Disclosure of User's Profile in Personalized Search for Enhanced Privacy

MANOJ KUMAR. K
Department of Computer Science and Engineering,
Sathyabama University,
Chennai, Tamil Nadu 600119, India
kandalamanojkumar@gmail.com

Abstract: - Personalized Web Search (PWS) is a search technique for providing better search results, viewing user history, and has been enticing much responsiveness recently. However, effective personalized search requires gathering and accumulating user data, which often raise severe concerns of privacy intrusion for many users. These concerns have become one of the key barriers for organizing personalized search applications, and how to do privacy preserving search is a prodigious challenge. Indeed, evidences show that users' unwillingness to disclose their private data throughout search has become a major obstacle for the PWS. Sensitive User data is unprotected in the database records. In this paper, examining the issue of privacy preservation in personalized search and representing stages of privacy preservation. Encoding Algorithm is used for encoding the users' profile data & it is mapped by unique alphanumeric identity using random unique identity generator algorithm. User identity to server is fully disclosed and privacy is preserved in web search without negotiating both personalization and privacy.

Key-words: web search, encoded user profile, Personalization, Random unique identity, Privacy preservation

1. Introduction

Search engines have been effectively organized to serve user's information needs, they are far from optimal. A major absence of current search engines is that they follow the model of "one size fits all" and are not adaptive to distinct users. This causes intrinsic non-optimality as is seen clearly in the following two cases: (1) Diverse users may use exactly the same query (e.g., "Apple") to search for different information (e.g., Apple Fruit or the Apple Company), but existing search engines return the same results for these users. (2) A user's information needs will alter over time. The same user may use "Apple" sometimes to mean the Fruit Apple and occasionally to mean the Company. Current web search engines are not capable to distinguish such a scenario. Evidently, without using user search history and/or the search circumstance of a user it is difficult for a web search engine to understand which intellect "Apple" refers to in a search query. In directive to enhance search exactitude, must use user search queries and personalize search results affording to each individual user [10]. In broad, personalized search is considered as one of the most favorable techniques to break the constraint of present search engines and advance the search results

Even though the desirability of personalized search, does not seen large scale uses of user data. But, users are not contented with the lack of preservation of user data privacy [8, 11]. Google, AOL, Bing for example, has deployed a personalized search by using user private data and exposing in the server. There is an intrinsic rigidity amid providing personalized web search and preservation of privacy since personalized search entails collecting a lot of user history. Explicitly, in mandate to personalize queries, a user data must be assembled to precisely build a user's information requirements. Web search engine requires accurate user search history & data which rises the threat of privacy infringement. Regrettably, such discreetly collected personal user data can easily expose an extent of user's private life [1]. Privacy concerns intensifying from the lack of privacy protection for such data, for occasion the AOL user history scandal [11], not only increase dread among individual users, and also lessen the data providers' eagerness in contribution to personalized service. Entire User private data is exposed in server and search engine providers are explicitly using users' data for their profits. By this entire user private data is exposed. For preventing it, different altitudes of privacy preservation steps of framework is taken further through this paper.

2. Related Work

The existing system of personalized web search is having runtime profiling support and user decision to personalize or not. In the current system, entire user profile is stored with entire expose to search engine server. Online user decision to personalize the each query is available. A user profile is usually generalized for under once offline, and accustomed alter altogether

keywords from an identical user. Such an approach actually has drawbacks given the variability of queries. One proof rumored in is that profile-based personalization might not even facilitate to boost the search quality for a few impromptu queries, however revealing user profile to database and violating user's privacy. The existing strategies don't take under consideration the customization of privacy needs [4]. This most likely makes some user privacy to be overprotected whereas others insufficiently protected. For instance, in, all the sensitive topics square measure detected victimization associate degree absolute metric referred to as perturbation supported the knowledge theory, presumptuous that the interests with fewer user support of document square measure additional sensitive [7]. Sadly, very little previous work will effectively address individual privacy wants throughout the generalization.

Many personalization techniques need reiterative user interactions once making personalized search results. They typically refine the search results with some metrics that need multiple user interactions, like rank marking, average rank, and so on. Entire user profile is exposed to server.

