Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 3, March 2015, pg.508 – 512

RESEARCH ARTICLE

A New Technique to Optimize User's Browsing Session using Data Mining

Neha D. Anandpara[1], Protik Ganguly[2], Vidhi Doshi[3], Prof. Govind Wakure[4]

[1]Information Technology, Mumbai University, India

[2]Information Technology, Mumbai University, India

[3]Information Technology, Mumbai University, India

[4]Information Technology, Mumbai University, India

[1] neha.anandpara1@gmail.com; [2] protik.ganguly@gmail.com;
[3] vidhidoshi1610@gmail.com; [4] govind.wakure@mctrgit.ac.in

Abstract— *From a wide range of options, an ambiguous query would return different search results for different users when they submit it to a search engine. However, ambiguous query/topic submitted to search engine doesn't satisfy user information needs every time, because different users may have different information needs on diverse aspects upon submission of same query/topic to search engine. So deducing search goals as per the needs of variety of users becomes complicated. The evaluation and interpretation of user search goals can be very useful in optimizing search engine performance and users' need of information. This paper proposes a new method for deducing user search goals by analysing user click logs from the search result tray. The proposed approach is used to discover different user search goals for a query by clustering the user feedback sessions. Feedback sessions are constructed from click-through logs of various search options. Pseudo-items are first generated in this method to represent feedback sessions in a better way for clustering. Finally, pseudo-items are clustered to discover different user search goals and show them with some keywords. Then these user search goals are used to restructure the web search results by "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals.*

Keywords— *User search goals, feedback sessions, pseudo-items, restructuring search results, classified average precision*

I. INTRODUCTION

Queries return the information needed by the user which when submitted to the search engine in a web application. However, sometimes queries may not satisfy users' specific information needs since several ambiguous queries may cover a wide range of topic and different users may want to get information on different aspects when they submit the same query. For example, "parachute" is submitted as a query by the user to the search engine, some users probably will be interested to extract information about hair oil while some users will wish to gather information regarding the hot air balloon. Therefore, it is essential to capture different user search goals in information retrieval. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can further lead to improvement of search engine performance and user experience. Some advantages are summarized as follows:

- According to user search goals web search results can be restructured by grouping the search results with the same search goal; thus, users with different search goals can easily find what they want.

- Keywords used to represent User search goals can be utilized in query recommendation; thus, queries can be formed more precisely by the user with the help of suggested queries.
- The distributions of user search goals can also be useful in applications such as re-ranking web search results that contain different user search goals.

Clustering the results of our search is an efficient method to organize search results, which allows a user to find a way in the needed documents quickly. Our motive is to determine different user search intents for a query and depict each search result/intent with some keywords automatically. To determine the user information automatically at different point of view with user given query and collects the similar search result with URL first we collect similar feedbacks sessions from user click-through logs of different search engines. Then, map feedback sessions to pseudo-items which shows user information needs. At last, these pseudo documents are clustered using "Apriori" clustering algorithm to deduce user search intents. Since the evaluation of clustering is also an important problem, we also propose a novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results and also helps us to optimize the parameter in the clustering method when inferring user search intents.

II. PROPOSED SYSTEM

In this section, we have illustrated the basic operations involved in the approach we are trying to propose to determine user search intents by clustering pseudo- documents. The flow of the proposed system design is as shown in Fig. 1.

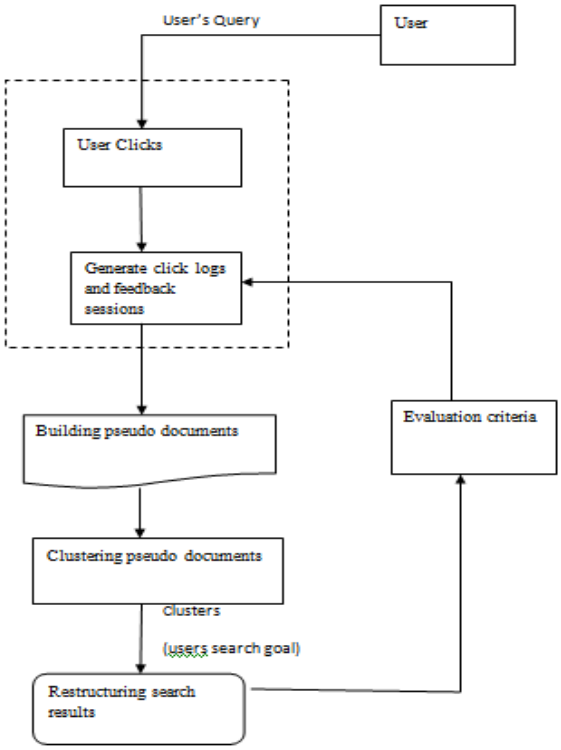

Fig.1: Flow of the Proposed System

A. Clickthrough data

There are multiple number of queries and user clicks in web environment. In this framework, user clicks are recorded in user clickthrough data that represent implicit relevance feedback. The clickthrough data is stored in user logs which are used to personalize user experience in web search. In general, when query is written, the user usually scans the links to documents in a result list from top to bottom. The user clicks on the links to the documents that is an appropriate choice and skips other documents. Therefore, the proposed approach utilizes user clicks as relevant judgments to evaluate search precision since clickthrough data can be collected at low cost, it is possible to do large scale evaluation under this framework.

