Interface for search engine query recording

Fadhilah Mat Yamin, Wan Hussain Wan Ishak
Universiti Utara Malaysia
Email: fmy@uum.edu.my, hussain@uum.edu.my

Abstract Search engine is a popular tool to search for information on the World Wide Web. Knowing and learning of what and how searchers are looking for information is one of the promising challenges in information science. However, as search log are recorded at the hosting server, obtaining the data is a tedious process. This paper discuss an approach to record user query to create the search log at user own server. This approach applies redirecting strategy to redirect query to the hosting server. An interface is developed to capture the query and record all the necessary details which include the date, time, searchers IP, session, and the query. The extensive report is then generated listing all the search details and the time spent. The report is a good source for those who are conducting search log related research.

INTRODUCTION

Information Retrieval (IR) is science of retrieving relevant document from document databases [1]. Since, its introduction many tools have been developed for the ease of information retrieving. The popularity of Internet and World Wide Web (WWW) for information storing, sharing and broadcasting, increase the need for a better IR system. In addition, varieties of information type, format and structure stored in WWW increase the complexity of IR system and its locating and retrieving become a very tedious process.

In 1990, the first web search tool called Archie has been introduced [2]. Since then, many other search engines have been developed. To date, search engine such as Yahoo! Search, MSN Search and Google, have been a popular tool for searching information on the WWW.

Web search engine is an interface to search for information on the WWW. This search engine contains a script that link the interface with the document databases. Document databases contain the representation of a documents or information which might include text, graphic, video and audio. Typical web search engine inherit the typical IR system as shown in Figure 1. The figure shows typical IR system that consists of three main components; input, processor, and output.

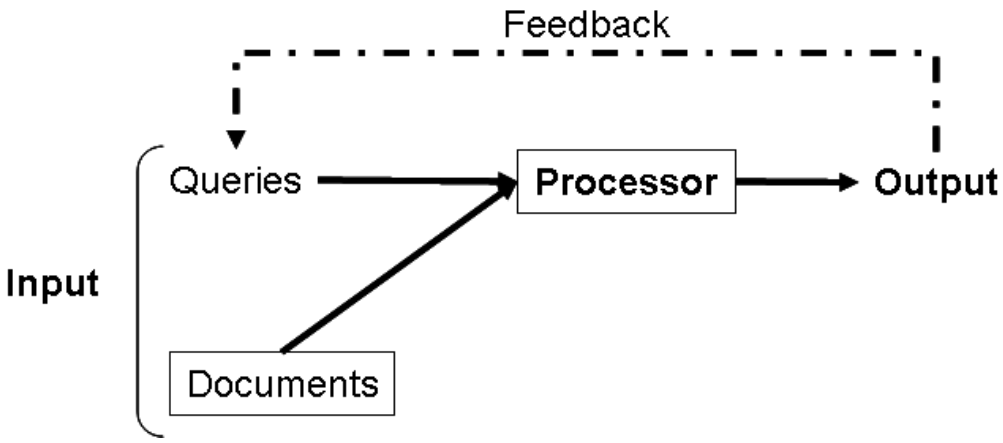


Figure 1: A Typical Information Retrieval System

The input consists of query and document. Both query and document are the representation of the original document such as a list of extracted words. The query is then process by the processor which is the engine that does the retrieval process such as structuring the information, matching and retrieving the match information. In order to best serve the users' need, a search engine must retrieve the most relevant information and then present that information in a friendly format [3]. Output is the outcome of the search system which usually is a set of citations, document reference number or hyperlink. The output serves as the feedback to user's query.

Searchers' activities and their interaction with the web search engine are usually recorded in the search engine server log. The log contains searchers details such as IP and session ID and information searched which includes the search streams, terms, operators and etc. Search log provide valuable information to the researchers that study the searchers searching behaviour, searchers search patterns, usage mining and etc. However obtaining the log is "expensive" as it was not intended for public access. Therefore, alternative method is proposed to record the usage and creating researcher's own search log.