2.1. Runtime Profiling

Previous works on profile-based PWS in the main concentrate on improving the search utility. The essential plan of those work is to alter the search consequences by affecting to, often indirectly, a user profile which exposes a private information goal. Within the remainder of this section, by reviewing the previous solutions to PWS on 2 aspects, namely the illustration of profiles, and also the live the efficiency of personalization. Numerous profile depictions are offered within the literature to facilitate totally different personalization ways. Previous systems exploit tenure of lists [5] or basket of words [2] to characterize their profile. However, utmost up-to-date works build profiles in ranked structures owing to their stronger descriptive ability, higher tendency, and higher access potency. The bulk of the ranked representations are created with existing weighted topic hierarchy/graph, like ODP1 [1], [10], [12], [5], Wikipedia, and so on. Another add [10] builds the hierarchical profile mechanically via term-frequency analysis on the user information. In this planned framework, do not focus on the enactment of the user summaries. Essentially, this context will doubtless adopt any ranked representation supported a taxonomy of data. As for the performance measures of PWS within the literature, Normalized Discounted accumulative Gain could be a common live of the effectiveness of data retrieval system. It's reinforced a social grouped significance measure of item sites contained by the result forms, and is, therefore, illustrious for its high price in express feedback collection. Having a tendency to use the typical exactitude metric, planned by Dou et al. [1], to live the efficiency of the personalization in this context. In the meantime, this effort is illustrious from preceding revisions since it conjointly proposes 2 prognostic metrics, particularly personalization effectiveness and disclosure threat, on a summary occurrence deprived of wishing for user response .In the existing approach, the framework is used for runtime profiling, online user decision for each query to whether personalize or not. In which, the user has to take decision for each and every time to while giving queries to search engine whether to store that particular query in server or not. If any user sensitive data exists a metric used to identify and it won't store in server. They proposed it using two greedy algorithms.

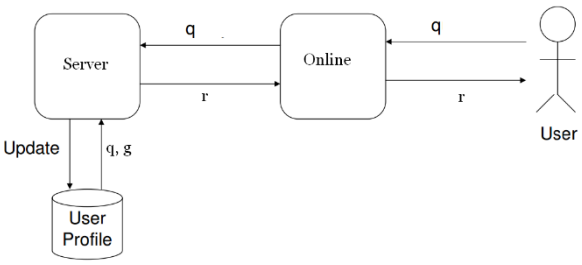

Fig. 1. Server with Exposed User Profile.

2.2. Preservation of Privacy

Generally there are unit 2 categories of privacy protection problems for PWS. One category includes these indulgence privacy as the empathy of a isolated, as defined in [2]. The other includes those think about the sensitivity of the information, chiefly the user summaries, visible to the server. Representative mechanisms within the literature of protective user identifications (class one) try and solve the privacy drawback on totally different levels, together with the unique identity, the cluster identity, no distinctiveness, and personal information exclusion [3]. Resolution to the first level is proven to fragile [11]. The third and fourth levels area unit impractical attributable to high value in communication and cryptography. In this stages of privacy, first and second levels are privileged for the user privacy in personalized web search. Therefore, the prevailing efforts concentrate on the first and second level. Mistreatment in this method, the association among the keyword and single user is fragmented. The useless user profile (UUP) protocol [8] is planned to shuffle queries among a multitude of users concern by them. Indeed the result of any article cannot outline an explicit individual. Viejo and Castell-Roca use heritage social networks instead of the third party to produce a distorted user profile to the online program. Within the theme, every user acts as a pursuit activity of its neighbors. They can select to characterize the request on behalf of issuing of it, otherwise onward it to dissimilar neighbors. The shortcomings of current solutions at school one is that the high value introduced attributable to the collaboration and communication. In UPS

framework, they implemented user decision to store query in server or not. Anyway, in that approach also entire user private information is exposed in server. Through the usage of that private user data, there is lot of advantages for web search providers commercially by using user private life [9]. There are lot of criticisms on search providers like google, yahoo, Bing, AOL regarding user privacy in web search and usage of user private information.

2.3. Drawbacks

The existing runtime profiling and customized internet search is a typically approach which doesn't resolve the entire problem related to privacy. Each and every time user decision is to be taken whether to store the query or not in server. In some case [8], user may accidently stores his sensitive information. Anyway entire user profile is revealed in the database records. Eavesdropping attacks can easily occurred, by invading the server and capturing entire user data. Exposure of user's identity with mapped user search histories. Usage of user private data by database administrators for research and development and commercial purpose for their own benefits by search engine providers. All the sensitive topics square measure detected mistreatment an absolute metric known as disruption supported the data theory. Indeed, the level of privacy protection is not attained in the current personalized privacy approach.

