ISSN:0975-9646

Privacy –Enhanced Web Personalization

Rupali Keshavrao Aher, Akshay Rajdhar Adik

Department of Computer Engineering,
MCERC,Nashik, India.

Abstract- **for users with individual information goals web personalization is used to improve search quality by customizing search results, based on the personal data of user provided to the search engine. Users are not comfortable with disclosing private preference information to search engines, but if there is gain in service or profitability to the user then privacy can be compromised. Thus, there should be a balance between the search quality and privacy protection. A PWS framework called User Customizable Privacy Preserving Search (UPS) generalizes profiles by queries when the user specifies privacy requirements. Runtime generalization is used for providing a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of disclosing the generalized profile. Two greedy algorithms, namely Greedy discriminating power algorithm (GreedyDP) and Greedy Information Loss algorithm (GreedyIL), are used for runtime generalization. GreedyIL algorithm achieves high efficiency than the GreedyDP algorithm. Online prediction mechanism is used for deciding whether personalizing a query is beneficial. Session attacks like eavesdrops attacks are controlled.**

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

I. INTRODUCTION

The web search engine is widely used by the users for searching useful information on the web. But the amount of information on the web grows continuously so it becomes very difficult for web search engines to find information that satisfies user's individual needs. Due to the enormous variety of user's contexts and backgrounds, as well as the ambiguity of texts, search engines return irrelevant results that do not meet the users real intentions. For providing better search results a general category of search techniques, personalized web search (PWS) is used. To figure out the user intention behind the issued query, user information has to be collected and analyzed.

There are two types of solutions to the PWS

1) Click-log-based method:

This is a straightforward method. The click-log based methods uses clicked pages in the users query history. But it has strong limitation that it can only work on repeated queries from the same user [2].

2) Profile-based methods:

Profile-based methods can be used effectively for almost all sorts of queries, but under some circumstances the results are unstable [2]. It improves the search experience with complicated user-interest models generated from user profiling techniques.

There are pros and cons for both types of PWS techniques, but profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users. It is usually gathered implicitly from query history[3],[4],[5], browsing history[6],[7], click-through data[8],[9],[2], bookmarks[10], user documents[3],[11], and so forth. Unfortunately, such implicitly collected personal data can easily disclose a span of user's private life. Privacy issues are raised from the lack of protection for such data, for instance the AOL query logs scandal [12], raise panic among individual users, and also dampen the data-publishers enthusiasm in offering personalized service. So the privacy concerns have become the major barrier for wide proliferation of PWS services.

Existing system have a privacy-preserving personalized web search framework UPS. User specifies the privacy requirements and according to the requirements user profiles are generalized. The problem of privacy-preserving personalized search is formulated as δ-Risk Profile Generalization, by using two conflicting metrics, personalization utility and privacy risk, for hierarchical user profile. Two simple and effective generalization algorithms, GreedyDP and GreedyIL are developed, which support runtime profiling. GreedyDP tries to maximize the discriminating power (DP), and the GreedyIL attempts to minimize the information loss (IL). To enhance the stability of the search results and to avoid the unnecessary exposure of the profile an inexpensive mechanism is used for deciding whether to personalize a query in UPS. UPS allows customization of privacy needs; and it does not require iterative user interaction.

II. RELATED WORK

User profiles disclose the individual information goals so to improve the search quality, profile based PWS refers the user profile. Term list/vectors [6] or bag words [3] are used previously to represent the profile. Hierarchical structures are commonly used to build the profiles as they provide higher access efficiency, stronger descriptive ability, and better scalability. Hierarchical profiles build automatically by using term frequency analysis of the user data [11]. Weighted topic hierarchy/graph such as ODP [2][13][15], Wikipedia [15][16] are used for constructing hierarchical profiles. Normalized Discounted Cumulative Gain (nDCG) is a common measure of the effectiveness of an information retrieval system but it requires high cost in explicit

feedback collection. Other metrics of personalized web search rely on clicking decisions, including average rank [4][9], Rank Scoring and Average Precision[19][11] which reduces human involvement in performance measuring. To measure the effectiveness of the personalization in UPS we used average precision metric [2], and two predictive metrics, personalization utility and privacy risk on a profile instance without requesting for user feedback.

