PROFILE BASED PERSONALISED WEB SEARCH

[1]BHAMINI U, [2]SALMAN MOHAMMED, [3]VARSHA LAKSHMAN, [4]D. ASHOK

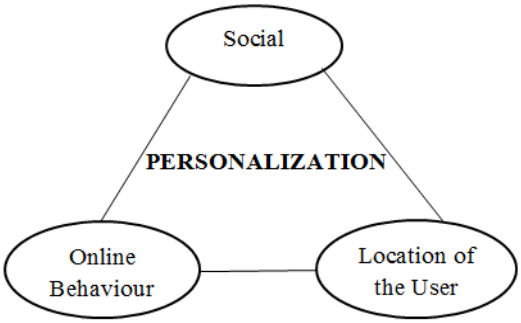
[1,2,3,4] Computer Science Department SRM University, Chennai, India

Abstract- Personalised Web Search is a search technique that is used for improving the relevance of the search results of various search engines in the internet. This may also leave the user experience disappointment when relevant results of proper quality are not obtained. The reason behind this may be due to the huge contexts and backgrounds that the query may belong to. This problem can be overcome only when the user shares quite a lot of his/her personal information. However, they might be reluctant to do so. Hence, we propose a framework where the overall user search experience is improved by enhancing the quality of the search results, keeping the privacy of the user also in mind. This is achieved by adaptively generalizing profiles and thereby maintaining the privacy requirements of the user. We also propose a mechanism for runtime profiling and a method of online decision to decide whether a user search query can be personalized or not.

Keywords- profile, personalized web search, privacy, dynamic profiling, encrypting, greedy algorithm

I. INTRODUCTION

The web search engine now holds a very important part in all our day-to-day lives. Each and every one of us use web search engines to retrieve both useful and sometimes unwanted information from the web. The world wide web is home to huge amounts of data that vary across all genres. It is due to this reason that users sometimes experience failure when the set of results obtained for their search query is absolutely irrelevant with respect to what they expected to see. Therefore, as a remedy, the concept called 'Personalized Web Search' is used. In personalized web search, information about the user is collected. This information is used to predict the overall online behavior of the user and the user's search results are filtered based on the findings. Search personalization can broadly be classified into two categories namely context oriented and individual oriented [1]. Context oriented PWS includes factors such as the nature of information available, the information that is being used at that point of time, the applications used by the user and so on. Whereas, individual oriented PWS includes other factors such as the goals of the user, prior and tacit knowledge about the user and his interests and so on. The factors that affect PWS can be broadly divided into two categories namely spatial factors and temporal factors [1]. Spatial factors deal with the search queries that were previously used by the user, pages visited, browsing behavior and so on while temporal factors deal with user's shift in interests and so on.



This leaves us with two basic categories of personalization – The click-log based methods and the Profile based methods. Click-log based methods are straightforward and personalize results based on spatial factors whereas Profile based methods personalize results based on temporal factors. Click-log based methods are efficient only when repeated queries are obtained from the same user. This is a big loophole in its applicability [2]. This problem doesn't exist in Profile-based PWS. It aims to improve the user search experience based on the information gathered from the profile of the user. It can be effective for queries of all sorts, but has its own disadvantages. The user profile can be generated from their query history [3], [4], [5], browsing history [6], [7], previous clicks [8], [9], [2], bookmarks of the user [10], and their documents [3], [11], and so on. Such data reveal much of a user's private life and hence breaches the security and privacy of the user.

There are many searches that use Profile-based Personalization. They do not support runtime profiling [11], [12]. The profile of a user is generalized offline, and once generalized, the same profile is used to personalize all the queries from that user. This technique has a lot of drawbacks and doesn't improve the quality of the search. They also do not take the user's privacy into account, revealing much of the user's private life. The existing system also assumes that certain topics with lesser user documents are more sensitive to the user than other topics that have more documents based on them. This method is called surprisal and can be doubted using various examples. It doesn't work in all scenarios. All the above stated problems can be overcome only by having repeated interactions with the user, thereby enhancing the quality of results of the user with each query. Various metrics like ranking of webpages, average rank for a webpage etc. are used to determine the sensitivity of the user to a certain topic. However, this is not advisable because it reveals too much of private information of the user and can cause various threats. It also requires a huge time for processing the profile of the user each time the user engages in a

new search query. In short, the disadvantages of the existing system of personalization can be summed up as follows:

- Existing profile-based PWS methods do not support runtime profiling.
- There is a big breach in privacy of the user and their privacy requirements are not taken into concern.
- The techniques require iterative interactions with the user which may involve a lot of processing, adding to the inefficiency of the whole system.
- The topics of sensitivity to the user are detected using old user documents and need not be correct everytime.

