iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** III **Month of publication:** March 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Privacy Protection in Personalized Web Search

Anuja Agnihotri[1], Jyoti Kale[2], Priyanka Patil[3], Prof. Sheetal Thakare[4]
[1]*BE, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai-400614*
[2]*Assistant Professor,Department of Computer Engineering, Bharati Vidyapeeth College of Engineering,Navi Mumbai-400614*

Abstract — Since the content in Internet is growing rapidly, the search provider users demand accurate search result as per their need. One of the options available to users is personalized web search which presents search result based on the personal data of user provided to the search provider. However, users' unwillingness to share their private information during search has become the major barrier for personalized web search. 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. Two greedy algorithms namely GreedyDP and GreedyIL are used for runtime generalization. Also, an online prediction mechanism to decide whether to personalize a query or not is provided in this paper.

I. INTRODUCTION

The web search engine has gained a lot of popularity and importance for users seeking information on the web. Since the contents available in web is very vast and ambiguous, users at times experience failure when an irrelevant result of user query is returned from the search engine. Therefore, in order to provide better search result a general category of search technique Personalized Web search is used. For a given query, a personalized web search can Provide different search results for different users or Organize search results differently for each user, based upon their interests, preferences, and information needs. Personalized web search differs from generic web search, which returns identical research results to all users for identical queries, regardless of varied user interests and information needs personalized search refers to search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided. In personalized web search, user information is collected and analyzed in order to find intention behind issued query fired by user. There are two categories of PWS, namely click-log-based and profile-based. The click-log based methods are straightforward— they simply impose bias to clicked pages in the user's query history. This strategy has been performing well but it work on repeated queries from same user which is a strong limitation to its applicability. While profile-based methods improve the search experience generated from user profiling techniques. Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances. There are both advantages and disadvantages for both type of PWS technique, profile based PWS is more effective for improving search result. The user profile is made from information gathered from query history, browsing history, click-through data bookmarks, user documents and so forth. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life.

II. RELATED WORKS

In this section, we overview the related works. We focus on the literature of profile-based personalization and privacy protection in PWS system. Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. We review the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization. Two classes of privacy protection problems for PWS is identified. One class treats privacy as identification of individual. Other considers data sensitivity as the privacy. This paper provides personalized privacy protection in PWS. A person can specify the degree of privacy protection for her/his sensitive values by specifying "guarding nodes" in the taxonomy of the sensitive attribute. Thus, this paper allows user to customize privacy requirements in hierarchical user profiles.

A. User Profiling

To provide personalized search results to users, personalized web search maintains a user profile for each individual. A user profile stores approximations of user tastes, interests and preferences. It is generated and updated by exploiting user-related information. Such information may include:
Demographic and geographical information, including age, gender, education, language, country, address, interest areas, and

other information; Search history, including previous queries and clicked documents. User browsing behaviour when viewing a page, such as dwelling time, mouse click, mouse movement, scrolling, printing, and bookmarking, is another important element of user interest. Other user documents, such as bookmarks, favourite web sites, visited pages, and emails.

B. Server-Side And Client-Side Implement

Personalized web search can be implemented on either server side (in the search engine) or client side (in the user's computer or a personalization agent).For server-side personalization, user profiles are built, updated, and stored on the search engine side. User information is directly incorporated into the ranking process, or is used to help process initial search results. The advantage of this architecture is that the search engine can use all of its resources, for example link structure of the whole web, in its personalization algorithm. Also, the personalization algorithm can be easily adapted without any client efforts. This architecture is adopted by some general search engines such as Google Personalized Search. The disadvantage of this architecture is that it brings high storage and computation costs when millions of users are using the search engine, and it also raises privacy concerns when information about users is stored on the server For client-side personalization, user information is collected and stored on the client side (in the user's computer or a personalization agent), usually by installing a client software or plug-in on a user's computer. In client side, not only the user's search behaviour but also his contextual activities (e.g., web pages viewed before) and personal information (e.g., emails, documents, and bookmarks) could be incorporated into the user profile. This allows the construction of a much richer user model for personalization. Privacy concerns are also reduced since the user profile is strictly stored and used on the client side. Another benefit is that the overhead in computation and storage for personalization can be distributed among the clients. A main drawback of personalization on the client side is that the personalization algorithm cannot use some knowledge that is only available on the server side (e.g., Page Rank score of a result document). Furthermore, due to the limits of network bandwidth, the client can usually only process limited top results.

C. Challenges Of Personalized Search

Privacy is an issue. Personalized web search, especially server-side implement, requires collecting and aggregating a lot of user information including query and click through history. A user profile can reveal a large amount of private user information, such as hobbies, vocation, income level, and political inclination, which is clearly a serious concern for users. This could make many people nervous and feel afraid to use personalized search engines. A personalized web search will be not well received until it handles the privacy problem well.

It is really hard to infer user information needs accurately. Users are not static. They may randomly Search for something which they are not interested in. They even search for other people sometimes. User search histories inevitably contain noise that is irrelevant or even harmful to current search. This may make personalization strategies unstable.

