www.ijecs.in
*International Journal Of Engineering And Computer Science ISSN:2319-7242*
*Volume 4 Issue 3 March 2015, Page No. 11100-11103*

# A Study on Securing Privacy In Personalized Web Search

***Priyanka Deulkar [1], Dr.A.D.Gawande[2]***

[1]*Sipna College Of Engineering & Tech , SGBAU university, India*
*priyanka_deulkar@rediffmail.com*

[2]*Professor, HOD, Dept. of Computer Science,*
*Sipna College Of Engineering & Tech , SGBAU university, India*
*adgawande@rediffmail.com*

**Abstract:** ***Millions of users wants to get some information on web search engines. The growing use of search engines enables us to mainly describe the information that we seek. However, the major pitfall of generic search engine is that they returns the same list of results to user which can be irrelevant for users need. To address these problem, personalized search is considered to be encouraging solution as it provides relevant search results as per users information need and interest. We study securing privacy in PWS which captures user personal information and generates user profile and outputs relevant list of results. For web searching, user profiles are must for effective results. But the use of this profile to find interest is a breach to secure privacy. To conquer this issue, securing privacy is necessary. Hence, we study the existing methods for security of privacy in personalized web search and its efficacy.***

**Keywords** : Personalized Web Search, User Profile, Privacy, Generalization.

## 1. Introduction

Search Engines are the miracle of internet. The amount of information on the web is increasing day by day. With the volume and scope of information made accessible by search engines, people use search engines to investigate or address nearly every aspect of life. Regardless of the growing use of internet, many search engines produce list after list of irrelevant links of web pages. When same query is submitted by different users, a generic search engine outputs the same result, nevertheless of user information. This irrelevance is mostly because of extensive variety of users indeterminate query. Let's see, for a query "*kingfisher*" one user wants to travel by this flight or wants to compare fare details with others while other user may use same query to get information about kingfisher bird. So different users may use exactly the same query to get the information. So generic search engine is unable to distinguish such cases and returns the same results to all users. To solve these problems , web search engines need to be personalized.

Personalization is the process of deciding - given a large set of possible choices - what has the highest value to an individual. This adds both utility and warmth to a web application, as users find what they seek faster and feel "recognized" by the sites. Personalized web search (PWS) tailors the search experience specifically to match user interest by incorporating the information about the individual beyond the specific query. Lidan [1] has described this as a search technique category which gives relevant search results differently for each user, incorporating their requirement, interest. It utilizes user information and search context in learning to which sense the query refers. There are variety of applications to which we can apply personalization and variety of devices on which this personalized information can be delivered. Many personalization systems are based on some types of user profiles, which is a data instance of user model. It may include demographic information such as name, age, email etc and may also contain area of interest. In order to construct user profile, information may be gathered explicitly or implicitly. Different techniques can be implied to build user profile.

## 2. Literature Survey

Now we overview the pre-work which is done on PWS and user profile. Various implementations and research has been done on PWS. The personalized privacy protection concept was first introduced by Xiao in Privacy Preserving Data Publishing. In paper [1] a new personalized approach has been developed that uses online decision on the query personalization. This approach overcame previous problems and now supports

1. Online profiling. This allows separate user profile and hence improves search results.
2. Considers customization of privacy requirement, and also supports personalized anonymity.
3. Do not incur iterative user interaction while creating search results .This pose a new challenge in terms of efficiency.

To fulfill these features , the framework customizes user generalization on user-level and query-level which allows user to add his/her requirement and varies the generalization based on the query contents respectively.

Gang in [2], proposed a UPS (User customizable privacy preserving search) framework in which proxy generalizes user profile according to user specified privacy requirement and query content while submitting query . Then query along with this generalized user profile is sent together to PWS server. The results are personalized with profile and delivered back to proxy which then shows results to user. The key facility of securing privacy is online generalizer which maintains offline and online phases.

Paper[1],[2] focuses on the literature of *profile-based personalization* which improves the search utility and *privacy-protection in PWS*. The aim of user profiling is to collect the information about the user and his/her interests. In related paper [3], user profiling process consist of Data Collection, Profile Constructor and Application or Algorithm which utilize the user profile information in order to provide personalized services. User identification is crucial for system that constructs user profile. Many research has been done on constructing this user profile in different manner such as Weighted Keyword profile, Semantic Network profiles, Concept profiles [3], Search History, and many more. Yabo Xu in [4] has implemented Split and BuildUP algorithms to construct user profile hierarchically. Through this profile, user can control the information which will be publicly available to server.

