Feedback Session Based User Search Goals Prediction: A Survey

H. G. Chetan
4th sem M.Tech., Department of CS&E
Adichunchanagiri Institute of Technology
Chikmagalur, India
chetanhgcool@gmail.com

S. Sharath
4th sem M.Tech., Department of CS&E
Adichunchanagiri Institute of Technology
Chikmagalur, India
sharuyashassu@gmail.com

***Abstract*—Different users may have different search goals when they submit it to a search engine. Many activities on the web are driven by high-level goals of users, such as "plan a trip" or "buy some product". The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. This paper presents results from an exploratory study that focused on analyzing selected search sessions from a search engine log by clustering feedback sessions. URL is considered to prepare pseudo-document to represent feedback session for clustering. Based on clicking the URLs, scoring is done. New Criterion called Classified Average Precision (CAP) is proposed to evaluate the performance of the inferred user search goals. In this paper, we are interested in exploring the role and structure of user goals in web search.**

Keywords— feedback session; URL; clustering; CAP

I. INTRODUCTION

In web search applications, queries are submitted to search engines to represent the information needs of users. However, sometimes queries may not exactly represent users specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query.

A Web is a collection of inter-related files on one or more Web servers. Web mining is the application of data mining technique; it is used extract knowledge from Web data. Web data is Web content data (text, image, record), Web structure data (hyperlinks, logs) and Web usage data (http logs, app server logs).

A Web server usually registers a log entry, or Weblog entry, for every access of a Web page. It includes the URL requested, the IP address from which the request originated, and a timestamp. Based on the Weblog records, we have to construct the feedback session. Because Weblog data provide information about what kind of users will access what kind of Web pages. This session consists of URL's and click sequence and it focus on user search goals. Only using a feedback session we do not understand the user search goals exactly. Based on the feedback session, construct the pseudo document for analyzing the accurate result. This pseudo document consists of keywords of URL's in the feedback session. This is called as enriched URL's.

The enriched URL's are clustered and form a pseudo document. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have a high similarity in comparison to one another but are very dissimilar to object in other clusters. After constructing the pseudo document the Web search results are restructured based on the documents collection detail.

II. USE OF WEB MINING

Web mining is mining the Weblog records to discover user access patterns of Web pages. In developing techniques for Web usage mining, we may consider the following

- It is encouraging and exciting to imagine the various potential applications of Weblog file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge can be discovered from the large raw log data.
- The available URL, time, IP address, and Web page content information, a multidimensional view can be constructed on the Weblog database, and multidimensional OLAP analysis can be performed to find the top N users, top N accessed Web pages, most frequently accessed time periods, and so on which will help discover potential customers, users, markets, and others
- Data mining can be performed on Weblog records to find association patterns, sequential patterns, and trends of Web accessing.

III. PROBLEMS IN INFORMATION RETRIEVAL BASED ON USER SEARCH GOALS

- Irrelevant or ambiguous queries.
- What users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.
- Mismatching of vocabularies.
- Only identifies whether a pair of queries belongs to the same goal or mission and does not care what the goal is in detail.
- How we represent the information with selected keyword?
- How document and query representations are compared to calculate the weight?

IV. LITERATURE REVIEW

R. Jones and K.L. Klinkner proposed "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs" [8]. Their method is to detect search goal and mission boundaries for automatic segmenting query logs into hierarchical structure. User may issue number of queries to search engine in order to accomplish information need/tasks at a variety of granularities. Their method identifies whether a pair of queries belongs to the same goal or mission and does not consider search goal in detail.

Uichin Lee and Zhenyu Liu proposed "Automatic Identification of User Goals in Web Search". Their work is based on the Web query assigned by the user's analysis the goal [2], the goal identification is used to improve quality of search results. In existing system with use the manual query log investigation to identify the goals. This proposed system use automatic goal identification process. The human-subject study strongly indicates the automatic query goal identification. It can use two tasks like as past user click behavior and anchor link distribution for goal identification combining these two tasks can identify 90% goal accurately.

