VOL. 6, NO. 10, OCTOBER 2011 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

A SURVEY ON SEMANTIC WEB MINING BASED WEB SEARCH ENGINES

S. Latha Shanmuga Vadivu[1], M. Rajaram[2], and S. N. Sivanandam[3]
[1]Tamilnadu College of Engineering, Coimbatore, India
[2]Anna University, Tirunelveli, India
[3]Karpagam Institutions, Coimbatore, India
E-Mail: latha_tce@yahoo.co.in

ABSTRACT

With the advancement of the World Wide Web (WWW), the information search has been developed to be a major business segment of a global, competitive and money-making market. Search engines are the basic tool of the internet, from which related information can collected according to the specified query or keyword given by the user. A perfect search engine is the one which should travel through all the web pages in the WWW and should list the related information based on the given user keyword. In spite of the recent developments on web search technologies, there are still many conditions in which search engine users obtains the non-relevant search results from the search engines. Many web specialists have confirmed that no search engine in the world is perfectly up to date. It is also proved that no two search engines which index the similar Web information, and none of them are searching the information in the same manner. This paper also focuses on the survey of many web search engines which are proposed by various authors.

Keywords: semantic web, world wide web (www), search engine, information retrieval (IR).

1. INTRODUCTION

Searching is one of the common used operation on the Internet. Search engines is a tool of searching, are extremely popular and recurrently used sites [1]. Consider a card catalogue in a library, there exist a lot of huge and helpful information, but it's physically not possible to check all the books personally for finding particular information. In the same case, there are billions of pages on the Web; it is not possible to search all the related websites manually. So, it necessitates the need for search engines.

The web search engine generates a new challenge for information retrieval. The quantity of information on the web is mounting at a fast pace, in addition to that the quantity of new users inexpert in the art of web research are increasing. The searching operations are becoming more and more difficult with the growth of Web. The traditional search engine provides the list of dissimilar search results for the user specified keyword. Unwanted search results are created and listed by conventional Web search engines is a healthy problem in Web information retrieval (IR).

Computerized search engines that depend on keyword matching generally returns too many poor quality matches. A web search engine is intended to look for information on the World Wide Web and FTP servers [2]. The search outcomes are normally presented in a list of results and are frequently called hits. The information may comprise of web pages, images, textual data and other types of files. In addition, a few search engines mine data existing in databases or open directories.

Search engines are system programs that pass through the Web, collect the text of Web pages and formulate it possible to search for them. Remember that no search engine gathers data from the entire Web. In truth, web specialists approximated that the biggest search engines cover only 15% of the World Wide Web. Also all web specialists accepted that no search engine in the world is perfectly up to date. All the search engines are rough individualists, no two search engines which index the similar Web information, and none of them are searched the information in the same manner. It is very important to take a look at the help menu before using any search engine.

On the other hand, regardless of the recent development on web search technologies, there are still many circumstances in which search engine users are reflected with inappropriate search results. One of the most important reasons for this complexity is that web search engines have problem in recognizing users' exact search interest with the specified initial query. Also, this is because of the ambiguity that happens naturally in the variety of language itself, and that no context structure is presented to search engines. Conversely, inexpert Web search engine users are habitually not clear of the precise terms that best correspond to their specific information requirements. In the most horrible case, users are still incapable of creating exact queries representing their specific information need. Therefore, it is necessary to learn users' search patterns [3] and distinguish their search interests to provide the exact information required by the user. There are many approaches available in the literature for web search engines which gives many ideas and techniques to provide the required information collected from the World Wide Web.

2. LITERATURE SURVEY

Yan Li *et al*., [1] has done research on Web mining-based intelligent search engine. At present, the World-Wide Web has grown to a distributed information space with almost 100 million workstations and a number of billion pages, which generates the troubles for people

trying to find necessary information amongst the vast amount of information available. The search engine is a extremely an important tool for people to find information on the Internet, but low-exactness and low-recall prevails extensively in existing search engines. Through the quick improvement of the Internet, efficient and precise sharp search engines based on Web mining tools have become important. The author developed the hypothesis and correlative conception of the search engine and then the description, categorization and the application of Web mining are illustrated. In conclusion, significant applications of Web mining approach in sharp search engines are discussed in detail.

