Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.399 – 406

Anonymizing Personalized Web Search using GreedyIL Algorithm

Snehal R. Kawalkar [1], Prof. P. L. Ramteke[2]
[12]Department of CS & IT, H.V.P.M. College of Engineering & Technology, Amravati University, India
[1] kawalkar.snehal@gmail.com; [2] pl_ramteke@rediffmail.com

Abstract— Personalized web search (PWS) is efficient in improving the search quality on internet. But users generally hesitate to disclose their personal information during search which has become a hindrance in wide increase in PWS. We model user preferences into user profiles. We proposed a PWS framework called CPS while performs profile generalization without affecting users' privacy requirements. This generalization maintains a balance between personalization and privacy risk while a generalized profile is exposed. GreedyIL algorithm improves the efficiency of the generalization using heuristics based on numerous answers. This paper models preference of users as hierarchical user profiles. It proposes a framework called UPS which generalizes profile at the same time maintaining privacy requirement specified by user. It has been found that UPS framework is one of the efficient techniques which guarantees the user privacy and retrieves the contents as per user requirement accurately

Keywords— Personalized web Search, Privacy protection, Search engine, Profile, GreedyIL

I. INTRODUCTION

This document is a template. An electronic copy can be downloaded from the conference website. For questions on paper guidelines, please contact the conference publications committee as indicated on the conference website. Information about final paper submission is available from the conference website.

The web search engine has gained a lot of popularity and importance for users seeking information on the web. However, users' experiences are sometimes bad when search engines return results that do not match with its needs. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.

The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward—they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well, it can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models 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.

Although there are pros and cons for both types of PWS techniques, the profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered implicitly 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.

Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

II. LITERATURE REVIEW

Previous works has focused on improving search result on profile- based PWS. Many representations for profile are available, some of them are term lists/vectors or bag of words to represent their profile while recent work create profile in hierarchical structure. The hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as Wikipedia or the hierarchical profile is generated via term-frequency analysis on the user data. UPS framework can adopt any hierarchical representation.

Two classes of privacy protection problems for PWS are identified. One class treats privacy as identification of individual. Other considers data sensitivity as the privacy. Typical literature works in for class one try to solve the privacy problem on different levels, which includes the pseudoidentity, the group identity, no identity, and no personal information. The first level solution is proved too fragile and the third and fourth levels are impractical because of high cost in communication and cryptography. Therefore, the existing efforts focus on the second level. Online anonymity for PWS provides anonymity by generating a group profile of k users. Using this approach, the relation between the query and a single user is broken. The useless user profile (UUP) protocol shuffle queries among a group of users who issue them. As a result no entity can profile a certain individual. The shortcoming of class one solution is the high cost.

In Class two solutions, users only trust themselves and don't tolerate the exposure of their complete profiles to anonymity server. Krause and Horvitz employ statistical techniques to learn a probabilistic model, and then use this model to generate the near-optimal partial profile. Privacy Enhancing personalized web search proposed a privacy protection solution for PWS based on hierarchical profiles. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted subtree of the complete profile.

The literature review covers the background, latest development of and related techniques for profile-based personalization and privacy protection in PWS system.

M. Spertta and S. Gach, systematically examined the issue of privacy preservation in personalized search. The four levels of privacy protection is distinguished, and analyze various software architectures for personalized search. This work showed that client-side personalization has advantages over the existing server-side personalized search services in preserving privacy, and envision possible future strategies to fully protect user privacy.

Y. Xu, K. Wang, G. Yang proposed the notion of online anonymity to enable users to issue personalized queries to an un-trusted web service while with their anonymity preserved. The challenge for providing online anonymity is dealing with unknown and dynamic web users who can get online and offline at any time. Introduces the notion of online anonymity to ensure that each query entry in the query log cannot be linked to its sender and an algorithm that achieves online anonymity through the user pool is proposed. This approach can be extended to deal with personally identifying information that may be contained in the query. The method is also applicable to general web services where there is a need to anonymize the query, with or without personalization.