SEARCH LOG AND RELATED STUDIES

Search log is a timestamps electronic record of interactions between searcher and the web search engine [4]. The information contains in the log that can be used to study and understand the searchers searching behaviour [4,5]. According to Wang et al [6], search log are accurate, unobtrusive, longitudinal, transactional, temporal and can be automatically collected and processed. In addition, Wang et al [6] also highlighted some disadvantages of search log such as the data being open to interpretation (accurate or not), privacy concern, and the vast amount of data gathered can be difficult to manage. However, these advantages are not the main issue in search log research. It depends on the method and skill of the researcher to analyze the log and to ensure that it is as accurate as it should be.

Studying searchers behaviour through web search log is vital especially to the search provider. The analysis provides in-depth understanding of the searchers requirement. Furthermore, it is well understood that the technology will fail if it does not reflect the user's need [5]. Thus, improvement can be made either improving the web search facility or improving the document database.

Currently, there are many studies incorporate search log to keep track user searching behaviour such as [5, 6,7,8] and etc. Some of the search dimensions are shown in Table 1.

Table 1: Study on the search log

Study	Search Dimensions
Madden et al., (2007)	Search length, search depth, intensity
Choo et al., (2000)	Menu choices, button bar selection, and keystroke action
Wang et al., (2000)	URL visited, continuous screen shots with a time stamp on a video tape, verbalization of thoughts recorded on the same video tape
Jansen (2006)	User identification, date, the time, search URL

SEARCH INTERFACE AND RECORDING

Google has been one of the popular primary search engine and top ranking world wide [9]. Its success was primary based on its research and underlying technology such as *PageRank* [10,11]. In order to study the searcher search behaviour, the searcher's search activities need to be recorded in the search log. According to [4], search log involves three major stages:

- Collection: the process of collecting the interaction data for a given period in a transaction log.
- Preparation: the process of cleaning and preparing the transaction log data for analysis
- Analysis: the process of analyzing the prepared data

To record the search activities, an interface has been developed. The interface is a layer between the searcher and the Google. As defined by [6], interface is a layer between the user and the system that facilitates human computer communication. Interface is like a bridge and it effectiveness plays a crucial role in the success of the interaction [5].

Figure 2 shows a model of the proposed search interface. Search interface consist of search interface engine and reporting module. Query entered by the searcher will be stored into a database and forwarded to Google. The query will not be modified. It will be forwarded as it is. Figure 3 shows the main page of the search interface. Figure 4 shows the example of the search results returned by Google. Besides query, the search interface will also record user's IP and session ID. The date and time of the search activity will be recorded at the database server.