Semantic Web based intelligent search engine (SWISE) was developed by F. Shaikh *et al*., [4]. The majority of the search engines look for keywords to respond the queries from users. The search engines usually search web pages for the required information. On the other hand they filter the pages from searching needless pages by means of advanced algorithms. These search engines can respond topic wise queries effectively by providing state-of-art algorithms. But they are weak in responding sharp queries from the user due to the dependence of their results on information accessible in web pages. The most important goal of these search engines is answering these queries with close to precise results in little time with the help of much researched algorithms. The author accepts that still these search engines are weak in answering intellectual queries using this technique. They either demonstrate imprecise results with this technique or show precise but unreliable results. The user cannot have a fulfillment with these outcomes due to lack of hope on blogs etc. To get the required results search engines necessitate searching for pages that preserve such data at several places. This needs together with domain knowledge in the web pages to assist search engines in responding intelligent queries. The layered representation of Semantic Web offers solution to this difficulty by offering tools and technologies to facilitate machine readable semantics in present web contents.

Yi Jin *et al*., [5] performed the research of search engine based on semantic web. Search engines plays a key role in the achievement of the Web, search engines assists several internet users to quickly find appropriate information. However the unsettled troubles of existing search engines have created the way for the growth of the semantic Web. In the environment of semantic Web, the search engines are supposed to be more helpful and well-organized for searching the appropriate Web information. In this paper the author have formulated the architecture of a semantic search engine, and this paper shows how the fundamental elements of the semantic search engine can be used in the fundamental task of information retrieval. And then an enhanced algorithm based on TFIDF approach was developed to assure the retrieve information resources in a more effective way.

Yufei Li *et al*., [6] developed a relation-based search engine in semantic web. Due to the growth of the Web, information "Big Bang" has occurred on the Internet.

Search engines have turned out to be one of the most supportive tools for gathering helpful information from the Internet. However, instead of considering about the semantics of information, the machine on the existing web cares only about the location and display of information. Because of this limitation of the existing web, the search outcomes by even the most accepted search engines cannot generate acceptable results. The expansion of the next generation web, semantic Web, will turn the condition entirely. In this paper, the author developed a prototype relation-based search engine, "OntoLook," which has been applied in an essential semantic Web environment in lab. The author also presented its system architecture and evaluated the key technique.

T. Tomiyama *et al*., [7] developed concept-based web communities for Google™ search engine. The main goal of this paper is to build up an intelligent computer system with some deductive capabilities to theoretically group, match and rank pages according to predefined linguistic formulations and regulations defined by experts or based on a set of recognized homepages. The conceptual fuzzy set (CFS) technique will be utilized for intellectual information and knowledge recovery through conceptual matching of both text and links. The chosen query doesn't want to match the decision criteria accurately, which gives the system an additional human-like behavior. The technique supports the intelligent information and knowledge retrieval with the help of Web-connectivity-based clustering.

Hang Cui *et al*., [8] suggested hierarchical structural approach to improving the brows ability of web search engine results. Web users have been mostly depending on Web search engines to discover information of interest on the Web. Still, two main problems remain with conventional Web search engines: the brows ability of searching outputs and the capacity of Web coverage. The extended ranked list appearance of search results, which is extensively adopted by the industry, adds a layer of uncertainty to users, particularly when the number of matches returned from search engines can simply go beyond ten thousand levels. The authors proposed an agent system based on hierarchically structural method for arranging Web search results coupled with a metasearch method for Web searching. The metasearch method would assist in extracting the greatest of the Web from a bigger Web coverage; and this ontological approach is expected at presenting a method to classify search results in a semantic hierarchical association, and let users to discover objective information in an interactive manner.

Kyung-Joong Kim *et al*., [9] presented a personalized web search engine using fuzzy concept network with link structure. Many researches have been done on link-based search engines like Google and clever. These search engines uses link structure to discover accurate result. Generally, a link-based search engine creates advanced-quality results than a text-based search engine. However, they have difficulty in producing the result fit to a specific user's preference. Personalization is necessary to support a more suitable result. Many

approaches are available; among that the fuzzy concept networks depending on a user profile can characterize a user's subjective interest properly. In this paper, the author develops another search engine that utilizes the fuzzy concept network to personalize the outputs from a link-based search technique. The fuzzy concept network depending on a user profile rearranges five outputs of the link-based search engine, and the system offers a personalized elevated-quality result. Experimental observations with three subject's points out that the system developed searches not only appropriate but also personalized Web pages on a user's preference.

