

# Personalized Web Search Protecting Privacy Using Greedy Algorithm

Reshma Y. Totare<sup>1</sup> Meenakshi Rathod<sup>2</sup> Prajakta Raskar<sup>3</sup> Pooja Modak<sup>4</sup> Venuvanshi Bhujbal<sup>5</sup>

<sup>1,2,3,4,5</sup>Department of Information Technology  
<sup>1,2,3,4,5</sup>AISSMS IOIT

**Abstract**— The amount of information on the World Wide Web is growing rapidly; search engines must be able to retrieve information according to the user's preference. Current web search engines are built to serve all users, independent of the special needs of any individual user. Personalization of web search is to carry out retrieval for each user incorporating his/her interests. Every user has a distinct background and a specific goal when searching for information on the Web. Thus the goal of Web search personalization is to tailor search results to a particular user based on that user's interests and preferences. However, effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. Indeed, these concerns have become one of the main barriers for deploying personalized search applications, and how to do privacy-preserving personalization is a great challenge. Thus, a balance must be struck between search quality and privacy protection. Hence, privacy protection in PWS applications that model user preferences as hierarchical user profiles is proposed using a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Along with Personalized Search and Privacy Protection the Custom Search functionality will also be provided so that the users get relevant information.

**Key words:** Information, Personalization, PWS-Personalized Web Search, Search Engine, 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 search engine returns the same result, regardless of who submitted the query. This may not be suitable for users with different information needs. For example, for the query "apple", some users may be interested in documents dealing with "apple" as "fruit", while some other users may want documents related to Apple computers. 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 "cooking" or the category "fruit" is associated with the query "apple", then the user's intention becomes clear [1]. 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 [2]. Personalized web search differs from generic web search, which returns identical research results to all users for identical queries, regardless of varied user interests and information needs [2]. Despite the attractiveness of personalized search, we have not yet seen large scale uses of personalized search services. This is not because such services are not available, but likely because users are not comfortable with the lack of protection of user privacy [5, 6]. Indeed, there is an inherent tension between providing personalized search and privacy

preservation since personalized search requires collecting and aggregating a lot of user information. Specifically, in order to personalize search, a user profile or user model must be constructed to accurately represent a user's information need. To build a precise user profile, a lot of user information including query and clickthrough history is often aggregated [3].

Unfortunately, such type of collected personal data can easily reveal entire scope of user's private life. Protecting privacy issues rising from the lack of protection for such data, for example the AOL query logs scandal, not only raise panic among individual users, but also downs the data publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide use of PWS services[4]. Thus there appears to be a dilemma: high-accuracy Web search requires accurate user modelling which increases the risk of privacy infringement. Indeed, the privacy concern is one of the major barriers in deploying serious personalized search applications, and how to achieve personalized search while preserving users' privacy is In this paper, we systematically examine the issue of privacy preservation in personalized search [3].

## II. APPROACH

Nowadays, for every user query, users enter that query in search engine's provided on web and get's a result for that query really fast. But the query results are not that accurate or relevant to the query. Personalized Web Search (PWS) is a general category of search techniques aiming at providing better search results, which are provided for individual user's needs. The solutions to PWS can be categorized into two types click-log-methods and profile-based-ones. The click-log based methods are straightforward—they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well, it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. In our, we have introduced one more method, that is the third method - custom-search-based method. In this method, the user can fire it's query without being logged in and he or she will be provided with the same functionality as explained in the first two methods above. The condition is that the IP (Internet Protocol) or system of the user that he or she is using without being logged in should always be same. And the user can select a option , according to which the user can decide that whether the click-log-method and profile-based history should be mapped or added to the search history on IP -level based search ,that is custom search.

The target is to provide search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided. And our main goal is to provide

protection to the user's personal data gathered implicitly from query history, browsing history, click-through data, bookmarks, user documents and so forth in Personalized Web Search using UPS framework and also customize the search.

### III. BLOCK DIAGRAM OF THE SYSTEM

The proposed system consists of simple, efficient and privacy preserving model which ensures good suggestions as well as promises for effective and relevant information retrieval.

