International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 7,July 2015

Privacy Security in Personalized Web Search

[1] Mrs. Priyanka Deulkar,[2]Dr.A.D.Gawande
[1][2] M.E.2nd yr Sipna college of Eng & Tech Amravati , HOD Dept Of Comp Science Sipna College Of Eng & Tech,[1]priyanka_deulkar@rediffmail.com [2]adgawande@rediffmail.com

Abstract- **One user hundred needs on internet. As more and more queries are being searched on the web, it is increasingly difficult to let the search engine know in what context user wants to search. Coping with ambiguous queries has long been an important part in the research of Information Retrieval, but still remains to be a challenging task. Personalized search has recently got significant attention to address this challenge in the web search community, based on the premise that a user's general preference may help the search engine disambiguate the true intention of a query. In this paper, we implement an algorithm that returns relevant results to users based on their preferences keeping sensitive data more secure. Our experiments show that user's sensitive preferences can be preserved accurately from attacks. Our scheme provides an affordable overhead while offering privacy benefits to the users.**

Keywords--**Category, Personalized Web Search, Sensitive User Profile**

I. INTRODUCTION

As the amount of information on the Web increases rapidly, it creates many new challenges for Web search. When the same query is submitted by different users, a typical generic search engine returns the same result, regardless of user's intention for that query. This may not be suitable for users with different information needs. For example, for the query "caterpillar", some users may be interested in documents dealing with "caterpillar" as "insect" , while other users may want documents related to caterpillar products. One way to disambiguate the words in a query is to associate a small set of categories with the query. For example, if the category "animal" or the category "insect" is associated with the user of query "caterpillar", then the user's intention becomes clear. To solve these problems, web search engines need to be personalized. Personalized systems address the overload problem by building, managing, and representing information customized for individual users. This customization may take the form of filtering out irrelevant information and/or identifying additional information of likely interest for the user. Research into personalization is ongoing in the fields of information retrieval, artificial intelligence, and data mining, among others.

In this paper our personalization relies on rich user profiles. User profile is data instance of a user model which can be applied to adaptive interactive systems. To receive personalized web services, the user has to provide personal information and preferences, in addition to query to web service. So these user profiles ,description of user interest can be used by search engine to provide personalized search results. In [1] Susan has mentioned approaches for user identification as software agents, login, enhanced proxy servers, cookies and session id. We are using login method which is reliable and user can identify themselves during login, the identification is generally accurate.

Personalization is the process of deciding - given a large set of possible choices what has the highest value to an individual. Personalized web search (PWS) tailors the search experience specifically to match user's interest by incorporating the information about the individual beyond the specific query. Lidan [2] has described this as a search technique category which gives relevant search results differently for each user, according to user's interest. The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods have strong limitations as they work on repeated queries only from the same user In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling 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 [4],[5],browsing history, click-through data[6], bookmarks [7], user documents, and so forth. But the profile-based methods become more unstable when users search history grows. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life [8]. However, the detailed personal information provided for personalized web service could identify the sender of sensitive queries, thus compromise users privacy. So to overcome this privacy threat we are asking user to specify sensitivity for the selected interest which they do not want to expose.

To better understand how a user may be identified in personalized web services, let us consider concrete example. Suppose that the user xyz submits a query q={insects, study} to a vertical search engine like animal-

knowledge, which provides specialized search on insect information. Wishing to get personalized results, user xyz registered his personal information d on date of birth, gender, zip code as required in the online registration form. Each query leaves a trace <u,q,t> on the site's query log. In our implementation, we are encrypting the user's identity by secure random number . As a secondary use, the query log is published to travelling company for data mining research. In the following discussion, the attacker refers to a party that has access to the query log and seeks to re-identify the (sensitive) queries of xyz, called the target. Usually, the attacker has some sort of relationship with xyz, e.g., colleagues, neighbors, friends, enemies, etc.

Consider the following two ways of re-identification.

- Re-identification through personal information
- Re-identification through approximate query time

In the above search scenario, a personalized query has two parts <u,q>. The query q contains query terms on which the user wants to get results. This part is unstructured and contains sensitive information, meaning that the user does not want to be identified as the sender of the query. The personal information <u>contains user data and other preference information.