Development of a self-adaptive Web search engine was done by W. Zhang *et al*., [2]. Since the Web progresses towards the direction of offering greater extent information, finding the preferred information competently becomes an extremely significant problem. Web search engines are very helpful information search tools in the Internet. Existing Web search engines generates search results in accordance with the search terms and the needed information collected by them. While the selection of the search outputs cannot influence the future ones, they possibly will not cover most people's interests. The search results can be influenced by the response information created by users' accessing list. Thus it allows the search engines to offer self-adaptability.

L. Ahuja *et al*., [10] suggested a new expert web search engine for Web Environment. It applies a knowledge engineering based approach for the expansion of this expert system. To be aware of the fundamental functioning of search engine, a variety of Web Forums and Blogs have been considered. In this paper, the author proposes an Intelligent Agent and Interaction Agent dependent knowledge base of Search Engine. This knowledge based Search Engine technique will be more helpful in knowledge management and knowledge reuse. At user level, it can be utilized for predicting finest search results to the user and at organizational level it will be more helpful for drawing a variety of conclusions for managing quality database for enhanced application use.

A new graphical interface for web search engine was proposed by A. Amato *et al*., [11]. The main purpose of this paper is to develop a novel human computer interface for Web search engines. Even though there has been a remarkable development introduced in the Web search engines, their human interface still remain unexpectedly depends on a textual sorted list. The location of a site in this listing expresses its distance from the user's query. The authors developed a Graphical User Interface (GUI) for search engines depending on the geomorphologic metaphor. This interface enables the user to know about the semantic distribution of the Web sites retrieved by the search engine. The suggested interface is executed as a browser plug-in and it is capable to work with all the recent search engines.

Sungchae Lim *et al*., [12] proposed a hierarchical cache scheme for the large-scale web search engine. Numerous researches has been done over many years to provide solution for the technical challenge concerning the Web search engine, such as crawling Web documents, better performance indexes, and ranking systems using hyperlink analysis. In this approach, the authors proposed a distributed architecture for the query processing method and its hierarchal cache scheme. This approach is based on the commercial Web search engine intended to answer 5 million user queries against over 6.5 million Web pages per day. With the help of the hierarchal cache scheme, a section of query results is kept in multi-level caches with the intention that excessive I/O or CPU time is not utilized for query processing. It is possible to lessen around 70% of the server costs by using this scheme.

E.D. Sciascio *et al*., [13] has designed and implemented the Web-search engine based on computation tree logic. In this paper, the authors presented the design of a entire Web engine for text search and retrieval, which is based on CTL. The temporal logic is assumed to define the syntax of a structural query to be development on the graph model of a Web site. The influence of this query language relies on its capability to look for the text that the user desires not only inside one single document but on the complete structure of the site. The method performs as a search engine that offers the option to pose a query of mounting level of complexity. The outcome of the experiments proved to be selective and accurate enough concerning user expectations.

D. Bollegala *et al*., [3] suggested a Web Search Engine-based approach to measure semantic similarity between words. Computing the semantic resemblance among the words is a significant component in different tasks on the web such as relation extraction, community mining, document clustering, and automatic Meta data extraction. Even with the effectiveness of semantic resemblance measures in these applications, perfectly measuring semantic resemblance between two words remains a difficult task. The authors proposed a semantic resemblance measure using page counts and text snippets obtained from a Web search engine for two words. In particular, the authors define a variety of word co-occurrence measures using page counts and combine those with lexical patterns extracted from text snippets. To recognize the various semantic relations that exist among two given words, proposed a new technique called pattern extraction algorithm and a pattern clustering algorithm. The best possible grouping of page counts-based co-occurrence measures and lexical pattern clusters is studied using support vector machines (SVM). The developed search engine provides better performance than the previously proposed web-based semantic resemblance measures on three benchmark datasets showing a high correlation with human ratings. Furthermore, the proposed semantic resemblance measure considerably improves the accuracy in a community mining task.

D. Celik *et al*., [14] proposed a semantic search agent approach: finding appropriate semantic Web services based on user request term. This paper recommends a searching method to determine semantic Web services satisfying user requirements. The growth in Web services and need of semantic base in search

mechanisms of UDDI make it complicated for users to find an essential Web service. It is proposed to formulate a semantic search agent (SSA) to find out necessary Web services from Web. The system utilizes OWL-S for illustrating the semantics of Web services and determines an appropriate semantic Web services with the help of these semantic descriptions. OWL-S permits semantic description of Web services and therefore, the semantic search agent is capable to recognize predefined concepts of semantic Web services extract required information and decide on the requirement of a service for a user. The semantic search agent intermingle with user and semantic Web services.

