International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 9 5543 - 5548

Analyzing Privacy Protection in Personalized Web Search

Prateek C. Shukla[1]
Student of BE Information Technology
BVCOE & RI, *Nasik, India*
Savitribai Phule Pune University
E-mail: prateekshukla40@gmail.com

Yogeshwar J. Shirsath[3]
Student of BE Information Technology
BVCOE & RI, *Nasik, India*
Savitribai Phule Pune University
E-mail: yjshirsath007@gmail.com

Tekchand D. Patil[2]
Student of BE Information Technology
BVCOE & RI, *Nasik, India*
Savitribai Phule Pune University
E-mail: teks712@gmail.com

Dnyaneshwar N. Rasal[4]
Student of BE Information Technology
BVCOE & RI, *Nasik, India*
Savitribai Phule Pune University
E-mail: drdrasaln@gmail.com

Prof. S. R. Jadhav[5]
Assistant Prof. Dept. of Information Technology
BVCOE & RI, *Nasik, India*
Savitribai Phule Pune University
E-mail: srj4041@gmail.com

***Abstract*—**Personalized Web Search (PWS) is very effective in improving the quality of search services on the internet. The information on internet has increased day by day and user demand for the accurate result, for the accurate result the user has option to PWS. PWS works on the basis of information that user provide to search provider, the current result based on that information. This paper model makes use of hierarchical user profiles, it simultaneously maintaining privacy protection required by the user. Greedy DP (Discriminating Power) & Greedy IL (Information Loss) are used for runtime generalization and it have online prediction that query requires personalization or not.

Keywords- *Privacy protection, personalized web search, profile, GreedyIL, GreedyDP*

*****

I. INTRODUCTION

Many of ordinary people use web search engine for their query. The web search engine has become important portal for the users. However, users get all related information about the query, the web search engine gives the irrelevant results. These irrelevant results are due to the different users. The reason of these irrelevant results is the ambiguity of query, large variety of user's contexts and background. Personalized web search (PWS) provides the better search result. Personalization is an attempt to find most of the relevant documents using information about user's goals, knowledge, preferences, navigation history, etc. [3]. The PWS resolves the ambiguity of the query in order to reduce ambiguity. Ex: The profile of interest will allow to distinguish what the user asked about "apple" ("Fruit", "cell phone") really requires. The PWS revealing hidden treasures. The profile permits to bring around surface most relevant document that can be hidden on the far side prime result page. Ex: Owner for Google humanoid Pages touching on each would be most fascinating. The solutions of PWS are often categorized into 2 sorts, specifically click-log-based technique and profile primarily based. The click-log based ways that is easy they just impose bias to clicked pages at intervals the user's question history. This that could be a strategy is extraordinarily indisputable to perform consistently and considerably well [1], it'll exclusively work on continual queries from a similar user that could be a robust limitation confining its pertinency. Also, profile-based technique virtually improves the search expertise with sophisticated user-interest models generated from user identification techniques. Profile-based strategies are generally probably effective for pretty much all types of queries, however area unit reported to be unstable underneath some circumstances [1]. though there are a unit professionals and cons for each styles of PWS techniques, the profile-based PWS has incontestable a lot of effectiveness in up the standard of net search recently, with enhanced usage of non-public and behavior info to profile its users, that is typically collected implicitly from question history [2], [3], browsing history, click-through knowledge bookmarks, user documents then forth. Sadly, such implicitly collected together personal knowledge will simply reveal a gamut of user's non-public life [6]. Privacy problems rising from the shortage of protection for such knowledge, for example the AOL question logs scandal [11], not raise panic among individual users, however can jointly dampen the data-publisher's enthusiasm in providing personalized service. In fact, privacy considerations became the most important hurdle for wide proliferation of PWS services [12].

A. Motivation:

In order to shield user privacy in profile-based PWS, researchers ought to contemplate two contradicting effects

