A Modified Approach For Inferring User Search Goal Using Feedback Session

Harshal Kekan
Department Of Information Technology
RMD. Sinhagad School of Engineering
Pune,India
E-mail: harshalkekan128@gmail.com

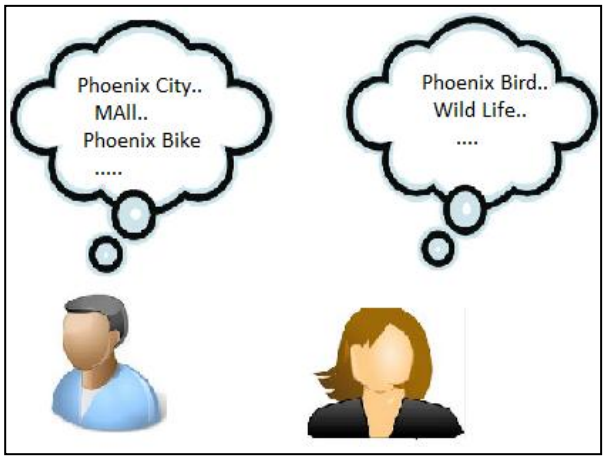
Prof. Dhara T. Kurian
Department Of Information Technology
RMD. Sinhagad School of Engineering
Pune,India
E-mail: dtkurian@sinhgad.edu

***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- Psudo-document,Classified average precision , Feedback session, Ambiguous query

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 moment 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 movie or album". So it becomes essential for the user to develop a technique for categorizing such ambiguous results. We treat user query as a source to reach the desired information. 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 matches to the user's intensions. Different intension of different user to search is depicted in the following figure.



Figur 1. Goal text. different user have different goals in their mind.

This results or intentions have no correlation. Here these keywords are named as goal text which reflects the user information.

Figure.2 Different Result for Query

II. LITERATURE SURVEY

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 [1][4] 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 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 [4] [5]. 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. Query classification before retrieval is applied in . Before gathering the documents information query classification is performed. It is nothing but the pre-retrieval of the query[6]. Author proposes three different mechanisms to classify the obtained results . Search engine content the data of the intension, and history of the user and his query. Here main focuse in the collection of the query using the search users search history log. Zealous algorithm is used to preserve the privacy of the user. This concept contain the privacy preserving in the clicked log, query and goals of the users[1] [4] [7]. 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[9].

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[2]. 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 Strategy Method" or "click Through log" method[9]. 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[9]. 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.

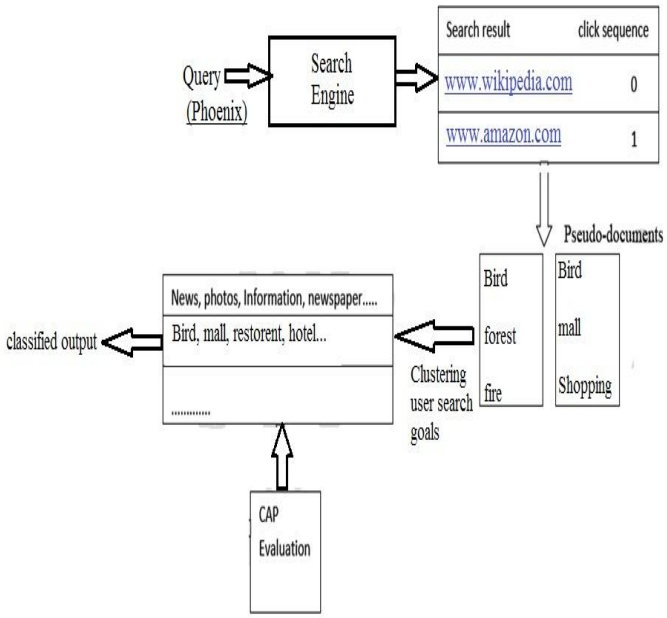


Figure 3. System Architecture

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.

Suppose their query is 'jaguar'. The user A wants the data regarding "jaguar animal" and user B wants the data regarding "Jaguar Vehicle". Clustering is basically a method that will produce different outputs to both the users according to their clicks or feedback. Psudo-documents are created depending upon this feedback provided by the user[9]. After Once the pseudo documents are created, with the help of these documents clustering of the user search result is done. Then at the last applying Cap evaluation technique the restructured output is displayed.

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. Basically in feedback session clicked URL's tell what user expecting and unclicked show what user do not care about feedback session should include the last URL last URL clicked in the session[3]. This feedback is nothing but the clicked status 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 represent the user interests and the unclicked URLs represent 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 processing the unclicked URLs are also stored with their status being '0' in the database. The feedback session is least concerned 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. We can also directly represent feedback session using binary vector representation but it will not give more information to the user. So new way to represent the feedback session is Psudo-document creation[9].

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[9].

