Proceedings of National Conference on Recent Advances in Computer Science & Engineering (NCRACSE-2017)

International Journal for Modern Trends in Science and Technology
Volume: 03, Special Issue No: 01, February 2017
ISSN: 2455-3778
http://www.ijmtst.com

Enhancing Personalized Web Search with Privacy Protection

Mallela Ratna Raju[1] | Jeevan Ratnakar Kondru[2] | Kotapati Upendra[3]

[1,2,3]Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India.

To Cite this Article
Mallela Ratna Raju, Jeevan Ratnakar Kondru and Kotapati Upendra, "Enhancing Personalized Web Search with Privacy Protection", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Special Issue 01, 2017, pp. 131-137.

ABSTRACT

Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting userspecified privacy requirements. Our 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. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that GreedyIL significantly outperforms GreedyDP in terms of efficiency.

KEYWORDS: *Privacy protection, personalized web search, utility, enhancing, risk,Taxonomy, profile.*

I. Introduction

The web search engine has long become the most important portal for ordinary people looking for useful information on the web. Actually the personalized web search means the ability to identify the different needs of different people who issue the same text query for web search. Yahoo uses this concept in 1998. At present 80% of the people prefer to use personalized web search engines.As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.When people mention the term "search engine", it is often used generically to describe both crawler-based search engines and human-powered references. In fact, these two types of search engines gather their listings in radically different ways and therefore are inherently different.

Crawler-based search engines are good when you have a specific search topic in mind and can be very efficient in finding relevant information in this situation. However, when the search topic is general, crawler-base search engines may return hundreds of thousands of irrelevant responses to simple search requests, including lengthy documents in which your keyword appears only once.Human-powered directories are good when you are interested in a general topic of search. In this situation, a directory can guide and help you narrow your search and get refined results. Therefore, search results found in a human-powered directory are usually more relevant to the search topic and more accurate. However, this is not an efficient way to find information when a specific search topic is in mind.

As the size of the Internet continues to grow the users of search providers continually demand search results that are accurate to their needs. Personalized Search is one of the options available to users in order to sculpt search results returned to them based on their personal data provided to the search provider. This raises concerns of privacy issues however as users are typically uncomfortable revealing personal information to an often faceless service provider on the Internet. This paper aims to deal with the privacy issues surrounding personalized search and discusses ways that privacy can be enriched so that users can become more comfortable with the release of their personal data in order to receive more accurate search results.

The solutions to personalized web search can generally be categorized intotwo types, namely click-log-based methods and profile-basedones.

The click-log based methods are straightforward—they just force predisposition to clicked pages in the client's inquiry history. Despite the fact that this system has been showed to perform reliably and impressively well [10], it can just take a shot at rehashed inquiries from the same client, which is a solid confinement keeping its pertinence. Interestingly, profile-based techniques enhance the inquiry involvement with muddled client investment models created from client profiling systems.

Profile-based techniques can be possibly compelling for just about assorted types of inquiries, however are accounted for to be shaky under a few circumstances [10].The profile-based personalized web search has showed more adequacy in enhancing the nature of web pursuit as of late, with expanding use of individual and conduct data to profile its clients, which is normally assembled certainly from inquiry history [11], [3], [4], scanning history [5], [8], navigate information [7], [6], [10] bookmarks [9], client records [11], [1], et cetera. Tragically, such certainly gathered individual information can undoubtedly uncover an array of client's private life. Security issues climbing from the absence of assurance for such information, for example the AOL inquiry logs embarrassment [2] raise alarm among individual clients, as well as hose the information distributer's energy in offering customized administration. Actually, security concerns have turned into the real obstruction for wide expansion of personalized web search administrations.We propose a PWS system called UPS that can adaptively sum up profiles by questions while regarding userspecified protection prerequisites. Our runtime speculation goes for striking a harmony between two prescient measurements that assess the utility of personalization and the security danger of uncovering the summed up profile. We display two voracious calculations, to be specific GreedyDP and GreedyIL, for runtime speculation.

II. Literature Survey

Personalized web search is an attempt to find most relevant documents using information about user's goals, knowledge, preferences, navigation history, etc.R. Larsen: With the growth of DL even a good query can return not just tens, but thousands of "relevant" documents.A user's profile is a collection of information about the user of the system.This information is used to get the user to more relevant information. Common term for user models in either IR or IF.Views on user profiles in IR community are Classic - a reference point and Modern - simple form of a user model. The benefits of personalized web search are Resolving ambiguity and Revealing hidden treasures. The components of web search as shown in fig. 1

