IPHV8I1037X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .I
ISSN Online: 2319 – 9253
Print: 2319 – 9245

ANNONYMIZATION OF USER PROFILES BY USING PERSONALIZED WEBSEARCH

[#1]M.SRAVANALAXMI, M.Tech Student,
[#2]RAVI MATHEY, Professor & HOD,
Dept of CSE,
VIDYA JYOTHI INSTITUTE OF TECHNOLOGY, HYDERABAD, TELANGANA, INDIA.

Abstract: With increasing number of websites the Web users are increased with the massive amount of data available in the internet which is provided by the Web Search Engine (WSE). The aim of the WSE is to provide the relevant search result to the user with the behavior of the user click were they performed. WSE provide the relevant result on behalf of the user frequent click based method. From this method no assurance to the user privacy and also no securities were providing to their data.

We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. Our runtime 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. We present two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that Greedy IL significantly outperforms Greedy DP in terms of efficiency.

Keywords: *Privacy Protection, Personalized Web Search, Utility, Risk, Profile.*

I.INTRODUCTION

The web search engine has gained a lot of popularity and importance for users seeking information on the web. Since the contents available in web is very vast and ambiguous, users at times experience failure when an irrelevant result of user query is returned from the search engine. Therefore, in order to provide better search result a general category of search technique Personalized Web search is used. In personalized web search, user information is collected and analyzed in order to find intention behind issued query fired by user.

There are two categories of PWS, namely click-log-based and profile-based. The click-log based methods are straightforward— they simply impose bias to clicked pages in the user’s query history. This strategy has been performing well but it work on repeated queries from same user which is a strong limitation to its applicability. While profile-based methods improve the search experience 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. There are both advantages and disadvantages for both type of PWS technique, profile based PWS is more effective for improving search result. The user profile is made from information gathered 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.

Many search engines like Google, Yahoo provide a relevant and irrelevant data to the user based on their search. To avoid the irrelevant data the technique called Personalized Web Search (PWS) were arise. Inferring user search goals is very important in improving search-engine relevance and personalized search [3, 4]. This is based on the user rofiles based on the click through log and the feedback session [5]. These data were generated from the frequent query requested by the user, history of query, browsing, bookmarks and so on.

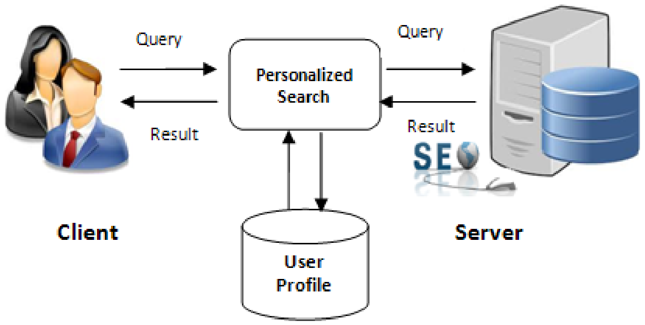


Fig 1: Personalized Search Engine Architecture

By these methods personal data were easily reveal. While many search engines take advantage of information about people in common, or regarding particular groups of people, personalized search based on a user profile that is unique to the individual person. Research systems that personalize search outcomes model their users in different ways. The Personalized Web Search provides a unique opportunity to consolidate and scrutinize the work from industrial labs on personalizing web search using user logged search behavior context. It presents a fully anonymized dataset, which has anonymized user id, queries based on the keywords, their terms of query, providing URLs, domain of URL and the user clicks. This dispute and the shared dataset will enable a whole new set of researchers to study the problem of personalizing web search experience. It decreases the likelihood of finding new information by biasing search results towards what the user has already found. By using these methods privacy of the user might be loss because of clicking the relevant search, frequently visited sites and providing their personal information like their name, address, etc. in this case their privacy might be leak. For this privacy issue, many existing work proposed a potential privacy problems in which a user may not be aware that their search results are personalized for them [6, 7]. It affords a host of services to people, and several of these services do not necessitate information to be grouped about a person to be customizable. While there is no warning of privacy assault with these services, the stability has been tipped to errand personalization over privacy, yet when it comes to search [8]. That approaches does no protect privacy issues rising from the lack of protection for the user data. To providing better privacy we propose a privacy preserving with the help of greedy method by providing the hybrid method of the discriminating power and prevent the information loss.

