www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 4 April 2015, Page No. 11518-11523

Intelligent Query Search

Prof.M.R.Kharde, Priyanka Ghuge ,Kurhe Prajakta,Cholake Shital,Maghade Rachana
Department Of Computer Engineering,PREC,LONI
Pdghuge15@gmail.com
manojkharde@gmail.com

***Abstract*— Now a day's use of internet is increasing rapidly. For broad topic each new user may have his different user goals. Hence the inference and analysis of the user search goals can improve the efficiency of the search engine and also reduce the time needed to search the query as unwanted data can get hide from the user and user get only his goal oriented search results. Currently everyone is searching on the internet and internet provides you ambiguous result of same things as it contains lot of information. In proposed method system will provide the information related to the user goals. In this paper we have discover a novel framework to discover the user goals by clustering the user search goals and then new approach to generate the pseudo document to represent the clustering effectively. At the end we have proposed novel approach CAP to calculate the performance of the search engine.**

Keywords- Search Engine, Hidden Web Crawler, Query Optimization, Search engines, Metadata, document frequency, term weights

I. INTRODUCTION

Internet is the most easiest and rapid source of information that can be. The search engines crawl the entire databases and provide all the information relevant to the query entered. But the availability of many ambiguous objects or information available associated with the same name or category creates lot of confusion for the internet users. In the search engine query are submitted to the search engine and search engine retrieves the information needed to the user. The major problem with the search engines is that it is least concerned with the user specific interests and therefore gathers all the information from the internet and presents it to the user.Its the user who has to face the problems in categorizing the obtained results. For an example consider the query "the Kite" , the search engine will provide the data regarding "the kite that we fly in the sky" and "the kite bird" and "the kite film". So it becomes necessary for the user to develop a technique for categorizing such ambiguous results. We treat user query as a source to reach the desired information. In many websites the search engine are widely used for finding the user need. As it's the digital world and internet is on fingertips of the users i.e. through mobile phones or Tabs the size of the query goes on reducing as the exposure to enter the longer queries are not provided. i.e. normally two or three words. And ultimately such queries give an ambiguous results. Results do not exactly match to the user's intensions. Many times different search engine produces different search result. So that non useful results arises and those are fail to satisfy the user's expectations. So consequently we reach to a conclusion that we need to design a technique that would be proven to be beneficial to the user at its searching side. Thus we define user's need with the name of cluster. It will ultimately result in improving the performance of search engine. We can able to redesign the result by grouping the needs of the user at different time. The user need can assigned by a word on which the clustering will be done. Depending upon the clustering the result are ranked [3] [5]. For better searching, many methods were invented to make searching more effective like classification of query, recognition of search results, and session limit detection. However, this method has limitations since the number of different clicked URLs of a query may be small [6]. Other works analyse the search results returned by the search engine when a query is submitted. Different intension of different user to search is depicted in the following figure.

Figure 1. Goal Text. Different user have different goals in their mind.

This results or intentions have no correlation. Here this are named as goal text which reflects the user information. Therefore, there is no particular syntax or pattern in which user can specify his intentions to the search engines, and its very well known that query formulation is a bottleneck issue in the usability of search engines. Most text classification research focuses on classifying documents, which contain enough terms to adequately train machine learning approaches. The task of classifying web queries is different in that web queries are short, providing very few inherent features [7]. Therefore, most approaches use the documents retrieved by a query as features to classify it. For example, the user has entered a query 'phoenix' in Google search engine. Basically it should produce the results for phoenix as a bird. But it is displaying the result of a shopping mall in pune. The expected result is found to user but it is not ranked as a first result. Many times user have to search for many pages of search results to find his need. Every time user wanted to submit query 'phoenix' it will firstly shows the result of mall instead of bird.

