ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 4, Issue 9, September 2015*

# Privacy protection in personalized web search

**Khwaja Aamer[1], Dr. A. S. Hiwale[2]**

Department of Information Technology, MIT college of Engineering, Pune, India[1]

HOD, Department of Information Technology, MIT college of Engineering, Pune, India[2]

**Abstract**: Search engines play important role in fulfilling user's desire to reach relevant results. Search engine treats all users in same category and follow one profile fits all strategy, neglecting the needs of different kinds of user. Personalized web search (PWS) allow users to get only those results which are relevant to them by generating the most relevant results to each user according to their needs based on their profile, interest and needs. For getting benefit of personalized web search user profile is maintained which includes personal information. This user information can become biggest threat because most of the users are not comfortable to share their personal information in fear of hacking. To protect user information, privacy in PWS is required. To do so our algorithm provides balance between privacy by generating fake queries and quality by using online profiler. Experiment shows that accuracy and privacy protection is achieved.

**Keywords**: Accuracy, fake queries, online decision, privacy protection, personalized web search.

## I. INTRODUCTION

Web search has become the integral part of life and almost everyone using search engines to find the information easily and in effective way. Web as an information system contains web pages without much distinction. There are trillions of web pages and user needs only few of them, but related. Typical search engine can give same results to the same query submitted by different users. For example the word run has 606 different meanings which create ambiguity or the bank could mean financial institution and riverside [6]. Search engines fail to provide relevant results to users in case of ambiguous words. To serve this need personalized web search is introduced.

Personalized web search is an attempt to provide people better search results. When the user is signed-in, search engine will be able to provide most relevant results according to user profile [8]. . For example, a user always searches for [sports] and click on results from espn.com, the search engine will rank espn.com higher on the results for next time user search. On the other hand, when I'm searching for schedule about interviews in India but it might show me pune job result first, instead of jobs in India because I frequently click on punejobs.com, Google might show me this result first, instead of the Big Red soda company or others.

Profiles contain data related to user's interest and past searches to track them and to improve ranking algorithms against the size of web. This data is stored either client side or server side called query logs. This data is then extracted by applying data mining techniques and search methods to characterize areas of interest or demographic aspects (e.g., age, gender or nationality).

Personalized web search saves the time but raises privacy issues [9]. As user information may contain sensitive information like his belief and views about some organization which needs to be protected from going into wrong hands because bunch of queries can identify the user. For example the case of Thelma Arnold, user of the AOL's WSE, who was identified by her searches, after AOL leak search data of 650000 users submitted over a three-month period. To protect the real identity of user all queries were hidden behind a pseudonym. Single query might not reveal user identity however, the aggregation of hundreds of queries was enough to identify and profile her [10]. To address the problem of privacy protection many approaches are taken such as cryptographic solution or solution by using Dissociating Privacy Agent [7].

## II. LITERATURE REVIEW

We now look at the existing methods and terminologies used in the prior work.

Lidan shou, et al.[1] discusses the privacy issues in personalized web search environment. It stores user preferences in hierarchical order. Author propose runtime generalization which helps in keeping balance between personalization utility and privacy risk of exposing information. The framework called UPS (User Customizable Privacy-preserving Search) contains online profiler running at client side; it takes online decision on whether to personalize query or not to improve privacy and quality. Online profiler improves the search quality for ad hoc queries but privacy protection is the problem. To resolve this problem customization of privacy requirements is introduced which allow users to adjust the privacy according to need.

Alexandre Viejo et. al. [2] proposed a method called single party scheme focus on preserving user's private and personal information by generating fake queries and without any external entity. This scheme also take care the two contradicting parameters, privacy and quality without any change at server side. Author proposes a scheme where m fake queries are generated submitted to the server with original query. The method of generating fake queries is same as GooPIR [22] or TrackMeNot [23] but point of focus here is that the generated fake queries are almost

**DOI 10.17148/IJARCCE.2015.49119**

similar to the original query which improves quality. The fake query generation is base on knowledge base which controls the distance between original and fake queries. This paper uses Open Directory Project (ODP) hierarchy which is the most comprehensive and largest hierarchy [26]. The main advantage of this scheme is that user is given freedom to select the number of fake queries to be generated and freedom to decide distance between fake and original queries.

