ISSN:0975-9646

A Survey on Supporting Privacy Protection in Personalized Web Search

Chanda Agarwal[1], Prof.Dr.Suhasini Itkar[2]

M.E. Scholar, Computer Engineering, PES Modern College of Engineering, Pune, India [1]

Head of Computer Department, Computer Engineering, PES Modern College of Engineering, Pune, India [2]

***Abstract*— In today's Internet world, web search engines such as Google, Yahoo, Microsoft Live Search, etc. are widely used to find certain information from a huge database in a minimum amount of time and with minimum effort. However, all these search engines also pose a privacy threat to the user. In order to address this privacy threat, we have proposed User customizable Privacy-preserving Search(UPS) approach. In this survey paper, we have briefly described the working of UPS with its possible challenges. In the literature survey, we have discussed single database PIR protocol,Private web Search, and privacy enhanced PWS with their advantages and limitations. Also we have explained different personalization approaches.**

***Keywords*: Personalized Web Search(PWS), Private Information Retrieval (PIR), Search EngineWeb Search, Wrapper**

I. INTRODUCTION

Search engines (Google,Yahoo etc.)have become one of the most important tools for ordinary people who are looking for useful information on the World Wide Web. However, many a times, people do not get the relevant information on their topic of interest. Instead, they get irrelevant information due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts.It is a challenging task to find the exact information relevant to the search since it is scattered around the World Wide Web.According to areputed survey, 84% of the Internet users usea web search engine at least once in a day. Among the different search engines, Google is the most used. Google improves its performance by simply storing a record of each visited site and also by storing the past web search history of each user.

Personalized web search (PWS) is a general category of search techniques which is capable to provide better search results on particular topic, which satisfy the user need. It is generally categorized into two types, viz. click-log-based method and profile-based method.

The click-log based method isquite simple and straightforward. This method simply tracks the clicked pages in the user's query history. The performance of this strategy is consistently and considerably well, but it can only work on repeated queries from the same user, which is also its strong limitation.

On the other hand, profile-based methods generally improve the search experience with complicated user-interest model. One can generate the user interest models from user profiling techniques. Profile-based methodis potentially effective for almost all types of queries, but with one important limitation - it is reported to be unstable under some circumstances.

In order to protect the user privacy in profile-based PWS, we need to consider two effects during the search process on particular topic. First, we have to improve the search quality with the personalization utility of the user profile. Second, we need to hide the privacy contents existing in the user profile to place so that we can have low privacy risk.

Generally, people are willing to compromise privacy if supplying user profile to the search engine generates better search quality and relevant information. But in an ideal situation, we can get a significant gain by personalization at the expense of only a small portion of the user profile. Thus, we can achieve the user privacy and also protect it without compromising the personalized search quality.

II. SYSTEM ARCHITECTURE

A. UPS-User Customizable Privacy-Preserving Search

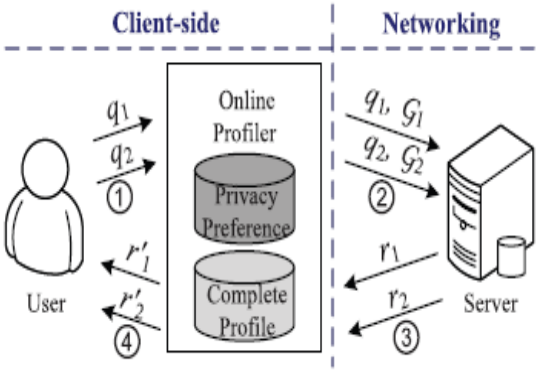

Figure 2.1: System Architecture of UPS-User customizable Privacy-preserving Search.

Figure 2.1 shows the system architecture of UPS. In this approach, we generally assume that the queries to be provided to the search engine do not contain any sensitive information. We aim to protect the privacy of each individual user profile while providing the best search quality.

UPS consists of a non-trusted search engine server and a number of clients(users). Each client (user) accessing the search service trusts no one. Online profiler is the keycomponent for privacy protection which is implemented as a search proxy running on the client machine. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified privacy requirements represented as a set of sensitive-nodes.

The framework works in two different phases, the offline and online phase, for each user.
In offline phase, first a hierarchical user profile is constructed and then customized with the user-specified privacy requirements.
In online phase, when a user issues a query, it has to go through the following steps:

1. When a user issues a query say, qi on the client, the proxy generates a user profile in runtime in thelight of query terms. This output is nothing but a generalized user profile Gi which satisfies the privacy requirements. The generalization process can be measured by considering personalization utility and the privacy risk - both these factors are defined for the user profiles.
2. The query and the generalized user profile generated in step 1 are sent to the PWS server.
3. Personalized search result with the profile is delivered to the proxy.
4. Finally, the proxy presents the raw results to the user, or it re-ranks the personalized search result with the complete user profile.

