# A Framework for Protecting Privacy of Users in Personalized Web Search

# بيئة عمل لحماية خصوصية المستخدمين في مواقع البحث الشخصية


المدرس المساعد
سعد علي محمد الفضلي
Saad_ff@yahoo.com

المدرس المساعد
علي محمد حميد الصفار
Ali_alsaffar88@yahoo.com

كليـــة الإمـــام الكـــاظم للعلـــوم الإســـلامية الجامعـــة



**Abstract:**

The world has become a small village due to the invent of Internet and related technologies. People of all walks of life started using Internet services for communication and exchange of information. Search engines became very popular as they can provide requiredinformation to people of all fields. Especially Google is used widely as search engine. Other search engines like Yahoo are also used similarly. The problem with search engines is that they support personalized web search to keep track of users' actions with respect to search key words and the results provided. However, they do lack in security to protect the privacy of online users. This is the main cause of concern. Not only personalized web search but also privacy to the identity of users is very important. Privacy is nothing but non-disclosure of identifies of users who perform search operations on different sensitive topics as well. In this paper we proposed a framework and an algorithm to ensure that the privacy of the users is not lost. We also built a prototype application that demonstrates the proof of concept. The empirical results are encouraging.



**Index Terms** –Search engines, web search, personalized web search, privacy protection.


## INTRODUCTION

Internet and World Wide Web (WWW) changed the way information is stored and retrieved. People of all walks of life started

using WWW for sharing information and resources. Now it became common practices to use search engines in order to obtain domain specific and general information. Search engines like Yahoo and Google are well known for extracting useful information from web. People of different sectors need their related information. Search engines provide a common platform for any kind of information search. Many search engines may not need to have the users' identity for searching. In fact, it is not mandatory for users to get authentication for searching. In fact, users over WWW can directly perform search operations in Google and other search engines.

However, Google and Yahoo kind of search engines are now providing personalized search. When user logs in and performs search, the search history is saved and associated with the user's profile. This is called profile based personalized web search. This will help the user to reuse the search history in future without spending much time. This will save time and effort on part of the users. However the PWS is throwing challenges to privacy of users. User identity is disclosed to adversaries who can infer sensitive information from the search history associated with users. To prevent this security problem, many researchers came with different measures to know the utility of the PWS.

In this paper, we proposed a framework and two algorithms are used [1] for implementing the framework. The remainder of the paper is structured as follows. Section II presents review of literature. Section III presents the proposed framework. Section IV provides prototype implementation. Section V presents experimental results while section VI provides the conclusions and future work.

## RELATED WORKS

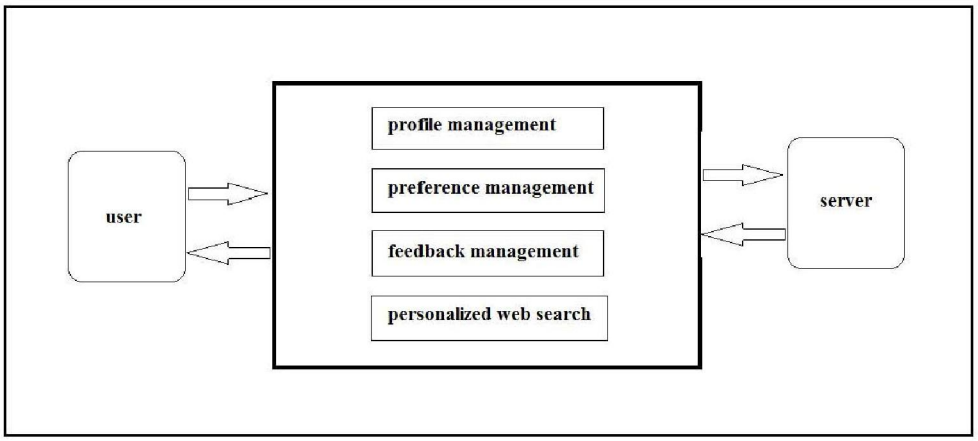
This section review relevant literature. Personalization has been around for some time. Profile based personalization improves utility of search. When individual's profile is maintained and the search process is associated with the profile. Hence is it is known as personalized web search (PWS). This approach has two important advantages such as representation of profiles and also effectiveness