II.RELATED WORK

In [9] this paper, author study this problem and provide some preliminary conclusions. It presents a large scale evaluation framework for personalized search based on query logs and then evaluates with the click and profile based strategies. By analyzing the results, author reveals that personalized search has significant improvement over common web search on some queries but it has little effect on other queries. Author also reveals that both long term and short-term contexts are very important in improving search performance for profile-based personalized search strategies. In this paper, author tries to investigate whether personalization is consistently effective under different situations. The profile-based personalized search strategies proposed in this paper are not as stable as the click-based ones. They could improve the search accuracy on some queries, but they also harm many queries. Since these strategies are far from optimal, author will continue his work to improve them in future [10]. It also finds for profile-based methods, both long-term and short-term contexts are important in improving search performance. The appropriate combination of them can be more reliable than solely using either of them. From the author [11], they studied how to exploit implicit user modeling to intelligently personalize information retrieval and improve search accuracy. Unlike most previous work, it emphasizes the use of immediate search context and implicit feedback information as well as eager updating of search results to maximally benefit a user. Author presented a decision-theoretic framework for optimizing interactive information retrieval based on eager user model updating, in which the system responds to every action of the user by choosing a system action to optimize a utility function. Author propose [12] specific techniques to capture and exploit two types of implicit feedback information: (1) identifying related immediately preceding query and using the query and the corresponding search results to select appropriate terms to expand the current query, and (2) exploiting the viewed document summaries to immediately re-rank any documents that have not yet been seen by the user. Using these techniques, author develops a client side web search agent (UCAIR) on top of a popular search engine (Google) without any additional effort from the user. From the [13] author have explored how to exploit implicit feedback information, including query history and click-through history within the same search session, to improve information retrieval performance. Using the KLdivergence retrieval model as the basis, author proposed and studied four statistical language models for context sensitive information retrieval, i.e., FixInt, BayesInt, OnlineUp and BatchUp. It uses TREC AP Data to create a test set for evaluating implicit feedback models.

The current work can be extended in several ways: First, it has only explored some very simple language models for incorporating implicit feedback information. It would be interesting to develop more sophisticated models to better exploit query history and click through history. For example, this may treat a clicked summary differently depending on whether the current query is a generalization or refinement of the previous query. Second, the proposed models can be implemented in any practical systems. It currently develops a client-side personalized search agent, which will incorporate some of the proposed algorithms. Author will also do a user study to evaluate effectiveness of these models in the real web search. Finally, author should further study a general retrieval framework for sequential decision making in interactive information retrieval and study how to optimize some of the parameters in the context-sensitive retrieval models.

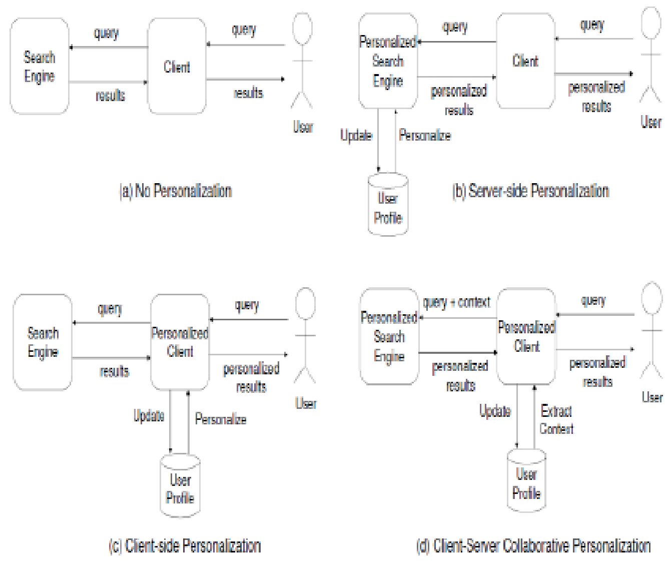
This paper [14] was motivated by two emerging trends: web users want personalized services and web users want privacy. One challenge is that personal information must be made anonymous under the assumption that the participating parties, including the web service, are not completely trusted, due to systematic collection of personal information in addition to queries. Another challenge is the online and dynamic nature of web users. Author proposed the notion of online anonymity to protect web users and proposed an approach to maintain online anonymity through time. This approach makes use of a third party called the user pool and it does not require the user pool to be trusted. The simulation study on real US demographics showed promising results: it is feasible to achieve personalization for reasonable privacy settings.