To secure privacy, we introduce an untrusted third party called key pool, which assigns unique key to individual user. So instead of sending <u,q> to the web service directly, the user first generates u' through the key pool and then sends <u',q> to the web service, where u' is some generalization of u. The web service possess the query q and the generalized personal information u', but cannot identify user from u' because u' has been generalized.

II. LITERATURE SURVEY

Many research has been done on Personalized Web Search (PWS). But existing approaches do not take into account the sensitivity of data. The personalized privacy protection concept was first introduced by Xiao in Privacy Preserving Data Publishing. In paper [2] a new personalized approach has been developed that uses online decision on the query personalization. The previous work on PWS mainly focuses on user profile to provide personalized search results to individuals.

Profile-Based Personalization

In order to construct an individual user's profile, information may be collected explicitly or implicitly. The basic idea of [2],[3] works is to tailor the search results by referring a user profile that reveals an individual information goal. 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[14] or bag of words[9] to represent their profile. However, most recent works build profiles in hierarchical structures. The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP[7], Wikipedia[10] or so on. In Paper [8] F.Qiu proposed a framework to investigate personalized web search problem and learn user's topic preference vector on user's earlier history without user intervention. Basically he proposed different user models to formalize user interests on web pages and then correlate them with user's clicks on search result. Based on this correlation they described algorithm to fetch user interest. But this solution is not so feasible as user's interest will no longer be private.

Our implementation is based on hierarchical technique which can identify user on login basis which is most reliable method to identify user.

Securing Privacy in PWS

Different users have completely different requirements to secure their privacy. Thus the level of privacy protection may need to be tuned for different users to accommodate different preferences for the tradeoff of personalization and privacy protection. Paper[11] has defined and analyzed different levels of privacy protection in PWS.

One main drawback of existing work is that they build 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.[12] 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 sub-tree of the complete profile. Unfortunately, this work does not address the query utility, which is crucial for the service quality of PWS. Paper [13] has examined 3 processes of mapping a new user query to a set of categories as User profile only, General Profile Only, both user and general profile. For comparison, our approach takes both the privacy requirement and the query utility into account.

Contribution

Our main contributions are summarized as following:

- We implement a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements.
- We also generate logs of user searches and clicked

results for future research studies. When generating log files we are securely encrypting personalized key with random number for security purpose.

- We implement the solution where users themselves are able to set their own privacy levels for user profiles to improve the search quality.
- We develop new algorithm *Topic-Similarity* which personalizes results as per user profile.

III. IMPLEMENTATION

Main objective of this research aims at securing the privacy in individual user profiles whereas retaining their usefulness for PWS. In the existing framework, researchers have not focused on user identification as in user can be easily identified by his/her identity. And also we generalizes the search result as per users interest and generate query log with randomly generated key.

We are implementing the concept of personalized anonymity i.e. a user can rate sensitive topics as either 0 means less sensitive or 1 means highly sensitive. We implement PWS framework called UPS (User customizable Privacy-preserving Search) that can adaptively generalize profiles by queries with respecting user specified privacy requirements in formal way.

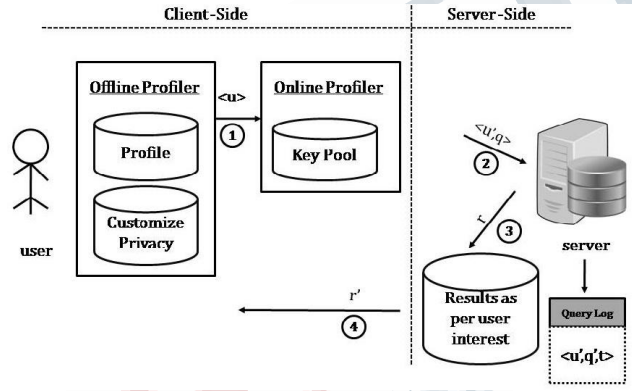

Figure 1: System Architecture Of UPS

In above Fig. UPS consists of a non-trusty search engine server and a number of clients. Each client accessing the search service trusts no one but himself/herself. The key component for privacy protection is an offline and online profiler implemented as a search proxy running on the client machine itself. The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows:

Before user submits a query q on the client, the online profiler generates a unique key for the user and generalizes user profile u'.

The query and the generalized user profile <u',q> are sent together to the PWS server for results. The search results are personalized with the profile and delivered back to the query proxy. and also generates encrypted query log and update generated user profile with the search history.

Procedure

In Paper[2],[3] Generalization procedure works in four steps. We are implementing the same procedure in different manner as follows.