There are two ways to build the Pseudo-document 1) representing the URLs in the feedback session: In this method we have to extract the enriched URLs with additional content by extracting title and snippet. Then have to calculate TF-IDF value for these snippet and title.2) Forming pseudo-document based upon URL representations: Here F_{fs} (feature representation vector) indicates the importance of the term in feedback session. F_{fs} is the pseudo-document they are looking for. Basically it will represent the user desire and what user don't want. It will represent the goal text in the user mind[10]

C. K_Means Algorithm:

The results that are to be clustered are categorised according to the pseudo documents created in the previous module1. And the most important part of the entire process is carried out in the clustering phase. The K-Means clustering algorithm is processed in following way:

Let Y = {y1,y2,y3,……..,yn} be the set of data points and V = {v1,v2,……..,vc} be the set of centers

.

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

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

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

4) Recalculate the new cluster center using mean formula

5) Here recalculate the distance between each data point and new obtained cluster centres.

6) After all processes if no data point was reassigned then have to stop, otherwise 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 method is useful to determine the best cluster amongst the number of clusters. This aids in maintain in the metric of user search results. This will helps to determine user search goals are inferred properly or not. By using CAP we can choose the best cluster among all the present clusters . In the cap we are getting information from the user clicked, clicked means relevant and unclicked means irrelevant. This will help to evaluate the restructured result is providing the goal oriented result or not[9].

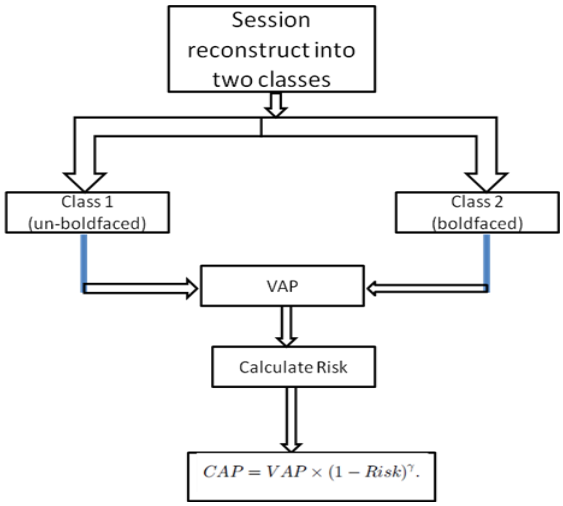


Fig.4 Working of CAP

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. 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.

F. Image Classification:

Images plays an important role while searching a query over internet. Many search results are bounded with image

results. Image makes searching easier and it improves speed of searching. Many search engines retrieves image result along with query results. But these images are not displaying in a ranking manner. In this system we are going to achieve ranking of images on the basis of relevancy. Here we are going to tag image with its name. Then according to user interest we will create clusters. And then according to user request images will be reranked.

IV.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.

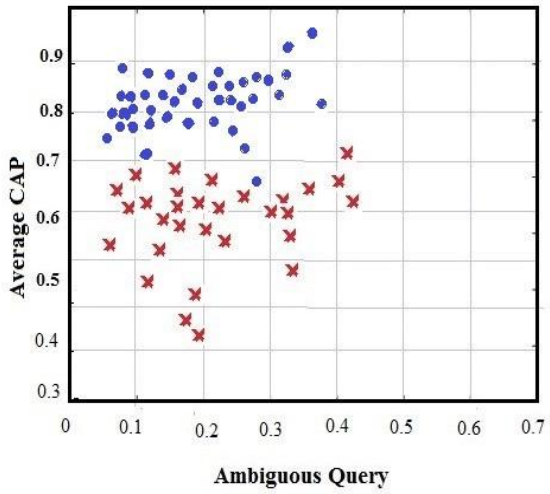


Fig.5 Analysis of user query

V. 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 The complexity of proposed method is very less as compared to other methods and we can use this method in reality easily. Thus by using the proposed method user can find what he want conveniently.

ACKNOWLEDGMENT

I take this opportunity to thank Prof. Dhara T. Kurian, Head of the information technology, for her encouragement and guidance. I also want to thank my guide Prof. Dhara T. Kurian for her continuous help and generous assistance. She helped me in a broad range of issues from giving me direction, helping to find solutions to problems, outlining requirements and always having the time to see me. I also extend sincere thanks to all the staff members for their valuable assistance. Last but not least I am very thankful to my class-mates for all their valuable suggestions and support.

.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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:.

[8] 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.

[9] Harshal Kekan, Prof. Dhara T. Kurian "A Survey On User Search Goal Inferring System" International Journal of Computer Engineering and Applications, ISSN 2321-3469 Volume VII, Issue III, Part I, September 14

[10] Zheng Lu, Hongyuan Zha, Xiaokang Yang, ,Weiyao Lin and Zhaohui Zheng "A New Algorithm for Inferring User Search Goals with Feedback Sessions" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013