From this approach [15, 16] they requires users to contribution the server full access to personal information on Internet, which break users' privacy. In this paper, author inspects the possibility of accomplish a balance between users' privacy and search quality. First, an algorithm is provided to the user for collecting, abbreviation, and organizing their personal information into a hierarchical user profile, where general terms are ranked to higher levels than explicit terms. Through this profile, users control what section of their private information is uncovered to the server by adjusting the minDetail threshold. An additional privacy measure, expRatio, is proposed to approximation the amount of privacy is exposed with the specified minDetail value. Yet, this paper is an exploratory work on the two features: First, author deal with unstructured data such as personal documents, for which it is still an open problem on how to define privacy. Secondly, author try to bridge the conflict needs of personalization and privacy protection by breaking the premise on privacy as an absolute standard. Also, author believe that an enhanced balance between privacy protection and search quality can be achieved if web search are personalized by allowing for only revealing those information associated to a specific query. It performs less protection for the user data and they were no assured for the user data and their profile information's.

In this paper [17] the author studied the existing generalization methods are insufficient because they cannot assurance privacy protection in all cases, and frequently acquire redundant information loss by performing too much generalization. In this paper, author proposes the idea of personalized secrecy, and develops a new generalization structure that takes into account customized privacy necessities. This technique successfully avoid privacy intrusion even in scenarioswhere the existing approaches fail, and results in generalized tables that permit accurate aggregate analysis. This work [18] lays down a solid theoretical foundation for developing substitute generalization strategies. For instance, the greedy algorithm presented in this paper is not optimal, in the sense that it does not necessarily achieve the lowest information loss.

III. SOFTWARE ARCHITECTURE FOR PERSONALIZED SEARCH

For Web search applications, server-client architecture, as shown in Fig.2 (a), is commonly adopted, where a client (often the web browser) sends queries to a server (the search engine). The search engine analyzes the user information need, looks up its index structure of documents, and returns a ranked list of search results to the client for a user to view. A search engine generally stores user search logs for various kinds of purposes including personalization and anti-spam. Thus it is to the interest of search engines not to remove the search engine logs automatically. Indeed, they tend to keep the search engine logs indefinitely. There are three kinds of software architectures that expand the basic server-client model of Web search to support personalized search. Their main differences lie in where personally identifiable informationP (U) is stored and how it is exploited for personalization. In this section, we describe these three kinds of software architectures and analyze what levels of privacy preservation can be achieved with these different architectures.



(a) No Personalization

(b) Server-side Personalization

(c) Client-side Personalization

(d) Client-Server Collaborative Personalization

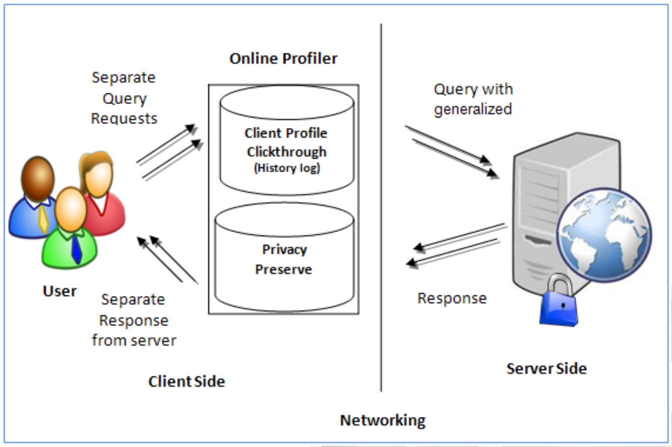


Fig.2. Software Architecture of Personalized Web Search.

