Survey on Privacy Preservation in Personalized Web Environment

Esmita Gupta
Dept. of Information Technology
VIT, Mumbai, India
esmita.g@gmail.com

Prof. Deepali Vora
Dept. of Information Technology
VIT, Mumbai, India
Deepali.vora@vit.edu.in

***Abstract*—** Personalized web search (PWS) is a general category of search techniques aiming at providing different search results for different users or organize search results differently for each user, based upon their interest, preferences and information needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query. However, users are uncomfortable with exposing private information during search which has become a major barrier for the wide proliferation of PWS. Search engines should provide security mechanism such that user will be ensured of its privacy and its information should be kept safe. Many personalization techniques are giving access to achieve personalization of user's web search. Search engines can provide more accurate and specific data if users trust search engine and provide more information. But users should be ensured that their private information should be kept safe. In this paper we will discuss on different techniques on personalized web search and securing personalized information.

Keywords*-Personalized web search, Personalization Techniques, Privacy, Information Retrieval*

*****

I. INTRODUCTION

Nowadays Internet is widely used by users to satisfy various information needs. However, ambiguous query/topic submitted to search engine doesn't satisfy user information needs, because different users may have different information needs on diverse aspects upon submission of same query/topic to search engine. So discovering different user search goals becomes complicated. The evaluation and depiction of user search goals can be very useful in improving search engine relevance and user knowledge [1].

Personalized Web search is a technique in order to provide better search results. It is a promising way to improve search quality by customizing search results for people with different information goals. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts.

Apart from the personalized results, there is need of security in the personalized web search. Users are not keen to disclose their information during web search. This has become major issue in profiling the user in personalized web search. There should be a mechanism which considers profiles according to information provided by user. Actually more the search engine knows about user, more accurate search results will be obtained by search provider. But users cannot trust on search engine that information provided by user is not misused. Search engines can provide more accurate and specific data if users trust search engine and provide more information. Hence, search engines should provide security mechanism such that user will be ensured of its privacy and its information should be kept safe.

In personalized web search, user information is collected and analyzed in order to find intention behind issued query fired by user. Typically search is performed by providing queries to retrieval system in form of set of words. If different users enter same query, the system will produce same results without considering the user. But search results should be produced by taking the user in the equation, so that different users can get different search results for same query. By keeping track of user's personal information and interests.

II. LITERATURE SURVEY

Personalized web search (PWS) differs from generic web search, which returns identical research results to all users for identical queries, regardless of varied user interests and information needs. PWS can be categorized into two types; one is click-log-based methods and other profile-based ones.

The click log based methods are based on just selecting the clicked pages in the user's query history. The main drawback of this method is that it works on repeated set of queries by the users only.

Profile based method has more effectiveness in improving the quality of web search with increasing usage of personal and behavior information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data, bookmarks, user documents and so forth. [2]. The main drawback of this method is that it requires the user personal data to be send to the server, hence this privacy issue makes the user uncomfortable.

To provide personalized search results to users, personalized web search maintains a user profile for each individual [3].These profiles can be used in various ways to create an

environment of personalized search. Some of the methods to help in inferring user's information needs are:

- Personalized Search Based on Content Analysis

Personalized web search can be achieved by checking content similarity between web pages and user profiles [3]. When the user issues a query, each returned snippet/documents are filtered or re-ranked and classified. Chirita et al. [3] [4] use the ODP (Open Directory Project) hierarchy to implement personalized search. In [5], a user profile is built as a vector of distinct terms and is constructed by aggregating past user click history [3]. Shen et al. [6] first use language modeling to mine immediate search contextual and implicit feedback information [3]. Teevan et al. [7] and Chirita et al. [8] exploit rich models of user interests, built from both search-related information, and other information about the user. [3]

- Personalized Web Search Based on Hyperlink Analysis

Most generic web search approaches rank importance of documents based on the linkage structure of the web. A large group of these works focuses on personalized PageRank. PageRank, proposed by Page and Brin [9], is a popular link analysis algorithm used in web search. The fundamental motivation underlying PageRank is the recursive notion that important pages are those linked-to by many important pages [3].

Qiu and Cho [10] develop a method to automatically estimate a user's topic preferences based on Topic-Sensitive PageRank scores of the user's past clicked pages. The topic preferences are then used to bias future search results.

- Community-based Personalized Web Search

Some approaches that personalize search results for the preferences of a community of like-minded users [3] is known as community or collaborative based search. In community-based personalized web search, when a user issues a query, search histories of users who have similar interests to the user are used to filter or re-rank search results [3]. Sugiyama et al. [5] use a modified collaborative filtering algorithm to constructed user profiles to accomplish personalized search. Sun et al. [11] proposed a novel method named CubeSVD to apply personalized web search by analyzing correlations among users, queries, and web pages in clickthrough data. Smyth et al. [12] show that collaborative web search can be efficient in many search scenarios when natural communities of searchers can be identified [3].

- Server-Side and Client-Side Implement

Personalized web search can be implemented on either server side (in the search engine) or client side (in the user's computer or a personalization agent).

For server-side personalization, user profiles are built, updated, and stored on the search engine side [3].

For client-side personalization, user information is collected and stored on the client side (in the user's computer or a personalization agent), usually by installing a client software or plug-in on a user's computer [3].

Despite of having various advantages of personalized search, there is no large-scale use of personalized search services currently. Personalized web search faces several challenges that hinder its real-world large-scale applications:

- Privacy is an issue.
- Users are not static.
- Queries should not be handled in the same manner with regard to personalization [3].