In J. Castelli-Roca, A. Viejo and J. Herrera presents a novel protocol Useless User Profile (UUP) protocol, specially designed to protect the users' privacy in front of web search profiling. System provides a distorted user profile to the web search engine. Also offers implementation details, computational and communication results that show that the proposed protocol improves the existing solutions in terms of query delay. The protocol also provides an affordable overhead while offering privacy benefits to the users.

In X. Xiao and Y. Tao**,** presented a new generalization framework based on the concept of personalized anonymity. This technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata.

J. Teevan, S.T. Dumais, and D.J. Liebling, examines variability in user intent using both explicit relevance judgments and large-scale log analysis of user behavior patterns. They characterize queries using a variety of features of the query, the results returned for the query, and people's interaction history with the query. Using these features, the authors build predictive models to identify queries that can benefit from personalization.

X. Shen, B. Tan, and C. Zhai, Information retrieval systems (e.g., web search engines) are critical for overcoming information overload. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users, resulting in inherently non-optimal retrieval performance.

In 2007, Z. Dou proposed Average Precision metric, to measure the effectiveness of the personalization in UPS.

Susan T. Dumais introduces a search algorithm that considers user's prior interactions with a wide variety of content, to personalize their current web search. Rather than relying on the unrealistic assumption that people will precisely specify their intent when searching, it pursues techniques that leverage implicit information about the user's interests.

Lidan Shou and Gang Chen et al, uses hierarchical user structure for modeling user interests. The system provides generalization of user profile with use of an online profiler at the client side.

III. PROPOSED SYSTEM

The proposed system seems to be more effective for privacy protection. An online profiler is designed in this system, which can adaptively generalize profiles by queries while respecting user specified privacy requirements. The online profiler is at the client side where the complete user profile is stored along with the specified sensitive topics. Runtime generalization aims at providing search efficiency along with privacy protection of user profiles. Online generalization avoids unnecessary privacy disclosure and also removes topics irrelevant to the current query. Overgeneralization causes ambiguity in personalization, leading to poor search results. In this section, the procedures carried out for each user during two different execution phases, namely the offline and online phases. Generally, the offline phase constructs the original user profile and then performs privacy requirement customization according to user-specified topic sensitivity. The subsequent online phase finds the Optimal -Risk Generalization solution in the search space determined by the customized user profile. The online generalization procedure is guided by the global risk and utility metrics. The computation of these metrics relies on two intermediate data structures, namely a cost layer and a preference layer defined on the user profile.

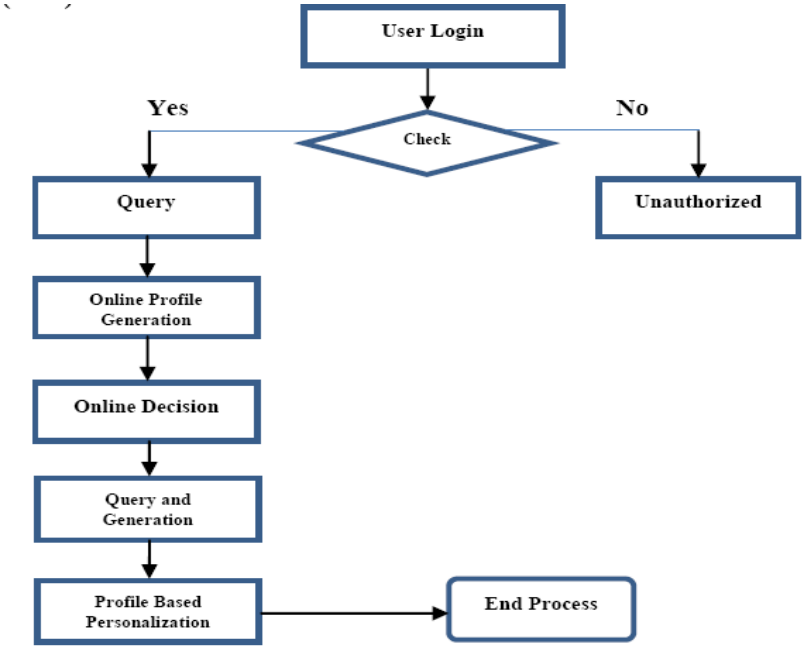

Fig. 1 dataflow diagram

Advantages:

1. It enhances the stability of the search quality.
2. It avoids the unnecessary exposure of the user profile.
3. The framework allowed users to specify customized privacy requirements via the hierarchic al profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality.

