

An Innovative Method for Concluding User Search Intention Using Feedback Session

Vipul D. Punjabi

Department of Computer Engineering, R. C. Patel Institute of Technology, Shirpur, Maharashtra, India

ABSTRACT

With the advent of computers, it became possible to store large amounts of information and finding useful information from such collections became a necessity of today's world. For a broad-topic and ambiguous query, different users may have different search goals. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. Hence, in this paper we have tried to improve the efficiency of the search engine by combining inferring user search goal with feedback sessions. Initially, we composed a framework to implement different user search goals for an ambiguous query with the help of clustering the proposed feedback sessions. The user needs are reflected efficiently through the feedback sessions built by the user click-through logs. Second, to represent better feedback sessions for clustering we generate the pseudo document. Finally, we proposed a new criterion "Classified Average Precision (CAP)" to evaluate the performance of inferring user search goals. Therefore, when users submit their 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.

Keywords: Query logs, clicked un-clicked, pseudo-document, clustering, feedback session, Classified Average Precision (CAP).

I. INTRODUCTION

With the increasing size and popularity of the Internet, there exist a billion static web pages and some commercial search engine service have tens of millions of queries per day [3]. This made necessary to evolve an acute need of automatic methods to organize this data [3]. After organizing of data another important challenge is to display the accurate and relevant queries reply to the user. Web pages on different aspects of the same topic are mixed together in the returned list. The user has to shift through a long list to locate pages of interest [6]. However, sometimes queries may not exactly represent users' specific information needs.

For an example, if a user raises a query "gladiator". It is hard to determine the user's search intent, i.e., whether the user is interested in the history of gladiator, famous gladiators, or the film gladiator [5]. Such confusion can

be sort out in various steps of inferring and analysis approach.



Figure 1: The examples of the different user search goals and their distributions for the query "the sun" by our experiment.

This paper uses the inference and analysis of user search goals to improve the results of search engines. We aim at discovering the number of diverse user search goals

for a query and depicting each goal with some keywords automatically.

II. PROBLEM DEFINITION

We first propose a novel approach to infer user search goals for a query by clustering our proposed feedback sessions. The feedback session is defined as the series of both clicked and un-clicked URLs and ends with the last URL that was clicked in a session from user click-through logs.

Then, we propose a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords. 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.

Framework of our Approach

The framework of our approach is detailed explain in Fig 2. Our framework consists of two parts which is separated by the dashed line. The part above line shows, all the feedback sessions of a query which are initially extracted from user click-through logs and mapped to pseudo-documents. Following this the user search goals are concluded by clustering these pseudo-documents with some keywords. As the exact number of searches are unknown to us optimal value will be determine by the bottom part. In the bottom part, the original search results are restructured based on the user search goals inferred from the upper part. Then, we evaluate the performance of restructuring search results by our proposed evaluation criterion CAP. And the evaluation result will be used as the feedback to select the optimal number of user search goals in the upper part.

III. ALGORITHM

In this paper the main approach is to build such a search engine results that it is time consuming and relevant to the users. Therefore the framework of our approach is divided into 4 sections.

3.1 Representation of feedback session

In this section, first we proposed feedback session and then the proposed pseudo-documents to represent feedback session.

3.1.1 Feedback sessions

In this paper, we focus on inferring user search goals for a particular query. Generally, a session for web search is a series of successive queries to satisfy a single information need and some clicked search results [11]. Therefore, the single session containing only one query is introduced, which makes it different from the conventional session. Also, the feedback session in this paper is based on a single session, although it can be extended to the whole session.

Search results	Click sequence
www.imdb.com/title/tt0172495/	3
www.exovedate.com/the_real_gladiator.html	0
en.wikipedia.org/wiki/Gladiator	1
www.rottentomatoes.com/m/gladiator/	0
www.history.com/.../10-things-you-may-not-know-about-roman-gladiat	2
www.gladiatorgarageworks.com/	0
https://soundcloud.com/gladiatormusic	0
www.bbc.co.uk/history/ancient/romans/launch_gms_gladiator.shtml	0
www.bbc.co.uk/history/interactive/games/gladiator/index_embed.shtml	0
www.miniclip.com/games/gladiator-arena/en/	0

Figure 2 : A feedback session in a single session. “0” in click sequence means “unclicked.” All the 10 URLs construct a single session. The URLs in the rectangular box construct a feedback session.

The proposed feedback session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. Fig. 3 shows an example of a feedback session and a single session. In Fig. 3, the left part lists 10 search results of the query “the sun” and the right part is a user’s click sequence where “0” means “unclicked.” The single session includes all the 10 URLs

in Fig. 3, whereas the feedback session only includes the seven URLs in the rectangular box. The seven URLs consist of three clicked URLs and four unclicked URLs. Normally, since users will scan the URLs one by one from top to down, we can consider that besides the three clicked URLs, the four unclicked ones in the rectangular box have also been browsed and evaluated by the user and they should reasonably be a part of the user feedback. The result appears from rectangular box is that URLs having non zero number is the user desired whereas the URLs having zero is what user does not want. It should be considered that the unclicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not.

