Review on Providing Privacy Protection in Personalized Web Search

Miss. Snehal R. kawalkar, Prof. P. L. Ramteke

Abstract— **General web search engine are generally used for gathering huge information from web. However these internet users generally does not require generic model. Every web users wants specific information that he wants. This system will attempt to improve customized web search and additionally privacy of search query. Customer's Profile gives a basic data to performing customized web search. Personalized web search (PWS) is ability to identify different needs of different people who issue the same text query for web search and to carry out data retrieval for each and every user as a part of his interests. In Web searching, user profiles are main source for better retrieval effectiveness but using a user profile to find interest is violation of privacy. To overcome this privacy protection is necessary. System propose a PWS framework called UPS, it has been found that UPS framework is one of the efficient techniques which guarantees the user privacy and retrieves the contents as per user requirement accurately. We use GreedyIL algorithm which improves the efficiency of the generalization using heuristics based on numerous answers.**

Index Terms— Personalized web Search, Privacy Protection Profile, Search engine.

I. INTRODUCTION

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.

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.

Although there are pros and cons for both types of PWS techniques, the 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, which is usually gathered implicitly from query history , browsing history , click-through data , bookmarks, user documents, and so forth. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life.

Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

A. Motivations

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. A few previous studies suggest that people are ready to compromise privacy by supplying user profile to the search engine yields better search quality. In an ideal case, significant gain can be obtained by personalization at the expense of only a small (and less sensitive) portion of the user profile, namely a generalized profile. Thus, user privacy can be protected without compromising the personalized search quality. In general, there is a balance between the search quality and the level of privacy protection achieved from generalization. Unfortunately, the previous works of privacy preserving PWS are far from optimal. The problems with the existing methods are explained in the following observations

1. The existing profile-based PWS do not support runtime profiling. 1. Whether to personalize the query (by exposing the profile) and 2. What to expose in the user profile at runtime. To the best of our knowledge, no previous work has supported such feature.
2. The existing methods do not take into account the customization of privacy requirements.
3. Many personalization techniques require iterative user interactions when creating personalized search results.
4. Online prediction mechanism for deciding whether personalizing a query is not addressed.

Manuscript received Nov, 2015.

Miss. Snehal R. Kawalkar*, Department of computer science & Information Technology, H.V.P.M. College of Engineering & Technology Amravati, India.*

Prof. P. L. Ramteke*, Department of computer science & Information Technology, H.V.P.M. College of Engineering & Technology, Amravati, India.*

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR

B. Contributions

The above problems are addressed in our UPS (literally for User customizable Privacy-preserving Search) framework. The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified (customized) privacy requirements represented as a set of sensitive-nodes.

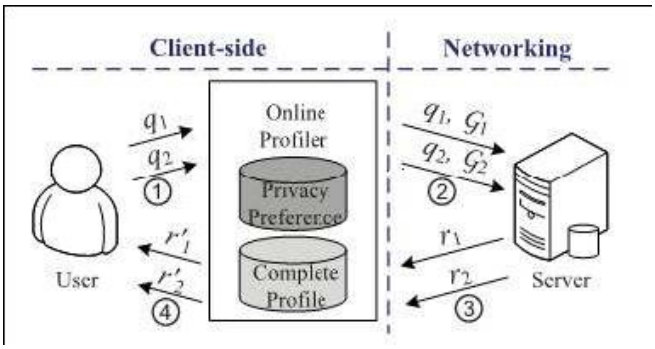

Fig.1 System architecture of UPS

The Figure1 contains 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 userspecified privacy requirements. The online phase handles queries as follows

1. 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.

2. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.
3. The search results are personalized with the profile and delivered back to the query proxy.
4. Finally, the proxy presents the raw results to the user, or re-ranks them with the complete user profile.

UPS is distinguished from conventional PWS in that it

1) Provides runtime profiling, which in effect optimizes the personalization utility while respecting user's privacy requirements
2) Allows for customization of privacy need
3) Does not require iterative user interaction.

II. Literature Review

In this section, many profile representations are available in the literature to facilitate different personalization strategies.

In 2007, Y. Xu, K. Wang, B. Zhang, and Z. Chen [1] build the hierarchical profile automatically via term-frequency analysis on the user data. He proposed UPS framework,

In 2007, Z. Dou [2] proposed Average Precision metric, to measure the effectiveness of the personalization in UPS.

In 2013, S.Vanitha [3] proposed a technique extracts the web pages based on two methods. The information extraction is based on User Profile and Click through data. The main advantage of this system is that it is possible to extract personalized web pages. An efficient user profile can improve the search engines performance by identifying the individual interest.

M. Spertta and S. Gach [4] User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles collect user information through proxy servers (to capture browsing historieson a personal computer). Both these techniques require participation of the user to install the proxy server or the bot.) or desktop bots (to capture activities

Krause and Horvitz [5] 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.

X. Shen, B. Tan, and C. Zhai, [6] Information retrieval systems (e.g., web search engines) are critical for overcoming information overload. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users, resulting in inherently non-optimal retrieval performance.

The concept of personalized privacy protection is first introduced by Xiao and Tao [7] in Privacy-Preserving Data Publishing (PPDP). A person can specify the degree of privacy protection for her/his sensitive values by specifying "guarding nodes" in the taxonomy of the sensitive attribute. Motivate by this, we allow users to customize privacy needs in their hierarchical user profiles.

