International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 1 Issue 1 –AUGUST 2013

Clustering User Search Queries Using Textual Similarity

P. Siva Prasad[#1], Chekka Raju[*2]

[#]*PG Scholar, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

[1] *prasadsiva_17@yahoo.com*

[*]*Assistant Professor, Dept. of CSE, VRS & YRN College of Engg. & Technology, Chirala, AP*

***Abstract*— Internet users are using the complex task oriented works on the web is goes on increasing in recent years. The complex tasks are booking for travels, managing finances and planning for purchase the new things. The successful previous work, they usually break down the tasks into a few co-dependent steps and issue multiple queries around these steps repeatedly over long periods of time. For the better support, user s in the long term information quests on the web, the search engine keep track on their queries and clicks while searching in the online. In this paper, we examine the issues in the histories queries on the search engine. In order to overcome the issues, we introduce this paper. In this we propose an organizing user search historical queries into groups in a dynamic automated mode. By using this mode, the queries are searched by automatically. So, automatically identifying queries groups is very useful for a number of different search components and applications for the users. The propose scheme automatically identifies a query suggestions, result ranking, query alterations, sessions and collaborative search. In our approach deals with that rely on textual similarity or time thresholds. We also propose that leverages search query logs. We experimentally study the performance of different techniques, and showcase their potential, especially when combined together**

***Keywords*— User history, search history, query clustering, query reformulation, click graph, task identification, time thresholds.**

I. INTRODUCTION

In recent years, personalization of the search engines is in active research field and the user profile edifice is an important property of any personalization system. To express the customization has been widely used to personalize the look and content of the websites is very beautiful manner. The personalized search engines approaches focus on the implicitly building and exploiting the user profiles. Many companies provide the marketing data for the search engines utilizes more and more, when it is compared to the direct navigation and web links. As search engines perform a larger role in commercial applications, the desire to increase their effectiveness grows. However, search engines are affected by problems such as ambiguity and results ordered by web site popularity rather than user interests.

Recently, the number of users and richness of information increases, the complexity of the tasks also increasing. The users on the web are no longer content with issuing the simple navigational queries. Various studies show that the 20% of the queries are navigational in nature. And the rest of the 80% queries are transactional and informational queries. This is due to the users on the web undergoes many complex tasks oriented goals such as booking for travels, managing finances and planning for purchase the new things. In this task the users need many queries to complete the tasks. Each step in the complexity tasks needs one or more steps queries and each query results in one or more clicks on the relevant pages. Because of it requires many steps to complete the task, it creates many issues.

One of the important step towards the queries on the search engine the enabling services and features helps the users during their complex on search quest online is the capability to identify and group related queries together. Recently, the some of the many search engines introduced the “Search History”. It is used to track the users for their online searches and it is recorded by queries and clicks towards a particular page or links. And these histories consist of four queries arranged in order of time of occurrence in the reverse order with their corresponding clicks. If we want to search their history, users can manipulate it by manually editing and organizing related queries and clicks into groups, or by sharing them with their friends. While these features are helpful, the manual efforts involved can be disruptive and will be untenable as the search history gets longer over time.

In this query grouping allows the search engines for the better understand a user’s session and effectively the user’s search experience according to their needs. . Once query groups have been identified, search engines can have a good representation of the search context behind the current query using queries and clicks in the corresponding query group. Query group will help to improve the qualities of the key management of search engine such as query suggestions, collaborative search, sessionization, result ranking and query alterations. It also assists to promoting the task level into collaborative search. At any instance, set of queries groups created by the expert users, we can select the ones which is highly relevant to the current user’s query activity and recommended them to their needs.

To overcome the problems in the organizing a user's search history, we propose that the set of query groups in this paper. In this each query groups is a collection of queries by the same user that are relevant to each other around a common information n needs of the users on the web. There set of queries group dynamically updated the issues of the users and new groups are created over the time. Organizing [38], the query groups within a user's history is challenging for a number of reasons. First, related queries may not appear close to one another, as a search task may span days or even weeks. This is further complicated by the interleaving of queries and clicks from different search tasks due to users' multitasking, opening multiple browser tabs, and frequently changing search topics. To achieve more effective and robust query grouping, we do not rely solely on textual or temporal properties of queries.

We make the following contributions as the main, in this paper:

First, we motivate and propose a method to perform query grouping in a dynamic fashion. Secondly, we investigate how signals from search logs such as query reformulations and clicks can be used together to determine the relevance among query groups. And finally, we show through the experimental evaluation the effectiveness and its secure level of our proposed scheme.

The rest of the paper is organized as follows. In Section II, we discuss about the related work of the paper. In Section III formally introduces our proposed scheme in the search engines. In Section IV we summarize about the algorithm. In Section V, we present the full simulation study of the proposed system. Finally, we conclude the paper and discuss future work in Section VI.

