ISSN: 2321-8169
6643 - 6646

Providing Customized Requirements for Privacy Preserving In Web Search Engines

D. Veerendra
Computer Science Engineering
Baba Institute of Technology and Sciences
Andhra Pradesh, India
e-mail: d.veerendra20@gmail.com

A. V. S. Pavan Kumar
Asst.Prof, Computer Science and Engineering
Baba Institute of Technology and Sciences
Andhra Pradesh, India
e-mail: avspavankumar@bitsvizag.com

Abstract— Web search tools are generally used to get information from web servers. These web crawlers use client profiles and as of late sought information to give indexed lists, so here there is no security insurance for client information. We give an framework that can assist clients with customizing their protection necessities. The protections prerequisites gave by client are utilized here for querying the Web server with the hunt keys given by client. In this methodology we can ready to conceal the protection information of client from web search servers. The procedure of modifying protection necessities will happen in offline and will be utilized dynamically .The calculations utilized here will give speculation inquiries expected to look by safeguarding security prerequisites gave. The Experimental results will prove that this Framework will ensure privacy of client.

Keywords- *Privacy protection, personalized web search, utility, risk, profile*

*****

I. INTRODUCTION

The web search plays a vital role for looking information on the web. Sometimes these may return irrelevant results due to variety of users, contexts. Personalized web search is a general technique used to provide better search results.

There are two methods for performing personalized web search. First one is Click-log based and second is Profile-based. Click-log based methods simply focus on clicked pages in user's history. Profile-based methods works with profile created for each user based on profiling techniques. Profile-based method will work more efficiently when compared to click-log based. This Profile-based search will provide all users data stored in history, bookmarks, and documents. So, it is transferring privacy data of user to servers. This has become the main barrier for wide proliferation of PWS services. To protect privacy in profile-based search, we need to hide data in query requests, but ensure search results will not give irrelevant data.

The framework works in two stages, offline and online. During offline user profile is developed and added with user specified privacy requirements. The online phase handles queries as

1. When a user provides query ‘q’ on the client, the framework generates a user profile in light of query terms. It generates a generalized G profile satisfying privacy requirements.
2. The query and generalized profile are sent to Web search server for results.
3. The search results are sent to query proxy.
4. Finally, proxy provides the same search results or else re-ranks them with user profile.

UPS can be distinguish from conventional PWS in three areas.1) providing runtime profiling which optimizes personalization utility with respecting privacy requirements.2) allowing customization of privacy requirements 3) it does not require iterative user interactions.

We propose a framework which can generalize profiles according to user-specified privacy requirements.

By the definition of two conflicting metrics, personalization utility and privacy risk, for hierarchical user profile, we define problem of privacy-preserving personalized search. We define two algorithms Greedy DP and Greedy IL to support runtime profiling. We also provide mechanism for client to decide whether to personalize or not.

II. RELATED WORK

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

Pprofile-Based Personalization

Profile-Based searches are mainly implemented to improve the search quality. The solutions to PWS is done by representation of profiles and measure of effectiveness of results.

There are many strategies used to develop these profile searches,earlier they used lists/vectors to represent the profile.But recent works build profiles based on hierarchical structures due to their strong abilities,better efficiency.To reduce the human involvement in performance measuring, researchers also propose other metrics of personalized web search that rely on clicking decisions,including Average Precision (AP) [19], [10], Rank Scoring [13],and Average Rank.

To measure the performance of the proposed framework,we used metrics Average Precision..

A. Privacy Protection in PWS System

There are two areas we need to ensure privacy in PWS systems, privacy in identifying individuals, and second exposing of user profile data to PWS server.

Identifying individual's problem can be solved by including pseudo identity, no identity and no personal information. Using this approach, the bond between query and user is broken. As a result we cannot identify a certain individual. So instead of relying on third-party assistance we provide a mechanism which provides user profile from client side. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted sub tree of the complete profile. The profile provided in this framework will work as per degree of sensitivity provided by user in form of guarding nodes in the taxonomy.

Queries with smaller keys will work distinctly and they benefit with personalization, while with higher values will not. Therefore use of personalization in some queries will become questionable. Therefore we propose a prototype of UPS, collectively with a greedy algorithm to support profiling based on metrics of personalization utility and privacy risk.

This paper is enhancement for previous work, which is proposed as prototype of UPS, together with GreedyDP to support online profiling. In this paper we extend implementation of UPS and also the metric of personalization utility to capture three new observations.

III. CREATING CUSTOMIZED USER PROFILE AND PRIVACY REQUIREMENTS

User profiles in this framework are represented as hierarchical structure. The profile is constructed based on taxonomy which satisfies assumptions. The Repository R indicates a total domain of human knowledge, and ‘t’ represents a selected topic then the ‘t’ can be found in R, with the sub tree subtr(t,R).

The repository is indicated as publicly available and can be used by anyone as the background knowledge. Each topic (t belongs R) is associated with a repository support, denoted by supR(t). If we consider each topic to be the result of a random walk from its parent topic in R .

