Web Search and Recommendation-based on User Interest for PWS

[1]Arun R. Desai, [2]Mr. Pankaj Chandre
[1]P.G. Student, [2]Assistant Professor
[1]Department of Computer Networks,
[1]Flora Institute of Technology, Pune, India

Abstract **- Numerous personalization approaches have been investigated but it is still unclear whether personalization is reliably effective on dissimilar queries for different users, and under different search contexts. Personalized web search (PWS) has established its effectiveness in increasing the quality of several search services on the web. This paper proposes a personalized web search (PWS) framework known as User customizable Privacy-preserving Search (UPS) that can adaptively specify profiles by queries while regarding user quantified privacy requirements. The system goals at striking a balance among two predictive metrics that estimate the utility of personalization as well as the privacy risk of exposing the generalized profile. The greedy algorithm namely GreedyIL is presented for runtime generalization. Additionally, paper provides an online web age prediction mechanism for deciding whether personalizing a query is helpful. Additionally, this paper proposes a Personalized Web page Recommendation model (PWR) through collaborative filtering and a topic-aware Markov model. Topic-aware Markov model is used to widely applied to learn users' navigation behaviors for predicting the next step while surfing the Web.**

IndexTerms **- Privacy preservation, personalized web search, recommendation, profile privacy risk, user profile.**

I. INTRODUCTION

Personalization has been an active research area in the last some years and construction of user profile is an important component of any personalization scheme. Explicit customization has been generally used to personalize the look and content of several web sites, personalized search [11] methodologies focus on indirectly building and developing user profiles. Corporations that make available marketing data report that search engines are used progressively as referrals to web sites, compared to direct navigation and web links. As search engines make a larger role in commercial applications, the desire to increase their effectiveness grows. However, search engines are affected by difficulties such as ambiguity and outcomes ordered by web site popularity rather than interests of user.

Though various information retrieval methods (for instance, web search engines applications and digital library systems) have been effectively installed, the present retrieval systems are far from optimal. A key deficiency of present retrieval schemes is that they usually lack of user modeling and are not adaptive to individual users. This characteristic non-optimality is seen openly in the subsequent two cases: (1) Different users can use the identical query (e.g., "Java") to search for dissimilar information (for example, the Java island located in Indonesia or the Java programming language), however existing IR methods return the identical results for these users. Without considering the actual user, it is difficult to know which sense "Java" refers to in a query. (2) A user's data needs can change over time. The similar user can use "Java" sometimes to mean the Java Island in Indonesia and some other times to mean the programming language. It would be impossible to recognize the correct sense without recognizing the search context.

So as to optimize retrieval accuracy, there is need to model the user suitably and personalize search according to every individual user. The main objective of user modeling for information retrieval is to accurately model a user's information requirement, which is, inappropriately, a very problematic task. Indeed, it is hard for a user to exactly define what his/her information necessity is.

The web search engine has become the maximum important portal for normal people observing for valuable information on the web. Though, users might experience failure when search engines return irrelevant results that do not meet their real meanings. Such irrelevance is mostly due to the enormous contexts of users and backgrounds, in addition to the ambiguity of texts. PWS is a common group of search methods aim to provide improved search results that are personalized for needs of individual user. As the outcome, user information has to be collected and examined to understand the user purpose behind the delivered query.

The way out to PWS can usually be characterized into two categories, viz. click-log-based approaches and profile-based ones. The click-log based approaches are straightforward they just impose bias to clicked pages in the history of user's query. Though this approach has been established to perform consistently as well as considerably well [1], it can simply work on repeated queries from the identical user, which is a strong drawback restricting its applicability. In contrast, profile-based approaches improve the search knowledge with problematical user-interest models created from user profiling methods. Profile-based approaches can be possibly effective for majority kinds of queries, but are described to be unstable under some conditions [1]. Though there are pros as well as cons for both types of PWS methods, the profile-based PWS has established extra effectiveness in improving the quality of web search with increasing usage of personal and behavior data to profile its users,

which is typically assembled implicitly from query history [2], [3], [4], click-through data [7], [8], [1] bookmarks [9], browsing history [5], [6], documents of user [2], [10], and so forth.

