Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Towards Efficient and Precise Privacy Defence System in Personalized Search

Shiva Soni, M. Nirmala

Department of Computer Science and Engineering
Aurora Technological and Research Institute
Hyderabad, India

Abstract*—Personalized Web Search (PWS) is different from the normal web search. Personalized Web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. Privacy protection which model user preferences as hierarchical user profiles in PWS applications. A PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Various research works already has been done towards how to protect user profile and personalization on query, but very few authors discussed about the protection of query as query may also have sensitive information. We know that adversaries may attack based on query analysis which may be passed by user during the personalized search. In this regard we have proposed and implemented the method to protect the user query along with existing approach to protected user profile. Our goal is to complete the current query with 'k' other fake queries using logical OR propositions approach called obfuscator. After receiving the raw search results, we applied post filtering approach to eliminate the number of irrelevant results from the retrieved raw search results, which was introduced due to fake queries. After the result analysis we found that proposed approaches may provide the complete protection during the personalized web search. At the same time search result quality may be bit compromised which is ignorable.

***Keywords: -** Web search engine, Personalized web search, Obfuscator, Post-filtering*

I. INTRODUCTION

Data Mining is the process of extracting information from large data set and transform into data set which is understandable and useful. World Wide Web (WWW) is largest, commonly used and most accessible source of information. Day-by-day the web pages on Internet are growing rapidly. Web structures are large as well as sophisticated and users often miss the goal of specified queries or receive ambiguous and sometimes unwanted results when they try to navigate through them. The search engines finds out the relevant web pages according to the query specified by user. While searching users might experiences failure when search engines returns unwanted as well as irrelevant results that do not meet their searched query expectation. Such type of irrelevance of results is largely because of the enormous variety of user's contexts and backgrounds, the ambiguity of texts and some type of confusion of queries. Therefore, in order to provide better search result a general category of search technique Personalized Web search is used. In personalized web search, user information is collected and analyzed in order to find intention behind issued query fired by user. Although personalized Search has been proposed for many years and many personalization strategies have been investigated, it is still unclear whether personalization is consistently effective on different queries for different users, and under different search contexts.

There are mainly two types of personalized web search they are Click-log-based and Profile-based personalized web search.

A. Click-Log-Based Method

In this approach, personalization is carried out on the basis of clicks made by user. The data recorded through clicks in query logs, simulates user experience. The web pages frequently clicked by user in past for a particular query is recorded in the history and score is computed for particular web page and based on that web search results are provided. This method will perform consistent and considerably well when it is works on frequent queries. When a never asked query is entered by user; it will not provide any precise search results, which is the main drawback of this method.

B. Profile Based Personalization

The basic idea of these works is to tailor the search results by referring to a user profile, implicitly or explicitly which reveals an individual information goal. Many profile representations are available in the literature to facilitate different personalization techniques.

- *Lists / vectors or bag of words*: Earlier techniques utilize term lists/vectors or bag of words to represent their profile. It is the simple representation in information retrieval system. Here a text is represented as the bag of its words, disregarding grammar and even word order [8]. But it keeps multiplicity of those words. In each vector the second entry will be the count of that word.

- *Hierarchical representation:* Most recent works build user profiles in hierarchical structures. The reason is their stronger descriptive ability, better scalability, and higher access efficiency. Majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP, Wikipedia, and DMOZ and so on. Using the term-frequency analysis on the user data, the hierarchical profile can be build automatically also.[9]

Although there are pros and cons for both types of PWS techniques, the profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, 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. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life. Privacy issues are rising from the lack of protection for such data. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services [9].

II. BACKGROUND AND RELATED WORK

Personalization of web search is to carry out retrieval for each user incorporating his/her interests. For a given query, a personalized Web search can provide different search results for different users or organize search results differently for each user, based upon their interests, preferences, and information needs. There are many personalized web search algorithms for analyzing the user interests and producing the outcome quickly; User profiling, Hyperlink Analysis, Content Analysis and collaborative web search are some of the instances for that kind of algorithm.

Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Many recent research efforts have focused on this area. The profile creation involves, [6], (1) collecting information from users: All searches, for which at least one of the results was clicked, were logged per user. (2) Creation of user profiles. Two different sources of information were identified for this purpose: all queries submitted for which at least one of the results was visited and all snippets visited. Two profiles were created out of either queries or snippets. (3) Evaluation: The profiles created were used to calculate a new rank of results browsed by users. The average of this rank was compared with Google's rank. Many approaches create user profiles by capturing browsing histories through proxy servers or desktop activities through the installation of bots on a personal computer. These require participation of the user to install the proxy server or the both.

The user profile is constructed by observing the information from the browser web page cache. The user desktops contains large amount of personal data; richer profiles can be built using this data. Most of the prior efforts in creating user profiles use frequently occurring document words to represent the profile. The following problems may occur due to this kind of profile creation, [5]. (1) Irrelevant words ,(2) Polysemy and synonymy ,(3) Size of the profile, (4) The profile content may represent a mixture of recreational needs of the user, information and transactional.

In [5], creating a user profile using Wikipedia requires the following three steps. (1) Web pages are mapped to Wikipedia concept, (2) Hierarchical profile created from this concept. Concept in profile is tagged in two ways. First, whether the target is transaction or recreational. Second, how recent the users are in that topic.

In [1], users have to register personal information such as their interests, age, and so on, beforehand, or users have to provide feedback on relevant or irrelevant judgments, ratings on a scale from 1 (very bad) to 5 (very good), and so on. These types of registration, collecting feedback, or ratings are consumes some specific time and users prefer easier methods.

In [3] & [4], a search process involving many such interaction cycles, a user thus potentially reveals the following three kinds of personal information: 1. *User identity:* This could be a personal user ID in the case when the user has to register an account, or the IP address of the machine that the user is using. 2. *Queries:* This includes all the queries the user has submitted to the search engine. 3. *Viewed results:* This includes all the viewed web pages by the user.

The personalized web search is takes place in three ways: (1) Server side personalization, (2) Client side personalization and (3) client-server cooperative personalization.

A. Server side personalization

Level II privacy protection can be achieved. But when the search engine uses the user login ID to collect user information, this method will not achieve Level II privacy protection; when the search engine only uses the IP address to aggregate the user information, this method works. Sometimes, search engines group users randomly or according to some criteria before they release the search engine logs. Then we will also have Level II privacy protection to those third parties which receive the search engine logs. It is impossible to implement Level III or Level IV privacy protection if personalization is done on the server side.

B. Client side personalization

A client-side personalized search agent can do query expansion to generate a new query before sending the query to the search engine. The sensitive contextual information is generally not a major concern since it is strictly stored and used on the client side and the overhead in computation and storage for personalization can be distributed among the clients. A main drawback of personalization on the client side is that the personalization algorithm cannot use some knowledge that is only available on the server side (e.g., PageRank score of a result document).

C. Client-server cooperative personalization

The user profile is still stored on the client side, but the server also participates in personalization. When a query is given to the search engine then the client extracts contextual information from the user profile. The combination of extracted information from the profile and query is sends it to the search engine. The search engine then does personalization with the received context. The contextual information sent to the server specifies the user's search preferences (e.g., query expansion terms, topic weight vector). This architecture provides the same level of privacy protection as server-side personalization. However, the personally identifiable information collectable on the server side is less than in the case of pure server side personalization.

III. PROBLEM STATEMENT

Earlier work done by authors [9] has discussed about how to protect user profile and personalization on query, but they have not discussed about the protection of query. We know that adversaries may attack based on query analysis which may be passed by user during the personalized search. In this regard we are proposing the method to protect the user query along with existing approach which is to protected user profile. Our goal will be to complete the current query with k other fake queries using logical OR propositions. After receiving the raw search result, a post filtering approach may be used to eliminate the number of irrelevant result from the raw search result, which is introduced due to fake queries. Hence above proposed work may provide the complete query protection during the personalized web search.

