Privacy Protection in Personalized Web Search- A Survey

Greeshma A S.*
M.Tech Student
Dept. of CSE & Kerala University
Thiruvananthapuram

Lekshmy P. L.
Assistant Professor
Dept. of CSE & Kerala University
Thiruvananthapuram

Abstract— ***Current web search engines are built to serve all users, independent of the needs of an individual user. Personalization technologies that offer powerful tools to users that enhance their experience in varieties of search engines. 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 privacy protection is necessary. This survey investigates the several privacy preserving techniques and provides the idea about the new efficient method in the future. The main goal of this work is to assure the privacy guarantee to the user who is involved in the personalized web search. To do this several mechanism which is related to the privacy protection is investigated in this paper. Among the different methodologies that are discussed, it has been found that UPS framework is one of the efficient techniques which guarantees the user privacy and retrieves the contents as per user requirement accurately.***

Keywords— Generalization, Personalization, User profile, GreedyIL, GreedyDP

I. INTRODUCTION

Web search engines are very important in web life. Web search engines are built for all users and not specified for any individual user. Generic web search engines cannot identify the different needs of different users, if user enter improper keyword or ambiguous keywords and lack of users ability to express what they want are some challenges faced by generic web search engines. To address this issue we should personalize these results. As it is becoming an important aspect, to provide such environments, different techniques and approaches have developed. But at the same time security of personalized web searches has also gained significance, in which the user's personal or private information cannot be disclosed through web searches.

User's hesitation to disclose their private information during search has become major issue on personalization technologies. For example system that are personalize some advertisements according to physical location of user or their search history, introduces new privacy challenges that may discourage the wide adoption of personalization technologies. Personalized web search is proving its effectiveness but also raising matter of privacy and securing personal information. Many personalization methods have been exposed and been in practice. But it is not sure that those methods will make sure their efficiency in dissimilar queries for different users.

The solutions to PWS can generally be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods are straightforward; they simply impose bias to clicked pages in the user's query history. Although this strategy has been demonstrated to perform consistently and considerably well, It can only work on repeated queries from the same user, which is a strong limitation confining its applicability. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. Profile-based methods can be potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances. The two contradicting effects [4] during the search process to be considered. Improve the search quality with the personalization utility of the user profile and the need to hide the privacy contents existing in the user profile to place the privacy risk under control. This survey investigates the several privacy preserving techniques and provides idea about the new efficient method in the future. The main goal of this work is to assure the privacy guarantee to the user who is involved in the personalized web search.

II. BACKGROUND ON PERSONALIZED WEB SEARCH

There are mainly two types of personalized web search they are Click-log-based and Profile-based personalized web search.

A. Click-Log-Based Method

Here, personalization is carried out on the basis of clicks made by user. The data recorded through clicks in query logs, simulates user experience. The web pages frequently clicked by user in past for a particular query is recorded in the history and score is computed for particular web page and based on that web search results are provided. This method will perform consistent and considerably well when it is works on frequent queries. When a never asked query is entered by user; it will not provide any precise search results, which is the main drawback of this method.

B. *Profile Based Personalization*

The basic idea of these works is to tailor the search results by referring to a user profile, implicitly or explicitly which reveals an individual information goal. Many profile representations are available in the literature to facilitate different personalization techniques.

- *Lists / vectors or bag of words*: Earlier techniques utilize term lists/vectors or bag of words to represent their profile. It is the simple representation in information retrieval system. Here a text is represented as the bag of its words, disregarding grammar and even word order [3]. But it keeps multiplicity of those words. In each vector the second entry will be the count of that word.
- *Hierarchical representation:* Most recent works build user profiles in hierarchical structures. The reason is their stronger descriptive ability, better scalability, and higher access efficiency. Majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP, Wikipedia, and DMOZ and so on. Using the term-frequency analysis on the user data, the hierarchical profile can be build automatically also.

III. Privacy Protection in PWS

There are two classes of privacy protection problems for PWS in general. One class includes those works, treat privacy as the identification of an individual. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server.

A. *Identification Of An Individual*

Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudo-identity, the group identity, no identity, and no personal information [13]. Solution to the first level is proved fragile. The third and fourth levels are impractical due to high cost in communication and cryptography. So the existing efforts focus on the second level.