Queries should not be handled in the same manner with regard to personalization. Personalized search may have little effect on some queries. Some work investigates whether current web search ranking might be sufficient for clear/unambiguous queries and thus personalization is unnecessary. Personalized search has little effect on queries with high user selection consistency. A specific personalized search also has different effectiveness for different queries. It even hurts search accuracy under some situations. For example, topical interest-based personalization which leads to better performance for the query ''mouse,'' is ineffective for the query ''free mp3 download.'' Actually, relevant documents for query ''free mp3 download'' are mostly classified into the same topic categories and topical interest-based personalization has no way to filter out desired documents. Topical interest-based personalized search methods are difficult to deploy in a real world search engine. They improve search performance for some queries, but they may hurt search performance for additional queries.

III. EXISTING SYSTEM

The existing profile-based Personalized Web Search do not support runtime profiling. A user profile is typically generalized for only 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 evidence reported in is that profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk. The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. For example, in, all the sensitive topics are detected using an absolute metric called surprisal based on the information theory, assuming that the interests with less user document support are more sensitive. However, this assumption can be doubted with a simple counterexample: If a user has a large number of documents about "sex," the surprisal of this topic may lead to a conclusion

that "sex" is very general and not sensitive, despite the truth which is opposite. Unfortunately, few prior works can effectively address individual privacy needs during the generalization. Many personalization techniques require iterative user interactions when creating personalized search results. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on. This paradigm is, however, infeasible for runtime profiling, as it will not only pose too much risk of privacy breach, but also demand prohibitive processing time for profiling. Thus, we need predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

A. *Disadvantages*

All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.
The existing profile-based PWS do not support runtime profiling.
The existing methods do not take into account the customization of privacy requirements.
Personalization techniques require iterative user interactions when creating personalized search results.

IV. PROPOSED SYSTEM

We propose a privacy-preserving personalized web search framework UPS (User customizable privacy preserving search, which can generalize profiles for each query according to user-specified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, we formulate the problem of privacy-preserving personalized search as Risk Profile Generalization, with its NP-hardness proved. We develop two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. While the former tries to maximize the discriminating power (DP), the latter attempts to minimize the information loss (IL). By exploiting a number of heuristics, GreedyIL outperforms GreedyDP significantly. We provide 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.

A. *Advantages*

It enhances the stability of the search quality.
It avoids the unnecessary exposure of the user profile.
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.

V. METHODOLOGY

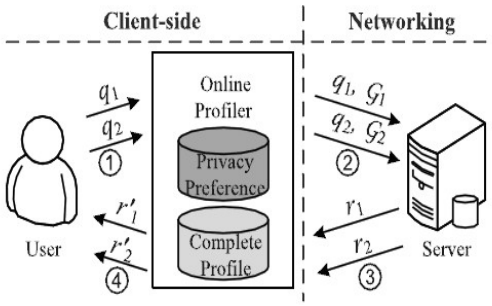


Fig. 1

UPS consists of a nontrusty search engine server and a number of clients. Each client (user) accessing the search service trusts no one but himself/ herself. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified (customized) privacy requirements represented as a set of sensitive-nodes. The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile

is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows: When a user issues a query qi on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile Gi satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search. The search results are personalized with the profile and delivered back to the query proxy. Finally, the proxy either presents the raw results to the user, or re-ranks them with the complete user profile. UPS is distinguished from conventional PWS in that it provides runtime profiling, which in effect optimizes the personalization utility while respecting user's privacy requirements; allows for customization of privacy needs; and 3) does not require iterative user interaction.

VI. GREEDY ALGORITHM

A greedy algorithm is an algorithm that follows the problem solving heuristic of making the locally optimal choice at each stage with the hope of finding a global optimum. Greedy algorithm considers easy to implement and simple approach and decides next step that provide beneficial result. In many problems, a greedy strategy does not produce an optimal solution, but a greedy heuristic yields locally optimal solutions that approximate a global optimal solution in a reasonable time. This collection should be from several and different kind of sources so that it can be made more robust in training.

A. Greedydp Algorithm

It works in a bottom up manner. Starting with the leaf node, for every iteration, it chooses leaf topic for pruning thus trying to maximize utility of output. During iteration a best profile-so-far is maintained satisfying the Risk constraint. The iteration stops when the root topic is reached. The best profile-so-far is the final result. GreedyDp algorithms require recomputation of profiles which adds up to computational cost and memory requirement.

B. Greedyil Algorithm

GreedyIL algorithm improves generalization efficiency. GreedyIL maintains priority queue for candidate prune leaf operator in descending order. This decreases the computational cost. GreedyIL states to terminate the iteration when Risk is satisfied or when there is a single leaf left. Since, there is less computational cost compared to GreedyDP, GreedyIL outperforms GreedyDP.

C. Calculating Probabilities

Probabilities of word occurring in spam and genuine emails are calculated. Then spam probabilities of words are calculated

VII. CONCLUSION

A client side privacy protection framework called UPS i.e. User customizable Privacy preserving Search is presented in the paper. Any PWS can adapt UPS for creating user profile in hierarchical taxonomy. UPS allows user to specify the privacy requirement and thus the personal information of user profile is kept private without compromising the search quality.