To protect user privacy in profile-based PWS, researchers have to consider two contradicting effects through the search procedure. 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 [10], [12] suggest that people are willing to compromise privacy if the personalization 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 tradeoff between the search quality and the level of privacy protection achieved from generalization. Previous works on profile-based PWS mainly focus on improving the search utility. The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP1 [1], [14], Wikipedia [15], and so on.

With the rapid growth of the Web, it becomes more and more difficult for Web users to find useful information. In particular, a Web user often wanders aimless on the Web without visiting pages of his/her interests, or spends a long time to find the expected information. Web page recommendation is thus proposed to address this problem. It aims to understand the users' behaviors, and guide users to visit pages of their interests at a specific time. An essential task of Web page recommendation is to understand users' navigation behaviors from their Web usage data, and devise a model to predict what pages the users are more likely to visit at the next step.

The rest of paper is organized as follows: Section II gives the essential literature survey. Section III addresses existing. Section IV introduces the proposed system architecture. Section V describes proposed system setup and section VI describes expected results. Section VII concludes the paper.

II. RELATED WORK

In the literature review the topical methods over secure data retrieval are going to discuss.

Z. Dou et al. [3] presented a large-scale evaluation framework for personalized search based on query logs, and then evaluate five personalized search strategies (including two click-based and three profile-based ones) using 12-day MSN query logs. Click-based personalization strategies presented in [3] are straightforward and stable though they can work only on repeated queries. The profile-based personalized search strategies proposed in this paper are not as stable as the click-based ones.

J. Teevan et al. [4] explore rich models of user interests, built from both search-related information, such as previously issued queries and previously visited Web pages, and other information about the user such as documents and email the user has read and created. This technique shows that it is possible to provide effective and efficient personalized Web search using a rich and automatically derived user profile. This system must improve ability to personalize search.

M. Spertta and S. Gach [5] implemented a wrapper for Google to examine different sources of information on which to base the user profiles: queries and snippets of examined search results. The scheme able to demonstrate that information readily available to search engines is sufficient to provide significantly improved personalized rankings. The concept hierarchy is static and best results occurred when conceptual ranking considered only one concept from the query-based profile, and two from the snippet-based profile.

B. Tan et al. [6] introduces statistical language modeling based methods to mine contextual information from long term search history and exploit it for a more accurate estimate of the query language model. The mixture model used in this paper is quite simple. Found through study of different cutoffs in search history that although recent history is more important, remote history is also useful, especially for recurring queries.

X. Shen et al. [7] presented a decision theoretic framework and develop techniques for implicit user modeling in information retrieval. The system develops an intelligent client-side web search agent (UCAIR) that can perform eager implicit feedback. The time complexity for decision making increases as data increases.

X. Shen et al. [8] proposed several context sensitive retrieval algorithms based on statistical language models to combine the preceding queries and clicked document summaries with the current query for better ranking of documents. Click through history mechanism substantially improve retrieval performance without requiring any additional user effort. Privacy issues for this system are rising from the lack of protection for such data.

III. EXISTING APPROACH

To protect user privacy in profile-based PWS, researchers have to consider two contradicting effects during the search process. On one contrary, researchers attempt to increase the quality of search with the personalization utility of the user profile. On the other contrary, they require to hide the privacy contents that are existing in the user profile to place control on the privacy risk. The issues with the existing methods are described in the following observations:

The existing profile-based PWS do not maintaining runtime profiling.

A user profile is normally generalized for simply once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries. One suggestion reported in [1] is that profile-based personalization may not even assistance to increase the search quality for particular ad hoc queries; however exposing profile of user to a server side has put the user's privacy at risk. An improved approach is to create an online choice on:

1. whether to personalize the query (by exposing the user's profile) and
2. What to describe in the user profile at runtime.