when search is carried out. Vectors [2] and bag of words [3] used earlier for representing profiles. In Open Directory Project (ODP) most of the work was based on weighted topic hierarchy [4], [5], [6], [7]. Similar kind of research was carried out with respect Wikipedia [8], [9]. As per the literature, the common mechanism used in the PWS is Normalized Discounted Cumulative Gain [10]. Average rank is used in [4] and [11]. Average precision is used in [12] and [13]. These metrics are used in the literature in order to measure the effectiveness of the personalization utility.

There are two problems of classes identified in PWS. One class makes use of treaty privacy as explored in [14] while the second class is related to the sensitivity of data. There are different levels in solving the privacy issues in PWS. First level explored in [15] was proved fragile. The third level and fourth level also proved to be not suitable practically. Thus many efforts were made on second level [16], [17]. Online anonymity approach was explored in [18] besides useless user profiles in order to secure identity of users. Privacy preserving data publishing was explored in [19]. The works found in [20] and [4] explored on different effects of the personalization. They could also identify privacy issues with personalization in web search. In [20] effort was made to classify queries so as to take measures in ensuring privacy. GeedyDP algorithm was proposed in [21] to deal with privacy issues in PWS. It made use of heuristics in order to have improved efficiency in privacy protection. In this paper we built a framework to protect user identifies in PWS.

## PROPOSED SYSTEM

This section provides details of the proposed framework. The framework provides overview of the operations carried out in order to protect privacy of users. With respect to PWS performed with different search engines over WWW, it is understood that the identity disclosure along with sensitive search keywords have potential that adversaries can exploit them for many reasons. Preventing such identity disclosure is the aim of this paper. The framework shows provision for profile management, preference management, feedback management and personalized web search.




The profile management takes care of user profile and its modifications. The preference management can help in choosing personalized preferences with respect to privacy. The feedback management module is responsible to take feedback from users from time to time and use it in improving the search experience to end users. The GreedyPD and GreedyIL algorithms proposed in [1] were used to have prevented privacy disclosure problems. Besides the algorithms, our approach makes use of users' feedback from time to time to incorporate that into the main business intelligence of the framework. The personalized web search modules keeps track of history of search and helps users to track history and also find the results of old searches with ease. The underlying algorithms protect the queries from different issues pertaining to privacy.

## PROTOTYPE IMPLEMENTATION

The prototype application is implemented using Microsoft.NET platform. It is a simulation of an application that makes use of local search and provides search history. Backend used is SQL Server. The application makes use of the algorithms such as GreedyDP and GreedyIL [1] in order to achieve the functionality of the proposed framework. The implementation contains many users such as owner, admin and client. They have specific operations to be performed. Owner is responsible to prove identity by authentication, select

template for generating profile, and create the site with proper content and managing the same. He also sends the created template to server.

Admin user is responsible to analyze the nature of templates, view template search log, view owners and clients, block and unblock templates, activate and deactivate users. The client user is responsible to register with proper identity, search for the resource with proper text, resource obtained from server and search text ranked by the constant value and prioritized by one level upward. Figure 2 shows the data flow of owner.