Kenneth Wai et.al.[3] proposed a personalized search engine for mobile which stores user preferences by using clickthrough data. The advantage above all previous work is that user location is considered and according to the location results are delivered. For example user if is in Pune and searching for hotels then the search engine must show him nearby hotels in pune. Author proposed client server architecture called Personalized Mobile Search Engine (PMSE), privacy is protected by restricting the clickthrough data at client side. The architecture divides the tasks among PMSE server and PMSE client such as heavy tasks performed at server and light weight tasks performed at client which ensure fast and efficient search.

Rakesh Kumar et. Al.[4] uses domain knowledge with user's browsing history which search improves performance. Domain knowledge is used to store information about different categories and browsing history is used as user profile. Learning component is heart of the proposed framework which learns user choice by user's browsing history and enhanced user profile is built and relevant web pages are suggested to user based on enhanced user profile. O shafiq et. al.[10] proposed community aware personalized web search where user activities are captured through social networking profile. This approach has many drawbacks which are addressed by enhanced user profile.

S. vanitha[5] proposes new approach for personalized web search where every user has two user profiles and using the two user profile one more effective user profile is created. The approach works in simple way. If user already has a profile then this profile is analyzed and if the terms are missing then two user profiles are generated and the last step is that these profiles are compared and the most effective user profile is built based on the two user profiles. Using two user profiles helps getting more personalized web pages but handling two user profiles is chaos.

## III. PROPOSED WORK

In this section, we focus the working of framework and overview to explore the same. The paper proposes a framework for securing identity of a user in personalized web search.

### A. Algorithm

```
        Input: user search queries
        Output: personalized search results
1. Accept query from user
2. Generalize user profile
3. Check for sensitive query
4. If (query=sensitive)
   Perform prune leaf operation
5. Generate fake queries
6. Send query to the server
7. Rerank results according to profile
8. Display result
```

### B. Architecture

The framework consists of multiple users and mistrust/doubtful search engine server. The most crucial component here is online profiler running as a search proxy on client side. The online profiler upholds the complete user profile in hierarchy. The complete user profile is divided into two parts 1.Generalized data 2.sensitive data generalized data is a data which user wants to share or disclose without any hesitation.

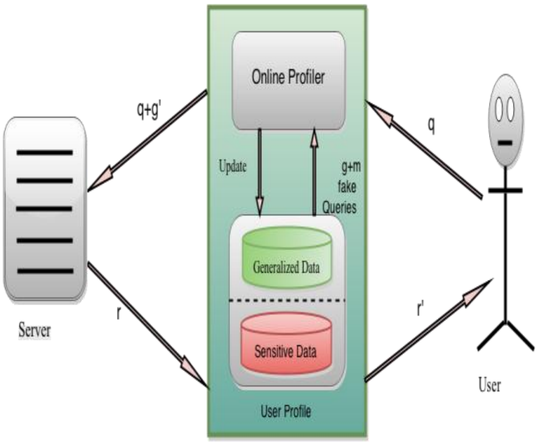

Figure 1: Proposed Architecture

Sensitive data is the data which user doesn't want to share or this data is private for user. The generalized data or sensitive data may be different for different user based on their requirements; WSEs must provide privacy to user according to his/her requirements. To meet this goal framework uses user specified privacy requirements which allow users to define what is sensitive for him/her. Figure shows the architecture.

When user issues a query q, online profiler generates user profile in runtime according to query and result of this step generalized user profile. After generating generalized profile the query and generalized profile along with m fake queries are sent to the mistrust server. Server evaluates the user profile and transport results to the online profiler according to the user profile.

In the last stage online profiler reranks the result according to user's privacy requirements and deliver results to the user.

## IV. METHODOLOGY

### A. Personalization

Search engines can give unwanted results to user because user entered queries may be ambiguous or incomplete. To get better search results user profile is maintained and

DOI 10.17148/IJARCCE.2015.49119

based on that personalized results are shown to user. The user profile is created in hierarchical structure which provides efficiency scalability and ….

B. Privacy

User profile personalized the search but becomes threat to user privacy. User search pattern is stored in user profile which can expose user's daily needs and habits. To provide balance between search quality and the privacy user is given the choice to scale the privacy and quality according to his/her need. User is given choice to decide what is sensitive for him or what is not. The sensitive queries of user are protected and kept at client side to avoid exposure of sensitive query.

C. User Profile

The user profiles are built in hierarchical structure. prerequisites are met by the generalization process to handle the user profile, achieved by preprocessing the user profile. The user profile is rooted subtree. At first, the parent node of profile is taken into consideration. In second step inherited properties are added to the user profile. Sensitive queries are kept at user side and rest queries are sent to server. The sensitive queries are pruned by online profiler to avoid risk.