To the best of knowledge, no prior study has supported such feature.

The existing approaches do not allow for the customization of privacy requirements.

This perhaps makes nearly user privacy to be overprotected while other privacies insufficiently protected. For instance, in [10], all the sensitive subjects are identified using an absolute metric named surprisal based on the information theory, supposing that the interests with less user document support are extra sensitive.

Many personalization methods require iterative user communications once creating personalized search results.

They typically improve the search results with particular metrics which need multiple user communications, for example rank scoring [13], average rank [8], etc. This example is, though, infeasible for runtime profiling, as it will not simply pose too much privacy breach risk, but moreover demand prohibitive processing time for profiling. Therefore, predictive metrics is needed to measure the search quality as well as breach risk after personalization, deprived of incurring iterative user interaction.

Solution on this issue is User customizable Privacy-preserving Search and Recommendation (UPSR) framework. The framework assumes that the queries do not hold any sensitive data, and targets at protecting the privacy in different user profiles while holding their usefulness for PWS. Framework also integrates personalized Web page recommendation paradigm to predict the pages that Web users are intent in devoid of explicitly asking for them.

IV. PROPOSED ARCHITECTURE

UPS involves of a non-trusty search engine server including with a number of clients. Each client (or user) retrieving the search service beliefs no one but himself/herself. The key elements for privacy protection are an online profiler implemented as a search proxy running on the client machine itself. The proxy preserves both the whole user profile, in a hierarchy of nodes by means of semantics, and the user-specified (or customized) privacy requirements signified as a set of sensitive-nodes.

Architecture Overview

The main contributions of proposed framework are summarized as following:

- The framework proposes a privacy-preserving personalized web search UPS, which can simplify profiles for every query allowing to user-specified privacy necessities.
- The problem of privacy-preserving personalized search is formulated by relying on the definition of two conflicting metrics, specifically personalization utility and privacy risk, for hierarchical user profile.
- The effective generalization algorithm is developed, GreedyIL, to support runtime (online) profiling. While the previous attempts to maximize the discriminating power (DP), the final goes to minimize the information loss (IL). By developing a number of heuristics, GreedyIL outdoes GreedyDP considerably. The main problem of GreedyDP is that it requires re-computation of all candidate profiles (together with their discriminating power and privacy risk) generated from attempts of prune-leaf. This causes significant memory requirements and computational cost.
- An inexpensive mechanism for the client is provided to decide whether to personalize a query in UPS.
- A collaborative filtering framework is exploited for personalized Web page recommendation. Collaborative filtering is a common method for personalization in numerous applications on the Web.

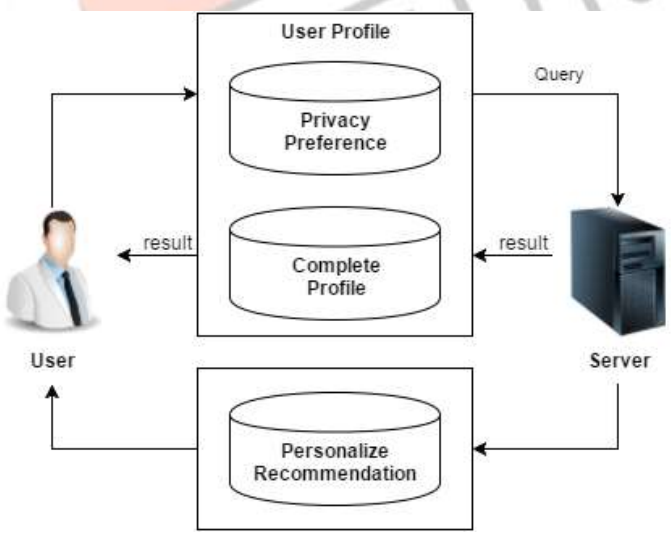

***Fig 1**: Proposed Architecture*

Architecture Explanation

The proposed system consists of four entities:

- Profile-Based Personalization.
- Privacy Protection in PWS System.
- Generalizing User Profile.