The Proposed system consists of four modules:

- Profile-Based Personalization
- Generating User Profile
- Online Decision
- Privacy Protection in PWS system

A. Profile-Based Personalization

Personalization is the process of presenting the right information to the right user at the correct instant. In order to study on a user, systems must gather personal data, investigate it, and accumulate the consequences of the analysis in a user profile. Data can be composed from users in two traditions: unambiguously, for instance ask for comment such as preferences or ratings; or perfectly, for instance detect user behaviors such as the time spent reading an on-line document. The accessible profile-based PWS do not hold runtime profiling. A user profile is usually inclusive for only one time offline, and utilized to personalize all query from a similar user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries. The existing methods do not take into account the customization of privacy requirements. This possibly creates several user privacy to be overprotected while others insufficiently protected. For example, all the sensitive topics are detected using an absolute metric called surprised based on the information theory, supposing that the interests with less user document support are more sensitive.

We propose a method to personalize the search results of a user based on their profile information. The two main mechanisms used for this purpose are: a profile generator that is used to create the profile of the user based on the inputs and preferences given by the user and an algorithm that ranks the search results based on the preferences and interests of the user.

B. Generalizing User Profiles

Generalizing the profile of the user is the most important step in the entire framework. Since our proposal focuses mainly on the privacy requirements of the user, the generalized profile of the user has a very important role to play in the entire process. The process of generalization is done by using a parent profile and an inherited profile. The parent profile is the original profile of the user that contains all the details of the user. The inherited profile contains the profile with the necessary privacy requirements of the user.

C. Online Decision

The personalization of web search leads to lot of unwanted information of the user being shown to the server, with not much improvement in the quality of the search results. This puts the privacy of the user in risk. To make sure that the user gets relevant and efficient results even for distinct queries, we propose an online decision method. Here, it is decided if a query should be personalized or not. If the query is distinct, i.e., it is very different and not much related to the preferences of the user, the profile that is created during runtime (the generalized profile) of the user is discarded and the query is sent to the server without a user profile.

D. Privacy Protection in PWS system

A PWS framework called UPS that generalize profiles for each query according to user-specified privacy requirements. Personalized web search is picking up more prevalence. However keeping up privacy is not kidding issue in personalized web search. As personalizing search obliges assembling and preparing of user information, which prompts privacy issue. This is turning into the principle obstruction in conveying personalized web search applications. Personalized obscurity is a security procedure which is executed to give privacy in personalized web search in which individual can determine level of privacy. Anonymizing user profile is additionally method by which privacy of user can be kept up.