One class of Privacy protection problem for PWS treats privacy as the identification of an individual [18]. It try to solve the privacy problem on different levels, pseudoidentity, the group identity, no identity, and no personal information. Due to the high cost in communication and cryptography the third and fourth levels are impractical. First level solution is proved to fragile [12]. By generating a group profile of k users [19] and [20] provide online anonymity on user profiles. To shuffle queries among a group of users who issues them useless user profile protocol is proposed [21] So that entity cannot profile a certain individual. It assumes the existence of a trustworthy third-party anonymizer. Instead of third party to provide a distorted user profile to the search engine Viejo[21] use the legacy social network.

Other class considers the sensitivity of the data, particularly the user profiles disclosed to the PWS server. Users only trust themselves and cannot tolerate the disclosure of their complete profiles on anonymity server. Third party assistance or collaborations between social network entries is not required. To generate the near–optimal partial profile Krause and Horvitz employ statistical techniques to learn a probabilistic model. But it builds the user profiles as a finite set of attributes and the probabilistic model is trained through predefined frequent queries. Privacy protection solution given by Xu et al [10] is based on hierarchical profiles. Generalized profile is obtained as a rooted subtree of the complete profile using a user specified threshold. But it does not address the query utility which is important for the service quality of UPS. Personalization have different effect on different queries [2], distinct queries are more benefited while larger click-entropy value queries are not. To classify queries by their click entropy Teevan et al. [22] collect a set of features of the query. Based on a client-side solution UPS framework differentiate distinct queries from ambiguous ones using the predictive query utility metric.

In the previous work [23] the prototype of UPS is proposed together with a greedy algorithm GreedyDP which support online profiling based on predictive metrics of personalization utility and privacy risk. In this paper metric of personalization utility captures three new observations. Evaluation model is refined to support user-customized sensitivities. New profile generalization algorithm GreedyIL is proposed.

III. EXISTING SYSTEM

The existing profile-based Personalized Web Search does not support runtime profiling. User profile is generalized only once offline, and used to personalize all queries from a same user. Such "one profile fits all" strategy has drawbacks for the variety of queries. Also, the existing profile-based personalization does not even help to improve the search quality for some ad hoc queries. The existing methods do not take into account the customization of privacy requirements. In existing system, all the sensitive topics are detected using an absolute metric called surprisal based on the information theory which assumes that the interests with less user document support are more sensitive. However, this assumption can be doubted with a simple example: If a user has a large number of documents about "sex," the surprisal of this topic may lead to a conclusion that "sex" is very general and not sensitive, despite the truth which is opposite.

Iterative user interactions are required in many personalization techniques for creating personalized search results. Search results are refined with some metrics such as rank scoring, average rank, and so on. This is infeasible for runtime profiling, since it pose too much risk of privacy breach, and also require processing time for profiling. Therefore, we need predictive metrics to measure the search quality without iterative interaction of user.

Disadvantage:

i. All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.
ii. The existing profile-based PWS do not support runtime profiling.
iii. The existing methods do not take into account the customization of privacy requirements.
iv. Personalization techniques require iterative user interactions when creating personalized search results.

IV. PROPOSED SYSTEM

This paper proposes a privacy- preserving personalized web search framework called UPS i.e. User customizable Privacy- preserving Search, that generalize profile for every query as per user privacy specification. Based on personalization and privacy risk metric, this paper formulates Risk Profile Generation, with its NP- hardness proved. It develops two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. GreedyDP maximize the discriminating power (DP) while GreedyIL minimize the information loss (IL). This paper also provides a mechanism for the client to decide whether or not to personalize a query in UPS. This decision is made before each runtime profiling to enhance the stability of the search results.

