Survey on the Use of Feedback Sessions for Inferring User Search Goals

Mr. Ajinkya A. Godbole
ME Student, Department of Computer Engineering, P.V.P.I.T., Bavdhan, Savitribai Phule Pune University, Pune, Maharashtra, India
ajinkyagdbl@gmail.com

Mrs. V. S. Nandedkar
Assistant Professor, Department of Computer Engineering, P.V.P.I.T., Bavdhan, Savitribai Phule Pune University Pune, Maharashtra, India
vaishu111@gmail.com

Abstract **-** Largest source of web traffic are search engines. Search engines are being used by different kind of users for different purpose. When users are searching something they have a different search goal in mind. Thus the queries are mostly ambiguous one. In order to improve search engine relevance and thus user experience inference and analysis of user search is required. To get the best results it is needful to capture different user search goals. This paper first talks about the different ways of inferring user search goals. Then insights of new approach has been discussed. A new algorithm firstly specifies a framework to analyze user search goals by clustering feedback sessions. There should be a proper way to represent these feedback sessions. In the second step of this algorithm pseudo-documents are prepared to represent feedback sessions. With this original results are restructured. This in turn is used to select optimal user search goals.

Keywords- *search engines, user search goals, feedback sessions, pseudo-documents*

*****

I. INTRODUCTION

Search engine is the most important application in today's internet. User needs some information and thus queries to internet inorder to get the result. Most of the times these queries are ambiguous. Means user is expecting information in one topic is not returned by the search engine as search engine interprets the query differently. For example, when the query is "gladiator". It is hard to determine what user is expecting in result in such scenarios as query is ambiguous. It is hard for a search engine to decide if the user is interested in history of a gladiator or list of famous gladiators or the film gladiator.

Without looking at the context of search, search engine suggests many queries with very low accuracy. Thus it is required to capture user search goal. Information need is nothing but a user's desire to satisfy his/her need. In order to improve user search goals the inference and analysis of goals have a lot of advantages. First advantage is web search results can be restructured [7], [4], [5] according to user search goals by grouping the search results with the same search goal. Another advantage is the usage of keywords to represent user search goals in the query suggestion [8], [9], [10]. Third advantage would be reranking of web search results can also be done with the distribution of user search goals.

User search goals can be represented in following three classes: Query classification, Search Result Reorganization and Session Boundary Detection. In Query classification, classification is done depending upon some predefined classes. User goals are either navigational or informational. In case of navigational user goal user has web page in mind. In case of informational user does not have any particular page in mind. In case of search result reorganization user try to recognize search result. This is done either by learning aspects of queries by analyzing the clicked URLs or by analyzing search results returned by a search engine. In third method the main aim is to detect session boundaries. Feedback session ends with the last URL clicked in a session and contains both clicked and unclicked URLs.

The rest of the paper is organized as follows: Section II is about the previous and current methods in use and Section III talks about the proposed system. Finally paper ends with the conclusion in section IV.

II. LITERATURE SURVEY

A. Automatic identification of user goals:

Uichin Lee, Zhenyu Liu, Junghoo Cho [2], proposed automatic identification of user search goals. Majority of queries have a goal which is predictable was the statement of them. Classification of query goals based on two types:

A1. Navigational queries

In case of navigational user has web page in mind. User may have visited that site before or predicts that site may exist.

A2. Informational queries

In case of informational user does not have any particular page in mind. User also may intend to visit different pages to know about the topic. In this type user keeps on exploring webpages. User does not have a guarantee which page is going to have correct required answer.

For the prediction of user goal two features are used:

1. Past user-click behavior:

In case of navigational, users has a result in the mind and will click on that result. So, user goal can be identified by Observing the past user-click behavior.

2. Anchor-link distribution:

If the user is associating query with website then links with the anchor will point to respective websites. So potential goal of the query can be identified by observing destinations of the links with the keyword of the query.