B. Advantages of UPS Over PWS:

1) UPS provides runtime profiling due to which it is possible to optimize the personalization utility while respecting user's privacy requirements.
2) UPS also allows for customization of privacy needs.
3) In UPS, we does not require any iterative user interaction.

III. LITERATURE SURVEY

A. Single-Database PIR Protocols

Private Information Retrieval (PIR) can be described as submitting a query to a web search engine while preserving the user privacy.Here, a user can retrieve information from the database but the server which holds the database does not have any knowledge about the data which is requested by the user. In this case, the server is nothing but the web search engine and the database can be treated as the web pages that the web search engine stores.
The first PIR protocol which was designed by Chor et al. [2, 3] isbased on several servers which holds the same database. These servers cannot communicate between them. The main drawback of this proposal is that it cannot work with a single server.

Pros and Cons of the Single-database PIR:

(1) Single-database PIR schemes are not useful for large databases. In PIR,the database is usually designed as a vector. Here, we can consider a scenario where the user wants to retrieve the value of the j^{th} component of the vector while keeping the index j hidden from the server. Let us assume that database contains n number of different items. A PIR protocol will access all the records in the database in order to find the value of j^{th} record. Here, we note that if a user has access onlyto a part of the database, it will be easier for the server to know the real interests of this user. The cost of accessing all records in the database is O(n).
(2) In PIR, when we are accessing any records in the database, it is assumed that the user must know their physical location. But this situation is not realistic because the database is not managed by the user.
(3) Here in PIR scheme, it is assumed that the server collaborates with the user but this assumption is false because the server has no motivation to protect or preserve the privacy of the users. Users themselves have to take care of their own privacy. Users cannot expect any protection of thier privacy from the web search engine.

B. PWS - Private Web Search

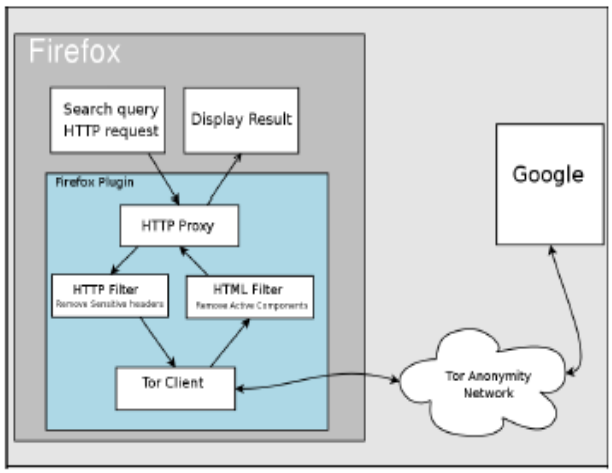

Figure 3.1: System Architecture of PWS

Figure 3.1 describes the system architecture of PWS [4]. PWS provides privacy to the user. It is nothing but aFirefox plug-in which runs a Tor Anonymity client and an HTTP proxy. When the user want to execute a query, a connection is made to the HTTP proxy. The proxy filters the HTTP request, then sends this request to the search engine (e.g. Google) over the Tor Anonymity network. Once the search is completed through the search engine, the proxy receives the response from Tor, then filters the HTML to remove all active components, and finally gets the answer back to Firefox for display.

Pros and Cons of PWS:

1. PWS gives better assurance of removing all the active components and undesired labels even if the user has written them while typing the query.
2. PWS generally gives control over the functionality to the user. There is a possibility that the functionality will break if some feature is disabled.
3. PWS can be used to minimize the information that users need to submit to a search engine.
4. PWS protects the users against various attacks that involve active components and timing information.
5. If changes are made to the search engine's HTML code, there is a possibility that the HTML filter may fail to produce an output.

C. Privacy-Enhancing Personalized Web Search System Architecture:

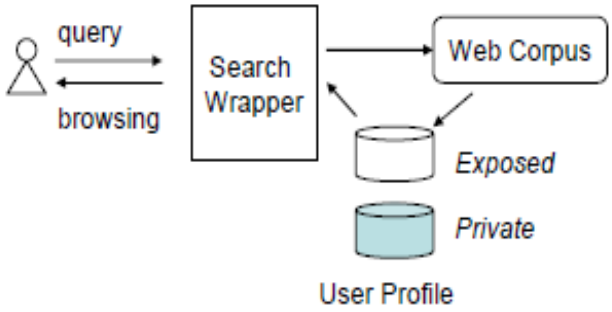

Figure 3.2: System Architecture of Privacy enhancing Personalized web search.