IV. EXISTING AND PROPOSED SYSTEM

In above problem statement, we have identified the requirements of the modifying the existing framework based on the proposed solution by the author for personalized web search and the shortcoming in the existing solution. In this section we are explaining in brief about the existing system, and our proposed enhanced system.

A. Existing system

Earlier existing profile-based Personalized Web Search model do not support runtime profiling. A user profile 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.

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. 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. However, this assumption can be doubted with a simple counter example: If a user has a large number of documents about “sex,” the surprisal of this topic may lead to a conclusion that “sex” is very general and not sensitive, despite the truth which is opposite. Unfortunately, few prior works can effectively address individual privacy needs during the generalization.

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, we need predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

Recently few authors [9] proposed a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, author formulated the problem of privacy-preserving personalized search as Risk Profile Generalization, with its NP-hardness proved.

Author developed two simple but effective generalization algorithms to support runtime profiling. While the former tries to maximize the discriminating power (DP), the latter attempts to minimize the information loss (IL).

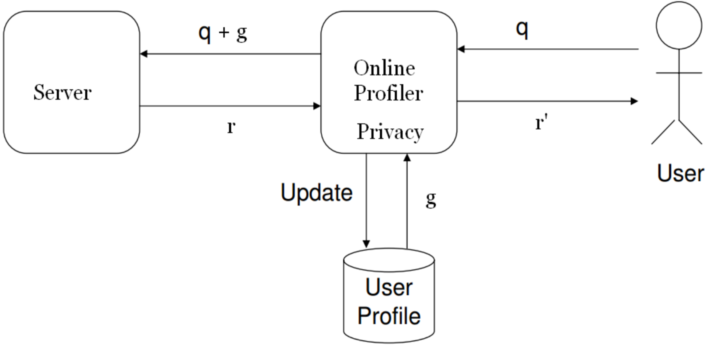

Figure1: Author proposed and adopted architecture

Author also provided an inexpensive mechanism for the client to decide whether to personalize a query in UPS. This decision can be made before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile.

Advantages:

1. It enhances the stability of the search quality.
2. It avoids the unnecessary exposure of the user profile.

Disadvantages:

1. In the above work done does not concentrate on the protection of user query which is floating over the network.

B. Proposed system solution

In our proposed enhance system architecture; we have proposed two additional approaches along with the current work done in paper [9]. In proposed enhanced architecture we added obfuscator for fake query generation along with original query forwarded by search user and post filtering approach. Brief description about those approaches we have given in the next section. Below is our proposed enhanced architecture for personalized web search.

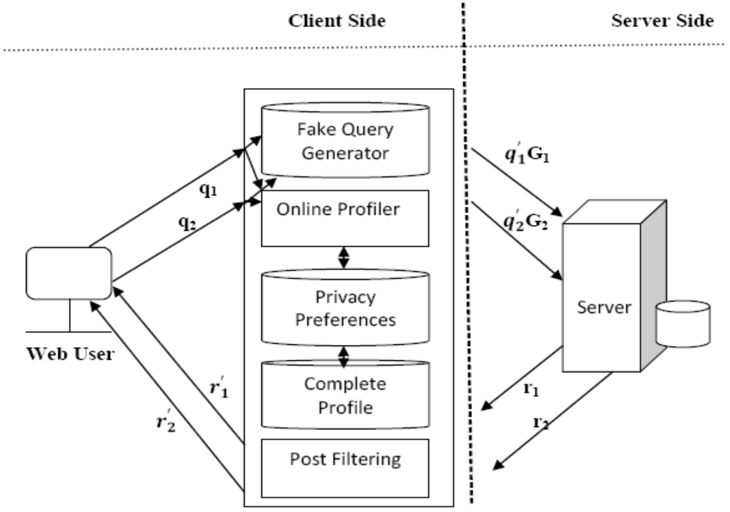

Figure2 Proposed Enhanced System architecture of UPS-PWS

V. MODULES DESCRIPTION AND IMPLEMENTATION

The Enhance proposed architecture for the UPS in PWS consist of various modules that are profile-based personalization, privacy protection in PWS system, generalizing user profile, online decision, attack controller, online fake query generation/Obfuscator, and post filtering. Explanation of modules functionality in detail is given below.