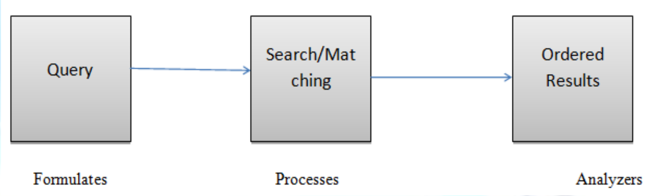

Fig. 1 Components of web search

Z. Dou, R. Melody, and J.-R. Wen, Although customized quest has been proposed for a long time and numerous personalization methods have been examined, it is still vague whether personalization is reliably powerful on diverse questions for distinctive clients, and under distinctive hunt settings. In this paper, we ponder this issue and give some preparatory conclusions. M. Spertta and S. Gach, User profiles, depictions of client investments, can be utilized via web search tools to give personalizedsearch results. Numerous methodologies to making client profiles gather client data through intermediary servers (to catch skimming histories) or desktop bots (to catch exercises on a PC). Both these procedures oblige investment of the client to introduce the intermediary server or the bot. B. Tan, X. Shen, and C. Zhai, Long-term look history contains rich data around a client's inquiry inclination, which can be utilized as hunt setting to enhance recovery execution. X. Shen, B. Tan, and C. Zhai, Information recovery frameworks (e.g., web crawlers) are discriminating for overcoming data over-burden. A real inadequacy of existing recovery frameworks is that they for the most part need client displaying and are not versatile to individual clients, bringing about characteristically non-ideal recovery execution.

A. Existing System

The existing profile-based Personalized Web Search do not support runtime profiling. A client profile is regularly summed up for just once disconnected from the net, and used to customize all inquiries from a same client indiscriminatingly.Such "one profile fits all" methodology absolutely has disadvantages given the mixture of questions.One confirmation reported in is that profile-based personalization may not by any means help to enhance the quest quality for some impromptu questions, however presenting client profile to a server has put the client's security at danger.

The current techniques don't consider the customization of protection necessities. This likely makes some client protection to be overprotected while others deficiently secured. For example, in, all the delicate themes are discovered utilizing a flat out metric called surprisal focused around the data hypothesis, accepting that the hobbies with less client record backing are more sensitive. However, this assumption can be doubted with a simple counterexample: If a user has a large number of documents about “sex,” the surprisal of this topic may lead to a conclusion that “sex” is very general and not sensitive, despite the truth which is opposite. Unfortunately, few prior work can effectively address individual privacy needs during the generalization.

Numerous personalization procedures oblige iterative client cooperations when making customized indexed lists. They generally refine the query items with a few measurements which oblige numerous client cooperations, for example, rank scoring, normal rank, etc. This ideal model is, be that as it may, infeasible for runtime profiling, as it won't just posture an excessive amount of danger of protection break, additionally request restrictive handling time for profiling. Along these lines, we require prescient measurements to gauge the pursuit quality and rupture chance after personalization, without acquiring iterative client cooperation.

In the existing system, 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 [10], 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 poten-tially effective for almost all sorts of queries, but arereported to be unstable under some circumstances.

The problems with the existing methods are explained in the following

a) All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.

b) Itdo not support runtimeprofiling.

c) Do not take into account thecustomization of privacy requirements.

d) Generally there are two classes of privacy protection problems for PWS. One class includes those treat privacy as the identification of an individual, as described. The other includes those consider the sensitivity of the data, particularly the user profiles, exposed to the PWS server.

e) Many personalization techniques require iterative userinteractions when creating personalized search results.

III. PROBLEM DEFINITION

When using a Personalized Search service such as the ones mentioned [10] [11], how is the user and the search provider to ensure the privacy and protection of a user’s identity and information that is supplied to the service? As mentioned before[8], if a user does not trust the search provider, than the user is simply not going to either a) divulge sufficient personal information in order to optimize his/her search results or b) will not use the personalized search feature at all. Also, users may have concerns regarding the security surrounding the storage of their data. After a user’s personal information has been given away, the onus is on the search provider to ensure that the information remains private and does not fall into the hands of people or organizations with malicious intentions for that data.

Our work aims at providing protection against a typical model of privacy attack, namely eavesdropping. As shown in Fig. 2, to corrupt Alice's privacy, the eavesdropper Eve successfully intercepts the communication between Alice and the PWS-server via some measures, such as man-in-the middle attack, invading the server, and so on. Consequently, whenever Alice issues a query q, the entire copy of q together with a runtime profile G will be captured by Eve. Based on G, Eve will attempt to touch the sensitive nodes of Alice by recovering the segments hidden from the original H and computing a confidence for each recovered topic, relying on the background knowledge in the publicly available taxonomy repository R.Note that in our attack model, Eve is regarded as an adversary satisfying the following assumptions:

Knowledge bounded. The background knowledge of the adversary is limited to the taxonomy repository R. Both the profile H and privacy are defined based on R.

Session bounded. None of previously captured information is available for tracing the same victim in a long duration. In other words, the eavesdropping will be started and ended within a single query session.

