International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 10, Volume 2 (October 2015) www.ijirae.com

Privacy Preservation in Personalized Web Search

Mr. Arun Desai*
Computer Networks
Flora Institute of Technology, Pune

Mr Pankaj Chandre
Computer Networks
Flora Institute of Technology, Pune

Abstract— In recent years, personalized web search (PWS) has demonstrated effectiveness in improving the quality of search service on the Internet. Unfortunately, the need for collecting private information in PWS has become a major barrier for its wide proliferation. 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. This paper proposed a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. This 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.

Keywords—disruption-tolerant network (DTN); Ciphertext-policy attribute-based encryption (CP-ABE); attribute-based encryption; secure data retrieval;

I. Introduction

Huge amount of information gets added to the Web every day. Publicly visible text creation is of the order of 10 GB per day and private text creation (including user email, IM messages, tags, reviews etc) is of the order of 3 terabytes per day. This rapidly increasing scale of the web is in many ways limiting the utility of the web. There is a high level of noise beginning from spam and ending with a lot of uninteresting, irrelevant and duplicated content. Search engines and other forms of ranking are unable to keep up with this. Recently, search engines have started showing Wikipedia links as the top search result because ranking has become very hard.

Personalized search is a promising way to improve the accuracy of web search, and has been attracting much attention recently. However, effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. Indeed, these concerns have become one of the main barriers for deploying personalized search applications, and how to do privacy-preserving personalization is a great challenge. The web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.

As the amount of information on the web continuously grows, it has become increasingly difficult for web search engines to find information that satisfies users' individual needs. Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Many recent research efforts have focused on this area. Most of them could be categorized into two general approaches: Re-ranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine. A good personalization algorithm relies on rich user profiles and web corpus. However, as the web corpus is on the server, re-ranking on the client side is bandwidth intensive because it requires a large number of search results transmitted to the client before re-ranking. Alternatively, if the amount of information transmitted is limited through filtering on the server side, it pins high hope on the existence of desired information among filtered results, which is not always the case. Therefore, most of personalized search services online like Google Personalized Search and Yahoo! My Web adopt the second approach to tailor results on the server by analyzing collected personal information, e.g. personal interests, and search histories.

Nonetheless, this approach has privacy issues on exposing personal information to a public server. It usually requires users to grant the server full access to their personal and behaviour information on the Internet. Without the user's permission, gleaning such information would violate an individual's privacy.

Personalized web search is a promising technique to improve retrieval effectiveness. However, it often relies on personal user profiles which may reveal sensitive personal information. The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well, it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances.

Despite the attractiveness of personalized search, we have not yet seen large scale uses of personalized search services. This is not because such services are not available, but likely because users are not comfortable with the lack of protection of user privacy. However, to the best of our knowledge, it has not been widely adopted by users yet. Indeed, there is an inherent tension between providing personalized search and privacy preservation since personalized search requires collecting and aggregating a lot of user information. Specifically, in order to personalize search, a user profile or user model must be constructed to accurately represent a user's information need. To build a precise user profile, a lot of user information including query and click through history is often aggregated. However, from a user's privacy perspective, such a user profile can reveal a gamut of user's private life such as political inclination, family life, and hobbies, which is clearly a serious concern for users. Thus there appears to be a dilemma: high-accuracy Web search requires accurate user modelling which increases the risk of privacy infringement. Indeed, the privacy concern is one of the major barriers in deploying serious personalized search applications, and how to achieve personalized search while preserving users' privacy is. 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, this paper reviews the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization.

The rest of paper is organized as follows: Section II overviews the essential background. Section III addresses contribution. Section IV introduces the system architecture. Section V describes assumptions and security Requirements. Section VI concludes the paper.

II. Literature review

Profile-Based Personalization

Earlier techniques utilize term lists/vectors [5] or bag of words [2] to represent their profile. However, most recent works build profiles in hierarchical structures due to their stronger descriptive ability, better scalability, and higher access efficiency.

The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP [1], [14], [3], [15], Wikipedia [16], [17], and so on. Another work in [10] builds the hierarchical profile automatically via term-frequency analysis on the user data.

As for the performance measures of PWS in the literature, Normalized Discounted Cumulative Gain (nDCG) [18] is a common measure of the effectiveness of an information retrieval system. It is based on a human graded relevance scale of item-positions in the result list, and is, therefore, known for its high cost in explicit feedback collection. 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 [3], [8].

Privacy Protection in PWS System

Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual, as described in [20].

Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudo-identity, the group identity, no identity, and no personal information. Solution to the first level is proved to fragile [11].

The third and fourth levels are impractical due to high cost in communication and cryptography. Therefore, the existing efforts focus on the second level. Both [21] and [22] provide online anonymity on user profiles by generating a group profile of k users. Using this approach, the linkage between the query and a single user is broken.

In [23], the useless user profile (UUP) protocol is proposed to shuffle queries among a group of users who issue them. As a result any entity cannot profile a certain individual. These works assume the existence of a trustworthy third-party anonymizer, which is not readily available over the Internet at large.

Viejo and Castell_a-Roca [24] use legacy social networks instead of the third party to provide a distorted user profile to the web search engine. In the scheme, every user acts as a search agency of his or her neighbours. They can decide to submit the query on behalf of who issued it, or forward it to other neighbours. The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication.

