". International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014.

[16] Ch. D. Manning, P. Raghavan and H. Schütze ."Introduction to Information Retrieval". Book Introduction to information retrieval Cambridge university press New York, NY ,USA , 2008, ISBN:05218657199780521865715.

[17] R. Baeza-Yates and E. Davis ."Web page ranking using link attributes" . In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters , 2004, PP.328-329.

[18] H Jiang et al."TIMERANK: A Method of Improving Ranking Scores by Visited Time". In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[19] J.Jayanthi, Dr.K.S.Jayakumar. "An Integrated Page Ranking Algorithm for Personalized Web Search". International Journal of Computer Applications (0975 – 8887), Volume 12– No.11, January 2011.

[20] K.-J. Kim and S.-B. Cho . "Personalized mining of web documents using link structures and fuzzy concept networks". Applied Soft Computing, Volume 7, Issue 1, January 2007, Pages 398–410.

[21] Guangyu Zhu and Gilad Mishne." Clickrank: Learning session-context models to enrich web search ranking". TWEB,6(1):1, 2012.

[22] Ahmad Kayed, Eyas El-Qawasmeh, and Zakaryia Qawaqneh."Ranking web sites using domain ontology concepts".Information & Management, 47(7-8):350–355, 2010.

[23] Yajun Du and Yufeng Hai. "Semantic ranking of web pages based on formal concept analysis". Journal of Systems and Software, 86(1):187–197, 2013.

[24] Wei Wang, Sujian Li, Jiwei Li, Wenjie Li, and Furu Wei." Exploring hypergraph-based semi-supervised ranking for query-oriented summarization". Inf. Sci., 237:271–286, 2013.

[25] N. Bhushan, K. Rai." Strategic Decision Making Applying the Analytic Hierarchy Process". http://www.springer.com/978-1-85233-756-8, 2004.

[26] A. Hliaoutakis, "Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline", Master's thesis, Technical University of Crete,Greek, 2005.

[27] A. Budanitsky and G. Hirst," Evaluating WordNet-based measures of semantic distance", Computational Linguistics, vol.32,1, March 2006.

[28] P. Resnik., "Using information content to evaluate semantic similarity". In Proceedings of the 14th international Joint Conference on Artificial Intelligence, 448–453. Montreal, Canada, 1995.

[29] J. J. Jiang and D.W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,", Proc. ROCLING X, 1997.

[30] D. Lin, "An Information-Theoretic Definition of Similarity,". Proc.Int'l Conf. Machine Learning, July 1998.

[31] R. Rada, H. Mili, E. Bichnell, and M. Blettner, "Development and Application of a Metric on Semantic Nets,", IEEE Trans. Systems,Man, and Cybernetics, vol. 9, no. 1, pp. 17-30, Jan. 1989.

[32] C. Leacock., M. Chodorow, "Combining local context and WordNet similarity for word sense identification," In Fellbaum, C., ed., WordNet: An electronic lexical database, pp. 265-283. MIT press. 1998.

[33] Z. Wu . , M. Palmer, "Verb semantics and lexical selection," In 32nd Annual Meeting of the Association for Computational Linguistics, pp. 133–138,1994.

[34] Strube , S.P. Ponzetto, "Wikirelate! Computing Semantic Relatedness Using Wikipedia," Proc. Nat'l Conf. Artificial Intelligence (AAAI '06), pp. 1419-1424.2006.

[35] E. Iosif ,A. Potamianos, "Unsupervised Semantic Similarity Computation between Terms Using Web Documents". IEEE transactions on knowledge and data engineering, vol. 22, no.11, November 2010.

[36] A. Awasthi ,S.S. Chauhan . "Using AHP and Dempster–Shafer theory for evaluating sustainable transport solutions". Environ. Model. Softw.,2011, 26, (6), pp. 787–796

[37] H. Hama ,Thi Thi Zin ,P. Tin "Optimal Crawling Strategies for Multimedia Search Engines". Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, September 12-September 14,2009.

[38] C. Castillo ."effective web crawling". SIGIR Forum, ACM Press, Volume 39, Number 1, New York, NY, USA, p.55-56 (2005)"

[39] S. Pathak, S. Mitra." A New Web Document Retrieval Method Using Extended-IOWA (Extended-Induced Ordered Weighted Averaging) Operator on HTML Tags". IOSR Journal of Computer Engineering (IOSR-JCE),e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, PP 65-74,www.iosrjournals.org,2014.

[40] J. Deng, L. Chen." Web Documents Categorization Using Fuzzy Representation and HAC". In: Proceedings of the IEEE First International Conference, vol. 2, 2000, pp. 24-28.

[41] Y. Bassil , P. Semaan ." Semantic-Sensitive Web Information Retrieval Model for HTML Document". European Journal of Scientific Research, ISSN 1450-216X, vol. 69(4), 2012.

[42] Y.Lv and C. Zahi. "Positional relevance model for pseudo-relevance feedback." In SIGIR, pages 579-586,2010.

[43] Z. Li, M. A. Sharaf, L. Sitbon, X. Du and X. Zhou." CoRE: A Context-Aware Relation Extraction Method for Relation Completion", IEEE TRANSACTION ON KNOWLEDGE AND DATA ENGINEERING,2013.