International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Privacy Protection in Personalized Web Search

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

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

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

Abstract: ***Web search engines are very important part in web life. Web search engines are built for all users and not for any individual user. Generic web search engines cannot identify the different needs of different customers, if user enter improper keyword, ambiguous keywords and lack of users ability to express what they need are some challenges faced by generic engines. We should personalize search results to address this issue. Personalized web search (PWS) is ability to identify different needs of different people who issue the same text query for web search and to carry out data retrieval for each and every user as a part of his interests. In Web searching, user profiles are main source for better retrieval effectiveness but using a user profile to find interest is violation of privacy. To overcome this problem privacy protection is required. Here, we have discussed existing methods for privacy protection and effectiveness in personalized web search.***

Keywords: privacy, personalized search, profile based search, search engine

1. Introduction

Web is made up of 60 trillion individual pages and its constantly growing, to find document of our need we follow link from page to page. To deliver best results programs and formulas are written, algorithms look to understand what do you mean, by checking spelling, search methods, synonyms after checking all possible clues most relevant document from index is delivered to the user. People are getting more dependent on WSE's for information needs People use web search for many reasons like for finding queries of daily need or business issues or for getting information about entities, web search engine sorts information out of millions of pages and send results to the users. Because of large size of web or amount of information continuously increasing user may get thousands of results which may be related or not related i.e queries submitted by different type of user with different need may get same results. The features of the query submitted by the user are; like in complete, short and ambiguous. For example for the query "bat" some users like sports men, cricket lover may be interested in documents related to "bat" as "cricket bat" while some other users like scientist or biology professor may want documents related to "bat bird". If same results are delivered to both the users it will create problems to find the actual content which user wants.

The user clicks one or more documents that look relevant and skips those documents that the user is not interested in.[4] 68% of the users click a search result within the first page of results and 92% click a result within the first three pages [5]. Therefore, WSEs must put the links that are more interesting for the users in the first result page. It is the need to deliver related contents to user based on user profile its challenge when different user search for similar query in different context, in this era of technology users expect WSE to be intelligent and serve results according to their needs where our general search engine failed. The solution is personalized web search (PWS), personalizing web search (PWS) is a technique which provide better search results according to individual's need. PWS motivates to concentrate more on creating interactive content, high quality content but raises reasonable concerns about privacy. User may be uncomfortable to expose personal information, which lead to being increasingly identifiable and can release personal information of user. Most efforts have ignored privacy to enhance utility, as these are two contradicting effects, to improve search quality user should compromise on search utility or vice versa.

For personalizing online services implicit and explicit methods can be used. In explicit personalizing methods users specify there topic of interest that reside on server or client on the other hand implicit personalizing methods user is not aware about his/her information collection which includes user location, clicks and search activities. To capture user's interest for personalization two methods are used namely click based and profile based. Click through is simple, gathers data generated by user click i.e. repeated queries from same user. Profile based method maintains complete user profile to form user interest models; these are effective for all queries by same user. Even though these user profile can't identify users directly but they can achieve identification by recovering IP address linked to bunch of queries or with his/her name, national Id etc. single query might not reveal identity of a user, bunch of queries might cause this situation. An example of this situation is the case of Thelma Arnold, user of the AOL's WSE, who was identified by her searches, submitted over a three-month period. All these queries were hidden behind a pseudonym to protect the real identity of the user. However, the aggregation of hundreds of queries was enough to identify and profile her [14]. User profile contains sensitive and personal information which pose serious privacy threat to user. WSE's are not proper to use for privacy instead user should use privacy preserving mechanism to prevent exposing information.

2. Attacks

The evolving personalization introduces a attack surface for all those want to steal user information, regardless of their

intents personalization either log their users to track the activities on the site or by using session cookie even when users are logged out. Attackers may also use past history which incorporated by personalization to customize the content.

2.1 Pollution Attacks

If a user visits a web page containing exploit, these services allow third party to alter the personalized content with the objective of altering users choice set[6]. The attacker alters the user's history using false clicks, using cross-site request forgery as the attack vector. Attacker can model personalization algorithms to affect the choice set of user or based on some knowledge attacker can inject a seed to the users choice set.

- Visit based: to promote particular product, the attacker visit the Amazon and retrieve the related product and that URL can be used as a seed to promote product.
- Search Based: to promote particular product, the attacker use natural language toolkit to extract keyword of that product and combination of these keywords can be used as a seed.

2.2 Historiographer

An attack that reconstruct the history of web searches conducted by users [7]. Historiographer uses the cookies that transmitted on the network and uses that cookie to exploit the choice set of user to alter users search history. It reconstructs the user history by enclosing sensitive and protected information from non sensitive data this attack is much more powerful than the eavesdropping attack.

3. Existing Methods

We now overview the existing personalize web search and terminologies used in the prior work Lidan shou, et al.[1] explained the security and privacy challenges in PWS environment. PWS has generated significant interest in both the world, but it is yet an evolving paradigm. Essentially, it aims to combine the utility search model and privacy with the evolutionary development. Many doubts exists in IT communities about how a PWS differs from existing web search and how these differences affect its adoption. He proposed a new web search personalization approach that uses online profiler as a key component; UPS which can foster generalize profiles by queries. User profiles either learnt from historical activities or specified by themselves [1][4]

The novel features supported in this paper are

1. It supports runtime profiling i.e. "one profile fits all" strategy is replaced by online profiler which considers separate profile for each user. It helps to improve the search quality and privacy by taking online decision on whether to personalize a query or not.
2. Takes into account the customization of privacy requirements. It effectively addresses individual privacy needs.
3. Not require iterative user search while creating personalized search results.

