iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 Issue: XI Month of publication: November 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Privacy Protection Using CPS Framework in Personalized Web Search

Priyanka Zaware[#1], Satish R. Todmal[#2]
[#1]Department of Computer Engineering, Imperial college of Engineering,
Savitribai Phule Pune University, India
[#2] Head of Department, Department of Information Technology Engineering, Imperial college of Engineering,
Savitribai Phule Pune University, India

***Abstract*— Personalized web search (PWS) is efficient in improving the search quality on internet. But users generally hesitate to disclose their personal information during search which has become a hindrance in wide increase in PWS. We model user preferences into user profiles. We proposed a PWS framework called CPS while performs profile generalization without affecting users' privacy requirements. This generalization maintains a balance between personalization and privacy risk while a generalized profile is exposed. We propose two algorithms namely GreedyDP and GreedyIL helping us in runtime generalization. We provide an online prediction mechanism to decide whether personalizing a query is helpful.**
***Keywords*— Personalized Web Search, Privacy protection, Utility metric, Privacy risk, Profile**

I. INTRODUCTION

Now a day, the web search engine becomes a very important gateway for common people looking for useful information on the internet. Sometimes these search engines gives users unrelated information that do not complete their real requirements. This irrelevance is due to a very large number of user contexts as well as very much open text information. Personalized web search is a general type of search techniques which provide better search results and meet individual customer needs. As the expense, by collecting and analyzing user information we know the user intention behind the input query.

Personalized web search is of two types, one is click log based method and another is profile based method. In the click log based method - they simply done inclination to clicked pages in the query history. Although this strategy has been worked well [1], a Click log based method can work only on repeated queries. This is the limitation of this strategy. Profile based methods are effective for all types of queries, but are seen to be unstable under some conditions. Profile based methods improve search results with complicated user models created from user profiling method.

There are advantages and disadvantages for both types of PWS methods, the profile based PWS is more effective in improving the quality of web search. By using personal and behavior information to profile users, which are normally gathered from user browsing history [2], [3], user query history [4], [5], [6], bookmarks[7], click through data [1], [8], [9], user documents [4], [10] and so forth. Unluckily this indirectly collected personal information can easily disclose a user's private life completely. Privacy issues not only create anxiety in users but also reduce enthusiasm in offering personalized search.

For protecting user's privacy in personalized web search researchers have to consider two things in search process. One is, they have to improve search quality using personal query of the user. Another thing is, they have to hide private information available in the user's profile in order to keep privacy risk under control. But a few previous studies [10], [11] show that people are ready to compromise privacy if they get better search results by supplying user profile. In an ideal case, significant amount of information can be obtained at the cost of only small part of user profile called generalized profile. In this way user privacy can be protected without any compromise. There is a balance achieved between search quality and privacy protection get from generalization.

The existing works of privacy preserving PWS are not satisfactory. The problems with the existing system are as follows:

1. The old profile based PWS do no generalize profile at runtime. A user profile is generalized only once and that is in offline mode and queries used to personalize are from the same user. One issue reported is [1] that sometimes profile based personalization method does not support ad hoc queries. Online profiling is the better approach but no previous work has supported this.
2. The existing system does not consider modification in privacy requirements. In this user privacies are not protected properly.
3. While creating personalized search results many existing personalization techniques require repeated user interactions.

All the above problem of the system are resolved in our CPS (Client customizable Privacy-preserving search) framework. The framework works with the assumption that the queries are without any sensitive information, aiming to protect not only the privacy of individual users but also retaining their usefulness for PWS.

Framework generally works in two phases' offline phase and online phase.

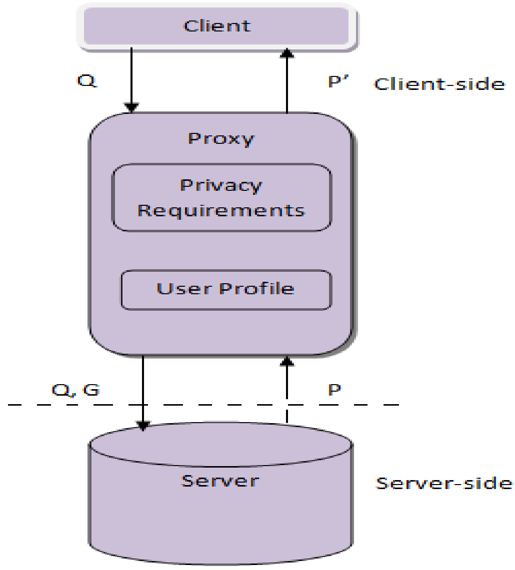

Fig. 1. Architecture of CPS

Offline phase: During this phase a hierarchical user is constructed and customized with user-specified privacy requirements.