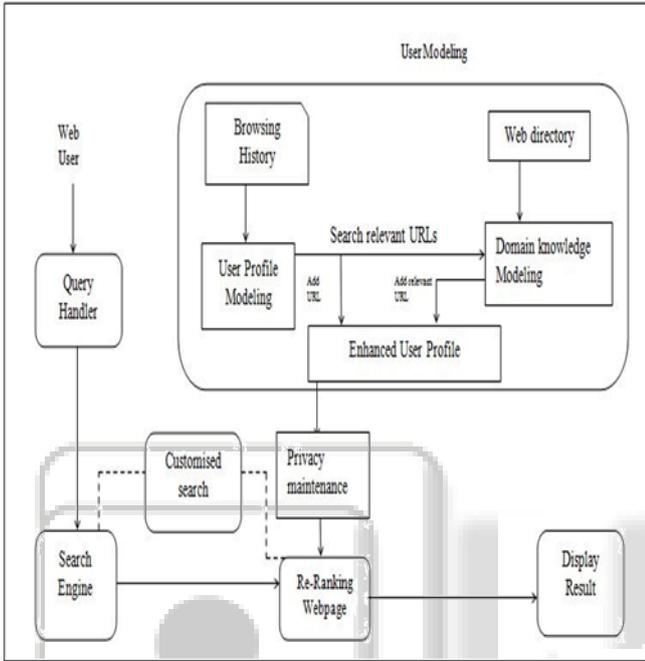


Fig. 1: Architecture

#### A. Working of the System:

To achieve personalization the proposed system generates an enhanced user profile as follows; it considers user's profile (based on user's weblog navigation browsing history) and Domain Knowledge. Using a Domain Knowledge, the system stores information about different domains/categories. Information obtained from User Profile is classified into these specified categories. The learning agent learns user's choice automatically through the analysis of user navigation/browsing history, and creates/updates enhanced User Profile conditioning to the user's most recent choice.

Once the user inputs query, the system provides good suggestions for personalized web search based on enhanced user profile. Further the proposed model makes good use of the advantages of popular search engines, as it can re-rank the results obtained by the search engine based on the enhanced user profile.

For privacy maintenance the proposed model uses UPS (User customizable privacy preserving search). UPS consist of anon-trusty search engine server and a number of clients. 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. The proposed system makes use of UPS (User customizable privacy

preserving search) to maintain the privacy and achieve personalization.

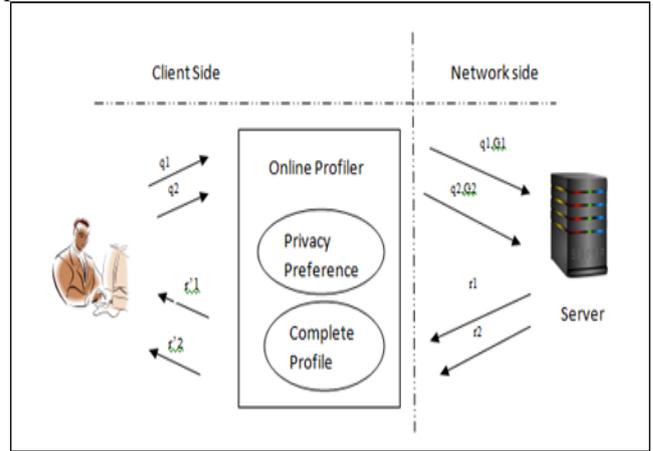


Fig. 2: UPS Framework

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.

UPS framework works in two phases

- Offline phase
- Online phase

In offline phase hierarchical user profile is constructed and customized with the user-specified privacy requirements.

The Online phase works as follow:

- 1) The user issues a query  $q_i$  on the client; the proxy generates a user profile in runtime. The output of this step is a generalized user profile  $G_i$  satisfying the privacy requirements.
- 2) The query and the generalized user profiles 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.

### IV. MODULES

- Profile-Based Personalization.
- Privacy Protection in PWS System.
- Generalizing User Profile.
- Online Decision.
- Customized Search Results

#### A. Profile-Based Personalization:

This project 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.

**B. Privacy Protection in PWS:**

We propose a PWS framework called UPS that can generalize profiles in for each query according to user-specified privacy requirements. We develop two simple but effective generalization algorithms (GreedyDP and GreedyIL) for user profiles allowing for query-level customization using our proposed metrics.

**C. 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.

**D. 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.

**E. Customized Search:**

This project provides a customized search to the user. It shows relevant information to the user depending on the ranking of the page in its history.

