International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Personalizing Search Based on User Profile by Using Anonymization

Abhilasha V. Biradar[1], K. B. Sadafale[2]

[1]Dept. of Information Technology, Sinhgad College of Engineering, Pune, India

[2]Professor, Dept. of Information Technology, Sinhgad College of Engineering, Pune, India

Abstract: ***The search engine becomes the most important gateway for ordinary people who are looking for useful information on the web. In spite of, users might sense failure when search engines return inappropriate results that do not meet their real objectives. Such inappropriateness is largely due to the variety of users' contexts and backgrounds, as well as the uncertainty of texts. Personalized web search (PWS) is a type of search techniques which aims at providing better search results, which are restricted to individual user needs. The existing web search does not support runtime profiling. A user profile is mostly generalized for only once offline and used to personalize all queries from the same user. The existing methods do not concern for the customization of privacy requirements. Because of that, some user privacy is overprotected while others partly protected. The proposed UPS framework generalizes profiles for each query given to user-specified privacy specification. Online generalization on user profiles is performed to protect the personal privacy without compromising the search quality. K-anonymization technique is used to anonymize attributes of users profile like age and zip code. This proposed technique improves the security of user profile.***

Keywords: Personalized Web Search(PWS), Generalization, K-anonymization.

1. Introduction

World Wide Web (WWW) is very popular and commonly used internet's information retrieval service. Now-a-days commonly used task on the internet is web search. The user gets a variety of related information for their queries. To provide more relevant and effective results to the user, personalization technique is used. For a given query, a personalized Web search (PWS) can provide different search results for different users or organize search results differently for each user, based on their interests, preferences, and information needs. The solutions to PWS can be categorized into two types, namely click-log-based methods and profile-based ones. The click-log based methods simply impose bias to clicked pages in the user's query history. It can only work on repeated queries from the same user, which is a strong limitation of this method. In contrast, profile-based methods improve the search experience with complicated user-interest models generated from user profiling techniques. User profiles may include demographic information, e.g., name, age, country, education level, etc, and may also represent the interests or preferences of either a group of users or a single person. There are two types of the user profile: Short-term profiles represent the user's current interests where as long-term profiles indicate interests that are not subject to frequent changes over time. Besides the personalized results, security required in the personalized web search. Users are not interested to expose their information during the web search. This has become a significant concern in profiling the user in personalized web search. There should be a mechanism which recognizes profiles according to information given by the user. Hence, search engines should give security mechanism such that user will assured of its privacy and its information should be kept secure.

The rest of the paper is organized as follows: Section II summarizes the related work. Section III presents the proposed methodology. In Section IV results and discussion are described, and Section VI states the conclusion and future scope.

2. Related Work

In [1], author proposes a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user specified privacy requirements. The 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. Here two greedy algorithms are presented, namely GreedyDP and GreedyIL, for runtime generalization.

In [2], two simple but effective generalization algorithms for user profiles are developed which allows query-level customization using proposed metrics and also provide an online prediction mechanism based on query utility for deciding whether to personalize a query in UPS.

Both [3] and [5] provide online anonymity 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.

Krause and Horvitz in [4] employ statistical techniques to learn a probabilistic model and then use this model to generate the near-optimal partial profile.
In [6], the useless user profile (UUP) 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 at large.

Teevan et al. [7] collect a set of features of the query to classify queries by their click entropy. He first examined the

variability in user intent for a large number of queries using both implicit and explicit measures. Then the study was carried out to show variation in the implicit measures predicts variation in the explicit measures, and look at what other factors can account for variation in the implicit measures.

Gan et. al [8] suggested that search queries can classified into two types, content (i.e., non-geo) and location (i.e., geo). A classier was built to classify geo and non-geo queries, and the properties of geo queries studied in detail. It was found that a significant number of queries were location queries focusing on location information.

P. Palleti et al. [9] by using probabilistic query expansion author developed personalized web search. In this approach, the authors developed a personalized Web search system applied at proxy which changes to user interests perfectly by generating user profile with the use of collaborative filtering.

