Privacy Preserving of Contextual user Profiles in Search Engine Repository

S. Haripriya[1]
PG Scholar,
Department of Information Technology,
E.G.S.Pillay Engineering College,
Nagapattinam

R. Indumathi[2]
PG Scholar,
Department of Information Technology,
E.G.S.Pillay Engineering College,
Nagapattinam

V .M. Suresh[3]
Assistant Professor,
Department of Information Technology,
E.G.S.Pillay Engineering College,
Nagapattinam

Abstract: **Retrieving the most relevant information for the Web becomes difficult because of the huge amount of documents available in various formats. One approach to satisfy the requirements of the user is to personalize the information available on the Web, called Web Personalization. PWS is the present techniques has proved that increases the quality of searching the services on web but the user privacy is the major problem in the wide proliferation of PWS. In the proposed system, implementing the String Similarity Match Algorithm (SSM Algorithm) for improving the better search quality results. To address this privacy threat, current solutions propose new mechanisms that introduce a high cost in terms of computation and communication. And present a novel protocol specially designed to protect the users' privacy in front of web search profiling. Personalized search is promising way to improve the accuracy of web search.. It aims on runtime generalisation and customization of user profile, thus providing privacy and improving the quality of search services.**

Keywords: Privacy protection, personalized web search, utility, risk, profile

1. INTRODUCTION

Communication networks enable us to reach a very large volume of information in a minimal amount of time. Furthermore, that huge quantity of data can be accessed at any time and any place with a capable device (e.g. a laptop, a PDA, etc.) and an Internet connection. Nowadays, it is pretty common to access easily to both resources. In the future, it will be even easier. However, useful information about a specific topic is hidden among all the available data and it can be really challenging to find it since that information can be scattered around the Word Wide Web.

Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to do this hard job for us. The 84% of the Internet users have used a web search engine at least once. For the 32%, web search engines are an essential tool to address their everyday duties [1]. Among the different search engines, Google is the most used in the US with a 43.7% of the total amount of searches performed in 2006 [2]. Google improves its performance (it gives personalized search results) by storing a record of visited sites and past searches submitted by each user [3] (Web History). Those searches can reveal a lot of information from individual users or the institutions they work for. For example, let us imagine an employee of a certain company A. This employee uses Google to obtain information about a certain technology. If a company B, which is a direct competitor of A, knows this situation, it can infer that this technology will be used in the new products offered by A. This knowledge gives to B an important advantage over A. Another example of this situation occurs when a person is applying for a certain job. In this case, if the employer knows that the applicant has been looking for information regarding a certain disease, she can use this knowledge to choose another person for the job. In both examples, the attacker (the entity who gets some advantage over the other) benefits from the lack of a privacy-preserving mechanism between the user and the web search engine.

1.1 Contributions

In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling. In this we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. In the Existing System, Greedy IL and Greedy DP algorithm, it takes large computational and communication time.

For generalize the retrieved data by using the background knowledge. Through this we can resist the adversaries. Privacy protection in publishing transaction data is an important problem. A key feature of transaction data is the extreme sparsity, which renders any single technique ineffective in anonymizing such data. Among recent works, some incur high information loss, some result in data hard to interpret, and some suffer from performance drawbacks. This paper proposes to integrate generalization and compression to reduce information loss. However, the integration is nontrivial. We propose novel techniques to address the efficiency and scalability challenges.

1.2 Problem Statement

In the Existing System, they presented a client-side privacy protection framework called UPS for personalized web search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. They proposed two greedy algorithms namely GreedyDP and GreedyIL, for the online generalization. It achieves quality search results while preserving user's customized privacy requirements. It also improves effectiveness and efficiency. But in the Existing system, it uses only the generalization concept. It degrades the performance of existing system. For this we are going implement and extend the process by using some other properties such as exclusiveness and to make a system capable to capture a series of queries. In the Existing System, it has a high cost in terms of computation and communication. Existing System have three system architectures. In these three components has been used. There are server, client and proxy. Client information's are shared to the proxy. In the proposed system, information's has
exclusiveness. It cannot be shared to the privacy. When the searched information's are generalized and then only information's are stored in the history. Only hided information's are stored into the history. String Similarity

Match Algorithm (SSM Algorithm) is better than the greedy algorithm. It achieves more accuracy in search results.

