Click-through Data Analysis in Web Search Engine

Prof.B.D.Shendkar, Akshay Shirbhate,Vikas Yadav, Radhika Sadistap, Vaishali Sutar

Department of Computer Engineering ,SITS, Narhe, Maharashtra, India

ABSTRACT

While using a search engine, different users might look for different search results when they submit it to a search engine .If the query is ambiguous or covering a broad topic, it is difficult to guess the users interest. So we are proposing an approach which will analyse the search log and provide us users intent . Generally the generic search engines fail to properly outline the user's interest. So these search engines are not adaptable to individual user. But in our approach user interest is used in the clustering process to achieve personalization effect. The goal of personalized IR (information retrieval) is to return search results that better match the user intent. So to implement this we work in two phases. First, to make use of feedback session to discover users goal. Second, we find the similarities between the first clicked URL and the other URLs available about the query and display the most similar URLs. By doing this we can achieve specialized searching .

Keywords-*Information Retrieval, Restructuring, Implicit Feedback Mechanism, Ranking, Click-through Data.*

1. INTRODUCTION

Click-through data:-
The data collected from the users clicks and corresponding URLs is called as Click-through data.

In this paper we have proposed the new the study is all about user profiling based approach for on-going behaviour while surfing the internet for user friendly environment user profiling based approach for on-going user behaviour. In our proposed work we have taken user click through data to obtain the users interest in specific domain. User click through data is then used to re-rank the results obtained.
With the development of World Wide Web, web search engines have contributed a lot in searching information from the web. They help in finding information on the web quick and easy. Search engines do not consider specific needs of user and serve each user equally. It is difficult to let the search engine know what we the user actually wants.

Generic search engines are following the "one size fits all" model which is not adaptable to individual users. This might not be appropriate for users which require information. While searching for the information from the web, users need information based on their interest. For the same keyword two users might require different piece of information. This fact can be explained as follows: a biologist and a programmer may need information on "virus" but their fields are is entirely different. Biologist is searching "virus" that is a microorganism and programmer is searching for the malicious software. For this type of query, a number of documents on distinct topics are returned by generic search engines. Hence it becomes difficult for the user to get is considered as a promising solution to handle these problems, since different search results can be provided depending upon the choice and information needs of users.

2. LITERATURE SURVEY

Now a day's, multiple Web Usage Mining algorithms have been proposed to mining user navigation behaviour. In the following we examine some of the most significant navigation pattern mining systems and algorithm in web usage mining area that can be compared with our system.
In a paper published in May 8 ,2007, by Liyun Ru from China, it is stated that traditional evaluation methods rely on much human efforts and are therefore quite time-consuming. With click- through data analysis, we proposed an automatic search engine performance evaluation method.
This method generates navigational type query topics and answers automatically based on search users' querying and clicking behaviour. Experimental results based on this engine's user logs show that the automatically method gets a similar evaluation result with traditional assessor-based ones.
While our paper uses a user profile so no need of separate data analyser. The user click-through makes a path what user wants and the related suggestion will be shown according to their ranks.

According to one more paper which was published on July 23,2007 by author Rossie jones ,Ben Carterette

The web is a highly dynamic environment: documents disappear or become outdated, new documents appear, the query distribution changes. An important, but often overlooked, part of search engine design is evaluation. In order to know whether one ranking function is better than another, we need to evaluate them over a common set of queries .Each time a user enters a query and clicks on a result, he/she is making, in some sense, a "relevance judgment". Of course, these are very noisy: users get distracted, fail to specify their query well enough, change their interests, and so on. But looked at in aggregate, they may provide valuable information about the relevance of each document.

Thorston Joachims published a paper on Oct 31st 2011. This paper presents an approach to automatically optimizing the retrieval quality of search engines using click-through data.
Also it helps learning retrieval functions by analysing which links the users click on in the presented ranking.

And paper presented an approach to mining log files of WWW search engines with the goal of improving their retrieval performance automatically. Intuitively, a good information retrieval system should present relevant documents high in the ranking, with less relevant documents following below. Taking a Support Vector approach, the resulting training problem is tractable even for large numbers of queries and large numbers of features. And presents the information related to the query according to their ranking higher the rank, the priority will be higher following the lower ranks .

Where as in our paper we are going to suggest relevant URLs list related to the query will be shown on the screen ,having a user profile for business analytics for user friendly to show some related suggestion studying his/her history in the user profile and this will help the many organisation like online shopping.

