Privacy Protection in Personalized Web Search Using UPS

Detty Susan George

Abstract-**Personalized Web Search (PWS) is introduced to improve the quality of search service on internet that helps in an effective and efficient information retrieval. But many previous studies prove that users are unwilling to disclose their personal details or information's while searching which has emerged as a major problem for explosion of PWS. Here it deals with privacy protection in PWS application where user preferences are modeled in a hierarchical manner. In this a PWS framework is proposed called UPS (User Customizable Privacy Preserving Search) that can simplify user profiles using queries while giving preference to user specified privacy requirements.**

Index Terms—**Privacy protection, personalized web search, service, user profile**

I. INTRODUCTION

The web search engine has emerged as one of the most important gateway for common peoples who look for useful information on the internet. But sometimes the users may encounter failure when the search engines returns inappropriate results that do not meet their real intentions. Such insignificance is mainly due to the massive diversity of users' backgrounds and contexts. Personalized web search (PWS) is a search technique aiming at providing improved search results, which are adapted for individual user requirements. The click-log-based methods and profile-based ones are two categories of PWS solution. The click-log based methods works in a manner by going through clicked pages in the user's query history. A strong constraint on its applicability is that it can work effectively only on repeated queries from the same user. But profile-based methods improve the search experience by profiling the user-interest. Therefore the more valuable method for all types of queries is the profile based method. The experimental outcome discovered that UPS can attain quality search results preserving user's customized privacy requirements.

A. Why Privacy Protection needed?

During the search process it considers two contradicting effects in order to provide privacy protection in user profile based PWS. Considering personalization utility of the user profile which attempt to improve the search quality. On the other hand, they need to hide the privacy contents that exist in the user profile to control the privacy risk.

Manuscript received March 02, 2015.

Detty Susan George, Computer Science& Engineering, M. G. University, MountZion College of Engineering Pathanamthitta, India, 9633404296.

Some previous studies recommend that people are agreeable to compromise privacy if the search engine yields better search quality. To address solution that helps to obtain personalization is by exposing a small (and less-sensitive) portion of the user profile, namely a generalized profile. This helps to protect privacy in user profile, without giving up the personalized search quality. In common, there is a transaction between the level of privacy protection and the search quality achieved from generalization.

II. ADVANTAGES OF UPS OVER CONVENTIONAL PWS

It allows profiling of user profiles at runtime, which optimizes the personalization utility giving preference to the user's privacy requirements; the privacy needs of user is customized; and iterative user interaction is not required.
Main achievements of this paper are highlighted as follows:
Here it proposes a UPS framework enables the user profile to be generalized for each query based on the privacy requirement as specified by the user.
Risk Profile Generalization with its NP-hardness is one of the problems of privacy preserving personalized search which is solved using two conflicting metrics for hierarchical user profile, namely personalization utility and privacy risk.
To support runtime profiling two simple and effective generalization algorithms are used, Greedy DP and Greedy IL, where the first one tries to maximize the discriminating power (DP), the second one attempt to reduce the information loss (IL).
It provides an inexpensive mechanism for the client to personalize a query in UPS to boost the constancy of the search results while avoiding the unnecessary disclosure of the profile.

III. WORKING OF UPS

The UPS framework assumes that the queries received from the user do not have any sensitive information, and helps at protecting the individual user profiles privacy while preserving their usefulness for PWS. The main components of UPS framework are a non trusty search engine server and many numbers of clients. Each individuals or client (user) accessing the search service do not trust anyone else except themselves. The major component or element for privacy protection in UPS framework is an online profiler which is implemented as a search proxy that runs over the client machine itself. The role of the proxy is to maintain the complete user profile, in a hierarchical structure of nodes including its semantics, and as well as the user-specified or customized privacy requirements which is represented as a set of sensitive-nodes.

For each user, the framework adopts mainly two stages, namely the offline and online phase. In the offline phase, a user profile is constructed in a hierarchical manner and the user-specified privacy requirements are customized or personalized.

In the online phase the queries are handled as follows:

1. When a user issues a query on the client side, the proxy will generates in the light of query terms a user profile in runtime. As a result a generalized user profile fulfilling the privacy requirements is produced. Here the generalization process is guided by two contradictory metrics, namely the privacy risk and the personalization utility where both are defined for user profiles.
2. Consequently, the generalized user profile and the query profile are sent collectively to the PWS server for customization or personalized search.
3. Then result returned by the search is customized with the user profile which is delivered back to the query proxy.
4. Lastly, the proxy either presents the result to the concerned user, or re ranks the result with the complete user profile. In personalized web services UPS adopts a hierarchical structure for each individual user profile.

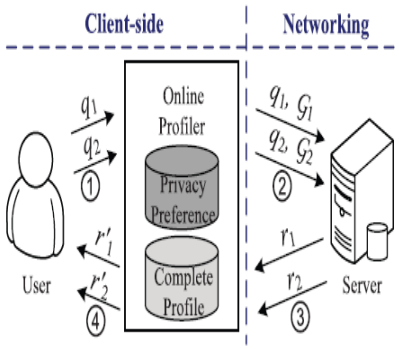

Fig: System architecture of UPS

Based on the availability of a public accessible taxonomy, denoted as R, the profile is constructed. The customized privacy requirements in the user profile can be represented with a number of sensitive-nodes (topics), which can introduce privacy risk to the user when exposed to the server. To address the problem of exposing user profile here it proposes a technique, which detects and removes a set of nodes X from user profile H, such that the privacy risk introduced by revealing user profile is always under control. The sensitive nodes S and Set of nodes X are normally different. For simplicity of description, it can be assumed that all the sub trees of user profile H rooted at the nodes in X do not overlap with each other. This process is known as the generalization, and the resultant output G is said as the generalized profile. The offline generalization might generate an output that may contain many topic branches, which are inappropriate to a query. A good solution requires online generalization, depending on the queries given. Online generalization helps to removes noisy topics that are irrelevant to the current query and also prevents unnecessary privacy disclosure of user profile. The personalization utility is to be monitored or controlled during the generalization is an important factor. By using the running example, the user profiles can be generalized into a smaller rooted sub trees. If overgeneralization occurs it may cause uncertainty in the personalization, and ultimately this can result in the poor search results. This issue can be addressed using runtime generalization which monitors the utility safely.