In [12], Krause and Horvitz employ statistical techniques to learn a probabilistic model, and then use this model to generate the near-optimal partial profile. One main limitation in this work is that it builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries. These assumptions are impractical in the context of PWS. Xu et al. [10] proposed a privacy protection solution for PWS based on hierarchical profiles. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted subtree of the complete profile. Unfortunately, this work does not address the query utility, which is crucial for the service quality of PWS.

The concept of personalized privacy protection is first introduced by Xiao and Tao [25] in Privacy-Preserving Data Publishing (PPDP).

Teevan et al. [26] collect a set of features of the query to classify queries by their click entropy. While these works are motivated in questioning whether to personalize or not to, they assume the availability of massive user query logs (on the server side) and user feedback.

III. Proposed System

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. Unfortunately, the previous works of privacy preserving PWS are far from optimal. Thus this paper proposes a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements. In this there develop two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. In this UPS framework, we differentiate distinct queries from ambiguous ones based on a client-side solution using the predictive query utility metric.

IV. Proposed Architecture

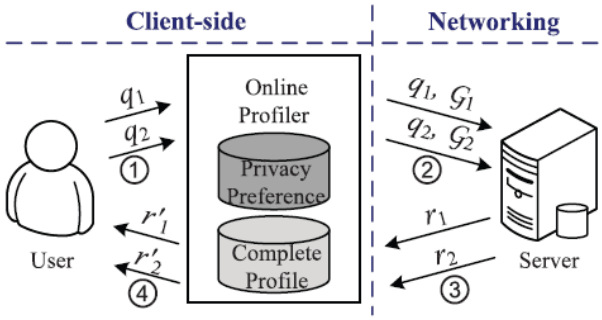

Figure1. Architecture of UPS.

As in figure 1, The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The offline phase handles queries as follows:

A. When a user issues a query qi on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile Gi satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles.

B. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.

C. The search results are personalized with the profile and delivered back to the query proxy.

D. Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile.

V. Assumptions and Security Requirements

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.

> *Assumption1: The repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. that is, given any human recognizable topic t, a corresponding node(also referred to as t) can be found in R, with the subtree subtr(t, R) as the taxonomy accompanying t.*

If we consider each topic to be the result of a random walk from its parent topic in R, we have the following recursive equation:

$$sup_R(t) = \sum_{t' \in C(t,R)} sup_R(t').$$

This equation can be used to calculate the repository support of all topics in R, relying on the following assumption that the support values of all leaf topics in R are available.

> *Given a taxonomy repository R, the repository support is provided by R itself for each leaf topic.*

Assumption 2 can be relaxed if the support values are not available. In such case, it is still possible to "simulate" these repository supports with the topological structure of R. That is, supR(t) can be calculated as the count of leaves in subtr(t, R).

VI. System Construction

A.UPS procedure

Specifically, each user has to undertake the following procedures in our solution:
1. Offline profile construction,
2. Offline privacy requirement customization

B.Offline-1. Profile Construction

The first step of the offline processing is to build the original user profile in a topic hierarchy H that reveals user interests. Let's assume that the user's preferences are represented in a set of plain text documents, denoted by D. To construct the profile, take the following steps:

1. Detect the respective topic in R for every document d ϵ 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 ϵ T with its document support from D, and then compute

$$sup_{\mathcal{H}}(t) = \sum_{t' \in C(t,\mathcal{H})} sup_{\mathcal{H}}(t').$$

supH(t) of other nodes of H with

There is one open question in the above process how to detect the respective topic for each document d ϵ D.

C.Offline-2. Privacy Requirement Customization

This procedure first requests the user to specify a sensitive-node set $S \subset \mathcal{H}$s, and the respective sensitivity value sen(s) > 0 for each topic s ϵ S. Next, the cost layer of the profile is generated by computing the cost value of each node t ϵ H as follows:
1. For each sensitive-node, cost(t) = sen(t);
2. For each nonsensitive leaf node, cost(t) = 0;
3. For each nonsensitive internal node, cost(t) is recursively given by in a bottom-up manner:

$$cost(t) = \sum_{t' \in C(t,\mathcal{H})} cost(t') \times Pr(t' \mid t).$$

Till now, we have obtained the customized profile with its cost layer available.

VII. Conclusion

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. This paper also refers two greedy algorithms, namely GreedyDP and GreedyIL, for the online generalization [1].

Acknowledgment

This study has been supported by author Junbeom Hur and Kyungtae Kang member of IEEE, ACM.

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

[14] 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.

[15] 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.

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

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

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

[19] K. Ja¨rvelin and J. Keka¨la¨inen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," Proc. 23rd Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), pp. 41-48, 2000.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[27] 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.

[28] 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, 011.

[29] J. Conrath, "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy," Proc. Int'l Conf. Research Computational Linguistics (ROCLING X), 1997.

[30] D. Xing, G.-R. Xue, Q. Yang, and Y. Yu, "Deep Classifier: Automatically Categorizing Search Results into Large-Scale Hierarchies," Proc. Int'l Conf. Web Search and Data Mining (WSDM), pp. 139-148, 2008