The above assumptions seem strong, but are reasonable in practice. This is due to the fact that the majority of privacy attacks on the web are undertaken by some automatic programs for sending targeted (spam) advertisements to a large amount of PWS-users. These programs rarely act as a real person that collects prolific information of a specific victim for a long time as the latter is much more costly. The sample taxonomy repository and sample user profile figures are observed in supporting privacy protection in personalized web search [12].

If we consider the sensitivity of each sensitive topic as the cost of recovering it, the privacy risk can be defined as the total (probabilistic) sensitivity of the sensitive nodes, which the adversary can probably recover from G. For fairness among different users, we can normalize the privacy risk with

$$\sum_{s\in S} sen(s)$$

which stands for the total wealth of the user. Our approach to privacy protection of personalized web search has to keep this privacy risk under control.

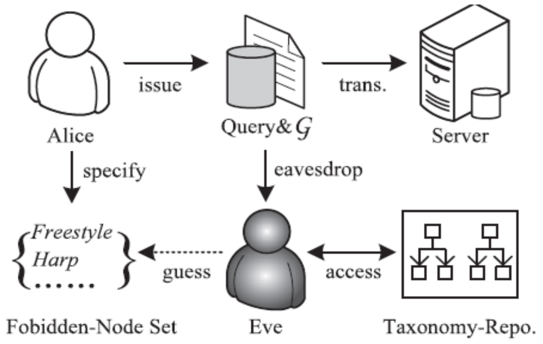

Fig. 2 PWS Attack model

Definition 1 (USER PROFILE/H).A user profile H, as a hierarchical representation of user interests, is a rooted subtree of R. The notion rooted subtree is given in Definition 2.

Definition 2 (ROOTED SUBTREE). Given two trees S and T , S is a rooted subtree of T if S can be generated from T by removing a node set X _ T (together with subtrees) from T , i.e., S =rsbtr(X,T).

IV. PROPOSED WORK

The above problems are addressed in our UPS (User customizable Privacy-preserving Search) framework. The framework assumes that the queries do not contain any sensitive information, and aims at protecting the privacy in individual user profiles while retaining their usefulness for PWS.The framework works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase, a) The search results are personalized with the profile and delivered back to the query proxy. b) The query and the generalized user profile are sent together to the PWS server for personalized search.

We propose a privacy-preserving personalized web search framework UPS, which can generalize profiles for each query according to userspecified privacy requirements. Relying on the definition of two conflicting metrics, namely personalization utility and privacy risk, for hierarchical user profile, we formulate the problem of privacy-preserving personalized search as Risk Profile Generalization, with its NP-hardness proved. The main advantages of this approach are

a) It enhances the stability of the search quality.
b) Increasing usage of personal and behaviour information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data bookmarks, user documents, and so forth.

c) The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, UPS also performed online generalization on user profiles to protect the personal privacy without compromising the search quality.

d) It avoids the unnecessary exposure of the user profile.

We develop two simple but effective generalization algorithms, GreedyDP and GreedyIL, 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 outperforms GreedyDP significantly.We provide an inexpensive mechanism forthe client to decide whether to personalize a query in UPS[8]. 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. Our extensive experiments demonstrate the efficiency and effectiveness of our UPS framework.

A. Greedy Algorithm

A greedy algorithm is a mathematical process that recursively constructs a set of objects from the smallest possible constituent parts. Recursion is an approach to problem solving in which the solution to a particular problem depends on solutions to smaller instances of the same problem. Greedy algorithms look for simple, easy-to implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit. Such algorithms are called greedy because while the optimal solution to each smaller instance will provide an immediate output, the algorithm doesn't consider the larger problem as a whole. Once a decision has been made, it is never reconsidered.

The advantage to using a greedy algorithm is thatsolutions to smaller instances of the problem can be straightforward and easy to understand. The disadvantage is that it is entirely possible that the most optimal short-term solutions may lead to the worst long-term outcome.

Greedy algorithms are often used in ad hocmobile networking to efficiently route packetswith the fewest number of hopsand the shortest delay possible. They are also used in machine learning, business intelligence, artificial intelligence and programming.

B. GreedyDP Algorithm

The first greedy algorithm GreedyDP works in a bottom-up manner. Starting from G0, in every i^{th} iteration, GreedyDP chooses a leaf topic for pruning, trying to maximize the utility of the output of the current iteration, namely G_{i+1}. During the iterations, we also maintain a best-profile-so-far, which indicates the G_{i+1} having the highest discriminating power while satisfying the risk constraint. The iterative process terminates when the profile is generalized to a root-topic. The best-profile-so-far will be the final result (G^*) of the algorithm.The GreedyDP algorithm is observed in supporting privacy protection in personalized web search[12]. The main problem of GreedyDP is that it requires recomputation of all candidate profiles (together with their discriminating power and privacy risk) generated from attempts of prune-leaf on all. This causes significant memory requirements and computational cost.