Jiang Huiping [15] put forth information retrieval and the semantic web. Information Retrieval towards the semantic web has turn out to be one of the inspirations of semantic web because it was initiated by Berners-Lee. The author presents a semantic web search technique to improve the efficiency and accuracy of information retrieval for unstructured and semi-structured data. For the purpose of increasing the system's scalability, the author utilizes RDF knowledge dependent to accumulate metadata in the proposed systems. Additionally, the author presents a Ranking assessor to compute the similarity among data with semantic data for quick and accurate data retrieval. Outstandingly, the system provides accurate solution to accurate question with the involvement of Ranking Predictor. Contrasting to existing techniques, one more significant thought presents in this paper is that the author utilizes a Search Arbiter to review whether the doubt is cleared by Keyword-dependent Search Engine or Ontology Search Engine that is dependent on whether there is not adequate ontology knowledge or not. In addition, key methods are conversed in this paper. It is supposed that the proposed technique can appropriately be utilized to semantic web for information retrieval.

R. Singh *et al*., [16] proposed SCHISM-A Web search engine using semantic taxonomy. The most of the existing search engines produce a large list in response to a user query. This outcome is usually ranked with the help of ranking criteria like page rank or relevancy to the query. On the other hand, this list is exceptionally difficult to users, because it expects the user to glance into every page successively in a comprehensive way to determine the appropriate data. As an outcome, most users just look for initial little Web pages on the list. Therefore various other related data can be ignored. The clustering technique is will provide requirements to deal with this difficulty. As an alternative of a sequential list, it clusters the search outcome into groups and labels these with delegate words for every group. These labeled groups of search outcome are uncovered to users. The clustering technique offers advantages by means minimized size of data supplied to the end users.

Web caching in semantic web based multiple search engines was formulated by M. Rajaram *et al.*, [17]. The World Wide Web is an international information space. Due to the remarkable development of information presented to end users with the advancement of the Web, search engines play a key role. Because of their general-purpose technique, it is constantly less infrequent that obtained result sets offer a burden of ineffective pages. The future generation Web architecture, characterized by the Semantic Web, offers the layered architecture probably permitting overcoming this limitation. The ontology for several search engines is coded such that in this search engine for single query the last result is obtained from multiple search engines. Clustering can be done after receiving the user query outcome. In this clustering approach the user query outcomes is sorted in the a to z form, the numerous search engines have been developed, which permit growing information retrieval accuracy by developing a key content of Semantic Web resources, i.e., relations. The web cache optimization can be utilized in search engine to obtain fast retrieval of user query outputs. In this paper, the authors have used web cache optimization depends on eviction technique for semantic web search engine. Also, analization of both merits and demerits of a few existing Web cache replacement techniques comprising lowest relative value algorithm, least weighted usage algorithm and least unified-value algorithm is performed. Depending upon the analysis of results, the authors have used least grade replacement (LGR) algorithm in this technique which takes recency, frequency, ideal history, and document size into account for Web cache optimization. When the set is filled, the least grade document will be substituted, but its grade will be stored in an ideal-history grade collection for future references.

3. PROBLEMS AND DIRECTIONS

There are various limitations and drawbacks in most of the above discussed existing techniques. Hence, novel and efficient techniques are necessary to overcome those limitations and drawbacks. Researchers may concentrate on the following areas to provide the best Web search engine.

Semantic web

The Semantic Web is well-known for being a web of Semantic Web Documents (SWD); but, the structure or growth of the semantic web is not fully known. Most of the available search engines, offers poor support to accessing the web of result’s and no effort is made to acquire the benefits of the structural and semantic information programmed in SWDs. The Semantic Web will recommend the approach for solving the problem of existing approaches at the architecture level. Every page in semantic web contains semantic metadata that holds additional informations about the Web page.

Multiple search engine

Multiple search engines permits to make use of many search engines at the same time. Some of the search engines are predominantly efficient and complicated. But no search engine is entirely comprehensive. In many search engines, they possibly will use only a small

database from which set of outputs are retrieved or those databases are not updated regularly.

