ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Improving the Quality of Search in Personalized Web Search

P.Ramya
PG Scholar,
Dept of CSE,
Chiranjeevi Reddy Institute of Engineering & Technology,
AP, India.

S.Sravani
Assistant Professor,
Dept of CSE,
Chiranjeevi Reddy Institute of Engineering & Technology,
AP, India.

Dr.G.Prakash Babu
Professor,
Dept of CSE,
Chiranjeevi Reddy Institute of Engineering & Technology,
AP, India.

Abstract:

Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study 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. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that GreedyIL significantly outperforms GreedyDP in terms of efficiency.

1. INTRODUCTION:

Searching is one of the common factor to know the information from the internet. Internet is one of the service providers, which provide the search result to the user with the help of the Web search engine (WSE) [1]. It employ by storing information about many web pages. WSE is a tool which allows the web user for finding information from the World Wide Web. WSE is one of the software that searches for and identifies the content or item from the web engine or web server or web database with correspond keywords or character specified by the user and finding particular sites on the World Wide Web [2]. Data search and information retrieval on the Internet has located high demands on search engines. Many search engines like Google, Yahoo provide a relevant and irrelevant data to the user based on their search. To avoid the irrelevant data the technique called Personalized Web Search (PWS) were arise. Inferring user search goals is very important in improving search-engine relevance and personalized search [3, 4]. This is based on the user profiles based on the click through log and the feedback session [5]. These data were generated from the frequent query requested by the user, history of query, browsing, bookmarks and so on. By these methods personal data were easily reveal. While many search engines take advantage of information about people in common, or regarding particular groups of people, personalized search based on a user profile that is unique to the individual person. Research systems that personalize search outcomes model their users in different ways.

The Personalized Web Search provides a unique opportunity to consolidate and scrutinize the work from industrial labs on personalizing web search using user logged search behavior context. It presents a fully anonymized dataset, which has anonymized user id, queries based on the keywords, their terms of query, providing URLs, domain of URL and the user clicks. This dispute and the shared dataset will enable a whole new set of researchers to study the problem of personalizing web search experience. It decreases the likelihood of finding new information by biasing search results towards what the user has already found.

By using these methods privacy of the user might be loss because of clicking the relevant search, frequently visited sites and providing their personal information like their name, address, etc. in this case their privacy might be leak. For this privacy issue, many existing work proposed a potential privacy problems in which a user may not be aware that their search results are personalized for them [6, 7]. It affords a host of services to people, and several of these services do not necessitate information to be grouped about a person to be customizable. While there is no warning of privacy assault with these services, the stability has been tipped to errand personalization over privacy, yet when it comes to search [8]. That approaches does not protect privacy issues rising from the lack of protection for the user data. To providing better privacy we propose a privacy preserving with the help of greedy method by providing the hybrid method of the discriminating power and prevent the information loss.

1.1 Motivations:

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 [10], [12] 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. In [9] this paper, author study this problem and provide some preliminary conclusions.

It presents a large scale evaluation framework for personalized search based on query logs and then evaluates with the click and profile based strategies. By analyzing the results, author reveals that personalized search has significant improvement over common web search on some queries but it has little effect on other queries. Author also reveals that both long term and short-term contexts are very important in improving search performance for profile-based personalized search strategies. In this paper, author tries to investigate whether personalization is consistently effective under different situations. The profile-based personalized search strategies proposed in this paper are not as stable as the click-based ones. They could improve the search accuracy on some queries, but they also harm many queries. Since these strategies are far from optimal, author will continue his work to improve them in future [10].

It also finds for profile-based methods, both long-term and short-term contexts are important in improving search performance. The appropriate combination of them can be more reliable than solely using either of them. From the author [11], they studied how to exploit implicit user modeling to intelligently personalize information retrieval and improve search accuracy. Unlike most previous work, it emphasizes the use of immediate search context and implicit feedback information as well as eager updating of search results to maximally benefit a user. Author presented a decision-theoretic framework for optimizing interactive information retrieval based on eager user model updating, in which the system responds to every action of the user by choosing a system action to optimize a utility function.

Author propose [12] specific techniques to capture and exploit two types of implicit feedback information: (1) identifying related immediately preceding query and using the query and the corresponding search results to select appropriate terms to expand the current query, and (2) exploiting the viewed document summaries to immediately re-rank any documents that have not yet been seen by the user.

