# A Survey Paper on Supporting Privacy Protection in Personalized Web Search

**Avan Wansing[1], Neha Pandey[2], Kamruddin Rogangar[3]**

[123]*Dept. of Information Technology, ICOER Pune, Savitribai Phule Pune University, Maharashtra, India*

---***---

**Abstract -** *Internet of things is in its mount in today's world. Web exploring is the most common task perform on the internet. The web search engines are the most important tool of the internet, search engines are the place from where an individual can collect the relevant information and search according to keyword given by the user. The data on the wed are increasing day by day very dramatically. The user has to spend a lot of time on the net for finding the data in which they are interested. The irrelevant result may irritate the user and hence, the efficiency of the query search should be improved. To improve the search, personalized web search framework has demonstrated to retrieve the data on the user's interest. In this paper, the users profile is protected from handling privacy threats using generalisation techniques with Greedy Discriminating Power and Greedy Information Loss Algorithm. A user profile is protected by using PWS which models the user preferences as in hierarchical structure. In this, a PWS framework called UPS focus on providing protection against any model of privacy attack. UPS framework can generalize profiles by the entered queries. The use of Greedy algorithms has significantly improved the efficiency and effectiveness of search result.*

***Key Words***: Privacy protection, Personalized Web Search, UPS Framework, Generalisation.

## 1. INTRODUCTION

The search engines are an essential hatch for the peoples for getting the useful required data. But, the user generally faces the failure and find the result they got is unnecessary or improper which are irrelevant of their intention. The regular web search engines give the similar log of results without considering of who enter the query. Hence, the requirement of personalized web search is arises which give the appropriate output to the user. Personalized web search i.e PWS is a search technique which mainly focuses on providing a quality search result, as per an individual needs. So, for this purpose, the users information has to be gathered and studied so that the perfect results required by the client behind the entered query is to be given to the client. The solution for this is personalized web search (PWS).It is mainly categorized into two types, one is a clicked-log-base method and other is profile base method. The clicked-log-base method is very simple and straightforward. This method accomplishes the search which is based on clicked logged pages in the user's query history. As this method has been proved to achieve consistently and considerably well [6], it can work on repeated queries from the same user only which is a huge limitation as well as restricted for certain application. On the other hand, the profile-based method has improved the search experience the techniques, the profile based PWS has proved its high effectiveness to make better the quality of web search, with increase in the uses of users personal information to build its user profile, which is generally collected implicitly with the help of searched query[6], browsing history[7], clicked data, bookmarks[6] and so on. Unfortunately, such type of collected users' personal data can be easily revealed to the outside world or can be hacked for misuse entire scope of user's private life. Lack of protection for the data can rise the privacy protection issues.

## 2. RELATED WORK

In [1] Z. Dou, R. Song, and J.-R. Wen et al. Personalization strategies had been proposed and investigated for many years but it's miles nonetheless doubtful whether or not the strategy is always effective on distinctive queries for special customers, under one of a kind search context. In [1], they have investigated whether personalization is continuously effective below distinctive conditions. They advanced an evaluation framework based totally on question logs to allow big scale assessment of personalized search. Click entropy an easy size on whether the question must be personalized. Click primarily based personalization strategies can paintings on repeated queries. The benefits revealed that the personalization has different effectiveness on different queries and both short term and long term context improve the search performance. On the other side, because of a large-scale evaluation of search contexts, the framework may be time-consuming and complex to handle.

In [2] A. Krause and E. Horvitz et al. Online offerings, for example, web search, news portals, and e-commerce applications confront the test of giving amazing support of an expansive, heterogeneous client base. To overcome such problem an effort has been introduced by introducing methods to personalize services based on special knowledge about users and their context. Researchers and organizations have sought after explicit and implicit methods for customizing online administrations. An approach for explicitly optimizing the utility-privacy tradeoff in personalized services such as web search. Privacy concerns show super-modularity; the more private information we accrue, the faster sensitivity and the risk of identifiability grow.

[2]A. Krause et al demonstrated how can efficiently find a provably near- optimal utility-privacy tradeoff and evaluated methodology on real-world web search data. The common belief is that the principles and methods employed in the utility-theoretic analysis of tradeoffs for web search have applicability to the personalization of a broad variety of online services. In [2] found that significant personalization can be achieved using only a small amount of information about users with the limitation that the system is dependent on the log of user search activity.

