International Journal of Engineering Innovation & Research
Volume 5, Issue 3, ISSN: 2277 – 5668

Privacy Protection for Personalized Web Search

Manjula H S **Basamma Patil** **Srinidhi K S** **Pooja B**

***Abstract* – In web search engines privacy protection has become more serious now a days. The main problem of privacy protection in web search is discussed, with a special focus on IP-address based personalized web search. The main goal is to break the linkage between users' identities and their issued queries so as to prevent privacy breaches. It provides a strong privacy guarantee in web search. The main idea of this privacy model is to protect user's search activities within a social peer group. Social peer group contains a set of individual users. From search engines perspective, search queries that are issued by users from the same peer group cannot be linked uniquely to individuals within the same group. Experimental results show that our methods achieve high efficiency in practice.**

***Keywords* – Privacy, Personalized Web Search, Social, Peer Group.**

I. INTRODUCTION

Web search engines have become an indispensable component for millions of users to search desired information on the web. They gather tremendous amounts of users' personal information. Although such information can be used to provide personalized web search which improves the accuracy of search results greatly, the intensive usage of users' personal information in web search engines also raises terrifying privacy threats to users.

Information related to users' search activities such as IP address, search queries, and click-through data are all captured and maintained by web search engines using search logs. A few real-life examples indicate that detailed user profiles are constructed from search logs [1].

Privacy breach in web search engines has introduced more threats to individuals. There is an extremely high demand of effective privacy protection mechanisms in web search. There are two major questions related to privacy breach in web search. Firstly, who issued the search query? Secondly, what is the search query about? From the individual's point of view, if the answers to each of the two questions are identified by some malicious attackers, there is no big privacy concern. However, if a strong linkage between a user who issued the search query and the content of the search query is uniquely identified, the user's search activity is undoubtedly under risk.

The linkage between a user and his/her search query should be well protected. If even the search engine companies could not correctly recover the linkage, user's privacy is strongly protected. In this paper, we focus on breaking the linkage between users' identities and their issued queries so as to prevent private information to be disclosed by any parties. Several existing studies focus on hiding the true IP address of the users who issued a query. These methods cannot provide personalized searches which require the original IP address.. Once the true IP address is completely anonymized, the personalized search result cannot be returned. Thus, can we develop techniques to provide strong privacy protection guarantees for search engine users without compromising the personalized search performance?

The major contribution of this work is a novel privacy framework with guaranteed privacy protection in IP address-based personalized web search. The main idea of privacy framework is to protect user's search activities within a social peer group. A peer group represents a social group of individuals who share similarities. The queries from the similar peer group will be submitted to web search engines together. Users from the same peer group cannot be linked to individual users within the group.. This framework consists of an online peer grouping step that dynamically constructs a peer group for each user, and an information obfuscation step which protects each individual user in the crowd. Also provides a practical privacy model that will share similar characteristics of *l*-diversity in privacy preserving data publishing of relational data to provide a strong privacy guarantee in personalized web search.

The rest of the paper contains: - Review on some related studies is provided in Section 2.Privacy protection framework for personalized web search is presented in Section 3. A practical privacy model with strong privacy protection guarantee is also discussed in this section. In Section 4, some strategies to efficiently formalize peer groups which serve as core foundations of the proposed privacy protection framework is discussed. A systematic empirical study conducted on the AOL search log data set is reported in Section 5. Section 6 concludes the paper.

II. RELATED WORK

Privacy has become a more serious concern in many applications. One of the privacy related problems is publishing relational data for public use [2], which has been extensively studied in the recent years. The major objective of privacy preserving data publishing research is to hide sensitive knowledge from the data while maintaining the utility of data for various data analysis tasks [3]. Several privacy models, such as *k*-anonymity [4], *l*-diversity [5], and their variations have been proposed for the purpose of privacy protection. Other than relational data, some other types of data such as social networks and search log data also suffer from privacy breaching concerns. Recently, *k*-anonymity and *l*-diversity [6] have been successfully extended to address privacy issues in social networks [7, 8] and search logs.

The Private Information Retrieval model [9] is considered to be the perfect private solution to address the privacy breaching issues in web search. Due to its high complexity and the inability of personalized search,

Private Information Retrieval does not have practical usage. Some recent studies [10] try to obfuscate the search query itself. Randomly generated keywords are injected into the actual query to hide the real search intent. These methods rely on a thesaurus for generating queries which is not practical in the web search scenario. In addition, linkages between users' identities and their queries are maintained in the search logs, which in-fact still poses great privacy threats to associated individuals.