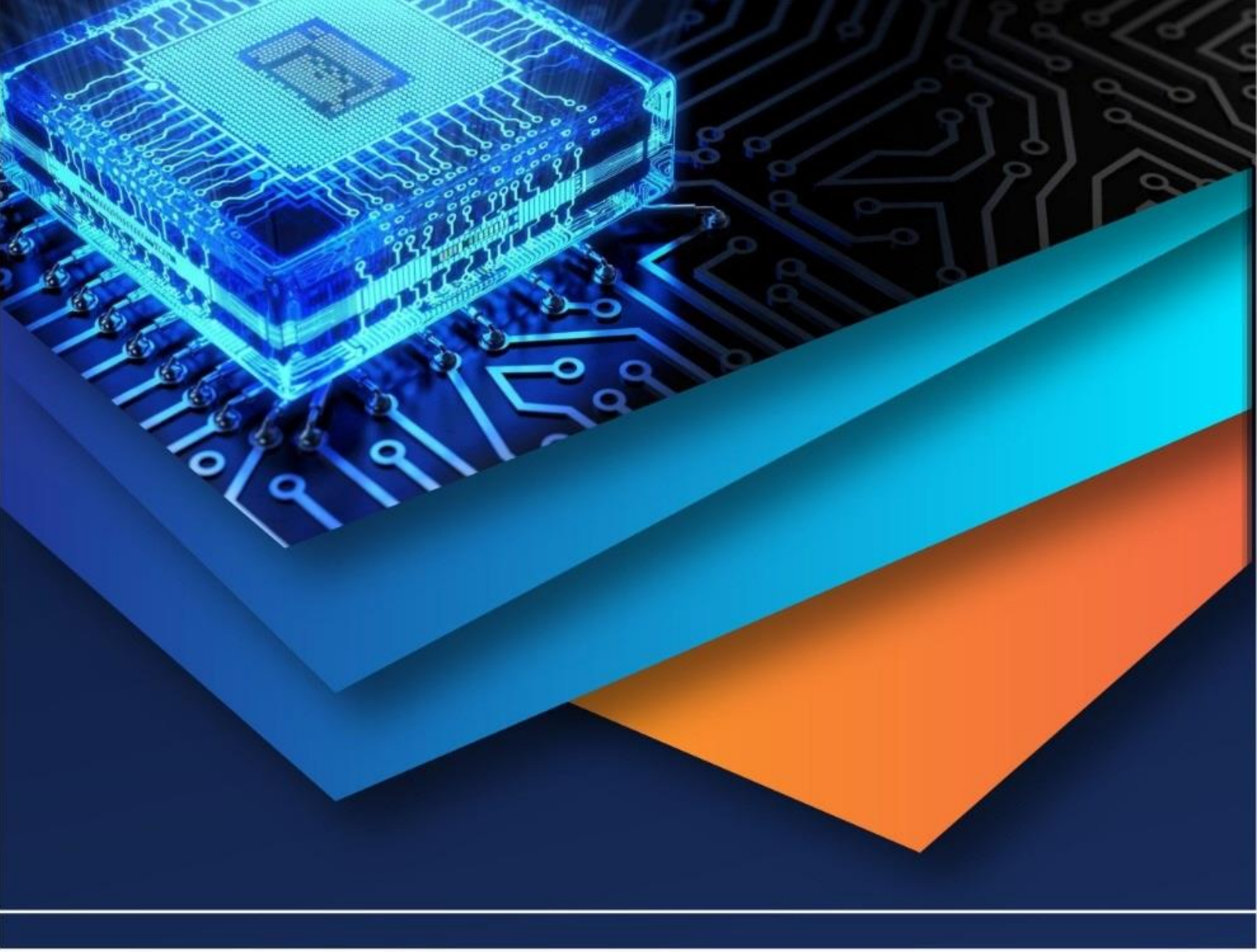
VIII. ACKNOWLEDGEMENT

I would like to take this opportunity to express my gratitude towards all the people who have in various ways, helped in the successful completion of my project. This work was influenced by countless individuals whom we were fortunate enough to meet during our project duration. I must convey my gratitude to Prof. Sheetal Thakare and Seminar coordinator Prof. Rahul Patil for giving me the constant source of inspiration and help in preparing the project, personally correcting my work and providing encouragement throughout the project. We would like to thank our H.O.D. Prof. D.R.Ingle and our respected principle Dr. M.Z. Shaikh. We express our thanks to senior friends for extending their support. I also thank all my faculty members for steering me through the tough as well as easy phases of the project in a result oriented manner with concern attention

REFERENCES

[1] 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.

[2] 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.
[3] M. Spertta and S. Gach, "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8]K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.
[9] 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.
[10] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.
[11] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.
[12] 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.
[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] 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.
[15] A. Viejo and J. Castell_a-Roca, "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines," Computer Networks, vol. 54, no. 9, pp. 1343-1357, 2010.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY