A Real Time Visual Exploratory Search Engine for Information Retrieval in a Cloud

Mohammed Najah Mahdi, Abdul Rahim Ahmad, and Roslan Ismail

Abstract—**Search engines employed under cloud computing have been crucial in enabling users in retrieving information. This has allowed users to access wide volume of information that is significantly impossible under the tradition information collection that involved manual search of the intended information. The main focus of this study is on improving the traditional search engine by providing a novel framework for people to obtain reliable, personalization and graphical representation search results. This study proposes and develops a Visual Exploratory search engine solution extremely good in understanding semantic similarity computation which can access multiple Metadata based search engines at the same time and generating real-time relationship graphs for a better visual exploratory search in a cloud computing environment. The study proposes a Research Model based on which the proposed prototype would be developed.**

Index Terms—**Cloud computing, exploratory search, meta-search, search engine.**

I. Introduction

Searching the internet has become a part of everybody's life. The development in technology and the various technological tools have changed the way in which the users search the web. Search engine technology has become ubiquitous, providing a standard interface to the endless amount of information that the web contains [1]. According to a statement by Levene [1] search engines have been delivering a simultaneous stream of innovative results and satisfying their users by providing the accurate results which they are looking for. Most of the search engines ensure this by implementing advanced information retrieval algorithms and other distributed architectures.

Search and navigation technologies are central to the smooth operation of the web and it is hard to imagine finding information without them. Understanding the computational basis of these technologies and the models underlying them is of paramount importance both for IT students and practitioners [1], [2].

II. Problem Statement

Searching the web and retrieving the information has become one of the complex processes in the recent days. A number of new websites appear every day and old ones change frequently or disappear over a period of time and in general the content is emergent rather than planned. This clearly states that the results generated from the search engines are not stable and the users have to use different strategies to satisfy them self. Basex claims that "According to our latest research Information Overload costs the U.S. economy a minimum of $900 billion per year in lowered employee productivity and reduced innovation. Despite its heft, this is a fairly conservative number and reflects the loss of 25% of the knowledge worker’s day to the problem. The total could be as high as $1 trillion" [3].

White and Roth [2] emphasize that there is a tremendous need to support the search behaviors of the users in a more detailed manner where by the algorithms should be able to cater both in technological and human contexts to achieve the best of information retrieval. Whereas researchers like Tang [4] have mentioned that still the present do search engines are unclear and the users have to figure out the semantic relationships by themselves by applying various efforts.

Another challenge that is experienced under the traditional search engine framework is the contradictory, low quality, noisy and unreliable content [5]. Information retrieval will reasonably be of high quality content if the documents of a given collection are accurate and authoritative. In addition, the technique design of the collection should be able to incorporate the potential of content that is of low quality. However, the nature of democracy that is experienced in web creation has led to a mass that is of poor quality and noisy fundamentally [6]. This is because the traditional search engines fail in ensuring that typical documents quality assumption is not done in isolation by synthesizing a large number of documents that are of low quality in order to provide results of best set [7]. Even though link-based approaches such as PageRank estimate web pages quality, the use of web link structure cannot optimally guarantee quality of the web page [8]. An optimal solution will be applying information of different pages for correlation purpose and within the page.

Duplication of hosts is another major problem that is faced under the traditional search engine framework. Predicting the potential of a search engine been duplicated under the traditional framework is a problematic work [7]. This is even complicated in locating the duplicate hosts serving similar content. This is because the artifact of domain name system (DNS) under the traditional framework of search engines can allow resolution of two hostnames under one physical engine.

Apart from the issues mentioned above the user visual exploratory search is constrained to the relationship graphs inferred off-line based on prior domain knowledge which is an disadvantage in itself and this needs to be produced in real time mainly for time inference of relationship graphs for

Manuscript received October 30, 2014; revised March 5, 2015.

Mohammed Najah Mahdi is with the College of Graduate Studies, University Tenaga National, Malaysia (e-mail: moh12004@gmail.com).

Abdul Rahim Ahmad and Roslan Ismail are with the College of Information Technology, University Tenaga National, Malaysia (e-mail: abdrahim@uniten.edu.my, roslan@uniten.edu.my).