As a result, the user will get satisfied only by searching in many search engines to make sure that the whole thing was covered under a specific topic. Consequently, a multi-search engine possibly will reduce the difficulty of visiting a variety of different sites. Few search engines might be more powerful in searching a particular topic and some other search engine would be more stronger in some other topics. So, to generate a most powerful search engine, it is necessary to integrate those search engines and create a multiple search engine.

4. CONCLUSIONS

In recent business trends, development of search engines is an important business across the world. Internet beginners can start using internet only with the help of search engines. Also the internet experts still depend on the search engines to look for the informations like textual data, images, videos, etc. So it is very true that almost all the internet users depend on the search engines to find the relevant information according to their needs. The web specialists have confirmed that all the Web search engines are not up-to-date and also approximated that the biggest search engines cover only 15% of the World Wide Web. Many search engine techniques have discussed in the literature survey will help the researchers to develop the perfect search engine which provides only the relevant information and must satisfy all the users in all the field of information.

REFERENCES

[1] Yan Li, Xin-Zhong Chen and Bing-Ru Yang. 2002. Research on Web mining-based intelligent search engine. International Conference on Machine Learning and Cybernetics. 1: 386-390.

[2] W. Zhang, B. Xu and H. Yang. 2001. Development of a self-adaptive Web search engine. Proceedings of 3rd International Workshop on Web Site Evolution. pp. 86-93.

[3] D. Bollegala, Y. Matsuo and M. Ishizuka. 2010. A Web Search Engine-based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering. PP (99): 1.

[4] F. Shaikh, U.A. Siddiqui, I. Shahzadi, S.I. Jami, and Z.A. Shaikh. 2010. SWISE: Semantic Web based intelligent search engine. International Conference on Information and Emerging Technologies (ICIET).

[5] Yi Jin, Zhuying Lin and Hongwei Lin. 2008. The Research of Search Engine Based on Semantic Web. International Symposium on Intelligent Information Technology Application Workshops, (IITAW '08). pp. 360-363.

[6] Yufei Li, Yuan Wang and Xiaotao Huang. 2007. A Relation-Based Search Engine in Semantic Web. IEEE Transactions on Knowledge and Data Engineering. pp. 273-282.

[7] T. Tomiyama, R. Ohgaya, A. Shinmura, T. Kawabata, T. Takagi and M. Nikravesh. 2003. Concept-based Web communities for Google™ search engine. 12th IEEE International Conference on Fuzzy Systems (FUZZ '03). 2: 1122- 1128.

[8] Hang Cui and O.R. Zaine. 2001. Hierarchical structural approach to improving the browsability of Web search engine results. Proceedings of 12th International Workshop on Database and Expert Systems Applications. pp. 956-960.

[9] Kyung-Joong Kim and Sung-Bae Cho. 2001. A personalized Web search engine using fuzzy concept network with link structure. IFSA World Congress and 20th NAFIPS International Conference. 1: 81-86.

[10] L. Ahuja and E. Kumar. 2010. Development of expert search engine for web environment. 2nd IEEE International Conference on Information Management and Engineering (ICIME). pp. 288-291.

[11] A. Mato, V. Di Lecce and V. Piuri V. 2007. A New Graphical Interface for Web Search Engine. IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems (VECIMS 2007). pp. 42-46.

[12] Sungchae Lim and Joonseon Ahn. 2008. A Hierarchical Cache Scheme for the Large-scale Web Search Engine. 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing (SNPD '08). pp. 925-930.

[13] E.D. Sciascio, F.M. Donini, M. Mongiello and G. Piscitelli. 2004. Design and implementation of a Web-search engine based on computation tree logic. Proceedings of the 12th IEEE Mediterranean Electro technical Conference (MELECON). 2: 705-708.

[14] D. Celik and A. Elgi. 2005. A semantic search agent approach: finding appropriate semantic Web services based on user request term(s). ITI 3rd International Conference on Information and Communications Technology, Enabling Technologies for the New Knowledge Society. pp. 675-687.

[15] Jiang Huiping. 2010. Information Retrieval and the semantic web. International Conference on Educational and Information Technology (ICEIT). 3: 461-463.

[16] R. Singh, D. Dhingra and A. Arora. 2010. SCHISM-A Web search engine using semantic taxonomy. IEEE Potentials. 29(5): 36-40.

[17] M. Rajaram and S.L.S. Vadivu. 2010. Web caching in Semantic Web based multiple search engines. IEEE International Conference on Computational Intelligence and Computing Research (ICCIC). pp. 1-7.