A. Server-Side Personalization

For server-side personalization as shown in Fig.2 (b), the personally identifiable information P (U) is stored on the search engine side. The search engine builds and updates the user profile either through the user's explicit input (e.g., asking the user to specify personal interests) or by collecting the user's search history implicitly (e.g., query and click through history). Both approaches require the user to create an account to identify him. But the latter approach requires no additional effort from the user and contains richer description of user information need. The advantage of this architecture is that the search engine can use all of its resources (e.g, document index, common search patterns) in its personalization algorithm. Also, the client software generally requires no changes. This architecture is adopted by some general search engines such as Google Personalized. Currently most personalized search systems with server-side personalization architecture require the user to give consent before his/her search history can be collected and used for personalization. If the user gives the permission, the search engine will hold all the personally identifiable information possibly available on the server side. Thus from the user perspective, it even does not have level I privacy protection.

Since many users fear its potential privacy infringement by search engines, this has hindered the wide adoption of personalization with this architecture. However, if the search engine decides to voluntarily replace the user identity ID (U) with a pseudo user identity IDP(U), Level I privacy protection can be achieved. When the search engines release the search engine logs to the public or a group of researchers, they generally replace user identity ID (U) by a pseudo user identity IDP(U). To the third parties receiving these search engine logs, which may use it for personalized search purpose, the user will have Level I privacy protection. If the user decides to use a proxy to communicate with the search engine, all user information going through the same proxy will be combined in a user profile. Through this method, privacy protection can be achieved. However, this method does not always work: When the search engine uses the user login ID to collect user information, this method will not achieve privacy protection; when the search engine only uses the IP address to aggregate the user information, this method works. Sometimes, search engines group users randomly or according to some criteria before they release the search engine logs. Then the user will also have privacy protection to those third parties which receive the search engine logs. It is impossible to implement privacy protection if personalization is done on the server side.

B. Client-side Personalization

For client-side personalization as shown in Fig.2(c), the personally identifiable information is always stored on a user's personal computer. As in the case of server-side personalization, the user profile can be created from user specification explicitly or search history implicitly. The client sends queries to the search engine and receives results, which is the same as in the general web search scenario. But given a user's query, a client-side personalized search agent can do query expansion to generate a new query before sending the query to the search engine. The personalized search agent can also re-rank the search results to match the inferred user preferences after receiving the search results from the search engine. With this architecture, not only the user's search behavior but also his contextual activities(e.g., web pages viewed before) and personal information (e.g., emails, browser bookmarks) could be incorporated into the user profile, allowing for the construction of a much richer user model for personalization. The sensitive contextual information is generally not a major concern since it is strictly stored and used on the client side. Another benefit is that the overhead in computation and storage for personalization can be distributed among the clients.

A main drawback of personalization on the client side is that the personalization algorithm cannot use some knowledge that is only available on the server side (e.g., Page Rank score of a result document). UCAIR adopts the client-side personalization. With proxy functionality applied to the client side, Level II privacy protection can be achieved. If the client side uses an anonymous network such as Tor to communicate with the search engine, privacy protection can also be achieved. In order to achieve privacy protection, additional cooperation of the search engine would be needed as we described.