The proposed privacy model is based on the concept of peer groups. Peer group, which represents a social group of individuals who share similarities, has been an important concept in social science research. The analysis of peer groups has been applied in many areas, such as stock analysis [11], collaborative information sharing [12], distributed computing and cyber network structure [13, 14]. However, the analysis of peer group has not been used for the purpose of privacy protection in web search.

III. Privacy Protection Framework and Privacy Model

In this section, the framework for protecting user's privacy in personalized web search is first discussed. Then, a practical privacy model with strong privacy protection guarantee in the web search scenario is discussed.

A. The Framework of Privacy Protection in Web

A web search activity usually involves interactions between a user (client) and a web search engine (server). Our methods address the problem of privacy preserving web search at the client side by formalizing a peer group for each web user.

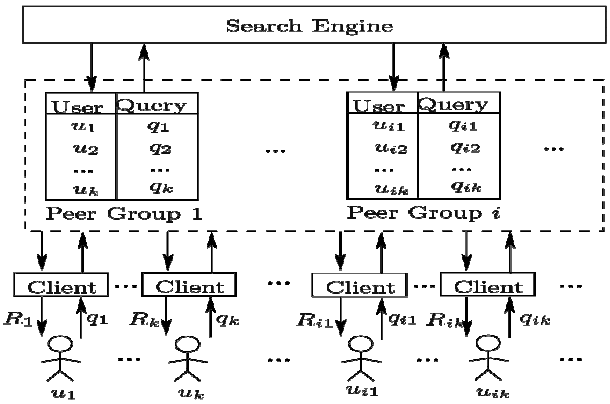

Fig. 1. The framework of privacy protection in personalized web search (ui: user; qi: query issued by ui; Ri: ranked result for qi)

Figure 1 presents the framework of our methods. While each user issues their own query as usual from their clients (e.g., web browsers), an automatic online grouping protocol will be applied to cluster users into peer groups. The queries from the same peer group will be submitted to web search engines together. The grouping protocol can be achieved by a software plug-in designed specifically for those web browsers.

While this framework does not completely hide users' actual search queries, however, from the search engine's point of view, it is equally plausible that an individual issues one of the queries in the same peer group. As a result, even the search engine cannot correctly infer which user issued what search query with 100% accuracy. Thus, the user's privacy in web search is well protected.

B. The Privacy Model in Web Search

According to the *l*-diversity model for relational data [9], if queries from the same peer group have the same search intent, a linkage is still able to be constructed between a user and search intent. Thus, the privacy model is defined by considering the diversity of queries.

Definition 1 (*l*-Diversity Search Privacy). *A user u*1 *issuing a query q*1 *has l-Diversity Search Privacy, if*

*1. The peer group G of u*1 *has at least l* − 1 *other distinct users, denoted as*

$$G = \{u1, u2, \ldots, ul, \ldots\};$$

2. Let I(*G*) *be the most frequent search intent of queries in G, thus, I*(*G*) /|*G*| ≤ 1/ *l ;*

*3. u*1 *only appears in one peer group G at any time.*

In general, *l*-Diversity Search Privacy has a similar property of *l*-diversity. If a user *u* satisfies *l*-Diversity Search Privacy, search engines could not determine a linkage between *u*'s identity and *u*'s search intent with a confidence higher than 1/*l*. The larger the value of *l*, the stronger the privacy guarantee.

C. Discussions on the Privacy Framework

As the proposed privacy framework has an information obfuscation step to break the linkage between a user's identity and his/her search queries, it is necessary to consider whether this would affect the quality of web search performance. On the search engine side, each time it receives a group of user identities and search queries. If the size of a group is *l*, there exist *l*2 different combinations of user identities and search queries. To ensure that the actual personalized search results are always generated, search engines need to conduct searches for all of these *l*2 combinations. On the client side, the plug-in will only present to the users the personalized search results of the original IP address. Therefore, the quality of personalized search is not affected at all. It is interesting to see if a balance between the overheads on the search engine side and the search quality can be achieved. We leave this as a future research direction [15].

In the *l*-Diversity Search Privacy model, the grouping protocol knows all the search queries and their corresponding IP address mapping. The grouping protocol should be robust and reliable. Any mapping information should not be leaked. As an interesting future research direction, we plan to investigate practical security and encryption-based technique to enhance the security of the grouping protocol.

IV. Online Construction of Peer Groups

As the formalization of peer groups serves as core foundations to protect individual's privacy in web search, an online formalization of peer groups is a necessity. Not surprisingly, in the current information era, millions of users are issuing queries to search engines at any time. An online grouping procedure needs to be conducted to form peer groups instantly. The details of the algorithms will be discussed in this section. In practice, users may stop issuing queries, and new users will start to issue queries. Thus, when a user does not satisfy the l -Diversity Search Privacy anymore, a reconstruction of peer groups is triggered automatically.