II. PROPOSED SYSTE

In this paper, we propose a system that preserves the privacy of the user while they get personalized results for their search queries on the web. It is based on a framework that generalizes the profile of the user for every distinct query based on the privacy specifications mentioned by the user. The generalization is done using two algorithms namely DisPower and InfoLoss. They are both greedy algorithms that support runtime profiling. DisPower is used to maximize the Discriminating Power that is used to distinguish between two similar search results. The more the discriminating power is, the more relevant the results will be. On the other hand, InfoLoss algorithm is used to minimize the Information Loss that is involved when a profile is generalized and results are produced. The architecture of the system that is proposed in this paper is as follows.

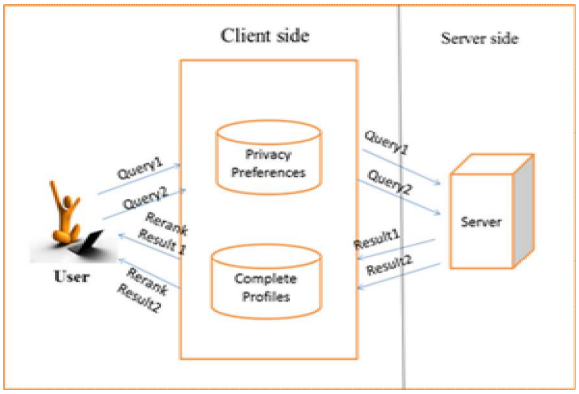


Fig. 2.1: Proposed System Architecture

As shown in Fig. 2.1, the framework proposed in this paper consists of a user submitting his/her search query to the server that contains the database of results that fulfills the user's request. Here, we use an interface called the online profiler that is used to generalize the profile of the user. The details of the user are not explicitly shown to the server. The privacy requirements of the user are implemented and the privacy-tailored profile of the user is only sent to the server. In the user's device, the online profiler is implemented in the form of a search proxy that maintains the profile of the user in the form of a graph which hierarchical nodes. The graph contains the privacy specifications of the user in the form of sensitive nodes. The online profiler is used to cater to the privacy requirements of the user and to preclude from showing the complete profile of the user. The server receives the query with the online profile of the user (with privacy requirements). This works in two phases namely online and offline. In the offline phase, the hierarchical graph is created using the inputs provided by the user. The privacy specified by the user is also marked in it. The online phase is used when the user inputs a search query. In this phase, the proxy user profile is created and the search query is forwarded to the server. The output of this phase is the generalized user profile that is created after catering to the privacy needs of the user. The search result is provided by the server, which is personalized and sent back to the proxy profile. The proxy then filters the results based on the profile of the user and his/her interests. This way, personalized web search is implemented keeping in mind the privacy requirements of an average user. The user is also given the authority to customize their privacy requriements according to their needs.

The Proposed system consists of four modules:

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

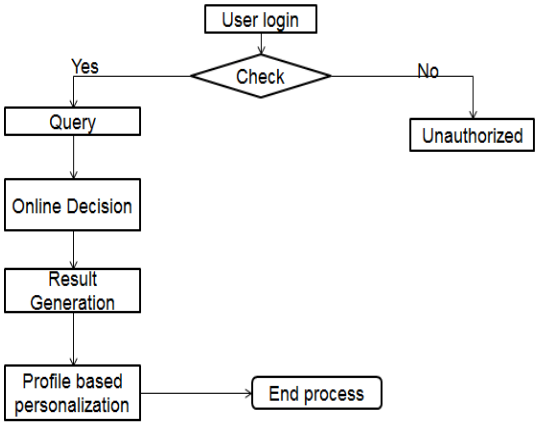
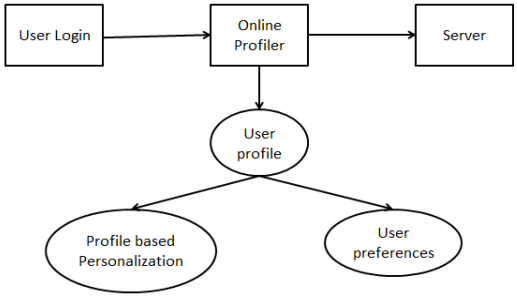


Fig. 2.2 : System flow diagram

3.1 Profile-Based Personalization



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.