In order to improve performance of the web search we need to take care of privacy issues in personalized web search. As personalizing search requires gathering and processing of user information, which leads to privacy issue. This is becoming the main obstacle in deploying personalized web search applications.

Adequate work has been proposed inorder to maintain privacy in personalized web search:

- Chaum proposed in [18] the use of an anonymity network which consists of several routers that act as anonymizers. It is a technique which is based on public key cryptography that allows an electronic mail system to hide who a participant communicates with as well as the content of the communication - in spite of an unsecured underlying telecommunication system.
  Drawback: The main drawback of this approach was that the process of submitting a query to the WSE and receiving the answer through an anonymous channel is very time-consuming.
- Krause and Horvitz in [19] employ statistical techniques to learn a probabilistic model, and then use this model to generate the near-optimal partial profile. They had introduced and explore an economics of privacy in personalization, where people can opt to share personal information, in a standing or on-demand manner, in return for expected enhancements in the quality of an online service.
  Drawback: Limitation in this work was that it builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries. These assumptions are impractical in the context of PWS.

- Xu et al. in [15] proposed a privacy protection solution for PWS based on hierarchical profiles. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted subtree of the complete profile. These profiles summarize a user's interests into a hierarchical organization according to specific interests. Two parameters for specifying privacy requirements are proposed to help the user to choose the content and degree of detail of the profile information that is exposed to the search engine.
  Drawback: The main drawback of this approach was that this work does not address the query utility, which is crucial for the service quality of PWS.
- Xiao and Tao proposed Privacy-Preserving Data Publishing (PPDP). A person can specify the degree of privacy protection for her/his sensitive values by specifying "guarding nodes" in the taxonomy of the sensitive attribute. [13]
  Drawback: The greedy algorithm presented in this paper was not optimal and also did
  not support runtime profiling.
- Teevan et al. collect a set of features of the query to classify queries by their click entropy. He first examined the variability in user intent for a large number of queries using both implicit and explicit measures. Then study was carried out to show variation in the implicit measures predicts variation in the explicit measures, and look at what other factors can account for variation in the implicit measures. Queries are characterized using a variety of features of the query, the results returned for the query, and the query's interaction history. Using these features predictive models were built to identify the queries that will benefit most from personalization, and explore which features are the most valuable for prediction [14].
  Drawback: This works motivates in questioning whether to personalize or not to, they assume the availability of massive user query logs and user feedback.

All the existing profile-based Personalized Web Search does not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries. Profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk. [17]

The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. [2] For example, in, all the sensitive topics are detected using an absolute metric called surprisal based on the information theory, assuming that the interests with less user document support are more sensitive.

Many personalization techniques require iterative user interactions when creating personalized search results. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on. This paradigm is, however, infeasible for runtime profiling, as it will not only pose too much risk of privacy breach, but also demand prohibitive processing time for profiling.

Thus, there is a need of predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

III. CONCLUSION

This paper presents a survey report of different methods to help in inferring user's information needs of Personalized Web Search. It also covers issues like need of personalized web search, how personalized web search can be implemented, what are challenges in it, privacy and security issue of it and existing system of personalized web search.

This paper also gives a survey report of different ways to maintain privacy in personalized web environment. It also tells about the drawbacks of the existing privacy issues.

The future scope of our paper will be to overcome the existing system drawbacks and design a framework to maintain a complete privacy of the users so that they can work without any fear of working in personalized web environment.

REFERENCES

[1] Charudatt Mane, Pallavi Kulkarni, "A Novel Approach to Discover User Search Goals Using Click through Data", International Journal of Computer Science and Information Technologies, Vol. 5 (1), 2014.

[2] Lidan Shou, He Bai, Ke Chen, and Gang Chen, "Supporting Privacy Protection In Personalized Web Search", IEEE transactions on knowledge and data engineering vol:26 no:2 year 2014.

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

[4] 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.

[5] Sugiyama K., Hatano K., and Yoshikawa M. Adaptive web search based on user profile constructed without any effort

ISSN: 2321-8169
1787 – 1790

from users. In Proc. 12th Int. World Wide Web Conference, 2004, pp. 675–684.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Qiu F. and Cho J. Automatic identification of user interest for personalized search. In Proc. 15th Int. World Wide Web Conference, 2006, pp. 727–736.

[11] Sun J.-T., Zeng H.-J., Liu H., Lu Y., and Chen Z. CubeSVD: a novel approach to personalized web search. In Proc. 14th Int. World Wide Web Conference, 2005, pp. 382–390.

[12] Smyth B., Coyle M., Boydell O., Briggs P., Balfe E., Freyne J., and Bradley K. A live-user evaluation of collaborative web search. In Proc. 19th Int. Joint Conf. on AI, 2005.

[13] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc.ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006

[14] J. Teevan, S.T. Dumais, and D.J. Liebling, "To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 163-170, 2008.

[15] 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.

[16] Y. Zhu, L. Xiong, and C. Verdery, "Anonymizing User Profiles for Personalized Web Search," Proc. 19th Int'l Conf. World Wide Web (WWW), pp. 1225-1226, 2010.

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

[18] D. Chaum, Untraceable electronic mail, return addresses, and digital pseudonyms, Commun. ACM 24 (2) (1981) 84–90.

[19] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services", Journal of Artificial Intelligence Research 39 (2010) 633-662.