DOI: 10.7763/IJFCC.2015.V4.388

online visual exploratory search services to users, when users input keywords that do not exist in off-line relationship graphs. In this scenario, the powerful cloud computing capability will play important roles. Making use of exploratory search with good semantic similarity computation annotations can overcome the existing issues mentioned above.

III. Objectives

1) To investigate the role of Metadata, Exploratory research and Semantic Similarity in generating better results from a visual exploratory search engine solution
2) To propose and develop a real time Visual exploratory search engine solution in cloud;
3) To develop an algorithm which is extremely good in understanding semantic similarity computation which can access multiple Meta data based search engines at the same time in cloud.
4) To propose a visual exploratory search engine solution based on the cloud computing model by providing a reliable, personalization and graphical result framework. The system infers the semantic results of Google, Gigablast and other search engine.
5) To develop real-time inference of relationship graphs for online visual exploratory search services to users, when users input keywords that do not exist in our off-line relationship graphs.

IV. Research Questions

1) What are the various challenges associated in relation to information retrieval from traditional search engines?
2) What are the various factors which influence the development of a Visual exploratory search engine in a Cloud?
3) What is the role of Metadata, Exploratory research and Semantic Similarity in generating better results from search engine?
4) What is the role of Cloud computing in developing a Visual exploratory search engine?
5) What are the relationship-based exploratory search and what the key advantages over traditional keyword-based lookup search in a variety of ways including rich semantics, learning and discovering, more relevance, personalization, and natural interaction.

V. Purpose of the Study

The purpose of this study is to propose and develop a Visual exploratory search engine solution extremely good in understanding semantic similarity computation which can access multiple Meta data based search engines at the same time and generating real-time relationship graphs for a better visual exploratory search in a cloud computing environment.

VI. Significance of the Study

The proposed study has both practical and theoretical significance. Theoretically the study contributes towards the understanding of Meta data, and Meta data related search engines by adding more information to the existing literature. Apart from this the study also focuses on semantic similarity and exploratory research related to information retrieval which is another theoretical contribution. The study contributes practically to the existing search engines by proposing a research model towards a visual exploratory search engine in a cloud environment which is real time. Studies in the past have developed such applications but these studies were offline and not real time based and part from that studies in the past had limitations in relation to semantic similarities which would also be enhanced in this study. In short the proposed Meta data search engine would be better in terms of results generation with better semantic similarity relationship among keywords and real-time graphs.

VII. Meta Data and Search Engines

Meng and Yu [9] state that a Meta data based search engine is a tool which supports unified access to various search engine systems at the same time. Basically in a Meta data based search engine there is a simple user interface to make the query which is again sent to multiple search engines and the returned responses are retrieved and merged in real time organized in a presentable manner. The process is deceptively simple, for a great number of technical, informatics and design issues need to be solved in order to make a practical meta-search engine Fig. 1 explains the architecture of Meta data search engine.

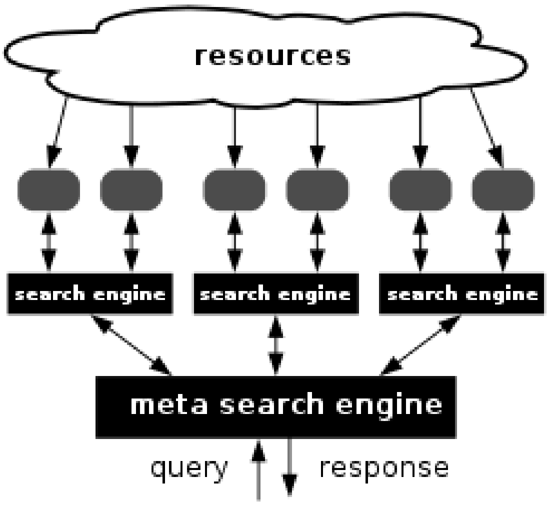

Fig. 1. Architecture of meta data search engine.