Offline -1 User Profile Construction:-

Personalization is playing an increasingly important role in creating better Internet experiences. An important aspect of personalization is creation of a user profile. The user profile could be created on the client PC or on an Internet server. Both these methods have different advantages. Client side profiles offer better privacy, a more complete view of the user data. Server side profiles enable collaborative filtering and profile portability.

The goal of User-profile based personalization [1] is to collect information about the subjects in which a user is interested.

Offline-2 Customize Privacy Requirement:-

User profiles are build in hierarchical manner. A user may associate one or more categories to his/her profile manually. For example, a user may first select one or more categories in the hierarchy before submitting his/her query. By utilizing the selected categories, a search engine is likely to return documents that are more suitable to the user. Customized privacy requirements are given by specifying a number or sensitivity for sensitive topics in their profile which user don't want to disclose. This process of specifying the sensitivity to topic is called as forbidding[3]. In our work, When user selects sensitive topic from taxonomy he/she has to rate sensitivity for selected topic as either 0 or 1. Higher the sensitivity, more sensitive it is and will not be generated either in generalized user profile or in log files.

Online-1 Query-Topic Mapping:-

After submitting query q, we retrieve the documents similar to q using conventional approach. These documents are then grouped together. The relevance method used in this framework is simple and fast to evaluate. We propose a novel technique to map a user query to a set of categories, which represent the user's search intention. We have developed new algorithm that searches query with user's specified category (if any). This set of categories can serve as a context to disambiguate the words in the user's query.

Online-2 Profile Generalization:-

Specifying a sensitive node is not enough for privacy protection, as rooted subtree may still leave to severe privacy disclosure. A generalized profile is a rooted subtree obtained in effect by removing a node set S(U) from Topic. The output of forbidding on the profile only releases its parent node. This procedure generalizes the seed profile G_0 in a cost-based iterative manner relying on the privacy and utility metrics. In addition, this procedure computes the discriminating power for online decision on whether personalization should be employed. The generalization technique can seemingly be conducted during offline processing without involving user queries.

Below table summarizes all the symbols used here:

Table I: Symbol Used

Symbol	Description
t	Topic/Category
$s(t)$	The sensitive topic
u'	Generalized User Profile

The workflow for generalizing user profile of our implementation is shown below :

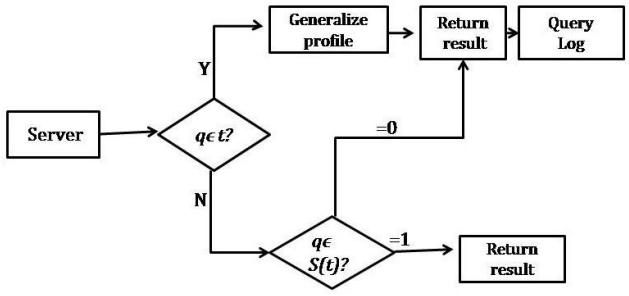

Figure 2: Workflow for personalizing search result

After submitting query if query is among the selected category or topic and not sensitive then the user profile will be generated and results will be returned to user. When user selects or clicks on any of the given link then the query log will be generated with the tuple as

$$\langle ID, key', query, time, clickedURL \rangle$$

Where *ID*= User Id,
key'= encrypted key, This key is securely encrypted for every query (by same user also),
query = query user has asked for,
time = time at which query was fired,
clickedURL= *URL name* user has visited After getting results.

If the query is sensitive which user do not want to disclose his identity (during offline-2 phase) then system will check the sensitivity for that topic. If it is 0 (less sensitive) then still it will not be generalized in profile but will be written in query log for further research. But if it is marked as highly sensitive by rating it as 1 then it will be completely hidden from generalized user profile as well as from query log.

To achieve this privacy we are using cryptographic tools for key generation and key encryption.

ALGORITHM

To implement security and receive relevant results we have developed new algorithm as *Topic-Similarity* which returns personalized relevant results as per user topic selected. The algorithm is written below. If user issues query which belongs to topic domain repository *R* then it iterates over all the selected topics and its parents (line no 3-4) and get documents related to category. It also checks if the query belongs to sensitive topics denoted as $(s(t))$ (line no 5). Suppose, if query is not indexed as topic then it will search in whole database for the documents and accordingly generalizes user profile.

