ISSN: 2321-7782 (Online)
Volume 2, Issue 5, May 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

An Efficient Query Processing for Personalized Web Search Engine

K. Mala[1]
IInd Year – M.E. CSE
Shivani Engineering College
Trichy, Tamil Nadu – India

M. A. Hema[2]
M.E(CSE), Assistant Professor
Shivani Engineering College
Trichy, Tamil Nadu – India

Abstract: ***Mobile based search engine the major problem is that the interaction between mobile users and search results are managed by small numeral of factors in the mobile phones. In order to manage these problem collect user query and their relevant result to satisfy the user profile according to the interest. To perform this by observing the different types of concepts in the personalized web search engine (PWSE), it captures the user preferences concepts by mining click through data. In Personalized web search engine (PWSE) preferences of each user are ordered in ontology based model and each user profiles are ranked with the use of multi-facet for future search results. The search result can be classified into location and content based concepts based on their importance information. Improve the PWSE result by investigate methods to develop normal query travel patterns from the location and clickthrough data to further enhance the personalization effectiveness of PWSE. By introducing an association rule mining algorithm collect the different travel patterns by original search engine result in each and every query of user from the original personal mobile search engine profile. Association rule learning is used for finding the interesting query travel pattern results from each user query in PWSE search engine From this query related patterns of the user to identify strong rules discovered in databases using different measures of interestingness. They introduced association rules for discovering regularities between normal patterns and query related patterns in the personalized mobile search engine result.***

Keywords: ***Association rule mining, Click-through data, concept, location search, mobile search engine, ontology, personalization and user profiling.***

I. INTRODUCTION

Search engine plays a major important way to search the applicable information from the web. Though, the investigate results acquire may not forever be cooperative to the user, as investigate engine fail to be familiar with the user purpose behind the query. Because the exacting word might denote numerous things in different contexts and the predictable background can be strong-minded by the user unaccompanied. For picture, particular a investigate keyword “apple”, a consumer might be penetrating intended for fruit apple or for apple computer. Characteristic search engines provide alike set of consequences without bearing in mind of who submit the query. Consequently, the obligation arises to have personalized web search organization which gives yield suitable to the user as extremely ranked pages to manipulate this type of issues personalized user profile based system are proposed in previous work personalized web searches have been developed.

In personalized search (ps) how to adeptly attain user’ information requirement is a key difficulty. Therefore it is tremendous to achieve user’s need simply from the user given query or keywords. In web search system the main difficulty is that doesn’t imagine concerning the difference surrounded by personality user needs. To overcome this difficult by integrating information the meta-search engine in mobile surroundings. Meta search engine with personalized helps individuals search problem to find the important information according to user's interest none of the previous work support the result based on concept and location based results. It either considering the location or concepts in single manner not produces both results

simultaneously by observe the requirements of dissimilar type of users, there personalized web search engine (pwse) which represent disparate type of concepts in disparate ontology's. It categorize the user information into both content and location based concept from user given query with personal search engine result it adopt the meta search engine approach which relies on marketable search engines such as google, yahoo to achieve a actual search. The client is responsible for receiving the user's requests, conveyance the needs to the pwse server, display the return consequences lastly collecting their click throughs in organizes to obtain their personal preferences. The PWSE server is responsible for performing the task to the main search engine and as well as rank the results according the different user and their similar query based results in the server side and return result to the client side in PWSE. To distinguish the diversity of the concepts associated with a query and their relevancies' to the user's preferences, dissimilar entropy measures are introduce to equilibrium the weights among the content and location concepts. In PWSE doesn't support to develop usual travel patterns that is search related patterns and query patterns beginning the GPS. To perform this use of association rule mining to collect query patterns from the initial stage of the search process and it further enhance click through data result in the personalization efficiency of PWSE.

II. Related Work

PWSE uses "concepts" to model the interests and preferences of a user. Since location information is important in mobile search, the concepts are further classified into two different types, namely, content concepts and location concepts. The concepts are modeled as ontologies, in order to capture the relationships between the concepts. To observe that the characteristics of the content concepts and location concepts are different. Thus, to propose two different techniques for building the content ontology and location ontology. The ontologies indicate a possible concept space arising from a user's queries, which are maintained along with the clickthrough data for future preference adaptation. In PWSE, to adopt ontologies to model the concept space because they not only can represent concepts but also capture the relationships between concepts. Due to the different characteristics of the content concepts and location concepts first discuss our method to mine and build the content ontology from the search results. It presents our method to derive location ontology from the search results.