C. Client-Server Cooperative Personalization

For the client-server cooperative personalization as shown in Fig.2 (d), it is a compromise between the previous two kinds of architectures. The user profile is still stored on the client side, but the server also participates in personalization. At query time, the client extracts contextual information from the user profile, and sends it to the search engine along with the query. The search engine then does personalization with the received context. Compared with client-side personalization, this architecture has an advantage of allowing for the use of a search engine's internal resources. The contextual information sent to the server specifies the user's search preferences (e.g., query expansion terms, topic weight vector). It is extracted from the user profile (e.g., the weight vector can be learned from search history), and is only relevant to a particular query. Therefore, it is a condensed version of the whole user profile (generally a few terms or a weight vector from a user's search history), thus the architecture can minimize the personal information obtained by the search engine. A main drawback is that the condensed contextual information may not be as powerful as the whole user profile. We have not yet seen any personalization products in this category, probably due to the relatively complex architecture. This architecture provides the same level of privacy protection as server-side personalization. However, the personally identifiable information collectable on the server side is less than in the case of pure server-side personalization.

IV. PRIVACY PROTECTION IN CURRENT WEB SEARCH SYSTEMS

Currently, there are a variety of search engines on WWW-general search engines such as Google and Yahoo!, meta-search engines such as dogpile and ixquick, special search engines such as cluster search engine vivisimo, and personalized search systems such as UCAIR. In this section, we analyze privacy protection for some of these typical search paradigms.

A. Autonomous Search Engines

When people do web search with an autonomous search engine such as Google, Yahoo, or MSN, both the IP address and query terms are stored on the search engine side unless the user uses a proxy or anonymous communication system additionally. Although Google has a strict and clear privacy policy, the personally identifiable information P (U) is stored on Google severs and the users have no full control of their personal information. According to the levels of privacy protection described, it does not even satisfy privacy protection unless the user applies some privacy protection measures to strengthen the privacy protection themselves. Users are generally not comfortable with counting on others to protect their privacy. Recent history has witnessed several privacy infringement incidents when some companies accidentally or willingly had violated such trust and were facing bankruptcy courts, civil subpoenas or lucrative mergers.

B. Meta Search Engines

There are quite a few Meta search engines on the Web such as Dogpile, Look smart and ix quick. A meta-search engine sends user requests to several autonomous search engines and re-ranks search results returned from each one. When people use the Meta search engines, autonomous search engines only receive all user queries from the single Meta search engine. Thus there is the privacy protection to those underlying autonomous search engines. However, there is no automatic privacy protection for the users of these Meta search engines, which is the same as the scenario when people directly use autonomous search engines. Interestingly, the Meta search engine ixquick guarantees that it removes the IP addresses of users and keep no other unique identity. Thus ID (U) of personally identifiable information is not stored on the server side although TEXT(N) still is. It provides Level III privacy protection for the users of this Meta search engine, but ixquick has no personalization functionality.

C. Client-side Personalized Search Tools

There are also some client-side personalized search tools such as UCAIR]. These client-side personalized search tools are installed on a personal computer and build rich user profiles for individual users. They communicate with autonomous search engines when they do web search. Authors have designed and developed a privacy-preserving personalized search system (UCAIR),which resides on the client side and greatly alleviates the privacy concerns while doing personalized search. A user's personal information including user queries and click through history resides on the user's personal computer, and is exploited to better infer the user' information need and provide more accurate search results. UCAIR is implemented as a web browser plug-in 8.

The software architecture of the system is as Fig.3 As shown in Fig.3; the UCAIR personalized search system has three major components: (1) the implicit user modeling module captures a user's search context and history information, including the submitted queries and any clicked search results and infers search session boundaries. (2) The query modification module selectively improves the query formulation according to the current user model. (3) The result re-ranking module immediately re-ranks any unseen search results when-ever the user model is updated. For example, when the user clicks on a search result to view the corresponding web page, UCAIR would assume that the clicked result summary is appealing to the user and thus reflect the user's information need. It would immediately re-rank the not-yet-viewed results based on the viewed summaries and attempt to pull up results that match the clicked summaries well while pushing down those results that are originally ranked high, but do not match the clicked summaries well.