II. Related Works

In this section, we briefly discuss the works which is similar techniques as our approach but serve for different purposes.

Mirco Speretta [39], in this paper he propose User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles collect user information through proxy servers (to capture browsing histories) or desktop bots (to capture activities on a personal computer). Both these techniques require participation of the user to install the proxy server or the bot. In this study, we explore the use of a less-invasive means of gathering user information for personalized search. In particular, we build user profiles based on activity at the search site itself and study the use of these profiles to provide search engine, we were able to collect information about individual user search activities. In particular, we collected the queries for which at least one search result was examined, and the snippets (titles and summaries) for each examined result

Devang Karavadiya and , Purnima Singh [40], . In this paper, we study the problem of organizing a user's historical queries into groups in a dynamic and automated fashion. Automatically identifying query groups is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. We experimentally study the performance of different techniques, and showcase their potential, especially when combined together.

While we are not aware of any previous work that has the same objective of organizing user history into query groups, there has been prior work in determining whether two queries belong to the same search task. In recent work, Jones and Klinkner [4] and Boldi et al. [5] investigate the search-task identification problem. More specifically, Jones and Klinkner [4] considered a search session to consist of a number of tasks (missions), and each task further consists of a number of subtasks (goals). They trained a binary classifier with features based on time, text, and query logs to determine whether two queries belong to the same task. Boldi et al. [5] employed similar features to construct a query flow graph, where two queries linked by an edge were likely to be part of the same search mission.

Our work differs from these prior works in the following aspects. First, the query-log based features in [4], [5] are extracted from co-occurrence statistics of query pairs. In our work, we additionally consider query pairs having common clicked URLs and we exploit both co-occurrence and click information through a combined query fusion graph. Jones and Klinkner [4] will not be able to break ties when an incoming query is considered relevant to two existing query groups. Additionally, our approach does not involve learning and thus does not require manual labeling and retraining as more search data come in; our Markov random walk approach essentially requires maintaining an updated query fusion graph. Finally, our goal is to provide users with useful query groups on-the-fly while respecting existing query groups. On the other hand, search task identification is mostly done at server side with goals such as personalization, query suggestions [5], etc.

Some prior work also looked at the problem of how to segment a user's query streams into "sessions." In most cases, this segmentation was based on a "time-out threshold" [21], [22], [23], [24], [25], [26], [27]. Some of them, such as [23], [26], looked at the segmentation of a user's browsing activity, and not search activity. Silverstein et al. [27] proposed a time-out threshold value of 5 minutes, while others [21], [22], [24], [25] used various threshold values. As shown in Section 5, time is not a good basis for identifying query groups, as users

may be multitasking when searching online [3], thus resulting in interleaved query groups.

The notion of using text similarity to identify related queries has been proposed in prior work. He et al. [24] and Ozmutlu and C¸ avdur [28] used the overlap of terms of two queries to detect changes in the topics of the searches. Lau and Horvitz [29] studied the different refinement classes based on the keywords in queries, and attempted to predict these classes using a Bayesian classifier. Radlinski and Joachims [30] identified query sequences (called chains) by employing a classifier that combines a timeout threshold with textual similarity features of the queries, as well as the results returned by those queries. While text similarity may work in some cases, it may fail to capture cases where there is "semantic" similarity between queries (e.g., "ipod" and "apple store") but no textual similarity. In Section 5, we investigate how we can use textual similarity to complement approaches based on search logs to obtain better performance.

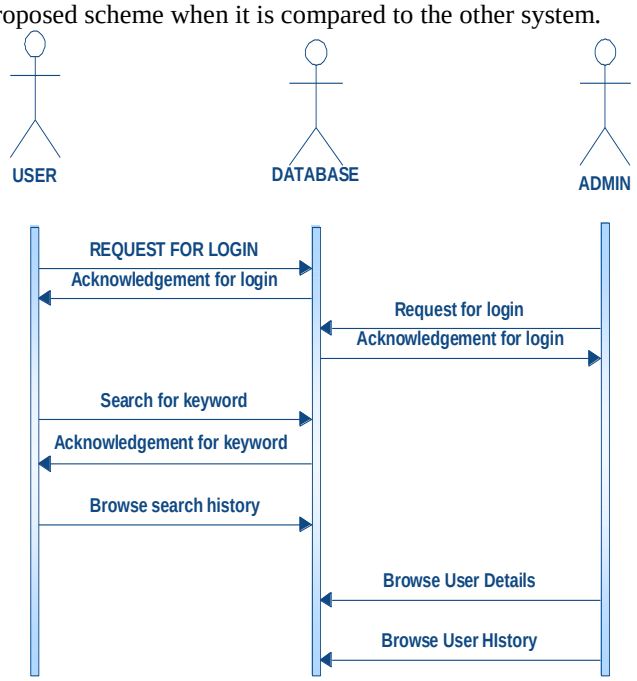
The problem of online query grouping is also related to query clustering [13], [31], [6], [7], [32]. The authors in [13] found query clusters to be used as possible questions for a FAQ feature in an Encarta reference website by relying on both text and click features. In Beeferman and Berger [6] and Baeza-Yates and Tiberi [7], commonly clicked URLs on query-click bipartite graph are used to cluster queries. The authors in [31] defined clusters as bicliques in the click graph. Unlike online query grouping, the queries to be clustered are provided in advance, and might come from many different users. The query clustering process is also a batch process that can be accomplished offline. While these prior work make use of click graphs, our approach is much richer in that we use the click graph in combination with the reformulation graph, and we also consider indirect relationships between queries connected beyond one hop in the click graph. This problem is also related to document clustering [33], [34], with the major difference being the focus on clustering queries (only a few words) as compared to clustering documents for which term distributions can be estimated well.