*D. Online Decision*

The personalization is provided on profile-based which degrades the quality of search and would risk user's privacy while exposing profile to server. To solve this problem user's are given a choice to select privacy level according to his needs. i.e user will decide to personalize or not. In case of distinct query whole runtime profiling is aborted and query will be sent excluding profile.

*E. Fake query generation*

Based on the sensitivity level user can choose how many fake queries he wants to issue. More the number of fake queries more the privacy. Algorithm allows users to customize their privacy requirements or to decide whether to personalize a query or not in efficient manner.

## V. EXPERIMENTAL SETUP

The framework is implemented on a Dual core 2.10-GHz CPU and 2 GB RAM running windows7. The GreedyDP and greedyIL algorithms are implemented in JAVA. Search results are retrieved from Google search engine. Online profiler runs as a proxy at client side, not inside the search engine due to practical reason.

For each query, results are retrieved from Google search engine and then reranked by the online profiler and then delivered to the user [11].

## VI. RESULTS

Figure 2 shows the fake queries generated by the system. The Y-axis shows the number of fake queries generated while using personalized web search engine and on the X-axis, we had plot the queries for which the fake queries are generated.

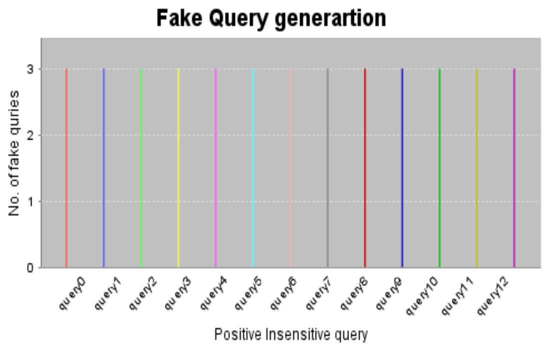

Figure 2: Fake query generations

We consider that this k (no. of fake queries) value offers a correct trade-off between privacy and quality of distorted user profiles. A small k will generate fake terms very close to the original ones which implies weak privacy protection, while a large k will produce fake terms far away from the authentic ones which in turn generates a useless user profile, k = 3 is an average value which may produce a balanced behavior.

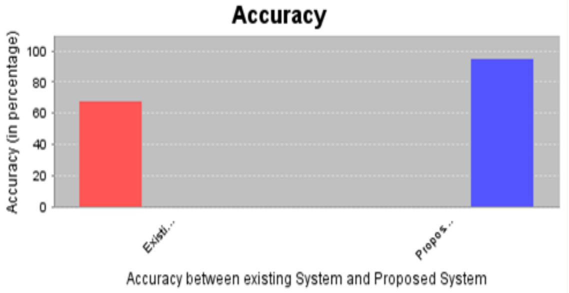

Figure 3: Accuracy of Proposed Algorithm

Figure 3 shows comparison between the accuracy for the greedyIL algorithm and the proposed algorithm. For measuring the accuracy NDCG methodology is used where user will give the input about whether the query is relevant for him or not. Based on the results obtained by different users the accuracy is measured in percentage. Thus, the accuracy is improved by approximately 20% using proposed algorithm.

The figure 4 and 5 shows comparison between the performances for the greedyIL algorithm and the proposed algorithm. The x-axis shows the comparison of the both filters and on the Y-axis, we had plots the number of queries searched by the user. The sensitive queries sent to the server is less in proposed algorithm as compare to GreedyIL algorithm which shows the privacy achieved is higher than the existing algorithm.

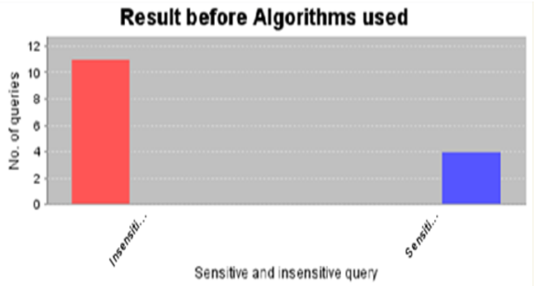

Figure 4: Privacy of GreedyIL Algorithm

DOI 10.17148/IJARCCE.2015.49119

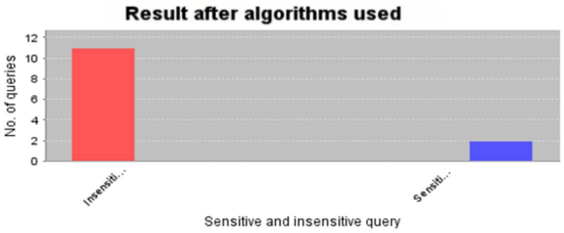

Figure 5: Privacy of Proposed Algorithm

Fig shows the sensitive queries sent to server using greedyIL is 4 but it is reduced to 2 in proposed algorithm which increases the user privacy.