Preceding studies comprehends mainly interest on manual query-log investigation to recognize Web query goals. U. Lee et al. approached "Automatic Identification of User Goals in Web Search" [7]. They studied the "goal" at the back based on a user's Web query, so that this goal can be used to get better the excellence of a search engine's results. Their proposed method identifies the user goal automatically with no any explicit feedback from the user.

Zamir et al. proposed "Grouper: A dynamic clustering interface to web search results" [9] and used Suffix Tree Clustering (STC) to identify set of documents having common phrases and then create cluster based on these phrases or contents. They used documents snippets instead whole document for clustering web documents. However, generating meaningful labels for clusters is most challenging in document clustering. So, to overcome this difficulty, in [3], a supervised learning method is used to extract possible phrases from search result snippets or contents and these phrases are then used to cluster web search results.

T. Joachim's demonstrated "Optimizing Search Engines Using Click through Data" [4]. This approach is automatically optimizing the retrieval quality of search engine using click-through data stored in query logs and the log of links the users clicked on in presented ranking. By using support vector machine (SVM) approach, for learning ranking functions in information retrieval.

T. Joachim's et al. achieved "Accurately Interpreting Click through Data as Implicit Feedback" [5]. Their contribution is on examining the reliability of implicit feedback generated from click-through data in World Wide Web search. The author approached strategy to automatically generate training examples for learning retrieval functions from observed user behavior. The user study is intended to examine how users interrelate with the list of ranked results from the Google search engine and how their behavior can be interpreted as significance judgments. Implicit feedback can be used for evaluating quality of retrieval functions [6].

V. CLASSIFICATION, PREDICTION AND CLUSTERING

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology.

Figure 1 represents the flow chart of the proposed system. When the queries are submitted by the user, queries are taken has an input and fed to the search engine. Here two searching methods are used, if the feedback session allows the input queries it undergoes further procedure. If the input queries are nit fed to the feedback, then queries are fed to the normal search. After searching of both methods the search results will be displayed.

- Numeric prediction is the task of predicting continuous (or ordered) values for given input. For example, we may wish to predict the salary of college graduates with 10 years of work experience, or the potential sales of a new product given its price.
- The process of grouping a set of physical or abstract objects into classes of *similar* objects is called clustering. A cluster is a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Although classification is an effective means for distinguishing groups or classes of objects, it requires the often costly collection and labeling of a large set of training topples or patterns, which the classifier uses to model each group.
- Dissimilarities are assessed based on the attribute values describing the objects, often, distance measures are used. In this paper we use k-means clustering technique for constructing pseudo documents. K-means clustering is a centroid based technique. Classification and prediction are two forms of data analysis that be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large whereas classification predicts categorical labels, prediction models continuous-valued functions. It uses the preprocessing technique such as data cleaning, relevance analysis, data transformation and reduction. It provides the accuracy, scalability, robustness, speed and interpretability.

