International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

A novel approach for infer user search goal for query by clustering feedback session represented by psuedo documents

Chethan R A
P.G Scholor,
Department of Computer Science
Rajeev Institute of Technology,
Hassan,Karnataka
suryachethan@yahoo.com

Dr. Vasanth G
Professor and Head,
Department of Computer science
Govt.Engineering College,K.R.Pet,
Karnataka
gvasanth_ss@yahoo.co.in

Venkatesh Prasad B S
Assistant Professor,
Department of Computer science
Govt.Engineering College,K.R.Pet,
Karnataka
venkyp25@gmail.com

***Abstract* - There have been recent interests in studying the "goal" behind a user's Web query, so that this goal can be used to improve the quality of a search engine's results. For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, a novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. 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 unclicked 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 pseudo documents 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 inference.**

I. INTRODUCTION

In web search based applications user enters the query in the website to search the efficient information. The needs of the information may differ from each user and goal to achieve the user need are still becomes difficult. Because the user given queries may not understandable by system or it becomes less sometimes queries may not exactly represented by users. To achieve the user specific information needs many uncertain queries may cover a broad topic and dissimilar users may want to get information on different point of view when they submit the same query. User information need is to desire and obtain the information to satisfy the needs of each user. To satisfy the user information needs by considering the search goals with user given query. We cluster the user information needs with different search goal .Because the interference and examination of user search goals with query might have a numeral of advantages by improving the search engine significance and user knowledge. So it is necessary to collect the different user goal and retrieve the efficient information on different aspects of a query. Capture different user search goals in information retrieval outcome becomes changes than the normal query based information retrieval. In this paper, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. 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 unclicked 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. We also demonstrate that the proposed evaluation criterion can help us to optimize the parameter in the clustering method when inferring user search goals.

II. LITERATURE REVIEW

The problem of clustering investigate results has been investigate in a numeral of previous works. All of the previous work apply clustering algorithms which first group documents into similar groups according to content similarity, and produce expressive summary for clusters. Though, these summaries are often illegible which construct it difficult for Web users to recognize relevant cluster.

R. Baeza-Yates, C. Hurtado, and M. Mendoza et.al [1] propose a method that, given a query submitted to a search engine, suggests a list of related queries. The related queries are based in previously issued queries, and can be issued by the user to the search engine to tune or redirect the search process. The method proposed is based on a query clustering process in which groups of semantically similar queries are identified. The clustering process uses the content of historical preferences of users registered in the query log of the search engine. The method not only discovers the related queries, but also ranks them according to a relevance criterion.

S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder et.al[2] defined Topical classification of web queries has drawn recent interest because of the promise it offers in improving retrieval effectiveness and efficiency. However, much of this promise depends on whether classification is performed before or after the query is used to retrieve documents. They examine two previously unaddressed issues in query classification: pre vs. post-retrieval classification effectiveness and the effect of training explicitly from classified queries vs. bridging a classifier trained using a document taxonomy. Bridging classifiers map the categories of document taxonomy onto those of a query classification problem to provide sufficient training data. They find that training classifiers explicitly from manually classified queries outperforms the bridged classifier by 48 % in F1 score. Also, a pre-retrieval classifier using only the query terms performs merely 11 % worse than the bridged classifier which requires snippets from retrieved documents.

H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li et.al.[3] proposed Query suggestion plays an important role in improving the usability of search engines. Although some recently proposed methods can make meaningful query suggestions by mining query patterns from search logs, none of them are context-aware -- they do not take into account the immediately preceding queries as context in query suggestion.

In this project, they propose a novel context-aware query suggestion approach which is in two steps. In the offline model-learning step, to address data sparseness, queries are summarized into concepts by clustering a click-through bipartite. Then, from session data a concept sequence suffix tree is constructed as the query suggestion model. In the online query suggestion step, a user's search context is captured by mapping the query sequence submitted by the user to a sequence of concepts. By looking up the context in the concept sequence suffix tree, our approach suggests queries to the user in a context-aware manner.

H. Chen and S. Dumais et.al. [4] proposed and developed a user interface that organizes Web search results into hierarchical categories. Text classification algorithms were used to automatically classify arbitrary search results into an existing category structure on-the-fly. A user study compared our new category interface with the typical ranked list interface of search results. The study showed that the category interface is superior both in objective and subjective measures. Subjects liked the category interface much better than the list interface, and they were 50% faster at finding information that was organized into categories. Organizing search results allows users to focus on items in categories of interest rather than having to browse through all the results sequentially.

