

DATA AWARE QUERY SUGGESTION BY USER CLICK THROUGH FEEDBACK

Dr. P.Perumal

Associate Professor,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, India

N. Archana Reddy,

UG Student,

Department of Computer Science and Engineering,
Sri Ramakrishna Engineering College, Coimbatore, India

Abstract: Given an ambiguous query, different users may have different search goals and the accurate links may not be produced. The current existing system does not look through the complete description and content of every link. A novel approach is proposed to infer user search goal by analyzing search engine query logs. First, producing a feedback session which are constructed from user click through logs and can efficiently the information needed by the user. Second the removal of the stop and stemming words from the given query and produce the related links. Third, by using administrator login and user login, storing the user's history though used from different systems and the mail is sent regarding the recent viewed link to the user's mail id. Generating the pseudo documents to represent feedback session for clustering. Finally the classified average precision is used to restructure and evaluate the performance of inferring user search goals.

Keywords: *User search goal, feedback session, pseudo-documents, classified average precision*

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 user's 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. Defining the user search goals as the information on different aspects of a query that user groups want to obtain.

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 in improving search engine relevance and user experience. Some advantages are summarized as follows. First, restructuring web search results. According to user search goals by grouping the search results with the same search goal.

Thus the users with different search goals can easily find what they want. Second, user search goals represented by some keywords can be utilized in query recommendation thus; the suggested queries can help users to form their queries more precisely. Third, the distributions of user search goals can also be useful in applications such as reranking web search results that contain different user search goals.

In this paper, aiming at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First proposing 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 unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs. Then, proposing a novel optimization method to map feedback sessions to pseudo-documents which can efficiently reflect user information needs.

At last, clustering these pseudo-documents to infer user search goals and depict them with some keywords. Since the evaluation of clustering is also an important problem, also proposed a novel evaluation criterion classified average precision (CAP) to evaluate the performance of the restructured web search results. Then also demonstrate that the proposed evaluation criterion can help us to optimize the parameter in the clustering method when inferring user search goals.

Hence proposing framework to infer different user search goals for a query by clustering feedback sessions. Then demonstrating that clustering feedback sessions is more efficient than clustering search results or clicked URLs

directly. Moreover, the distributions of different user search goals can be obtained conveniently after feedback sessions are clustered.

Proposing a novel optimization method to combine a pseudo-document, this can effectively reflect the information need of a user. Thus, finally can tell what the user search goals are in detail. Then proposing a new criterion CAP to evaluate the performance of user search goal inference based on restructuring web search results. Thus, determining the number of user search goals for a query. The rest of the paper is organized as follows: The framework is described In Section 2, pseudo-documents in Section 3, tag extraction in Section 4, user search and user logs in Section 5, the Map feedback sessions to pseudo documents is proposed in Section 6 and the experimental result and analysis in Section 7 and finally the section 8 concludes the paper.

II. FRAME WORK

Our framework consist administrator login and tag extraction and also the user registration and login. The query entered is searched and the related links are provided after removing stop and stem words are. Once the link is clicked the count keeps increasing in the feedback session [1].

All the feedback sessions of a query are first extracted from user click-through logs and mapped to pseudo-documents. Then, user search goals are inferred by clustering these pseudo-documents and depicted with some keywords. Since the framework do not know the exact number of user search goals in advance, several different values are tried and the optimal value will be determined by the feedback.

The original search results are restructured based on the user search goals inferred from the upper part [1]. Then, evaluating 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.

III. EXPERIMENT

The secret key, date of birth and nick name attributes are taken as the factors of Eigenvector. With the above details the customer details are cross checked and verified in order to detect the outliers. Further the user is granted access only if

the attributes are matching else the user is redirected to the home and the access will be denied. This is a very effective method for fraud detection using osPCA.

A. FEEDBACK SESSION

In this paper, focusing on inferring user search goals for a particular query. Therefore, the single session containing only one query is introduced, which distinguishes from the conventional session [1]. Meanwhile, the feedback session in this paper is based on a single session, although it can be extended to the whole 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.

It is motivated that before the last click, all the URLs have been scanned and evaluated by users [3]. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. Each feedback session can tell what a user requires and the user does not care about it. 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.

IV. PSEUDO DOCUMENTS

Since feedback sessions vary a lot for different click-through and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation method is needed to describe feedback sessions in a more efficient and coherent way.

There can be many kinds of feature representations of feedback sessions [6]. Representing the URLs in the feedback session. In the first step, first enrich the URLs with additional textual contents by extracting the titles and snippets of the returned URLs appearing in the feedback session. In this way, each URL in a feedback session is represented by a small text paragraph that consists of its title and snippet.

Then, some textual processes are implemented to those text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Forming pseudo-document based on URL representations. In order to obtain the feature representation of a feedback session, proposed an optimization method to combine [5] both clicked and

unclicked URLs in the feedback session. It is worth noting that people will also skip some URLs because they are too similar to the previous ones.

In this situation, the “unclicked” URLs could wrongly reduce the weight of some terms in the pseudo-documents to some extent [6].

A. CLUSTERING PSEUDO DOCUMENTS

Clustering the pseudo-documents by K-means clustering which is simple and effective. Since it do not know the exact number of user search goals for each query, setting K to be five different values (i.e., 1; 2; . . . ; 5) and perform clustering based on these five values, respectively [5].

The optimal value will be determining 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 keyword 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 getting the number of the feedback sessions in each cluster, the useful distributions of user search goals can be obtained simultaneously [4].

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 goal.

Each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document is F_{fs} . The similarity between two pseudo-documents is computed as the cosine score of F_{fs_i} and F_{fs_j} , as follows [3]:

$$\begin{aligned} Sim_{i,j} &= \cos(\mathbf{F}_{fs_i}, \mathbf{F}_{fs_j}) \\ &= \frac{\mathbf{F}_{fs_i} \cdot \mathbf{F}_{fs_j}}{|\mathbf{F}_{fs_i}| |\mathbf{F}_{fs_j}|} \end{aligned} \quad (1)$$

And the distance between two feedback sessions is

$$Dis_{i,j} = 1 - Sim_{i,j}. \quad (2)$$

Clustering pseudo-documents by K-means clustering which is simple and effective. Since that do not know the exact number of user search goals for each query, set K to be five

different values (i.e., 1, 2, . . . 5) and perform clustering based on these five values, respectively. After clustering all the pseudo-documents, each cluster can be considered as one user search goal [5].

The center point of a cluster is computed as the average of the vectors of all the pseudo-documents in the cluster. Since it 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 goal.

V. TAG EXTRACTION

The administrator is the responsible for maintaining all the information in the server. The data uploading based on the user personalization and the group related information are uploaded by the administrator in the server. The admin has the information about the list of user processing in this application. The regular updating of query related results are added and maintained by the administrator in the server. Administrator will add tags to URLs.

A tag tree for the page rooted in the <HTML> tag is constructed. Each node represents a tag in the HTML page and its children are tags enclosed inside it. Each internal node of the tag tree has a tag string, which includes all tags of its descendants, and a tag path. Once the URL is entered the keyword and description is extracted from the source code of the web page and the tag is added to the sql 2005.

VI. USER SEARCH AND USER LOGS

User can enter the URL of search engine then login if the user gives correct username, password then it transfer to search web page. In the search page user can give the query to search information. Stop and stemming words of user's query has been removed by using Codec then the keywords of the query have been used for search. Web databases generate query result pages based on a keyword. In this a tag tree for the page rooted in the <HTML> tag is constructed.

Automatically extracting the data from these query result pages is very important for many applications, such as data

integration, which need to cooperate with multiple web databases. Search engine logs record the activities of web users, which react the actual user's needs or interests when conducting web search. They generally have the following information, text queries that users submitted, the URLs that they clicked after submitting the queries, and the time when they clicked.

Search engine logs are separated by sessions. A session includes a single query and all the URLs that a user clicked after issuing the query. Our idea of using search engine logs is to treat these logs as past history, learn users' interests using this history data automatically, and represent their interests by representative queries. For example, in the search logs, a lot of queries are related to car and this reacts that a large number of users are interested in information about different car users are probably interested in different aspects of car.

Some are looking for renting a car, thus may submit a query like car rental some are more interested in buying a used car, and may submit a query like used car and others may care more about buying a car accessory, so they may use a query like car audio. By mining all the queries which are related to the concept of car, can learn the aspects that are likely interesting from a user's perspective. User may give ambiguous query and have different search goals.

Discovering the user search goals by collecting the last URL that was clicked in a session by the user for a query. User click logs are dynamically updated. Each user search goals has depicted with some other words. Then these keywords are clustered. Search history and IP address of system in which user performed search are sent to corresponding user's mail. Search history includes clicked or viewed URLs of user [3].

The query result is important while searching the data. The exact information should be obtained based on the user query and user search goals. Here proposing query relevance based on Web search logs. Our measure of relevance is aimed at capturing queries that have induced the users to click on similar sets of pages [4]. URLs that are clicked by user are stored in feedback sessions and the history of user search is also maintained. Prediction process helps to optimize the data to get the exact search results.

User can view searching information based on information available in the server as well as the related data extracted by the several users.

VII. MAP FEEDBACK SESSIONS TO PSEUDO-DOCUMENTS

Since feedback sessions vary a lot for different click-through and queries, it is unsuitable to directly use feedback sessions for inferring user search goals. Some representation method is needed to describe feedback sessions in a more efficient and coherent way. There can be many kinds of feature representations of feedback sessions, when the query is submitted, and "0" represents "unclicked" in the click sequence [6].

The binary vector [0110001] can be used to represent the feedback session, where "1" represents "clicked" and "0" represents "unclicked." However, since different feedback sessions have different numbers of URLs, the binary vectors of different feedback sessions may have different dimensions. 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.

For a query, users will usually have some vague keywords representing their interests in their minds. They use these keywords to determine whether a document can satisfy their needs. Naming these keywords "goal texts". However, although goal texts can reflect user information needs, they are latent and not expressed explicitly. Therefore, introducing pseudo-documents as surrogates to approximate goal texts. Thus, pseudo-documents can be used to infer user search goals.

In this paper, proposing a novel way to map feedback sessions to pseudo-documents. The Restructuring Web Search Results Since search engines always return millions of search results, it is necessary to organize them to make it easier for users to find out what they want. Restructuring web search results is an application of inferring user search goals. Introducing how to restructure web search results by inferred user search goals at first. Then, the evaluation based on restructuring web search results will be described.

The inferred user search goals are represented by the vectors and the feature representation of each URL in the search results can be computed. Then, can categorize each URL into a cluster centered by the inferred search goals. In this paper, performing categorization by choosing the smallest distance between the URL vector and user-search-goal vectors. By this way, the search results can be restructured according to the inferred user search goals [4].

In order to apply the evaluation method to large-scale data, the single sessions in user click-through logs are used to minimize manual work. Because from user click-through logs, can get implicit relevance feedbacks, namely “clicked” means relevant and “unclicked” means irrelevant.

A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks. AP is the average of precisions computed at the point of each relevant document in the ranked sequence.

VIII. EXPERIMENTAL RESULT AND ANALYSIS

The following figure 1 shows the performance analysis of existing and proposed work.

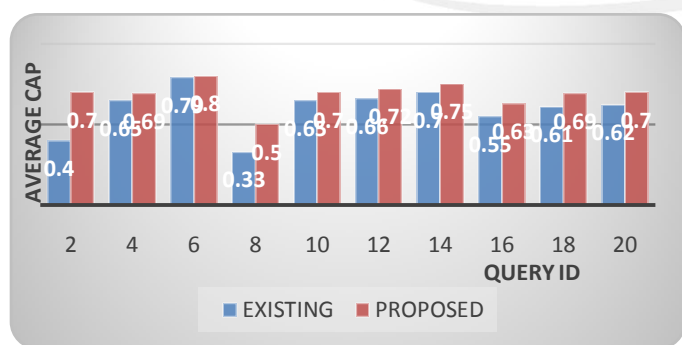


Figure 1: Result Analysis of Proposed System

IX. CONCLUSION

In this paper, using the clustering method to provide an efficient webpage. The web page are generally heterogeneous and dynamic because of which is needed to use a method to produce better search results. The administrator, web search logs and web search results can be taken together in a sensible order to get the results from feedback session [8]. The feedback session stores the complete data on the click counts and is dynamically updated once the user clicks the link. The feedback session produces pseudo documents with the

keywords .The pseudo documents are then clustered to produce the result.

By using the administrator login the required tag can be extracted along with the keywords and description which helps in searching based on content as well as on the link. The user can use personal id and password to login before searching a link or the content in the link [8].

This process of login before search helps to store the complete history of the user search which is stored in feedback session and even if the user searches from various systems with different IP address is still stored in feedback session that provides better search results.

Once the user search by using the content search they can either just view the content described in source code of the webpage or download it as a data file. The recent visit link is displayed at the end of the search page to keep the user updated about their recent search link and a mail is sent to the user mail id regarding their recent viewed link.

This method ensures security as only the particular user can look through the search history and any other user using the same system cannot find out about the search history of the user. The method of clustering in data mining is used in this system which helps in producing better search results.

X. ACKNOWLEDGMENT

We thank our Director, Principal and the management of Sri Ramakrishna Engineering College for providing lab

XI. REFERENCES

- [1] Zheng Lu, Hongyuan Yang, “ A New Algorithm for Inferring User Search Goals with Feedback Sessions,” Text Mining,pp. 87-94, March 2013.
- [2] B. Poblete and B.-Y Ricardo, “Query-Sets: Using Implicit Feedback and Query Patterns to organize Web Documents,” Proc. 17th Int’l Conf. World Wide Web (WWW ’08), pp. 41-50, 2008.
- [3] 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.

- [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] 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.
- [6] 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.
- [7] T. Joachims, "Evaluating Retrieval Performance Using Click through 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 Click through Data," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