Different users have completely different requirements, so for that the level of privacy protection need to be used to accommodate preferences for the trade-off between profile-based personalization and privacy-protection. X. Shen in [5] simplified securing privacy on the levels as pseudo-identity, group identity, no identity, no personal identity. To secure privacy, sensitivity is very important. He simplified the privacy concern in web search. The searching takes into consideration the communication between user (U) and search engine (S) with

- *Search*: U submits query (q) to S and S would return search results $R=\{R_1,R_2,...,R_n\}$to user.
- *Browse*: A U selects to view result $R_i \in R$, and then S would take U to the content of Ri.

In this process, a user thus shows *user identity* which could be his personal ID or IP address, *Queries* which includes all the queries he has asked for and *viewed results* that includes all viewed pages by him. X. Shen has also described and analyzed software architecture of PWS systems from securing privacy perspective. And shown that while securing privacy, client-side personalization is efficient over existing server-side personalized search services as former constructs richer user model for personalization. While Ji-Rong in [6] mentioned that Google personalized search uses server side personalization. But it requires high computation cost as well as raises the concern of privacy as user information is on server side. Though client side personalization distributes the overhead of computation cost and storage among clients, its downside is that this algorithm cannot use knowledge which is available on server side.

Feng Qiu [7] proposed a framework to investigate personalized web search problem and learn topic preference vector on users earlier history without user intervention. Basically he proposed different user models to formalize user interests on web pages and then correlate them with users clicks on search result. Based on this correlation they described algorithm to fetch user interest. But this solution is not so feasible as users interest will no longer be private. Another approach implemented in [8] uses similarity function that checks how strongly two words are related. As related words are fairly close to each other than unrelated words, they assumed two words co-occurring within window size are related. Window size parameter specifies the maximum distance between a pair of words for consideration of co-occurrence.

## 3. Problem Statement

This study paper focuses on problem of privacy protection in various aspect. As we surveyed, Paper [2] implemented UPS which allows user to specify customized privacy requirement. User information is analyzed to infer the user intention behind the issued query. Two Greedy algorithms are also proposed for online generalization. GreedyUtility is used to find optimal generalization in view of computational hardness. This algorithm can maximize the query utility while maintaining the exposure probability below user specified threshold. Gang proposed another algorithm, GreedyPerformance which is only used in experiment to analyze the trade-off between risk and search quality. An online prediction mechanism is provided for deciding whether personalizing a query is beneficial. Utility of personalization and the privacy risk of exposing the generalized profile metrics are used to analyze the system. The following drawbacks are identified from the existing system.

- Session based query attacks are not handled.
- Query weightage is not considered.

Privacy Preserving problem is also focused in [9] but in completely different manner through graph based model. The researchers have proven that the personalized protection implementation problem is NP-hard even with simple optimal objectives.

## 4. Existing Methods for generalization implementation

Many research has been done on the generalization of privacy of sensitive data. Xiao kui , Yufei Tao in [10] presented generalization framework on personalized anonymity. Their technique considers customized privacy requirement which prevents privacy intrusion and results in
generalized tables. Gang in [2] presented generalization in UPS. This implementation is done in *offline* and *online* phases.

- During the offline phase, A user profile is built in hierarchical manner by collecting information such as name, age etc. from user. There are various techniques implemented for user profile creation. paper[11] points to Rocchio-Based methods to learn user profile.
- In next stage of the same phase, user can customize his/her interest and can also specify sensitive topics which user is not keen to expose. Persuade by this, in our proposed system we allow user to specify the rating for his/her sensitive topic. Higher the degree more sensitive it is. When the offline phase is finished, we have received the complete user profile with customize requirements.

When user submits query q, it goes into the online processing which consists of two steps.

- In first step of online phase, the user query is mapped to topic.
  Paper [11] focuses on mapping between query and category or topic in three processes as:

- Using *User Profile Only* in which k-nearest neighbor (kNN) computes similarity between the query and each category from Document-Category and Document-Tree.
- Using *Generalized User Profile Only* which uses Pseudo-Linear Least Square Fit (PLLSF) to compute generalized profile as it has highest average accuracy but it is computationally expensive.
- Using *both profiles* which computes similarity for every category.