throughout the search method. On one aspect, they're creating a trial to boost the search quality with the personalization utility of the user profile [5]. And on the opposite aspect, they need to cover the privacy contents existing within the user profile to position the privacy risk in restraint. a number of previous studies [12] recommend that individuals are willing to compromise with their privacy if the personalization by providing user profile to the computer programme yields improved search quality. In a perfect case, vital gain is get by personalization at the expense of solely a bit (less-sensitive) portion of the user profile, particularly a profile that I generalized. Thus, user privacy would be protected while not compromising with their personalised search quality. Generally, there is trade-off between the search quality and therefore the privacy protection achieved from generalization. Sadly, the sooner works of privacy protective PWS are removed from optimum. The issues with the present strategies are explained within the following observations:
1. the present profile-based PWS don't support runtime identification. A user profile is especially generalized for the authors are with the school of engineering and Technology, Zhejiang University one time offline, and accustomed change all queries from a same user indiscriminatingly. The profile which follows "one profile fits all" strategy actually has drawbacks given the range of queries. One proof rumored in [1] is that profile-based personalization might not additionally facilitate to raise the search quality for a few unplanned queries, the exposing user profile to a server has placed the user's privacy in danger. a stronger approach is to form an internet call on a whether or not to change the question (exposing the profile) and b. what to reveal within the user profile at runtime. To the most effective of our data, no previous work has supported such feature.
2. The present strategies don't take under consideration the customization of privacy wants. This could in all probability making few user privacy to be overprotected whereas others insufficiently protected. In an example, in [10], all the sensitive topics are detected victimisation associate degree absolute metric known as perturbation supported the data theory, assuming that the interests with very less user document support ar a lot of sensitive. But, this assumption will have doubt with an easy counterexample: If a user encompasses a sizable amount of document like "sex," the perturbation of this subject cause a conclusion that "sex" is extremely comman and not sensitive, despite the reality that is in opposite side. Sadly, some previous task will effectively address individual's privacy that is required in the generalization
3. Several personalization techniques would like repetitious user interactions whereas making personalised search results. They sometimes refine the search results with few metrics which require multiple user interactions, like rank grading [13], average rank [8], and so on. This paradigm is, however, isn't optimum for runtime identification, because it will not solely cause an excessive amounts of risk of privacy breach, and however additionally demand preventive process period to identification. Thus, we want prognosticative metrics to live the search quality and breach risk once personalization, with none abundant acquisition repetitious user interaction.

B. Contribution:

The issues which are addressed above is present in our UPS (User Customizable Privacy-preserving Search) framework. It is taken as assumption that the queries does not contain any sensitive information, and set the target aim to the protection of privacy in particular users profiles and holding utility for PWS. As given in Fig.1, UPS have no accurate computer program server and various purchasers. Every user i.e. client approaching the search service have no faith on other except himself/herself. Web profiler strengthens as a proxy search running on the client's machine is the key element for Privacy Protection. The entire user profile is maintained by the proxy, whiles a hierarchy of nodes with their own languages, and also customized privacy. The framework works particularly in two phases such as offline and online section for each and every user. The offline section for each and every user.

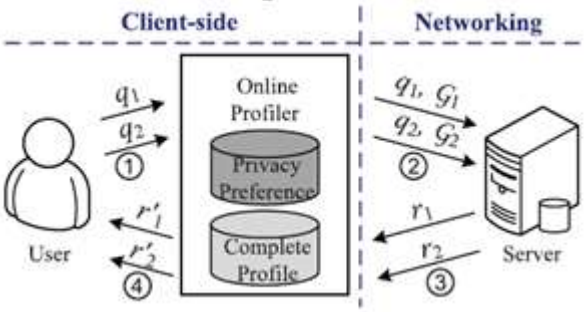


Fig. 1. System architecture of UPS.

The offline section, user profile is built hierarchically and also customized with user-specified privacy requirement. The online section queries are handled as given below:
1) If a user finds problem, then a query qi on the client side, a user profile is generated by the proxy in runtime under the light-weight of the term query. A generalized user profile Gi satisfying the privacy needs, may be the output of this particular step. The method of generalization is directly a target-hunting by taking up two conflicting metrics, specifically the personalization utility and also the risk of privacy with outlined for user profiles.
2) Thereafter, the generalized user profile and the query are sent with PWS search for the personalized search.
3) The results which are searched is Personalised with particular profile and that profile is delivered back to the query proxy.
4) Lastly, the proxy can present the raw result to the user i.e. client or it can re-rank them with the whole user profile.
Standard PWS is differentiated from UPS (1) Runtime identification are provided, that is its impact optimizes the personalization utility as respecting the privacy requirement by the user. (2) Personalization of privacy requirement is permitted. (3) Re-iterative user interaction is needed. Important contribution are as follows:-

- We suggest privacy-preserving personalized web search framework UPS, in which profiles are generalized for all query containing user-specified privacy requirements.
- Depending on the definition of two conflicting metrics, particularly personalization utility and privacy risk, for user profile which hierarchical, we

set to formulate the issue of privacy preserving personalized search as risk profile generalization and with its NP-hardness tested.