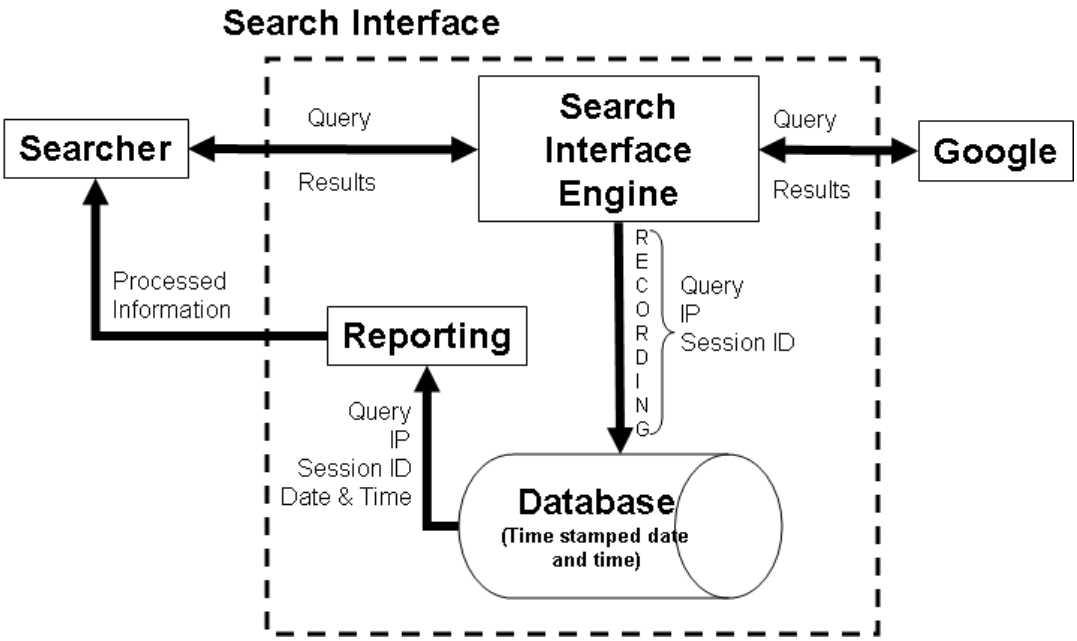


Figure 2: A Model of Search Interface

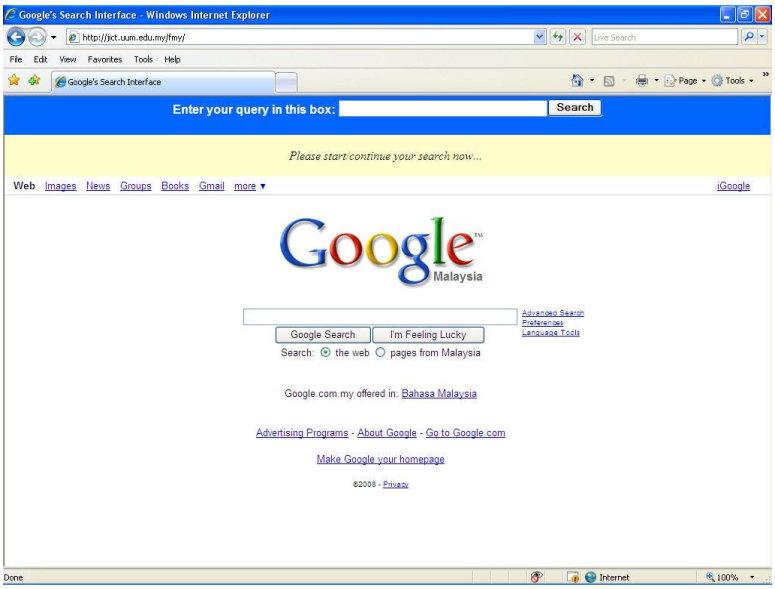


Figure 3: Main Page of the search interface

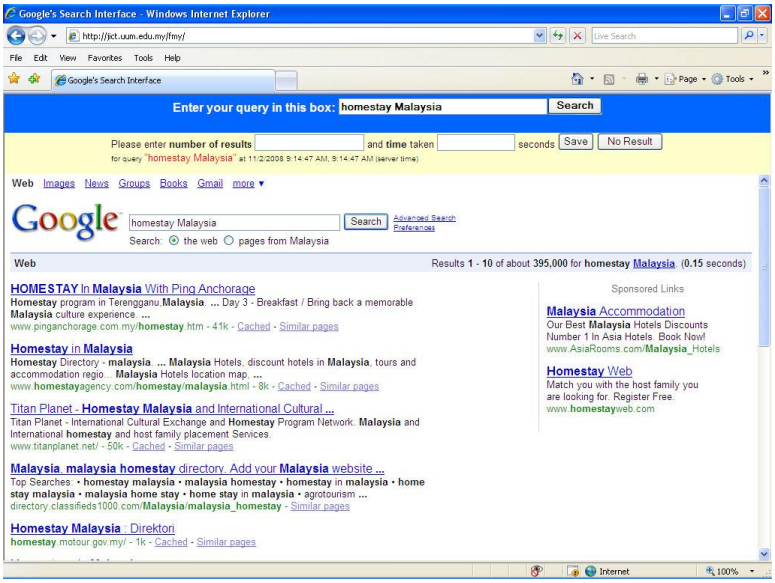
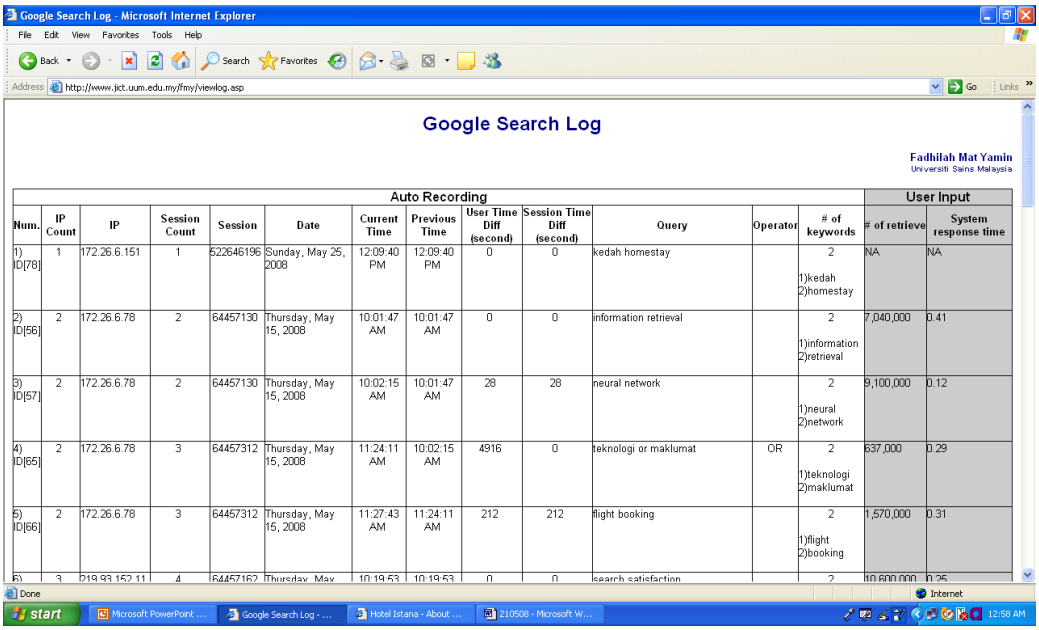


Figure 4: Example of the search result

In a reporting module (Figure 5), the engine will retrieve back all the data and process the query to extract the search term, operators, and the query length (i.e number of terms or words). If the searcher repeating the search activity. The module will calculate the search time different for the same session and overall search activities. This data will provide information on the length of search activity. Table 1 describe the attributes in the search log.



Auto Recording													User Input	
Num.	IP Count	IP	Session Count	Session	Date	Current Time	Previous Time	User Time Diff (second)	Session Time Diff (second)	Query	Operator	# of keywords	# of retrieve	System response time
1) ID[78]	1	172.26.6.151	1	522646196	Sunday, May 25, 2008	12:09:40 PM	12:09:40 PM	0	0	kedah homestay		2 1)kedah 2)homestay	NA	NA
2) ID[56]	2	172.26.6.78	2	64457130	Thursday, May 15, 2008	10:01:47 AM	10:01:47 AM	0	0	information retrieval		2 1)information 2)retrieval	7,040,000	0.41
3) ID[57]	2	172.26.6.78	2	64457130	Thursday, May 15, 2008	10:02:15 AM	10:01:47 AM	28	28	neural network		2 1)neural 2)network	9,100,000	0.12
4) ID[65]	2	172.26.6.78	3	64457312	Thursday, May 15, 2008	11:24:11 AM	10:02:15 AM	4916	0	teknologi or maklumat	OR	2 1)teknologi 2)maklumat	637,000	0.29
5) ID[66]	2	172.26.6.78	3	64457312	Thursday, May 15, 2008	11:27:43 AM	11:24:11 AM	212	212	flight booking		2 1)flight 2)booking	1,570,000	0.31
6)	3	219.93.152.11	4	64457162	Thursday, May	10:19:53	10:19:53	0	0	search satisfaction		2	10,600,000	0.25