Zamir and Etzioni [15] introduced a Suffix Tree Clustering (STC) which first identifies sets of documents that split common phrases, and after that create clusters according to these phrases. Web search engines challenge to satisfy users' information needs by standing web pages with reverence to queries. But the realism of web search is that it is frequently a procedure of querying, learning, and reformulating. A sequence of interactions among user and search engine can be essential to satisfy a solitary information need [16]

C.-K Huang, L.-F Chien, and Y.-J Oyang et.al[5] proposes an effective term suggestion approach to interactive Web search. Conventional approaches to making term suggestions involve extracting co-occurring key terms from highly ranked retrieved documents. Such approaches must deal with term extraction difficulties and interference from irrelevant documents, and, more importantly, have difficulty extracting terms that are conceptually related but do not frequently co-occur in documents. they present a new, effective log-based approach to relevant term extraction and term suggestion. Using this approach, the relevant terms suggested for a user query are those that co-occur in similar query sessions from search engine logs, rather than in the retrieved documents. In addition, the suggested terms in each interactive search step can be organized according to its relevance to the entire query session, rather than to the most recent single query as in conventional approaches.

Lee et al. [10] consider user goals as "Navigational" and informational" and categorize queries into these two classes. Li et al. [11] define query intents as "Product intent" and "Job intent" and they try to classify queries according to the defined intents.

T. Joachims et.al [6]explore and evaluate strategies for how to mechanically produce training example for learning retrieval functions from experiential user behavior. Yet, implicit feedback is more hard to interpret and potentially noisy and did lots of works on how to use implicit feedback to get better the retrieval quality [7], [8].

Jones and Klinkner [9] predict goal and mission boundaries to hierarchically segment query logs. However, their method only identifies whether a pair of queries belong to the same goal or mission and does not care what the goal is in detail.

Hua-Jun Zeng et.al [14] suggested a query based search results for user goal and the rank list of documents return by a certain Web search engine, it first extracts and ranks most important phrases as candidate cluster names, base on a regression model learned beginning human labeled training data.

III. METHODOLOGY

In this paper first the section are majorly divided into two parts user query based information are extracted, user search goals are conditional by clustering these pseudo-documents and depicted with some keywords and then the original query based information are extracted from the web pages from that restructure the web pages based on user profile Then, we evaluate the performance of restructuring search results by evaluation criterion Average Precision, Voted Average Precision and Classified Average Precision. In first step of the process is the collection of the web pages with similar query .For example when the user Given query as sun then collect all the log files that related to the web pages based on query with link pages clicked by user .Before that copy all the links and copy the contents from the link that contains information about the link pages .After these process finished then only we Map the feedback session of the each use

A. Feedback Sessions

The inferring user search goals for a particular query. Therefore, the single session containing only one query is introduced, which distinguishes from the conventional session. 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. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user feedbacks. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Consequently, for inferring user search goals, it is additional efficient to examine the feedback sessions than to examine the investigate consequences or clicked URLs in a straight line. To represent the feedback session efficiently some demonstration methods needed, because each and every user based search goal feedback sessions are differs and their corresponding log files also changed. Represent a feedback session to Pseudo-Documents with Binary vector technique to characterize a feedback session search consequences are the URLs return by the search engine when the question "the sun" is submits, and "0" represent "unclicked" in the click sequence. The binary vector [0110001] can be second-hand to symbolize the feedback session, where "1" represent "clicked" and "0" represents "unclicked.