Figure 3.2 gives an overview of the proposed Privacy enhancing Personalized web search [5]. Here, the authors have provided an algorithm for the users to automatically build a hierarchical user profile that represents the user personal interests.

The general interests of user are put on a higher level - on the contrary, specific interests of the user are put on a lower level. In this work, authors have exposed only a little portion of the user to the search engine in order to protect the user privacy.

On the server side, the authors have developed a search engine wrapper which is used to incorporate a partial user profile with the results returned from a search engine. After getting the results from partial user profiles and search engine, these results are combined and ranked. Lastly, the customized results are delivered to the user by the wrapper.

The Privacy-Enhancing Personalized Web Search consists of three parts:

1 . Building a hierarchical user profile from available source data by using a scalable algorithm.
2. Offering privacy parameters to each user in order to determine the content and amount of personal information that will be revealed.
3. Using search engine wrapper to combine and rank the search engine results with the help of the partial user profile.

Advantages:

1. This approach provide a scalable way to build a hierarchical user profile automatically on the client side. This approach also give us an easy way to protect and measure privacy.
2. In this approach, one needs to grant the server full access to the user's personal information. This sometimes violates user privacy.

IV. DIFFERENT PERSONALIZATION APPROACHES

A.Social- Based Personalization

Nowadays, there is tremendous growth of social networks systems which has created a very large online repository of information. According to a survey, social networking sites such as Facebook, Twitter, LinkedInetc.together have a combined user base of more than 1000 million users all over the world. All these sites store important information about their users such as users' real names, email addresses, list of friends, personal pictures, audio, video and much more.

It has become a challenging task to implement privacy of userpersonalization in social networking for the following reasons:

1. Social networking sites includes highly sensitive information such as personal messages or pictures.
2. There is a need to focus on privacy of the users as well as their connections.
3. Leakage of users' personal information may cause embarrassment to them.

B. Profile Based Personalization

Profile based personalization is the collection of information about person's activities and their experiences based on those activities. This approach generally focuses on improving the search utility.

C. Location-Based Personalization

Location-based personalization is also considered in order to preserve the location information of user.

Recently, location aware services are becoming more popular and use widely. Their development has been triggered due to more advancement and adoption of GPS enabled mobile phones and Wi-Fi positioning technologies.

D.Client-Based Personalization

In client side personalization, users' can store their personal data only at the client side i.e. at their own computers,laptops or other mobile devices. In this approach, users' can have more control on their data and perceive less privacy issues and risks.

V. CONCLUSION

In this work, we have presented a client-side privacy protection framework called UPS for personalized web search. UPS can be easily adopted by any PWS that captures user profiles in a hierarchical taxonomy. This UPS framework allows users to specify customized privacy requirements via the hierarchical profiles. Here,it is also possible to perform online generalization of user profiles to protect their personal privacy without compromising the search quality.

REFERENCES

[1]. LidanShou, He Bai, Ke Chen and Gang Chen " Supporting privacy protection in personalized websearch" IEEE transaction on knowledge and data engineering vol:26 No:2 year 2014.

[2] R. Ostrovsky, W.E. Skeith III, "A survey of single-database PIR: techniques and applications", Lecture Notes in Computer Science 4450 (2007) 393–411.

[3] B. Chor, N. Gilboa, M. Naor, "Private information retrieval by keywords, Technical Report TR" CS0917, Department of Computer Science, Technion, 1997.

[4] F. Saint-Jean, A. Johnson, D. Boneh, J. Feigenbaum, "Private web search", in: Proceedings of the 2007 ACM workshop on Privacy in electronic society –WPES'07, 2007, pp. 84–90.

[5] Y. Xu, B. Zhang, Z. Chen, K. Wang, "Privacy-enhancing personalized web search", in: International World Wide Web Conference, 2007, pp. 591–600.

[6] Kalyani R Kshirsagar" Survey on Privacy Protection in Personalized Web Search" In : international Journal of Computer Technology &Applications,Vol 5 (6),1974-1977.

[7] X. Shen, B. Tan, C.X. Zhai, "Privacy protection in personalized search", ACM SIGIR Forum 41 (1) (2007) 4–17.

[8] J. Castellı´-Roca, A. Viejo, and J. Herrera-Joancomartı´, "Preserving User's Privacy in Web Search Engines," Computer Comm., vol. 32, no. 13/14, pp. 1541-1551, 2009.

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

[10] G. Chen, H. Bai, L. Shou, K. Chen, and Y. Gao, "Ups: Efficient Privacy Protection in Personalized Web Search," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information, pp. 615-624, 2011.