3. ARCHITECTURE

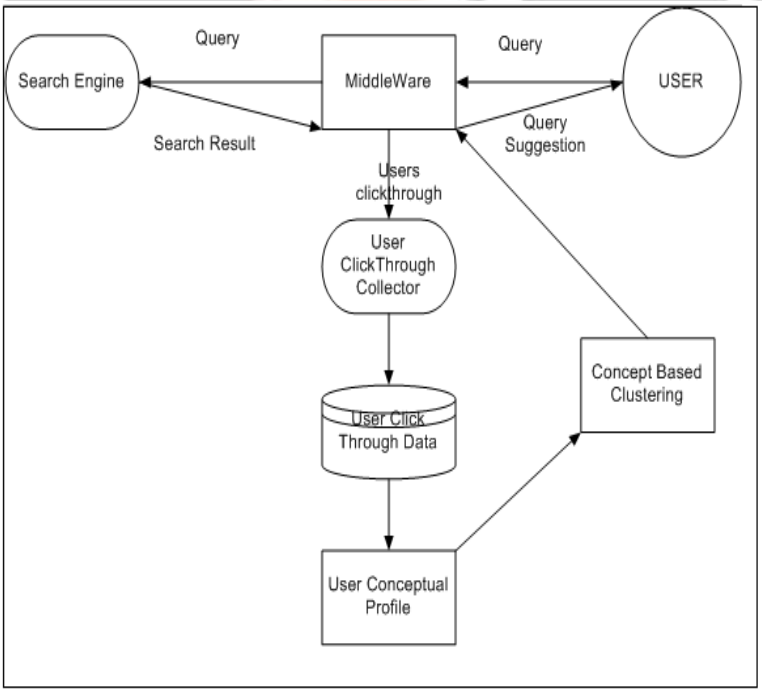

Fig 2.1. Architecture of proposed system

The architecture is mainly the assembly of above shown components. The user and the search engine are the components between which the query is processed to get the desired search results. The middleware represents the assembly of following three components- Click- through collector, click through repository ,user conceptual profile. Based on the architecture implementation is described in following part.

4. IMPLEMENTATION

The above figure[2.1] gives us an idea about the work flow of the system. We can see that there are two ends which are search engine and the user. These two ends are bridged by the middleware .The middleware is something which carries out all the core processing in the system. Firstly the user submits the query to the search engine. But actually the query is handed over to the middleware. This query is passed through the click-through data collector. This collector collects the queries and the users corresponding clicks in that particular session. The after a proper arranging the data ,it is stored in the database and called as click through data. This click through data is responsible for deducing a user conceptual profile based on the information provided by the click-through collector. There may be different algorithms and methods to find this users conceptual profile. These methods might be the implicit feedback by the user or study of the click-through data or any other methods. After getting the conceptual profile we make use of clustering based approach and it helps finding the most likely URLs which will be relevant to the user. From the results we got from the clustering process will be used to suggest the query or URL to the user. This whole process will be repeated every time user uses the search engine. And every time the feedback of users clicks will be updated in the click through data database. Note that , in the early sessions ,the suggestions might not be accurate. But we can fix this by suggesting random URLs related to the query..(but only for the some initial sessions).

5. EXPECTED RESULTS

In this paper following are the expected outcomes from our system:-

1) As we are making use of the Google API to retrieve the data as per users request, we expect the API to produce proper output. This will be ensured by the proper linking of the API to our search engine.
2) We expect the GUI(Graphical user interface) to be user friendly.
3) The search results should be available in as less as possible time.
4) The suggestion of the URLs should be such that user will be happy to use it.
5) Every personal information should be secure.
6) The search engine should not affect the other functionalities which are used by the user in any browser or in any API.

6. CONCLUSION

In this paper we are creating a personalised search engine which will effectively suggest relevant URLs and maintain and monitor the users profile in a secure manner. The methodology is found to be solvable (NP problem). And this paper successfully demonstrate and achieve the desired goal .

7. REFRENCES

[1] Leuski A. and Allan J. Improving Interactive Retrieval by Combining Ranked List and Clustering. Proceedings of RIAO, College de France, pp. 665681, 2000.

[2] T.Joachims, "Evaluating Retrieval Performance Using Click-through Data," Text Mining, J. Fran G Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.

[3] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feed-back," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

[3]Kenneth Wai-Ting Leung, Wilfred Ng, and Dik Lun Lee ,"Personalized Concept-Based Clustering of Search Engine Queries", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 11, NOVEMBER 2008.

[4]S .Preetha, K.N. Vimal Shankar "**PERSONALIZED SEARCH ENGINES ON MINING USER PREFERENCES USING CLICKTHROUGH DATA**",ICICES2014 - S.A.Engineering College, Chennai, Tamil Nadu, India.

[5]Kenneth Wai-Ting Leung, Dik Lun Lee, ***"Deriving concept-based user profiles for search engine logs",*** IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 01, JULY 2010

[6] T. Saracevic, Evaluation of evaluation in information retrieval, Proceedings of the 18th ACM SIGIR, 1995

[7]U. Lee, Z. Liu and J. Cho, Automatic Identification of User Goals in Web Search, in the 14th WWW Conference, 2005.