- We derive two effective algorithms known as GreedyDP and GreedyIL in order to support identification during runtime. Earlier, it was trying to maximize the discriminating power (DP) and the second algorithm makes the reduction of information loss (IL).
- We are giving a mechanism which is able to be afforded for the user i.e. client in order to make decision whether to personalize the query or not in UPS. Generally, this call is created right before every identification during runtime to boost the steadiness of result which is searched to avoid the redundant exposure of profile.

II. LITERATURE SURVEY

Z. Dou, R. Song, and J.-R. Wen, although personalized search has been planned for several years and plenty of personalization methods are investigated, it's still unclear whether or not personalization is systematically effective on completely different queries for various users, and beneath completely different search contexts. During this paper, we tend to study this drawback and supply some preliminary conclusions [1]. M. Spertta and S. Gach, User profiles, descriptions of user interests, are often utilized by search engines to supply personalized search results. Several approaches to making user profiles collect user data through proxy servers (capturing browsing histories) or desktop bots (capturing activities on a private computer). Each these techniques need participation of the user to put in the proxy server or the bot [3]. B. Tan, X. Shen, and C. Zhai, Long-term search history contains large data of user's search preferences, which may be used as search context to enhance retrieval performance. Information retrieval systems is important for overcoming data overload. A significant deficiency of existing retrieval systems is that they often lack user modeling is not depending on adaptive to individual users, leading to inherently non-optimal retrieval performance [4]. The existing profile-based customized web Search don't support runtime profiling. A user profile is usually generalized for once offline, and wont to modify all queries from a same user indiscriminatingly. This one profile fits all method has actually drawbacks given the variability of queries. One proof reported in is that profile based customization might not even facilitate to improve the search quality for a some ad hoc queries, although to get possessing user profile to a server has place the user's privacy in hazard. The existing ways don't take under consideration the customization of privacy needs. This feasibly makes some user privacy to be overprotected whereas others insufficiently protected [7], [9]. For instance, in all the sensitive topics are detected exploitation an absolute metric known as surprisal supported the knowledge theory, making assumption that the interests with less user document support are a lot of sensitive. However this assumption will be doubted with an easy counter example: If a user contains a number of documents regarding word sex and it is finds that sex is very general and not sensitive, despite the reality that is frontal. Unfortunately, few previous work will effectively address individual privacy wants throughout the generalization [10]. Many customization techniques need repetitive user interactions once making customized search results. They ideally filter the searching results with metrics that require multiple user interactions, like rank rating, average rank, and so on. This paradigm is, however, unfeasible for runtime profiling, because it won't solely create an enormous amount of hazard of privacy breach, however additionally demand prohibitory time interval for profiling. Thus, we want predictive metrics to measure the search quality and breach risk when personalization, while not acquisition repetitive user interaction [12]. We offer a privacy-preserving customized web search framework UPS, which may generalize profiles for every query per user-specified privacy needs [1]. Hoping on the definition of two conflicting metrics, specifically customization utility and privacy hazard for tree structure of user profile, we have a tendency to formulate the matter of privacy preserving in customized search as hazard Profile Generalization, with its NP-hardness verified [2], [3]. We are implementing two easy however effective generalization algorithms, such as GreedyDP and GreedyIL, in order to support runtime profiling. Whereas the last tries to maximize the differentiating power (DP) and the second algorithm makes an attempt order to reduce the information loss (IL). By feating variety of heuristics, GreedyIL outperforms GreedyDP rank [12], [14]. We provide a cheap mechanism for the client to come to a decision whether or not to modify a query in UPS. This judgment is created before every runtime profiling to boost the steadiness of the given search results in order to avoid the extra exposure of the profile [15].

III. SYSTEM OVERVIEW

To generalize user profile at client side by using greedy algorithm. This profile generalization depends on two metrics: a)using information from user's profile b) due to risk arrived protection of that information. Our main goal is that there should be less risk to disclose the sensitive information present in the profile as per the user's expectations as well as to improve better search results .