A. CONTENT ONTOLOGY

Content concept extraction method first extracts all the keywords and phrases (excluding the stop words) from the web-snippets2 arising from q. If a keyword/phrase exists frequently in the web-snippets arising from the query q, it would treat as an important concept related to the query, as it coexists in close proximity with the query in the top documents. The following support formula, which is inspired by the well-known problem of finding frequent item sets in data mining [8], is employed to measure the importance of a particular keyword/phrase ci with respect to the query q:

$$Support(ci)=\frac{sf(ci).|ci|}{n}$$

where sf(ci) is the snippet frequency of the keyword/ phrase ci (i.e., the number of web-snippets containing ci), n is the number of web-snippets returned and |ci| is the number of terms in the keyword/phrase ci. If the support of a keyword/phrase ci is higher than the threshold s (s = 0:03in our experiments), we treat ci as a concept for q. To adopt the following two propositions to determine the relationships between concepts for ontology formulation:

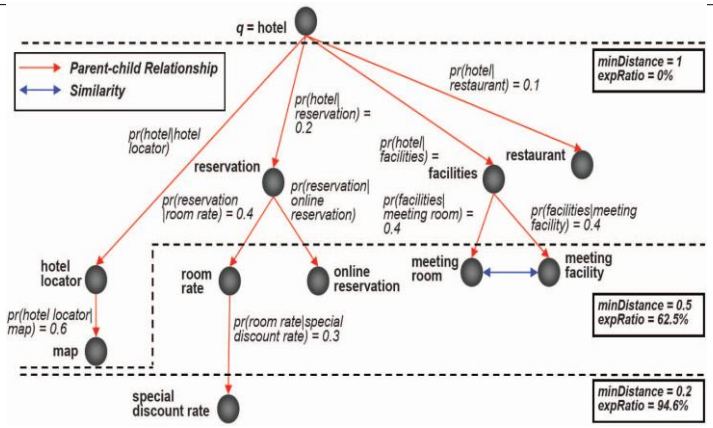

Fig 1.Ontology for q="hotel"with p=0.2,0.5,1.0.

- Similarity. Two concepts which coexist a lot on the search results might represent the same topical interest. If coexist(ci,cj)>$1 ($1 is a threshold), then ci and cj are considered as similar.
- Parent-child relationship: More specific concepts often appear with general terms, while the reverse is not true. Thus, if pr(cj|ci) > $2 ($2 is a threshold), to mark ci as cj's child. For example, the more specific concept "meeting facility" tends to occur together with the general concept "facilities," while the general concept "facilities" might also occur with concepts such as "meeting room" or "swimming pool," i.e., not only with the concept "meetin facility."

Fig. 1 shows an example content ontology created for the query "hotel," where content concepts linked with a one sided arrow(⟶) are parent-child concepts, and concepts linked with a double-sided arrow (⟷) are similar concepts. Fig. 1 shows the possible concept space determined for the query "hotel," while the clickthrough data determine the user preferences on the concept space. In general, the ontology covers more than what the user actually wants. The concept space for the query "hotel" consists of "map," "reservation," "room rate,"..., etc. If the user is indeed interested in information about hotel rates and clicks on pages containing "room rate" and "special discount rate" concepts, the captured clickthrough favors the two clicked concepts. Feature vectors containing the concepts "room rate" and "special discount rate" as positive preferences will be created corresponding to the query "hotel." As indicated in Fig. 1, when the query is issued again later, these feature vectors will be transmitted to the PWSE server an transformed into a content weight vector to rank the search results according to the user's content preferences.

B. LOCATION ONTOLOGY

The approach for extracting location concepts is different from that for extracting content concepts. To observe two important issues in location ontology formulation. First, a document usually embodies only a few location concepts, and thus only very few of them co-occur with the query terms in web-snippets. To alleviate this problem, to extract location concepts from the full documents. Second, the similarity and parent-child relationship cannot be accurately derived statistically because the limited number of location concepts embodied in documents. Furthermore, many geographical relationships among locations have already been captured as facts. Thus, we obtain about 17,000 city, province, region, and country names and create a predefined location ontology among these locations. Organize all the cities as children under their provinces, all the provinces as children under their regions, and all the regions as children under their countries. The statistics of our location ontology are provided in Table I.