1) Feedback sessions: Feedback sessions are considered as users' implied search results. A session for web search is a sequence of successive queries to satisfy single need and some clicked results. But to infer user search intents for a particular query, single session is considered. A single session includes both clicked and

unclicked URLs which are evaluated by user and session ends with the last clicked URL. In each feedback session, clicked URLs depict the users' information needs while the unclicked URLs depict what users are not interested in. Due to the presence of multiple feedback sessions in user clickthrough logs, examining feedback sessions to deduce user search intents is better than examining clicked URLs or search results.

B. Building pseudo-items

The query's intended result is not obtained by just the URLs. To obtain the accurate result and information needed, we enhance each URL with additional text content by extracting the keywords of URLs appearing in feedback session

C. Clustering pseudo-items with Apriori

Input: The dataset (D) and min_s.
Output: The frequent dataset.
1. x = 1;
2. Find frequent set, Lx from Cx, the set of all candidate datasets;
3. Form Cx+1 from Lx;
4. x = x+1;
5. Repeat 2-4 until Cx is empty;
Step 2 is called the frequent dataset generation step. Step 3 is called as the candidate dataset generation step. Details of these two steps are described below.
Frequent itemset generation
Scan D and count each dataset in Cx, if the count is greater than min_s, then add that itemset to Lx.
Candidate dataset generation
For k = 1, C1 = all datasets of length = 1.
The join step:
Cx = x-2 way join of Lx-1 with itself.
If both {a1,..,ax-2, ax-1} & {a1,.., ax-2,
ax} are in Lk-1,
then add {a1,..,ax-2, ak-1, ax} to Cx.
The items are always stored in the sorted
order.
The prune step:
If any non-frequent (x-1) subset found in {a
a3…….....ax} then discard this set.

In order to cluster pseudo-items with Apriori, the important factor is to define the distance measure between two data points and defining the number of clusters. Apriori algorithm is used to cluster pseudo-items because of its simplicity and effectiveness. Apriori clustering results in good quality performance for document clustering. After clustering all pseudo-items, each cluster denotes user search goal i.e. intention of user. The combination of item sets, or features, from the result of association rule mining has many patterns with several groups of items. Users have to set threshold of minimum support value to limit the result that shows only groups of item sets related to the specified criteria. The results are also filtered by minimum confidence value that is to be specified by user corresponding to user's requirement and usage. Association results include group of related features called "item set" that are considered in each frequent item set, for example, examine two related features in co-occurrence type is called two-item set. Apriori is a structure to count candidate item sets efficiently. It generates candidate item sets of length x from the x-1 item sets and avoids expanding all the item set's graph. Then it prunes the candidates which have an infrequent sub pattern. The candidate set contains all frequent x-length data sets. After that, it scans the transaction database to determine frequent item sets among the candidates. With Apriori technique the algorithm can decrease time processing in generating fewer groups of item sets and avoid infrequent candidate item sets expansion.

D. Restructuring web search results

Web search results are reorganized on the basis of discovered user search intents. Then categorize each URL into a cluster, centered with user search intents by selecting smallest distance between user search goal and URL.

E. Evaluation criterion

The performance of restructured (clustered) web search results and original search results is evaluated by using parameters like Average Precision (AP) , Voted AP (VAP) which is AP of the class having more clicks, Risk to avoid wrong classification of search results and Classified AP (CAP). If user got correct classified results with higher CAP value, this value is used to optimize the number of clusters of user search goals.

1) *Average precision (AP):* It is calculated according to given user feedbacks. AP is the average of precisions computed at the point of each clicked document in the ranked sequence of user feedback.

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

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 R_r is the number of relevant retrieved documents of rank r or less.

2) Voted AP (VAP): It is calculated for purpose of restructuring of search results classes i.e. different clustered results classes. It is same as AP and calculated for classes which have more clicks.

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

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

3) Risk: Sometimes VAP will always be highest value because each URL from single session is classified into the single class no matter whether users have different search goals or not. So, there should be a risk to avoid wrong classification search results into too many classes. It evaluates the normalized number of clicked URL pairs that are not in the same class.

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

Where m is number of clicked URLs and d_{ij} is 0 if pair of clicked URLs belongs to same class otherwise d_{ij} is 1.
If the pair of the i^{th} clicked URL and the j^{th} clicked URL are not categorized into one class, dij will be 1; otherwise, it will be 0. $(C_m)^2 = m(m-1)/2$ is the total number of the clicked URL pairs.

4) Classified AP (CAP): New criterion Classified AP (CAP) is extension of VAP by using above Risk. It combines AP of class having more clicks and risk of wrong classification. It is used to evaluate performance of restructured search results.

$$CAP=VAP*(1-Risk)$$

III. CONCLUSION

In this paper, a new technique has been proposed to deduce user search goals for a query by using Apriori algorithm for clustering its feedback sessions represented by pseudo-items. By using proposed system, the inferred user search goals can be used to restructure web search results. So, users can find exact information quickly and very efficiently. The discovered clusters of query can also be used to assist users in web search.

Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods.

Acknowledgement

We authors, would like to thank our guide for this project and paper**, Prof. Govind Wakure**, IT Dept., RGIT, Mumbai for his constant support and encouragement.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*.ACM Press, 1999.
[2] 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.
[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] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin, Zhaohui Zheng, "*A New Algorithm for Inferring User Search Goals with Feedback Sessions*", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, pp.502-513, 2013.
[5] J.-R Wen, J.-Y Nie, and H.-J Zhang, "*Clustering User Queries of Search Engine,*" Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.