Dragut *et al.* [10] mentions that some of the technical issues related to Meta data based search engines are creating a universal search engine query interface from the individual based search engine interfaces. Apart from that Dragut [10] also mentions that search engines should ensure the queries submitted should be understood and meaningful results are provided in a timely manner.

The concept of metadata is essential in a number of disciplines, including Database Architecture and Library Science. In a cloud environment, service introspection is a must have requirement to establish meaningful relationships between cloud service consumers and providers. Metadata exchange refers to the automated exchange of information about a service offering during service setup and operations. Automated, standardized metadata exchange will become an essential ingredient in a functioning, scalable cloud service ecosystem. Under the current state of the art, when an application is implemented using outsourced cloud components there is inherently less transparency about the service components when services cross organizational boundaries as in private clouds, or even more so company boundaries. This circumstance makes it difficult to implement manageability policies for a composite application made of outsourced service components [11].

VIII. Cloud Computing

Different clients have different computing power such as iPad, Mobile Laptops, and PC and so on, so it was decided to put the servers on the cloud environment. The search results obtained from the cloud server need further post-processing before they are delivered to and displayed on the client smart phone and other devices. Since these pre-processing and post-processing steps of semantic relationship graphs need substantial amount of computing resources, the computing capacity of mobile smart phone is currently incapable of handle those computing-intensive steps. Therefore, we move these computing processes to the cloud servers. With the powerful computing capacity and high scalability of cloud computing services, our framework will achieves the real-time processing.

IX. Semantic Web

Semantic Web (SW) has been identified as the next generation vision of web that will allow computers to comprehend web pages information beyond a mere presentation to the end users [12]. This will transform the World Wide Web (www) to an intelligent web that is capable of enabling machines to understand and exchange information. Thus, probability of retrieving relevant information by users will significantly increase [13]. Semantic web initial waves will be experienced through web changes that will simplify internet surfing significantly. Nevertheless, implementation of the semantic web is likely to face difficulties due to different lexical processing between machines and human beings. The early use of words by human beings in their life enables them to make conclusions out of incomplete and irrelevant information that is different in machines [14]. Accordingly, it will require documents in computers that interpret necessary connection of words and logic to ensure understanding in computers.

Consequently, the web will require instruments that will enable the system to function automatically through cooperation and learning abilities. Software agents that possess automatic cooperation and learning abilities will be utilized in semantic web [15]. Owing to the large size of World Wide Web in a single cloud environment size, it will require a number of agents to retrieve information sought by the users. Each of the agents is designed to execute certain task. In addition, each of the agents assigned communicates to other agents to ensure it offers services beyond it capacity or collaborating as a group to execute complicated tasks.

X. Exploratory Research

Exploratory research is generally defined as information seeking problem context which is largely multi-faceted and multi-tactical. During the initial stages the term exploratory research was largely applied to scientific discovery but off late this has also been applied in exploratory tactics and used in information seeking preferences [16]. In exploratory search people usually submit a tentative query to get them near relevant documents then explore the environment to better understand how to exploit it, selectively seeking and passively obtaining cues about where their next steps lie. For the current study an exploratory search would be used for the purpose of metadata which would provide enough resource to carry out the project. For the purpose of the project a number of multiple search engines would beutilised like Google, Baidu, Infoseek, Alsta Vista and a special search engine which would be designed based on our Meta data engine. The following Fig. 2 explains the exploratory research process.

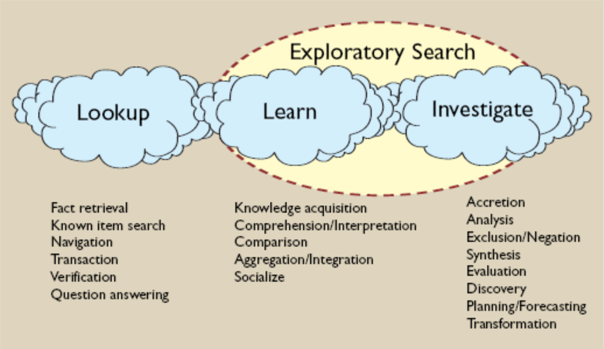

Fig. 2. Exploratory research (Marchionini, 2006).

