ISSN:0975-9646

Categorising Web Search Results from Search Engine Logs

Mr.K.Morarjee [#], Ms.Bijam.Mownika [*]

[#] *Department of Computer Science and Engineering*
CMR Institute of Technology
Kandlakoya,Hyderabad,India

[*] *Department of Computer Science and Engineering*
CMR Institute of Technology
Kandlakoya,Hyderabad,India

***Abstract*— With exponential expansion of the Internet, it has turned out to be increasingly difficult to discover information. In the recent times, research on inferring user goals in support of text search has received great concentration. Inferring user search goals is extremely significant in getting better search engine significance as well as user experience. A search engine has to consider not only significance of every individual document, however in addition, how pertinent the document is in light of other recovered documents. The inferences as well as examination of user search goals can contain a lot of advantages in getting better search engine relevance as well as user experience. We put forward a new approach to infer user search goals in support of a query by clustering projected feedback sessions. A new approach has been projected to infer user search goals in support of a query by means of clustering its feedbacks sessions symbolized by pseudo-documents. All feedback sessions of a query are initially extracted from logs of user click-through and mapped towards pseudo-documents**.

***Keywords*— Feedback Sessions, User Search Goals, Pseudo-documents, Voted Average Precision, Risk, Classified Average Precision.**

I. INTRODUCTION

In this contemporary world popular search engines receives millions of queries and collects large amounts of user behavior data. This huge amount of data gets accumulated in the search engines as search log data on the web server and as browse log data on the client machine. A Web search engine consists of three programs called Crawler,Indexer and query engine [1]. The Fig.1 shows the structure of web search engine.

The web page cache consists of all the web pages collected by the search engine crawler program from different web servers on the internet. Then indexes are created for all those web pages of the cache by the search engine Indexer program. The query engine takes the user query, consults with the Indexed data pool and the web page cache and a log of previous queries and produces a result page which the user is searching.

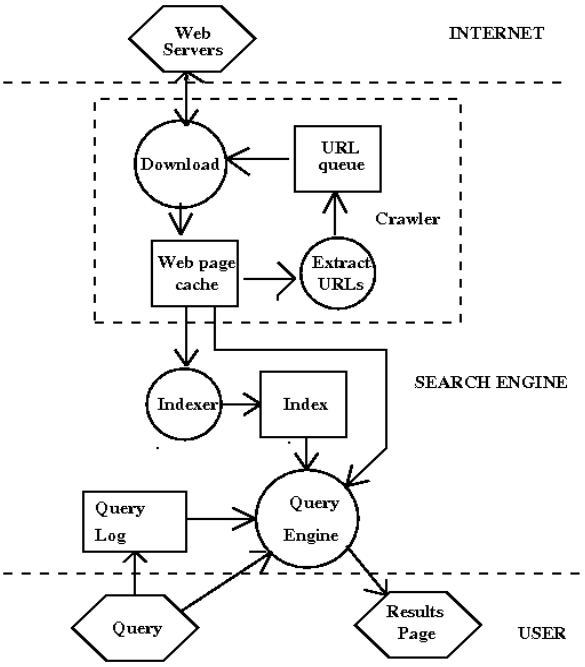

Fig.1 Structure of web search engine

II. METHODOLOGY

The below Fig.2 shows our proposed system

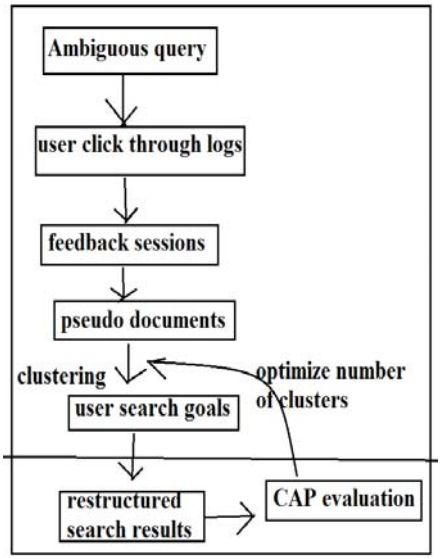

Fig.2 An Overview of proposed system

All feedback sessions of a query are initially extracted from logs of user click-through and mapped towards pseudo-documents [2] [3]. Goals of user search are inferred by means of clustering pseudo-documents and described by several keywords. As we do not recognize precise number of user search goals earlier, quite a few different values are attempted and most advantageous value is determined by feedback. The original search results are reorganized based on goals of user search. Performance of restructuring search results was evaluated by projected assessment criterion Classified Average Precision [4]. And the assessment result is used as the feedback to choose the most favourable number of user search goals.

