An Effective Analysis of Search Goals by Mapping Pseudo Documents

K.Preethi[1], S.Shinija[2], J.A.Dhinesh Joseph[3]

[*1,*2]U.G Students, B.E CSE, Alpha College of Engg, Chennai, T.N, India.
[#3]Assistant Professor, Dept of CSE, Alpha College of Engg, Chennai, T.N, India.

[1]preethikathir29@gmail.com

Abstract – Different user have different search goals when they submit query to the search engine.These query are analyzed to get revelant information.This paper describes a novel approach to conclude the user search goals by analyzing the query logs.First is to discover different search goals by clustering the feedback sessions Second is to generate pseudo documents.Finally "Classified Average Precision" used to evaluate the performance to conclude user search goals.

Keywords – User search goals, feedback sessions, pseudo-documents, classified average precision

I. INTRODUCTION

IN web search ,queries are submitted to search engines to represent the information which is needed by users. Sometimes queries may not exactly represent what a user needs.For example, when the query "the sun" is submitted to a search engine, some users want to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun..Therefore, it is necessary to capture different user search goals for information retrieval.Results that are obtain from search engine after query submission should satisfies his/her needs.User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages.Some advantages are ,First restructuring web search results according to user search goals by grouping the search results with the same search goal; thus, users can easily find what they want. Second user search goals represented by some keywords.Finally the distributions of user search goals can also be useful in applications such as reranking web search results. Due to its advantages, many works about user search goals analysis have been investigated. They can be described as three classes: query classification, search result reorganization, and session boundary detection. In the first class, users attempt to infer their goals and intents by predefining some specific classes and performing query classification accordingly.In the second class, people try to restructure web search results. In the third class user aim to detect session boundaries, that is user identifies whether a pair of queries belongs to the same goal or not

II. RELATED WORKS

A. Automatic Identification of User Goals in Web Search

Based on the Web query assigned by the user's analysis the goal, 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. In 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.

B.Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents

Document representation model (DRM) is based on the implicit user feedback. Implicit user feedback is mean that the feedback from weblog. Document representation model is obtained from search engine queries. The main objective of this DRM is to achieve the better results using non-supervised tasks such as clustering and labeling obtained from search engine queries. Users are motivated for document representation. Based on the clicked queries the term provide the better choice of feature from the user's point of view. This model represent the frequency query patterns called as query set model.
The query set model reduces the 90% the number of features needed for represent the set of documents, then improve 90% the quality of results.

C.Learning Query Intent from Regularized Click Graphs

Improve the query intent classifier using a click graphs, this method is critical for vertical and general purpose search services offered by user interface. In existing they use query classification for improving feature representation of queries. In proposed we focus on completely orthogonal approach for enriching feature representation. The main objective is to increasing the amounts of training data using semi-supervised learning with click graphs. Based on the click graph we understand the unlabeled queries from those of labeled ones. Moreover we regularize the learning with click graphs using content based classification to avoid the error labels. We define the effectiveness of our algorithms using two different application (product intent and job intent classification). Using this both applications weexpands the training data and leading to improvements in classification performance. An additionally finding the large amount of training data based and classifiers using query words as features.

III. DESIGN AND IMPLEMENTATION

A. Framework of Our Approach

Framework consists of two parts upperpart and bottom part. In the upper part all feedback of the query are first extracted from user click-through logs. Thus feedback is not alone useful for finding user search goals so novel optimization method used to combine the enriched URLs in the feedback sssion to form pseudo-documents. After constructing the pseudo-documents mapping is done between the both feedback session and pseudo-documents. Then, user search goals are inferred by clustering the pseudo-documents and with some keywords. Since we do not know the exact number of user search goals in advance, so several different values are tried and optimal value is determined from bottom part. In the bottom part he original search results are restructured based on user search goals inferred from the upper part. Then, we evaluate the performance of restructuring web search results by "classified Average Precision". This evaluation results is used to select the optimal value of user search goals in the upper part.

