ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 8 , August 2016

Security Enhancement in Personalized Web Search

Rupali Keshavrao Aher , Akshay Rajdhar Adik

Department of Computer Engineering, MCERC, Nashik, India

Software Engineer, Pune

ABSTRACT: For clients with individual data objectives, web personalization is utilized to enhance look up quality by altering query items, in view of the individual information of client shared to the web search tool. Clients are not happy with uncovering private inclination data to web crawlers, yet in the event that there is gain in search quality or productivity then protection can be traded off. Accordingly, there ought to be a harmony between the inquiry quality and security assurance. Security assurance can be achieved by using hierarchical client profile model in PWS applications.

A PWS framework called UPS that can adaptively generalize profiles by queries when the client specifies privacy requirements. Runtime generalization aims at striking a harmony between predictive metrics that evaluate the average precision of information and the privacy risk of exposing the generalized profile. Offline generalization and online generalization algorithm are used for static and runtime generalization. Online prediction mechanism is used for deciding personalizing a query is beneficial. Discriminating power is used for the prediction of generalization decision. It also considers the preferences of interest mentioned in user profile and sensitivity of information defined by user while generalizing the user Profile. Session attacks like hijacking, eavesdrops, injections are controlled.

KEYWORDS: Privacy protection, personalized web search, utility, risk, profile

I. INTRODUCTION

The web search engine is widely used by the users for searching useful information on the web. But the amount of information on the web grows continuously so it becomes very difficult for web search engines to find information that satisfies user's individual needs. Due to the enormous variety of user's contexts and backgrounds, as well as the ambiguity of texts, search engines return irrelevant results that do not meet the user's real intentions. For providing better search results a general category of search techniques, personalized web search (PWS) is used. To figure out the user intention behind the issued query, user information has to be collected and analyzed.
There are two types of solutions to the PWS

1. Click-log-based method: This is a straightforward method. The click-log based methods uses clicked pages in the users query history. But it has strong limitation that it can only work on repeated queries from the same user [2].
2. Profile-based methods: Profile-based methods can be used effectively for almost all sorts of queries, but under some circumstances the results are unstable [2]. It improves the search experience with complicated user-interest models generated from user profiling techniques.

There are pros and cons for both types of PWS techniques, but profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users. It is usually gathered implicitly from query history[3],[4],[5], browsing history[6],[7], click-through data[8],[9],[2], bookmarks[10], user documents[3],[11], and so forth. Unfortunately, such implicitly collected personal data can easily disclose a span of user's private life. Privacy issues are raised from the lack of protection for such data, for instance the AOL query logs scandal [12], raise panic among individual users, and also dampen the data-publishers enthusiasm in offering personalized service. So the privacy concerns have become the major barrier for wide proliferation of PWS services.

Existing system have a privacy-preserving personalized web search framework UPS. User specifies the privacy requirements and according to the requirements user profiles are generalized. The problem of privacy-preserving personalized search is formulated as δ-Risk Profile Generalization, by using two conflicting metrics, personalization utility and privacy risk, for hierarchical user profile. Two simple and effective generalization algorithms, GreedyDP

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 8 , August 2016

and GreedyIL are developed, which support runtime profiling. GreedyDP tries to maximize the discriminating power (DP), and the GreedyIL attempts to minimize the information loss (IL). To enhance the stability of the search results and to avoid the unnecessary exposure of the profile an inexpensive mechanism is used for deciding whether to personalize a query in UPS. UPS allows customization of privacy needs; and it does not require iterative user interaction.

II. RELATED WORK

User profiles disclose the individual information goals so to improve the search quality, profile based PWS refers the user profile. Term list/vectors [6] or bag words [3] are used previously to represent the profile. Hierarchical structures are commonly used to build the profiles as they provide higher access efficiency, stronger descriptive ability, and better scalability. Hierarchical profiles build automatically by using term frequency analysis of the user data [11]. Weighted topic hierarchy/graph such as ODP [2][13][15], Wikipedia [15][16] are used for constructing hierarchical profiles. Normalized Discounted Cumulative Gain (nDCG) is a common measure of the effectiveness of an information retrieval system but it requires high cost in explicit feedback collection. Other metrics of personalized web search rely on clicking decisions, including average rank [4][9], Rank Scoring and Average Precision[19][11] which reduces human involvement in performance measuring. To measure the effectiveness of the personalization in UPS we used average precision metric [2], and two predictive metrics, personalization utility and privacy risk on a profile instance without requesting for user feedback.