3. Proposed System

In the proposed system, for avoiding the exposure of user identity in existing system while storing user history, the keywords submitted by user in database are mapped by unique alpha numeric identity also called as pseudo identity[3] for each user & group of users. During the user registration, after submitting user profile data proximately a unique identity for that particular user is generated using random unique identity generator algorithm and also entire user profile information is encoded using the base encoding algorithm. Through this proposed system, database analysts can track that particular user & there is no loss for search engine providers for their research. Personalized Web search feature is active via the unique identity of that anonymous user. Particular user history is exposed to the server for sake of personalization and viewing user history but it is mapped with the unique alpha numeric identity. So that if the man in middle attacks occurs also they cannot identify the user profile and they cannot decode. Two algorithms used in the proposed system are random unique identity generator algorithm, Base encoding algorithm and also generally greedy algorithm used for searching [1]. By this proposed approach, search engine providers' commercial usage of user data with their exposed profile can be mostly reduced, In the Fig. 2. A user submits a query q after user registration to the search engine in middle online profiler checks the encoded user profile and updates the query in user profile and it is mapped by pseudo identity while viewing search history, then sends the requested query q and generalized profile g to search engine server after that search results r will be displayed to the user.

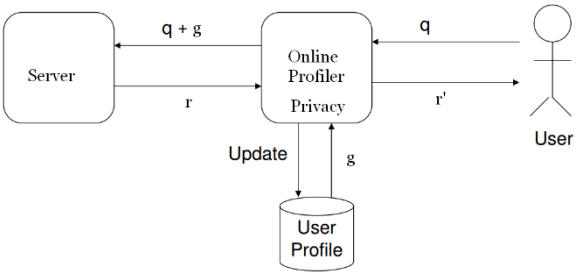

Fig. 2. Server with user profile mapped by unique identity

3.1. User Profile Encoding Algorithm

User Profile Encoding Algorithm is used as entire user data encoded in supplementary format mechanism.

Description

User profile encoding algorithm is considered to convert any binary records, a cluster of bytes, into a stream of 128 printable types. There are character set map by the user profile encoding.

The User profile encoding process is takes place in following steps:

Step 1 - Split the input bytes stream into chunks of 3 bytes. **Step 2** – Split 24 bits of each 3-byte chunk into 4 sets of 6 bits. **Step 3 -** Plot each cluster of 6 bits to 1 printable character, based on the 6-bit value using the Base128 set map.
Step 4 - If the last 3-byte chunk has only 1 byte of input, pad 2 bytes of nil (\0000). Later converting it as an ordinary chunk, overrule the preceding 2 appeals with 2 equal signs (==), so the cracking procedure identifies 2 bytes of zero were prolonged. **Step 5 -** Uncertainty the previous 3-byte chunk has only 2 bytes of input records, pad 1 byte of zero. After encrypting it as a normal chunk, override the last 1 character with 1 equal signs (=), so the decoding process knows 1 byte of zero was expanded.
Step 6 - Otherwise return (\r) and new line (\n) are implanted into the output character. They will be unnoticed by the decoding approach.

Algorithm (User profile encoding)

Input: Entire user data submitted through forms while registration

Output: Encoded user information in server, user data is disclosed.

```
String encryptedString = null;   //Encoding
byte[] encodedBytes =
Base128.encodeBase128(unencryptedString.getBytes());
encryptedString = new String(encodedBytes);
String decryptedText=null;      //Decoding
byte[] encodedBytes =encryptedString.getBytes();
decryptedText = new
String(Base64.decodeBase64(encodedBytes));
```

Summary

The User Profile Encoding algorithm is simple to implement and using this mechanism while submitting user data to database servers, it transforms the entire user profile into encoded format and that transformed user data is stored in the servers. By this approach, user identity and private information is disclosed. Some non-sensitive information can be exposed. In this way the entire mechanism takes place in the proposed system.

3.2. Random Unique Identity Generator Algorithm

Random unique identity generator algorithm (RUIG) is used for generating a unique alphanumeric identity for each user and it is mapped to the particular user profile.

Description

This Algorithm is designed to generate a unique alphanumeric identity for individuals and it is mapped to that particular user profile. So, that this random identity plays a key role for recognition of that user profile in the server. Anyway the entire user profile is encoded so by this unique identifier it will mapped to the all user activities.