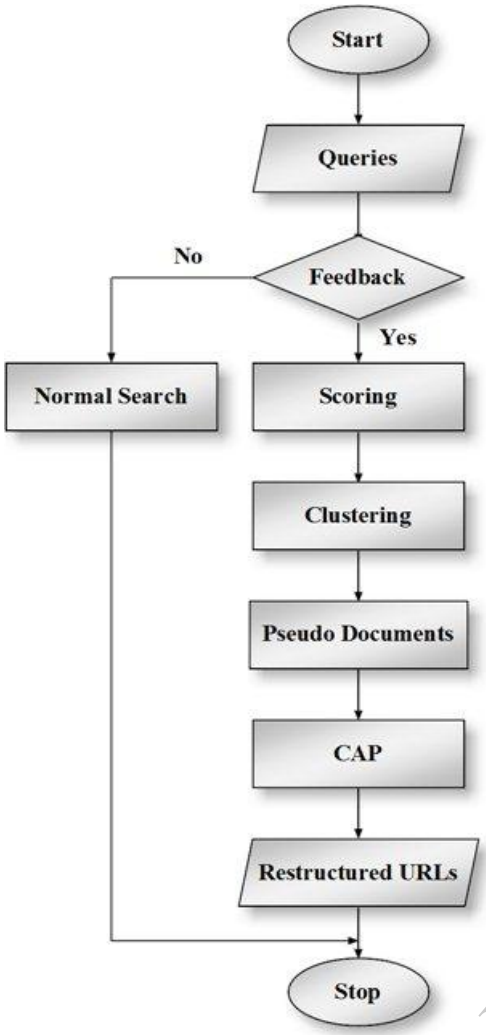

Fig. 1. Flow chart of the proposed system

As shown in Figure1, when the queries are submitted by the user, queries are taken has an input and fed to the search engine. Here two searching methods are used, if the feedback session allows the input queries it undergoes further procedure. If the input queries are nit fed to the feedback, then queries are fed to the normal search. After searching of both methods the search results will be displayed. The comparative study is listed in the Table 1. In this paper the context in QS consist of two steps like

1. In offline process the learning step is used to address the data, queries are converted into concepts by a technique called clustering, a click through bipartite. Based on session data a sequence suffix tree is constructed for the QS model.

2. In online process the query suggestion is used to capture the user search results by mapping with the query sequence submitted by the user. This approach provides to the user in a context-aware manner. It is also called as Context-Aware Concept-Based Approach (CACBA).

TABLE I. COMPARATIVE STUDY

SI. no	Title	Techniques used	Benefits	Drawbacks
1.	Learning Query Intent from Regularized Click Graphs	Semi-supervised click graph	Improve classification performance	Impact of seed queries and faceted query classification
2.	Feedback and Query Patterns to Organize Web Documents	Non supervised tasks	Improve the quality as 90%	A broader comparison with online directory
3.	Context-Aware QS by Mining Click-Through and Session Data	Offline model learning and online QS step, concept sequence suffix	Coverage and quality of suggestions	Larger coverage area
4.	Generating Query Substitutions	Query pair algorithm	Increase coverage and effectiveness	Machine translation techniques
5.	Varying Approaches to Topical Web Query Classification	Pre vs. post retrieval classification	QC is outperforms bridging a document taxonomy as 48%	Multiple approaches to improve the performance
6.	Learn from Web Search Logs to Organize Search Results	Commercial search engine log data and clustering	Better result organization and meaningful labels	Informative feedback information from user
7.	Automatic Identification of User Goals in Web Search	User click behavior and Anchor link distribution	Using goal identification task to achieve 90% of accurate results	Potentially-biased dataset

VI. CONCLUSION

As the Web and its usage continues to grow, so grows the opportunity to analyze Web data and extract all manner of useful knowledge from it. The past five years have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it. In this paper, we have made the literature review of user search goals using feedback session and pseudo document. First we construct a feedback session to analysis the user search goal from the Weblog record. It cannot provide the accurate result. This proposed system includes the pseudo document to provide the accurate results. Based on the pseudo document we have to restructure the Web search results.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from the Adichunchanagiri Institute of Technology, Chikmagalur, through its strategic initiative. The authors also acknowledge their Head of the Department, who gave guidelines for preparing this work. Last but not least the authors acknowledge support from their parents and from their friends.

REFERENCES

[1] Zheng Lu, Student Member, IEEE, Hongyuan Zha, Xiaokang Yang, Senior Member, IEEE, Weiyao Lin, Member, IEEE, and Zhaohui Zheng, "New algorithm for inferring user search goals with feedback session,", vol. 2, 1999, pp. 329-351

[2] Uichin Lee and Zhenyu Liu, "Automatic Identification of User Goals in Web Search," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), 2006, pp. 310-317

[3] 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.

[4] 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.

[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] 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.

[7] U. Lee, Z. Liu, and J. Cho, "Automatic Identification of User Goals in Web Search," Proc. 14th Int'l Conf. World Wide Web (WWW '05), pp. 391-400, 2005.

[8] R. Jones and K.L. Klinkner, "Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-708, 2008.

[9] O. Zamir and O. Etzioni. "Grouper: A dynamic clustering interface to web search results," Computer Networks, 31(11-16), pp.1361- 1374, 1999.

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

[11] Cao. H, 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.

[12] Jones. R, B. Rey, O. Madani, and W. Greiner, "Generating Query Substitutions," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 387-396, 2006.

[13] Li. X, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.

[14] Wang. X 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