Journal of Education and Vocational Research
Vol. 3, No. 6, pp. 199-203, June 2012 (ISSN 2221-2590)

Prototype Model Agent-Based Search Engine for Researchers and Scientists

*Naveed Dastgir, Muhammad Naeem Ahmed Khan
Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad, Pakistan
*chheena@acrologix.com

Abstract: Now days, the Word Wide Web (WWW) covers most of the information channels across the world and is the cheapest resource to publish the corporate as well as individual information. The key advantage of WWW is that the published information instantly becomes available around the globe. Since the information load on the Internet is increasing day by day, therefore, it is causing serious troubles for the researchers and scientists to retrieve the targeted and relevant information from the huge bulk of data. While surfing the Internet, precious time of researchers and scientists is wasted due to browsing of the irrelevant and unnecessary material. Such an approach of information retrieval results in overlooking the important contents. In this paper, we look into the elementary components of a customized search engine in order to develop an agent-based search engine tool that may help researchers and scientists to find their desired information in an efficient manner and with minimal clicks of pointer. Through this tool, the researchers and scientists will be able to find a summarized report against their search phrase along with the search details, which will provide them a solid background to their research subject based on history available on the WWW. The important aspect of this search engine is that it retrieves the subject-specific material from the designated websites in accordance with the user-defined criteria.

Keywords: *Wild Wide Web (WWW), Search Engine, Efficient Search, Context Driven Search, Web Crawler*

I. Introduction

Online academic and research resources are increasing day by day on the World Wide Web (WWW) and this huge amount of information that could add up the knowledge of researchers/scientists in their specific area of interest raises new challenges for researchers/scientist. To find the relevant information from the huge pile of data banks not only wastes a lot of their precious time but also results in the hassle of sifting through the extraneous stuff. Researchers/scientist search and examine many web pages, which occasionally may not be of their interest, and such situations increase the degree of research complexity as well as consumption of precious time. Web-based search engines like Google, Yahoo, Bing, Swoogle etc. (Ding et al., 2004) though help researchers to find the desired information, but the search engines return enormous information — most of which is irrelevant — and researchers/scientists have to go through it in order to sift the more specific and meaningful stuff out of it. Hence, search engines alone do not appear to be a suitable and subtle approach. To discover an appropriate solution for these challenges, we intend to propose a conceptual structure of an agent-based search engine that is capable of minimizing the search time and enhances the search preciseness by sorting and summering the web information and making it available to the researchers in the form of a short and an abridged report. The idea to develop an agent-based search engine for researchers and scientist is based on the concept that it would be capable to produce the summery report against the query generated by the researcher/scientist in accordance with the context of his/her area of interest along with the geographical location related to the sought information. This search engine would be capable to provide multi-step results; in the first step, it is envisaged to provide the summery of the search result and in the next step it is anticipated to provide the detail itemized results of the search query; thus, decreasing the search time. Further, the search engine sorts the information according to *location*, *technology*, *institution* and *research centers* etc. This approach may also help researchers to find the gaps and the perspective future work in a specific research area along with the available/existing knowledge. Agent-based search engine for researcher and scientists first finds the information available on WWW by crawling through the entire web. The retrieved information is then evaluated both quantitatively and qualitatively by employing specific search functions. Based on the search results, the sorting functions organize the information in the desired order and the report preparation functions format the information in the shape of

shortened report. Since data, mining techniques are based on some context (Lohani & Jeevan, 2007) and we have applied this principle in formulating the structure of the search engine. The agent-based search engine organizes the search results report based on researcher's context by applying data mining techniques. The context of required information helps the search engine to find the appropriate and meaningful information from the bulk data and filters the insignificant information automatically. The paper is organized into eight sections. First and second sections describe the background and core concepts of search engines. In section III, context of the search engine is provided. In section IV, resource-ranking algorithms are discussed and in section V, the expected results generated by the search engine are discussed. The authority of research resources are discoursed in section VI. The related work appeared in the contemporary research is summarized in section VII and the final section elaborates the prospective future work and concludes the research work.

2. Search Crawler and Spider

A *spider* in the search engine is used to find the information resources (i.e., websites) and a *crawler* creeps through the required information with the individual information resource. In this scenario, one more tool is required to analyze the substantiation and relevance of the resource; and, in fact, this is the first step towards sorting and sifting the required information. The resource that does not contain the information as per the supplied criteria is dropped from the search results. The search engines available on the web also provide the application program interface (API) to facilitate the searching functionality. Google can be considered as a one of such tool to support information retrieval (Tho et al., 2007). The online available APIs could be used instead of building the new crawler and spider, but there is a caveat that the underlying techniques for most of such APIs have never been published or made public (Cho & Tomkins, 2007). Therefore, a new crawler/spider is always needed to be fabricated while developing a search engine to ensure reliability of the online contents.

Context of Search: The context-based search raises many issues in the search patterns. For researchers and scientists, context is very important since it plays a pivotal role to sift data in accordance with their respective area of interest. An overview of different context issues related to this study is given below.