TABLE I
Statics of the Location Ontology

No.of Countries	7	Total No.of Nodes	16899
No.of Regions	190	Country-Region Edges	190
No.of Provinces	6699	Region-Province Edges	1959
No.of Towns	10003	Province-City Edges	

The predefined location ontology is used to associate location information with the search results. All of the keywords and key-phrases from the documents returned for query q are extracted. If a keyword or key-phrase in a retrieved document d matches a location name in our predefined location ontology, it will be treated as a location concept of d. For example, assume that document d contains the keyword "Los Angeles." "Los Angeles" would then be matched against the location ontology. Since "Los Angeles" is a location in our location ontology, it is treated as a location concept related to d. Furthermore, explore the predefined location hierarchy, which would identify "Los Angeles" as a city under the state "California." Thus, the location "/United States/California/Los Angeles/" is associated with document d. If a concept matches several nodes in the location ontology, all matched locations will be associated with the document. Similar to the content ontology, the location ontology together with clickthrough data are used to create feature vectors containing the user location preferences. They will then be transformed into a location weight vector to rank the search results according to the user's location preferences.

C. ASSOCIATION RULE MINING (ARM)

Association rule learning is a well-liked and well research technique for discover attractive relatives among variables in PWSE search engine. It is planned to distinguish physically powerful rules exposed in search engines using dissimilar procedures of interestingness. Based on the conception of well-built rules, Rakesh Agrawal et al. Introduce association rules for discovering regularities among products in important transaction data record by point-of-sale (POS) systems in supermarket. Association rules are typically compulsory to keep happy a user-specified smallest amount support and a user-specified minimum confidence at the similar instance. Association regulation generation is typically opening up interested in two divide steps: Primary, minimum support is practical to discover all frequent query patterns in a user clickthrough data. Next frequent query item sets and the minimum confidence restraint be second-hand to appearance rules.

To discover a query traveler's interest extract beginning search based user click all the way through files whilst the personal user search the consequences on or after mobile. Whilst user enter query base path or traversal patterns are recognized initially and after that make frequent item set with the intention of number of instance the user click thorough files and find the majority significant travel patterns in the clickthrough files. This investigate focus on the travelers who use mobile search contain the majority frequent based links in together location and concept based ontology previous to so as to discover the frequent item set that is additional numeral of period user look for the comparable web pages or concept and location. Beginning this compute the support and confidence standards of the click through files and the majority relevant regular query patterns results are considered as consumer the majority important concepts and location then yet again go on the concept to rank the feature for both content and location ontology. Association Rule Mining (ARM) query travel pattern to explore for go target that is user concept consequences practical data mining and association rules method to investigate the association among travelers' profile and their transactions in the data .After this examine the identify majority important pattern to investigate the outcome and can amplify opportunity for the competitive operations of tourism firm to respond the travelers' demand effectively.

Specified a set of user click through is measured as set of items I= {i1,i2,i3,….im} and a record of transactions with travel patterns DB= {t1,t2…tn} where ti={Ii1,Ii2,…Iip}, $p \leq m$ and, if $A \subseteq I$ with K = |A| is called a k-itemset or simply an itemset. Let a database D be a multi-set of subsets of I as shown. Each $T \in DB$ supports an itemset $A \subseteq I$ if $A \subseteq T$ holds. An

association rule is an expression A => B, where A, B are item sets and $X \cap Y = \emptyset$ holds. Number of transactions T supporting an item A w.r.t DB is called support of A, $Supp(A) = |\{T \in DB \mid A \subseteq T\}| / |DB|$. The strength or confidence (c) for an association rule A => B is the ratio of the number of transactions that contain A ∪ B to the number of transactions that contain A, Conf (A → B) =Supp (A ∪ B)/ Supp (A). After that the query patterns are analyzed with the help of the patterns in association rule mining. PWSE first performs the functions with RSVM training at the PC server side with user profile based preferences. Lastly clickthrough data says the precise user preference on the search results are stored into PWSE server and protects the user privacy. Personalized query collects the query travel patterns from the original query patterns in ARM, after that query based results uses a concept and location based contents by maintain the ontology that the result returned from the search result based on user interest in query travel patterns by considering both location and concept information. It is classified into two categorize content concepts and location concepts with patterns returned from ARM. Observe the significant individuality of the content concepts and location concepts on dissimilar perspective of each user.

III. LITERATURE SURVEY

A. *APPLYING COTRAINING TO CLICKTHROUGH DATA FOR SEARCH ENGINE ADAPTATION*

The author [7] has been proposed to applying cotraining to clickthrough data for search engine adaptation.

