International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Supporting Privacy Protection in Personalized Web Search with Secured User Profile

Archana R.Ukande[1], Nitin Shivale[2]

[1]Department of Computer Science, BSITR, Wagholi, Pune, India

[2]Assistant Professor, Department of Computer Science, BSITR, Wagholi, Pune, India

Abstract: ***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. 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. For improving better search quality the String Similarity Match Algorithm (SSM Algorithm) can be implemented with the proposed. Current solutions propose new mechanisms that introduce a high cost in terms of computation and communication, to address this privacy threat. Personalized search is a promising way to improve the accuracy of web search, also it is attracting much attention recently. Effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. These concerns have become one of the main barriers for deploying personalized search applications, and privacy-preserving personalization is a great challenge. Adversaries are tried to resist in proposed system with the help of broader background knowledge (i.e. 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. With the help of this mechanism, privacy can be achieved.***

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

1. Introduction

Novel protocol is proposed specially designed to protect the users' privacy in front of web search profiling. Adversaries are tried to resist in proposed system with the help of broader background knowledge i.e. knowing 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 DP and, Greedy IL algorithm it takes large computational and communication time.

For generalize the retrieved data by using the background knowledge [1], [5], [3], [7] through this adversaries can be avoided. Privacy protection in publishing transaction data is an important problem. A key feature of data transaction is the extreme scarcity, which renders any single technique ineffective in anonymizing such data. Among recent works, some suffer from performance drawbacks, some incur high information loss and some result in data hard to interpret. This approach proposes to integrate generalization and compression to reduce information loss. However, the integration is non-trivial. Novel techniques are proposed to address the efficiency and scalability challenges. A few previous studies [8], [9] suggest that people are willing to compromise privacy if the personalization by supplying user profile to the search engine yields better search quality

2. Literature Review

A. Privacy Protection In Personalized Search

In privacy protection, analytically observe the concern of privacy preservation in personalized search [10]. Here discriminate and describe four levels of privacy protection, and analyze numerous software architectures for personalized search. It shows that client-side personalization has advantages over the existing server-side personalized search services in preserving privacy in this situation; personalized web search cannot be done at the individual user level, but is possible at the group level. This may reduce the effectiveness of personalization because a group's information need explanation is used to model an individual user's information need.

However, if the group is appropriately constructed so that people with similar interests are grouped together, it has much richer user information to offset the sparse explanation of individual user information requirements. Thus the search performance may essentially be improved because of the availability of more information from the group profile [11] and [12]. In this circumstance, personalized web search cannot be done at the distinct user level, but is possible at the group level. This may reduce the effectiveness of personalization because a group's information need description is used to model an individual user's information need. However, if the group is properly constructed so that people with comparable interests are grouped together, it may have much richer user information to offset the sparse explanation of distinct user information needs.

Thus the search performance may really be better because of the accessibility of more information from the group profile

a. Advantages

1) The architecture has an advantage of allowing for the use of a search engine's internal resources.
2) It improves the accuracy of web search.

b. Disadvantages

1) It does not fully protect user privacy.
2) They were not discussed different levels of privacy protection provided by search engines depending on a

user's preference for the tradeoff between the privacy concern and the improved search service quality.

B. Implicit User Modeling For Personalized Search

In implicit user modeling for personalized search [2], explicated how to infer a user's interest from the user's search context and practice the conditional implied user model for personalized search. A decision speculative basis and develop methods for implicit user exhibiting in information retrieval. They developed an intelligent client-side web search agent (UCAIR) that can achieve eager implicit feedback, e.g., query development established on prior queries and instant result re-ranking established on search show that search agent can progress search accuracy over the popular Google search engine. In this paper, described how to make and update a user model based on the instant search context and implicit feedback information and use the model to improve the accuracy of ad hoc retrieval.

In order to extremely benefit the user of a retrieval system through implicit user modeling, offered to perform "eager implicit feedback". Those is, as soon as experimental any new piece of evidence from the user, and update the system's certainty about the user's information need and respond with improved retrieval outcomes based on the updated user model. A decision-theoretic basis for enhancing interactive information retrieval based on eager user model updating, in which the system replies to each achievement of the user by choosing a system exploit to enhance an efficacy function.

