Securing identity in personalized web search

Khwaja Aamer
Department of Information Technology,
MIT college of Engineering,
Pune, India
aamerit121@gmail.com

Dr. A. S. Hiwale
HOD, Department of Information Technology,
MIT college of Engineering,
Pune, India
ashiwale@gmail.com

Abstract: **Search engines are crucial for getting information from World Wide Web. These search engines are created for all users neglecting their individual needs and pursue the "one size fits all" model which is not flexible to individual users. Personalized web search (PWS) generates the most relevant results to each user according to their needs based on their profile. while Web searching, user profiles play very important role for better retrieval effectiveness but using a user profile to find interest poses some threat to privacy ,violation of privacy. To overcome this problem privacy protection is essential. Here, in this paper the framework generalize user profiles by queries according to user's privacy requirements. The framework also takes online decision on whether to personalize query or not. User privacy is protected by generating fake queries and by keeping the sensitive data at client side.**

Index terms: fake queries, online decision, privacy protection, personalized web search

I. INTRODUCTION

With today's modern era information age, the Internet can enable individuals to access information more easily. Web is formed up of 60 trillion pages and its perpetually growing, to get the required document user have to search multiple pages[10]. Modern era web search engines trying to overcome this situation and trying hard to deliver the results required by user. To achieve better search results multiple programs and formulas are written, algorithms are implemented to understand what do user mean, by using spell checker, search methods, metonym and after analyzing all probable clues most relevant document from index is delivered to the user[10].

The Web Search Engines (WSE) are become key factor in information searching and gained popularity but WSE present information mix of images, web pages and other files from many sources, there are fairly high chances that most of the information provided by WSE is irrelevant. This irrelevance of results may be due bulk amount of data or due to some incomplete or ambiguous query entered by user i.e. the WSE is not able to figure out different type of users and their query patterns[11]. It makes user to waste more time to deal with the irrelevant information in which they are not interested. For example the query "bat" is ambiguous because some users like sports men, cricket fan may be keen for documents related to "bat" as "cricket bat" while some of the other users like scientist or biology professor may want documents related to "bat bird"[12]. If same results are delivered to both the users it will create problems to find the actual content which user wants.

The solution is personalized web search (PWS), personalizing web search is a methodology which gives better search results according to individual's need. Personalization access the information that is relevant to user search query and algorithms decide which queries are irrelevant. There are two types of personalization 1. Click based 2. Profile based

Click based method gathers user information by carefully observing clicks in search results and learns the sites one favoring. Ex if anyone searching links about cricket, the system learns about users like and ranking boost is given to cricket query and user will see more cricket listing. In profile based method a user profile is maintained which reveal user information goal and collect data about user's activities to improve search utility

A. Privacy

Building user profile may assure quality but it degrades the privacy. When user searches for some query the user profile may get visible to another user or some middle man who wants to steal information. Different methods can be used to expose user information from his profile include, IP address of the computer, browser cookies and browser search bars[9]. To protect user privacy proper profile building is essential because each user has different information goals. Survey shows that 80% consumers are interested in personalization; yet, only 32% willing to share information [8]. Individuals may have different privacy requirements for example, consider a cricket fan will be comfortable to share his interest in cricket if it helps him to get schedule, news on cricket but he will be not comfortable airing his purchase pattern or work-out, at the same time another user might not have any problem to show his daily work out.

B. The benefits of paper

This paper objects at providing privacy as well as quality; bridging the gap between personalization and

privacy protection. It provides an environment where users can decide their own privacy setting based on structured user profile.
The paper offers scalability to build hierarchical user profile. To protect user privacy, framework would prune the sensitive nodes from hierarchical profile. Sometimes it might not necessary to prune sensitive nodes.
The paper offers online decision to personalize query or not i.e. to handle distinct queries which are entered by user sometimes. This allows user profile to halt, and only distinct query is sent to server without profile.

II. LITERATURE REVIEW

We now look at the existing methods and terminologies used in the prior work.
Lidan shou, et al.[1] explained the security and privacy issues in personalized web search environment. PWS has gained significant popularity in short time, but it is yet an emerging technology. Essentially, it aims to combine the utility search model and privacy with the evolutionary development. This paradigm still not clear and has many doubts in IT communities about the working, its effectiveness, working and how differences going to work. Author proposed a new framework for web personalization that uses proxy server as online profiler; UPS which can foster generalize profiles by queries. User profiles either learnt from historical activities or specified by themselves [1][4]

The novel features backed by paper are

1. Runtime profiling: i.e. "one profile fits all" strategy is thrown out and replaced by online profiler, which favor separate profile for every user. It takes online decision on whether to personalize a query or not, which improves the search quality and privacy.
2. Deals with the customization of privacy requirements to address the privacy needs of individual.
3. Iterative search is not required during personalized search results creation.

Alexandre Viejo et. al. [2] proposed a method called single party scheme that takes care of both privacy and quality at the same time without any change at the server side. The proposed scheme generates m fake queries and these fake queries with authentic one are sent to the server to hide the original query from middle man. The main advantage of this paper is that the similarities between original query and fake query are taken into account thus the achieved quality of service is high. Fake queries are generated based on knowledge base which shortens the distance between fake and original query and user is given a freedom to select distance between original and fake queries to achieve desired level of quality and privacy. After acquiring queries from Open Directory Project (ODP), all queries (fake and original) are submitted to the web search engine.