Assumption2 will not be taken into consideration if there are no support values available. That is, $sup_R(t)$. can be calculated as the count of leaves in subtr(t,R). In this we create a model for domain of human knowledge. In this model, repository R is a hierarchical partitioning of universe.

Definition1: A user profile is a hierarchical representation of user interests, is a rooted sub tree of R.

Definition2 : A Rooted sub tree can be calculated based on two trees S and T, S is a rooted sub tree of T and if S can be generated from T by removing a node set. Although a user profile H inherits from R a subset of nodes and links, it will not duplicate the repository supports.

Customized privacy requirements are specified based on number of sensitive nodes in the profile. In this user can choose the sensitive nodes which need not to be sent to server while querying. we define a mechanism where user can provide sensitivity node which should not be exposed to server. As the sensitivity values indicate the privacy values, these are directly removed from sub trees rooted at all sensitive nodes.

A. Attack Model

We should also ensure about the protection against a privacy attack, called as eavesdropping. Here eavesdropper successfully intercepts communication between the client and PWS-server via some measures called as man in the middle attack. Here Alice issues a query q, then entire copy of q and runtime profile G are captured by Eavesdropper. Based on G, Eve will try to touch the sensitive nodes of alice by recovering the segments hidden from the original.

In Attack model Eve is satisfying the assumptions. Knowledge bounded represents the background knowledge of the adversary limited to taxonomy repository. Here both profile and privacy are defined based on R. Session bounded represents the captured information available for tracing the same victim in long duration. The above assumptions are strong but are reasonable in practice.

If we consider the sensitivity of topics as cost of recovering it,then privacy risk is defined as total sensitivity of nodes,which can recover from G. Our approach for privacy protection of PWS is to keep privacy under control.

IV. GENERALIZED USER PROFILE

To address the problem of forbidding, we propose technique which identifies and removes a set of nodes ‘X’ from H, such that the privacy risk will always be under control. We should also ensure that all the sub trees of ‘H’ rooted at the nodes in ‘X’ do not overlap each other. This process is called generalization and the generated output is called generalized profile. This generalization technique can be conducted in offline processing without involving user queries. It is not practically possible to do this in offline processing due to two reasons.

1) The output which gets from offline generalization may contain topic branches,which may not be relevant to a query.Efficient solution requires online generalization,which depends on queries.Online generalization not only avoids irrelevant privacy topics,but also removes noisy topics that are not relevant to current query.

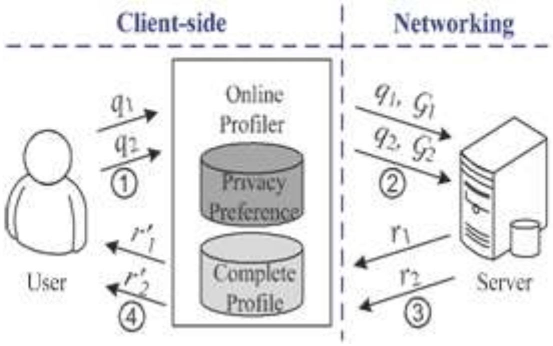
Example,given a query which consists of all the data { Adults, Privacy, Database, Develop, Arts, Sports,Computer science,Instrument}, there we can provide a node sets to be removed as {Adults,Arts, Instrument} then these are removed from the root set and sends to the server.

2) Monitoring the personalizing utility is important during generalization. These generalized profiles are the smaller rooted sub trees. Overgeneralization will cause ambiguity in the personalization and eventually may lead to poor search results. Monitoring would be possible only if we perform generalization at runtime

Note that metric risk(q,G) and util(q,G); G only depend on the instance of G and the query q as they are implemented to predict the privacy risk and personalization utility of G on q, without any user feedback.

V. PROPOSED SYSTEM

All the problems in the PWS system are addressed in this Framework. This Framework assumes that these queries do not have any sensitive information and protects the privacy in user profiles.



As illustrated in Fig, Framework consists of a no trusty search server and a number of clients. Each client accessing this service trusts no one else except himself. The key component for privacy protection is an online profiler implemented as a search proxy running on client machines. The proxy maintains both complete user profile as well as privacy requirements represented as set of sensitive-nodes.

Here we present the procedures which are carried for users in two phases, namely offline and online phases. Offline phase constructs user profile and then performs privacy requirement customization according to user specified topic. Each user will undergo following procedures in our solution

Profile construction:

This process will take place in offline phase. To construct the profile we take the following steps.

1) Detect the respective topic in R for every document d 2 D. Thus, the preference document set D is transformed into a topic set T.
2. Construct the profile H as a topic-path trie with T i.e., H=trie (T).
3. Initialize the user support supH(t) for each topic t 2 T with its document support from D, then compute SupH(t) of other nodes of H with (4).

Privacy Requirement Customization:

This process requests the user to provide sensitive node set and the respective sensitive-node value sen(S)>0 for each topic s belongs S.
1. For each sensitive-node,cost(t)=sen(t);
2. For each no sensitive leaf node, cost(t)=0;

Query Mapping:

The purpose of query-mapping is to compute a rooted sub tree of H,which is called a seed profile, so that all topics relevant to q will be there.
1. Findthe topics in R that are relevant to q.
2.Overlapping R(q) with H to obtain the seed profile G_0 which is also a rooted sub tree of H.