To ensure privacy many solutions have been proposed such as Private information retrieval (PIR) [12] oblivious transport (OT) protocol [13] these protocols provide confidentiality of results but deployment of these are not realistic in real world because they require support from provider.

To resolve the problems new scheme proposed by Alexandre Viejo et. al. [2] that generate m fake queries and submit together with authentic one, this architecture consider similarity between original and fake queries therefore the quality of service achieved is high. This paper proposed generation of fake queries on Knowledge base to minimize the distance between fake queries and authentic queries. According to his/her desired level of privacy and quality user will select the distance between original query and fake query. Finally all queries are submitted to WSE.

In 2013 Kenneth Wai [3] et.al proposed a personalized web search engine that mines their click through data to capture the user's preferences**.** The paper [3] focus on a new kind of search engine, which focus on recognizing the results according to users location, GPS is used to position user location. PMSE is application designed on Google android platform to personalize results for mobile search engine. The architecture uses ontology based approach to organize user preferences which can be used for adaptation of personalized ranking function. To protect privacy in client server model proposed by author the information is restricted in the user profile and click through data is collected and stored by client locally, whereas tasks like re-ranking and concept extraction are performed at server.

The paper [4] focuses on improving effectiveness, to measure the effectiveness Zhicheng Dou et. al.[4] used 12 days of Windows Live query logs to evaluate five personalized search algorithms algorithm. Algorithms use either click based or topical interest based approach. Prior to this work the method used for evaluating the performance of personalized search is, each user issues certain number of queries and decide whether results are relevant or not. Unfortunately, there are some drawbacks in this method.

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. These user clicks are utilized as relevance judgment to evaluate search accuracy. This method can rerank most relevant documents higher in the list, so user would be more satisfied. Results show that, click-based personalization algorithms worked well. This framework is more useful for evaluating precision when experimenting with large number of queries.

A novel technique is proposed by Fang Liu et.al [5] to improve PWS for retrieval effectiveness it gather user profiles from users' search histories and for better retrieval effectiveness in web searching. Two profiles are maintained namely *user profile* and *general profile*. User profile is maintained by each and every user by themselves while general profile uses i.e. "one profile fits for all" terminology. These two profiles are combined and web search is conducted based on both queries, proves effective and efficient. This paper works in two steps, first step is to

categories a user queries by mapping into set of categories, and second step is to utilize both the query and its context to retrieve Web pages. Then tree model structure is proposed to represent users search history.

4. Comparative Study

The background work discussed in section 3 has some advantages and disadvantages. The UPS proposed in [1] to protect user privacy in personalized web search which supports runtime profiling but it assumes that queries does not contain any sensitive information. To overcome this we need to design a framework which can clearly differentiate between sensitive and generalized user profile.

The paper [8] uses search knowledge created by search communities as the basis of relevance model i.e. the queries submitted and the results they selected by particular community but it will not suite for all communities. As compare to community based search the method proposed in [9] User profile is maintained by each and every user by gathering user profiles from user's search histories and used for better retrieval effectiveness in Web searching. But it issues privacy problems.

To protect user profile from altering, the paper [2] generates fake queries with authentic one which gives extra privacy and does not require any changes at server site. The drawback of this scheme is that because of unproven abstraction layers it introduce security risk and more skills are required.

There are many algorithms proposed for retrieval effectiveness, person level reranking[10], BuildUP[8]. Paper[2] also proposes some algorithms but have many limitations like they work only for repeated queries and does not prohibit privacy issues and location based results.

To improve the quality of search results and the location based results PSME proposed in [3], which uses the GPS. The GPS location helps to improve retrieval effectiveness. It represents different content in different ontologies. The privacy is maintained by allowing user to control information exposed to server. the limitation is that It tries to minimize user involvement which results in synthesizing user queries from given queries which can result in different output because of different search behaviors.

5. Conclusion

This paper provides a review on personalized web search and the related security concepts. The PWS techniques are developed remarkably in the last decades. A variety of techniques have emerged to increase search effectiveness and to protect privacy using multiple algorithms. Different methods conclude that privacy preservation is not handled well. UPS framework which is proposed to provide privacy for each user, uses the online profiler to take online decision on whether to personalize a query or not. This framework can significantly reduce the risk of attack and performs better as compared to others.

Refrences

[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] 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

[5] 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

[6] Xinyu Xing, Wei Meng, Dan Doozan, Alex C. Snoeren†, Nick Feamster, and Wenke Lee "Take This Personally: Pollution Attacks on Personalized Services"Georgia Institute of Technology and UC San Diego

[7] Claude Castelluccia1, Emiliano De Cristofaro2, Daniele Perito1 "Private Information Disclosure from Web Searches(The case of Google Web History)" Information and Computer Science, University of California, Irvine

[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] 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

[11] S. Yu, T. Thapngam, S. Wei, and W. Zhou, "Efficient web browsing with perfect anonymity using page prefetching," in Algorithms and Architectures for Parallel Processing (C.-H. Hsu, L. Yang, J. Park, and S.-S. Yeo, eds.), vol. 6081 of Lecture Notes in Computer Science, pp. 1{12, Springer Berlin Heidelberg, 2010.

[12] W. Ogata and K. Kurosawa, "Oblivious keyword search," Journal of complexity, vol. 20, no. 2, pp. 356{371, 2004

[13] .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

Author Profile

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, 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.