The information on the World Wide Web is growing without bound. Users may have very diversified preferences in the pages they target through a search engine. It is therefore a challenging task to adapt a search engine to suit the needs of a particular community of users share similar interests. Ranking SVM in a Co training Framework. Essentially, the RSCF algorithm takes the clickthrough data containing the items in the table II.

TABLE II
Clickthrough for the query "HOTEL"

Doc	Search Result	Ci	Li
D1	Hotels.com	Room rate	International
D2	Japan Hotel.net	Reservation, Room rate	Japan
D3	Hotel Wiki	Accommodation	international
D4	US Hotel Guides	Map, room rate	USA, California
D5	Booking.com	Online reservation	USA
D6	JAL Hotels	Meeting room	Japan
D7	Shinjuku Prince	Facility	Japan, Shinjuku
D8	Discount Hotels	Discount rate	International

Search result that have been clicked on by a user as an input, and generates adaptive rankers as an output. By analyzing the clickthrough data, RSCF first categorizes the data as the labeled data set, which contains the items that have been scanned already, and the unlabelled data set, which contains the items that have not yet been scanned. The labeled data is then augmented with unlabelled data to obtain a larger data set for training the rankers. The RSCF algorithm produces better ranking results than the standard Ranking SVM algorithm. Analyzing clickthrough data is a useful means to understand users' target preference in the returned search results, since it conveys partial relative relevance judgments on the links that a user has browsed through. However, if we consider a collection of clickthrough data generated from a large number of users, the clickthrough data may be too diversified for inferring the best results for a query across the entire user group. On the other hand, we observe that for a particular community of users who share interests in the same domain Computer Science, their behavior is relatively similar and the inference of preference can be much more reliable. Clickthrough data was used to optimize the ranking in search engines.

B. *TO IMPROVE PERSONALIZED SEARCH*

The author[5] has been proposed to discovering and using groups to improve personalized search.

Web search personalization algorithms improve the Web search experience by using an individual data that is topical categories marked interesting, query history, or term vectors of previously viewed content to identify the results that are the most relevant to that individual. a searcher's query can be modified to reflect a particular interest, or results may be reranked so that personally relevant results appear higher in the list. Personalized Web search takes advantage of information about an individual to identify the most relevant results for that person. A challenge for personalization lies in collecting user profiles that are rich enough to do this successfully. One way an individual's profile can be augmented is by using data from other people. To better understand whether groups of people can be used to benefit personalized search, the similarity of query selection, desktop information, and explicit relevance judgments across people grouped in different ways.

The groupings are explored fall along with two dimensions, the longevity of the group member relationship, and how explicitly the group is formed. Find that some groupings provide valuable insight into what members consider relevant to queries related to the group focus, but that it can be difficult to identify valuable groups implicitly. An explicit task based group is one where group members are overtly collaborating on a specific task. Group membership can also be inferred such group called implicit groups. An implicit task based group may be formed from people who appear, based on their actions, to be conducting the same task. Interest based groups are identified based on mailing list membership; the membership is explicit, but the inference that members share an interest is implicit. Based on general interest to participants, the queries were chosen via examination of logs of queries issued. Although the queries were intended to be interesting, they were not created by the participants. Instead, participants had to create an intent for each query evaluated. By allowing people to decide whether they wanted to evaluate a particular query, sought to allow them to only work with queries and associated results that were meaningful.

Personally relevant explicit judgments for the same queries would have been difficult to collect using only self generated queries, since it would require different participants to coincidentally issue the same query on their own. By exploring the potential for using information from a group of related users to enhance the personalization of Web search results. Analyzing the similarity of query choices, relevance judgments, and personal content for several categories of explicitly and implicitly defined task based and trait based groups. And found that several of the groups at the explicit end of the spectrum were similar in many respects when considering queries related to their group's theme, but that for off-theme queries such groups were less cohesive. Implicitly defined groups also lacked cohesion with respect to our three similarity metrics. The effectiveness of grouping and personalization technique combines personal and group content to improve Web rankings, for different group query combinations.

C. PRIVACY ENHANCING PERSONALIZED WEB SEARCH.

The author[10] has been proposed to privacy enhancing personalized web search to improve the search quality. As the amount of information on the web continuously grows, it has become increasingly difficult for web search engines to find information that satisfies users individual needs. Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Most of them could be categorized into two general approaches are Reranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine. A good personalization algorithm relies on rich user profiles and web corpus. However, as the web corpus is on the server, reranking on the client side is bandwidth intensive because it requires a large number of search results transmitted to the client before reranking.