B. Pseudo-documents

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. To obtain the feature representation of a feedback session, we propose an optimization method to combine both clicked and unclicked URLs in the feedback session. Finally, every URL's title and snippet are generated by a Term Frequency-Inverse Document Frequency(TF-IDF) vector, correspondingly

$$T_{u_i} = \{T_{W_1}, T_{W_2}, \dots\dots, T_{W_n}\}^T \longrightarrow (1)$$
$$S_{u_i} = \{S_{W_1}, S_{W_2}, \dots\dots, S_{W_n}\}^T \longrightarrow (2)$$

Where

T_{u_i}- TF-IDF vectors of the URL's title
S_{u_i} - are the TF-IDF vectors of the URL's snippet .
u_i- i^{th} URL in the feedback session.
W_j={1; 2; . . . ; n} -j^{th} term appear in the enriched URLs. Each term in the URL is defined as a word or a numeral in the vocabulary of document collections. t_{wj} and s_{wj} characterize the TF-IDF significance of the jth term in the URL's title and snippet, correspondingly. Taking into consideration that URLs' titles and snippets have dissimilar significances, we symbolize the enriched URL by the weighted sum of T_{ui} and S_{ui} , namely,

$$F_{u_i} = T_{u_i}\omega_t + S_{u_i}\omega_s = \{f_{w_1}, f_{w_2}, \dots\dots, f_{w_n}\}^T \rightarrow (3)$$

Where F_{ui} means the feature representation of the i^{th} URL in the feedback session, and weights of the ω_t titles and $\omega_s$the snippets, respectively. In order to obtain the feature demonstration of a feedback session, suggest an optimization method to merge both clicked and unclicked URLs in the feedback session. Attain such a F_{f_s} with the purpose of the calculation of the distance between$F_{f_s}$and each F_{uc_m} is minimize and the sum of the distance between F_{f_s} and each F_{uc_l} is maximize. Based on the supposition that the terms in the vectors are self-governing, perform optimization on each dimension separately,

$$F_{f_s} = \left[f_{f_s}(\omega_1), \dots\dots f_{f_s}(\omega_n)\right)]^T \longrightarrow (4)$$

Infer user search goals and represent them with a number of significant keywords. Then the similarity between the pseudocuments is evaluated as the cosine similarity score

$$Sim_{i,j} = \cos\left(f_{fs_i}, f_{fs_j}\right) = \frac{f_{fs_i} f_{fs_j}}{|f_{fs_i}||f_{fs_j}|} \longrightarrow (5)$$

$$Dis_{i,j} = 1 - Sim_{i,j} \longrightarrow (6)$$

C. Inferring Pseudo-documents

The proposed pseudo-documents, we can infer user search goals. In this section, we will describe how to infer user search goals and depict them with some meaningful keywords. As each feedback session is represented by a pseudo-document and the feature representation of the pseudo-document. 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 and perform clustering based on these five values, respectively. 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.

In this investigate we cluster pseudo-documents by K-means clustering which is straightforward and efficient. Because we not recognizable with the precise figure of user search goal for every query, we position K to be five different values.

$$F_{center_i} = \frac{\sum_{k=1}^{C_i} F_{fs_k}}{C_i}, \left(F_{fs_k} \subset Cluster\ i\right) \longrightarrow (7)$$

Where F_{center_i} - ith cluster's center and Ci is the numeral of the pseudo-documents in the ith cluster. F_{center_i}e to finish the investigate goal of the ith cluster. Finally, the conditions with the highest values in the ar F_{center_i}d-hand as the keywords to represent user search goals, it is a keyword based explanation is that the extracted keywords be able to in addition be utilized to form a more significant query in query suggestion and thus can represent user information needs most effectively.

5. Evaluation of Search Result

If user search goals are inferred properly, the search results can also be restructured properly, since restructuring web search results is one application of inferring user search goals. Therefore, we propose an evaluation method based on restructuring web search results to evaluate whether user search goals are inferred properly or not. In this section, we propose this novel criterion "Classified Average Precision" to evaluate the restructure results. Based on the proposed criterion, we also describe the method to select the best cluster number. Restructuring web search results is an application of inferring user search goals. 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. The inferred user search goals are represented by the vectors in (7) and the feature representation of each URL in the search results can be computed by (1) and (2). Then, we can categorize each URL into a cluster centered by the inferred search goals.

IV. EVALUATION CRITERIA

In this paper the major part is the evaluation of the results from the experiments with classification results from each user search goal inference us major problem , since user search goals are not predetermined and there is no ground truth. It is necessary to develop a metric to evaluate the performance of user search goal inference objectively. In this section finally the accessible pseudo documents based clustering Measure the performance of the system with parameters like Classified Average Precision (CAP), Voted AP (VAP) which is the AP of the class including more clicks namely, risk to avoid classifying search results and average precision (AP).

.