Figure.2 Different Result for Query

II. LITERATURE SURVEY

Query classification before retrieval is applied in [13]. Before gathering the documents information query classification is performed. It is nothing but the preretrival of the query. Author proposes three different mechanisms to classify the obtained results. [14] Aglomotive clustering of the query does not apply to the search history of the user with same query. For this author used a clickthrough data to along with the clicked sequence and clicked URLs. The main aim of proposed method is to find the users clicked data and divide this data into clusters. The second paper is content aware query suggestion by mining click through and session data. The main motto behind the query suggestion is to improve the performance of the search engine and increase the efficiency of the search engine, Although there are some query suggestion methods are there but no one of them is depends on the context of the query. In the previous methods this methods only check the context is belongs to the cluster or not. If contents are same then this method gives the output results in the search engine. In the zealous algorithm it creates the histogram [12] of the search results and the result having values below to the threshold are discarded and threshold upper than the threshold are considered in the search goals. This basically eliminates the UPLs with not having high threshold value. The user goals but this method not provide the accurate result. Basically query are submitted to the search engine and depends upon the history of search results information will provide to the user.

A query may combination of keyword or it may be some phrase or well-formed natural language [12]. Once a user query is input to the search engine the list of documents is presented to the user with a document title. Then it generate a histogram on the basis of threshold values.

A.Privacy preserving algorithm

Search engines are Mainly included the search keywords and files or phrases and the resulted URLs. In this paper author focuses in the collection of the query using the search users search history log. They display the frequent items in ZEALOUS [12]. Search log S, positive numbers m is the input. Zealous algorithm is used to preserve the privacy of the user. This paper

contain the privacy preserving in the clicked log, query and goals of the users. In the zealous algorithm in comprises of two phase in the first phase zealous calculate the histogram and in the second phase remove the items from histogram having range below the threshold.

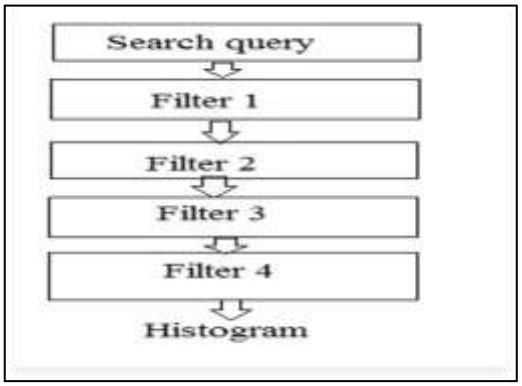

Figure: 3 Flow of Zealous algorithm

Disadvantages of this method is, that it never contains the users feedback because of this it may contain the more noisy data. The fourth paper is depends on the pre and post method. The pre method comprises of the users feedback and in the seconds i.e. post method the users URLs are restructured. This method first compare the documents and then combines this documents. The database containing the search keyword or phrase that have similar contents and the explicit query are classified according to the trained datasets. This contains the effective training data and search is done according to that training data. For the post retrieval of the data we have used the vector support machine. In the all existing methods user not getting his wanted search results because each method have some of the limitations. So we have proposed novel method to find the user goals and cluster the URLs according to the user goals. The proposed method is described in the following sections.

III. PROPOSED SYSTEM

We propose a very interesting and efficiently workable mechanism that aids in improving the search results obtained from the search engines. For an example, unless you meet a person for first time u cannot say that you have met him. i.e. if you don't have any feedback about any object may it be living thing or non living thing, you cannot pass your opinion about the same. Similarly, we know that search engines pre-requisites are that it works only after you trigger.So to trigger you search result optimization you need to present a method, that method used here is "Feedback Startegy Method" or "click Through log" method.