Figure 5: Example of search log

Table 1: Description of the attributes

Attributes	Description	
IP Count	IP label/counter	Automatically recorded and extracted by the search interface engine
IP	User unique IP number as the identification ID	
Session Count	Search session label/counter	
Session	Search session ID. Unique session ID will be recorded for each search task	
Date	Date of the search	
Current Time	Time of the search	
Previous Time	Last search event perform by the same searcher (based on the IP)	
User Time Diff (second)	Different of time spend on search for the same searcher	
Session Time Diff (second)	Different of time spend on search for the same searcher under the same session	
Query	Searcher's query	
Operator	Operator extracted from the query	
# of keywords	Number of keyword in the query	
# of retrieve	Number of retrieved document results	Key in by searcher manually (see Figure 4)
System response time	The system response time	

CONCLUSION

The search interface proposed in this paper is to be used to study the searcher behaviour. The searcher will be asked to perform a search for a given topic. As the search log provides rich information on the search activity, the data will provide the insight on how the searcher performing the search and the strategy used in order to achieve the search goal.

The use of search interface to study the search behaviour is an alternative approach as the search log is very expensive to be obtained. There might exist some delay between the actual search time and the time recorded at the server which might reflect the speed. However, this limitation can be ignored as the concern is not the speed of the search task but the overall time spends on searching.

REFERENCES

1. Rijsbergen, C. J. (1979). *Information Retrieval System*. London: Butterworths
2. Wikipedia (2007). Web Search Engine. Retrieved from http://en.wikipedia.org/wiki/Web_search_engine on 21 Oct. 2007.
3. Sahami, M., Mittal, V., Baluja, S., & Rowley, H. (2004). The Happy Searcher: Challenges in Web Informaiton Retrieval. *The Eighth Pacific Rim International Conference on Artificial Intelligence* (PRICAI).
4. Jansen, B.J. Search Log Analysis: What is, what's been done, how to do it. *Library & Information Science Research, 28*, pp: 407-432.
5. Jin, Z., & Fine, S. (1996). The Effect of Human Behavior on the Design of an Information Retrieval System Interface. *International*

Information & Library Review, 28, pp: 249-260.

6. Wang, P., Hawk, W.B., & Tenopir, C. (2000). Users' Interaction with World Wide Web Resources: An Exploratory Study Using a Holistic Approach. *Information Processing and Management*, pp: 229-251.
7. Choo, C.W., Detlor, B., & Turnbull, D. (2000). Information Seeking on the Web: An Integrated Model of Browsing and Searching. *First Monday, 5(2)*.
8. Madden, A.D., Eaglestone, B., Ford, N.J., & Whittle, M. (2007). Search Engines: A First Step to Finding Information: Preliminary Findings from a Study of Observed Searched. *Information Research, 11(2)*.
9. Notess, G. R. (2007). Search Engine Showdown Reviews. Retrieved from htpp://searchengineshowdown.com /reviews/ on 25 July 2007
10. Brin, S., & Page, L. (1998). The Anatomy of a Large Scale Hyper Textual Web Search Engine. WWW7 / Computer Networks 30(1-7): 107-117.
11. Austin, D. (2007). How Google Finds Your Needle in the Web's Haystack. Feature Column: Monthly Essays on Mathematical Topics. Retrieved from http://www.ams.org/featurecolumn /archive/pagerank.html on 21 Oct. 2007.