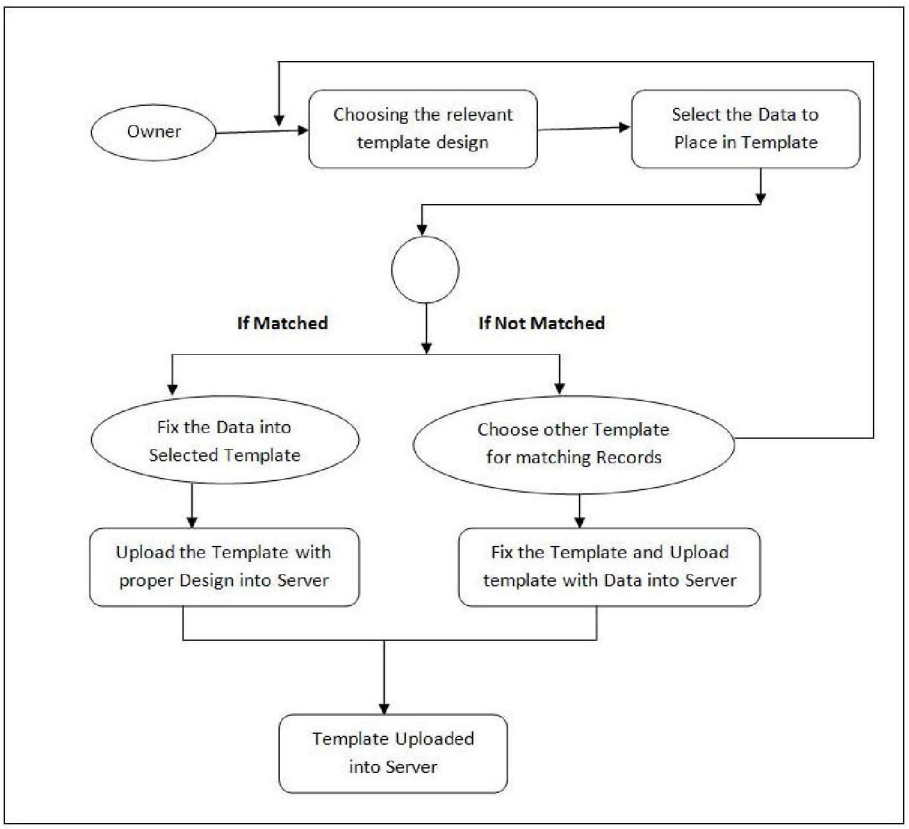



**Figure 2 – Data flow of owner**

As can be seen in Figure 2, it is evident that the owner data flow is provided. The owner is able to perform different activities that are in association with the server to which the application belongs to.