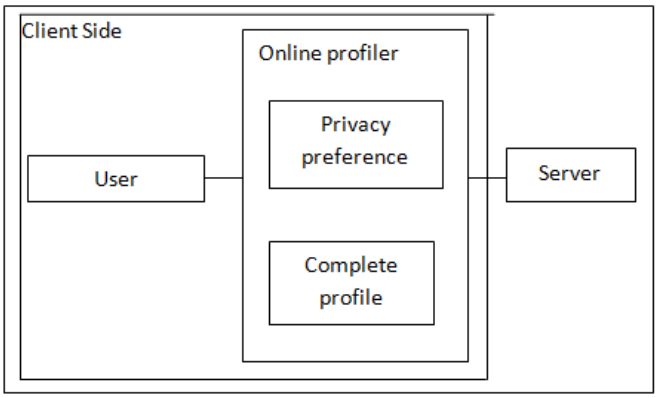


Fig.2 System Overview

Figure.2 shows the idea of complete system which is used to generate generalize profile at client side. Thus we can give freedom to user to select profile contents for searching required information.

Module Description

Personalize Web Search System with Privacy Protection system mainly consists of four modules.

1. Generation of offline profile
2. Privacy requirements of offline detection.
3. Generation and topic prioritization
4. Take online decision

1) Generation of offline profile

User uses any types of services provided by search engine like mail, map, You Tube, searching, bing, chat provided by different search engines. For accessing these services for first time registration of user is completed by filling user's profile. This user profile includes user's interests for registration, his or her personal information, photos, address, contact numbers, and much more. This detailed profile is generated only once. After profile filled up initial support for each topic or information present in the profile is calculated.

2) Privacy requirements of offline detection

For the support of information present in profile, threshold value of support is decided. If in present profile information any information's support is below of decided threshold then that respective topic is declared as sensitive topic. A set of all such sensitive topics is created. This is considered as risk of profile expositions which may be sensitive and interferes user's privacy.

3) Generation and topic prioritization

After generating offline profile only once, next step is to generalize that profile. It involves the creation of user profile every time according to user's requirement and sensitivity of information present in the topic. Whenever any query is generated by the user for any search, client profile generator computes a sub tree. For computing this sub tree.

Taxonomy depend on algorithm is used. Topics consists of present sub tree according to user's interest from the user's detailed profile. For this difference between selected sub tree support and risk factor is calculated, on the basis of individual profile data present in the list. If this difference is above threshold value the risk of maximum profile exposure will be reduced. Thus from the third step we get the customized profile from user's detailed profile.

4) Take online decision

After obtaining customized profile the decision is made by this step whether to send this profile to server for searching further information. If user is willingly ready to expose topics presented by generalized profile then this profile is finalized and profile is sent to server for search information accordingly.

IV. ALGORITHM

Greedy algorithm

A greedy algorithmic rule is associate degree algorithmic rule that follows the matter resolution heuristic of creating the domestically optimum alternative at every stage with the hope of finding a worldwide optimum. Greedy algorithmic rule considers simple to implement and straightforward approach and decides next step that give useful result. In several issues, a greedy strategy doesn't turn out associate degree optimum resolution, however a greedy heuristic yields domestically optimum resolutions that approximate a worldwide optimum solution in an exceedingly cheap time. We begin by introducing a brute-force optimum algorithm, which is proved to be NP-hard. Then, we have a tendency to propose 2 greedy algorithms, specifically the GreedyDP and GreedyIL.

A. The Brute-Force rule

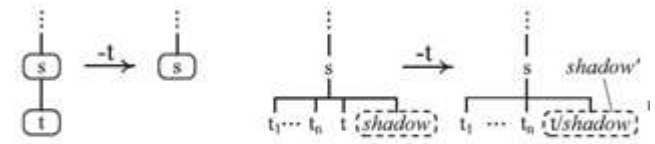
The brute-force rule exhausts all doable nonmoving subtrees of a given seed profile to seek out the optimum generalization. The privacy needs are respected throughout the exhaustion. The subtree with the optimum utility is chosen because the result. Though the seed profile G0 is considerably smaller than H, the exponential process quality of brute-force rule continues to be unacceptable.

B. GreedyDP algorithm

Fig.3 shows it works in an exceedingly bottom up manner. Beginning with the leaf node, for each iteration, it chooses leaf topic for pruning so making an attempt to maximize utility of output. Throughout iteration a best profile-so-far is maintained satisfying the danger constraint. The iteration stops once the basis topic is reached. The simplest profile-so-far is that the conclusion. GreedyDp algorithms need recomputation of profiles that adds up to process value and memory demand.

