An Efficient Approach for Retrieving Personalized User Goals

Deepa M, D Ravi, T K P Rajagopal

Abstract-- **The most important facility of a search engine is to retrieve relevant information as early as possible. For this understanding user search goal is necessary. For the same query the user information need may be different. In this paper we propose a method to infer user search goal by analyzing search engine query logs. User search goals are discovered by clustering proposed feedback sessions. Feedback sessions are constructed from user click-through logs and it reflect information need of users. Then propose novel approach to generate pseudo-documents which represent feedback session for clustering. We present a large-scale evaluation framework for personalized search based on query logs and then evaluate five personalized search algorithms (including two click-based ones and three topical-interest-based ones). It represents a significant improvement over generic Web search for some queries, while it has little effect and even harms query performance under some situations.**

Index Terms--**click-through data, personalization, feedback sessions**

I. INTRODUCTION

Searching is a process of querying, learning and reformulating. Search engines satisfy user information need through a series of interactions with the user and rank the results based on user feedback. For getting an exact result for an ambiguous query, different users may have different search goals when submit to the search engine. Some cases user information need cannot be represented because most ambiguous queries have different search categories. So identify the search category is an important task. Restructure the web search results according to user search goal by grouping the search results with the same search goal. Based on the user feedback his interested information is ranked on the top of the web search result.

II. EXISTING SYSTEM

One criticism of search engines is that when queries are issued, most return the same results to users. In fact, the vast majority of queries to search engines are short and ambiguous.

Different users may have completely different information needs and goals when using precisely the same query. For example, a biologist may query "mouse" to get information about rodents, while programmers may use the same query to find information about computer peripherals. It takes time for a user to choose which information he/she wants.

On another query of "free mp3 download," although most users find websites to download free mp3s, their selections can diverge: one may choose the website www.yourmp3.net, while another may prefer the website www.seekasong.com.

Most commercial search engines return roughly the same results for the same query, regardless of the user's real interest. Since queries submitted to search engines tend to be short and ambiguous, they are not likely to be able to express the user's precise needs. For example, a farmer may use the query "apple" to find information about growing delicious apples, while graphic designers may use the same query to find information about Apple Computer.

Most existing user profiling strategies only consider documents that users are interested in (i.e., users' positive preferences) but ignore documents that user's dislike (i.e., users' negative preferences). In reality, positive preferences are not enough to capture the fine grain interests of a user. For example, if a user is interested in "apple" as a fruit, he/she may be interested specifically in apple recipes, but less interested in information about growing apples, while absolutely not interested in information about the company Apple Computer. In this case, a good user profile should favor information about apple recipes, slightly favor information about growing apple, while downgrade information about Apple Computer. Profiles built on both positive and negative user preferences can represent user interests at finer details.

III. RELATED WORK

A. Understanding User Goals in Web Search

Previous work on understanding user web search behavior [1], has focused on how people search and what they are searching for, but not why they are searching. 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 generally believed, while a previously unexplored "resource seeking" goal may account for a large fraction of web searches. We also illustrate how this

ISSN: 2278 – 1323

knowledge of user search goals might be used to improve future web search engines.

B. Learning to Cluster Web Search Results

Organizing Web search results into clusters [2], facilitates users' quick browsing through search results. Traditional clustering techniques are inadequate since they don't generate clusters with highly readable names. In this paper, we reformalize the clustering problem as a salient phrase ranking problem. Given a query and the ranked list of documents (typically a list of titles and snippets) returned by a certain Web search engine, our method first extracts and ranks salient phrases as candidate cluster names, based on a regression model learned from human labeled training data. The documents are assigned to relevant salient phrases to form candidate clusters, and the final clusters are generated by merging these candidate clusters. Experimental results verify our method's feasibility and effectiveness.

C. Learn from Web Search Logs to Organize Search Results

Effective organization of search results [3], is critical for improving the utility of any search engine. Clustering search results is an effective way to organize search results, which allows a user to navigate into relevant documents quickly. However, two deficiencies of this approach make it not always work well: (1) the clusters discovered do not necessarily correspond to the interesting aspects of a topic from the user's perspective; and (2) the cluster labels generated are not informative enough to allow a user to identify the right cluster. In this paper, we propose to address these two deficiencies by (1) learning \interesting aspects" of a topic from Web search logs and organizing search results accordingly; and (2) generating more meaningful cluster labels using past query words entered by users. We evaluate our proposed method on a commercial search engine log data. Compared with the traditional methods of clustering search results, our method can give better result organization and more meaningful labels.