Online phase: This phase handles queries as below:

1. When a user issues a query Q on the client, a generalized user profile G is created by the proxy satisfying the privacy requirements.
2. The query Q and the user profile G are then sent together to PWS server for personalized web search.
3. Search results personalized according to the profile are sent back P to the query proxy.
4. At last, the proxy either presents raw results P to user or ranks them P' with the entire user profile.

Our main contributions are as listed below:

1. A privacy-preserving personalized framework CPS, which generalizes profiles for every query based on user specified privacy requirements.
2. The two metrics namely personalization utility and privacy risk are taken into consideration and we formulate the problem of privacy-preserving personalized search as δ risk profile generalization.
3. Two simple and efficient algorithms GreedyDP and GreedyIL are developed to facilitate runtime profiling.
4. The client can decide to personalize a query in CPS before each runtime profiling.

II. RELATED WORK

This section focuses on the literature of profile-based personalization and privacy protection in PWS system.

A. Profile-Based Personalization

For a better search results we use profile based personalization. To facilitate different personalization strategies many profile representations are available. Most of the hierarchical representations are constructed with weighted topic hierarchy. Our framework does not focus on the implementation of user profiles, it can efficiently implement any hierarchical representation based on knowledge taxonomy.

In order to reduce human participation in performance measuring, researchers have proposed other metrics of personalized web search like Average Precision [12], [10], Rank Scoring [13], and Average Rank [5], [9]. We use the Average Precision metric proposed by Dou et al. [1], which measures effectiveness of personalization in CPS. We propose two predictive metrics, namely metric of utility and metric of privacy, on a profile without requesting user feedback.

B. Privacy Protection in PWS System

There exist two classes of privacy protection problems for PWS. One class contains those which treat privacy as the identification of a user. The other class considers the data sensitivity, mainly user profiles, exposed to the PWS server.

Xu et al. [10] proposed a privacy protection mechanism for PWS system based on hierarchical profiles. A user-specified threshold obtains a generalized profile as a rooted subtree of the complete profile. An important property that differences our work from [10] is that we provide personalized privacy protection in PWS. Degree of privacy protection is specified by a user, specifying his/her sensitive values by specifying “guarding nodes” in taxonomy of the sensitive attribute. Users are allowed to customize privacy needs in their hierarchical user profiles. Queries having smaller click-entropies, like distinct queries generally benefit more from personalization, which is not the same for those with larger values. Since the latter may cause privacy disclosure, hence personalization becomes questionable for such queries.

In our CPS framework, a client side solution is used to distinguish distinct queries from ambiguous ones, this solution uses a predictive query utility metric. In this paper we provide a detail implementation of CPS. We refine the evaluation model of privacy risk and also provide a new profile generalization algorithm called GreedyIL.

III. PROPOSED APPROACH

To overcome problems with existing system, we have proposed new techniques for privacy protection in user profile generalization.

A. User Profile

Generally each user profile in CPS, adopts a hierarchical structure. Our profile is constructed based on public accessible taxonomy. As the taxonomy is considered to be publicly available, hence can be used by anyone as background knowledge. Taxonomies existed in the literature, for example, the ODP [1], [14], [5], [15], Wikipedia [16], [17], WordNet [18] and so on. A user profile is a hierarchical representation of user interests is a rooted subtree of the taxonomy.

B. Personalized privacy requirements

Personalized privacy requirements are specified with different sensitive topics in the user profile, which on disclosure to the server introduce privacy risk to the user. A user’s privacy concerns vary from one sensitive topic to another. A user may hesitate to share his/her personal interests to avoid various advertisements. For addressing the differences in privacy concerns, we allow the user an ability to specify sensitivity for each topic. Sensitivity values indicate a user’s privacy concerns, a simple privacy protection method is to remove subtrees rooted at all sensitive topics whose sensitivity values are greater than a threshold. This method is called forbidding.

