

# Innovative Privacy Preserving Search Framework for Personalized Web Search

Syeda Noushin Fathima

Department of CSE  
APSCE, Bangalore,  
Karnataka, India

Somasekhar.T

Department of CSE  
APSCE, Bangalore,  
Karnataka, India

**Abstract**— Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. The evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. For privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. This runtime generalization aims at striking a balance between two metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present greedy algorithm, namely GreedyIL, for runtime generalization.

**Keywords** — *Personalized Web search, Privacy Protection ,Risk , Profile.*

## I INTRODUCTION

Web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.

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 rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

To protect user privacy in profile-based PWS, researchers have to consider two contradicting effects during the search process. On the one hand, they attempt to improve the search quality with the personalization utility of the user profile. On the other hand, they need to hide the privacy contents existing in the user profile to place the privacy risk under control. A few previous studies suggest that people are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality. In an ideal case, significant gain can be obtained by personalization at the expense of only a small (and less-sensitive) portion of the user profile, namely a generalized profile. Thus, user privacy can be protected without compromising the personalized search quality. In general, there is a tradeoff between the search quality and the level of privacy protection achieved from generalization.

## II BACKGROUND AND RELATED WORK

In this section, we overview the related works. We focus on the literature of profile-based personalization

### 1. Profile-Based Personalization

Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. In the remainder of this section, we review the previous solutions to PWS on 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 list or bags of words to represent their profile. However, most recent works build profiles in hierarchical structures due to their stronger descriptive ability, better scalability, and higher access efficiency. The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP1, Wikipedia2 and so on. Another work in builds the hierarchical profile automatically via term-frequency analysis on the user data. In our proposed UPS framework, we do not focus on the implementation of the user profiles. Actually, our framework can potentially adopt any hierarchical representation based on a taxonomy of knowledge.

2. Existing system

To protect user privacy in profile-based PWS, researchers have to consider two contradicting effects during the search process. On the one hand, they attempt to improve the search quality with the personalization utility of the user profile. On the other hand, they need to hide the privacy contents existing in the user profile to place the privacy risk under control. A few previous studies suggest that people are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality.

In an ideal case, significant gain can be obtained by personalization at the expense of only a small (and less-sensitive) portion of the user profile, namely a generalized profile. Thus, user privacy can be protected without compromising the personalized search quality. In general, there is a tradeoff between the search quality and the level of privacy protection achieved from generalization.

- The existing profile-based PWS do not support runtime profiling
- The existing methods do not take into account the customization of privacy requirements.

III PRELIMINARIES

In this section, we first introduce the structure of user profile in UPS. Then, we define the customized privacy requirements on a user profile.

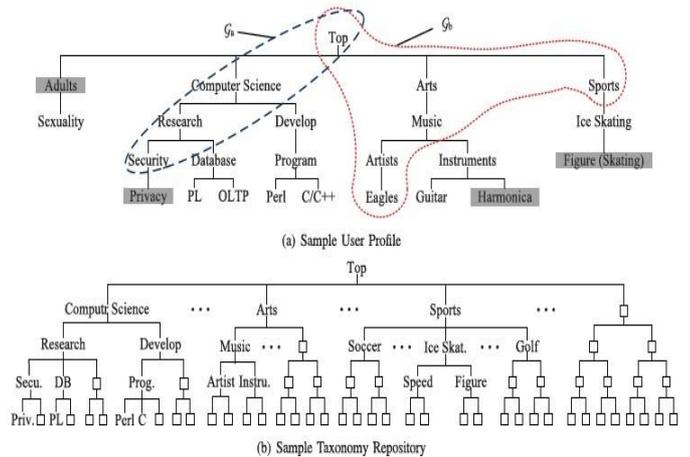
1. User Profile

Consistent with many previous works in personalized web services, each user profile in UPS adopts a hierarchical structure. Moreover, our profile is constructed based on the availability of a public accessible taxonomy, denoted as R, which satisfies the following assumption.

**Assumption 1.** The repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic t, a corresponding

node (also referred to as t) can be found in R, with the subtree  $subtree_t$ ; RP as the taxonomy accompanying t.

A diagram of a sample user profile is illustrated in Fig. a, which is constructed based on the sample taxonomy repository in Fig. b. We can observe that the owner of this profile is mainly interested in Computer Science and Music, because the major portion of this profile is made up of fragments from taxonomies of these two topics in the sample repository. Some other taxonomies also serve in comprising the profile, for example, Sports and Adults.



Taxonomy-based user profile

IV PROBLEM FORMULATION

1. Proposed system

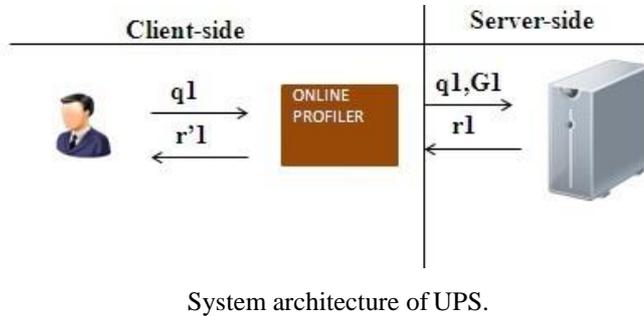
The above problems are addressed in our UPS (literally for User customizable Privacy-preserving Search) framework. The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS.

As illustrated in Figure of next section, UPS consists of a nontrusty search engine server and a number of clients. Each client (user) accessing the search service trusts no one but himself/ herself. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified (customized) privacy requirements represented as a set of sensitive-nodes.

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:

1. When a user issues a query q1 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 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.

2. Design goals

We propose 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, we formulate the problem of privacy-preserving personalized search as  $\delta$ -Risk Profile Generalization, with its NP hardness proved.

We develop a simple but effective generalization algorithm, Greedy IL, to support runtime profiling to minimize the information loss (IL).

3. The 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. In other words, the DP displays monotonicity by prune-leaf. Formally, we have the following theorem:

---

**Algorithm 1:** GreedyIL( $\mathcal{H}, q, \delta$ )

---

**Input :** Seed Profile  $\mathcal{G}_0$ ; Query  $q$ ; Privacy threshold  $\delta$   
**Output:** Generalized profile  $\mathcal{G}^*$  satisfying  $\delta$ -Risk

```

1 let  $\mathcal{Q}$  be the IL-priority queue of prune-leaf decisions;
   $i$  be the iteration index, initialized to 0;
  // Online decision whether personalize  $q$  or not
2 if  $DP(q, \mathcal{R}) < \mu$  then
3   Obtain the seed profile  $\mathcal{G}_0$  from Online-I;
4   Insert  $\langle t, IL(t) \rangle$  into  $\mathcal{Q}$  for all  $t \in T_{\mathcal{H}}(q)$ ;
5   while  $risk(q, \mathcal{G}_i) > \delta$  do
6     Pop a prune-leaf operation on  $t$  from  $\mathcal{Q}$ ;
7     Set  $s \leftarrow par(t, \mathcal{G}_i)$ ;
8     Process prune-leaf  $\mathcal{G}_i \xrightarrow{-t} \mathcal{G}_{i+1}$ ;
9     if  $t$  has no siblings then // Case C1
10      Insert  $\langle s, IL(s) \rangle$  to  $\mathcal{Q}$ ;
11    else if  $t$  has siblings then // Case C2
12      Merge  $t$  into shadow-sibling;
13      if No operations on  $t$ 's siblings in  $\mathcal{Q}$  then
14        Insert  $\langle s, IL(s) \rangle$  to  $\mathcal{Q}$ ;
15      else
16        Update the IL-values for all operations on
17         $t$ 's siblings in  $\mathcal{Q}$ ;
18      Update  $i \leftarrow i + 1$ ;
19    return  $\mathcal{G}_i$  as  $\mathcal{G}^*$ ;
20 return  $root(\mathcal{R})$  as  $\mathcal{G}^*$ ;

```

---

4. SYSTEM MODULES

• Profile-Based Personalization

This paper introduces an approach to personalize digital multimedia content based on user profile information. For this, 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.

• 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. We develop a simple but effective generalization algorithm (GreedyIL: minimize the information loss(IL)) for user profiles allowing for query-level customization using our proposed metrics. We also provide 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 our framework.

- **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.

- **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, we develop 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.

## V CONCLUSION

This paper presented a client-side privacy protection framework called UPS for personalized web search. UPS could potentially be adopted by any PWS that captures userprofiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality.

## VI 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] 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.
- [3] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.
- [4] 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.

- [5] 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.
- [6] X. Shen, B. Tan, and C. Zhai, "Implicit User Modeling for Personalized Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2005.
- [7] X. Shen, B. Tan, and C. Zhai, "Context-Sensitive Information Retrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.
- [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. Schu"tze, 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] 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.
- [11] K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.
- [12] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.
- [13] J.S. Breese, D. Heckerman, and C.M. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.
- [14] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlsch"utter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.
- [15] A. Pletschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999