Using these techniques, author develops a client side web search agent (UCAIR) on top of a popular search engine (Google) without any additional effort from the user. From the [13] author have explored how to exploit implicit feedback information, including query history and click-through history within the same search session, to improve information retrieval performance. Using the KLdivergence retrieval model as the basis, author proposed and studied four statistical language models for context sensitive information retrieval, i.e., FixInt, BayesInt, OnlineUp and BatchUp. It uses TREC AP Data to create a test set for evaluating implicit feedback models. The current work can be extended in several ways: First, it has only explored some very simple language models for incorporating implicit feedback information. It would be interesting to develop more sophisticated models to better exploit query history and click through history. For example, this may treat a clicked summary differently depending on whether the current query is a generalization

1.2 Contributions:

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 Fig. 1, 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 mclient 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.

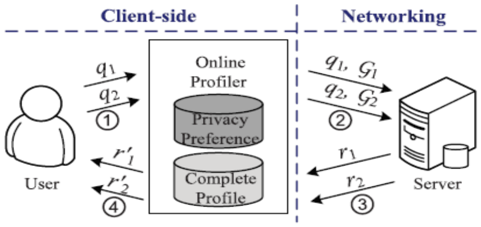


Fig. 1. System architecture of UPS.

Most of the existing works concentrate on server-side personalized search services in preserving privacy, it provide a less security to the user. To provide a security to the user from the profile-based PWS from the client side, many researchers have to deem two challenging effects during the search process of the user, (i) To increase the search quality by user profile and (ii) hide the privacy content to place the privacy risk under control. In many studies tells that user suggestions and their click based method is the helpful way to provide a personalized search and at the same time they have trouble with the loss of their privacy under their providing contents. Profile based method is an ideal case for providing the relevant search [18, 19]. Under this they were many drawbacks, it does not support on the runtime profiling, it can be based on the online and offline generalization, insufficiently protection of the data and require more iteration for obtaining relevant search.

2 RELATED WORKS:

In this section, we overview the related works. We focus on the literature of profile-based personalization and privacy protection in PWS system.

2.1 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 two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

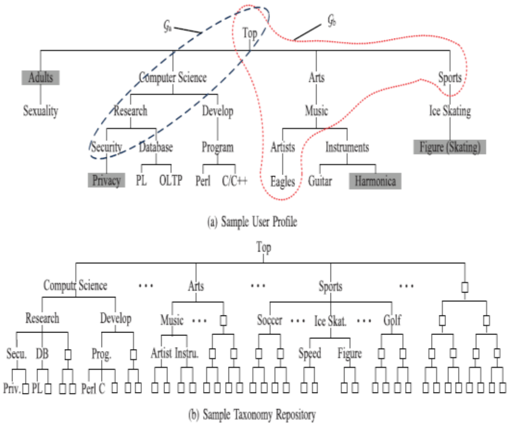
Indeed, the privacy concern is one of the major barriers in deploying serious personalized search applications, and how to attain personalized search though preserving users' privacy. Here we propose a client side personalization which deals with the preserving privacy and envision possible future strategies to fully protect user privacy. For privacy, we introduce our approach to digitalized multimedia content based on user profile information. For this, two main methods were developed: Automatic creation of user profiles based on our profile generator mechanism and on the other hand recommendation system based on the content to estimates the user interest based on our client side meta data. Above figure shows our proposed architecture which is builds in the client side mechanism and here we protect the data from the server, so only we provides a privacy to the client user. Every query from the client user were provided by the separate requests to the server, this hides the frequent click through logs or content based mechanism, from this user can protect the data from the server. In the same case our mechanism maintains the online profiler about the user hence it hides the click logs and provides a safeguard to the user data. After that, online profiler query were processed in the manner of generalization process, it is used to meet the specific prerequisites to handle the user profile and it is based on the preprocessing the user profiles. Our architecture, not only the user's search performance but also their background activities (e.g., viewed before) and personal information (e.g., emails, browser bookmarks) could be included into the user profile, permitting for the structure of a much richer user model for personalization.

3. PRELIMINARIES and PROBLEM DEFINITION

In this section, we first introduce the structure of user profile in UPS. Then, we define the customized privacy requirements on a user profile. Finally, we present the attack model and formulate the problem of privacy preserving profile generalization. For ease of presentation, Table 1 summarizes all the symbols used in this paper.