C. GreedyIL Algorithm

The GreedyIL algorithm improves the efficiency of the generalization using heuristics based on several findings. One important finding is that any prune-leaf operation reduces the discriminating power of the profile. In other words, the DP displays monotonicity by prune-leaf.

D. Proposed Approach

1. *Profile-based personalization:* This paper introduces an approach to personalize digital multimedia content based on user profile information. For this, two main mechanisms were developed; a profile generator that automatically creates user profiles representing the user preferences, and a content-based recommendation algorithm that estimates the user's interest in unknown content by matching her profile to metadata descriptions of the content. Both features are integrated into a personalization system.

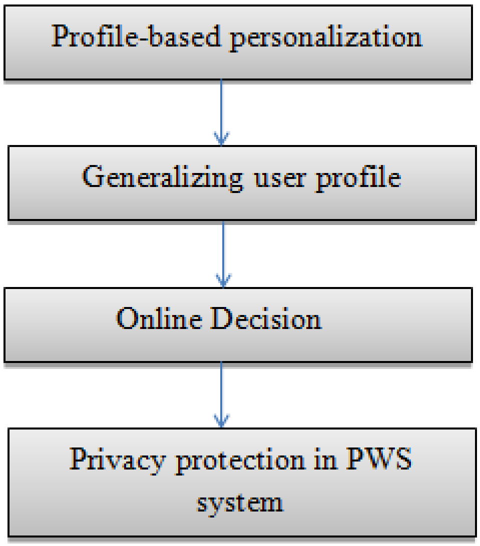
Profile-based personalization
Generalizing user profile
Online Decision
Privacy protection in PWS system

Fig.3: Proposed approach

2. *Generalizing user profile:* The generalization process needs to meet particular essentials to handle the client profile. This is attained by preprocessing the client profile. From the beginning, the methodology instates the client profile by considering the showed guardian client profile. The procedure adds the inherited properties to the properties of the neighborhood client profile. Thereafter the process loads the data for the foreground and the background of the map according to the described selection in the user profile.

Additionally, using references enables caching and is helpful when considering an implementation in a production environment. The reference to the user profile can be used as an identifier for already processed user profiles. It allows performing the customization process once, but reusing the result multiple times. However, it has to be made sure, that an update of the user profile is also propagated to the generalization process. This requires specific update strategies, which check after a specific timeout or a specific event, if the user profile has not changed yet. Additionally, as the generalization process involves remote data services, which might be updated frequently, the cached generalization results might become outdated. Thus selecting a specific caching strategy requires careful analysis.

3. *Online-Decision:* The profile-based personalization helps little or even diminishes the pursuit quality, while presenting the profile to a server would beyond any doubt chance the client's protection. To address this issue, we create an online component to choose whether to customize an inquiry. The fundamental thought is clear. in the event that a different inquiry is distinguished amid speculation, the whole runtime profiling will be prematurely ended and the question will be sent to the server without a client profile.

4. *Privacy Protection in PWS System:* We propose a PWS framework called UPS that can generalize profiles in for each query according to user-specified privacy requirements. Two predictive metrics are proposed to evaluate the privacy breach risk and the query utility for hierarchical user profile. We develop two simple but effective generalization algorithms for user profiles allowing for query-level customization using our proposed metrics. We also provide an online prediction mechanism based on query utility for deciding whether to personalize a query in UPS. Extensive experiments demonstrate the efficiency and effectiveness of our framework.

V. CONCLUSION AND FUTUREWORK

This paper presented a client-side privacy protectionframework called UPS for personalized web search. UPScould potentially be adopted by any PWS that captures userprofiles in a hierarchical taxonomy. The framework allowedusers to specify customized privacy requirements via thehierarchical profiles. In addition, UPS also performedonline generalization on user profiles to protect the personalprivacy without compromising the search quality. Weproposed two greedy algorithms, namely GreedyDP andGreedyIL, for the online generalization. Our experimentalresults revealed that UPS could achieve quality searchresults while preserving user's customized privacy requirements.The results also confirmed the effectiveness andefficiency of our solution.

For future work, we will try to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentiality, and so on), or capability to capture a series of queries (relaxing the second constraint of the adversary) from the victim. We will also seek more sophisticated method to build the user profile, and better metrics to predict the performance (especially the utility) of UPS.

REFERENCES

[1] 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.

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

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

[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] 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.

[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] Pavani Potnuri, Bharath Kumar Gowru and Venkateswara Rao Nadakuditi “Vampire Attacks: Draining Life From Wireless Adhoc Sensor Networks” in International Journal of Science Engineering and Advance Technology ISSN 2321-6905, Vol 3, Issue 11, NOVEMBER – 2015.

[9] 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.

[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] 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.

[12] 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.

[13] 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.