- Online anonymity: It works based on user profiles by generating a group profile of k users. Using this approach, the linkage between the query and a single user is broken.
- Useless user profile (UUP): This protocol is proposed to shuffle queries among a group of users who issue them. As a result any entity cannot profile a certain individual. These works assume the existence of a trustworthy third-party anonymizer, which is not readily available over the Internet all the time in large number.
- Legacy social networks**:** Instead of the third party to provide a distorted user profile to the web search engine, here every user acts as a search agency of his/her neighbors. They can decide to submit the query on behalf of who issued it, or forward it to other neighbors.

B. *Sensitivity Of Data*

The solutions in class two do not require third-party assistance or collaborations between social network entries. In these solutions, users only trust themselves and cannot tolerate the exposure of their complete profiles to an anonymity server.

(i) Statistical Techniques: To learn a probabilistic model, and then use this model to generate the near-optimal partial profile. One main limitation in this work is that it builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries. These assumptions are impractical in the context of PWS.

(ii) Generalized Profiles: Proposed a privacy protection solution for PWS based on hierarchical profiles. Using a user-specified threshold, a generalized profile is obtained in effect as a rooted sub tree of the complete profile.

C. *Issues*

The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication. The statistical methods builds the user profile as a finite set of attributes, and the probabilistic model is trained through predefined frequent queries in class two. These assumptions are impractical in the context of PWS and the generalized profile does not address the query utility, which is crucial for the service quality of PWS.

IV. Related works

Susan T. Dumais et al [3] introduces a search algorithm that considers user's prior interactions with a wide variety of content, to personalize their current web search. Rather than relying on the unrealistic assumption that people will precisely specify their intent when searching, it pursues techniques that leverage implicit information about the user's interests. This information is used to re-rank web search results within a relevance feedback framework. It explore rich models of user interests, built from both search-related information such as previously issued queries and previously visited web pages and other information about the user such as documents and email the user has read and created. The research suggests that rich representations of the user and the corpus are important for personalization but that it is possible to approximate these representations.

M. Spertta and S. Gach,[5] systematically examined the issue of privacy preservation in personalized search. The four levels of privacy protection is distinguished, and analyze various software architectures for personalized search. This work showed that client-side personalization has advantages over the existing server-side personalized search services in preserving privacy, and envision possible future strategies to fully protect user privacy.

Z. Dou, R. Song, and J. R Wen [6] studied personalization on dissimilar uncertainty queries for different users under dissimilar investigate background and present an important valuation structure for personalized search base on uncertainty logs, and then estimate five personalized search approach utilize 12-day MSN uncertainty logs. Here the consequences are examined and it is exposed that personalized search has important development over general web search on a number of query, but it also has tiny out come on some additional question. In addition, it also demonstrates that uncomplicated click-based personalization approach performs constantly and significantly well, even as profile-based ones are unbalanced in this research. Also discloses that both long-term and short-term contexts are very significant in humanizing search performance for profile- based modified search strategy.

Y. Xu, K. Wang, G. Yang proposed the notion of online anonymity [9] to enable users to issue personalized queries to an un-trusted web service while with their anonymity preserved. The challenge for providing online anonymity is dealing with unknown and dynamic web users who can get online and offline at any time. Introduces the notion of online anonymity to ensure that each query entry in the query log cannot be linked to its sender and an algorithm that achieves online anonymity through the user pool is proposed. This approach can be extended to deal with personally identifying information that may be contained in the query. The method is also applicable to general web services where there is a need to anonymize the query, with or without personalization.

In [12] J. Castelli-Roca, A. Viejo and J. Herrera presents a novel protocol Useless User Profile (UUP) protocol, specially designed to protect the users' privacy in front of web search profiling. System provides a distorted user profile to the web search engine. Also offers implementation details, computational and communication results that show that the proposed protocol improves the existing solutions in terms of query delay. The protocol also provides an affordable overhead while offering privacy benefits to the users. The proposed protocol submits standard queries to the web search engine. Thus, it does not require any change in the server side. In addition to that, this scheme does not need the server to collaborate with the user. This scheme also uses cryptographic building blocks such as Elgamal encryption, key generation, message encryption and decryption etc. for effective communication. The main idea of this scheme is that each user who wants to submit a query will not send her own query but a query of another user instead. At the same time, her query is submitted by another user. Using this approach, the web search engine cannot generate a real profile of a certain individual. The execution of queries may be delayed. The protocol assumes that, users follow the protocol correctly and no collision happens between entities, but in real it may be not the case.

