ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Supporting Privacy Protection in Personalized Web Search

Kamatam Amala
P.G. Scholar (M. Tech),
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

K.Rajasekhar Reddy
Assistant Professor,
Department of CSE,
Srinivasa Institute of Technology & Sciences,
Ukkayapalli, Kadapa, Andhra Pradesh.

ABSTRACT:

Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user-specified privacy requirements. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that GreedyIL significantly outperforms GreedyDP in terms of efficiency.

Index Terms:

Privacy protection, personalized web search, utility, risk, profile.

1. INTRODUCTION :

THE web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. 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.

2. LITERATURE SURVEY :

Z. Dou, R. Song, and J.-R. Wen, Although personalized search has been proposed for many years and many personalization strategies have been investigated, it is still unclear whether personalization is consistently effective on different queries for different users, and under different search contexts. In this paper, we study this problem and provide some preliminary conclusions.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

M. Spertta and S. Gach, User profiles, descriptions of user interests, can be used by search engines to provide personalizedsearch results. Many approaches to creating user profiles collect user information through proxy servers (to capture browsing historieson a personal computer). Both these techniques require participation of the user to install the proxy server or the bot.) or desktop bots (to capture activities

B. Tan, X. Shen, and C. Zhai, Long-term search history contains rich information about a user's search preferences, which can be used as search context to improve retrieval performance.

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.

EXISTING SYSTEM:

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 [1], 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 poten-tially effective for almost all sorts of queries, but arereported to be unstable under some circumstances .

DISADVANTAGES OF EXISTING SYSTEM:

The existing profile-based PWS do not support runtime profiling.The existing methods do not take into account the customization of privacy requirements.Many personalization techniques require iterative user interactions when creating personalized search results.Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual, as described. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server.

3. PROPOSED SYSTEM:

As shown in [Figure-1] UPS consists of number of clients/users and a server for fulfilling clients request. In clients machine, the online profiler is implemented as search proxy which maintains users profile in hierarchy of nodes and also maintain the user specified privacy requirement as a set of sensitive nodes. There are two phase, namely Offline and Online phase for the framework. During Offline, a hierarchical user profile is created and user specified privacy requirement is marked on it.The query fired by user is handled in the online phase as: When user fires a query on the client, proxy generates user profile in run time. The output is generalized user profile considering the privacy requirements. Then, the query along with generalized profile of user is sent to PWS server for personalized web search. The search result is personalized and the response is sent back to query proxy. Finally, the proxy presents the raw result or reranks them with user profile.

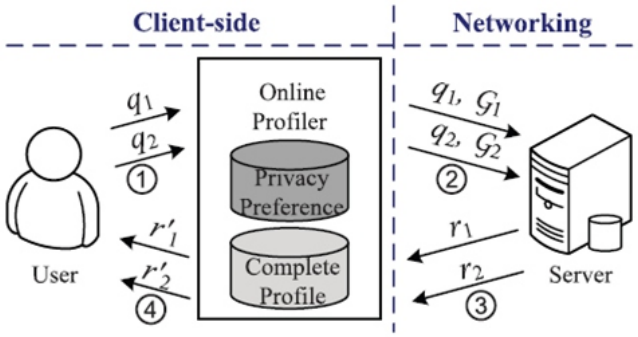

Figure 1. Architecture.

4.Greedy Algorithm :

A greedy algorithm is a mathematical process that recursively constructs a set Recursion of objects from the smallest possible constituent parts. is an approach to problem solving in which the solution to a particular problem depends on solutions to smaller instances of the same problem. Greedy algorithms look for simple, easy-to-implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit. Such algorithms are called greedy because while the optimal solution to each smaller instance will provide an immediate output, the algorithm doesn't consider the larger problem as a whole. Once a decision has been made, it is never reconsidered. The advantage to using a greedy algorithm is that solutions to smaller instances of the problem can be straightforward and easy to understand.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

The disadvantage is that it is entirely possible that the most optimal short-term solutions may lead to the worst long-term outcome. Greedy algorithms are often used in packets with the fewest number of hops machine learning, business intelligence (BI), artificial intelligence (AI) and programming. ad hoc mobile networking to efficiently route and the shortest delay possible.

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.

ADVANTAGES OF PROPOSED SYSTEM:

Increasing usage of personal and behaviour information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data bookmarks, user documents, and so forth?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.

5. SYSTEM ARCHITECTURE

A. Online profile:

The proposed idea also suggests that the queries issued are recommended that are related to the input query and also search for different issues. This redirects the search process to related information of interest to the users searching previously and also keeping track of the related queries issued by other users. The key component for privacy protection is an online profiler implemented as a search proxy that runs on client side. This proxy maintains both the complete user profile in a hierarchical structure with semantics, and the user-specified privacy requirements i.e. sensitive nodes. It works in two phases, namely the offline phase and the online phase. In the offline phase, hierarchical profile is constructed and then customized with the user-specified privacy requirements [1]. The online phase can be conducted as follows: • When query is generated the proxy generates a runtime user profile.

This process is guided by considering two conflicting metrics, personalization utility and privacy risk. • Then, the query and the generalized profile are sent together to the server.

• These results are then personalized with the profile and delivered back to the query proxy.