Personalized web search is a promising way to improve search quality by customizing search results for people with individual information goals. However, users are uncomfortable with exposing private preference information to search engines. On the other hand, privacy is not absolute, and often can be compromised if there is a gain in service or profitability to the user.

Thus, a balance must be struck between search quality and privacy protection. A scalable way formulated for users to automatically build rich user profiles. These profiles summarize a user's interests into a hierarchical organization according to specific interests. Two parameters for specifying privacy requirements are to help the user to choose the content and degree of detail of the profile information that is exposed to the search engine. The user profile improved search quality when compared to standard MSN rankings.

D. ONTOLOGY SUPPORTED PERSONALIZED SEARCH FOR MOBILE DEVICES.

Mobile web search represents a different paradigm in front of desktop search. When in a mobile environment, users need a time and cost efficient web search as well as obtaining significant results taking into account the context that surrounds him, all in a limited-interface device with a higher bandwidth cost. One important challenge to improve the web search for a proper mobile user experience is related to the mobile device interface and input method, because of most of actual devices has a 9 or 12 key keypad, and only minorities of them include a full physical or virtual keypad. An intelligent interface that helps users to write fewer letters with a recommender or auto completion system is a desirable component on limited keyboard devices. If these recommendations also consider context information, the recommended words will be more precise and effective. When in move, the user needs an efficient way to introduce query terms and receive more precise information.

It creates and manages the user and device profiles which consist of environmental data like time and place, user preferences, device capabilities description. The working prototype which also includes a user interface designed specifically for mobile devices which consider their limitations of screen and keypad. The use of ontologies and a thesaurus help us to offer better results according to our main objective in our recommendation system. The ontologies let us model the spatial, temporal, user dependent and device capabilities concepts easily .One important challenge to improve the web search for a proper mobile user experience is related to the mobile device interface and input method, because of most of actual devices has a 9 or 12 key keypad, and only minorities of them include a full physical or virtual keypad. An intelligent interface that helps users to write fewer letters with a recommender or auto completion system is a desirable component on limited keyboard devices.

IV. CONCLUSION

Users are given with predefined queries and topical interests, users have to synthesize the needed information from the given queries and topical interests and conduct the searches correspondingly. Ideally, a large scale user study should be conducted in which PWSE is subjected to real life use, users' behaviors are monitored transparently and satisfaction of the users is analyzed and compared with other systems, but a large scale is implemented.

References

1. Kenneth Wai-Ting Leung, Dik Lun Lee, and Wang-Chien Lee,"A Personalized Mobile Search Engine"2013.
2. B. Liu, W.S. Lee, P.S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," Proc. Int'l Conf. Machine Learning (ICML), 2002.
3. C.E. Shannon, "Prediction and Entropy of Printed English," Bell Systems Technical J., vol. 30, pp. 50-64, 1951.
4. E.Voorhees and D. Harman, TREC Experiment and Evaluation in Information Retrieval. MIT Press, 2005.
5. J. Teevan, M.R. Morris, and S. Bush, "Discovering and Using Groups to Improve Personalized Search," Proc. ACM Int'l Conf. Web Search and Data Mining (WSDM), 2009.
6. J.Y.H. Pong, R.C.W. Kwok, R.Y.K. Lau, J.X. Hao, and P.C.C. Wong, "A Comparative Study of Two Automatic Document Classification Methods in a Library Setting," J. Information Science, vol. 34, no. 2, pp. 213-230, 2008.
7. Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-Training to Clickthrough Data for Search Engine Adaptation," Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2004.
8. S. Yokoji, "Kokono Search: A Location Based Search Engine," Proc. Int'l Conf. World Wide Web (WWW), 2001.
9. W. Ng, L. Deng, and D.L. Lee, "Mining User Preference Using Spy Voting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.
10. Y. Xu, K. Wang, B. Zhang, and Z. Chen, "Privacy-Enhancing Personalized Web Search," Proc. Int'l Conf. World Wide Web (WWW), 2007.

AUTHOR(S) PROFILE

K.MALA received the B.E. degree in Computer Science and Engineering from Roever Engineering college, Anna Univercity, Chennai, Tamil, India and the M.E. Computer Science and Engineering from Shivani Engineering College, Anna Univercity, Chennai, Tamilnadu, India in 2012 and 2014.