Algorithm (RUIG)

Output: Unique Alphanumeric identity generated.

```
RUIG  r =  new RUIG

return r.nextInt(100000000); //Generating Random
unique number

String unique_id = input[0].substring(1,4)+getId();
```

Summary

RUIG is a simple algorithm for generating random unique identifiers for individuals while submitting entire user data. The user profile is clustered and it is mapped by this generated identity. This random identifier is mapped for all user activities instead of user profile. In the algorithm, extracting the substring from user name itself for alpha and the remaining numbers are generated uniquely without any similarity.

4. Implementation

By the combination of two mechanisms, additionally another greedy approach implementing the whole framework on search engines.

Eavesdropping Issue

In which there are few implementation issues regarding eavesdropping attacks for capturing user data and threat of decrypting the high secure algorithm [6]. These type of implementation issues can be recovered by the proposed algorithms. The encoding and decoding cost also reduced.

Modules Depiction

In the implementation part, there are mainly three modules They are user registration for submitting user data, user login for validating credentials and database administrator from server side where there is no expose of user private data. While user submitting data for registration, at first user profile encoding algorithm is started executing in parallel for encoding the user profile and instantly random alphanumeric identifier is generated and mapped with the encoded user profile. In the database administrator module, only encoded user profile data and unique identifier is visible, everything is disclosed. While user login, it validates credentials by decoding the user profile. After user successful sign in, indeed user giving any query to the search engine. Then the search results will be displayed by updating the user search history and it is mapped with identifier. The user's search history is stored in the server for personalization feature but that user submitted keywords and visited URLS are mapped with unique identifier. By this there is no scope of identity of particular user profile. After User Login, they can access search engine and viewing their user history. But the particular user identity is disclosed in database records by encoding algorithm and mapped by random unique identifier.

From this proposed system, user entire profile will be disclosed. Enhances the stability of search. The main aim to attain level of privacy protection is achieved by balancing both personalization search and privacy. Mistrust of user's data is became benefit for search engine provider commercially in

existing system. Spam messages, commercial advertisements and usage of user profile for their purpose can be reduced by this implemented proposed system. Greedy algorithm is used for the search ability and hierarchy of user profile for the personalization of the search results using particular user search history.

5. Experimental Results

In this section, presenting the experimental results of implemented proposed system. Different experiments are conducted on this proposed system. At first, there are thorough consequences of the metrics in each repetition of the proposed procedures. Secondly, looking at the effectiveness of the proposed user profile encoding algorithm, random unique identity generator and also in terms of responsiveness and search quality.

5.1. Experimental Structure

The Proposed System is implemented on a PC with a Core i7 2.90GHz Processor and 6 GB primary memory with an operating system Microsoft Windows 8.1. All the algorithms are implemented in java.

According to AOL data leak Scandal, which is recently published in different hosting sites, it is accessible. AOL query logs comprise over 1.5 crore keywords and 2.5 crore clicks of 6 lakh users over 90 days period. The format of data which appeared in their servers for database administrators are

{Email id, query, clicked url, time}

Where the first 2 fields indicate the email id of individual user and his delivered query to search engine. Third field indicates the clicked urls for that particular query by that particular user at timestamp time.

By this scam, the entire search engine providers are more focused on privacy of users. It attain the privacy preservation without revealing user's private life. For this already discussed some proposed algorithms in earlier sections. These algorithms are implemented experimentally using proposed privacy techniques.

5.2. Results of the Proposed Approach

In this experimental results section, by the implementation of proposed algorithms the encoded user profile and random unique identity generator are executed for enhanced privacy. The data format of user search history which is visible for the server side administrator are

{Unique id, query, Clicked urls, time}

Where the major difference between by the existing and proposed system is unique identity field. Which will mapped to that user's search history. Indeed, by the help of unique id also the database administrator and researchers can't get the user profile because the entire user profile is encoded using the user profile encoding algorithm. That encoded user profile is mapped with the unique identity and in user search history records also individual user history is mapped with the already assigned identity.

