International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 10, October 2015
Available at http://internationaljournalofresearch.org

A Novel Approach for Supporting Privacy Protection in Personalized Web Search by Using Data Mining

Kurcheti Hareesh Babu[1], K Koteswara Rao [2]
[1]PG Scholar,Dept of CSE, Rao & Naidu Engineering College, Ongole,Prakasam Dist,Andhra Pradesh
[2]Assistant Professor,Dept of CSE, Rao & Naidu Engineering College, Ongole,Prakasam Dist,Andhra Pradesh

ABSTRACT

Web search engines help the users to find useful information on the World Wide Web (WWW). But it has become increasingly difficult to get the expected results from the search engine as a large number of topics are being discussed on the web. Generally, each user has different information needs for his/her query. Therefore, the search results should be adapted to users with different information needs. Personalized web search is a way which provides customized search results for people with individual information goals. But this approach has private issues since it usually requires users to disclose their personal information during search and the users are reluctant to disclose their information. Here we study, how we can provide privacy in PWS applications. privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting userspecified privacy requirements. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that GreedyIL significantly outperforms GreedyDP in terms of efficiency.

Index Terms—Privacy protection; personalized web search; utility; risk; profile

I.**INTRODUCTION**

Web search engine has become the most important portal for users finding useful information on the web. As the amount of information on the web continuously grows, it has become increasingly difficult for the web search engines to find the information what a user is looking for. For example, for the query “office” some users may be searching for a vacant office space, while other users may be searching for popular Microsoft productivity software. Therefore, Web search results should adapt to users with different information needs. Personalized web search (PWS) is a promising way aiming at providing better search results, which are tailored for individual user needs [1].

The click-log based methods are straight forward methods because they simply impose bias to the clicked pages in the user’s query history. This strategy performs consistently and considerably well but it can work only on repeated queries from the same user, so this approach has limitations. The Profile-based methods improve the search quality by generating complicated user-interest models by using user profiling techniques. Profile-based methods prove to be effective for almost all sorts of queries, but they may become unstable under some circumstances. Though there are limitations in this approach, it has demonstrated more effectiveness in improving the quality of web search [2]. Searching is one of the common factor to know the information from the internet. Internet is one of the service providers, which provide the search result to the user with the help of the Web search engine (WSE). It employ by storing information about many web pages [3]. WSE is a tool which allows the web user for finding information from the World Wide Web. WSE is one of the software that searches for and identifies the content or item from the web engine or web server or web database with correspond keywords or character specified by the user and finding particular sites on the World Wide Web [4,5].

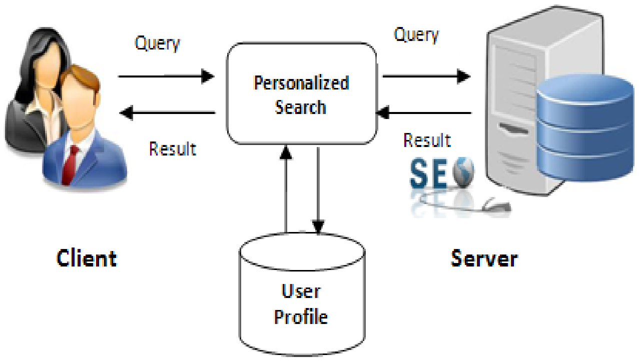

Figure 1: Personalized Search Engine Architecture

Data search and information retrieval on the Internet has located high demands on search engines. Many search engines like Google, Yahoo provide a relevant and irrelevant data to the user based on their search. To avoid the irrelevant data the technique called Personalized Web Search (PWS) were arise. Inferring user search goals is very important in improving search-engine relevance and

personalized search [6]. This is based on the user profiles based on the click through log and the feedback session. These data were generated from the frequent query requested by the user, history of query, browsing, bookmarks and so on. By these methods personal data were easily reveal. While many search engines take advantage of information about people in common, or regarding particular groups of people, personalized search based on a user profile that is unique to the individual person. Research systems that personalize search outcomes model their users in different ways [7]. The Personalized Web Search provides a unique opportunity to consolidate and scrutinize the work from industrial labs on personalizing web search using user logged search behavior context. It presents a fully anonymized dataset, which has anonymized user id, queries based on the keywords, their terms of query, providing URLs, domain of URL and the user clicks.

This dispute and the shared dataset will enable a whole new set of researchers to study the problem of personalizing web search experience. It decreases the likelihood of finding new information by biasing search results towards what the user has already found. By using these methods privacy of the user might be loss because of clicking the relevant search, frequently visited sites and providing their personal information like their name, address, etc. in this case their privacy might be leak. For this privacy issue, many existing work proposed a potential privacy problems in which a user may not be aware that their search results are personalized for them. It affords a host of services to people, and several of these services do not necessitate information to be grouped about a person to be customizable. While there is no warning of privacy assault with these services, the stability has been tipped to errand personalization over privacy, yet when it comes to search. That approaches does not protect privacy issues rising from the lack of protection for the user data. To providing better privacy we propose a privacy preserving with the help of greedy method by providing the hybrid method of the discriminating power and prevent the information loss.