II. SYSTEM ARCHITECTURE

2.1Definition of terms

2.1.1Data Mining

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.
Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software tools for analyzing data. It allows users to analyze data categorize it, and summarize the relationships identified. Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2.2.2 What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

2.2.3 How data mining work?

Data mining provides the link between transaction and analytical systems, Data mining software analyses relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

-

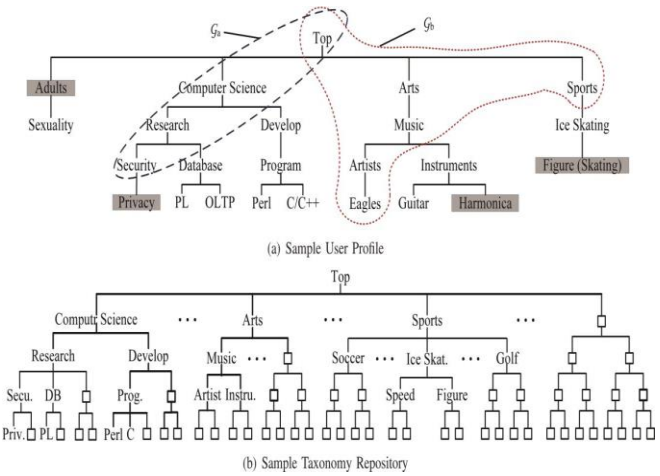

Fig 2.1: System Architecture

2.3 Data mining consists of five major elements:

Extract, transform, and load transaction data onto the data warehouse system.

- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

2.4 Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
- **Rule induction**: The extraction of useful if-then rules from data based on statistical significance.
- **Data visualization**: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

Existing System:

Algorithm Used—Greedy Information Loss Algorithm (Greedy IL)

In the Existing System, each user has to undertake the following procedures. 1. Offline profile construction,

2. Offline privacy requirement customization,
3. Online query-topic mapping, and

4. Online generalization.

Normally, user posts the query and retrieves the information from the server. In several systems, information is loosed due to the algorithm inefficiency. In this, Greedy IL algorithm minimizes the information loss during retrieving the information's. The advantage of GreedyIL over GreedyDP is more obvious in terms of response time. This is because GreedyDP requires much more computation of DP, which incurs lots of logarithmic operations. The problem worsens as the query becomes more ambiguous. For instance, the average time to process GreedyDP for queries in the ambiguous group is more than 7 seconds. In contrast, GreedyIL incurs a much smaller real-time cost, and outperforms GreedyDP by two orders of magnitude. GreedyIL displays near-linear scalability, and significantly outperforms Greedy

2.5 Algorithms for Proposed System

Step1: Detecting & removal of unwanted symbols

Step2: compute similarity calculation for user given word and word in database
Step3: In that similarity calculation, extract the features in the dataset.

Step4: Then estimate the ASCII difference for user given word and words in database

SteP5: The estimate the similarity values.

Step6:Then retrieve the most relevant documents based on the similar values

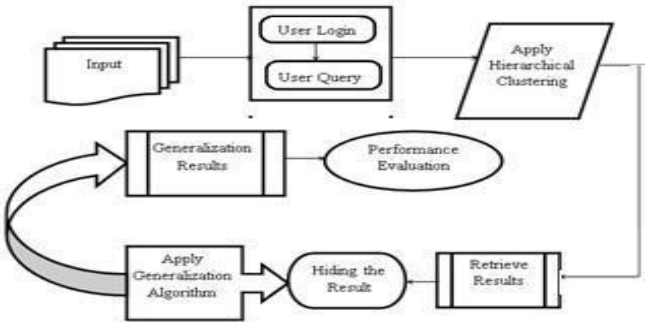

Fig.2.1 Steps in Existing sys

III. EXISTING SYSTEM

In the Existing Work, a client-side privacy protection framework called UPS for personalized web search was proposed. UPS could theoretically be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The context allowed users to stipulate customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. In this they proposed two greedy algorithms, namely GreedyDP and GreedyIL, for the online generalization. In this for query mapping process it has various steps to compute the relevant items.

