A REVIEW ON QUERY EXPANSION AND PROCESS OF SEMANTIC RANKING TO INFORMATION RETRIEVAL IMPROVE PERFORMANCE

Y.Raju[1], Dr D. Suresh Babu[2]

et *Geetanjali College of Engineering and Technology, IT Department, Hyderabad, India[1,] Departments of Computer Science, Head, Kakatiya Government College, Kakatiya University[2]*

raju.yeligeti@gmail.com[1], sureshd123@gmail.com[2]

Abstract:World Wide Web is a vital resource of data growing continuously without any hurdles and interruption. In the current days , it becomes increasingly difficult for users to fetch valuable data due to the continually rapid growth in data volume. This huge amount of data is making search more and more typical with traditional search engine as they return huge data for a given query which consisting of relevant as well as irrelevant data. As if user is getting huge wastage of time and the browser get overload problem. So, the users are not showing with searching the information by traditional search engine. As if the So the problem of re-ranking search pages or results has become one of the main problems in IR field. Currently searching methods are mainly based on keyword matching technique but this technique has some cons. 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 absolute and accurate solution for ranking documents when compared to the three common methods.

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

I. INTRODUCTION

World Wide Web is a vital resource of data growing continuously without any hurdles and interruption. In the current days , it becomes increasingly difficult for users to fetch valuable data due to the continually rapid growth in data volume. This huge amount of data is making search more and more typical with traditional search engine as they return huge data for a given query which consisting of relevant as well as irrelevant data. As if user. So, the users are not showing with searching the information by traditional search engine.

Search is the most popular and peculiar applications on the Web. The bulk of outdated retrieval systems usually make use of metadata keywords is getting huge wastage of time and the browsers get overload problem matching with the query. However, these systems do not take into account the semantic relationships between query terms and other concepts that might be significant to users which he needs. Thus, the addition of explicit semantics can improve the search process easy. 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 . This approach offers an enhancement to olden search as it allows retrieval to incorporate the underlying terms semantics. It improves the olden search that focuses on word frequency by trying to understand hidden meanings in the 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 various 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

This paper finds the 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 being 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 different kinds of solutions for improving the search engine.

- Expanding query taking into account the semantic meaning related to user's query terms.
- 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. Following are some of the simple techniques of query expansions

- Finding synonyms of words, and searching for the synonyms as well
- Finding all the various morphological forms of words by stemming each word in the search query
- Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results
- Re-weighting the terms in the original query
- Creating a dictionary of expansion terms for each terms and then looking up in the dictionary for expansion

There are many works in QE techniques, such as the mechanisms of relevance feedback and statistical term co-occurrence. 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.

Ranking methods are applied to arrange the documents in order of their relevance. Web mining techniques are applied in order to extract only relevant document 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).

There are various 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

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 been encrypted. The page linked to many pages with high Page Rank receives a high rank itself.

Weighted links rank (WLRank): is the modification of the standard page rank algorithm. 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 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 . 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.

III. ANALYTIC HIERARCHY PROCESS AHP)

As it can be seen in the material that follows, using the AHP involves the mathematical problem in hand. It is not uncommon for these judgments to number in the dozens or even the hundred or the last number . While the math can be done by hand or with a calculator, it is far more common to use one of several computerized methods for entering and synthesizing the judgments. The simplest of these involve standard spreadsheet software, while the most complex use custom software, often augmented by special devices for acquiring the judgments of decision makers gathered in a meeting room.

The procedure for using the AHP can be summarized as:

1. Model the problem as a hierarchy containing the decision goal, the alternatives for reaching it, and the criteria for evaluating the alternatives.
2. Establish priorities among the elements of the hierarchy by making a series of judgments based on pair wise comparisons of the elements. For example, when comparing potential purchases of commercial real estate, the investors might say they prefer location over price and price over timing.
3. Synthesize these judgments to yield a set of overall priorities for the hierarchy. This would combine the investors' judgments about location, price and timing for properties A, B, C, and D into overall priorities for each property.
4. Check the consistency for the judgments.
5. Come to a final decision based on the results of this process.

IV. SEMANTIC SIMILARITY

Semantics identifies concepts which allow extraction of information from data. For a machine to be able to decide the semantic similarity, intelligence is needed. It should be able to understand the semantics or meaning of the words. But a computer being a syntactic machine, semantics associated with the words or terms is to be represented as syntax.

synthesis of numerous judgments about the decision

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

Query expansion (QE) is the process of reformulating a seed query to improve retrieval performance in information retrieval operations. In the context of search engines, query expansion involves evaluating a user's input (what words were typed into the search query area, and sometimes other types of data) and expanding the search query to match additional documents. Query expansion involves techniques such as:

i. *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).

ii. *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.

B. WEIGHTING MODULE: consist of two parts.

(i) Building Tree Model :

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:

(ii) Weight terms AHP

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. Where the original query terms and their synonyms are in the same degree of importance, but their hyponyms terms have different degree.