Profile Generalization:

This profile generalizes the seed profile G_0 in accost-based manner depending on privacy and utility metrics. This procedure computes the discriminating power for online decision making.

We should also consider about eavesdropping while designing the framework.The eavesdropper successfully intercepts the communication between Alice and the server and soon. We must ensure the attack model by following the assumptions

Knowledge bounded.

The background knowledge of theadversary 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 longduration. In other words, the eavesdropping will be started and ended within a single query session.

VI. TECHNIQUES AND ALGORITHMS FOR GENERALIZATION

In this we introduce two metrics for generalization problem. Metric of Utility is to predict the search quality of theQuery q on a generalized profile G. We transform the utility prediction to the estimation of discriminating power of a given query q on profile G. Metric privacy is defined as the total sensitivity given in normalized form .The sensitive nodes here are pruned during the generalization and evaluate the risk of exposing the ancestors. This can be done using cost layer computing during offline.

Whether to personalize or not: Online mechanism is implemented to decide whether to personalize a query or not. We consider queries with good DP even the client does not expose any profile. There are benefits in making runtime decision, it enhances stability of search quality and avoids the unnecessary exposure of user profile.

Algorithm:

GreedyIL: To increase the efficiency GreedyIL algorithm is used [7].

Following terminologies are used in GreedyIL algorithm. G0: Seed profile

q :query

δ : Privacy Threshold.

G*: Generalized profile satisfying δ- Risk.

Q: IL-priority queue of prune-leaf decision.

i: Iteration index initialized to 0.

Input is G0, q, δ.

Output: G*.

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

```
If DP (q,R) < μ then do following:
Obtain the seed profile G0 from Online-1, Insert (t,IL(t)) into Q
   for all to ε T(q)
While risk (q,Gi) > δ do
Pop a prune-leaf operation on t from Q
Set s ←part (t,Gi)
Process prune leaf Gi→ Gi+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←i+1 returnGi as G*
return root(R) as G*
```

The Greedy IL Algorithm improves the efficiency of generalization based on findings. one important finding is prune-leaf operation reduces the discriminating power of profile. Algorithm1 shows the pseudo code of Greedy IL algorithm, it traces the information loss instead of discriminating power. It saves a lot of computational cost. It also avoids unnecessary iterations. To study the efficiency of proposed generalization, we perform Greedy IL algorithm on real profiles. The queries are randomly selected from their query logs. We present results in terms of average number of iterations and the response time of generalization. Scalability of proposed algorithms can be studied by the seed profile size and the data set size. We choose 100 queries from AOL query log, and take their respective R(q) as their seed profiles.

VII. CONCLUSION

In this paper, we present a client-side privacy mechanism for PWS systems. It can be used by any PWS that uses user profile in hierarchical manner. Our framework also performs online generalization on user profiles to protect the privacy without compromising on search quality. Two algorithms are proposed namely Greedy DP and Greedy IL for online generalization. Experiments proved that we can achieve quality search results by preserving customized privacy requirements. The results confirmed the effectiveness and efficiency of solution. In future, we also seek to provide more sophisticated method to build user profiles and better metrics to predict the performance.

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] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.

[5] X. Shen, B. Tan, and C. Zhai, “Privacy Protection in Personalized Search,” SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[6] Y. Xu, K. Wang, G. Yang, and A.W.-C. Fu, “Online Anonymity for Personalized Web Services,” Proc. 18th ACM Conf. Information andKnowledge Management (CIKM), pp. 1497-1500, 2009.

[7] Y. Zhu, L. Xiong, and C. Verdery, “Anonymizing User Profiles for Personalized Web Search,” Proc. 19th Int’l Conf. World Wide Web (WWW), pp. 1225-1226, 2010.

[8] J. Castellı´-Roca, A. Viejo, and J. Herrera-Joancomartı´, “Preserving User’s Privacy in Web Search Engines,” Computer Comm., vol. 32, no. 13/14, pp. 1541-1551, 2009.

[9] A. Viejo and J. Castell_a-Roca, “Using Social Networks to Distort Users’ Profiles Generated by Web Search Engines,” Computer Networks, vol. 54, no. 9, pp. 1343-1357, 2010.

[10] X. Xiao and Y. Tao, “Personalized Privacy Preservation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD),2006.

AUTHORS:

D. VEERENDRA holds a B.Tech certificate in Computer Science Engineering from the University of JNTU Kakinada. He presently Pursuing M.Tech (CST) department of Computer Science Engineering from Baba Institute of Technology and Sciences, Visakhapatnam.

A. V. S. PAVAN KUMAR is an Assistant Professor in the Department of Computer Science and Engineering in Baba Institute of Technology and Sciences. He awarded his M.Tech [Computer Science and Technology] degree from Gitam University. He is Pursuing Ph.D in Gitam University in the field of Data mining.