Most works on anonymization focus on relational data where every record has the same number of sensitive attributes. There are a few works taking the first step towards anonymizing set-valued or transactional data where sensitive items or values are not clearly defined. While they could be potentially applied to user profiles, one main limitation is that they either assume a predefined set of sensitive items that need to be protected, which are hard to done in the web context in practice, or only guarantee the anonymity of a user but do not prevent the linking attack between a user and a potentially sensitive item.

Another approach to provide privacy in web searches is the use of a general purpose anonymous web browsing mechanism. Simple mechanisms to achieve a certain level of anonymity in web browsing include: (i) the use of proxies; or (ii) the use of dynamic IP addresses.

3.1 Disadvantages

It has demonstrated the ineffectiveness or privacy risks of naive anonymization schemes. The utility of the data is limited to statistical information and it is not clear how it can be used for personalized web search. For retrieving the user query results, it takes high computational and communication time and also cost. Proxies do not solve the privacy problem. This solution only moves the privacy threat from the web search engine to the proxies themselves. A proxy will prevent the web search engine from profiling the users, but the proxy will be able to profile them instead. The renewal policy of the dynamic IP address is not controlled by the user but the network operator.

3.2 User Profile

Consistent with many previous works in personalized web services, each user profile in UPS adopts a hierarchical structure. Moreover, our profile is constructed based on the availability of a public accessible taxonomy, denoted as R, which satisfies the following assumption.

Assumption1. The repository R is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic t, a corresponding node (also referred to as t) can be found in R, with the subtree as the taxonomy accompanying t.

The repository is regarded as publicly available and can be used by anyone as the background knowledge. Such repositories do exist in the literature, for example, the ODP, Wikipedia, WordNet , and so on. In addition, each topic t 2 R is associated with a repository support, which quantifies how often the respective topic is touched in human knowledge. If we consider each topic to be the result of a random walk from its parent topic in R, we have the following recursive equation:

Equation (1) can be used to calculate the repository support of all topics in R, relying on the following assumption that the support values of all leaf topics in R are available. Assumption 2. Given a taxonomy repository R, the repository support is provided by R itself for each leaf topic.
In fact, Assumption 2 can be relaxed if the support values are not available. In such case, it is still possible to "simulate" these repository supports with the topological structure of R. That is, can be calculated as the count of leaves in

Based on the taxonomy repository, we define a probability model for the topic domain of the human knowledge. In the model, the repository R can be viewed as a hierarchical partitioning of the universe (represented by the root topic) and every topic t 2 R stands for a random event. The conditional probability (s is an ancestor of t) is defined as the proportion of repository support.

3.3 Generalizing User Profile

Now, we exemplify the inadequacy of forbidding operation. In the sample profile in Fig. 2a, Figure is specified as a sensitive node. Thus, only releases its parent Ice Skating. Unfortunately, an adversary can recover the subtree of Ice Skating relying on the repository shown, where Figure is a main branch of Ice Skating besides Speed. If the probability of touching both branches is equal, the adversary can have 50 percent confidence on Figure. This may lead to high privacy risk if is high. A safer solution would remove node Ice Skating in such case for privacy protection. In contrast, it might be unnecessary to remove sensitive nodes with low sensitivity. Therefore, simply forbidding the sensitive topics does not protect the user's privacy needs precisely.

To address the problem with forbidding, we propose a technique, which detects and removes a set of nodes X from H, such that the privacy risk introduced by exposing G ¼ is always under control. Set X is typically different from S. For clarity of description, we assume that all the subtrees of H rooted at the nodes in X do not overlap each other. This process is called generalization, and the output G is a generalized profile.