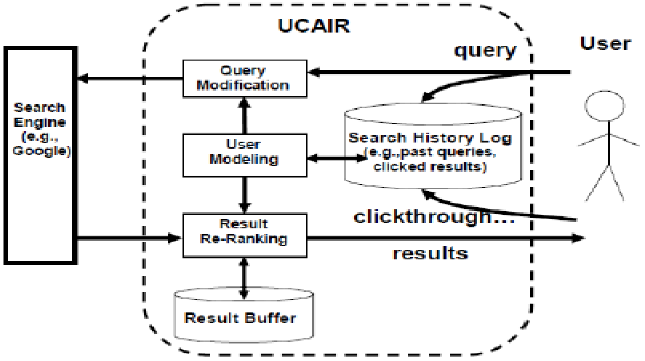


Fig.3. UCAIR architecture.

Thus when the user clicks on the \Back" button of the web browser or \Next" link of the search result page to view more results, the new results displayed would be different from the original results. When a user combines UCAIR with the Tor tool, it will be at the Level III privacy protection even though UCAIR communicates with a general search engine such as Google.

V.CONCLUSION

Web users were increases because of available of information's from the web browser based on the search engine. With the increasing number of user service engine must provide the relevant search result based on their behavior or based on the user performance. Providing relevant result to the user is based on their click logs, query histories, bookmarks, by this privacy of the user might be loss. For providing relevant search by using these approaches the privacy of the user may loss. Most existing system provides a major barrier to the private information during user search. That approaches does not protect privacy issues and rising information loss for the user data. For this issue this paper proposes client based architecture based on the greedy algorithm to prevent the user data and provide the relevant search result to the user in future it can include this work in mobile application.

We proposed two greedy algorithms, namely Greedy DP and Greedy IL, for the online generalization. Our experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution. For future work, we will try to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentiality, and so on),or capability to capture a series of queries (relaxing the second constraint of the adversary) from the victim.

REFERNCES:

[1] Lidan Shou, He Bai, Ke Chen, and Gang Chen, "Supporting Privacy Protection in PersonalizedWeb Search", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 2, February 2014.

[2] Z. Dou, R. Song, and J.-R. Wen, "A Large-Scale Evaluation andAnalysis of Personalized Search Strategies," Proc. Int'l Conf. WorldWide Web (WWW), pp. 581-590, 2007.

[3] J. Teevan, S.T. Dumais, and E. Horvitz, "Personalizing Search viaAutomated Analysis of Interests and Activities," Proc. 28th Ann.Int'l ACM SIGIR Conf. Research and Development in InformationRetrieval (SIGIR), pp. 449- 456, 2005.

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

[5] B. Tan, X. Shen, and C. Zhai, "Mining Long-Term Search Historyto 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 WebSearch Based on User Profile Constructed without any Effortfrom Users," Proc. 13th Int'l Conf. World Wide Web (WWW),2004.

[7] X. Shen, B. Tan, and C. Zhai, "Implicit User Modeling forPersonalized Search," Proc. 14th ACM Int'l Conf. Information andKnowledge Management (CIKM), 2005.

[8] X. Shen, B. Tan, and C. Zhai, "Context-Sensitive InformationRetrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACMSIGIR Conf. Research and Development Information Retrieval (SIGIR),2005.

[9] F. Qiu and J. Cho, "Automatic Identification of User Interest forPersonalized Search," Proc. 15th Int'l Conf. World Wide Web(WWW), pp. 727-736, 2006.

[10] J. Pitkow, H. Schutze, 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-EnhancingPersonalized 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]. Xu, Yabo, et al. "Privacy-enhancing personalized web search." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007

[14]. Xiao, Xiaokui, and Yufei Tao. "Personalized privacy preservation", *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*. ACM, -2006.

[15]. Shou, Lidan, et al. "Supporting Privacy Protection in Personalized Web Search." (2012): 1-1.

[16]. 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.