In a traditional retrieval model, the retrieval problem is often to match a query with documents and rank documents giving to their relevance values. As a result, the whole retrieval progression is a simple independent cycle of "query" and "result display". In the planned new recovery model, the user's search circumstance shows a significant role and the conditional implicit user typical is exploited directly to benefit the user. The novel retrieval model is thus essentially diverse from the traditional pattern, and is inherently more general.

a. Advantages

1) It expands search accuracy over the popular Google search engine.
2) The developed search cause thus can advance existing web search performance without any additional effort from the user.

b. Disadvantages

1) The search agent does not have control of the retrieval algorithm.
2) It should displayed summaries, but the document content is actually not.

c. IR Evaluation Method

IR evaluation method [4] is used for retrieving highly relevant documents. This paper proposes assessment approaches established on the use of non-dichotomous relevance judgments in IR investigates. It is maintained that evaluation methods should credit IR methods for their ability to retrieve highly relevant documents. This is desirable from the user point of view in modem large IR environments. The proposed methods are a novel application of P-R curves and average precision computations based on separate recall bases for documents of different degrees of relevance, and two novel measures cumulative computing gain the user obtains by examining the retrieval result up to a given ranked position.

Then demonstrate the use of these evaluation methods in a case study on the effectiveness of query types, based on combination of query structures and expansion, in retrieving documents of various degrees of relevance. Test was run with a best match retrieval system (In- Query I) in a text database consisting of newspaper articles. Results indicate that the tested strong query structures are most effective in retrieving relevant documents. The differences between query types are statistically significant and practically essential. More generally, the novel evaluation methods and the case demonstrate that non-dichotomous relevance assessments are applicable in IR experiments and allow harder testing of IR methods.

d. Advantages

1. The P-R curves demonstrate that the good performance of the expanded structured query types.
2. The best performance overall was achieved with expanded, facet structured queries.

e. Disadvantages

1. The DCV-based precision recall curves are better but still do not make the value gained by ranked position explicit.
2. The RHL alone is not sufficient as a performance measure.

C. Automatic Identification of User Interest

Automatic identification of user interest is done for personalized search [6]. Here a framework is proposed to investigate the problem of personalizing web search based on user s' past search histories without user efforts. Proposed a user model to formalize user's interests on web - pages and correlate them with user's clicks on search results .Based on this described correlation an intuitive algorithm to actually learn user's interests. Two different methods are proposed, based on different assumptions on user behaviors, to rank search results based on the user's interests learned.

The both theoretical and real-life experiments to evaluate our approach, In the theoretical experiment, found that for a reasonably small user search trace, the user interests estimated by our learning algorithm can be used to pretty accurately predict view based on importance of web pages, which is expressed by Personalized PageRank, showing that our method is effective and easily applicable to real-life search engines. In the real-life, we applied our method to learn the interests of 10 subjects contacted. The results

showed that, on average, our method per formed between 25%–33% better than Topic-Sensitive PageRank, which turned out to be much better than PageRank.

a. Advantages

1) The experiments show that user's preferences can be learned accurately even from personalized search based on user preference and small history data yields significant improvements over the best existing ranking mechanism in the literature.
2) PageRank is more Relevant than the global PageRank.

b. Disadvantages

1) It is not more users -specific information into consideration. The difficulties in doing this include integration of different information sources, modeling of the correlation between various information and the user's search behaviors, and efficiency concerns.
2) It does not design more sophisticated learning and ranking algorithms to further improve the performance of our system.

3. System Architecture

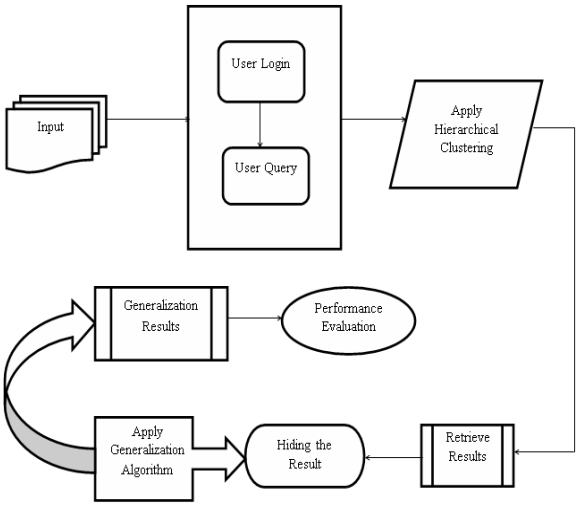

Figure 1: System Architecture

4. Modules

A. User Login

This is for user login page. In this module, user needs to register with unique user id and password based on that all further information of user will get stored.. In this module, users are entered after registering. After registering each user has unique id. After login, user can create personalized profile or can posts some queries to search.

B. User Profile Creation

Here to create user profile user needs to enter information against the category e.g. under Games category user can enter basketball, cricket, tennis etc. once user enters this information. Publicly available repository DMOZ is used for manual tagging and editing of topic to create hierarchical profile. Once information is retrieved form DMOZ it is stored in hierarchical format for further use based on user login

C. Keeping Track Of Sensitive Information

Along with profile creation user can specify sensitive information. It will not get shared to the web server, during data retrieval process. Also it removes the sensitive nodes from the user profile.to avoid the risks while sharing user profile.