The generalization technique can seemingly be conducted during offline processing without involving user queries. However, it is impractical to perform offline generalization due to two reasons:

1. The output from offline generalization may contain many topic branches, which are irrelevant to a query. A more flexible solution requires online generalization, which depends on the queries. Online generalization not only avoids unnecessary privacy disclosure, but also removes noisy topics that are irrelevant to the current query.

   For example, given a query qa ¼ "K-Anonymity," which is a privacy protection technique used in data publishing, a desirable result of online generalization might be Ga, surrounded by the dashed ellipse in Fig. 2a. For comparison, if the query is qb ¼ "Eagles," the generalized profile would better become Gb contained in the dotted curve, which includes two possible intentions (one being a rock band and the other being an American football team Philadelphia Eagles). The node sets to be removed are Xa ¼ fAdults; Privacy; Database; Develop; Arts; Sportsg, and Xb ¼ fAdults;Computer Science; Instrument; Ice Skatingg, respectively.

2. It is important to monitor the personalization utility during the generalization. Using the running example, profiles Ga and Gb might be generalized to smaller rooted subtrees. However, overgeneralization may cause ambiguity in the personalization, and eventually lead to poor search results. Monitoring the utility would be possible only if we perform the generalization at runtime.

We now define the problem of privacy-preserving generalization in UPS as follows, based on two notions named utility and risk.

The former measures the personalization utility of the generalized profile, while the latter measures the privacy risk of exposing the profile.

3.5Attack Model

Our work aims at providing protection against a typical model of privacy attack, namely eavesdropping. As shown in Fig. 3, to corrupt Alice's privacy, the eavesdropper Eve successfully intercepts the communication between Alice and the PWS-server via some measures, such as man-in-themiddle attack, invading the server, and so on. Consequently, whenever Alice issues a query q, the entire copy of q together with a runtime profile G will be captured by Eve. Based on G, Eve will attempt to touch the sensitive nodes of

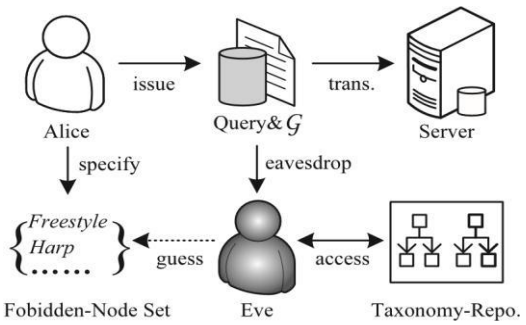

Fig. 3.Attack model of personalized web search.

Alice by recovering the segments hidden from the original H and computing a confidence for each recovered topic, relying on the background knowledge in the publicly available taxonomy repository R.

Note that in our attack model, Eve is regarded as an adversary satisfying the following assumptions:

Knowledge bounded. The background knowledge of the adversary is limited to the taxonomy repository R. Both the profile H and privacy are defined based on R.

Session bounded. None of previously captured information is available for tracing the same victim in a long duration. In other words, the eavesdropping will be started and ended within a single query session.

The above assumptions seem strong, but are reasonable in practice. This is due to the fact that the majority of privacy attacks on the web are undertaken by some automatic programs for sending targeted (spam) advertisements to a large amount of PWS-users. These programs rarely act as a real person that collects prolific information of a specific victim for a long time as the latter is much more costly.

If we consider the sensitivity of each sensitive topic as the cost of recovering it, the privacy risk can be defined as the total (probabilistic) sensitivity of the sensitive nodes, which the adversary can probably recover from G. For fairness among different users, we can normalize the privacy risk with which stands for the total wealth of the user. Our approach to privacy protection of personalized web search has to keep this privacy risk under control.

IV. PROPOSED SYSTEM

Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. However, these useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them. In the proposed system, we can implement the clustering algorithms for improving the better search quality results. It is retrieved by using the String Similarity Match Algorithm (SSM

Algorithm) algorithm. To address this privacy threat, current solutions propose new mechanisms that introduce a low cost in terms of computation and communication. In this paper we present a novel protocol specially designed to protect the users' privacy in front of web search profiling.