In 2010, Viejo and Castell_a-Roca [8] 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 neighbors. They can decide to submit the query on behalf of who issued it, or forward it to other neighbors. The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication.

J. Teevan, S.T. Dumais, and E. Horvitz [9] exploit rich models of user interests, built from both search-related information and other information about the user, including documents and e-mails that the user has read and created.

In 2010, K. Wang, L. Xiong, [10] 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.

III. Background on Personalized Web Search

There are mainly two types of personalized web search they are Click-log-based and Profile-based personalized web search.

A. Click-Log-Based Method

Here, personalization is carried out on the basis of clicks made by user. The data recorded through clicks in query logs, simulates user experience. The web pages frequently clicked

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR

by user in past for a particular query is recorded in the history and score is computed for particular web page and based on that web search results are provided. This method will perform consistent and considerably well when it is works on frequent queries. When a never asked query is entered by user; it will not provide any precise search results, which is the main drawback of this method.

B. Profile Based Personalization

The basic idea of these works is to tailor the search results by referring to a user profile, implicitly or explicitly which reveals an individual information goal. Many profile representations are available in the literature to facilitate different personalization techniques.

i. Lists / vectors or bag of words:

Earlier techniques utilize term lists/vectors or bag of words to represent their profile. It is the simple representation in information retrieval system. Here a text is represented as the bag of its words, disregarding grammar and even word order. But it keeps multiplicity of those words. In each vector the second entry will be the count of that word.

ii. Hierarchical representation:

Most recent works build user profiles in hierarchical structures. The reason is their stronger descriptive ability, better scalability, and higher access efficiency. Majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP, Wikipedia, and DMOZ and so on. Using the term-frequency analysis on the user data, the hierarchical profile can be build automatically also.

IV. Privacy Protection in Pws

There are two classes of privacy protection problems for PWS in general. One class includes those works, treat privacy as the identification of an individual. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server.

A. *Identification of an Individual*

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 fragile. The third and fourth levels are impractical due to high cost in communication and cryptography. So the existing efforts focus on the second level.

iii. Online anonymity: It works based 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.

iv. Useless user profile (UUP): This 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 all the time in large number.

v. Legacy social networks**:** Instead of the third party to provide a distorted user profile to the web search engine, here every user acts as a search agency of his/her neighbors. They can decide to submit the query on behalf of who issued it, or forward it to other neighbors.

B. *Sensitivity of Data*

The solutions in class two do not require third-party assistance or collaborations between social network entries. In these solutions, users only trust themselves and cannot tolerate the exposure of their complete profiles to an anonymity server.

*i. 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.

ii. Generalized Profiles: 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 sub tree of the complete profile.

C. Issues

The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication. The statistical methods builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries in class two. These assumptions are impractical in the context of PWS and the generalized profile does not address the query utility, which is crucial for the service quality of PWS.

V. Proposed System

The proposed system seems to be more effective for privacy protection. An online profiler is designed in this system, which can adaptively generalize profiles by queries while respecting user specified privacy requirements. The online profiler is at the client side where the complete user profile is stored along with the specified sensitive topics. Runtime generalization aims at providing search efficiency along with privacy protection of user profiles. 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. In this section, the procedures carried out for each user during two different execution phases, namely the offline and online phases. Generally, the offline phase constructs the original user profile and then performs privacy requirement customization according to user-specified topic sensitivity. The subsequent online phase finds the Optimal -Risk Generalization solution in the search space determined by the customized user profile. The online generalization procedure is guided by the global risk and utility metrics. The computation of these metrics relies on two intermediate data structures, namely a cost layer and a preference layer defined on the user profile.

Specifically, each user has to undertake the following procedures in our solution

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.

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR

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.

VI. Future Work and 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. For future work, we will try to resist adversaries with broader background knowledge, such as richer relationships among topics (e.g., exclusiveness, sequentiality, and so on), or capability to capture a series of queries from the Victim. We will also seek more sophisticated methods to build the user profile, and better metrics predict the performance (especially the utility) of the UPS.

Acknowledgment

I would like to thanks my Guide Prof. P. L. Ramteke and Principal Dr. A. B. Marathe, who provided me constructive and positive feedback during the preparation of this paper.

REFERENCES

[1] 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.
[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] S.Vanitha "A Personalized Web Search based on user profile and user clicks" International Journal of Latest Research in Science and Technology ISSN (Online):2278-5299 Volume 2, Issue 5: Page No.78-82,September-October 2013
[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] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research (2010), vol. 39, pp. 633-662
[6] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17 (2007)
[7] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[8] 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.
[9] 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 '05), pp. 449-456, 2005.
[10] 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.

Miss. Snehal R. Kawalkar, received B.E. degree in Information Technology from saint Gadge Baba Amravati university in 2012. She is currently pursuing Master's Degree in Computer Science and Information Technology from H.V.P.M's College of Engineering And Technology, Amravati, Maharashtra.

Prof. P. L. Ramteke, received the B.E.and M.E degree in Computer Science& engineering. He is currently working as a HOD and Associate Professor at H.V.P.M's college of Engineering and Technology, Amravati.

ISSN: 2278 – 7798 All Rights Reserved © 2015 IJSETR