Each feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

3.1.2 Map feedback sessions to pseudo documents

As there exists lot of feedback sessions for different clicks for queries, hence it is not suitable to directly use feedback sessions for inferring user search goals. There are many representation method to describe feedback sessions in a more efficient and coherent way. For example, Fig. 4 shows a popular binary vector method to represent a feedback session.

Combining Fig 3, using the clicked sequence of the user further the binary table is created where the non-zero number is assign value binary "1" and zero i.e. unclicked URLs is assign binary "0". Moreover, binary vector representation is not informative enough to tell the contents of user search goals. Therefore, it is improper to use methods such as the binary vectors and new methods are needed to represent feedback sessions.

Search results	Click Sequence	Binary vector
www.imdb.com/title/tt0172495/	3	1
www.exovedate.com/the_real_gladiator.html	0	0
en.wikipedia.org/wiki/Gladiator	1	1
www.rottentomatoes.com/m/gladiator/	0	0
www.history.com/.../10-things-you-may-not-know-about-roman-gladiat	2	1
www.gladiatorgarageworks.com/	0	0
https://soundcloud.com/gladiatormusic	0	0

Figure 3: The binary vector representation of a feedback session.

In this paper, we propose a new way to map feedback sessions to pseudo-documents, as illustrated in Figure. 4.

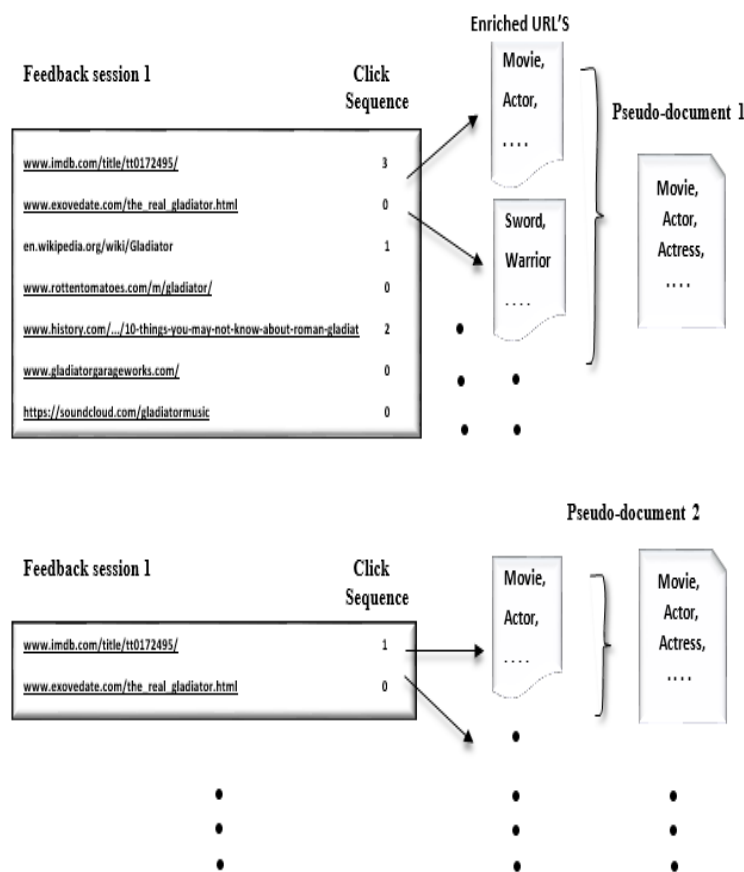


Figure 4 : Illustration for mapping feedback sessions to pseudo-documents.

3.1.3 Inferring user search goal by clustering pseudo document

We cluster pseudo-documents by K-means clustering which is simple and effective. Since we do not know the exact number of user search goals for each query, we set K to be five different values (i.e., 1; 2; . . . ; 5) and perform clustering based on these five values, respectively. The optimal value will be determined through the evaluation criterion. After clustering all the pseudo-documents, each cluster can be considered as one user search goal. The center point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster.

Finally, the terms with the highest values in the center points are used as the keywords to depict user search goals. Note that an additional advantage of using this key word based description is that the extracted keywords can also be utilized to form a more meaningful query in query recommendation and thus can represent user information needs more effectively. Moreover, since we can get the number of the feedback sessions in each cluster, the useful distributions of user search goals can be obtained simultaneously. The ratio of the number of the feedback sessions in one cluster and the total number of all the feedback sessions is the distribution of the corresponding user search.

IV. CONCLUSION

In this paper, a breaking new ground approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. Unlike previous methods, our approach considers not only the current query but also the recent queries in the same session to provide more meaningful suggestions. Moreover, we group similar queries into concepts and provide suggestions based on the concepts. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the un clicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to clustering pseudo documents to approximate goal texts in user minds. Finally, we propose the novel criterion "Classified Average

Precision" (CAP) is formulated to evaluate the performance of user search goal inference. Based on the proposed criterion, we also describe the method to select the best cluster number. Experimental results on user click-through logs from a commercial search engine demonstrate the effectiveness of our proposed methods. 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.

V. REFERENCES

- [1] 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.
- [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] 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.
- [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] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.
- [6] 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.