B.Feedback Sessions

A session for web search is a series of successive queries to satisfy the information needs. This paper focusing on inferring the search goals for a unique query. Feedback session is based on single or whole sessions. The feedback sessions consists of clicke and unclicked URLs. Each feedback session tells about what user exactly needs and what the user does not cares about.

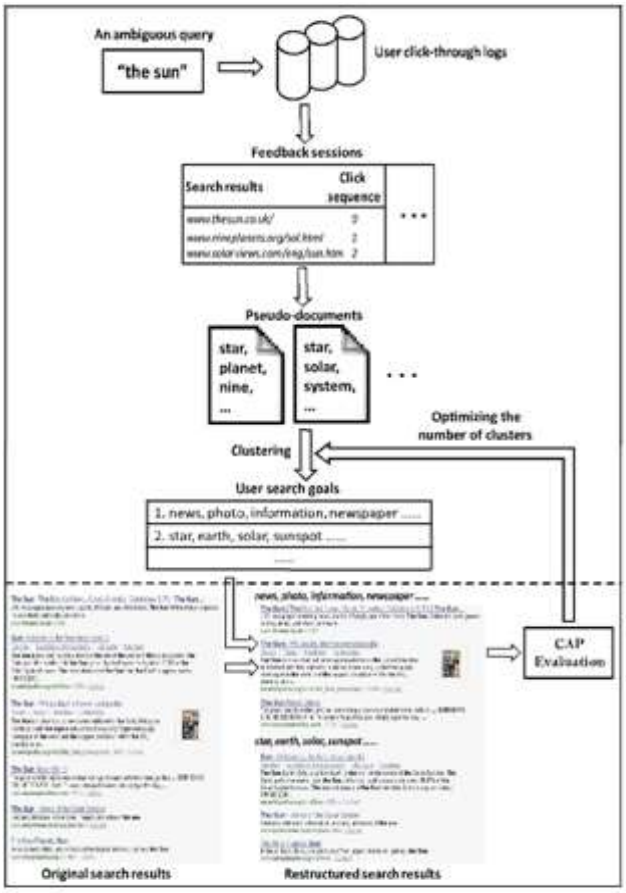

Fig .1. System Architecture

C.Mapping Feedback Sessions to Pseudo-Documents

It is unsuitable to directly use feedback sessions for inferring user search goals. So need to map the feedback sessions with pseudo-documents. For mapping "Binary Vector Method" is used, where "0" represents "unclicked" and "1" represents "clicked".

Search Results	Click Sequence	Binary Vector
www.thesun.co.uk/	0	0
www.nineplanet.org/sol.html	1	1
www.solarview.com/eng/sun.htm	2	1
En.wikipedia.org/wiki/sun	0	0
www.thesunmagazine.org/	0	0
www.space.com/sun/	0	0

Fig. 2. The binary vector representation

D.Building of pseudo-documents

1.Representing the URLs in the feedback session:

Enriching URLs with additional textual content by extracting the titles and snippets of th URLs appearing in the feedback sessions,then some textual processes are implemented such as transforming all letters to lowercase,stemming and removing stop words.It is represented by "Term Frequency-Inverse Documents Frequency"(TF-IDF).

$$\mathbf{T}_{ui}=[t_{w1},t_{w2,}\ldots\ldots\ldots\ldots t_{wn}]^{T}$$
$$\mathbf{S}_{ui}=[S_{w1,}S_{w2,}\ldots\ldots\ldots\ldots\ldots S_{wn}]^{T}$$

Where
T_{ui} and S_{ui} are TF-IDF vectors of URLs title and snippet.
ui means i^{th} URL in the feedback sessions.
t_{wj} and s_{wj} means TF-IDF value of the j^{th} term in the URLs title and snippet.

$$F_{ui}=_{wt}T_{ui}+_{ws}S_{ui}=[f_{w1},f_{w2},\ldots f_{wn}]^{T}$$

Where
F_{ui} means i^{th} URL in the feedback session
wt and ws are weights of the titles and snippets.