- **User Context Based Search:** User portfolio analysis defines the context of a user along with the user behavior; and based on user context, the appropriate results are presented to the user.
- **Geographical Based Searc:** User portfolio analysis also describes the user location. The resource analysis defines the resource location that sorts the searched results according to the geographical location of the researcher or scientist as well as the location where the resource is available. Moreover, researchers can also stipulate the specific geographical location to the search engine to seek the search result on the context of geographical location. Based on the specified area, the search engine can adjust the search results and displays them as per user's preferences.
- **General Search:** A user without proper portfolio is treated as a general/ordinary user in our proposed search engine structure. For the general users, all the sorting and filtering functions are disabled and an ordinary user only gets a general result report.

3. Results

The search results can be ranked based on clicks behavior and metadata available on the online resource (Cho & Tomkins, 2007). In the intelligent search engines, the users of common interests help each other to find the more appropriate results; therefore, knowledge extracted by one researcher can endorse the ranking of result for the other researcher (Birukov et al., 2005). Furthermore, the bookmarking and/or tagging could help the search engine for ranking the search results (Ding et al., 2004; Cho & Tomkins, 2007; Farooq et al., 2007). Along with all these techniques, context and geographical location of the researcher are also important parameters to sort the search results and generating appropriate report. The sorting algorithms endorse the search engine functionality by minimizing and ordering the corresponding results against the specific query.

Reporting: Reporting is the core part of search engine and is responsible for arranging the search results in a precise manner. In case of any inconsistency or omission in the searched information, the retrieval report would be misleading; therefore, more attention is required to be paid at the time of data collection and data

handling processes. Once the researchers submit a query to the search engine, the report is populated with different parameters e.g., area of excellence, area of deficiency, research centers and industry based on researcher's context e.g., geographical location etc.

Authenticity of Search Source: Some of the social media resources do not carry meaningful information for the researchers or scientists as there is always doubts about the authenticity of information due to the fact that most of the information do not provide proper source and citation of the data and do not identify either the group of intended users who can benefit from this information. The information is hardly filtered and approved by the authority and this information is usually presented based on current trends, therefore, these resources needs to be skipped while searching the consequential and specific information (Cho & Tomkins, 2007; Heymann et al., 2007).

Related Work: WWW has emerged as a biggest source of information as well as a vibrant platform to run the businesses across the world. Many attempts by the researchers are made to make the online system to provide meaningful information by developing efficient search engines. According to Peristeras et al., (2009), the governments around the world have made huge investments in information and communication technologies (ICT), but there are still many difficulties being faced to successfully run the automation application. Now e-government and e-participation refocus the governments towards their citizens and rule of business. There still exists a challenge for the governments how ICT can be beneficial to their citizens and businesses through better use of "Intelligent Technologies" such as the semantic web, service oriented architectures (SOAs), Web 2.0 etc. Researchers of social computing also propose that coupling of *conventional* intelligent technologies with the latest searching techniques are required to be improved e.g., Natural Language Processing (NLP) with the support of argument representation, visualization and opinion processing **(**Peristeras et al., 2009) can be more beneficial for efficient use of web services. NLP can provide more precise and accurate results, which can save a lot of time of researcher/scientist by elaborating the search results in a precise manner. Liliana (2010) conducted research on the intent and objective of a user while formulating the search query — which is an enormous challenge for the new information retrieval systems. The user intent behind the search query is needed to be studied with respect to different dimensions as the end results of most of the queries lie in one or two dimensions in addition to their relationships and dependencies. Hence, the query dimensions are also useful for searching information on the web to understand the user intents and requirements. In this context, most of the user intents in our proposed search engine model are addressed automatically because researchers and scientists have been considered as the sole users of the system. Durgin & Sherif (2008) endorsed that distribution of information skittered on the web requires autonomous evolution of each of the available resource for a declaratively object by encoding procedures of semantic web which could be understandable by the machine. Therefore, an ontology-based standard is required in the semantic web technology. Currently such functions are being provided through extensible mark-up language (XML) by using XML to present the syntax structure for the semantic web that comprises subject, verb and object. Uniform Resource Identifier (URI) presents the subject and verb, and by using URI, the uniqueness of an object is defined (Durgin & Sherif, 2008). Semantic web applications require more recent data to be collected against each query even though each user context requires crawling information afresh.

Cho (2007) researched on the intelligent agents for searching the contents on the social media along with the contextual challenges faced due to the search criteria. The information available on the social media is ordinarily not from the reliable resources as it mostly contains information of users who just spend spare time on such media and leave their own personal and subjective views that seriously lack authenticity. Most of such platforms carry information, which are based on the visitors' interest as well as hot topics related to politics and social issues. Such a situation results in posing a challenge for the search engine to draw any context out of the contents found on the social media (Cho & Tomkins, 2007). Social media contents are hardly of any concern or relevance to the researchers or scientists. Sivashanmugatn (2005) descried the scenario to build standards on the semantic web technology and to monitor the semantic web page. Semantic web pages change the contents on the basis of semantic process templates to accommodate the participation of activities, controls, calculations and conditions and present them into the interface as this defines the quality of service (QoS) of the current activates. In this regard, it is not recommended to use strong coupling between the web services and process, however, the services collaboration along with the semantic process