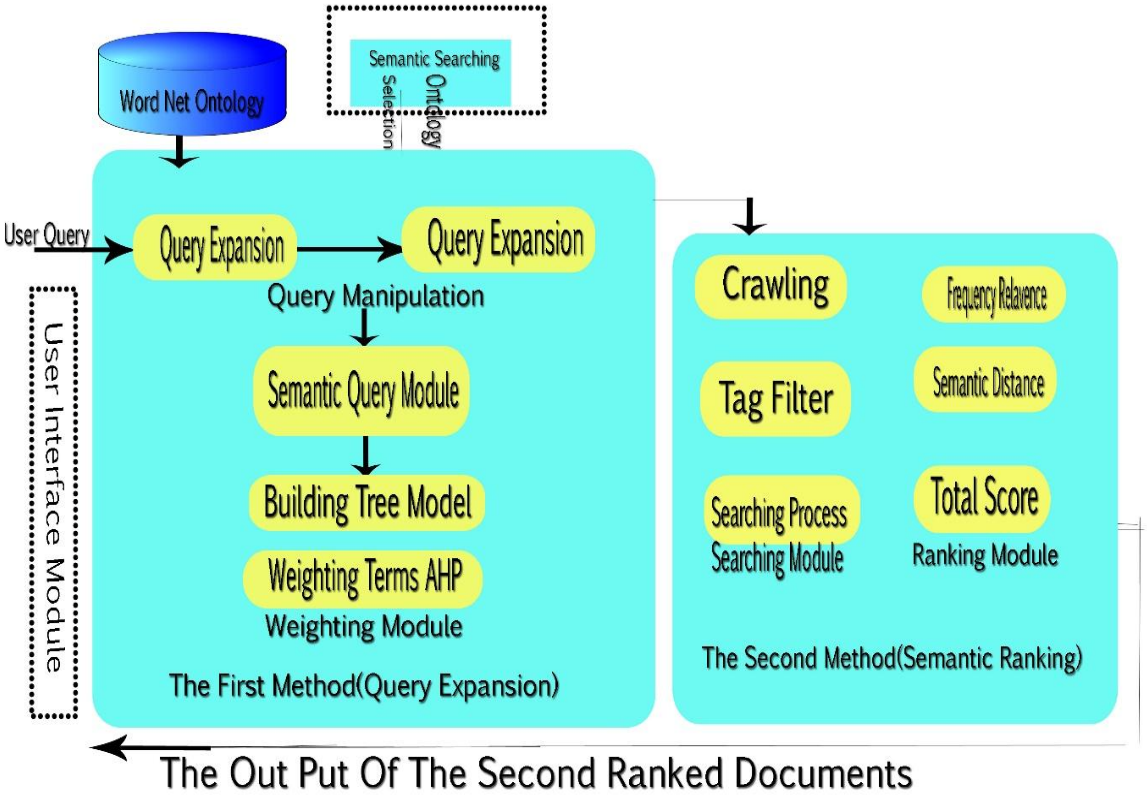


Fig. 1. Proposed System Architecture

C. 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

D. SEARCHING MODULE

(i) Crawling

Crawling the documents and indexing them .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 stop words and stemming them ,calculate the frequency of each term and storing them in database.

(ii) 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.

(iii) Searching Process

`Searching for documents that have query terms and their expansion and taking into account their frequency of each term found.

(E) RANKING MODULE

(i) 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.

(ii) Semantic Distance

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.

(iii) 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 the below

$$R\ (D) = W_1 * \sum_{i=1}^{3} SR\ (D) + W_2 * F\ (D)$$

Where, $\sum$ 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. 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 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 ontology's 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.

REFERENCE

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

[2] 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.

[3] 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).

[4] 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.

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

[6] 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

[7] 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.

[8] 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.

[9] 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

[10] H. HamaThi Thi ZinP. Tin"*Optimal Crawling Strategies for Multimedia Search Engines*". Fifth International Conference on Intelligent Information Hidingand Multimedia Signal Processing, September 12-September 14,2009.

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

[12] 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,

[13] J. Teevan, S.T. Dumais and E. Horvitz, "*Personalizing Search Via Automated Analysis of Interests and Activites*",In Proceedings of SIGIR, 2005.

[14] Boris Chidlovskii, Nathalie Glance and Antonietta Grasso, "*Collaborative re- ranking of search results*",In Proc. AAAI-2000 Workshop on AI for Web Search., 2000.

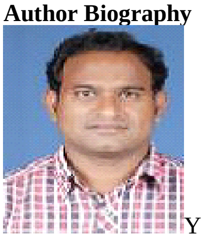
[15] Armin Hust, "*Query expansion methods for collaborative information retrieval*", Informatik - Forschung und Entwicklung, Volume 19, Issue 4, pp 224-238, July 2005.

[16] Lin Henxi, Gui-Rong Xue, Hua-Jun Zeng and Yong Yu, "*Using probabilistic latent semantic analysis for personalized web search*", In Proceedings of APWEB'05, 2005.

[17] X. Shen, B. Tan and C. Zhai, "*Context-sensitive information retrieval using implicit feedback*", In Proceedings of SIGIR 2005, page 4350, 2005.

[18] Prakasha S, Shashidhar Hr and Dr. G T Raju, "*A Survey on Various Architectures, Models and Methodologies for Information Retrieval*", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 1, 2013, pp. 182 - 194, ISSN Print:0976 – 6367, ISSN Online: 0976 – 6375.

Author Biography



Y Raju Working as an Associate professor in IT Dept at GCET, Hyderabad. He received M.Tech from JNTUH. Presently Pursuing Ph.D from JNTUH. His main research interest includes Data Mining, web personalization, neural networks, Information retrieval System and Artificial Intelligence. He has been published more than15 papers in International journals and conferences.

Dr.D.Suresh Babu is currently working Head, Department of Computer Science, Kakatiya Government College, Kakatiya University,Warangal India. He has received his Ph.D Degree in Computer science & Engineering from Acharya Nagarjuna University, Guntur, A.P., INDIA. His main research interest includes

Data Mining, neural networks, Information retrieval System and Artificial Intelligence. He has been involved in the organization of a number of conferences and workshops. He has been published more than 25 papers in International journals and conferences