Mobile search is increasing day by day but mobile devices have limited interfaces, I/O and narrow bandwidth. To ensure better mobile search many solutions have been proposed by various authors such as feng gui et.al.[6] , kunhui lin et.al.[7] but results are not as expected. To figure out the problems, Personalized Mobile Search Engine(PMSE) framework proposed by Kenneth Wai et.al.[3] that stores click through data to learn user preferences. The search engine deliver results based on users location (GPS is used to find users location) and click through data. The PMSE is client server architecture where computation intensive tasks are handled by PMSE server and low computation tasks are handled on mobile devices. The clickthrough data is stored at client side locally and information is restricted in user profile to protect the privacy.

Makvana K et.al [4] proposed a novel method that uses query reformulation and user profiling. It uses previous search to identify relevant results by analyzing web log file maintained in the server. It then rerank and proceed the user search result by calculating interest value of users retrieved links. The user interest values are generated from Vector Space Model. It maintains personalized web search agent which retrieve files from web logs and rerank the results according to web log but this web log poses threat to privacy.

Zhicheng Dou et. al.[5] focuses on improving effectiveness; author has used 12 days of Windows Live query logs on five algorithm to achieve best algorithm. Algorithms are based on both click based and interest based approach. Previous methods had some drawbacks which are addressed. The proposed framework uses Historical click based data [4] works on the principle, frequently clicked pages are more relevant than those seldom clicked by the user. For relevance judgment user clicks are utilized to evaluate search accuracy. The most relevant documents are reranked higher in the list, to give better results. This framework is more useful for evaluating precision when experimenting with large number of queries.

III. PROPOSED METHOD

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. 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

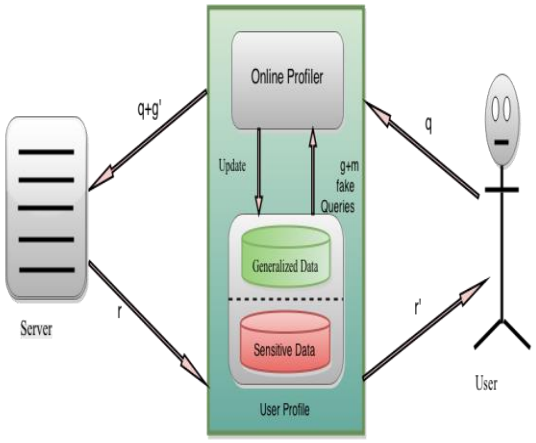

Figure 1: Proposed Architecture

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. 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.

V. RESULTS

A. *User Input*

The proposed system is implemented and generated following interface. First the registered user will enter the query in hierarchical order as shown in fig:1. This hierarchical structured query is saved into database and used later for user profile generation. User can specify sensitivity, 1 for sensitive query and 0 for generalized query to protect privacy.

B. *Popularity*

Each time user enter the query it is saved in to the database and rank of that query is increased based on the number of times query entered. For example movie/Hollywood/adult is most searched query with popularity 12 as shown in fig:2. These values are then used for giving personalized results.

Figure: Snapshot of a user searching query

user	query	sensitivity	popularity
xyz	sports/cricket/sachin	0	8
user1	mobile/samsung/galaxy	0	1
user1	sports/cricket/kapil	1	10
user1	sports/cricket/rohit sharma	0	1
user1	movie/bollywood/PK	0	9
user1	sports/cricket/sachin	0	9
xyz	movie/hollywood/adult	1	10
xyz	movies/bollywood/alone	1	3
xyz	Music/Classic/Lata Mangesgkar	0	6
xyz	movies/bollywood/roy	1	1

Figure: Snapshot of a user history stored according to popularity

VI. CONCLUSION:

In this paper, an approach is introduced to personalize web search results by using greedy algorithm. Previous methods have emerged to increase search effectiveness but privacy preservation is not handled well. First introduced an approach that takes online decision on distinct queries to personalize and then protects the user privacy by keeping sensitive data at client side and by generating fake queries. snapshot shows that the queries are properly ranked according to popularity based on user context. Proposed approach will display most relevant link at top of the retrieved result.

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]Makvana, K.; Shah, P.; Shah, P., "A novel approach to personalize web search through user profiling and query reformulation," *Data Mining and Intelligent Computing (ICDMIC), 2014 International Conference on* , vol., no., pp.1,10, 5-6 Sept. 2014
doi: 10.1109/ICDMIC.2014.6954221

[5]Zhicheng Dou; Ruihua Song; Wen, J.-R.; Xiaojie Yuan, "Evaluating the Effectiveness of Personalized Web Search," *Knowledge and Data Engineering, IEEE Transactions on* , vol.21, no.8, pp.1178,1190, Aug. 2009
doi: 10.1109/TKDE.2008.172

[6]Feng Gui; Adjouadi, M.; Rishe, N., "A Contextualized and Personalized Approach for Mobile Search," *Advanced Information Networking and Applications Workshops, 2009. WAINA '09. International Conference on* , vol., no., pp.966,971, 26-29 May 2009
doi: 10.1109/WAINA.2009.187

[7] Kunhui Lin; Jie Liu; Ming Qiu; Kaijie Guo, "Location-based personalized mobile search," *Computer Science & Education (ICCSE), 2014 9th International Conference on* , vol., no., pp.536,539, 22-24 Aug. 2014
doi: 10.1109/ICCSE.2014.6926519

[8] Yabo Xu , Ke Wang , Benyu Zhang , Zheng Chen, Privacy-enhancing personalized web search, Proceedings of the 16th international conference on World Wide Web, May 08-12, 2007, Banff, Alberta, Canada\

[9]Jaime Teevan Susan T. Dumais Eric Horvitz "Personalizing Search via Automated Analysis of Interests and Activities" , International Journal of Advanced Research in Computer Science Engineering and Information Technology Volume: 2 Issue: -Mar-2014,ISSN_NO: 2321-3337

[10] http://googleblog.blogspot.in/2009/12/personalized-search-for-everyone.html

[11] Kumar, R.; Sharan, A., "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

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