II.RELATED WORK

We focus on the literature of profile-based personalization and privacy protection in PWS system.

2.1 Profile-Based Personalization

Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. In the remainder of this section, we review the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization. Many profile representations are available in the literature to facilitate different personalization strategies. Earlier techniques utilize term lists/vectors or bag of words to represent their profile. However, most recent works build profiles in hierarchical structures due to their stronger descriptive ability, better scalability, and higher access efficiency [4,5]. The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP1. Another work builds the hierarchical profile automatically via term-frequency analysis on the user data. In our proposed UPS framework, we do not focus on the implementation of the user profiles. Actually, our framework can potentially adopt any hierarchical representation based on a taxonomy of knowledge. As for the performance measures of PWS in the literature, Normalized Discounted Cumulative Gain (nDCG) is a common measure of the effectiveness of an information retrieval system. It is based on a human graded relevance scale of item-positions in the result list, and is, therefore, known for its high cost in explicit feedback collection. To reduce the human involvement in performance measuring, researchers also propose other metrics of personalized web search that rely on clicking decisions, including Average Precision (AP) Rank Scoring, and Average Rank. We use the Average Precision metric, proposed by Dou et al., to measure the effectiveness of the personalization in UPS. Meanwhile, our work is distinguished from previous studies as it also proposes two predictive metrics, namely personalization utility and privacy risk, on a profile instance without requesting for user feedback [7].

2.2 Privacy Protection in PWS System

Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server. Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudoidentity, the group identity, no identity, and no personal information. Solution to the first level is proved to fragile. The third and fourth levels are impractical due to high cost in communication and cryptography [12]. Therefore, the existing efforts focus on the second level. Online anonymity on user profiles by generating a group profile of k users. Using this approach, the linkage between the query and a single user is broken. The useless user profile (UUP) protocol is proposed to shuffle queries among a group of users who issue them. As a result any entity cannot profile a certain individual. These works assume the existence of a trustworthy third-party anonymizer, which is not readily available over the Internet at large. Viejo and Castell_a-Roca use legacy social networks instead of the third party to provide a distorted user profile to the web search engine [9]. In the scheme, every useracts as a search agency of his or her neighbors. They can decide to submit the query on behalf of who issued it, or forward it to other neighbors. The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication [11]. How to exploit implicit user modeling to intelligently personalize information retrieval and improve search accuracy. Unlike most previous work, it emphasizes the use of immediate search context and implicit feedback information as well as eager updating of search results to maximally benefit a user [10]. Author presented a decision-theoretic framework for optimizing interactive information retrieval based on eager user model updating, in which the system responds to every action of the user by choosing a system action to optimize a utility function. Author propose specific techniques to capture and exploit two types of implicit feedback information:

(1) identifying related immediately preceding query and using the query and the corresponding search results to select appropriate terms to expand the current query, and

(2) exploiting the viewed document summaries to immediately re-rank any documents that have not yet been seen by the user.

III.PROPOSED SYSTEM

This paper proposes a privacy- preserving personalized web search framework called UPS i.e User customizable Privacy- preserving Search, that generalize profile for every query as per user privacy specification. Based on personalization and privacy risk metric, this paper formulates Risk Profile Generation, with its NP- hardness proved. It develops two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. GreedyDP maximize the discriminating power (DP) while GreedyIL minimize the information loss (IL). This paper also provides a mechanism for the client to decide whether or not to personalize a query in UPS. This decision is made before each runtime profiling to enhance the stability of the search results.

System architecture of UPS

UPS consists of number of clients/users and a server for fulfilling client's request. In client's machine, the online profiler is implemented as search proxy who maintains users profile in hierarchy of nodes and also maintain the user specified privacy requirement as a set of sensitive nodes. There are two phase, namely Offline and Online phase for the framework. During Offline, a hierarchical user profile is created and user specified privacy requirement is marked on it. The query fired by user is handled in the online phase as: When user fires a query on the client, proxy generates user profile in run time. The output is generalized user profile considering the privacy requirements. Then, the query along with generalized profile of user is sent to PWS server for personalized web search. The search result is personalized and the response is sent back to query proxy. Finally, the proxy presents the raw result or re ranks them with user profile.

The privacy concern is one of the major barriers in deploying serious personalized search applications, and how to attain personalized search though preserving users' privacy. Here we propose a client side personalization which deals with the preserving privacy and envision possible future strategies to fully protect user privacy. For privacy, we introduce our approach to digitalized multimedia content based on user profile information. For this, two main methods were developed: Automatic creation of user profiles based on our profile generator mechanism and on the other hand recommendation system based on the content to estimates the user interest based on our client side meta data.