D. Query Topic Mapping

When user enters text to search, after that it maps the string with categories present into user profile. If it finds matching between search string and category then it retrieves the user profile based on that with help of Greedy IL and removes the sensitive nodes from the profile. Query topic mapping it also considers approximate matching.

E. Retrieve User Profile In Privacy Manner

When user enters any string to search then it does query topic mapping and sends query and generalized profile for search. Here it assumes that query string do not contains any sensitive information. Based on that it retrieves and displays the result.

After this it stores the searched string in database for future use.

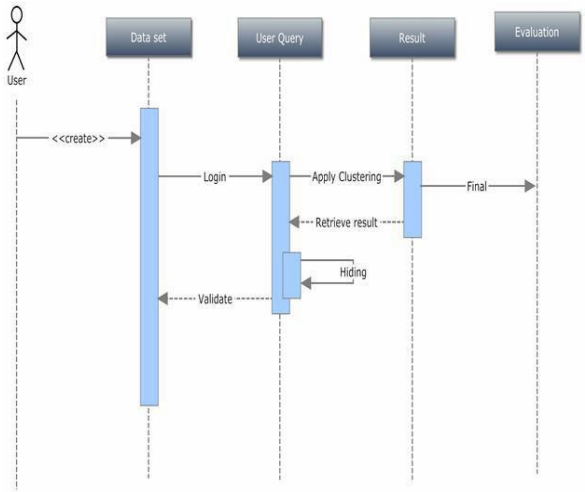

Figure 2: Sequence Diagram

5. Existing System

A. Methodology of Existing System

The previous works of privacy preserving PWS are far from optimal. The problems with the existing methods are explained in the following observations. The existing profile-based PWS do not support runtime profiling. Runtime profiling means user profile creation based on entered query at that point. A better approach is to make an online decision on

a) whether to personalize the query (by exposing the profile) and
b) what to expose in the user profile at runtime.

To the best of our knowledge, no previous work has supported such feature. When the user searches any information on to the web browser it gets shared over the network. If user wants to prevent any information getting shared over the network in that case the existing methods do

not take into account the customization of privacy requirements. It assumes that the interests with less user document support are more sensitive and vice versa that can't be correct in all situations.

When user enters anything to search, he expects the information based on his interest that might be not included into that search string. Because of that 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.

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.

B. Disadvantages:

Existing system 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. 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 pre-vent 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.

6. Proposed System

The above problems are addressed in our UPS (literally for User customizable Privacy-preserving Search) framework. The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS

Algorithm Used—Greedy Information Loss Algorithm (Greedy IL) 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.

Offline-1 : Profile Construction.

The first step of the offline processing is to build the original user profile in a topic hierarchy H that reveals user interests. We assume that the user's preferences are represented in a set of plain text documents, denoted by D.

Offline-2: Privacy Requirement Customization.

This procedure first requests the user to specify sensitive-nodes,

1. for each sensitive-node
2. for each no sensitive leaf node

When a query q is issued, this profile has to go through the following two online procedures:

Online-1: Query-topic Mapping.

Given a query q, the purposes of query-topic mapping are

1) To compute a rooted sub tree of H, which is called a seed profile, so that all topics relevant to q are contained in it; and
2) To obtain the preference values between q and all topics in H.

UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. In existing system proposed algorithms are greedy algorithms, namely GreedyIL, for the online generalization. In this for query mapping process it has various steps to compute the relevant items.

In the proposed system, clustering algorithms can be implemented 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. Privacy protection present a novel protocol specially designed to protect the users' privacy in front of web search profiling.

Advantages:

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

B. Methodology of Proposed System

After registration process user can create customized profile which gets stored into database with his credentials And when user searches for information then then based on query topic mapping it creates the generalized user profile for search as follows. With the help of Greedy IL algorithm.

1) Obtain seed profile from online 1 by mapping category
2) Identify sensitive items and maintain the list
3) Process pure leaf (t) operation till it contains elements
4) If t has siblings then no operation on siblings, merge it into shadow siblings
5) If t has no siblings then add those into generalized profile
6) Go to step 3.
7) Return generalized profile.