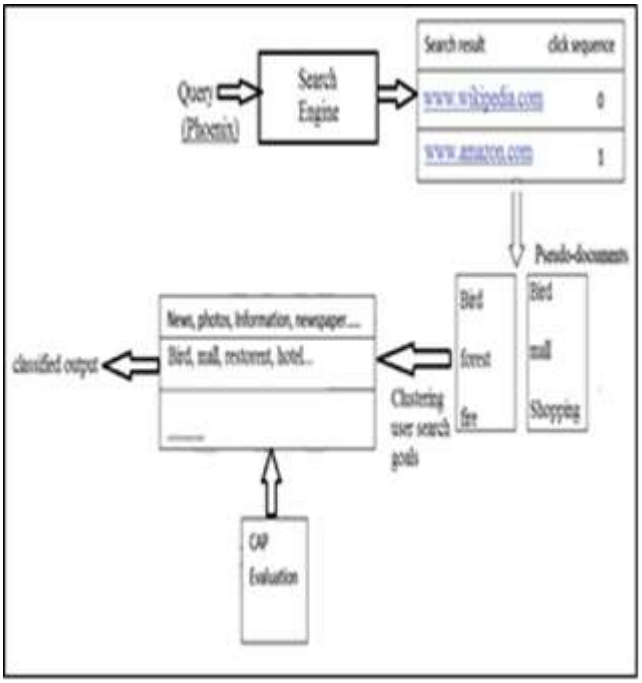

Figure 4. System Architecture

Our aim in designing this system is to enhance the search results according to user interests and reduce the overhead of surfing for further results from the noisy or unwanted data from the searched results. The proposed system initiates with user entering his.\/her short and ambiguous query to the browser. Browser then passes the query to the search engine to get the relevant information available over the internet and display it in the organised manner to the user. The user is now supposed to trigger the procedure of restructuring the obtained results by providing user clicks for the interested information. This user clicks are maintained in the logs and are valid only for a particular session. Once the user log has been created, the TF IDF values are computed and then the clustering procedure follows to obtain the restructured results. The restructured result are organised according to the user feedback from various clicks provided at the beginning of the session. Every user search the same query with different intensions.

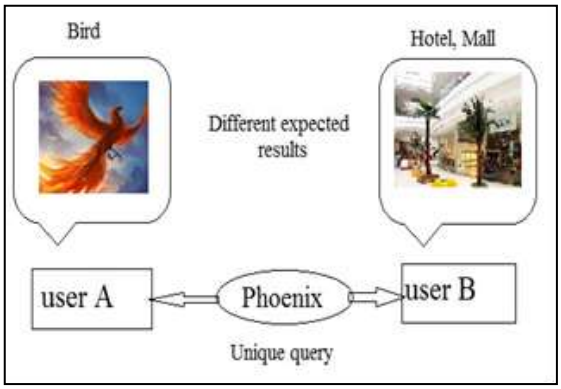

Figure 5. Different Requirements of User

For example if user A and B both typed same query in a search engine. Suppose their query is 'jaguar'. The user A wants the data regarding "jaguar animal" and user B wants the data regarding "Jaguar Vehicle". Then according to their click through logs and their searching. behaviour the clustering is done. Clustering is the process that will produce different outputs to both the users according to their clicks or feedback. Depending upon this feedback provided by the user the pseudo documents are created. After that, Once the pseudo documents are created, with the help of these documents clustering of the user search result is done. Then applying Cap evaluation technique the restructured output is displayed. This classified output is nothing but the expected result which user wants to search.

A .Feedback session:

The first module of the proposed system is the feedback session module where all the results displayed by the search engine is displayed without any client side processing. This feedback is nothing but the clicked ststus of the URLs displayed to the user. If the URL is clicked, correspondingly the database entry for that url is '1' and the unclicked URLs have the corresponding entry to be '0'. The user feedback i.e. the user clicks implicate the user interests and the unclicked URLs implicate the non interested information. The unclicked urls even tough are considered as non interested URLs as per user perspective, but there might be the case that the user might have missed some URLs relevant to the user interest and so for the further processings the unclicked URLs are also stored with their status being '0' in the database. The feedback session is least bothered about the sequence of the URLs clicked for clustering, but the sequence of the Clicked URLs matter a lot for the CAP Evaluation purpose. And so the clicked sequence is also stored during the every session.

B. Pseudo Documents:

The clicked urls and the unclicked urls are both processed by the TF IDF computing algorithm so as to get the frequent terms and frequent documents from the un structured result. The exact expansion of the term pseudo document can be defined as the conceptual category of the class that is created according to number of terms and documents found relevant to he user information interest that was triggered by the user;s feedback These collection of pseudo documents is then given as input to the Clustering module so as to cluster the results into well defined manner. So for improvising and evaluating the search restructured results, we define binary vector to store the polarity of the URLs i.e. clicked = 1 and unclicked = 0.

After the calculation of document frequency and URL weight the exact match of user's expected result is evaluated.