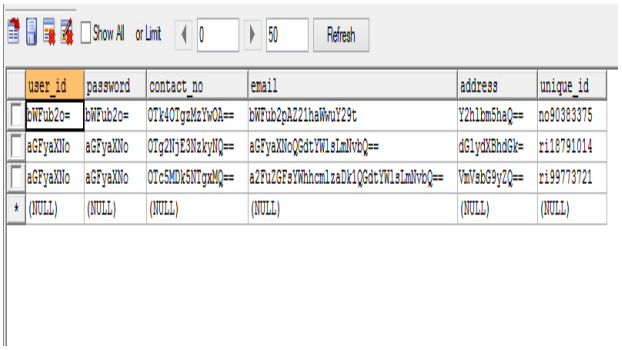

Fig. 3. Random Unique identity generated in the database and it is mapped with encoded user profile after issuing user data.

By undertaking different iterations of user data for the analysis of privacy concerns, effectiveness and search quality in the proposed system. In Fig. 3 after submitting the user data while registration, the entire user profile is encoded by implementing the user profile encoding algorithm through different character set mapping technique. As well as the entire encoded profile is mapped with one random alphanumeric identity which is generated runtime while submitting user data by using user id substring and random unique numbers, the identity will be generated. The effectiveness of the issuing data is determined. There is extremely good responsiveness.

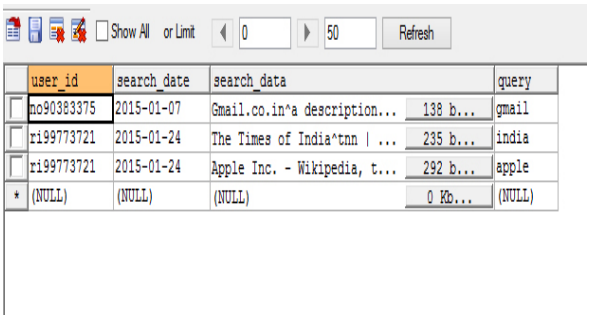

Fig.4. View of User Search History mapped with random unique identity in database

As Fig.4 shows that the preview of different user search histories mapped with their already assigned random

unique identities. The data format is similar as mentioned earlier. It is tested on different search engines and this framework is extremely impressive for attaining the new level of privacy of users in web search engines.

In the prospective of server administrator and researchers, there is no loss for them, they can use the user search histories which is mapped by identity by disclosing user profile. Privacy intrusive barrier is resolved by the proposed mechanism. The commercial usage of user data is reduced by this approach. By the visibility of user history only, the personalization feature can be applied using greedy techniques and generalization of the user search history for better search results.

6. Conclusion

This paper discussed a enhancement of privacy in web search by disclosing the users' identity in database. This proposed framework allowed to implement a user profile encoding approach and random unique identity generator apart from greedy technique for generalization of user history for search results improvement. Without compromising the both personalization and privacy, the new enhancement of privacy is attained. Additionally, the entire user profile encoding is done and it is mapped with random unique identity and also in user search history records. By this proposed approach, the privacy intrusion is reduced. Reduction of spams, unsubscribed mails to users for search engine commercial purpose by using user search queries. Risk of user private life is almost abridged.

For future work, there is ability to have a one more layer of privacy protection to be applied for the proposed framework. By the usage of cluster identity technique for the collection of individual identities. Through the wider background knowledge of the proposed mechanism, the cluster identity technique for additional layer of privacy will be helpful for the advancement in users' privacy in the web search engines.

References:

[1] S.L. Jany Shabu and Manoj Kumar.K, "Preserving User's Privacy in Personalized Search," International Journal of Applied Engineering Research (IJAER), Vol. 9, no. 22, pp. 16269-16276, 2014

[2] Lidan Shou, He Bai ; Ke Chen ; Gang Chen "Supporting Privacy Protection in Personalized Web Search," Volume:26, Issue:2 pp.453 – 467, 2014

[3] X. Shen, B. Tan, and C. Zhai, "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.

[4] B. Tan, X. Shen, and C. Zhai, "Mining Long-Term Search History to Improve Search Accuracy," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), 2006.

[5] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.

[6] 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.

[7] X. Shen, B. Tan, and C. Zhai, "Context-Sensitive Information Retrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[8] 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.

[9] J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turnbull, A.Edmonds, E. Adar, and T. Breuel, "Personalized Search," Comm. ACM, vol. 45, no. 9, pp. 50-55, 2002.

[10] 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.

[11] K. Hafner, "Researchers Yearn to Use AOL Logs, but They Hesitate," New York Times, Aug. 2006.

[12] A. Krause and E. Horvitz, "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662, 2010.