After the results returned from all used component search engines are collected, the meta-search system merges the results into a single ranked list. Most of the current generation search engines present more informative search result records (SRRs) of retrieved results to the user.

XI. Proposed Approach

The study proposes the following model towards the development of the metadata based search engine. The user initially queries the system engine that interacts with Multiple popular common search engines such as Alsta Vista , Yahoo Infoseek, Baidu, Google, Specifically, the proposed meta-search engine calls APIs provided by these search engines by inputting the keywords and retrieving the return links and web pages. In the third step, the engine performs effective statistical natural language processing on the returned web pages and websites, and construct semantic relationship graphs.

This is where semantic annotations are used and then the data is filtered based on the semantic annotations which the previously developed system was lacking. Specifically, for all of returned web pages from the meta-search results, we measured their Semantic similarity for Visual exploratory search. As a result, in the constructed relationship graph, the nodes are represented by the keywords, and the edges are defined by their semantic relationship strengths. The semantic search engine provides fine-grained access to the filtered documents (the search stage). The two components share a store component, which stores all data: document contents, and metadata instantiations, and intermediate results created by the analysis and annotation components. Little [17] recommended in his study the use of Semantic Framework and semantic annotations. This study was more relevant to information retrieval among Videos and this study proposes to adopt the same concept along with the study of Tang [4] which had certain limitations.

XII. Proposed Framework

Based on the literature the following Fig. 3 is the proposed framework for the study:

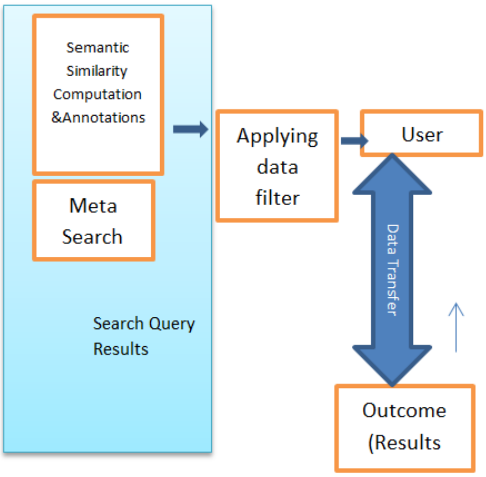

Fig. 3. Proposed model.

XIII. Findings

Based on the study conducted so far it was clearly identified that information retrieval from search engines is a complex process and a number of issues are surrounded to it. Traditional search machines that users have been using face a number of difficulties in improving or maintaining performance that is of quality [18]. These problems have been identified in numerous literature studies. One of the major challenges that search engines face is the existence of spam. Search results that are examined by the search engine users are highly the first page that is surfed. Empirical studies show that out of the 85% of the searches that are made by the users using the search engines; the first screen result is the only requested [5]. Consequently, a search result included in the first screen is likely to experience high traffic while a website excluded in the first screen that usually carries the top ten results will lead to a small proportion of users been linked.

XIV. Recommendations

The study recommends development of a Visual exploratory search engine solution extremely good in understanding semantic similarity computation which can access multiple Meta data based search engines at the same time and generating real-time relationship graphs for a better visual exploratory search in a cloud computing environment.

Studies in the past have developed such applications but these studies were offline and not real time based and part from that studies in the past had limitations in relation to semantic similarities which would also be enhanced in this study. In short the proposed Meta data search engine would be better in terms of results generation with better semantic similarity relationship among keywords and real-time graphs.

References

[1] M. Levene, *An Introduction to Search Engines and Web Navigation*, 2nd ed., John Wiley & Sons, Inc., October 18, 2010.

[2] R. White and R. Roth, *Exploratory Search: Beyond the Query-Response Paradigm, Synthesis Lectures on Information Concepts, Retrieval, and Services Morgan*, Claypool Publishers, 2009.

[3] J. Spira, "Information overload: Now $900 billion–What is your organization's exposure?" *Basex: Management Science for the Knowledge Economy*, 2008.

[4] J. Tang, X. Tang, and D. Chen, "A visual exploratory search engine solution based on cloud computing," in *Proc. 2012 Second International Conference on Cloud and Green Computing (CGC)*, 2012, pp. 368-374.