A. Average precision (AP)

In order to be appropriate the assessment method to large scale data, the solitary sessions in user click-through logs are second-hand to reduce physical work. Since beginning user click-through logs, we can get implied significance feedbacks, specifically "clicked" means applicable and "unclicked" means inappropriate. A probable evaluation principle is the average precision (AP) which evaluate according to user implicit feedbacks. AP is the average of precisions compute at the position of each applicable document in the ranked sequence

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

where N^+ is the numeral of applicable (or clicked) documents in the retrieved ones, r is the rank, N is the total numeral 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 or less.

B. Voted AP(VAP)

VAP of the modernized search result the AP of class 1, It is defined as ,

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

where N is the total numeral of retrieved documents with class label one , rel() is a binary function on the relevance of a given rank, and Rr is the number of relevant retrieved documents of rank r or less.

C. Classified Average Precision (CAP)

Extend VAP by introducing the above Risk and propose a new criterion Classified AP(CAP)

$$\mathrm{CAP} = \mathrm{VAP} * (1 - \mathrm{risk})^{\gamma}$$

Where γ is used to adjust the influence of Risk on CAP. CAP select the AP of the class with the aim of user is interested with the most clicks/votes and takes the risk of wrong classification into account.

D.Risk

VAP is still an unsatisfactory criterion. Taking into consideration an extreme case, if every URL in the click session is categorized into one class, VAP will forever be the highest value that is 1 no matter whether user contain so many investigate goals or not.

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

Consequently present be supposed to be a risk to avoid classify exploration results into too many classes by error. They propose the risk as the above.

IV. CONCLUSION

The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer user search goals by analysing search engine query logs. A framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. This approach to generate pseudo-documents to better represent the feedback sessions for clustering. A new criterion "Classified Average Precision (CAP)" is to evaluate the performance of inferring user search goals. Results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

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. We describe a framework for understanding the underlying goals of user searches, and our experience in using the framework to manually classify queries from a web search engine. Our analysis suggests that so-called "navigational" searches are less prevalent than seeking" goal may account for a large fraction of web searches. We also illustrate how this knowledge of user search goals might be used to improve future web search engines. 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. Therefore, besides the clicked URLs, the unclicked.

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]. 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

[3]. 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.

[4]. 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.

[5]. 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.

[6]. 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.

[7]. 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.

[8]. 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.

[9]. 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.

[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.

[11]. 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.

[12]. 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.

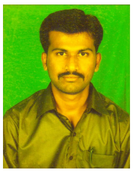
[13]. H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004

[14]. Zamir O., Etzioni O. Grouper: A Dynamic Clustering Interface to Web Search Results. In Proceedings of the Eighth International World Wide Web Conference (WWW8), Toronto, Canada, May 1999.

[15]. A. Spink, B. J. Jansen, and H. C. Ozmultu." Use of query reformulation and relevance feedback by Excite users" Internet Research: Electronic Networking Applications and Policy, 10(4):317–328, 2000

About the authors:

Dr. Vasanth G received the B.E degree in Computer Science from Gulbarga University Karnataka in 1995, also Master Degree in Computer Science and Engineering from Bangalore University in 2000,and the PH.D degree in Computer Networks from the Magadh University Bihar in 2008. He is a professor and Head of the Department of Computer Science at Government Engineering College Mandya. He has published 7 International Journals and presented 25 papers in international Conferences and 18 National Conference papers. He has also publishedone IEEE journal. He is having 19 years of teaching experience and also one of the recognized guide for Visvesvaraya Technological University. His main research intersts include computer networks,wireless adhoc networks,storage area networks and Data Mining.



Chethan R A received the B.E degree in Computer Science and Engineering from Visvesvaraya Technological University Karnataka in 2011.currently he is a post graduate student pursuing M.Tech in Computer Science and Engg in Rajeev Institute of Technology under Visvesvaraya Technological University Karnataka. His main research interest include Data Mining, Computer Networks, Wireless Sensor Networks and storage Networks.

Venkatesh Prasad B S received the B.E degree in Computer Science from Bangalore University, Karnataka in 2001, also Master Degree in Information Technology from Bangalore University in 2006. He is a Assistant professor in the Department of Computer Science at Government Engineering College mandya. His main research interests include computer networks, wireless sensor networks and Data Mining