Y. Zhu, L. Xiong, and C. Verdery et al [1] an optimal privacy notion to bound the prior and posterior probability of associating a user with an individual term in the anonymized user profile set is proposed. The authors proposes a novel bundling technique that clusters user profiles into groups by taking into account the semantic relationships between the terms while satisfying the privacy constraint. In this paper the problem of grouping user profiles (represented as a weighted term list) are studied, so that user privacy is sufficiently protected while the grouped profiles are still effective in enabling personalized web search. Anonymization goal is to prevent linking attacks that associate a user with an individual term in the anonymized user profile set.

A. Viejo and J. Castellia-Roca, propose a new scheme [11] designed to protect the privacy of the users from a web search engine that tries to profile them. The system uses social networks to provide a distorted user profile to the web search engine. The standard queries are submitted to the web search engine; thus it does not require any change in the server side. In this scheme, the server has no need to collaborate with the users. Delay of query execution is reduced here. Besides, the distorted profiles still allow the users to get a proper service from the web search engines. The proposed protocol preserves the privacy of the individuals who deal with a web search engine. In order to do that, it exploits the existence of neighborhoods of on-line users (social networks). In this way, a user generates queries and she can submit them directly to the WSE or she can forward them to her neighbors in the social network. The proposed system does not create groups for submitting queries. This represents a significant time reduction in comparison with other proposals in the literature. Also, anonymous channels are not used. But the proposed scheme uses a reward mechanism. Users who do not cooperate will be eliminated from the system.

These works come under class one considering, the privacy of an individual. The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication.

In [10] X. Xiao and Y. Tao**,** presented a new generalization framework based on the concept of personalized anonymity. This technique performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata. Generalization is a common approach to avoid the above problem, by transforming the Quasi-Identifier (QI) values into less specific forms so that they no longer uniquely represent individuals.

A table is k-anonymous if the QI values of each tuple are identical to those of at least k −1 other tuples. In general, k-anonymity guarantees that an individual can be associated with her/his real tuple with a probability at most 1/k. The motivations are that k-anonymity has several drawbacks. First, a k-anonymous table may allow an adversary to derive the sensitive information of an individual with 100% confidence. Second, a k-anonymous table may lose considerable information from the microdata. Third, k-anonymity does not take into account personal anonymity requirements. A novel privacy preserving technique that overcomes the above problems is proposed. The core of the solutions is the concept of personalized anonymity. A preference is formulated through a node in the taxonomy called guarding node. If null is specified underneath all the leaves, it can be published directly . Here direct protection between against the association between individuals and their sensitive values is provided. An algorithm for finding a generalized table that preserves a large amount of information in the microdata without violating any privacy constraints is also introduced. Utilizing several interesting problem characteristics, the algorithm optimizes the degrees of generalization on QI- and sensitive attributes, respectively.

In [2] Y. Xu, K. Wang, B. Zhang et al**,** presents a scalable way for users to automatically build rich user profiles. These profiles summarize a user’s interests into a hierarchical organization according to specific interests. Two parameters for specifying privacy requirements are proposed to help the user to choose the content and degree of detail of the profile information that is exposed to the search engine. A significant improvement on search quality can be achieved by only sharing some higher-level user profile information, which is potentially less sensitive than detailed personal information. The proper filtering of a user’s private information not only helps protect the user’s privacy but also may help improve the search quality. The key is distinguishing between useful information and noise, as well as striking balance between search quality and privacy protection. Offers a scalable way to automatically build a hierarchical user profile on the client side and also offers an easy way to protect and measure privacy. A search engine wrapper is developed on the server side to incorporate a partial user profile with the results returned from a search engine. Rankings from both partial user profiles and search engine results are combined. The customized results are delivered to the user by the wrapper.

J. Teevan, S.T. Dumais, and D.J. Liebling, [7] examines variability in user intent using both explicit relevance judgments and large-scale log analysis of user behavior patterns. They characterize queries using a variety of features of the query, the results returned for the query, and people's interaction history with the query. Using these features, the authors build predictive models to identify queries that can benefit from personalization. Large click entropy means many pages were clicked for the query, while small click entropy means only a few were. They explored this, using the variation in search result click-through to identify queries that can benefit from personalization. Drawing on explicit relevance judgments and large-scale log analysis of user behavior patterns, they also found that several click-based measures (click entropy and potential for personalization curves) reliably indicate when different people will find different results relevant to the same query. Here a number of additional factors are also explored, which influence these implicit measures, such as result churn, task, and result quality.