III. PROPOSED WORK

To overcome the problems in the organizing a user's search history, we propose that the set of query groups in this paper. In this each query groups is a collection of queries by the same user that are relevant to each other around a common information n needs of the users on the web. There set of queries group dynamically updated the issues of the users and new groups are created over the time. Organizing [38], the query groups within a user's history is challenging for a number of reasons. First, related queries may not appear close to one another, as a search task may span days or even weeks. This is further complicated by the interleaving of queries and clicks from different search tasks due to users' multitasking, opening multiple browser tabs, and frequently changing search topics. To achieve more effective and robust query grouping, we do not rely solely on textual or temporal properties of queries.

There are 3 contributions are proposed in the Organizing User Search Histories. We make the following contributions as the main, in this paper:

First contribution is to motivate and propose the method to perform query grouping in a dynamic fashion. Goal of our proposed system is to ensure the good performance while avoiding the break of existing user-defined query groups. Secondly, we study into how the signals from search logs such as query reformulations and clicks can be used to find the same query groups. So that, we examine the two potential ways of using clicks in order to enhance this process: 1) by fusing the query reformulation graph and the query click graph into a single graph that we refer to as the query fusion graph, and 2) by expanding the query set when computing relevance to also include other queries with similar clicked URLs. And finally, we show through the experimental evaluation the effectiveness and its secure level of our proposed scheme when it is compared to the other system.



IV. CONCLUSION AND FUTURE WORKS

In this paper we showed that the proposed scheme such that information can be used effectively for the task pf organizing user search histories into query groups. More specifically, we propose combining the two graphs into a

query fusion graph. We further show that our approach that is based on probabilistic random walks over the query fusion graph outperforms time-based and keyword similarity-based approaches. We also find value in combining our method with keyword similarity-based methods, especially when there is insufficient usage information about the queries. As future work, we intend to investigate the usefulness of the knowledge gained from these query groups in various applications such as providing query suggestions and biasing the ranking of search results.

V. REFERENCES

[1] j. teevan, e. adar, r. jones, and m.a.s. potts, "information re- retrieval: repeat queries in yahoo's logs," proc. 30th ann. int'l acm sigir conf. research and development in information retrieval (sigir '07), pp. 151-158, 2007.

[2] broder, "a taxonomy of web search," sigir forum, vol. 36, no. 2, pp. 3-10, 2002.

[3] spink, m. park, b.j. jansen, and j. pedersen, "multitasking during web search sessions," information processing and management, vol. 42, no. 1, pp. 264-275, 2006.

[4] 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), 2008.

[5] p. boldi, f. bonchi, c. castillo, d. donato, a. gionis, and s. vigna, "the query-flow graph: model and applications," proc. 17th acm conf. information and knowledge management (cikm), 2008.

[6] 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 (kdd), 2000.

[7] r. baeza-yates and a. tiberi, "extracting semantic relations from query logs," proc. 13th acm sigkdd int'l conf. knowledge discovery and data mining (kdd), 2007.

[8] j. han and m. kamber, data mining: concepts and techniques. morgan kaufmann, 2000.

[9] w. barbakh and c. fyfe, "online clustering algorithms," int'l j. neural systems, vol. 18, no. 3, pp. 185-194, 2008.

[10] lecture notes in data mining, m. berry, and m. browne, eds. world scientific publishing company, 2006.

[11] v.i. levenshtein, "binary codes capable of correcting deletions, insertions and reversals," soviet physics doklady, vol. 10, pp. 707-710, 1966.

[12] m. sahami and t.d. heilman, "a web-based kernel function for measuring the similarity of short text snippets," proc. the 15th int'l conf. world wide web (www '06), pp. 377-386, 2006.