One class of Privacy protection problem for PWS treats privacy as the identification of an individual [18]. It try to solve the privacy problem on different levels, pseudoidentity, the group identity, no identity, and no personal information. Due to the high cost in communication and cryptography the third and fourth levels are impractical. First level solution is proved to fragile [12]. By generating a group profile of k users [19] and [20] provide online anonymity on user profiles. To shuffle queries among a group of users who issues them useless user profile protocol is proposed [21] So that entity cannot profile a certain individual. It assumes the existence of a trustworthy third-party anonymizer. Instead of third party to provide a distorted user profile to the search engine Viejo[21] use the legacy social network.

Other class considers the sensitivity of the data, particularly the user profiles disclosed to the PWS server. Users only trust themselves and cannot tolerate the disclosure of their complete profiles on anonymity server. Third party assistance or collaborations between social network entries is not required. To generate the near–optimal partial profile Krause and Horvitz employ statistical techniques to learn a probabilistic model. But it builds the user profiles as a finite set of attributes and the probabilistic model is trained through predefined frequent queries. Privacy protection solution given by Xu et al [10] is based on hierarchical profiles. Generalized profile is obtained as a rooted subtree of the complete profile using a user specifiedthreshold. But it does not address the query utility which is important for the service quality of UPS. Personalization have different effect on different queries [2], distinct queries are more benefited while larger click-entropy value queries are not. To classify queries by their click entropy Teevan et al. [22] collect a set of features of the query. Based on a client-side solution UPS framework differentiate distinct queries from ambiguous ones using the predictive query utility metric.

In the previous work [23] the prototype of UPS is proposed together with a greedy algorithm GreedyDP which support online profiling based on predictive metrics of personalization utility and privacy risk. In this paper metric of personalization utility captures three new observations. Evaluation model is refined to support user-customized sensitivities. New profile generalization algorithm GreedyIL is proposed.

III. EXISTING SYSTEM

Existing system have a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to user-specified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, the problem of privacy-preserving personalized search is formulated as δ-Risk Profile Generalization, with its N P-hardness proved. Two simple but effective generalization algorithms, GreedyDP and GreedyIL are developed, to support runtime profiling. While the former tries to maximize the discriminating power (DP), the latter attempts to minimize the information loss (IL). By exploiting a number of heuristics, GreedyIL out performs GreedyDP significantly. An inexpensive mechanism is

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 8 , August 2016

provided for the client to decide whether to personalize a query in UPS. This decision can be made before each runtime profiling to enhance the stability of the search results while avoid the unnecessary exposure of the profile. UPS is distinguished from conventional PWS in that it 1) provides runtime profiling, which in effect optimizes the personalization utility while respecting users privacy requirements; 2) allows for customization of privacy needs; and 3) does not require iterative user interaction.

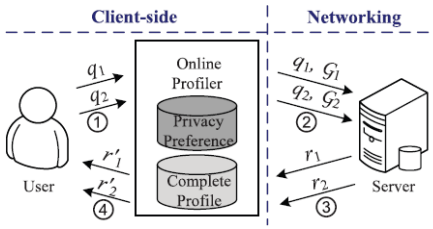

Figure 1: Existing System architecture

As illustrated in Fig.1.1, UPS consists of a nontrusty search engine server and a number of clients. Each client (user) accessing the search service trusts no one but himself/ herself. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. The proxy maintains both the complete user profile, in a hierarchy of nodes with semantics, and the user-specified (customized) privacy requirements represented as a set of sensitive-nodes.

Disadvantage:

i. All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.
ii. The existing profile-based PWS do not support runtime profiling.
iii. The existing methods do not take into account the customization of privacy requirements.
iv. Personalization techniques require iterative user interactions when creating personalized search results.

IV. **PROPOSED SYSTEM**

Proposed system works in two modes:

1. Offline mode
2. Online mode.

User first defines the small portion of the profile preferences and sensitivities and its intent. User intent is defined in Extensible Markup Language (XML) format.

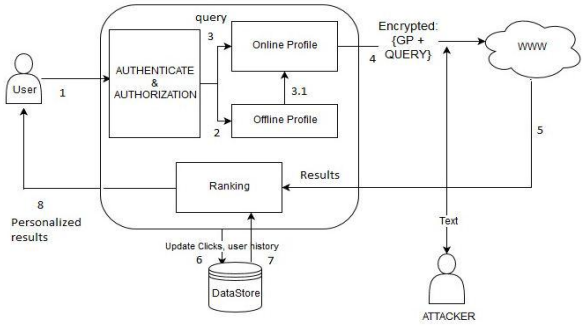

Figure 2: Proposed System Architecture

In proposed system, Users privacy protection is started by specifying the user's privacy profile definition and sensitivity data definition. Hierarchical profile is generated based on

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 3, Issue 8 , August 2016

user defined privacy requirements. When user logged in, offline profile generation is started by clicking button. Once the offline profile is created user can issue query. Personalized profile and user query is issued to PWS for personalized search. Search results are personalized and delivered back to query. PWS re-rank the search results based on the occurrences of query in document.