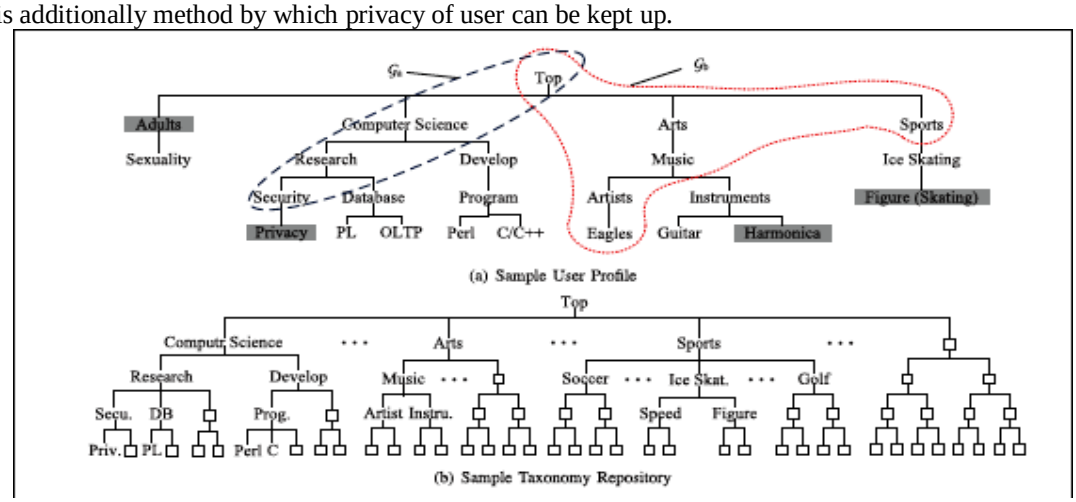

Fig. 2 Taxonomy Based user profile

IV. THE GREEDYIL ALGORITHM

The Greedy IL algorithm is used in the implementation concept to provide generalization of queries by which the generalized user profile is generated. The Greedy algorithm enhance the efficiency of generalization using heuristics on several findings. Another important findings is that prune leaf operation reduces the discriminating power of the profile.

If G' is a profile obtained by applying a prune-leaf operation on G, then $DP(q,\mathcal{G}) \geq DP(q,\mathcal{G}')$ Considering operation $\mathcal{G}_i \xrightarrow{-t} \mathcal{G}_{i+1}$ in the ith iteration, maximizing $DP(q,\mathcal{G}_{i+1})$ is equivalent to minimizing the incurred information loss, which is defined as $DP(q,\mathcal{G}_i) - DP(q,\mathcal{G}_{i+1})$

Symbol	Description
DP	Discriminating power
Gi	Generalized profile
q	Query

GET PROFILE (Greedy IL)
INPUT Query Q, Privacy Threshold δ
OUTPUT Profile P
1.Get Query Q.
2.Privacy Threshold δ [0-1]
3.Split query into words QW.
4.Find Seed Profile G. (Any category level contains the category name inside the query words).
5.CID ={}.
6.For i=1 to QW.Count
  a. Find Category Ids which contains QW[i] in their category names.
  b.Add Category Ids to CID
7.Next
8.If CID.Count>0
9.Create Profile P.
10.For j=1 to CID.Count
  a.While RiskFactor(Q,CID[j]) > δ
    i.P.CategoryId= CID[j]
    ii.P.CategoryName = Category Name of CID[j]
  a.End While
11.Next
12.Return P

V. RESULT AND ANALYSIS

Different users have different requirements of privacy protection. While some users may not want anyone else to know or hold any of their personal information, others may be willing to share some personal information for better search results or services. Thus the level of privacy protection may need to be tuned for Different users to accommodate Different preferences for the tradeoff of personalization and privacy protection. For that proposed system provide protection to privacy nodes and its url and non privacy query and url will maintain as per the user profile. The user can have better search results to access the previous results on the basis of searched urls. Proposed work shows session mechanism generated at login time so admin will know on which session perticular user login.

In fig. 3, shows the search page provided to each user where he can search his relevant information. When user enter a query and click on search button, system will ask whether to provide privacy or not. If yes

then content of that Query and clicked url will display on url's clicked privacy list box and that content will hide from other and visible to that particular user. If no then content of that Query and clicked url will display on advisor search list box shown in fig.4. In both cases previous search url will display of that query depending on particular user. Whatever the search has been done based on the profile display by topic shows in fig.5.