In [3] J. Castelli-Roca, A. Viejo, and J. Herrera-Joancomarti et al. Web search engines like Yahoo!, Google, Bing, etc. are widely used to find the particular amount of data among a large amount of data in a short amount of time. People over the globe use the web search engine for different purposes which are relevant to them. At the same time, needed information belongs to the specific topic is hidden among all the available data and it can be really difficult to find it since that information can be separated all over the World Wide Web. In fact, these useful things can also cause the privacy threats to the users, web search engines can profile the client by storing and analyzing the past queries requested by them. But to solve this privacy threats current mechanism introduces high cost in terms of computation and communication. In this paper, they produce a novel protocol designed to protect the user's privacy in front of web search profiling. Their system gives the duplicate or deformed user profile to the web search engines. [3] They offered implementation details and computational or communication results which show that the introduced protocol improves the existing solutions in terms of query delay. The limitation of the existing system was that the person or the entity can get some advantage over the other benefits from the absence of privacy protection mechanism between the user and the web search engine. So the problem of submitting the queries of the user to the search engine while preserving the privacy protection to the profile it can be term as Private Information Retrieval (PIR) problem. In PIR what happen is user can retrieve his values from the database while the server gets no information about the activity of the user. Simple methods to obtain the certain level of privacy to the web browsing includes the use of the proxies or the dynamic IP address. But proxy does not solve the privacy problem. The proxy can prevent the web search engines from creating the profile of the user, it can profile them instead.

In [4] X. Xiao and Y. Tao et al. did study on the generalization for preserving the privacy of the sensitive data which is daily produced by the users. The existing techniques concentrate on the each and every approach that cause the same amount of preservation for all the users without analyzing their original needs. This results in providing the insufficient protection to a group of people who actually need it while giving extreme privacy control to the group of people who doesn't need it. This system cannot guarantee the privacy protection in all cases this could lead to cause the unnecessary data loss by performing excessive use of generalization. At first, they make a concept that forms a new framework of computing privacy which takes into account the sensible information by an individual preference. Secondly, they analyze the theory behind their methodology and evaluate the formulae for quantifying the privacy which clearly show the scenarios where k-anonymity may make sure about safe data production. Finally, they evolved an algorithm for finding the generalized that keeps a huge amount of information in the microdata without breaking any privacy limits. The Greedy Algorithm divided into two categories, according to the constraint imposed on generalization. The first category includes "*full-domain generalization*" which undertake hierarchy on every QI attribute and all the partitions in the hierarchy needs to be at same level. The second category includes "*full-sub tree recording*" which drop the same level of hierarchy which mentioned earlier in the first category that causes unnecessary information lose.

## 3. EXISTING SYSTEM

For protecting the user privacy in the profile based personalized web search, examiners have to keep in mind two important and gainsay issue during the search process. The first point is that they try to make better the search quality with the personalization utility of the profile of the user. The second point is that they have to hide the privacy contents present in the user profile to place the privacy risks in control. However, some people are ready to compromise privacy if the search engines yield better search result by supplying the user profile. In similar condition, the significant rise can be achieved by personalization at the expense of only small part of the user profile i.e. generalized profile. There is give-and- take like the situation between the level of privacy protection and the search quality which is obtained from generalization. The issue with the existing method are explained in following remarks:

1. Profile-based Personalized Web Search has a disadvantage that it do not support runtime profiling. A user profile is typically generalized for only once offline and it may not even improve the search quality for some ad hoc queries, exposing user profile to a server has put the user's privacy at risk.
2. 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.

3. Most of the personalization techniques need repetition of user interaction when building up the personalized search results. The result with some metric which require multiple user interactions like rank scoring, average rank [8], and so on.

### 3.1 UPS Framework

To solve the above problem UPS (User customizable privacy-preserving search) is explained.[7] The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS. UPS consists of a number of users and typical web search engines server. Each user who is accessing the web search service trusts nobody but itself. The key element for privacy protection is an online profiler which is implemented as search proxy running on the user machine itself. The proxy maintains both the *complete user profile* [7], in a hierarchy of nodes with systematics and the [7] *user specified (customized) privacy requirements* represented as a set of *sensitive nodes*.

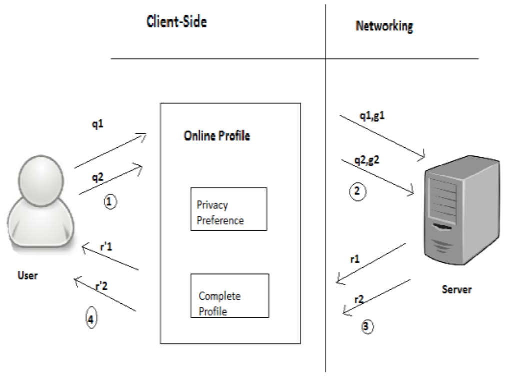


**Fig -1:** System architecture of UPS

The framework works in two phases, [7], [9] the *offline* and *online* phase, for each user. During the offline phase, a hierarchical user profile is build up and modified with the user-specified privacy requirements. The online phase handles queries as follows:

1. When a user issues a query *qi* on the client, the proxy generates a user profile in runtime in the light of *query terms*. The output of this step is a *generalized* user profile *Gi* satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles.

2. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.

3. The search results are personalized with the profile and delivered back to the query proxy. 4. Finally, the proxy either presents the raw results to the user or reranks them with the complete user profile. UPS is distinguished from conventional PWS in that it

a) Provides runtime profiling, which in effect optimizes the personalization utility while respecting user's privacy requirements;
b) Allows for customization of privacy needs; and
c) Does not require iterative user interaction

## 4. PROBLEMS AND ISSUES

In this section, the structure of user profile in UPS is introduced and presented the attack model and the problem of privacy preserving profile in generalization.

### 4.1 User Profile

The personalized web search is a framework where the user profile is protected during the search [9]. The profile is created with the help of detail information of users entered queries, browsing history, cookies and so on. As discussed earlier, the user profile can be generated in two phases, online and offline phase and a hierarchical structure is obtained. For instance, consider the following figure(2) which shows the general taxonomy of search from which the user profile is created showed in figure(3) with the sensitive topics.

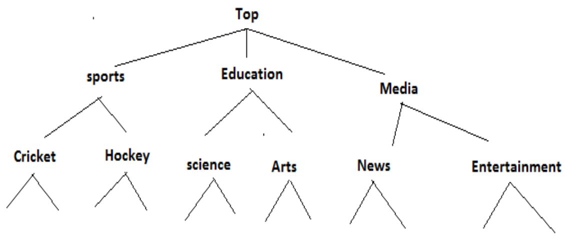


**Fig -2**: Taxonomy Repository.

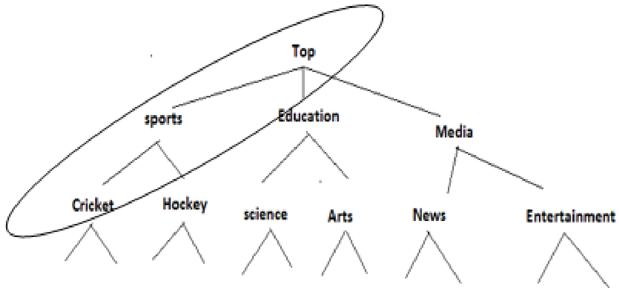


**Fig -3**: User's Profile creation from the taxonomy.

*Offline Phase*: The original user profile and customized privacy are constructed in the offline mode [9].

*Online Phase*: Query mapping and generalization of the profile is done in online phase [9].

### 4.2 Attack Model

The work is mainly focused at providing protection against a typical model of privacy attack, called eavesdropping. To corrupt Alice's privacy, the eavesdropper Eve successfully intercepts the communication between Alice and the PWS server via some measures, such as man attack, invading the server, and so on. Consequently, whenever Alice issues a query *q*, an entire copy of *q* together with a runtime profile *G* will be captured by Eve. Based on *G*, Eve will attempt to touch the sensitive nodes of Alice by recovering the segments hidden from the original H and computing a confidence for each recovered topic, relying on the background knowledge in the publicly available taxonomy repository *R*.

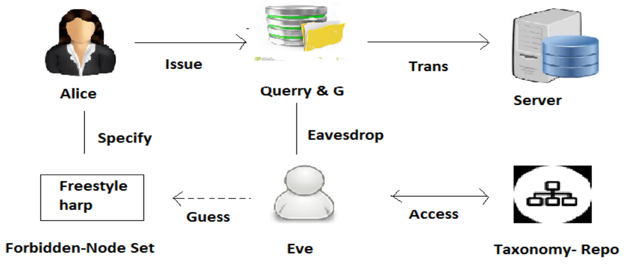


**Fig -4**: Attack model of personalized web search.

Note that in the attack model, Eve is considered as an adversary satisfying the following assumptions:

*Knowledge bounded*: The background knowledge of the adversary is limited to the taxonomy repository *R*. Both the profile H and privacy are defined based on *R* [7].

*Session bounded:* None of the previously captured information is available for tracing the same victim in a long duration. In other words, the eavesdropping will be started and ended within a single query session [7].

### 5. GENERALISATION

Generalization is an extension of context in a very less specific criteria. Generalization helps in avoiding the unnecessary privacy disclosure. Topics which are irrelevant to the current query are considered as noisy topics and they are removed. Generalization technique can be conducted during both online and offline process without actually involving users query. There are certain limitations of offline generalization such as it contains many branches which are irrelevant to queries, whereas online generalization provides flexible solutions.

### 5.1 Metric for Utility

The intention of the utility metric is to guess the search quality of the query *q* [7] on a generalized profile *G* [7]. The main reason for the use of utility metric is that the quality of search depends upon users search in the personalized web search engine [9].