Advantages:

i. It enhances the stability of the search quality.
ii. It avoids the unnecessary exposure of the user profile.
iii. 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.

V. SYSTEM ARCHITECTURE

This section introduces system architecture. The Block Diagram of system is shown in Fig 1 which gives the details of the system components.

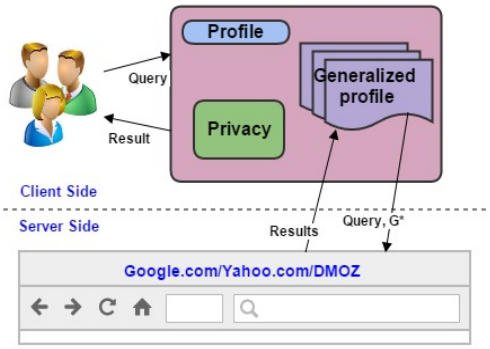

Fig.1 Architecture Diagram.

It consists of a nontrusty search engine and a number of clients. On the client machine online profiler is implanted as a search proxy for maintaining user profile, in hierarchy of nodes with semantics, and user specified privacy requirements represented as a set of sensitive nodes.

It works in two phases, offline and online phase. During offline phase based on user specified privacy requirement user profile is constructed. During online phase proxy generates a user profile runtime when the query is issued by the client according to user specified privacy requirements. Generalized user profile is created as a output.

Two metrics are used in generalization namely, personalization utility and the privacy risk. Then for personalized search the query and the generalized user profile are sent together to the PWS server. The search results are then delivered back to the query proxy and finally either give the raw results to the user or rerank them with the complete user profile.

VI. ALGORITHMIC STRATEGY

Greedy is an algorithmic paradigm that builds up a solution piece by piece, always choosing the next piece that offers the most obvious and immediate benefit.

For the online generalization two greedy algorithms, namely Greedy Discriminating Power and Greedy Information Loss are used. GreedyDP is used to maximize the discriminating power of the user profiles and Greedy Information Loss (GreedyIL) is used to minimize the information loss in user profiles.

1) GreedyDP :-
   It works in bottom-up manner, starting from the leaf node, for every iteration it chooses leaf topic for pruning to maximize the utility of output. Best profile having highest discriminating power is maintained during iteration, satisfying δ-risk constraint. When the root topic reached, iteration process stops, as a result we get the best profile.

2) GreedyIL :–
   It improves the generalization efficiency. Priority queue is maintained for candidate prune leaf operator in descending order, so the computational cost is decreased. When Risk is satisfied or when there is a single leaf left, iteration process stops. As the computational cost is decreased, GreedyIL algorithm achieves high efficiency than the GreedyDP algorithm.

VII. CONCLUSION

Client-side privacy protection framework called UPS is used to improve the search quality with the personalization utility of the user. UPS can be used by any PWS that captures user profiles in a hierarchical taxonomy. It allows users to specify customized privacy requirements via the hierarchical profiles. To protect the personal privacy without compromising the search quality UPS performed online generalization on user profile. For the online generalization two greedy algorithms, namely GreedyDP and GreedyIL are used. An experimental results show that UPS could achieve quality search results while preserving users customized privacy requirements.

REFERENCES

[1] Lidan Shou, He Bai, Ke Chen, and Gang Chen "Supporting Privacy Protection in Personalized Web Search",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING VOL:26 NO:2 YEAR 2014

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[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] 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.

[11] 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.

[12] K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006

[13] 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.

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

[15] 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.
[16] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.
[17] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.
[18] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.
[19] Y. Xu, K. Wang, G. Yang, and A.W.-C. Fu, "Online Anonymity for Personalized Web Services," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 1497-1500, 2009.
[20] 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.
[21] 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.
[22] J. Teevan, S.T. Dumais, and D.J. Liebling, "To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 163-170, 2008.
[23] G. Chen, H. Bai, L. Shou, K. Chen, and Y. Gao, "Ups: Efficient Privacy Protection in Personalized Web Search," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information, pp. 615- 624, 2011.