• Finally, the proxy sends back the results to the client. UPS differs from the conventional PWS since it provides runtime profiling which optimizes personalization utility, which performs customization on the sensitive data defined by the users, and does not require iterative user interaction. Again, for efficient browsing, it is required to find the ranks of the related queries and cluster them. Queries along with the text of their clicked URLs extracted from the web log are clustered. This is done on the basis of two notions:

• Similarity of the query. The similarity of the query to the input query.

• Support of the query. This is a measure of how relevant is the query in the cluster. It is measured with the support of the query as the fraction of the documents returned by the query that captured the attention of users (clicked documents).

It is estimated from the query log as well. The quality of service can be improved when the location of the users are closer [4]. So, if the users share more data with each other the services provided by the web will be accurate. The studies show that the user is biased when it comes to searching information on the web. It can be trusts-biased or quality-biased [3]. This shows that clicks should be interpreted relative to the order of abstracts and presentation. Some attempts are made to use implicit feedback [4]. The reading time is indicative of interest while reading new stories.

The reading time as well as number of times the user scrolls page can predict the relevance in browsing web. But it is generally considered that reading time varies between subjects and tasks, which makes it difficult to interpret. This difficulty can be resolved by the concept of eye-tracking. A general user approaches the results from top to bottom. It appears that users scan the viewable results before heading to scrolling It gives evidences about users' decision making and indicates that users' clicking decisions are influenced by relevant results.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

B. Session time-out:

An experiment can be conducted where the users are observed with their clicked URL and session lengths and then can be re-enacted. For further help, clicks can be observed and assessment of the user's objectives can be done to label each session. Each query and clicked URL ar assigned with ID number. A strength of this approach is that data is recorded without having an intervention and additionally we can observe large amount of users. There is a chance that the observer is biased to the user's goals but preliminary results show that the results achieved are reliable. The utility of adopting a hierarchical model for the grouping of user queries will allow us to more easily model what type of task the user may be doing when querying.

C. Attack Model:

The user profile should be protected from adversaries which try to hamper the privacy and sensitive nodes defined by the user by a typical attack, namely eavesdropping. As shown in the Fig. 2 the eavesdropper intercepts successfully the communication happening between the server and the user by a measure, such as man-in-the-middle attack, invading the server. Accordingly, whenever the user issues any query q, the entire copy of q along with the runtime profile of the user will be seized the attacker.

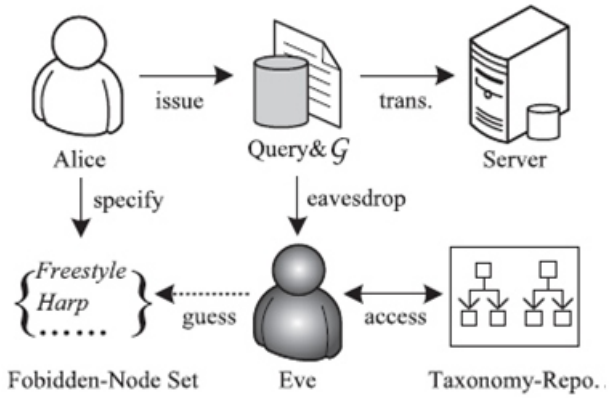

Fig. 2 Attack Model.

The attacker will then try to recover the hidden segments defined as private by the user. Now, the adversary is considered to satisfy the following assumptions: Knowledge Bound. The background knowledge of the adversary is limited to the entire information available on the web. Both the original user profile and the privacy are defined within this information.

Session Bound. Previously captured information is not available for tracing the same victim. The eavesdropping will be started and ended within a single session. These assumptions are strong but are reasonable in practice. This is considered since majority of attacks across the web happen by some automatic programs that sends advertisements (spam) to a wide range of users. An approach can be made to keep this privacy risk under control.

D. Generalizing user profile:

This technique can be considered during the offline phase processing without involving any of the user's queries. But however it is impractical to perform this in offline phase because:

• This output from the offline phase may contain many topics that are completely irrelevant to the particular query. This can be solve if profile is generalized in the online phase.

• It avoids unnecessary privacy disclosure to the adversaries and also avoids noisy topics which are irrelevant to the query.

• It is very important to monitor the personalization factor during generalizing. But overgeneralization may cause ambiguity. There are four phases in [1] which are used in generalization of the user profile. They can be explained inthe following manner:

• Offline profile construction: This is the first step of the offline processing wherein the original user profile is built in a topic hierarchy which reveals the interest o the user.

• Offline privacy requirement customization: This phase requests the user to specify sensitive nodes which the user considers to remain hidden from the world. When any query q is issued, this customized user profile goes through the online phases.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

6. RESULTS:

Figure 3. Admin login page

Figure 4. Users home window

7. CONCLUSION:

Web users were increases because of available of information's from the web browser based on the search engine. With the increasing number of user service engine must provide the relevant search result based on their behavior or based on the user performance. Providing relevant result to the user is based on their click logs, query histories, bookmarks, by this privacy of the user might be loss. For providing relevant search by using these approaches the privacy of the user may loss. Most existing system provides a major barrier to the private information during user search. That approaches does not protect privacy issues and rising information loss for the user data. For this issue this paper proposes client based architecture based on the greedy algorithm to prevent the user data and provide the relevant search result to the user in future it can include this work in mobile application.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.

[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.