B. Web query classification

Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen[3], proposed classification of web queries into target categories where there is no training data and queries are very short. Here there is no need of collecting training data as intermediate classification is used to train target categories and classifiers bridging. Following are internal classification approaches:

B1. Classification by exact matching

It has two categories defined. First is the intermediate taxonomy and the other is target taxonomy. Given a certain category in an intermediate taxonomy, we say that it is directly mapped to a target category if and only if the following condition is satisfied: one or more terms in each node along the path in the target category appear along the path corresponding to the matched intermediate category. For example, the intermediate category "Computers\Hardware \Storage" is directly mapped to the target category "Computers\Hardware" since the words "Computers" and "Hardware" both appear along the path Computers → Hardware → Storage

B2. Classification by SVM

Query classification with SVM consists of the following steps: 1) construct the training data for the target categories based on mapping functions between categories. If an intermediate category CI is mapped to a target category CT, then the Web pages in CI are mapped into CT; 2) Train SVM classifiers for the target categories; 3) For each Web query to be classified, use search engines to get its enriched features

B3. Classifiers by bridges

It is taxonomy-bridging classifier or bridging classifier by which target taxonomy and queries are connected by taking an intermediate taxonomy as a bridge. To reduce the computation complexity category selection is performed.

C. Reorganizing search results

Xuanhui Wang and ChengXiang Zhai[4], published a work on clustering of search results. This clustering organizes it and allows a user to navigate into relevant documents quickly. Two deficiencies of this approach make it not always work well: First is the clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective; and the second one the cluster labels generated are not informative enough to allow a user to identify the right cluster. In this paper, they propose to address these two decencies by following two steps:

1. Learning "interesting aspects" of a topic from Web search logs and organizing search results accordingly
2. Generating more meaningful cluster labels using past query words entered by users.

D. Clustering web search results

Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma[5], re-formalized the search result clustering problem as a salient phrases ranking problem. Thus they convert an unsupervised clustering problem to a supervised learning problem. Although a supervised learning method requires additional training data, it makes the performance of search result grouping significantly improve, and enables us to evaluate it accurately. This new algorithm has following four steps:

1. Search result fetching
2. Document parsing and phrase property calculation
3. Salient phrase ranking
4. Post-processing.

First the webpage of search results is returned by some web search engine. HTML parser then analyzes these webpages and result items are extracted. Phrases are ranked according to salience score. The top ranked phrases are taken as salient phrases. Then post processing is performed which filters out the pure stop words.

E. Session boundaries

Rosie Jones and Kristina Lisa Klinkner[6], published a work on session boundaries and automatic hierarchical segmentation of search topics in Query Logs. In this work they studie real sessions manually labeled into hierarchical tasks, and showing that timeouts, whatever their length, are of limited utility in identifying task boundaries, achieving a maximum precision of only 70%. They report on properties of this search task hierarchy, as seen in a random sample of user interactions from a major web search engine's log, annotated by human editors, learning that 17% of tasks are interleaved, and 20% are hierarchically organized. No previous work has analyzed or addressed automatic identification of interleaved and hierarchically organized search tasks. They proposes and evaluated a method for the automated segmentation of users' query streams into hierarchical units.

III. PROPOSED SYSTEM

Considering pros and cons of the existing approaches of inferring user search goals new method is required for finding out user's information need. Therefore, a new algorithm for inferring user search goals with the feedback sessions is effective in finding out user search goals. There are four modules of proposed system.

A. Capturing feedback sessions
B. Building pseudo-documents
C. Clustering pseudo-documents
D. Restructuring web search results

A. Capturing feedback sessions

A session for web search is a series of queries to fulfil a user's information need and some clicked search results. In proposed system main focus is on inferring user search goals for a particular query. So, the single session containing only one query is introduced. This distinguishes from the

conventional session. Also the feedback session in the system is based on a single session, although it can be extended to the whole session.

Feedback session consists of both clicked and unclicked URLs. This session ends with the last URL that was clicked in a single session. It is assumed that before the last click, all the URLs have been scanned and evaluated by users and along with the clicked URLs, the unclicked URLs before the last click are made a part of the user feedbacks. Feedback sessions are constructed with the click through logs. Each feedback session can tell what a user requires and what is not. It is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly for inferring user search goals.

B. Building pseudo-documents

Feedback sessions vary a lot for different click-through and queries. So, it is not recommended to directly use feedback sessions for inferring user search goals. In order to represent these feedback sessions some representation method is needed. This method should be a more efficient and coherent.

Feedback sessions can be represented in many ways. Binary vector method is one of them. For example if user searches "the sun" then "0" is used to represent the clicked URLs and "1" to represent the unclicked one. Binary vector formed can be like [0110001]. This is nothing but a representation of feedback session. But this is not that informative to understand the contents of the user search goals. Therefore, new methods are required to represent feedback session.

Proposed system has this new method "Pseudo-documents". These can be used to infer user search goals. The building of a pseudo-document includes two steps.

1. Representing the URLs in the feedback session:
In this step, titles and snippets of the returned URLs appearing in the feedback sessions are extracted and the URLs are enriched with this additional textual contents. In simple words, in a feedback session each and every URL is represented by a small text paragraph. This paragraph consists of its title and snippet. It is followd by some textual processes. These processed includes stemming and removing stop words and transforming all the letters to lowercases. At last Term Frequency-Inverse Document Frequency (TF-IDF) vector is used to represent each URL's title and snippet.

2. Forming pseudo-document based on URL representations:
In order to obtain the feature representation of a feedback session new system has an optimization method to combine both clicked and unclicked URLs in the feedback session. Let Ffs be the feature representation of a feedback session. Fucm be the feature representations of the clicked URLs and Fucl be the feature representations of the unclicked URLs. Then the pseudo-document documents are consturcted in such a way that the distances between Ffs and each Fucm is minimized and the sum of the distances between Ffs and each Fucl is maximized.

C. Clustering pseudo-documents

With the proposed pseudo-documents, system can infer user search goals. Each feedback session is represented by a pseudo-document and let Ffs be the feature representation of the pseudo-document. The similarity between two pseudo-documents is computed as the cosine score of Ffsi and Ffsj is

$$Sim(i,j) = cos(Ffsi, Ffsj)$$

and the distance between two feedback sessions is

$$Dis(i,j) = 1 - Sim(i,j)$$

where, i and j are two pseudo documents.

Clustering of pseudo-documents is done by K-means clustering which is simple and effective. Since the exact number of user search goals is not known for each query, K is set to the five different values (i.e., 1; 2; . . . ; 5) and clustering is done based on these five values. After clustering of all the pseudo-documents, each cluster is considered as one user search goal.

D. Restructuring web search results

Search engines returns millions of results. So, 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. Vectors are used to represent inferred user search goals. Each URL's feature representation is calculated and we can categorize each URL into cluster. This is done with the help of URL vector and user search goal vector. By choosing smallest distance between URL vector and user search goal vectors URL is categorized into a cluster and the user search goals are restructured. Evaluation criteria is Average precision (AP) and it evaluates according to user implicit feedbacks. It is computed at the point of each relevant document in ranked sequence.

IV. CONCLUSION

A novel approach has been proposed to infer user search goals. Feedback sessions are clustered and in order to make clustering effective, feedback sessions are represented by pseudo-documents. First to infer user search goals feedback sessions are considered to be analyzed rather than search results or clicked URLs. All the URLs are scanned and evaluated by users and along with the clicked URLs, the unclicked URLs before the last click are made a part of the user feedbacks. So these sessions can reflect user search goals more effectively. Second, feedback sessions are represented in the form of pseudo-documents. Pseudo-documents are mapped to feedback sessions to approximate goal text in user minds. Pseudo-documents has the URLs with extra text including titles and snippets. Based on these documents user search goals are discovered and denoted with some keywords. Finally performance of user search goals is evaluated. With this new

approach users can efficiently find what they want and satisfy their information need.

REFERENCES

[1] 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, MARCH 2013

[2] Uichin Lee, Zhenyu Liu, Junghoo Cho, "Automatic Identification of User Goals in Web Search" , Proc. 14th Int'l Conf. World Wide Web (WWW '05),pp. 391-400, 2005.

[3] Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen, "Building Bridges for Web Query Classification", Proc. 29th Ann. Int'l ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR '06),pp. 131-138, 2006.

[4] Xuanhui Wang and ChengXiang 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.

[5] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen 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.

[6] Rosie Jones and Kristina Lisa 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.