In [10], the author studied the existing generalization methods are insufficient because they cannot provide assurance privacy protection in all cases, and frequently acquire redundant information loss by performing too much generalization. In this paper, the author proposes the idea of personalized secrecy, and develops a new generalization structure that takes into account customized privacy necessities.

Chirita et al. [11] exploit rich models of user interests, built from both search-related information and other information about the user, including documents and e-mails that the user has read and created.

In [12], the authors discussing how metadata can be used to personalize search and then show that personalized search using ODP and other directory metadata is feasible already today. The author focused on two ways of doing this, first by directly using ODP and similar metadata, and second by biasing on and thus automatically extending these metadata to the whole web.
In [13], the author highlights the significance of studying the evolving nature of the Web personalization. Web usage mining is used to discover interesting user navigation patterns and can applied to many real-world problems, such as improving Web sites/pages, making an additional topic or product recommendations, user/customer behavior studies, etc. A Web usage mining system performs five major tasks: i) data gathering, ii) data preparation, iii) navigation pattern discovery, iv) pattern analysis and visualization, and v) pattern applications. Each task explained in detail, and its related technologies are introduced.

Susan T. Dumais et al. [14] introduces a search algorithm that considers the 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. 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.

Sugiyama et al. [15] proposed a system which monitors the user's browsing history and updates his/her profile whenever his/her browsing page changes. When the user submits a query the next time, the search results adapt based on his/her user profile. They have constructed each user profile based on the following two methods: (i) Pure browsing history, and (ii) Modified collaborative filtering.

3. Proposed Methodology

3.1 Overview

Architecture diagram explains the system overview, as to how the system works in real. As shown in fig.1, when a user issues a query q on the client, the proxy generates a user profile in runtime in the light of query terms. The output of this step is a generalized user profile G satisfying the privacy requirements. The generalization process guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles. Here, the proposed system takes user personal information for personalized web search (PWS), like their interests. If consider users age, postal code (Address), results can be retrieved based on the user's age category, like middle age group people what are they willing to search, so on so Pune people what they willing to search like that. But for the security of the personal profile information, data is anonymized like to k- diversity.

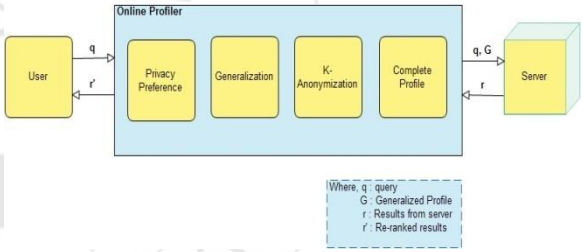

Figure 2: Proposed system architecture

This section includes the partitioning of the project into different modules. All these modules explained as follows:

1) Offline profile construction- The first step of the offline processing is to build the original user profile in a topic hierarchy H that reveals user interests. We assume that the user's preferences represented in a set of plain text documents, denoted by D. To construct the profile, we take the following steps:
   a) Detect the respective topic in R for every document d ϵ D. Thus, the preference document set D is transformed into a topic set T.
   b) Construct the profile H as a topic-path trie with T, i.e., H = trie(T).
   c) Initialize the user support $sup_H(t)$ for each topic t ϵ T with its document support from D, then compute $sup_H(t)$ of other nodes of H.
2) Privacy Requirement Customization- This procedure first requests the user to specify a sensitive-node set S $\subset$ H, and the respective sensitivity value sen(s) > 0 for each topic s ϵ S. Next, the cost layer of the profile is generated by computing the cost value of each node t ϵ H as follows:
   a) For each sensitive-node, cost(t) = sen(t);
   b) For each nonsensitive leaf node, cost(t) = 0;

c) For each nonsensitive internal node, cost(t) is recursively given in a bottom-up manner.