[5] C. Silverstein, M. Henzinger, and M. Moriez, "Analysis of a very large alta vista query log," *ACM SIGIR Forum*, vol. 33, no. 1, pp. 6-12, 1999.

[6] P. A. Buhler and J. M. Vidal. (2009). Towards adaptive workflow enactment using multiagent systems. *Information Technology and Management*. [Online]. Available: http://jmvidal.cse.sc.edu/papers/buhler03b.pdf

[7] D. Lewandowski. (2007). Web searching, search engines and information retrieval. [Online]. Available: http://eprints.rclis.org/6702/1/isu_preprint.pdf

[8] N. K. Salih, T. Zang, and A. A. Mohamed, "Autonomic management for multi-agent systems," *International Journal of Computer Science Issues*, vol. 8, no. 5, pp. 338-342, 2011.

[9] W. Meng and C. Yu, *Advanced Metasearch Engine Technology, Synthesis Lectures on Data Management*, San Rafael, CA: Morgan & Claypool, 2010.

[10] E. C. Dragut, W. Meng, and C. Yu, *Deep Web Query Interface Understanding and Integration, Synthesis Lectures on Data Management*, San Rafael, CA: Morgan & Claypool, 2012.

[11] D. S. Linthicum, *Cloud Computing and SOA Convergence in Your Enterprise-A Step-by-Step Guide*, Addison-Wesley, 2010.

[12] P. Hitzler, M. Krotzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*, New York: Wiley, 2011, ch. 1, pp. 11-12.

[13] T. Berners-Lee and J. Hendler, "The coming Internet revolution will profoundly affect scientific information," *Semantic Web*, vol. 410, pp. 1023-1024, 2001.

[14] H. Pascal. (February 13, 2013). Semantic data analytics-The key to challenging big data. [Online]. Available: http://knoesis.cs.wright.edu/faculty/pascal/pub/2013-02-Siemens-BIGDATA.pdf

[15] R. M. Fay and M. P. Sauers, *Semantic Web Technologies and Social Searching for Librarians*, New York: Library and Information Technology Association (U.S.), 2012.

[16] G. Marchionini, "Exploratory search: From finding to understanding," *Communications of the ACM*, vol. 49, no. 4, pp. 41-46, 2006.

[17] S. Little, K. Clawson†, A. Mereu, and A. Rodriguez, "Identifying and addressing challenges for search and analysis of disparate surveillance video archives," presented at 5th International Conference on Imaging for Crime Prevention and Detection ICDP-13, London, UK, Dec. 16-17, 2013.

[18] B. Chafik and K. Okba, "Agent approach for service discovery," *UbiCC Journal*, vol. 4, no. 3, pp. 509-515, 2009.

Mohammed Najah Mahdi was born in Baghdad, Iraq, in 1978. He received the B.E. degree in information technology engineering from the University of Baghdad, Iraq, in 2002, and the master degrees in information technology from the University of Malaya (MIT) Kuala Lumpur, Malaysia, in 2010 and currently he is doing his Ph.D at University Tenaga Nasional, Malaysia.

Abdul Rahim Ahmad received the B.Sc degree in computer science from the University of Queensland, Australia, in 1984, the master degrees in information technology from Loughborough University, UK in 1994 and a joint Ph.D from University of Nantes/Universiti Teknologi Malaysia in 2008. Currently he is the head of energy and environment at the Institute of Energy Policy and Research (IEPRe), University Tenaga Nasional, Malaysia. He has wide experience including in broadcasting, management and teaching. He has been active in IEEE Malaysia and was the chair for IEEE Computer Malaysia in 2013.

Roslan Ismail received the B.Sc degree in computer science from Universiti Putra Malaysia in 1989, the master degrees in computer science from Universiti Teknologi Malaysia in 1994 and the Ph.D from Queensland University of Technology in 2004. Currently he is the head of centre for Information and Network Security (COINS), University Tenaga Nasional, Malaysia. He has a broad working experience including in system development, auditor and teaching.