1. Profile-Based Personalization

In this module we introduce an approach to personalize digital multimedia content based on user profile information. A user profile is created explicitly. In the explicit manner the user need to interact with the system to create the user profile manually. The user need to input some own details (eg: Name, age, designation, area of interest, etc.,).The two main mechanisms were developed: a profile generator that automatically creates user profiles representing the user preferences, and a content-based recommendation algorithm that estimates the user's interest in unknown content by matching her profile to metadata descriptions of the content. Both features are integrated into a personalization system.

2. Privacy Protection in PWS System

We propose a PWS framework called UPS that can generalize profiles in for each query according to user-specified privacy requirements. Two predictive metrics are proposed to evaluate the privacy breach risk and the query utility for hierarchical user profile. Two simple but effective generalization algorithms are being used for user profiles allowing for query-level customization using proposed metrics. Also provided an online prediction mechanism based on query utility for deciding whether to personalize a query in UPS. Extensive experiments demonstrate the efficiency and effectiveness of framework.

3. Generalizing User Profile

The generalization process has to meet specific prerequisites to handle the user profile. This is achieved by preprocessing the user profile. At first, the process initializes the user profile by taking the indicated parent user profile into account. The process adds the inherited properties to the properties of the local user profile. Thereafter the process loads the data for the foreground and the background of the map according to the described selection in the user profile.

Additionally, using references enables caching and is helpful when considering an implementation in a production environment. The reference to the user profile can be used as an identifier for already processed user profiles. It allows performing the customization process once, but reusing the result multiple times. However, it has to be made sure, that an update of the user profile is also propagated to the generalization process. This requires specific update strategies, which check after a specific timeout or a specific event, if the user profile has not changed yet. Additionally, as the generalization process involves remote data services, which might be updated frequently, the cached generalization results might become outdated. Thus selecting a specific caching strategy requires careful analysis.

4. Online Decision

The profile-based personalization contributes little or even reduces the search quality, while exposing the profile to a server would for sure risk the user’s privacy. To address this problem, authors [9] developed an online mechanism to decide whether to personalize a query. The basic idea is straightforward. if a distinct query is identified during generalization, the entire runtime profiling will be aborted and the query will be sent to the server without a user profile.

5. Attack Controller

The attack controller is used to control query attacks. In this module the attacker details may be identified and the content attacked by malicious user. Once it is being identified the content recovery process may be initiated to recover from the attacked content. Session information is protected to control session based attacks.

6. Online Fake Query Generation

In order to protect queries that are attached with user profile generated by the online profiler for the given query, we need to modify the original query. We choose a basic functionality of search engines that allows the use of logical propositions in the query. Our goal is to complete the current query with k other fake queries using logical OR propositions. The challenge here is to generate fake queries that cannot easily be identified as such by the search engine. Toward this purpose, we aim at generating fake queries that are far from the user profile. The fake queries may be generated using information contained in the user profile or from dictionary. It takes words contained in the dictionary but not in the user profile. Consequently, obfuscated query, forged by combining the original query with fake queries [10]. This strategy misleads the adversary about the real identity of the requester query.

7. Post-filtering

The goal of the post-filtering is to remove all irrelevant answers introduced by the fake queries. A basic algorithm will analyze all the results returned by the search engine and remove results that contain words generated during the obfuscation step [10]. These irrelevant results are returned as an answer of the fake queries. Consequently, there are not interesting for the user.