C. K_MEANS ALGORITHM:

The results that are to be clustered are categorised according to the pseudo documents created in the previous module. And the most important part of the entire process is carried out in the clustering phase. The K- Means clustering algorithm is carried out in the

Search results	Click sequence
WWW.phoenix.edu/	0
WWW.phoenix.com/	1
Santara.deviantart.com	2
https://www.gov.in/	0
www.wearephonix.com	0
www.nasa.gov/mission_pages/phoenix/main/	0
www.phoenix.org.uk/	3
https://www.exeterphoenix.org.uk/	0
phoenix.craigslist.org/	0

Figure 6. Clicked Sequence

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Empirically select number of cluster to be created as 'k'

2) Compute the distance between each Pseudo Document value and the cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is the least of all the cluster centers..

4) Recompute the new cluster center using mean formula

5) Recompute the distance between each pseudo document and new obtained cluster centres.

6) If no data point was reassigned then stop, else repeat from step 3).

After following the above steps we will get the clustered output of the web URLs.

D. *CAP(CLASSIFIED AVERAGE PRECISION):*

Once the system is worked on , the Evaluation of the obtained restructured results and the efficiency is calculated using CAP. This novel method is useful to determine the best cluster amongst the number of clusters. This aids in maintainingss the metric of user search results. This will helps to determine user search goals are inferred properly or not. Depend on the criteria used in the CAP we also find out the best cluster. In the cap we are getting information from the user clicked, clicked means relevant and unclicked means irrelevant. This will help us to determine is user getting his goal oriented result or not.

E. Evaluation of re-designed web search results:

Finally we have invented new method to evaluate the search results. The restructuring of the search result is done till the user not getting his goal. This method helps user to reach to the final goal and get the noise free and appropriate data. This will also improve the efficiency of the search engine. This is the final stage of the proposed method. The proposed method is normally designed for the session only. Since the user search goal is not fixed, the evaluation of redesigned search result becomes more difficult. There is no approach invented yet to evaluate search queries. Therefore, we propose an evaluation method based on restructering web search results to evaluate whether user search goals are guessed properly or not. User search goals are represented by the vectors and the feature representation of each URL in the search results can be computed. Then, we are going to categorize each URL into a cluster cantered by the inferred search goals. In this we are doing categorization by selecting the smallest value between the URL vector and user-search-goal vectors. Categorization is done according to the user search goal vector and URLs. The main aim behind the restructuring the web result is to provide more accurate search result to the user and remove unwanted data till contain in the search results.

VI. EXPEREMENTAL ANALYSIS

During the experimental analysis of the system, we will pass the same query multiple times with the varying click feedback and accordingly compute the efficiency and accuracy of the obtained restructured results. We will compare the evaluation of the original results and the Obtained restructured results.

VII. CONCLUSION

We can conclude that the proposed system and the proposed mechanism for obtaining the Restructured results from the original results from the Clicked URLs as the feedback gives an efficient and highly accurate results as compared to state – of – art techniques. Both the clicked and the non clicked URLS and snippets are used to deduce the user interests so as to gain maximum accuracy as it may happen that user misses out the interested urls in the feedback process Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods. The complexity of proposed method is very low and we can use this method in reality easily. Thus by using the proposed method user can find what he want conveniently.

ACKNOWLEDGMENT

Authors would like to take this opportunity to express our profound gratitude and deep regard to our guide prof. M. R. Kharde for his valuable guidance, motivating feedback and constant encouragement throughout the duration of the project. His valuable suggestions were of acute importance throughout our project work. His firm guidelines kept me working harder to make this project in a much proper way. Working under his guidance was an extremely knowledgeable experience for me.

REFERENCES

[1] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in search Engines," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '00), pp. 407-416, 2000.

[3] S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, "Varying Approaches to Topical Web Query lassification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '08), pp. 875-883, 2008.

[5] H. Chen and S. Dumais, "Bringing Order to the Web: Automatically Categorizing Search Results," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[6] C.-K Huang, L.-F Chien, and Y.-J Oyang, "Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs," J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.

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

[8] T. Joachims, "Optimizing Search Engines Using Clickthrough Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

[9] D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

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

[11] H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.

[12] Naynaneni Lavanya and E. Sandhyarani ," A Comparative Study on Privacy by Search Engines while Publishing Search Logs" International Journal of Advanced Research in Computer Science and Software Engineering, doi. 8 , 2012.

[13] Steven M. Beitzel, "Varying Approaches to Topical Web Query Classification" SIGIR 2007.

[14] Doug Beeferman and Adam Berger, "Agglomerative clustering of a search engine query log", 2000.