To construct peer groups online, we model users' search activities as a sequence of (*ui, qi*) pairs, where *ui* is user's identification (e.g., IP address) and *qi* is a query. The peer group construction problem then becomes a sequence partitioning problem such that each partition should satisfy the privacy requirements in Definition 1.

Algorithm 1. The GreedyAdd algorithm
Input: a stream of users' search queries S = {(u1, q1), (u2, q2), . . , (ui, qi), . . .}
Output: a user group *G*;
1: let $G = \{u1\}$;
2: let *pointer* = 2;
3: **while** $|G| < l$ **do**
4: let *count* = 0;
5: **for** each $u \in G$ **do**
6: **if** $Sim(q, qpointer) > \delta$ **then**
7: *count* = *count* + 1;
8: **end if**
9: **end for**
10: **if** *count* = 0 **then**
11: let $G = G \cup \{upointer\}$;
12: let *pointer* = *pointer* + 1;
13: **end if**
14: **end while**
15: update *S*;
16: **return** *G*;

To determine whether two queries have different search intents, a straightforward solution is to calculate a similarity score between them. We adopt a similarity measure based on the Vector Space model due to its popularity. That is, each query is regarded as a term vector. A cosine similarity is calculated to measure the similarity between two queries. We develop a greedy solution to construct partitions from a sequence of (*ui, qi*) pairs. The major idea is to consider a variant of a traditional clustering problem: suppose *n* distinct users are issuing their own queries, we need to generate clusters of users (and their queries) $G = \{G1, G2, \ldots\}$ such that:

1. $\forall Gi \in G$, the size of *Gi*, denoted as $|Gi|$, satisfies $|Gi| \geq l$;
2. $\forall uj, uk \in Gi$, the similarity score of queries *qj, qk* issued by *uj, uk* (denoted as $Sim(qj, qk)$) satisfies $Sim(qj, qk) \leq \delta$, where δ is a parameter that determines whether two queries have different search intents.

The above problem is a variant of the *k-Gather Clustering* problem [1], which is NP-hard. However, in the web search scenario, the optimal solution is not necessary. In addition, millions of users may issue queries at the same time and new users and new queries will be issued continually. Taking the efficiency requirement into consideration, we develop the algorithm called GreedyAdd.

The details of the GreedyAdd algorithm are summarized in Algorithm 1. The algorithm starts by picking the top-1 user in the sequence of (*u, q*) pairs. Then, it scans the remaining sequence and keeps adding users who Once a peer group of l users is formalized, the queries are submitted to search engines together.

V. Experimental Results

We conduct some experiments using the well-known publicly released AOL search log data. The data set contains about 650,000 users over a 3-month period. We adopted this search log data for the simulation of users issuing queries to web search engines. Only user ids and their search queries are considered in the simulation experiment.

As discussed in Section 4, the quality of personalized web search is not affected at all using our proposed privacy protection framework. For the purpose of evaluation, one important efficiency measure we considered is the time delay for constructing the peer groups. Since a group of users and their queries are submitted together, some users who issued queries earlier may have to wait until the group is formed. To quantitatively evaluate this time delay, we use $p(ui, qi)$, the position of pair (*ui, qi*) in the sequence, as the time when *ui* issued a query *qi*. The largest position of a pair in the peer group *Gj* is denoted as *pGj*. Thus, the measure of time delay for *Gj* can be calculated as

$$Delay = \frac{\sum_{(ui,qi) \in Gj} \left| p(ui, qi) - pGj \right|}{|Gj|}$$

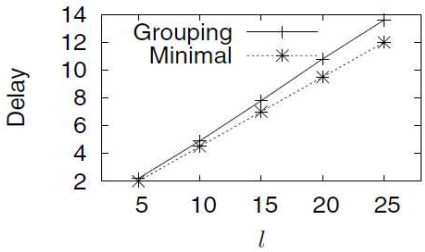

Fig. 2. The time delay in the simulation experiment

Figure 2 shows the average time delay for all the users in the AOL search log data. The X-axis represents the value of l – the size of peer groups, and the Y-axis represents the delay as calculated using Equation 1. For comparison, the calculation of the minimal time delay is done which refers to the case that each peer group contains a continuous set of (*ui, qi*) pairs in the sequence.

In general, the time delay due to grouping is quite small. When the value of *l* increases, the time delay increases as well. This is because the size of peer groups increases. Considering the fact that millions of users are issuing queries within a very short time period, the actual time delay for constructing peer groups in practice can be neglected.

VI. Conclusion

In this paper, we proposed a practical privacy model for protecting user's privacy in personalized web search. The general idea is to hide individual's search activities in a social crowd. Thus, the linkages between user's identity and user's queries are disconnected.