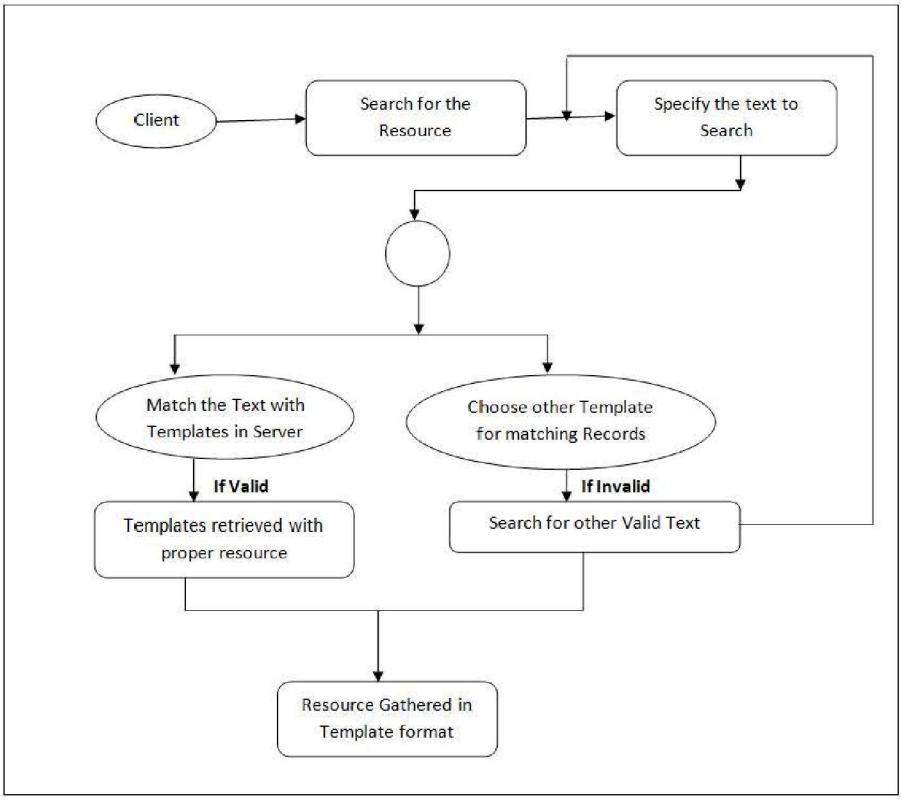



**Figure 3 – Shows data flow of client**

As shown in Figure 3, the client has different activities and verifications that can help in performing its duties pertaining to personalized search. The data is flows among processes and the template matching and resource availability in template format are considered. Both the client and owner are involved in activities that support PWS and the prevention of privacy attacks in the context of WWW search.

## EXPERIMENTAL RESULTS

We made experiments with the prototype application. The experiments are made in terms of average precision with different kinds of queries with Yahoo, ODP, scalability with GreedyDP and GreedyIL algorithms. The results also reveal the precision of privacy of the proposed framework.

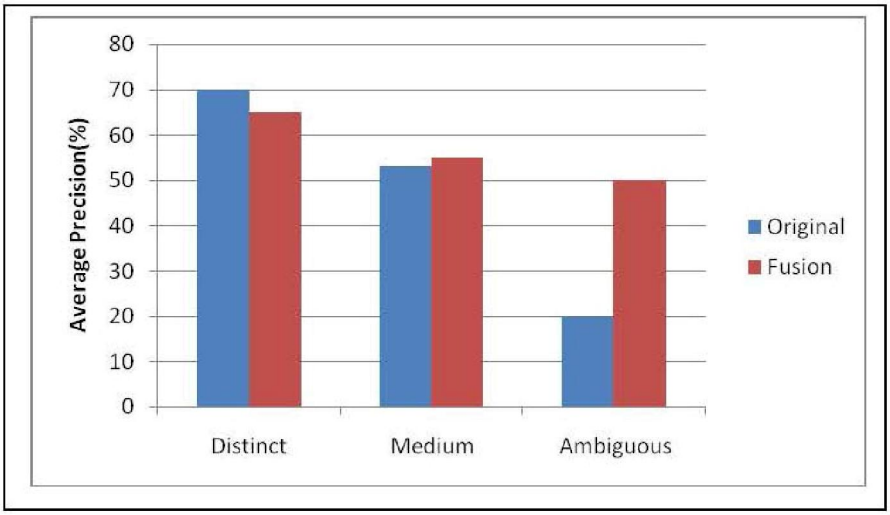



**Figure 1 – Effectiveness personalization on test queries (Yahoo)**

As shown in Figure 1, it is evident that the horizontal axis shows different types of queries such as distinct, medium and ambiguous. The vertical axis shows the average precision indicating the effectiveness of personalization. The results reveal that distinct queries exhibit high precision in both original and fusion models.

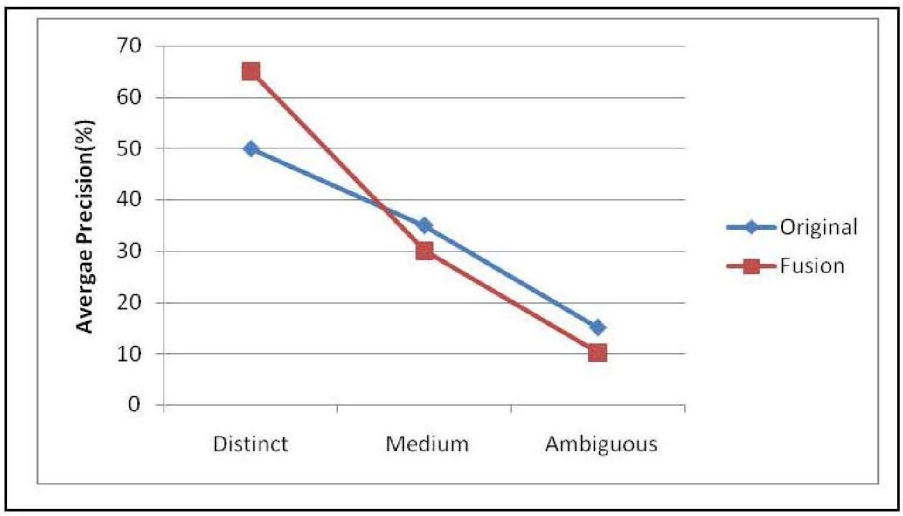



**Figure 2 – Effectiveness personalization on test queries (ODP)**

As shown in Figure 2, it is evident that the horizontal axis shows different types of queries such as distinct, medium and ambiguous. The vertical axis shows the average precision indicating the effectiveness of personalization. The results related to ODP and reveal that distinct queries exhibit high precision in both original and fusion models.

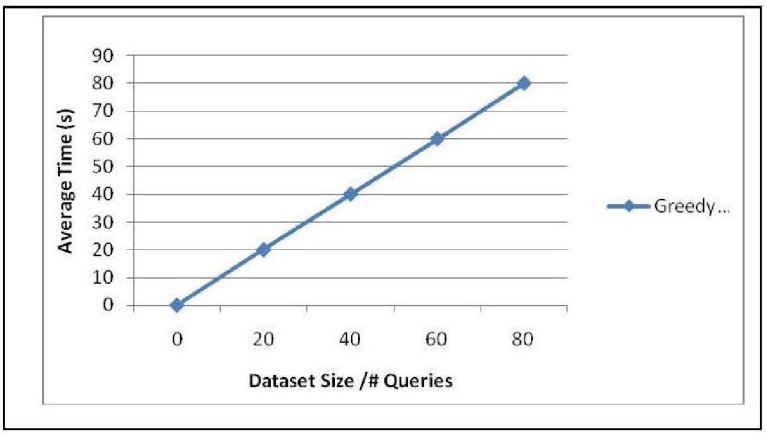



**Figure 3 – Performance with GreedyDP**

As shown in Figure 3, the results revealed that GreedyDP algorithm takes more time based on the size of the dataset. When size is increasing, its average time is increasing.

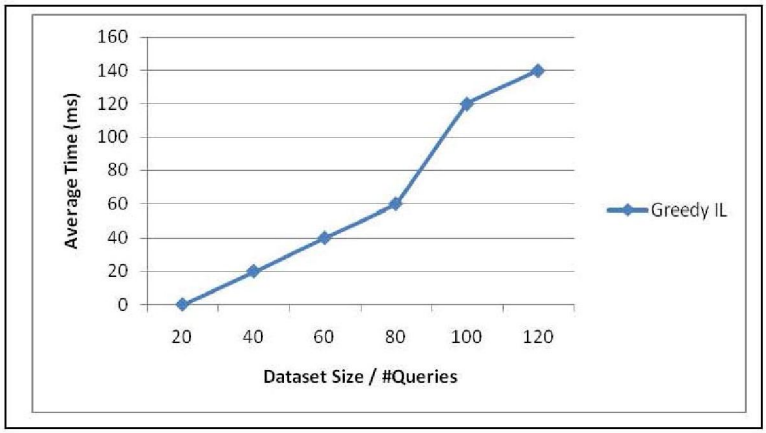



**Figure 4 – Performance with GreedyIL**

As shown in Figure 3, the results revealed that GreedyIL algorithm takes more time based on the size of the dataset. When size is increasing, its average time is increasing. However, when compared with GreedyDP, the GreedyIL has shown high performance as it takes very less time.

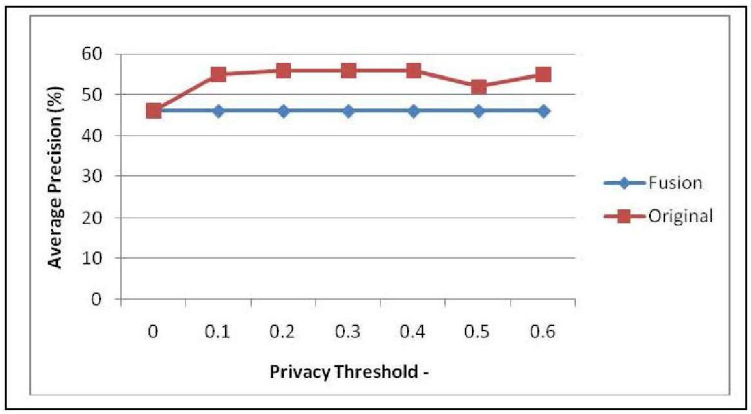



**Figure 5 – Effectiveness in privacy**

As shown in Figure 5, the results revealed that the average precision is high and continues without much change even if privacy threshold is changed.

## CONCLUSIONS AND FUTURE WORK

In this paper we studied the issues pertaining to the personalized web search. The issues were related to privacy of users. Sensitive information disclosure is the main problem with PWS though it provides plethora of benefits such as tracking search history and reusing it from time to time in future without spending much time again. Google, Yahoo and other search engines are providing history of users with respect to searching over web. This history has many benefits besides having privacy or identity disclosure issues. Many techniques came into existence to find out the effectiveness of PWS in terms of utility. Many measures were used in the literature to know the effectiveness of utility of PWS. However, they do lack in security to protect the privacy of online users. This is the main cause of concern. Not only personalized web search but also privacy to the

identity of users is very important. Privacy is nothing but non-disclosure of identifies of users who perform search operations on different sensitive topics as well. In this paper we implemented greedy algorithms which were recently proposed in[1] .We built a prototype application that demonstrates the proof of concept. Our empirical results revealed that the proposed framework with underlying algorithms was able to provide privacy to its users. This research can be extended further to enhance the privacy protection by combining different techniques and measures to know their effectiveness in future.

**الملخص**:

أصبح العالم كقرية صغيرة بسبب اختراع الانترنيت والتكنولوجيا ذات العلاقة . بدء الناس في كل مناحي الحياة يستخدمون خدمات الانترنيت للتواصل وتبادل المعلومات . وان محركات البحث أصبحت جدا شائعة لأنها تستطيع توفير المعلومات المطلوبة للناس لكل المجالات . خاصتا محرك البحث كوكل فانه يستخدم كمحرك بحث على نطاق واسع وأيضا محركات البحث الأخرى مثل الياهو الذي أيضا يعتبر شائعا . المشكلة في محركات البحث هي أنها تدعم شبكات البحث الشخصية لحفظ مسار إجراءات المستخدمين فيما يتعلق بالكلمات المفتاحية ونتائج البحث المقدمة .

مع ذلك فات هنالك ضعف في أمنية حماية الخصوصية لمستخدمي الانترنيت . هذا هو السبب الرئيسي للقلق ، ليس فقط شخصية شبكات البحث لكن حفظ خصوصية هوية المستخدمين ايضا جدا مهمة . الخصوصية هي لاشي سوى عدم الكشف عن هويات المستخدمين الذين ينفذون عمليات البحث حول مختلف المواضيع الحساسة . في هذا البحث نحن نقترح اطار بيئة عمل وخوارزمية لضمان حماية خصوصية المستخدمين لكي لا تنتهك، وايضا لقد قمنا ببناء نموذج تطبيق الذي يوضح إثبات المفهوم ، وان النتائج التجريبية كانت مشجعة .

**REFERENCES**

[1] Shou, L., Bai, H., Chen, K., and Chen, G. (2014). Supporting Privacy Protection in Personalized Web Search. IEEE, 26 (2), p1-15.

[2] 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.

[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] 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.

[5] 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.

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

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.
[17] 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.
[18] 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.
[19] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[20] 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.
[21] 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.