In [8] A. Krause and E. Horvitz, focused on the example of web search and formulate realistic objective functions for search efficacy and privacy. They demonstrate how to find a provably near-optimal optimization of the utility-privacy tradeoff in an efficient manner. A probabilistic model is used to generate an optimal user profile. It is evaluated on data drawn from a log of the search activity of volunteer participants. Here users preferences about privacy and utility are studied via a large-scale survey and are separately assessed, aimed at eliciting preferences about peoples willingness to trade the sharing of personal data in returns for gains in search efficiency. Also proved that a significant level of personalization can be achieved using a relatively small amount of information about users. Users become members of increasingly smaller groups of people associated with the same attributes.

In [4] Lidan Shou and Gang Chen et al, uses hierarchical user structure for modeling user interests. The system provides generalization of user profile with use of an online profiler at the client side. The system is expected to enhance the search efficiency with the personalization utility, along with the privacy protection of user profile contents. PWS framework called UPS (User Privacy Preserving Search) is introduced, which can adaptively generalize profiles by queries while respecting user specified privacy requirements is proposed. Runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. For generalization two greedy algorithms, namely GreedyDP and GreedyIL is used. An online prediction mechanism for deciding whether personalizing a query is beneficial or not ,is also proposed in this work.

V. **PROPOSED ENHANCEMENTS**

UPS framework seems to be more effective out of the methods discussed. For privacy protection, an online profiler is designed in this system, which can adaptively generalize profiles by queries while respecting user specified privacy requirements. The online profiler is at the client side where the complete user profile is stored along with the specified sensitive topics. Runtime generalization aims at providing search efficiency along with privacy protection of user profiles.

To prevent the information loss while performing runtime generalization, a greedy algorithm is used here. This work can be enhanced for complex query also. Location based search can be integrated along with the profile based personalization to retrieve faster results, based on a location without specifying the same in the query.

VI. CONCLUSIONS

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. The main goal of this work is to assure the privacy guarantee to the user who is involved in the personalized web search.

ACKNOWLEDGMENT

I am thankful to my guide Mrs. Lekshmy P L, Associate Professor of Computer Science and Engineering, for her guidance and encouragement for the paper work. I also acknowledge with grateful thanks the authors of the references and other literatures referred to in this survey.

REFERENCES

[1] Y. Zhu, L. Xiong and C. Verdery , “Anonymizing user profiles for personalized web search,” *Proc.19th Int’l conf.World Wide Web(WWW)* , pp.1225-1226, 2010.

[2] 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.

[3] J. Teevan, S.T. Dumais, and E. Horvitz, “Personalizing Search via Automated Analysis of Interests and Activities,” *Proc. 28th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, pp. 449- 456, 2005.

[4] Lidan Shou, He Bai, Ke Chen, and Gang Chen “Supporting privacy protection in personalized web search” *IEEE Transactions on knowledge and data engineering*, Vol. 26, No. 2, February 2014.

[5] M. Spertta and S. Gach, “Personalizing Search Based on User Search Histories,” *Proc. IEEE/WIC/ACM Int’l Conf. Web Intelligence (WI)*, 2005.

[6] Z. Dou, R. Song, and J.-R.Wen, “A Large-Scale Evaluation and Analysis of Personalized Search Strategies,” *Proc. Int’l Conf. World Wide Web (WWW)*, pp. 581-590, 2007.

[7] J. Teevan, S.T. Dumais, and D.J. Liebling, “To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent,” *Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR*), pp. 163-170, 2008.

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

[9] Y. Xu, K. Wang, G. Yang, and A.W.C Fu, “Online anonymity for personalized web services” *Proc.18th ACM conformation and knowledge management (CIKM)*, pp 1497-1500, 2009.

[10] X. Xiao and Y. Tao, “Personalized Privacy Preservation,” *Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD*), 2006.

[11] Viejo and J. Castella-Roca, “Using Social Networks to Distort Users’ Profiles Generated by Web Search Engines,” Computer Networks, vol. 54, no. 9, pp. 1343-1357, 2010.

[12] J. Castelli-Roca and A. Vijeo and J. Herrera-Joancomarti, “Preserving user’s privacy in web search engines, Computer *Comm.*vol.32,Vol.32,no.13/14, pp 1541-1551, 2009.

[13] X. Shen, B. Tan, and C. Zhai, “Privacy Protection in Personalized Search,” *ACM* SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.