D. Context Sensitive Information Retrieval Using Implicit Feedback

A major limitation of most existing retrieval models and systems is that the retrieval decision is made based solely on the query and document collection; information about the actual user and search context is largely ignored. In this paper, we study how to exploit implicit feedback information, including previous queries and click-through information, to improve retrieval accuracy in an interactive information retrieval setting. We propose several context sensitive retrieval algorithms based on statistical language models to combine the preceding queries and clicked document summaries with the current query for better ranking of documents. We use the TREC AP data [4],to create a test collection with search context information, and quantitatively evaluate our models using this test set. Experiment results show that using implicit feedback, especially the clicked document summaries, can improve retrieval performance substantially.

E. Beyond the Session Timeout: Automatic HierarchicalSegmentation of Search Topics in Query Logs

Most analysis of web search relevance and performance takes a single query as the unit of search engine [5], interaction. When studies attempt to group queries together by task or session, a timeout is typically used to identify the boundary. However, users query search engines in order to accomplish tasks at a variety of granularities, issuing multiple queries as they attempt to accomplish tasks. In this work we study real sessions manually labeled into hierarchical tasks, and show that timeouts, whatever their length, are of limited utility in identifying task boundaries, achieving a maximum precision of only 70%. We 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. We propose and evaluate a method for the automated segmentation of users' query streams into hierarchical units.

IV. PROPOSED SYSTEM

Although personalized search has been under way for many years and many personalization algorithms have been investigated, it is still unclear whether personalization is consistently effective on different queries for different users and under different search contexts. We present a large-scale evaluation framework for personalized search based on query logs and then evaluate five personalized search algorithms (including two click-based ones and three topical-interest-based ones. It represents a significant improvement over generic Web search for some queries, while it has little effect and even harms query performance under some situations.

Underlying idea of our proposed technique is based on concepts and their relations extracted from the submitted user queries, the web-snippets1 and the click-through data. Click through data was exploited in the personalized clustering process to identify user preferences: a user clicks on a search result mainly because the web-snippet contains a relevant topic which the user is interested in. Moreover, click through data can be collected easily without imposing extra burden on users, and thus providing a low-cost means to capture user's interest.

We proposing and studying seven concept-based user profiling strategies that are capable of deriving both of the user's positive and negative preferences. All of the users profiling strategies are query-oriented, meaning that a profile is created for each of the user's queries. The user profiling strategies are evaluated and compared with our previously proposed personalized query clustering method.

We extend the query-oriented, concept-based user profiling method proposed in to consider both users' positive and negative preferences in building users profiles. We

ISSN: 2278 – 1323

proposed six user profiling methods that exploit a user's positive and negative preferences to produce a profile for the user using a Ranking SVM (RSVM). Our proposed methods use an RSVM to learn from concept preferences weighted concept vectors representing concept-based user profiles. The weights of the vector elements, which could be positive or negative, represent the interestingness (or UN interestingness) of the user on the concepts. In the weights that represent a user's interests is all positive, meaning that the method can only capture user's positive preferences.

V. METHODOLOGY ADOPTED

A. Content Based Image Retrieval

Content based Image Retrieval Systems introduces the CBIR technique that helps to organize the digital image archives by their visual content. From this statement, anything ranging from image similarity function to a robust image annotation engine falls under the preview of CBIR. CBIR uses visual content called the features to search images from the large scale image databases according to the user request in the form of the query image. An algorithm called edge histogram is used to extract the features of the images. The content based retrieval can be of color, shape, texture. The fuzzy inference integrates the perfect feature of the image in the content based retrieval. CBIR is the most used technical aspect to be used in retrieval of image from the web services.

B. Image Retrieval Systems

Image Retrieval Systems Techniques, Promising Issues and other Trends focuses mainly on image retrieval for both text and images. Most image retrieval systems support random browsing , search and navigation of the images with the custom searches. QBIC standing for the query by the image content, is the first commercial content based image retrieval systems. Its system framework and techniques have profound effects on later image retrieval systems. Virage is a content-based image search engine developed at virage Inc. Virage supports visual queries based on the color, composition, texture, structure. It also supports arbitrary combinations of the above mentioned atomic queries.

C. Fuzzy Keyword Search

The Efficient Interactive Fuzzy keyword search uses traditional information systems that return answers after a user submits a complete query. Users often feel left in dark when they have limited knowledge about the underlying data, and have to use a try and see approach for finding information. A recent trend in supporting auto complete in these systems is the first step to solve the problems. The auto complete interfaces are extended by allowing the keywords to appear in multi attributes and finding relevance to matching the keywords.

D. Feature Weighted

The basic idea behind from the image attributes. The coefficient of the attributes for each image using correlation technique is calculated. The formula to calculate the correlation of the attributes in the image is: R(y,12…k)= √1-(1-ry1)^2.....(1-ryk...2(k-1)). Here ry1, ry2... ryk..(k-1) are the coefficient of single correlation. From correlation, the weights of the image based on the features are calculated.

E. Similarity Measurement

The matching procedure is found by means of similarity in the query image and the images in the database. The similar attributes are clustered together and the features are analyzed based on the need of the user.