- Online Decision.

Profile-Based Personalization

This paper presents an approach to personalize digital multimedia content based on user profile data. For this, two key mechanisms were developed: a profile generator that automatically creates user profiles on behalf of the user preferences, and a content-based recommendation algorithm that estimates the user's interest in unknown content by matching her profile to data descriptions of the content. Each option is integrated into a personalization system.

Privacy Protection in PWS System

This paper proposes a PWS framework called UPS that can simplify profiles in for every query rendering to user-specified privacy requirements. Two predictive metrics are proposed to evaluate the privacy breach risk as well as the query utility for hierarchical user profile. Two simple but effective generalization algorithms are developed for user profiles allowing for query-level customization using our proposed metrics. This paper also provides an online prediction mechanism based on query utility for deciding whether to personalize a query in UPS.

Generalizing User Profile

The generalization process has to encounter specific fundamentals to handle the user profile. This is accomplished by preprocessing the user profile. At initial stage, the process prepares the user profile by taking the designated parent user profile into account. The procedure adds the inherited properties to the properties of the local user profile. Afterward the procedure loads the data for the foreground as well as the background of the map rendering to the designated selection in the user profile.

Additionally, using references enables caching and is helpful when considering an implementation in a production environment. The reference to the user profile can be used as an identifier for previously processed user profiles. It allows performing the customization process once, but reusing the result multiple times. Though, it has to be made guaranteed, that an update of the user profile is too propagated to the generalization process. This wants specific update strategies, which check after a specific timeout or a specific event, if the user profile has not changed yet. Additionally, as the generalization process involves remote data services, which might be updated frequently, the cached generalization results might become outdated. Therefore selecting a specific caching strategy requires careful analysis.

Online Decision

The profile-based personalization contributes minute or even decreases the search quality, though exposing the profile to a server would for assured risk the user's security. To address this issue, this paper develops an online mechanism to choose whether to personalize a query. The simple idea is straightforward. The entire runtime profiling will be terminated and the query will be sent to the cloud server deprived of a user profile, if a distinct query is identified through generalization.

Personalizing Web Page Recommendation

In this paper, personalization into Web page recommendation is introduced by discovering users' profiles from browsing logs as well as measuring user similarities. The existing topic-aware Markov model captures together temporal and topical relevance of Web pages. We assume that two users are similar if they have visited many Web pages in common, or pages about relevant topics. We take the similarity between two sets of topics as the similarity between the two corresponding users. Computing the overlap between two sets is straightforward but cannot capture the relationship within similar topics. In contrast, we measure the similarity between each pair of topics and adopt a maximum weight bipartite matching algorithm to derive the similarities between Web users.

V. PROPOSED SYSTEM SETUP

Offline Search

Precisely, every user has to assume the subsequent procedures:
1. Offline profile construction
2. Offline privacy requirement customization

Offline-1. Profile Construction.

The main stage of the offline handling is to create the original profile of the user in a topic hierarchy *H* that releases user privacies. Assume that the partialities of user are indicated in a set of plain text document is denoted by *D*. To generate the profile, produce the subsequent stages:

1. Identify the specific topic in *R* for every document $d \in D$. Hence, the set of preference document *D* is changed into a set of topic *T*.
2. Generate the user profile *H* as a topic path trie with *T*, such as,
$$H = trie(T) \quad (1)$$
3. Initialize the user support $\sup_H(t)$ for every topic $t \in T$ through its document support from *D*, and then calculate $\sup_H(t)$ of supplementary nodes of *H* with equation
$$\sup_H(t) = \sum_{t' \in C(t,H)} \sup_H(t') \quad (2)$$

There is one open query in the above development process that how to detect the particular topic for every document $d \in D$.

Offline-2. Privacy Requirement Customization.