## VII. CONCLUSION

Today's era is of doing work with high efficiency but time consumption is also what matters. When it comes to business and dealing with data, quality and accuracy is what matters the most. Personalized web search can provide better search results. Online profiling is implemented to protect personal privacy without comprising search quality. The privacy is protected by keeping the sensitive data at client side itself and not sending the complete profile to server, by generating fake queries. The ideal framework abolish the burden of checking for document of users need and provides more secure environment for each query issued by user. The accuracy of the system is successfully improved approximately by 20%. The results shows better performance in terms of privacy protection, which increased by 20%, time taken to generate fake queries is reduced by 12%. In future work focus will be improving scalability effectiveness and accuracy or to enhance capability to capture a series of queries from the victim by extending algorithm.

## REFERENCES

[1] Lidan Shou; He Bai; Ke Chen; Gang Chen, "Supporting Privacy Protection in Personalized Web Search," *Knowledge and Data Engineering, IEEE Transactions on* , vol.26, no.2, pp.453,467, Feb. 2014 doi: 10.1109/TKDE.2012.201

[2] Viejo, A.; Castella-Roca, J.; Bernado, O.; Mateo-Sanz, J.M., "Single-party private web search," *Privacy, Security and Trust (PST), 2012 Tenth Annual International Conference on* , vol., no., pp.1,8, 16-18 July 2012 doi: 10.1109/PST.2012.6297913

[3] Leung, K.W.-T.; Dik Lun Lee; Wang-Chien Lee, "PMSE: A Personalized Mobile Search Engine," *Knowledge and Data Engineering, IEEE Transactions on* , vol.25, no.4, pp.820,834, April 2013 doi: 10.1109/TKDE.2012.23

[4] S.Vanitha "A personalized web search based on user profile and user clicks" International Journal of Latest Research in Science and Technology Volume 2, Issue 5: Page No.78-82,September-October 2013

[5] Rakesh kumar Kumar, "Personalized web search using browsing history and domain knowledge," *Issues and Challenges in Intelligent Computing Techniques (ICICT), 2014 International Conference on* , vol., no., pp.493,497, 7-8 Feb. 2014 doi: 10.1109/ICICICT.2014.6781332

[6] http://www.cam.ac.uk/research/features/our-ambiguous-world-of-words

[7] Marc Juarez, Vicenç Torra "DisPA: an Intelligent Agent for Private Web Search" Springer International Publishing 2015 Privacy Web search, Torra, Vicençdoi.1007/978-3-319-09885-2_21

[8] Radhika.M,V.Vijayac Chamundeeswari, R. Ramya devi "A survey on personalization of web search using generalized profile" Proceedings of 14th IRF International Conference, 28th September 2014, Chennai, India, ISBN: 978-93-84209-55

[9] Smyth, B., "A Community-Based Approach to Personalizing Web Search," *Computer* , vol.40, no.8, pp.42,50, Aug. 2007

10. Dou, Z., Song, R., and Wen, J.R. "A large-scale evaluation and analysis of personalized search strategies" In Proceedings of WWW "07, 581-590

[10] Khwaja Aamer; Dr. A.S.Hiwale "A survey on "privacy protection in personalized web search" , International journal of science and research Volume 3 Issue 12 Dec-2014, ISSN_NO: 2319-7064

[11] Khwaja Aamer, Dr. A.S.Hiwale "Securing identity in personalized web search", Spvryan's International Journal of Engineering Sciences & Technology (SEST), 2015.

## BIOGRAPHIES

**Khwaja Aamer** Research Scholar, MIT college of Engineering, University of Pune. He has received B.E. in Information Technology from BAMU University, Aurangabad. Currently he is pursuing M.E. in Information Technology from MIT college of Engineering, Pune University of Pune, Maharastra, India.

**Dr. Prof. A. S. Hiwale** received PhD in E&TC. He is working as Professor, Head of the Department in Department of Information Technology, MIT college of Engineering, Pune, India. He is having more than twenty five year experience. His research interest is Wireless, Digital Communication.

DOI 10.17148/IJARCCE.2015.49119