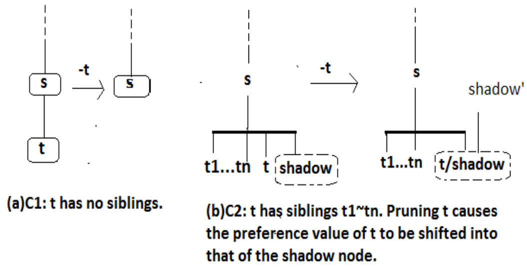
### 5.2 Online Decision

[7]The profile-based personalization contributes little or even reduces the search nice while exposing the profile to a server would for positive danger the user's privacy. To cope with this trouble, we expand an online mechanism to determinewhether or not to customize a question. The fundamental idea is honest- if a wonderful question is diagnosed at some point of generalization, the entire runtime profiling may be aborted and the question may be sent to the server without a person profile.

### 5.3 Generalization Algorithm

GreedyDP and GreedyIL, for runtime generalization. Where GreedyIL significantly outperforms GreedyDP in terms of efficiency. In the UPS, joint with a greedy algorithm i.e. Greedy DP [10] named as Greedy Utility to help online profiling based on predictive metrics of utility and privacy risk [10].

*5.3.1 GreedyDP Algorithm:* The first greedy algorithm GreedyDP works in a bottom-up manner. Firstly, introduce prune-leaf, which indicates the removal of a leaf topic *t* from a profile. Formally, denote by $G \xrightarrow{-t} G_{i+1}$ (shown in figure 5(a)) the process of pruning leaf *t* from *Gi* to obtain *Gi+1*. Obviously, the optimal profile *G** can be generated with a finite-length transitive closure of prune-leaf [7], [10]. Secondly, starting from *G0*, in every ith iteration, GreedyDP chooses a leaf topic tЄTGi (q) for pruning, trying to maximize the utility of the output of the current iteration, namely *Gi+1*. During the iterations, maintain the best profile-so-far, which indicates the *Gi+1* having the highest discriminating power while satisfying the δ- risk constraint [7],[10]. Finally, the iterative process terminates when the profile is generalized to a root topic. The best-profile-so-far will be the final result (*G**) of the algorithm [7], [10].
*5.3.2 GreedyIL Algorithm:* The GreedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings. One important finding is that any prune-leaf operation reduces the discriminating power of the profile [7], [10].



(a)C1: t has no siblings.

(b)C2: t has siblings t1~tn. Pruning t causes the preference value of t to be shifted into that of the shadow node.

**Fig -5:** Cases of prune-leaf on a leaf *t*

### 6. CONCLUSION AND FUTURE WORK

The search history and the search queries of the web user are saved by the web search engines. This saved data can be used by the user as to provide other relevant data for the user. User personal data i.e. browsing histories and the queries create the profile of the user by the engines and it should be protected to avoid the threats. UPS could be used by any typical PWS that takes users profiles in a hierarchical structure. The generalization algorithms, GreedyDP, and IL, which handles the privacy issues in PWS by offering user to control the amount of private data reveal to the web servers. The private parameters facilitate smooth control of privacy exposure while maintaining good ranking quality. In future, other privacy threats can be handled with efficient algorithm and can find smarter techniques to build the user profile, and better metrics to predict the performance of UPS.

## 7. REFERENCES

[1] 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.
[2] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.
[3] J. Castelli-Roca, A. Viejo, and J. Herrera-Joancomarti ´"Preserving User's Privacy in Web Search Engines," Computer Comm., vol. 32, no. 13/14, pp. 1541-1551, 2009.
[4] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[5] 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.
[6]F. Qiu and J. Cho, (2006) Automatic Identification of User Interest for Personalized Search," Proc. 15th Int'l 727-736.
[7] Lidan Shou, He Bai, Ke Chen, and Gang Chen (2014)Supporting Privacy Protection in Personalized Web Search, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 2.
[8] X. Shen, B. Tan, and C. Zhai, (2005) Context Sensitive Information Retrieval Using Implicit Feedback, Proc. 28th Ann. Int'l ACMSIGIR Conf. Research and Development Information Retrieval (SIGIR).
[9] N. Nivi, S Vanitha, R Saranya, B Sivaranjani, S Kavitha, "Preserving User's Profile Protection for Personalized Web Search", Vol. 3.
[10] V. Ramya, S. Gowthami, "Enhance Privacy Search in Web Search Engine using Greedy Algorithm", Int'l journal of Scientific Research Engineering & Tech. (IJSRET), Vol. 3. [11] A. Patil, M. Ghonge, M. Sarode, "User customizable Privacy-preserving Search Framework-UPS for Personalized Web Search", Int'l Journal of Research in Advert Technology, Vol.2.