C. GreedyIL algorithm

GreedyIL algorithmic rule improves generalization potency. GreedyIL operate priority queue for candidate prune leaf operator in descending order. This decreases the process value. GreedyIL states to terminate the iteration once danger is satisfied or once there is one leaf left. Since, there is less process value compared to GreedyDP, GreedyIL outperforms GreedyDP.



(a) C1: t has no siblings. (b) C2: t has siblings $t_1 \sim t_n$. Pruning t causes the preference value of t to be shifted into that of the shadow node.

Fig.3 Two Cases of prune-leaf on a leaf t.

Algorithm of projected style

The GreedyIL formula improves the potency of the generalization exploitation heuristics supported many findings. One necessary finding is that any prune-leaf operation reduces the discriminating power of the profile. In different words, the DP displays monotony-city by prune-leaf. The advantages of creating the above runtime call area unit, it enhances the steadiness of the search quality and it avoids the supernumerary exposure of the user profile. Therefore, GreedyIL is predicted to considerably outmatch Greedy DP. The steps for GreedyIL formula area unit

Step 1: If G' could be a profile obtained by applying a prune leaf operation on G, then DP (q, G) ≥ DP (q, G').
Step 2: Specifically, every candidate operator within the queue could be a tuple like op = (t, IL (t, Gi)), wherever t is that the leaf to be cropped by op and IL (t, Gi), indicates the IL incurred by Pruning t from Gi.
Step 3: The reiterative method will terminate whenever ϑ- risk is happy.
Step 4: The second term (TS (q, G) remains unchanged for any pruning operations till a Single leaf is left (in such case the only altemative for pruning is that the single leaf itself).
Step 5: In C1, t could be a node with no siblings, and In C2, t could be a node with siblings. The case C1 is straightforward to handle. However, the analysis of IL just in case C2 needs introducing a Shadow relation of t.
Step 6: whenever if we have a tendency to arrange to prune t, we tend to really merge t into shadow to get a replacement shadow leaf shadow0, beside the preference of t,
Step 7: Prune-leaf solely operates on one topic t. Thus, it doesn't impact the IL of different candidate operators in letter of the alphabet. Whereas just in case C2, pruning t incurs recomputation of the preference values of its relation nodes.
Step 8: Once a leaf topic t is cropped, solely the candidate operators pruning t's relation topics ought to be updated in letter of the alphabet. In general, GreedyIL traces the knowledge loss rather than the discriminating power. This protects lots of process value. The benefits increased Privacy Protection Framework is as follows:

- It enhances the steadiness of the search quality
- Improves the privacy protection against completely different sort of attacks
- It avoids the supemumerary exposure of the user profile
- It provides runtime identification

V. CONCLUSION

This paper describes a client-side personal protection framework called UPS for personalized web search. Any PWS that catches user profile UPS could not potentially adopted a hierarchical profile. The framework gives permission to user for specifying customized personal requirements. To protect the personal privacy without compromising the search quality UPS has to performed online generalization, there are two Greedy algorithm, first one Greedy DP, second one Greedy IL for the online generalization. While protecting user personal requirements our experimental results shows that UPS could achieve quality search result. The effectiveness and efficiency of our answers confirmed with the result. With broader background knowledge like a rich relationship among topics we will try to resist adversaries (example. Exclusiveness, Sequentiality etc) or a series of queries to capture (for the purpose of user profile we will also use more sophisticated method for performance of UPS)

VI. REFERENCES

[1] 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.
[2] 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.
[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] 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.
[13] J.S. Breese, D. Heckerman, and C.M. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.
[14] 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.
[15] A. Pretschner and S. Gauch, "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.

Prateek C. Shukla is engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under Savitribai Phule Pune University. His interest in the field of Data Mining.

Tekchand D. Patil is engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under Savitribai Phule Pune University. His interest in the field of Data Mining.

Yogeshwar J. Shirsath is engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under Savitribai Phule Pune University. His interest in the field of Data Mining.

Dnyaneshwar N. Rasal is engineering student of Information Technology at Brahma Valley College of Engineering And Research Institute, Nasik under Savitribai Phule Pune University. His interest in the field of Data Mining.

Prof. S. R. Jadhav, BE Computer Engg. Was educated at Savitribai Phule Pune University. Presently he is working as Prof. in Information Technology Department of Brahma Valley College of Engineering and Research Institute, Nasik, Maharashtra, India. He has presented papers at International conferences and also published papers in International Journals on various aspects of Computer Engineering and Networks. His areas of interest include Computer Networks Security and Advance Database.