There are several interesting future directions for our work, such as (1) how to extend the proposed privacy model to prevent privacy breaches which utilize individual's sequential search activities; (2) how to integrate users' click-through data to enhance the privacy model in web search. In addition, the proposed peer group formalization algorithm only considers whether user's queries share different search intents, it does not take into account whether users sharing similar social background should be grouped with high priority. We are also interested in exploring social profiles of users for more effective formalization of peer groups.

References

[1] Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., Zhu, A.: Achieving anonymity via clustering. In: Proceedings of the 25th ACM SIGMOD-SIGACT SIGART Symposium on Principles of Database Systems (PODS 2006), pp. 153–162. ACM, New York (2006)

[2] Blundo, C.: Private information retrieval. In: Encyclopedia of Cryptography and Security, 2nd edn., pp. 974–976 (2011)

[3] Bornhorst, N., Pesavento, M., Gershman, A.B.: Distributed beamforming for multiuser peer-to-peer and multi-group multicasting relay networks. In: ICASSP, pp. 2800–2803 (2011)

[4] Gkoulalas-Divanis, A., Verykios, V.S.: Hiding sensitive knowledge without side effects. Knowl. Inf. Syst. 20(3), 263 299 (2009)

[5] Jones, R.: Privacy in web search query log mining. In: Buntine, W., Grobelnik, M., Mladeni´c, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009, Part I. LNCS (LNAI), vol. 5781, p. 4. Springer, Heidelberg (2009)

[6] Kim, Y., Sohn, S.Y.: Stock fraud detection using peer group analysis. Expert Syst. Appl. 39(10), 8986–8992 (2012)

[7] Liu, K., Terzi, E.: Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD 2008), pp. 93–106. ACM Press, New York (2008)

[8] Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: Privacy beyond k anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering (ICDE 2006). IEEE Computer Society, Washington, DC (2006)

[9] Murugesan, M., Clifton, C.: Providing privacy through plausibly deniable search. In: Proceedings of the SIAM International Conference on Data Mining (SDM2009), pp. 768–779. SIAM (2009)

[10] Pang, H., Ding, X., Xiao, X.: Embellishing text search queries to protect user privacy. PVLDB 3(1), 598–607 (2010)

[11] Samarati, P.: Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE) 13(6), 1010–1027 (2001)

[12] Shtykh, R.Y., Zhang, G., Jin, Q.: Peer-to-peer solution to support group collaboration and information sharing. Int. J. Pervasive Computing and Communications 1(3), 187–198 (2005)

[13] Sweeney, L.: K-anonymity: a model for protecting privacy. International Journal on uncertainty, Fuzziness and Knowledge-based System 10(5), 557–570 (2002)

[14] Tsuneizumi, I., Aikebaier, A., Enokido, T., Takizawa, M.: A scalable peer-to-peer group communication protocol. In: AINA, pp. 268–275 (2010)

[15] Zhou, B., Pei, J.: Preserving privacy in social networks against neighborhood attacks. In: Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008), pp. 506–515. IEEE Computer Society, Cancun (2008)

Authors' Profiles

Mrs. Manjula H. S., received her M. Tech in computer networks from RV College of engineering, Bengaluru, Karnataka in 2013. She received her bachelor's degree in computer science and engineering from SIT, Tumkur, Karnataka in 2008.
She is currently working as an Assistant Professor in CSE Dept., SJB Institute of Technology, Bengaluru, Karnataka. Her areas of research include computer networks, information and network security.
Email: manjulahs.sit@gmail.com.

Mrs. Basamma Patil, received her M. Tech in computer science and engineering from M. S. Engineering College, Bengaluru, Karnataka in 2014. She received her bachelor's degree in information science from BEC, Bagalkot, Karnataka in 2009.
She is currently working as an Assistant Professor in CSE Dept. SJB Institute of Technology, Bengaluru, Karnataka. Her areas of research include networking, data mining and cloud computing.
Email: bupatil25@gmail.com.

Mrs. Srinidhi K.S., received her M. Tech in computer science and engineering from P.E.S. College of engineering, Mandya, Karnataka in 2014. She received her bachelor's degree in computer science and engineering from P.E.S. College of engineering, Mandya, Karnataka in 2009.
She is currently working as an Assistant Professor in CSE Dept. SJB Institute of Technology, Bengaluru, Karnataka. Her areas of research include data mining and artificial intelligence.
Email: srinidhi.ks27@gmail.com.

Miss. Pooja B., is currently a final year student of SJB Institute of Technology, Bengaluru, Karnataka. She is doing her bachelor's degree in computer science and engineering.
Her areas of research include computer networks, data mining and artificial intelligence.
Email: poojab546@gmail.com.