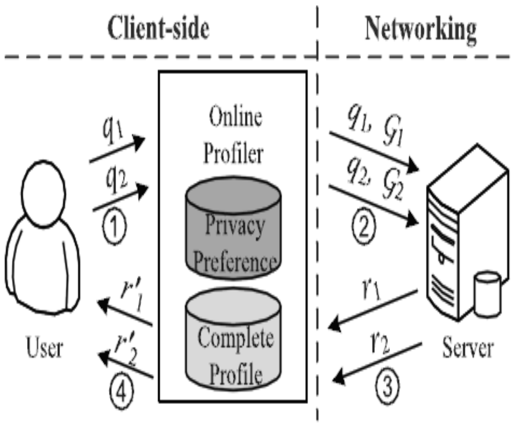

Figure 2: Proposed Architecture

Above figure shows our proposed architecture which builds in the client side mechanism and here we protect the data from the server, so only we provide privacy to the client user. Every query from the client user were provided by the separate requests to the server, this hides the frequent click through logs or content based mechanism, from this user can protect the data from the server. In the same case our mechanism maintains the online profiler about the user hence it hides the click logs and provides a safeguard to the user data. After that, online profiler query were processed in the manner of generalization process, it is used to meet the specific prerequisites to handle the user profile and it is based on the preprocessing the user profiles.

Our architecture, not only the user's search performance but also their background activities (e.g., viewed before) and personal information (e.g., emails, browser bookmarks) could be included into the user profile, permitting for the structure of a much richer user model for personalization. The sensitive contextual information is usually not a main aspect since it is strictly stored and used on the client side. A user's personal information including user queries and click logs history resides on the user's personal computer, and is exploited to better suppose the user' information require and provide a relevant search results.

Our proposed algorithm uses the greedy method based on the discriminating power and information loss protection to inherit the relations. Here it uses the inherited method to generalize the query. It allows performing the customization process to protect the data and use the User customizable Privacy-preserving Search framework addressed the privacy problems. This aims at protecting the privacy in individual user profiles.

Advantages:

1. It enhances the stability of the search quality.

2. It avoids the unnecessary exposure of the user profile.

3. The framework allowed users to specify customized privacy requirements via the

hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality.

IV CONCLUSION & FUTURE WORK

Web users were increases because of available of information's from the web browser based on the search engine. With the increasing number of user service engine must provide the relevant search result based on their behavior or based on the user performance. Providing relevant result to the user is based on their click logs, query histories, bookmarks, by this privacy of the user might be loss. For providing relevant search by using these approaches the privacy of the user may loss. Most existing system provides a major barrier to the private information during user search. That approaches does not protect privacy issues and rising information loss for the user data. For this issue this paper proposes client based architecture based on the greedy algorithm to prevent the user data and provide the relevant search result to the user in future it can include this work in mobile application.

V.REFERENCES

[1] "Personalized Web Search", W. M. P. VAN DER AALST Eindhoven University of Technology, Eindhoven.

[2] Chirita P.A., Nejdl W., Paiu R., and Kohlschu¨tter C. Using ODP metadata to personalize search. In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2005, pp. 178–185.

[3] Sugiyama K., Hatano K., and Yoshikawa M. Adaptive web search based on user profile constructed without any effort from users. In Proc. 12th Int. World Wide Web Conference, 2004, pp. 675–684.

[4] Shen X., Tan B., and Zhai C. Implicit user modeling for personalized search. In Proc. Int. Conf. on Information and Knowledge Management, 2005, pp. 824–831.

[5] Teevan J., Dumais S.T., and Horvitz E. Personalizing search via automated analysis of interests and activities. In Proc. 31st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2005, pp. 449–456.

[6] Chirita P.A., Firan C., and Nejdl W. Summarizing local context to personalize global web search. In Proc. Int. Conf. on Information and Knowledge Management, 2006.

[7] Page L., Brin S., Motwani R., and Winograd T. The pagerank citation ranking: bringing order to the web. Technical report, Computer Science Department, Stanford University, 1998.

[8]K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adap-tive Web Search Based on User Profile Constructed with-out any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

[9]X. Shen, B. Tan, and C. Zhai, "Implicit User Modeling for Personalized Search," Proc. 14th ACM Int'l Conf. In-formation and Knowledge Management (CIKM), 2005.

[10]X. Shen, B. Tan, and C. Zhai, "Context-Sensitive In-formation Retrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[11F. Qiu and J. Cho, “Automatic Identification of User Interest for Personalized Search,” Proc. 15th Int’l Conf. World Wide Web (WWW), pp. 727-736, 2006.

[12]J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turn-bull, A. Edmonds, E. Adar, and T. Breuel, “Personalized Search,” Comm. ACM, vol. 45, no. 9, pp. 50-55, 2002.

[13]Y. Xu, K. Wang, B. Zhang, and Z. Chen, “Privacy Enhancing Personalized Web Search,” Proc. 16th Int’l Conf. World Wide Web (WWW), pp. 591-600, 2007.

[14]K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.