[13] j.-r. wen, j.-y. nie, and h.-j. zhang, "query clustering using user logs," acm trans. in information systems, vol. 20, no. 1, pp. 59-81, 2002.

[14] fuxman, p. tsaparas, k. achan, and r. agrawal, "using the wisdom of the crowds for keyword generation," proc. the 17th int'l conf. world wide web (www '08), 2008.

[15] k. avrachenkov, n. litvak, d. nemirovsky, and n. osipova, "monte carlo methods in pagerank computation: when one iteration is sufficient," siam j. numerical analysis, vol. 45, no. 2, pp. 890-904, 2007.

[16] l. page, s. brin, r. motwani, and t. winograd, "the pagerank citation ranking: bringing order to the web," technical report, stanford univ., 1998.

[17] p. boldi, m. santini, and s. vigna, "pagerank as a function of the damping factor," proc. the 14th int'l conf. world wide web (www '05), 2005.

[18] t.h. haveliwala, "topic-sensitive pagerank," proc. the 11th int'l conf. world wide web (www '02), 2002.

[19] w.m. rand, "objective criteria for the evaluation of clustering methods," j. the am. statistical assoc., vol. 66, no. 336, pp. 846-850, 1971.

[20] d.d. wackerly,w.m. iii, and r.l. scheaffer, mathematical statistics with applications, sixth ed. duxbury advanced series, 2002.

[21] p. anick, "using terminological feedback for web search refinement: a log-based study," proc. 26th ann. int'l acm sigir conf. research and development in information retrieval, 2003.

[22] b.j. jansen, a. spink, c. blakely, and s. koshman, "defining a session on web search engines: research articles," j. the am. soc. for information science and technology, vol. 58, no. 6, pp. 862-871, 2007.

[23] l.d. catledge and j.e. pitkow, "characterizing browsing strategies in the world-wide web," computer networks and isdn systems, vol. 27, no. 6, pp. 1065-1073, 1995.

[24] d. he, a. goker, and d.j. harper, "combining evidence for automatic web session identification," information processing and management, vol. 38, no. 5, pp. 727-742, 2002.

[25] r. jones and f. diaz, "temporal profiles of queries," acm trans. information systems, vol. 25, no. 3, p. 14, 2007.

[26] a.l. montgomery and c. faloutsos, "identifying web browsing trends and patterns," computer, vol. 34, no. 7, pp. 94-95, july 2001.

[27] c. silverstein, h. marais, m. henzinger, and m. moricz, "analysis of a very large web search engine query log," sigir forum, vol. 33, no. 1, pp. 6-12, 1999.

[28]

[29] h.c. ozmutlu and f. c¸ avdur, "application of automatic topic identification on excite web search engine data logs," information processing and management, vol. 41, no. 5, pp. 1243-1262, 2005.

[30] t. lau and e. horvitz, "patterns of search: analyzing and modeling web query refinement," proc. seventh int'l conf. user modeling (um), 1999.

[31] f. radlinski and t. joachims, "query chains: learning to rank from implicit feedback," proc. acm conf. knowledge discovery and data mining (kdd), 2005.

[32] j. yi and f. maghoul, "query clustering using click-through graph," proc. the 18th int'l conf. world wide web (www '09), 2009.

[33] e. sadikov, j. madhavan, l. wang, and a. halevy, "clustering query refinements by user intent," proc. the 19th int'l conf. world wide web (www '10), 2010.

[34] t. radecki, "output ranking methodology for document- clustering-based boolean retrieval systems," proc. eighth ann. int'l acm sigir conf. research and development in information retrieval, pp. 70-76, 1985.

[35] v.r. lesser, "a modified two-level search algorithm using request clustering," report no. isr-11 to the nat'l science foundation, section 7, dept. of computer science, cornell univ., 1966.

[36] r. baeza-yates, "graphs from search engine queries," proc. 33rd conf. current trends in theory and practice of computer science (sofsem), vol. 4362, pp. 1-8, 2007.

[37] k. collins-thompson and j. callan, "query expansion using random walk models," proc. 14th acm int'l conf. information and knowledge management (cikm), 2005.

[38] n. craswell and m. szummer, "random walks on the click graph," proc. 30th ann. int'l acm sigir conf. research and development in information retrieval (sigir '07), 2007.

[39] heasoo hwang, hady w. lauw, lise getoor, and alexandros ntoulas-" organizing user search histories"- ieee transactions on knowledge and data engineering, vol. 24, no. 5, may 2012

[40] mirco speretta- b.sc. , udine university, udine, italy 2000- " personalizing search based on user search histories"

[41] devang karavadiyaa and , purnima singh-"user specific search using grouping and organization"- international journal of emerging trends & technology in computer science (ijettcs)