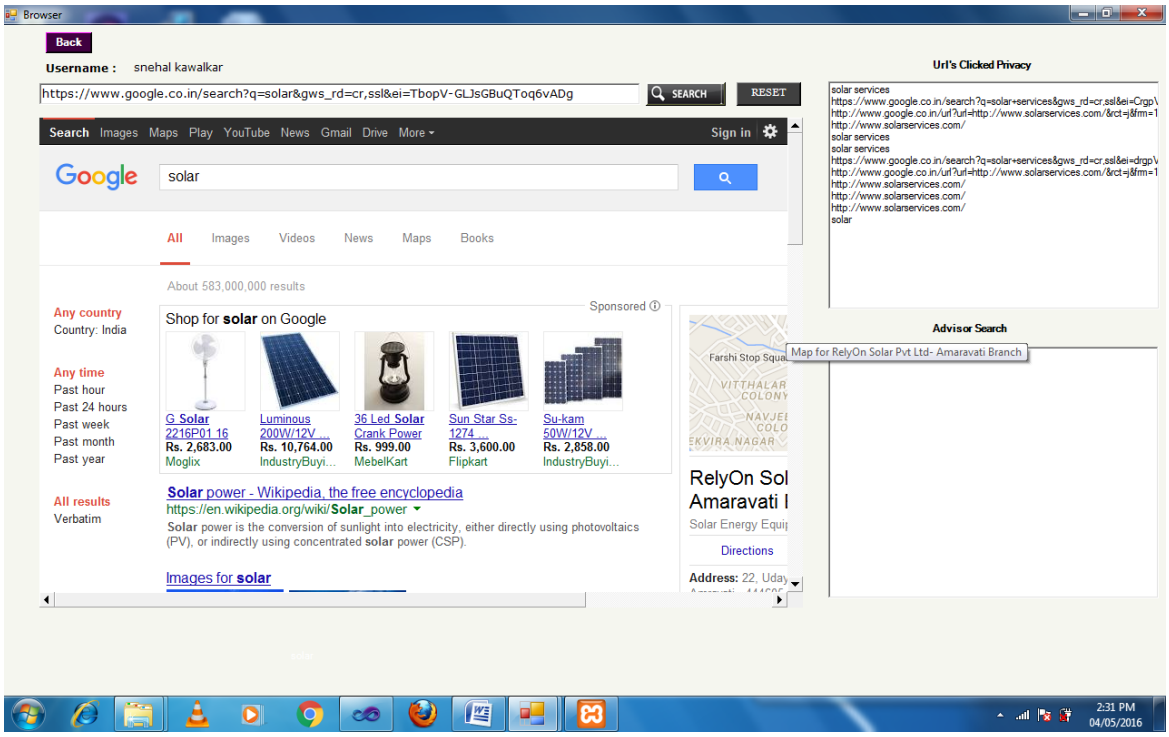

Fig. 3 search query by making privacy

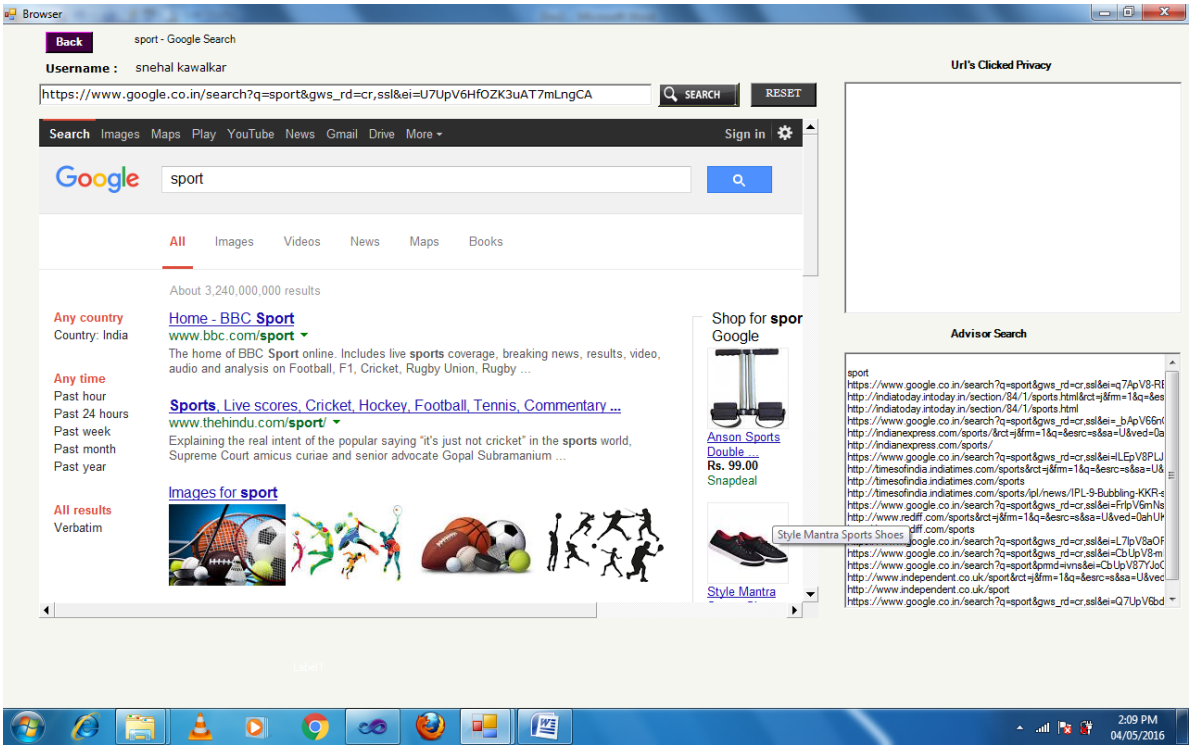

Fig. 4 search query by making no privacy

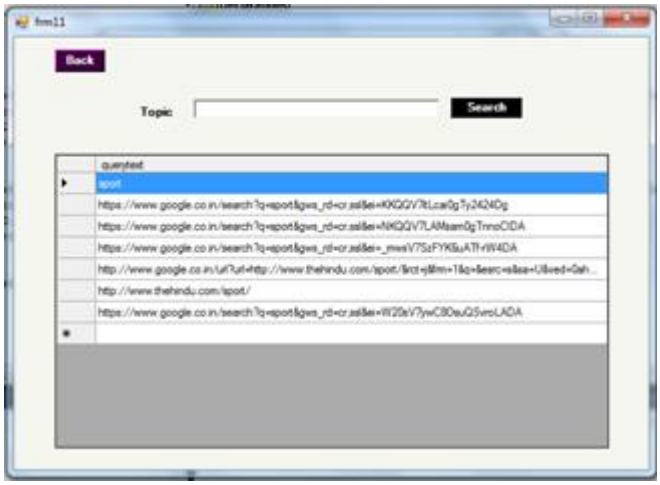

Fig. 5 Profile based search

VI. CONCLUSIONS AND FUTURE WORK

This paper 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. GreedyIL algorithm improves the efficiency of the generalization using heuristics based on numerous answers.

We proposed an inexpensive mechanism for the client to decide whether to personalize a query in UPS. This decision can be made before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile. Our proposed framework provided customized privacy requirements via the hierarchical profiles to the users. Through this profile, users control what portion of their private information is exposed to the server and the users can specify to which degree the content should be protected.

For future work, we will try to resist adversaries with broader background knowledge, such as richer relationships 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 methods to build the user profile, and better metrics predict the performance (especially the utility) of the UPS.

ACKNOWLEDGEMENT

I would like to thanks my Guide Prof. P. L. Ramteke and Principal Dr. A. B. Marathe, who provided me constructive and positive feedback during the preparation of this paper.

REFERENCES

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.
[3] Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010
[4] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," ACM SIGIR Forum, vol. 41, no. 1, pp.4-17, 2007.
[5] Y. Xu, K. Wang, G. Yang, and A.W.C Fu, "Online anonymity for personalized web services" Proc.18th ACM conformation and knowledge management (CIKM), pp 1497-1500, 2009.
[6] 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.
[7] Lidan Shou, He Bai, Ke Chen, and Gang Chen "Supporting privacy protection in personalized web search" IEEE Transactions on knowledge and data engineering, Vol. 26, No. 2, February 2014.
[8] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] S.Vanitha "A Personalized Web Search based on user profile and user clicks" International Journal of Latest Research in Science and Technology ISSN (Online):2278-5299 Volume 2, Issue 5: Page No.78-82,September-October 2013

[15] 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.

[16] B.Upender and Bathula Revathi, "Analysis on Supporting Privacy Protection in Personalized Web Search " INDIA / International Journal of Research and Computational Technology, Vol.7 Issue.2, ISSN: 0975-5662, April, 2015.