Advantages of proposed system:

Personalized search is able to unravel the aspects we need to optimize for users, instead of concentrating on search engines alone. Therefore, it motivates us to concentrate more on creating high quality, interactive content for the benefit of users rather than resorting to link acquiring efforts for boosting search engine rankings superficially. The focus has now shifted towards creating more engaging and rewarding experiences for the user.

V. EXPERIMENTAL RESULTS

Precision :

In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the query:
Precision = |{RelevantDocumentsg } {RetrievedDocuments}|÷|{RetrievedDocuments}|
Precision takes all retrieved documents into account, but it can also be evaluated at a given Cut-off rank, considering only the topmost results returned by the system. This measure is called precision at n or P@n. For example for a text search on a set of documents precision is the number of correct results divided by the number of all returned results. Precision is also used with recall, the percent of all relevant documents that is returned by the search. The two measures are sometimes used together in the F1 Score (or f-measure) to provide a single measurement for a system.
The following graph shows precision of direct search and our approach.

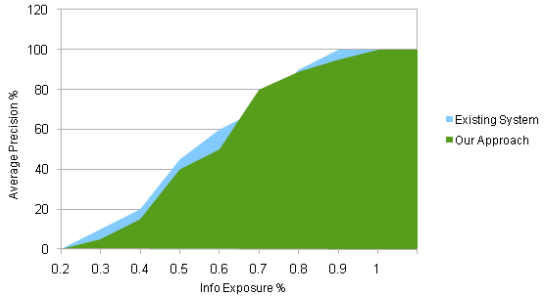

Figure 2: Precision of direct search and our approach

Recall :

Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved.
Recall = |{RelevantDocuments} {RetrievedDocuments}|÷|{RelevantDocuments}|
For example for text search on a set of documents recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. So it can be looked at as the probability that a relevant document is retrieved by the query. It is trivial to achieve recall of 100% by returning all documents in response to any query. Therefore, recall alone is not enough but one needs to measure the number of non-relevant
Documents also , for example by computing the precision. The following graph shows that precision value increases with every iteration, it indicates that suppose user enter search query for the first time the precision value will be minimal but after successive iteration for the same query it increases gradually.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 8 , August 2016

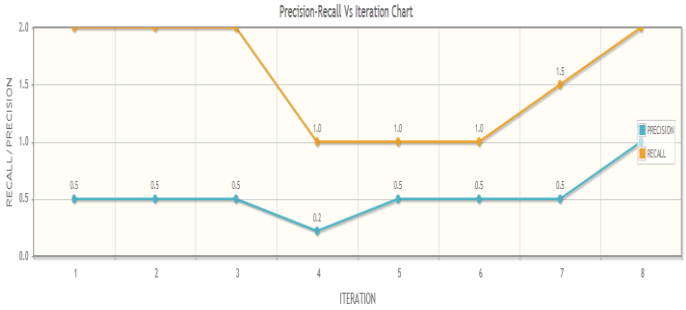

Figure 3: Average Precision Vs Iteartion

VI. CONCLUSION

Privacy -Enhanced Web personalization System can adaptively generalize profiles by queries while respecting user specified privacy requirements. Runtime generalization aims at striking a harmony between predictive metrics that evaluate the average precision of information and the privacy risk of exposing the generalized profile. Offine generalization and online generalization algorithm are used for static and runtime generalization. Online prediction mechanism used for deciding whether personalizing a query is beneficial. Discriminating power is used for the prediction of generalization decision. It also considers the preferences of interest mentioned in user profile and sensitivity of information defined by user while generalizing the user Profile. Session attacks like eavesdrops attacks are controlled.

REFERENCES

[1] 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 YEAR 2014

[2] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨ tter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.

[14] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.

[15] E. Gabrilovich and S. Markovich, "Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI), 2006.

[16] K. Ramanathan, J. Giraudi, and A. Gupta, "Creating Hierarchical User Profiles Using Wikipedia," HP Labs, 2008.

[17] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.

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

[19] Y. Xu, K. Wang, G. Yang, and A.W.-C. Fu, "Online Anonymity for Personalized Web Services," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 1497-1500, 2009.

[20] 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.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 3, Issue 8 , August 2016

[21] J. Castellı´-Roca, A. Viejo, and J. Herrera-Joancomartı´, "Preserving User's Privacy in Web Search Engines," Computer Comm., vol. 32,no. 13/14, pp. 1541-1551, 2009.

[22] 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.

G. Chen, H. Bai, L. Shou, K. Chen, and Y. Gao, "Ups: Efficient Privacy Protection in Personalized Web Search," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information, pp. 615- 624, 2011