Supporting privacy protection in personalized web search – A survey

S. Porna Sai[1], S. Udhaya[2], R. Suganya[3] and K. S. Sangeetha[4]

[1,2,3,4] Department of Information Technology, Sri Krishna College of Engineering and Technology, Kuniamuthur, Coimbatore, India

Pornasai4@gmail.com[1], udhayasukumaran2992@gmail.com[2], suganyarajuu@gmail.com[3], sangeethaks@skcet.ac.in[4]

Abstract

Objectives: This survey investigates the several privacy preserving techniques and provides the idea about the new efficient method in the future. The main goal of this work is to assure the privacy guarantee to the user who is involved in the personalized web search. To do this several mechanism which is related to the privacy protection is investigated.
Methods: In this manuscript, a survey has been done on the several privacy preserving mechanism that has been used to assure a users privacy guarantee and as well as accurate retrieval. Among the different methodologies that are discussed, 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.
Results: This survey expansively studies the problems in the privacy preserving mechanism in the personalized web search environment. The performance of the various methods is evaluated with numerous parameters like accuracy, failure rate and the privacy assurance. The survey analysis conducted were proves that.
Conclusion: the UPS framework is one of the efficient methodology which aims to retrieve the user wanted contents with the assurance of privacy guarantee.
Keywords: Personalization, utility, privacy protection, runtime generalization

1. Introduction

As the amount of information on the web continuously increases, it has become very difficult for web search engines to find information that satisfies individual needs of the user. Personalized search is a way to improve search quality by customizing search results for people with different information goals. Many recent researches 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 [1, 2, 3]. 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 adopts 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 behavior information on the Internet. Without the user's permission, gleaning such information would violate an individual's privacy [4, 5]. In particular, Canada launched information, i.e., age, race, income, evaluations, and even intentions to acquire goods or services from being released to outside is also evidenced by a recent survey conducted by Choicestream2 that the privacy fear continues to escalate although personalization remains something most consumers want. The number of consumers interested in personalization remains at a remarkably high 80%; however, only 32% of respondents were willing to share personal information in exchange for personalized experience, down from 41% in 2004. Recent coverage about identity thefts and online security breaches, i.e. AOL search query data scandal, even causes users to be more wary than ever on sharing their private information

Personal data, i.e. browsing history, emails, etc., are mostly unstructured, for which it is hard to measure privacy. In addition, it is also difficult to incorporate unstructured data with search engines without summarization. So, for the purpose of both web personalization and privacy preservation, it is necessary for an algorithm to collect, summarize, and organize a user's personal information into a structured user profile. Meanwhile, the notion of privacy is highly subjective and depends on the individuals involved. Things considered to be private by one person could be something that others would love to share. In this regard, the user should have control over which parts of the user profile is shared with the server.

2. Related works

In information retrieval [6], much research is focused on personalized search. Relevance feedback and query refinement harnesses a short-term model of a user's interests, and information about a user's intent is collected at query time. Personal information has also been used in the context of Web search to create a personalized version of PageRank. There are still approaches, including many commercially available information- filtering systems, which require users explicitly specify their interests. However, as pointed out, users are typically unwilling to spend the extra effort on specifying their intentions. Even if they are motivated, they are not always successful in doing so.

A majority of work focuses [7, 11] on implicitly building user profiles to infer a user's intention. A wide range of implicit user activities have been proposed as sources of enhanced search information. This includes a user's search history, browsing history, click-through data, web community, and rich client side information in the form of desktop indices. Our approach is open to all kinds of different data sources for building user profiles, provided the sources can be extracted into text. In our experiments data sources like IE histories, emails and recent personal documents were tested.

User profiles [8] can be represented by a weighted term vector, weighted concept hierarchical structures like ODP3, or other implicit user interest hierarchy. For the purposes of selectively exposing users' interests to search engines, the user profile is a term based hierarchical structure that is related to frequent term based clustering algorithms. The difference here is that the hierarchical structure is implicitly constructed in a top-down fashion. And the focus is the relationships among terms, not clustering the terms into groups.

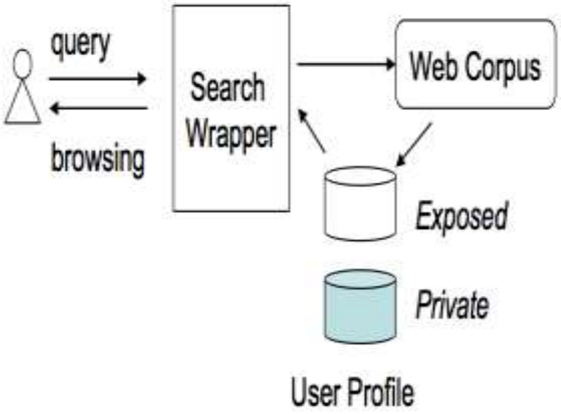
Privacy concerns [9, 12, 13] are natural and important especially on the Internet. Some prior studies on Private Information Retrieval (PIR), focuses on the problem of allowing the user to retrieve information while keeping the query private. Instead, this study targets preserving privacy of the user profile, while still benefiting from selective access to general information that the user agrees to release. To our knowledge, this problem has not been studied in the context of personalized search. One possible reason for this is that personal information, i.e. browsing history and emails, is mostly unstructured data, for which privacy is difficult to measure and quantify.

Some works on privacy issues [10, 14, 17] in the data mining community focus on protecting individual data entries while allowing information summarization. A popular way of measuring privacy in data mining is by examining the difference in prior and posterior knowledge of a specific value. This can be formalized as the conditional probability or Shannon's information theory. Another way to measure privacy is the notion of k-anonymity which advocates that personally identifying attributes be generalized such that each person is indistinguishable from at least k-1 other persons. In this study the notion of privacy does not compare information from different users, but rather the information collected over time for a single user. In addition, this study addresses unstructured data.

3. Challenges of Personalized search

Despite the attractiveness of personalized search, there is no large-scale use of personalized search services currently. Personalized web search faces several challenges that retard its real-world large-scale applications as shown in figure 1.

Figure 1. User privacy in personalized search engines



Privacy is an issue: Personalized web search, especially server-side implement, requires collecting and aggregating a lot of user information including query and click through history. A user profile can reveal a large amount of private user information, such as hobbies, vocation, income level, and political inclination, which is clearly a serious concern for users [15, 21]. This could make many people nervous and feel afraid to use personalized search engines. A personalized web search will be not well received until it handles the privacy problem well.

Infer user information: It is really hard to infer user information needs accurately. Users are not static. They may randomly search for something, which they are not interested in. They even search for other people sometimes. User search histories inevitably contain noise that is irrelevant or even harmful to current search. This may make personalization strategies unstable.

Query personalization: Queries should not be handled in the same manner with regard to personalization. Personalized search may have little effect on some queries. Some work [16] investigates whether current web search ranking might be sufficient for clear/unambiguous queries and thus personalization is unnecessary. Dou et al. [21] reveal that personalized search has little effect on queries with high user selection consistency. A specific personalized search also has different effectiveness for different queries. It even hurts search accuracy under some situations. For example, topical interest-based personalization, which leads to better performance for the query ‘‘mouse,’’ is ineffective for the query ‘‘free mp3 download.’’ Actually, relevant documents for query ‘‘free mp3 download’’ are mostly classified into the same topic categories and topical interest-based personalization has no way to filter out desired documents. Dou et al. [18] also reveal that topical interest-based personalized search methods are difficult to deploy in a real world search engine. They improve search performance for some queries, but they may hurt search performance for additional queries.

3.1. Existing models

The existing profile-based Personalized Web Search [19, 20] does not support runtime profiling. A user profile is generalized for only once offline, and it is used to personalize all queries from a same user indiscriminatingly. Such “one profile fits all” strategy certainly has drawbacks for given variety of queries. It has been seen that profile-based personalization may not even help to improve the search quality for some ad hoc queries, even though exposing user profile to a server has put the user’s privacy at risk.

The existing methods do not take into account the customization of privacy requirements. This makes some user privacy to be overprotected while others insufficiently protected. For example, the sensitive topics are detected using an absolute metric called surprisal based on the information theory, considering that the interests with less user document support are more sensitive. This assumption can be doubted with a simple counterexample: 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, where the conclusion is not true. A little prior work can effectively address privacy needs of user during the generalization.

Many personalization search techniques require multiple user interactions when creating personalized search results. The search results are usually refined with some metrics, which require iterative user interactions, such as rank scoring, average rank and many more. This paradigm is infeasible for runtime profiling, as it will pose too much risk of privacy breach and also demand prohibitive processing time for profiling. A predictive metrics is needed to measure the search quality and breach risk after personalization, without incurring multiple user interaction.

3.2. UPS framework

The personalized web search framework UPS can generalize profiles for each user query according to user-specified privacy requirements. The two important metrics, namely personalization utility and privacy risk are balanced using the hierarchical user profile. UPS makes use of two algorithms namely GreedyDP and GreedyIL for its runtime generalization. While the first algorithm tries to maximize the discriminating power (DP), the other attempts to minimize the information loss (IL). By exploiting a number of heuristics, GreedyPS outperforms GreedyDP significantly.

UPS provides an inexpensive mechanism for the client to decide whether to personalize a query or not. This decision can be made by the user before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile.