In this we propose and try to resist adversaries with broader background knowledge, such as richer relationship among topics. Richer relationship means we generalize the user profile results by using the background knowledge which is going to store in history. Through this we can hide the user search results. In the Existing System, Greedy IL and Greedy DP algorithm, it takes large computational and communication time.

Advantages

- It achieves better search results.
- It achieves the privacy results when applying the background knowledge to the user profiling results.
- It has less computational time and communicational time.
- It achieves better accuracy when compared with the Existing Works.

V. CONCLUSION AND FUTURE ENHANCEMENTS

Privacy protection in publishing transaction data is an important problem. This paper presented a client-side privacy protection framework called SSM for personalized web search. SSM could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, SSM also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. We proposed String Similarity Matching Algorithm, for the online generalization. Our experimental results revealed that SSM could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency of our solution.

Our proposed system gives better quality results and gives more efficiency. Privacy is too good when compared with the Existing system. In the Existing System, only generalization technique is used. Our String matching algorithm gives more accuracy when compared with the Greedy IL algorithm. Generalization and suppression technique achieves better privacy when compared with the existing system. In Future Work, we can implement the hierarchical divisive approach for retrieving the search results. It will gives better performance when compared with our proposed System.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 from the victim. We will also seek more sophisticated method to build the user profile, and better metrics to predict the performance (especially the utility) of UPS.

REFERENCES

[1] D. Fallows, Search engine users: internet searchers are confident, satisfied and trusting, but they are also unaware and naive, Pew/Internet & American Life Project (2005).
[2] D. Sullivan, comScore Media Metrix Search Engine Ratings, comScore, 2006. Available from: <http://searchenginewatch.com>.
[3] Google History, 2009. Available from: <http://www.google.com/history>.
[4] P. Agouris, J. Carswell, and A. Stefanidis, ``An environment for contentbased image retrieval from large spatial databases," ISPRS J. Photogram. Remote Sens., vol. 54, no. 4, pp. 263_272, 1999.
[5] M. Atallah and K. Frikken, ``Securely outsourcing linear algebra computations," in Proc. 5th ASIACCS, 2010, pp. 48_59.
[6] M. Atallah and J. Li, ``Secure outsourcing of sequence comparisons," Int. J. Inf. Security, vol. 4, no. 4, pp. 277_287, 2005.
[7] M. Atallah, K. Pantazopoulos, J. Rice, and E. Spafford, ``Secure outsourcing of scienti_c computations," Adv.
[8] Comput., vol. 54, pp. 216_272, Feb. 2001.
[9] D. Benjamin and M. Atallah, ``Private and cheating-free outsourcing of algebraic computations," in Proc. Conf. PST, 2008, pp. 240_245.
[10] E. Candès, ``The restricted isometry property and its implications for compressed sensing," Comptes Rendus Mathematique, vol. 346, nos. 9_10, pp. 589_592, 2008.
[11] E. Candès, J. Romberg, and T. Tao, ``Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Inf. Theory, vol. 52, no. 2, pp. 489_509, Feb. 2006.
[12] E. Candès and T. Tao, ``Decoding by linear programming," IEEE Trans. Inf. Theory, vol. 51, no. 12, pp. 4203_4215, Dec. 2005.
[13] E. Candès and T. Tao, ``Near-optimal signal recovery from random projections: Universal encoding strategies," IEEE Trans. Inf. Theory, vol. 52, no. 12, pp. 5406_5425, Dec. 2006.
[14] E. Candès and M. Wakin, ``An introduction to compressive sampling," IEEE Signal Proc. Mag., vol. 25, no. 2, pp. 21_30, Mar. 2008.
[15] (2009). Security Guidance for Critical Areas of Focus in
[16] Cloud Computing, [Online]. Available:http://www.cloudsecurityalliance.org
[17] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.
[18] 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.
[19] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.
[20] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.
[21] 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.
[22] 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.
[23] J. Castell´-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.
[24] A. Viejo and J. Castella-Roca, "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines," Computer Networks, vol. 54, no. 9, pp. 1343-1357, 2010.
[25] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[26] 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.