Fang Liu in [11] have used different approach to infer user search intention. To process query-level customized generalization, they have taken-out the category domain from query and then top "n" number of categories with next button are shown to user. If user interest is not among the shown result then the next button can be clicked to show preceding "n" categories. Different strategies can also be used for mapping i.e. that weights the candidates or term-distance mapping but those are unstable because it adopts the term-based profiles limitations.

- The final step of online phase is *cost-based generalization*. On the topic domain bounded by query and category in step1 of same phase, this generalizes seed profile in cost based manner. The algorithms used in [1][2] are based on the metric of utility and risk i.e. *util(q,G)* which predicts the potential gain of query q on generalized profile G and *risk(q,G)* which is the total sensitivity in G, given in a normalized form.

Online decision on personalizing queries are discussed in [2] which addresses the problem of reducing the search quality that may risks user privacy. They developed an online mechanism which personalizes queries on-the-fly. The advantage of online approach is that they improves search quality and avoids exposure of user-profile. This method is completely relied on client-side utility estimation of the query.

J. Teevan [12] believe that another promising approach to personalizing search is to infer user information goals automatically. He also studied overview of research done in information retrieval on how implicit measures can be used to aid search. The PS prototype uses a person's prior interactions with a wide variety of content to personalize that person's current Web search in an automated manner. Zhicheng[13] work has categorized personalized search methods into person level or group level. He proposed several re-ranking methods in both the levels. These strategies are used to re-rank search results by computing a personalized score for each page in the results returned to user on query. The study in [14] investigates the effectiveness of personalized search based on user profile constructed through user search histories. Their system architecture consist of GoogleWrapper, classifier from key concept. Another protocol i.e. Useless user protocol (UUP) was developed by Jordi Castella-Rocain [15] that allows user to submit query to web search engine with keeping user personal information secure. The idea of this protocol is that each user who wants to issue a query will send query of another user instead of sending his/her own query. At the same time, his/her query is submitted by another user. But this approach is not so feasible to use.

## 5. Proposed Approach

In our proposed system, we are building user profile hierarchically with user interest. If user specifies sensitivity for any topic then that are not allowed to appear in generalized user profile. We are focusing on the mapping approach taken by [2]. After submitting query q , we retrieve the documents similar to query using conventional approach. These documents are then grouped together. The relevance method used in this framework is simple and fast to evaluate. and will also check users last searches to get the relevant query meaning .We have used secure random number generation algorithm i.e.SHA1PRNG to ensure attacks from eavesdropper. Whenever the user profile is generated ,with that the unique random key will be assigned to user which can only be generated either by admin or by user himself/herself. And for searching the query on PWS, user has to enter the assigned query to search engine. Also we are generating log files for the searched query and user are asked to give explicit feedback about the results. Once the feedback is collected, the rated documents are extracted to enrich the use profile.

## Conclusion

This study paper presents the different approaches that have been implemented for personalizing web search. There is tremendous growth in the approaches taken to represent, construct and employ user profiles. These enabling techniques are key to providing user with accurate, personalized information services. As personalized search has different effectiveness for different kinds of queries, we believed that queries should not be handled in same manner with regards to personalization. We have also studied different generalization algorithm for online generalization which are implemented in related papers.

## References


[1] Lidan Shou, He Bai, Ke Chen, and Gang Chen, "Supporting Privacy Protection in Personalized Web Search, Knowledge and Data Engineering, IEEE Transactions on , vol.26, no.2, pp.453,467, Feb. 2014 doi: 10.1109/TKDE.2012.201.

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

[4] 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.

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

[6] Ji-Rong Wen, Zhicheng Dou, "Personalized web search".

[7] 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.

[8] H.R. Kim and P.K. Chan, "Learning implicit user interest hierarchy for context in personalization, " Proc. of International conference on Intelligent User Interfaces (IUI), Miami, Florida, 2003.

[9] Mingxuan Yuan Lei Chen Philip S. Yu, "Personalized Privacy Protection in Social Network," Proc. VLDB,2010.

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

[11] Fang Liu , Clement Yu , Weiyi Meng, "Personalized web search by mapping user queries to categories".

[12] J. Teevan, S. T. Dumais, and E. Horvitz. "Beyond the commons: Investigating the value of personalizing web search". In Proceedings of PIA '05, 2005.

[13] 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.

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

[15] 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.