templates for the required process could be beneficial (Sivashanmugatn et al., 2005). Search engine are required to be updated based on changes made on the web services, which are often periodical. Tho (2007) presented a survey of the research publications' searching. Birukov et al. (2005) recommended an agent based search engine using the techniques of data mining and users' behaviors. Dikaiakos (2009) contributed the idea of cloud computing as the future of computing. It is predicted that desktop PCs would transform into large data centers due to technological advancements, therefore, it is envisaged that investment on the hardware and software could be more beneficial. Heymann (2007) classified the spam handling functions related to the social media searches into three categories: detection, demotion and prevention. The best solution to contain spam is to delete the spam trough bookmarking. Hepp (2006) criticized publishers for not considering the semantic web or semantic web service publications as the part of publication and the author. Lond (2003) define the intelligent system with the help of computing by arguing that the approximation is a 'soft' concept and the capability to approximate for the purposes of comparison, pattern recognition, reasoning and decision-making is a manifestation of intelligence. It is therefore suggested to use soft computing to build the intelligent machines. Farooq et al., (2007) compared new and the old technologies i.e., social bookmarking and tagging which are used to retrieve the users' intended information from the bulk of data; and this particular research contributes towards the idea of our proposed search engine. However, a challenge remains there how to define a methodology that outlines detailed mechanism to process the information according to its meanings. Most of the researchers consider the web information as a data or metadata for the search engine along with some kind of information attenuation algorithms to furnish only the more specific search results.

4. Conclusion and Future Work

This search engine model proposed in this paper is meant for facilitating researchers and scientists to fetch the most relevant material from the net in accordance with the context of their search query. The search engine is also capable to analyze the status of censored topics. Search engine can also be of much use for marketing products by providing venders right information that helps avert overlooking of the product web pages by a crawler. The proposed model preserves its identity because some of search engines bypass certain web resources that carry useful information. Further, most of the search engines do not assign much weight to the authenticity of the information or prefer fetching data from authenticated web sources. As a prospective future work, we intend to develop the search engine in lines with the methodology outlined in this paper. The search engine will have dual functionality, firstly, it will serve the purpose of a search engine and secondly, it would be part of quality deficiency reporting and investigation system that aims as assisting users to find the deficiencies within the search results.

References

.

Birukov, A., Blanzieri, E. & Giorgini, P. (2005). Implicit: An Agent Based Recommendation System for Web Search. Association for Computing Machinery (ACM), Utrecht, Netherlands.

Cho, J. & Tomkins, A. (2007). Social Media and Search. IEEE. IEEE Computer Society.

Dikaiakos, D., Pallis, G., Katsaros, D., Mehra, P. & Vakali, A. (2009). Cloud Computing Distributed Internet Computing for IT and Scientific Research. IEEE INTERNET COMPUTING.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. C. & Sachs, J. (2004). Swoogle: a search and metadata engine for the semantic web. In: 13th ACM Conference on Information and Knowledge Management.

Durgin, K. & Sherif, S. (2008). The semantic web: a catalyst for future e-business. *Emerald Group Publishing Limited*, 37(1), 49-65.

Farooq, U., Song, Y., Carroll, J. M. & Giles, C. L. (2007). Social Bookmarking for Scholarly Digital Libraries. IEEE Computer Society, 1089-7801/072007.

Hepp, M. (2006). Semantic Web and Semantic Web Services Father and Son or Indivisible Twins? IEEE, IEEE INTERNET COMPUTING.

Heymann, P., Koutrika, G. & Garcia-Molina, H. (2007). Fighting Spam on Social Web Sites: A Survey of Approaches and Future Challenges. *IEEE Internet Computing*, 11(6), 36-45.

Lond, A. (2003). The role of soft computing in intelligent machines. *Physical and Engineering Sciences*, 361(1809).

Li, Q., Lau, W. H., Leung, W. C., Li, F., Lee, V., Wah, W. & Ashman, H. (2009). Emerging Internet Technologies for E-Learning.

Liliana, C. & Ricardo, L. (2010). Towards a Deeper Understanding of the User's Query Intent. Workshop on Query Representation and Understanding Geneva, Switzerland, SIGIR'2010.

Lohani, M. & Jeevan, V. K. J. (2007). Intelligent software agents for library applications Library Management. *Emerald Group Publishing Limited*, 28(3), 139-151.

Peristeras, V., Mentzas, G., Tarabanis, A. & Abecker, A. (2009). Transforming E-government and E-participation through IT. IEEE INTELLIGENT SYSTEMS, 1541-1672.

Sivashanmugatn, K., Miller, J. A., Sheth, A. P. & Verma, K. (2005). Framework for Semantic Web Process Composition. *International Journal of Electronic Commerce,* 9(2), 71-106.

Tho, Q. T., Fong, A. C. M. & Hui, S. C. (2007). A scholarly semantic web system for advanced search functions. *Emerald Group Publishing Limited*, 31(3), 353-364.