Algorithm: Topic-Similarity(u,q,t)

Input: A seed profile u, topics t, Query q
Output: A generalized user profile u', search result
1. Let R be the topic repository;
2. if *u* has $t \in R$ then
3. Iterate over the selected topics,
4. get parent category;
5. if $(q \in s(t))$,
6. if $(s(t) = 0)$,
7. insert <u,q> into log,
8. fetch documents,
9. return result;
10. else if $(q \in t)$ then
11. update user profile,
12. repeat 8-9;
13. else if u has not $t \notin R$,
14. repeat 11-13.
15. return u';

In existing framework[2],[3], GreedyIL (Greedy Information Loss) and GreedyDP(Greedy Discriminating Power) algorithms are used. GreedyDP bounds the search

space to the finite-length transitive closure of prune-leaf. It also requires much more recomputation of DP, which incurs lots of logarithmic operations. The problem worsens as the query becomes more ambiguous. In contrast, GreedyIL incurs a much smaller real-time cost, and outperforms GreedyDP by two orders of magnitude. The major advantage of this algorithm is that it do not require any prune-leaf operation.

EXPERIMENTS

We have experimented *Topic-Similarity* algorithm for UPS framework by conducting three experiments on UPS. In the first experiment, we study the detailed results without any topic selected by user. Second, we look at the effectiveness of the proposed query-topic mapping. And last experiment is for query with sensitive interest.

Data Set

For the experiment we have partitioned data set into three categories. For first two categories we are using "caterpillar" query as example.

Query without Interest specified

If use has not customized his requirement before issuing query then the user might get all results (unstructured) from server and his searched query will be updated in his profile.
Suppose if such user submits query as "caterpillar" then server will return all results about caterpillar (i.e. caterpillar insects , caterpillar shoe, caterpillar products etc).

Query with Interest

When user specifies his interest before issuing query he gets relevant result. In our example, if user selects interest as "root/hobbies/shopping" then on submit of "caterpillar" query he will get results related to "caterpillar shoes" followed by other caterpillar results. Likewise if another user selects interest as "root/Animals" then after submitting "caterpillar" query, he will get result for "caterpillar insects" followed by other caterpillar results. After getting result, when user clicks any of the given link, then clicked URL will be updated in query log.

Sensitive Queries

For sensitive queries, user has to make preferences as either 0 or 1. On issuing such queries, server would return result without generalizing profile and only 0th sensitive topic will be added in query log.

Considering, "Adults" as sensitive topic and if user searches anything related to this topic (which he is not willing to expose his identity) then neither that user's profile nor his searched log will be generalized.

CONCLUSION

To give relevant results according to user's interest with keeping his sensitive data secured, we have developed a simple algorithm. Our framework relies on profile based personalization and allows users to specify customized privacy requirements via the hierarchical profiles by specifying sensitivity. In addition, UPS performs online generalization on user profiles to protect the personal privacy without compromising the search quality. We have also used cryptographic tools to preserve security. Our experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirm the effectiveness and efficiency of the solution.

REFERENCES

[1] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli, "User profiles for personalized information Access" Springer LNCS 4321.

[2] 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.

[3] Lidan Shou, He Bai, Ke Chen, and Gang Chen, "Supporting Privacy Protection in Personalized Web Search," vol . 26, no. 2, Feb 2014.

[4] Hoashi, K., Matsumoto, K., Inoue, N., Hashimoto, K.: Document Filtering method using non-relevant information profile. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, July 24-28 (2000) 176-183.

[5] Kim, H., Chan, P. "Learning implicit user interest hierarchy for context in personalization. In: Proceedings of IUI' 03, Miami, Florida, January 12-15 (2003) 101-108.

[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] B. Tan, X. Shen, and C. Zhai, "Mining Long-Term Search History to Improve Search Accuracy," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2006.

[8] F. 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.

[9] J. Pitkow, H. Schutze, 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.

[10] Z. Dou, R. Song, and J.-R.Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590, 2007.

[11] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[12] 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

[13] Liu, F., Yu, C., Meng, W.: Personalized web search by mapping user queries to categories. In: Proceedings CIKM'02, Mclean, Virginia, November 4-9 (2002) 558-565.

[14] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

PUBLISHED PAPER

[1] "A Study On Securing Privacy In Personalized Web Search" in International Journal Of Engineering And Computer Science Volume 4 Issue 3 March 2015 .