C. User Profile Generalization

Removing topics with low sensitivity can be unnecessary. Hence, simply forbidding the sensitive topics do not protect the user’s privacy needs. In order to solve this problem with forbidding, we propose a new technique. This technique identifies and removes set of topics from user profile such that the privacy risk is under control. This process is called generalization, and the output of this process is a generalized profile.

Generalization is classified into offline generalization and online generalization. Offline generalization is performed without involving user queries. However it is impractical to perform offline generalization because the output in this process may contain topic branches irrelevant to a query. Online generalization avoids unnecessary privacy disclosure and also removes topics irrelevant to the current query. Overgeneralization causes ambiguity in personalization, leading to poor search results. The problem of privacy-preserving generalization in CPS is defined based on utility and risk. Utility measures the personalization utility of generalized profile, while risk measures the privacy risk of exposing the profile.

D. CPS Procedures

The procedures are carried out for each user during two different execution phases, namely offline and online phases.

1. Original user profile construction in offline phase – The original user profile is built in a topic hierarchy that shows user interests. User’s preferences are stored in a set of plaintext documents.
2. Privacy requirement customization in offline phase – This step takes sensitive topic and its sensitive value for each topic from the user. Customized profile is then obtained from these values.
3. Query-topic mapping in online phase – Query-topic mapping computes rooted subtree called ‘seed profile’ so that all topics related to query are included in it and obtains the preference values between a query and all topics in user profile.
4. Profile Generalization in online phase – This process generalizes the seed profile in a cost-based iterative manner depending on privacy and utility metrics. Also this process calculates the distinguishing power on online decision on whether personalization should be employed.

E. Generalization Metrics

1) Utility Metric: This metric predicts the search quality of the query on a generalized profile. We transform the utility prediction problem to the analysis of distinguishing power of a given query on a generalized profile. Similar assumption has been made in [11] to model utility, but this metric cannot be used in our problem settings, as we have a profile with hierarchical structure instead of flat one.

2) Privacy Metric: When a generalized profile is exposed the total sensitivity contained in it in normalized form is defined as privacy risk. If original profile is exposed, the risk of exposing all sensitive topics is maximum.

F. Profile Generalization Algorithms

1) Brute Force Algorithm: Most favourable generalization is created by generating all rooted subtrees of our seed profile by using Brute Force algorithm and the subtree with best utility is taken as the result.

2) GreeedyDP Algorithm: We apply this algorithm on generalized profile. We remove the leaf topic of this profile to generate optimal profile. Algorithm works in a bottom up manner. With the repeated iterations we generate the profile with maximum distinguishing power and satisfying δ risk constraint. And this is the final output of GreedyDP algorithm.

3) GreedyIL Algorithm: GreedyIL algorithm reduces the information loss. When δ risk is satisfied stop the iterative process and this reduces the computational cost. Then it simplifies the computation of information loss. It reduces the need of information loss recomputation.

IV. CONCLUSIONS

We propose CPS framework for personalized web search which do client side privacy protection. Any personalized web search having user profiles in hierarchical structure can use this framework. This framework provides some functions to the user like- to specify their privacy requirement, online decision making. CPS framework protects users personal privacy without compromising the search result. We use GreedyDP and GreedyIL algorithm for online generalization.

V. FUTURE WORK

We will try to find valuable relationship among the topics. We will try to find more appropriate metric to predict the performance and find better method to construct user profile.

VI. ACKNOWLEDGMENT

I take this opportunity to thank all in individuals for their guidance, help and timely support .It gives me great pleasure and immense satisfaction to present this paper. Which result of unwavering, Support expert guidance and focused direction of my guide Prof. Satish R. Todmal to whom I express my deep sense of gratitude and humble thanks, for valuable guidance throughout the work.

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] 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.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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..

[9] 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.

[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] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.

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

[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. Kohlschu¨ tter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[15] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.

[16] E. Gabrilovich and S. Markovich, "Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI), 2006.

[17] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.

[18] 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.