7. Results

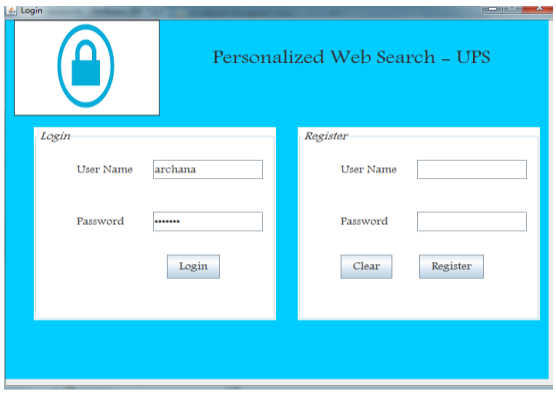

Figure 3: Home Page

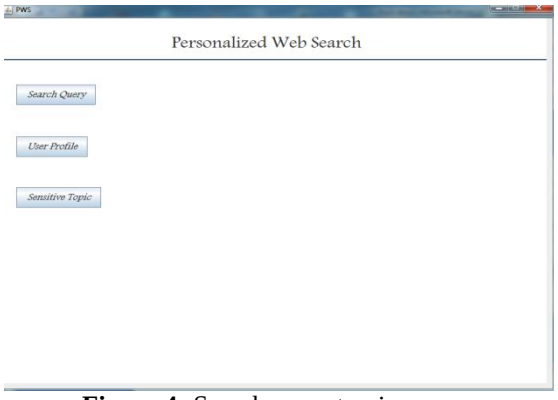

Figure 4: Search or customize menu

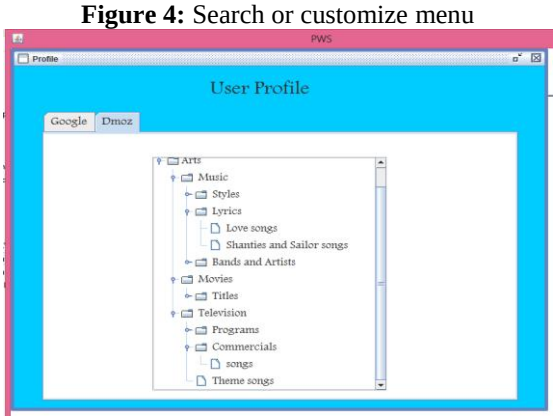

Figure 4: User Profile

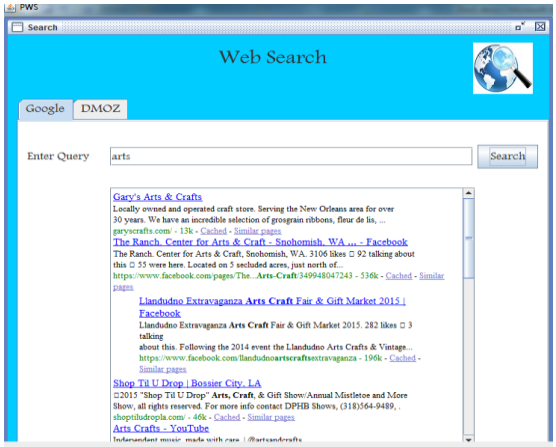

Figure 4: Search result

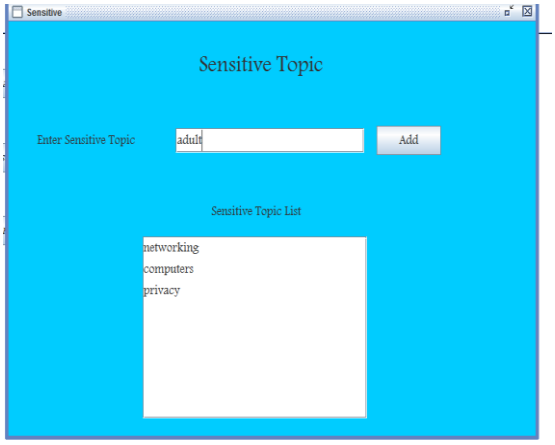

Figure 4: Customization for sensitivity

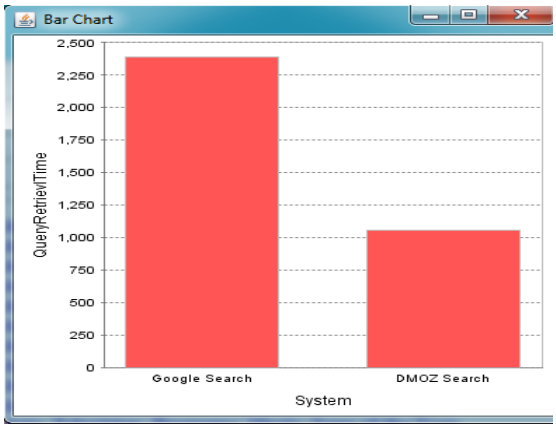

Figure 5: Result

8. Conclusion and Future Enhancement

For generalizing the retrieved data by using the background knowledge. Through this adversaries can be resists. 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 non-trivial. The novel techniques are proposed to address the efficiency and scalability challenges.

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.

9. Future Enhancements

Future work can be implemented for hierarchical divisive approach to retrieve the search results. It will gives better performance when compared with the our proposed System.

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] 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
[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. 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
[5] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨tter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 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] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.
[8] 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.pp. 1497-1500, 2009
[9] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.
[10] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007
[11] 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.