**V. ALGORITHM**

The project uses two algorithms for

**A. Greedy DP: Greedy Discriminating Power [7].**

This algorithm gives optimal solution hence called a Near Optimal Greedy Algorithm. For removal of leaf topic from profile we will introduce an operator  $---t-->$  This is called Prune leaf . We may have 2 cases for removal of leaf

**1) Case 1: When t Has No Siblings.**

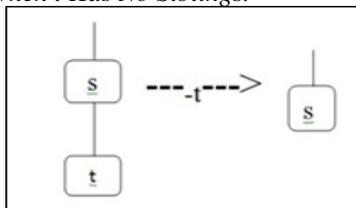


Fig. 3: t has no siblings

**2) Case 2: When t Has Siblings**

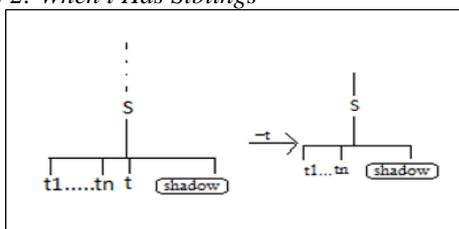


Fig. 4: when t has siblings

Optimal profile  $G^*$  is generated with finite length transitive closure of prune leaf. At  $i^{th}$  iteration, a leaf topic  $t$  for pruning is selected. During iterations the profile so far is maintained. Iteration terminates when profile is generalized to root topic.

The main problem with GreedyDP is that it requires a lot of computation of all candidate profile.

**B. GreedyIL: To Increase The Efficiency GreedyIL Algorithm Is Used [7].**

Following terminologies are used in GreedyIL algorithm.

- G0: Seed profile
- q:query
- $\delta$  : Privacy Threshold.
- G\*: Generalized profile satisfying  $\delta$ - Risk.
- Q: IL-priority queue of prune-leaf decision.
- i: Iteration index initialized to 0.
- Input is  $G_0, q, \delta$ .
- Output:  $G^*$ .

Following steps will be carried out for online decision whether to personalize  $q$  or not

```

If  $DP(q,R) < \mu$  then do following:
    Obtain the seed profile  $G_0$  from Online-1,
    Insert( $t, IL(t)$ ) into Q for all to  $\epsilon T(q)$ 
    While  $risk(q, G_i) > \delta$  do
    Pop a prune-leaf operation on  $t$  from Q
    Set  $s \leftarrow part(t, G_i)$ 
    Process prune leaf  $G_i \rightarrow G_{i+1}$ 
    If  $t$  has no siblings then //case 1
    Insert( $s, IL(s)$ ) to Q
    Else if  $t$  has siblings then //case2
    Merge  $t$  into shadow-sibling
    If No operation on  $t$ 's siblings in Q then
    Insert( $s, IL(s)$ ) to Q
    Else Update IL- value for all operations on  $t$ 's sibling Q.
    Update  $i \leftarrow i+1$ 
return  $G_i$  as  $G^*$ 
return root(R) as  $G^*$ 
    
```

**VI. CONCLUSION**

Personalized search is a promising way to improve search quality. However, this approach requires users to grant the server full access to personal information on Internet, which violates users' privacy. This paper proposes to provide fast and relevant search that is personalized using User Profile. Based on User Profile the system keeps on updating user profile and an enhanced user profile is created. This paper also provides information on User customizable Privacy preserving Search framework-UPS for Personalized Web Search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles.

**REFERENCES**

- [1] Fang Liu, Clement Yu, Weiyi Meng, Personalized Web Search for Improving Retrieval Effectiveness, University of Illinois at Chicago
- [2] Ji-Rong Wen, Zhicheng Dou, Ruihua Song, Personalized Web Search, Microsoft Research Asia, Beijing, China

- [3] XuehuaShen, Bin Tan, ChengXiangZhai, Privacy Protection in Personalized Search, University of Illinois at Urbana-Champaign
- [4] A.S. Patil, Prof M.M. Ghonge, Dr. M. V. Sarode, User Customizable Privacy-preserving Search Framework-UPS for Personalized Web Search, 2014.
- [5] C.-M. Karat, C. Brodie, and J. Karat. Usable privacy and security for personal information management. *Communications of the ACM*, 49(1):56-57, 2006.
- [6] S. Sackmann, J. Strker, and R. Accorsi. Personalization in privacy-aware highly dynamic systems. *Communications of the ACM*, 49(9):32-38, 2006.
- [7] Lidan Shou, He Bai, Ke Chen, and Gang Chen "Supporting Privacy Protection in Personalized Web Search".