3.2 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. The profile of the user is stored in the form of a graph that has a hierarchical structure. There are root nodes and child nodes for specific nodes. If a parent node is in privacy protection, all this children nodes are automatically privacy protected by the reverse need not be true. The privacy requirements specified by the user are indicated in the graph using a marking system. The inherited profile has all the attributes of the parent profile with additional attributes that are not present in the parent profile. If there is an update in the user profile, the change is propagated in the generalization process and the graph is modified accordingly. This process of generalization occurs when the user inputs a search query and the complete profile of the user should not be shown to the server. Here, the generalized profile with the privacy requirements of the user is only shown to the server. Hence, the privacy of the user is maintained.

3.3. 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.

3.4 Privacy Protection in PWS system

As mentioned before, the very essence of the proposal of this paper is the privacy of the user. The profile of the user is generalized for each query given by the user according to the privacy specifications of the user. We use encryption algorithms to protect the privacy of the user. For example, if a search result directs to a file that has to be downloaded by the user, they cannot do so unless they are an authenticated user. The server cannot view the file downloaded by the user as it is in an encrypted format. We use two encryption algorithms namely MD5 and Triple DES for this purpose. The file, when downloaded is automatically encrypted and saved.

When an authenticated user tries to access this file, it gets decrypted and the original contents of the file are shown to the user. This method has proven to be very effective against intruders and also protecting the privacy of the user because the files downloaded by the user can reveal a lot of the private information of the user.

III. ALGORITHMS USED

In this paper, we have used two different greedy algorithms for generalizing the user profile and encryption algorithms for file encryption. Greedy algorithm is a paradigm for problem solving that uses a heuristic where the optimal choice is chosen at each stage to find a globally optimal solution [14]. Greedy algorithm can be easily implemented and decides the next course of action that provides an optimum solution. A greedy heuristic helps to determine the locally optimal solution at various points thus helping us derive the globally optimum solution.

3.1 DisPower Algorithm

This algorithm follows a bottom up approach. In every iteration, it starts with the leaf node and then moves up the graph. Pruning of the graph is done from the leaf nodes, so as to maximize the use of the output as much as possible. The final iteration is that of the root node. The best profile that is obtained in the process is taken as the optimal solution. In this algorithm, recomputation of profiles is done to see the computational cost and memory requirement so as to choose the best profile among all.

3.2 InfoLoss Algorithm

This algorithm is basically used to improve the efficiency of the process of generalization. This algorithm maintains a queue on the basis of priority. It is used to prune leaf nodes in descending order. This basically cuts down the computational cost of every profile. This algorithm stops working, i.e., has the final iteration when only a single leaf node is left. Experiments have shown that the computational cost involved in InfoLoss algorithm is much lesser than that of DisPower algorithm and hence is proved to be better than DisPower algorithm for the purpose of profile generalization [13].

3.3 Encryption Algorithms

Encryption algorithms are used for the purpose of file encryption. When a user wants to download a file from any search result, they can do so only if they are an authenticated user. The algorithms MD5 and Triple DES are used for the purpose of encryption and the server cannot see the contents of the file due to this reason. This maintains the privacy of the user and thus much of his private information is not known to the server. MD5 and Triple DES algorithms have been chosen for this purpose because they are the strongest encryption algorithms and cannot be broken easily [15], [16].

CONCLUSION

A client side personalized web search that caters to the privacy needs of the user was proposed in this paper. Any personalized web search can adapt to this framework. This is helpful and advantageous because efficient search results are provided to the user based on their preferences in their profile. Also, their privacy needs are maintained and they can also be customized according to the varying needs of the user. We have used two algorithms namely DisPower and InfoLoss for online profile generalization and encryption algorithms for file encryption.

ACKNOWLEDGMENT

The authors would like to thank our Head of the department Mrs. Padmavathi in giving her useful comments and suggestions. We would also like to thank our guide Mr. D. Ashok for chaperoning us through various phases of the research and giving our ideas proper shape.

REFERENCES

[1] http://www.tmrfindia.org/eseries/ebookv2-c5.pdf
[2] 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.
[3] J. Teevan, S.T. Dumais, and E. Horvitz, “Personalizing Search via Automated Analysis of Interests and Activities,” Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), pp. 449-456, 2005.
[4] M. Spertta and S. Gach, “Personalizing Search Based on User Search Histories,” Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence (WI), 2005.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] A. Krause and E. Horvitz, “A Utility-Theoretic Approach to Privacy in Online Services,” J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.
[13] LidanShou, He Bai, Ke Chen, and Gang Chen, “Supporting Privacy Protection in Personalized Web Search,” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, pp. 453-468, 2014.
[14] http://en.wikipedia.org/wiki/Greedy_algorithm
[15] http://en.wikipedia.org/wiki/MD5
[16] http://en.wikipedia.org/wiki/Triple_DES

★★★