This procedure first requests the user to specify a set of sensitive node $s \in H$, as well as corresponding sensitivity value $sen(s) > 0$ for every topic $s \in S$. Resulting, the cost layer of the profile is complete by calculating the cost value of every node $t \in H$ as follows:

1. For every sensitive node, $\cos t(t) = sen(t)$;
2. For every single non-sensitive leaf node, $\cos t(t) = 0$;
3. For every one non-sensitive internal node, $\cos t(t)$ is recursively specified via in a bottom-up way:

$$\cos t(t) = \sum_{t' \in C(t,H)} \cos t(t') \times \Pr(t' | t) \qquad (3)$$

Thus far, customized profile with its cost layer accessible has been attained.

Online Search

Online: Query-topic Mapping

Assumed a query q, the drives of query-topic mapping are 1) to calculate a rooted subtree of H, which is known as a seed profile, with the intention of all topics relevant to q are enclosed in it; and 2) to find the preference values among q and all topics in H. This process is performed in the following steps:

1. Find the topics in R that are relevant to q.
2. Overlap $R(q)$ (it is usually a small fraction of R) with H to obtain the seed profile G_0, which is also a rooted subtree of H.

Online: Profile Generalization

This process generalizes the seed profile G_0 in a cost-based iterative manner depend on the privacy as well as utility metrics. This profile generation takes place by means of generation algorithm called as GreedyIL [1]. The GreedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings.

Topic-Aware MARKOV Model

In this paper a topic-aware Markov model is proposed to learn users' navigation behaviours. A topic-aware Markov model is used to captures both temporal and topical relevance of Web pages. Markov-model based methods consist of the subsequent steps.

Session partition.

A session is a sequence of pages ordered by access time also reflects user missions within a time interval t_θ (e.g., 30 minutes). Formally, a session with length m can be defined as $\langle p_1, p_2, ..., p_m \rangle$, where the time difference of pm and p1 does not exceed t_θ. After being sorted first by user ids and then by timestamps, the records in the browsing log L are partitioned into sessions.

State determination.

Each k-gram of a session is called a state, and k denotes the model's order. Formally, given a session $\langle p_1, p_2, ..., p_m \rangle$, the j^{th} state of a kth-order Markov model is $S_j^k = \langle p_j, p_{j+1}, ..., p_{j+1}, ..., p_{j+k-1} \rangle$, $1 \le j \le m-k+1$.

Conditional probability estimation.

Given a state S_j^k, the probability that the page pi is requested next by the active user is projected by the ratio of the frequency of S_j^k followed by the page p_i to the frequency of S_j^k, i.e.,

$$P(p_i | S_j^k) = frequency(S_j^k, p_i) / frequency(S_j^k)$$

Personalized Recommendation

This paper exploits a collaborative filtering framework for personalized Web page recommendation. Collaborative filtering is a popular technique for personalization in many applications on the Web.

Three stages are taken to resolve the problem of personalizing Web page recommendation.

First, for individual Web user u_i, $i = 1,2,...,N$, try to equal the prefix Φ with the conditions of his/her navigation behaviour model. The restricted probabilities of individual page p following Φ are projected as stated before, and p is taken as a candidate.

Second, the score of individual candidate is considered based on the idea of collaborative filtering. Definitely, for every candidate p, combine its probabilities in individual user's model with user similarities as follows.

$$k(p | a, \Phi) = \sum_{u_i \in U} sim(u_i, a) . \frac{frequency_{u_i}(\Phi, p)}{frequency_{u_i}(\Phi)} \qquad (4)$$

Finally, sort the candidates by their scores in descending order and recommend top-k pages to the active user a. Note that the candidate size may be smaller than k. If more recommendations are requested in this case, Φ will be augmented by one candidate page with the highest score to fetch more results.

VI. EXPECTED RESULTS

The scalability of the algorithms by varying 1) the seed profile size (i.e., number of nodes), and 2) the data set size (i.e., number of queries). The search results are re-ranked with the generalized profile output by GreedyIL over 30 target users. The final search quality is evaluated using the Average Precision of the click records of the users, which is defined as

$$AP = \sum_{i=1}^{n} \frac{i}{l_i.rank} / n \quad (5)$$

Where li is the ith relevant link at position rank identified for a query, and n is the number of relevant links.
Average time require for GreedyIL algorithm according to database is as given in graph below.

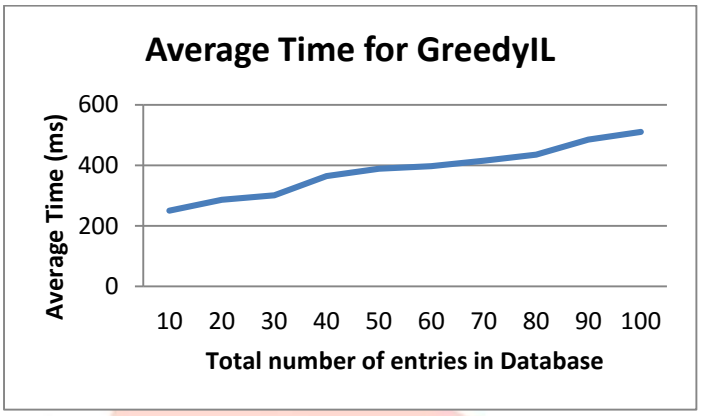

***Fig 2**: Time in (ms) for algorithm GreedyIL*

The web page recommendation according particular topic to user can shown as in graph given below. The topics are used in experiments are sports, arts Computer Science (CS), etc.

Categories	Total Documents	Actual Recommendation	Predicted by PWR
Sports	10	5	5
Arts	10	6	5
Computer Science	10	6	6
Books	10	7	5
Adults	10	4	4

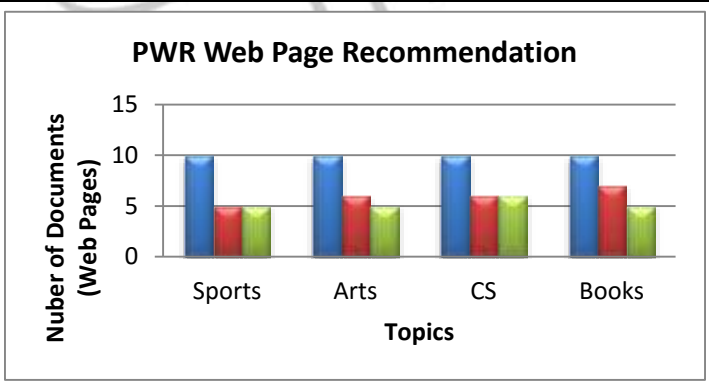

***Fig 3**: Recommendation to user*

VII. CONCLUSION

This paper presented a client-side privacy protection framework known as UPS for personalized web search and Personalize Web Recommendation. The framework permitted users to specify customized privacy requirements via the hierarchical profiles. Furthermore, UPS also achieved online generalization of user profiles to secure the personal privacy deprived of compromising

the search quality. This paper proposed greedy algorithm GreedyIL for the online generalization. In this paper, new model of personalized Web page recommendation (PWR) is studied to predict the Web pages that Web users are interested in devoid of explicitly asking for them. Taking the user similarities into consideration, personalized Web page recommendation is used to meet different preferences of Web users. Additionally, devised a novel model for learning the navigation patterns which is contribute to the topically coherent recommendations.

VIII. ACKNOWLADGEMENT

The author would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the Prof. Pankaj Chandre for his valuable guidance and constant guidelines also thank full the computer department staff of Flora Institute of Technology, Pune, and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] Lidan Shou, Gang Chen, and He Bai, Ke Chen,, "Supporting Privacy Protection in Personalized Web Search", IEEE Transactions On knowledge and data engg., vol 26, no. 2, 2014.

[2] Qingyan Yang, Ju Fan, Lizhu Zhou and Jianyong Wang,, "PersonalizingWeb Page Recommendation via Collaborative Filtering and Topic-Aware Markov Model," IEEE/ACM Transactions On Data Mining, 2010.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[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] 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.

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