ALGORITHM

1. A new cluster is created by specifying threshold.
2. Create_ new_ cluster=true
3. Assign d (i) to c(j)
4. For i form 2 to n
   a) A. For all clusters
      i. Calculate distance (d (i), c (j))
      ii. If (distance (d (i), c (j) <= threshold)
      And
   b) Create_new_cluster=false
   c) If (Create_new_cluster=True) then create a cluster k= k+1, c (k) =d (i)
5. Run the new cluster
6. End

VI. ARCHITECTURAL DIAGRAM

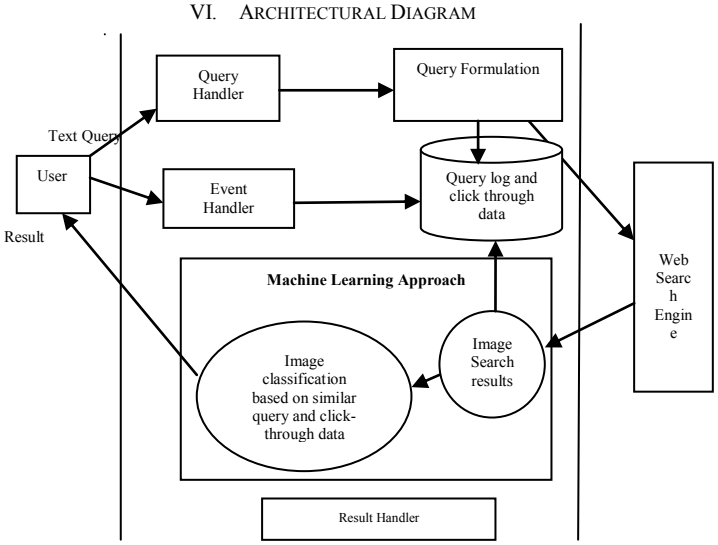

Fig 1: Architectural Diagram of Image Retrieval System

VII. CONCLUSION

The techniques make use of click through data to extract from Web-snippets to build concept-based user profiles automatically. We applied preference mining rules to infer not only users' positive preferences but also their negative preferences, and utilized both kinds of preferences in deriving

ISSN: 2278 – 1323

user's profiles. The user profiling strategies were evaluated and compared with the personalized query clustering method that we proposed previously. Apart from improving the quality of the resulting clusters, the negative preferences in the proposed user profiles also help to separate similar and dissimilar queries into distant clusters, which help to determine near optimal terminating points for our clustering algorithm.

ACKNOWLEDGEMENT

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this M.E project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

I express my warm thanks to the Head of Department Dr. Subadra madam, and Rajagopal sir for their support and guidance. I express my gratitude to my project guide Ravi sir and all the people who provide me with the facilities being required and conductive conditions for my project.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

AUTHORS BIOGRAPHY

Deepa M: She Completed UG from S H M Engineering College,Kadakkal Kollam under Kerala University, India. She is currently doing PG in computer Science and Engineering from Kathir College of Engineering, Coimbatore, India. She is an active member of IEEE student branch. Attended various International, National Conferences and Attended various work shop based on Image processing and Data mining and participated on national workshop on android application development. Her research interests are Image processing, Data mining**.**

Ravi.D: He received the MCA degree a from the IGNOU, New Delhi in the year 2003, the M.Phil from the Bhrathidasan University, Trichy in the 2005. the M.E (CSE) in Anna University of Technology, Coimbatore in the year 2011. He has 11 years of Teaching Experience. His area of Interest is in Web Mining and Personalization. He published 04 National Conference and 04 International Conference. He is a member of ISTE,IACSIT and IAENG.

T.K.P. Rajagopal M.A., M.Phil., M.C.A., M.Phil., M.E.(CSE), M.Tech.(IT), Ph.D'., is working as an Associate Professor in the Department of Computer Science and Engineering at Kathir College of Engineering, Coimbatore, Tamailnadu, India. He has 15 years of teaching experience. His research areas are Network Security, Data Mining and Digital Image Processing. He has published 3 books, 9 research papers in international journals and 2 research papers in national journals, Presented 17 papers in International Conference, 19 papers in National level conferences and has attended 3 QIP, 5 FTP, 10 Seminars and 33 Workshops. He is a life member of various professional societies like ISTE, DOEACC, CSTA, CSI, ACS, IAS, IACSIT, IIRJC, SDIWC, ISI, IoN, UACEE, theIRED, IAENG Society of Artificial Intelligence, Computer Science, Data Mining, Software Engineering, and Wireless Networks, Hong Kong, Serving as a reviewer in IJITCE, WCSIT, CSC-IJCSS, CSC-IJCN, IJCSIS, Scientific & Technical Committee & Editorial Review Board on Engineering and Natural Sciences in WASET.

ISSN: 2278 – 1323