Approach for Fake Query Generation and Post-Filtering:-

Require: The set (Dictionary) q which contains the query Q

```
Step1: while (QueryProtection)
Step2: q’← q U GenerateObfuscatedQuery()
Step3: end while
Step4: Send ToSearchEngine(q’) along with user profile U
Step5: R← GetResult()
Step6: for all result ∈ R do
Step7:  for all word ∈ result do
Step 8: if word ∈ q \{Q} then
Step 9: ELIMINATE (result)
Step10: break
Step11: end if
Step12: end for
Step13: end for
```

VI. RESULT AND DISCUSSION

We have implemented the proposed solution. We performed the execution initially with no query protection and then after we performed the execution using the query protection approaches. In query protection approach first we executed the obfuscation steps and found that after execution we received the raw result which comes along with original query results. We executed the post filtering steps and found that in the final results only those results are available for which query is forwarded by user. We have also computed the greedy discriminating power (gdp) and greedy information loss (gil) for the given set of user query with without query protection and with query protection and after analysis we realized that protection of query during the personalized search slightly decreases the greedy discriminating power and increases the information loss. Which may be ignorable, If query protection in necessary for user during the personalized search.

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system’s relationship to help user decision-making.

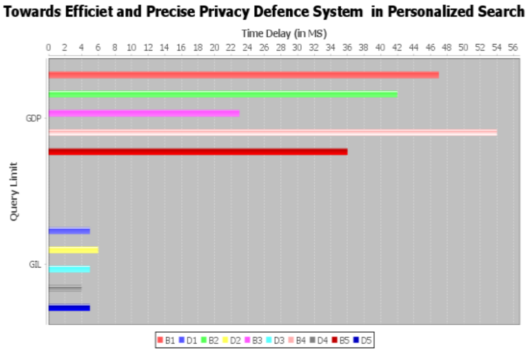

Figure:3 TEPPDS-PWS without Query Protection

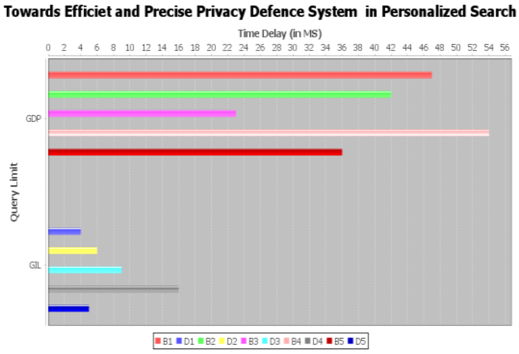

Figure:4 TEPPDS-PWS with Query Protection

VII. CONCLOUSION AND FUTURE WORK

Personalized web search (PWS) is used to improve the quality of various search services on the Internet. Privacy preserved PWS methods are used to protect the disclosure of personal information in search process. User customizable Privacy-preserving Search (UPS) framework is used to support privacy in search process.

In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. We used the existing greedy algorithms for the online generalization. We performed the query protection along with the user profile protection. We have used the obfuscation approach to protect user query. We performed the post filtering approach to get the user query results. We analyzed the experimental results and found that query protection approach performed better and preserving the query privacy. At the same time we found that it may bit compromised with the quality of search results after filtering. The quality of search results may further improve with various result reranking approaches.

REFERENCES

[1] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa," Adaptive Web Search Based on user Profile Constructed without Any Effort from Users" in WWW2007.

[2] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[3] A. Kobsa, "Privacy enhanced personalization", CACM vol. 50, no. 8, August 2007.

[4] Y. Xu, B. Zhang, Z. Chen and K. Wang, "Privacy enhancing-personalized web search", 16th Int'l conf. world wide web (WWW) pp. 591-600, 2007.

[5] Krishnan Ramanathan, Julien Giraudi, Ajay Gupta "Creating hierarchical user profiles using Wikipedia," inHPL-2008-127.

[6] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence(WI), 2005.

[7] Xuehua Shen, Bin Tan, ChengXiang Zhai, "Privacy Protection in Personalized Search", in ACM SIGIR Forum Vol.41 No.1 June 2007

[8] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search via Automated Analysis of Interests and Activities," *Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR),*pp. 449- 456, 2005.

[9] 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, February 2014.

[10] Albin Petit, Sonia Ben Mokhta, Lionel Brunie, "Towards Efficient and Accurate Privacy Preserving Web Search",MW4NG 14-Dec 8-12,2014, Bordeaux,France,2014ACM 978-1-4503-3222-4/14/12