2.Forming pseudo-documents based on URL representation:

It is a optimization method to combine series of both clicked and unclicked URLs in the feedback sessions. People will skip some URLs because they are similar to previous ones. In this situation the unclicked URLs will wrongly reduce the weight of some terms. So that we analyze the problem in three cases. Case 1: one term appears in clicked URLs and it does not appear in unclicked URLs, so that people would skip unclicked URLs because it does not contain important term. The weight of the term in pseudo-documents is set to highest value. Case 2:one term appears in both clicked and subset of unclicked URLs, in this case some unclicked URLs are skipped because of the duplication hence the weight of term will be reduced to some extent, but not to zero. Case 3 one term appears in both clicked and almost all unclicked URLs, in this case also people will skip because of duplication so that weight is set to zero.

IV INFERRING USER SEARCH GOALS BY CLUSTERING

The user search goals are inferred by clustering the pseudo documents and represent by using keyword.The clustering is done by K-means clustering.The similarity of pseudo-documents is determined by cosine function.
$Sim_{ij}=\cos(F_{fsi}, F_{fsj})$And the distance between two feedback sessions

$$Dis_{i,j}=1-sim_{i,j}$$

K-means clustering is simple and efficient.Exact number of user search goals is unknown so we set K as 5 which we perform clustering based on 5 values.After clustering each cluster is considered as a search goals.Therefore the center point of cluster is computed,the term with highest center point is used as a keyword.Thus advantage of using keyword is that represent user information needs more effectively.

V RESTRUCTURING WEB SEARCH RESULTS

When a query is submitted it results in millions of search results,so it is necessary to organize the results obtained from search engine.Restructuring web results makes easier for the user to find what they need.Performing categorization by choosing the smallest distance between URL and user search goals vectors.

VI EVALUATION CRITERION

The performance of restructured web search results is evaluated by "classified Average Precision"(CAP).

1.Average precision: It is calculated according to given user feedback. Average precision is computed at the point of each clicked document in the ranked sequence of user feedback.

$$AP=1/N+\sum_{r=1}^{N} rel(r)\frac{R_r}{r}$$

Where N+ is the number of clicked documents from total retrieved documents in single user feedback session,r is the rank,N is the total number of retrieved documents, rel() is a binary function on the relevance of a given rank, and Rr is the number of relevant retrieved documents of rank r.

*2.Voted average precision:*It is calculated for restructured search results classes.It is also similar to average precision.

$$VAP=1/Nc+\sum_{r=1}^{N} rel(r)\frac{R_r}{r}$$

Where Nc+ is the number of clicked documents from the class having maximum number of clicks.

3.Risk: It is calculated to avoid wrong classification of search results.

$$Risk=\sum_{ij=1(i<j)}^{m}\frac{d_{ij}}{c_m^2}$$

Where d_{ij} will be 1 or 0,m is the number of clicked URLs,$c_m^2=m(m-1)/2$.
4.Classified average precision: It is used to evaluate the performance of restructured search results.

$$CAP=VAP*(1-Risk)^{\Upsilon}$$

Where Υ is used to adjust the influence of risk on CAP.

VII. CONCLUSION & FUTURE WORK

In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions.First, we introduce feedback sessions that is to be analyzed to infer user search goals. Both the clicked URLs and the unclicked ones are taken into account to construct feedback sessions. Therefore,feedback sessions can reflect user information needs more efficiently. Secondly, we map feedback sessions to pseudodocuments to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods.The complexity of our approach is low and our approach can be used in reality easily. For each query, the running time depends on the number of feedback sessions.Therefore, the running time is usually short. In reality, our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently

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 Classification," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development (SIGIR '07), pp. 783-784, 2007.

[4] M. Pasca and B.-V Durme, "What You Seek Is what You Get: Extraction of Class Attributes from Query Logs," Proc. 20th Int'l Joint Conf. Artificial Intelligence (IJCAI '07), pp. 2832-2837, 2007.

[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] X. Li, 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.

[10] 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.