A. UPS Procedures

The offline and online phases are two different execution phases it carries out for each user. Generally, the original user profile is constructed during the offline phase and then performs customization of privacy requirement according to the user-specified topic sensitivity. Then subsequently in the online phase it identifies the solution for optimal Risk Generalization in the search space which can be determined by the customized user profile.

The global risk and utility metrics guides online generalization process. The computation of these metrics depends on two transitional data structures, namely a preference layer and a cost layer defined on the user profile. The cost layer is used to define the total sensitivity at risk caused by the disclosure of each node. From the user-specified sensitivity values of the sensitive nodes these cost values can be calculated offline. The preference layer is determined during the online phase when a query is issued to the client machine. It contains a value indicating the user's preference on query-related topic. These preference values are calculated depending upon a procedure called query topic mapping. In particular, each user has to follow the below procedures:

- profile construction in offline,
- customization of privacy requirement in offline,
- query-topic mapping in online, and
- generalization in online.

The generalization technique involves two critical metrics, namely metric of utility and metric of privacy. The search quality in revealing the user's intention of the query on a generalized profile can be predicted using the utility metric. Since the search quality depends largely on the execution of PWS search engine which is hard to predict therefore it not directly measured. In addition, to request user feedback on search results it is too expensive.

To propose the model of utility, here it introduces the concept of Information Content (IC), which describes how specific is a given topic. Now, the first module of the utility metric is developed called Profile Granularity (PG), which is the probability distributions of the topic domain with and without revealing the generalized profile of the KL-Divergence. The second component of utility is Topic Similarity (TS), which is used to measure the semantic similarity among the topics.

The total sensitivity contained in generalized profile, which is given in normalized form is defined by privacy risk of exposing the profile. If there exist distinct queries, to which the profile-based personalization may reduces the search quality or even responds little, while exposing the profile to a server. To personalize a query an online mechanism is developed to decide whether to personalize a query. The fundamental proposal is simple, when a distinct query is recognized during generalization, the complete runtime

profiling will be aborted and the query will be sent without a user profile to the server. The discriminating power helps to identify the distinct queries.

B. Implementation issues

There exist some open problems in the UPS process, this can be solved using a mechanism called an inverted-indexing mechanism for computing the query topic relevance. The publicly available repositories permit the editing as well as manual tagging on each topic. These topics contains textual data which consist of a document repository, which allows each leaf topic to identify its associated document set. Each document in document repository is assigned to one leaf topic only. Thus, it is possible to generate an inverted-index for each leaf topic, which contains entries such as term; doc id; topic id for all the documents. At the end, a hierarchy of inverted indices is obtained, where all the documents within the taxonomy will be contained in the inverted index file. Thus this structure enables each user to resourcefully process keyword search and retrieval. Specifically, the root index files are able to maintain the entire document set that can sustain term-based topic searching in repository.
During the Offline-1 procedure, it is needed to detect for each document the respective topic in repository. For this a naive method is to compute the relevance for each pair of document and their topic to repository with a discriminative naive Bayesian classifier (dnb). The topic that exhibits with the largest dnb value is considered the result of the search. But, if many of the topics in repository are not relevant to the documents then the naive method is inefficient.
Exploiting the user's click log to be the set of document will be a more efficient way (and the one used in this implementation). The click log contains entries such as query in the log and document clicked by the user after issuing a query. Thus, this allows reducing the necessity of computing the topics that are retrieved by the query from the topmost inverted index and then all documents relevant to the query are retrieved from the inverted index and their associated topics are obtained from the topic id. Then, the dnb value for each topic is computed.

During Online-1 the computation of query-topic relevance takes place. If a query is given, which is retrieved from inverted index then the documents relevant to query is determined using the conventional approach. Using their respective topics these documents are then grouped. The number of documents contained in each topic is computed from the relevance of each topic.

Thus the relevance metric used in the projected implementation is very easy and fast to estimate. More complicated versions can be used to easily replace it. Usually, Greedy IL is used to trace the information loss instead of the discriminating power. This helps to save a lot of computational cost, avoids redundant iterations and further simplifies the calculation of IL**.**

IV. CONCLUSION

Here it introduced a framework called UPS, a client-side privacy protection for personalized web search. UPS is able to be adopted by any PWS that models user profiles in a hierarchical taxonomy. Using this framework user can specify customized privacy requirements through the hierarchical profiles. Additionally, to protect the personal privacy without compromising the search quality, UPS also performs an online generalization on user profiles. For the online generalization it has proposed two greedy algorithms, namely Greedy DP and Greedy IL. The experimental results also discovered that UPS could attain quality search results and also preserve user's customized privacy requirements.

ACKNOWLEDGEMENT

I would like to extend my gratitude to the reference authors, as well as reviewer of our paper.

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. Gath, "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.