In the recent years there were a growing number of vertical search services accessible by means of a general-purpose search engine employing an integrated user interface. Such a service will make available additional pertinent and necessary results in support of in-domain web queries, however will build nonsense towards queries that are immaterial to that province. With exponential expansion of the Internet, it has turned out to be increasingly difficult to discover information. The manual nature of directory compiling procedure makes it not possible to contain as broad coverage as search engines, or to be appropriate similar structure towards intranet or else local files devoid of extra manual effort.

Diversifying search results frequently involves an exodus from autonomous document relevance supposition underlying eminent likelihood ranking principle in information retrieval. It is uncertain whether users will discover a specified document appropriate to their information require once previous documents by now satisfying this necessitate was observed [5]. Therefore, a search engine has to consider not only significance of every individual document, however in addition how pertinent the document is in light of other recovered documents. The recovered documents have to make available utmost coverage as well as lowest redundancy regarding likely aspects underlying a query.

A rising body of exploration is analysing users' universal Web searching features, by means of smaller number studies examining queries by users looking for multimedia information. In the recent times, research on inferring user goals or intents in support of text search has received great concentration. Inferring user search goals is extremely significant in getting better search engine significance as well as user experience.

A new approach has been projected to infer user search goals in support of a query by means of clustering its feedbacks sessions symbolized by pseudo-documents. All feedback sessions of a query are initially extracted from logs of user click-through and mapped towards pseudo-documents. Goals of user search are inferred by means of clustering pseudo-documents and described by several keywords.

By mining user click-through logs, we can get hold of two kinds of information such as click content information as well as click session information. A session within user logs of click-through is a progression of queries along with a series of clicks by user toward addressing particular information need.

III. RESULTS

The click through log data is collected from a commercial engine having queries with single session. In our approach we used a variant of the K-means method. The strategy is to first apply a hierarchical agglomeration algorithm which determines the number of clusters and finds an initial clustering and then improve iterative relocation to improve the clustering. User search goals are represented by the Centre points of different clusters. Each query is represented by keywords which represents its feature vector.

The following Table I shows representation of user search goals by keywords. This Inference method will infer the user search goal properly for other queries and depict them with meaningful keywords.

TABLE I

Ambiguous Query	**Four Keywords to depict user search goals from the query**
Flower	Photos,blooms,flowers,cards
	Plants,time,min,flow
Pogo	Pogotv,home,place,kids
	Cool,games,play,art
Vehicle	Motor,road,results,thing
Apple	Ipod,laptop,computers,news
	Chroncom,aapl,stock,price
Photoshop	Photoshop,inspiraf, photoshopcom,home
	Online,editor,photo,download,adobe
Hindu	Analysis,com,indepth,coverage
	Business,nation,http,publications
Training	Knowledge,practice,astd,associate,id
	Training,Wikipedia,acquis,acquisy

We compared our proposed method with the other methods of clustering search results [6] [7] and clicked URL's directly [8]. We have compared CAP of three methods and proposed clustering of feedback sessions method is more relevant for user search goals. The first method [6] [7] clusters the top 100 search results to infer user search goals by K-means clustering and selects the optimal K based on CAP criterion and the mean average VAP is 0.65 and mean average risk is 0.252 and average CAP is 0.816.

In the second method of clustering different clicked URL's directly [8], the mean average VAP is 0.805, mean average risk of 0.312 and average CAP is 0.860. Our method has resulted in mean average VAP is 0.892, mean average Risk is 0.289 and mean average CAP of 0.901. The proposed method has highest mean average CAP significantly higher than the other two methods as shown in the Table II below.

TABLE II

Method	Mean average VAP	Mean average Risk	Mean average CAP
Our Method	0.892	0.289	0.901
Method I	0.65	0.252	0.816
Method II	0.805	0.312	0.860

IV. CONCLUSIONS

The Feedback sessions as Pseudo documents are resampling which reflect user information needs by excluding noisy once and also feedback sessions are a combination of clicked and unclicked URL's which reflect user information needs more precisely, so our proposed method is more efficient.

REFERENCES

[1] Agichtein, E., Brill, E., Dumais, S., and Ragno, R. 2006. Learning user interaction models for predicting web search result preferences. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06. ACM, New York, NY, USA, 3–10.

[2] 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.

[3] 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.

[4] Zheng Lu, HongyuanZha, Xiaokang Yang, Weiyao Lin, and ZhaohuiZheng, 2013 "A New Algorithm for Inferring User Search Goals with Feedback Sessions" Published by the IEEE Computer Society.

[5] 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.

[6] 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.

[7] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.

[8] X. Wang and C.-X Zhai, "Learn from Web Search Logs to Organize Search Results," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.