4. Conclusion

This survey has made a study on client-side privacy protection framework called UPS for personalized web search. UPS framework could potentially be adopted by any PWS that gathers user profiles in a hierarchical taxonomy. The framework allowed users to specify their customized privacy requirements via the hierarchical profiles. In addition, UPS also performed runtime generalization on user profiles to protect their privacy without compromising the search quality. The experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. For future work we are focusing on ontological profile construction instead of hierarchical profile, as it helps to identify more accurate user interest.

5. References

1. J. Pitkow, H. Schuetze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel [2002] Personalized search, *Communications of the ACM*, Vol. 45(9), pp. 50-55.
2. P. Anick [2004] Using terminological feed back for Web search refinement: a log-based study, In *Proceedings of the 13 th International World Wide Web Conference (WWW)*, New York, New York
3. K.R. McKeown, N. Elhadad, and V. Hatzivassiloglou [2003] Leveraging a common representation for personalized search and summarization in a medical digital library, In *Proceedings of International Conference on Digital Library*
4. Glen Jeh and Jennifer Widom [2003] Scaling personalized web search, In *Proceedings of the 12 th International World Wide Web Conference (WWW)*, Budapest, Hungary.
5. T.H. Haveliwala [2002] Topic-sensitive PageRank, In *Proceedings of the 11th International World Wide Web Conference (WWW)*, Honolulu, Hawaii.
6. Paolo Ferragina, and Antonio Gulli [2005] A personalized search engine based on Web-Snippet hierarchical clustering, In *Proceedings of the 14th International World Wide Web Conference (WWW)*, Chiba, Japan.
7. P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter [2005] Using ODP metadata to personalize search, In the *Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil.
8. M. Speretta and S. Gauch [2004] Personalizing search based on user search history, In *Proceedings of International Conference of Knowledge Management (CIKM)*, Washington D.C.
9. K. Sugiyama, K. Hatano and M. Yoshikawa [2004] Adaptive Web search based on user profile constructed without any effort from users, In *Proceedings of the 13th International World Wide Web Conference (WWW)*, New York, New York.

10. J.Sun, H.Zeng, H.Liu, Y.Lu and Z.Chen.CubeSVD [2005] A Novel Approach to Personalized Web Search, In *Proceedings of the 14th International World Wide Web Conference (WWW)*, Chiba, Japan.
11. F. Qiu and J. Cho [2006] Automatic identification of user interest for personalized search, In *Proceedings of the 12th International World Wide Web Conference (WWW)*, Edinburgh, Scotland.
12. A. Kritikopoulos and M. Sideri [2003] The compass Filter: Search engine result personalization using web communities, In *Proceedings of Intelligent Techniques in Web Personalization (ITWP)*.
13. J.Teevan, S. T. Dumais and Eric Horvitz [2005] Personalizing search via automated analysis of interests and activities, In the *Proceedings of 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil.
14. B. Fung, K. Wang and M. Ester [2003] Hierarchical document clustering using frequent itemsets, In *Proceedings of SIAM International Conference on Data Mining*, San Francisco.
15. K. Wang, C. Xu and B. Ling [1999] Clustering transactions using large items, In *Proceedings of the 8th Conference on Information and Knowledge Management (CIKM)*, Kansas City, November.
16. W. Gasarch [2004] A survey on private information retrieval, *The bulletin of the European Association for Theoretical Computer Science (EATCS)*, Vol.82, 72, pp. 107.
17. R. Agrawal, and R. Skriant [2000] Privacy preserving data mining, In *Proceedings of the ACM SIGMOD Conference on Management of Data (SIGMOD)*, Dallas, Texas.
18. A. Evfimievski, J. Gehrke and R. Srikant [2003] Limiting privacy breaches in privacy preserving data mining, *In Proceedings of the ACM SIGMOD/PODS(PODS)*, San Diego, CA.
19. L. Sweeney [2002] k-anonymity: a model for protecting privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Vol.10 (5), pp. 557-570.
20. X. Shen, B.Tan, and C Zhai [2007] Privacy protection in personalized search, *Special Interest Group on Information Retrieval Forum*, Vol.41(1) pp. 4–17.
21. Z. Dou, R.Song and J.Wen [2007] A large-scale evaluation and analysis of personalized search strategies, In *Proceedings 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

The Publication fee is defrayed by Indian Society for Education and Environment (iSee). www.iseeadyar.org

Citation:

S. Porna Sai[1], S. Udhaya[2], R. Suganya[3] and K. S. Sangeetha[4] [2014] Supporting privacy protection in personalized web search – A survey. *Indian Journal of Innovations and Developments.* Vol 3 (3), pp.45-49.