3.1 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. The repository is regarded as publicly available and can be used by anyone as the background knowledge. Such repositories do exist in the literature, for example, the ODP [1], [14], [3], [15], Wikipedia [16], [17], WordNet [22], and so on. In addition, each topic t 2 R is associated with a repository support, denoted by supRðtÞ, which quantifies how often the respective topic is touched in human knowledge. If we consider each topic to be the result of a random walk from its parent topic in R,



(a) Sample User Profile

(b) Sample Taxonomy Repository

Fig. 2. Taxonomy-based user profile.

3.2 Attack Model:

Our work aims at providing protection against a typical model of privacy attack, namely eavesdropping. As shown in Fig. 3, to corrupt Alice's privacy, the eavesdropper Eve successfully intercepts the communication between Alice and the PWS-server via some measures, such as man-in-themiddle attack, invading the server, and so on. Consequently, whenever Alice issues a query q, the 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

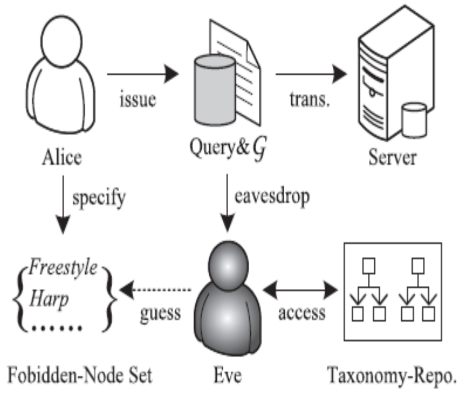


Fig. 3. Attack model of personalized web search.

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. Note that in our attack model, Eve is regarded 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. Session bounded. None of 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.

3.4 Generalizing User Profile:

Now, we exemplify the inadequacy of forbidding operation. In the sample profile in Fig. 2a, Figure is specified as a sensitive node. Thus, rsbtrðS; HÞ only releases its parent Ice Skating. Unfortunately, an adversary can recover the subtree of Ice Skating relying on the repository shown in Fig. 2b, where Figure is a main branch of Ice Skating besides Speed. If the probability of touching both branches is equal, the adversary can have 50 percent confidence on Figure. This may lead to high privacy risk if senðFigureÞ is high. A safer solution would remove node Ice Skating in such case for privacy protection. In contrast, it might be unnecessary to remove sensitive nodes with low sensitivity. Therefore, simply forbidding the sensitive topics does not protect theuser's privacy needs precisely.

To address the problem with forbidding, we propose a technique, which detects and removes a set of nodes X from H, such that the privacy risk introduced by exposing G ¼ rsbtrðX; HÞ is always under control. Set X is typically different from S. For clarity of description, we assume that all the subtrees of H rooted at the nodes in X do not overlap each other. This process is called generalization, and the output G is a generalized profile.

4 UPS PROCEDURES:

In this section, we present the procedures carried out for each user during two different execution phases, namely the offline and online phases. Generally, the offline phase constructs the original user profile and then performs privacy requirement customization according to user-specified topic sensitivity. The subsequent online phase finds the Optimal-Risk Generalization solution in the search space determined by the customized user profile As mentioned in the previous section, the online generalization procedure is guided by the global risk and utility metrics. The computation of these metrics relies on two intermediate data structures, namely a cost layer and a preference layer defined on the user profile. The cost layer defines for each node t 2 H a cost value costðtÞ _ 0, whichindicates the total sensitivity at risk caused by the disclosure of t. These cost values can be computed offline from the user-specified sensitivity values of the sensitive nodes. The preference layer is computed online when a query q is issued. It contains for each node t 2 H a value indicating the user's query-related preference on topic t. These preference values are computed relying on a procedure called query topic mapping.

5 CONCLUSIONS:

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 user profiles 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. We proposed two greedy algorithms, namely GreedyDP andGreedyIL, for the online generalization. Our experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution. For future work, we will try to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentiality, and so on), or capability to capture a series of queries (relaxing the second constraint of the adversary in Section 3.3) from the victim. We will also seek more sophisticated method to build the user profile, and better metrics to predict the performance (especially the utility) of UPS.

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.