Till now, we have obtained the customized profile with its cost layer available. When a query q is issued, this profile has to go through the following two online procedures:

3) The Greedy DP Algorithm- A more practical solution would be a near-optimal greedy algorithm. An operator $\xrightarrow{-t}$ called prune-leaf is introduced, which indicates the removal of a leaf topic t from a profile. Formally, the process of pruning leaf t from G_i to obtain G_{i+1} is denoted by $G_i \xrightarrow{-t} G_{i+1}$. Obviously, the optimal profile G* can be generated with a finite-length transitive closure of prune-leaf. The greedyDP algorithm works in a bottom-up manner. Starting from G_0, 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, a best profile-so-far, which indicates the G_{i+1} having the highest discriminating power is maintained. 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.
4) The GreedyIL Algorithm- The GreedyIL algorithm improves the efficiency of the generalization 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. Formally, if G' is a profile obtained by applying a prune-leaf operation on G, then DP(q, G) ≥ DP(q, G'). Considering operation $G_i \xrightarrow{-t} G_{i+1}$ in the i^{th} iteration, maximizing DP(q, G_{i+1}) is equivalent to minimizing the incurred information loss, which is defined as DP(q, G_i) – DP(q, G_{i+1}). This finding motivates to maintain a priority queue of candidate prune-leaf operators in descending order of the information loss caused by the operator. Specifically, each candidate operator in the queue is a tuple like op = {t, IL(t, G_i)}, where t is the leaf to be pruned by op and IL(t, G_i) indicates the IL incurred by pruning t from G_i. This queue, denoted by Q, enables fast retrieval of the best-so- far candidate operator.

A. Proposed Technique

In the context of k-anonymization problems, a database is a table with n rows and m columns. Each row of the table represents a record relating to a specific user of a proposed system, and the entries in the various rows need not be unique. The values in the various columns are the values of attributes associated with the users of the system.

The k-anonymity proposal focuses on two techniques: generalization and suppression, which, unlike other existing techniques, such as scrambling or swapping, preserve the truthfulness of the information.

1) Suppression: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. In the anonymized table below, we have replaced all the values in the 'Zip Code' attribute with a '*'.
2) Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' , etc. Generalization consists in substituting the values of a given attribute with more general values. To this purpose, the notion of a domain (i.e., the set of values that an attribute can assume) is extended to capture the generalization process by assuming the existence of a set of generalized domains. The set of original domains together with their generalizations is referred to as Dom. Each generalized domain contains generalized values, and there exists a mapping between each domain and its generalizations. For instance, postal addresses can be generalized to the street (dropping the number), then to the city, to the county, to the state, and so on. This mapping is stated using a generalization relationship $\leq_D$. Given two domains D_i and $D_j \in$ Dom, $D_i \leq_D D_j$ states that values in domain D_j are generalizations of values in D_i. The generalization relationship $\leq_D$ defines a partial order on the set Dom of domains, and is required to satisfy the following conditions:

C1: For all D_i, D_j, $D_z \in$ Dom:
$D_i \leq_D D_j$,$D_i \leq_D D_z$ => $D_j \leq_D D_z \vee D_z \leq_D D_j$
C2: all maximal elements of Dom are singleton.

Condition C1 states that for each domain D_i, the set of domains generalization of D_i is totally ordered and, therefore, each D_i has at most one direct generalization domain D_j. It ensures determinism in the generalization process. Condition C2 ensures that all values in each domain can always be generalized to a single value.

Thus leaking information that was not intended for disclosure is prevented.

4. Results and Discussion

Fig 2 shows the output of GreedyDP algorithm which shows the path of sensitive nodes of a user profile.

Figure 2: GreedyDP Screen

Fig 3 shows the outcome of a GreedyIL algorithm which causes minimum information loss.

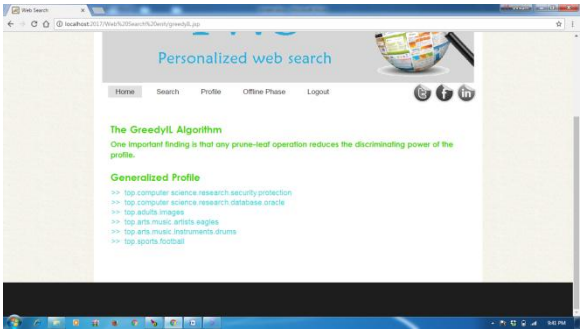

Figure 3: GreedyIL Screen

Fig 4 shows the generalization of user profile using GreedyDP and GreedyIL algorithm and anonymization on age and zip code.

Figure 4: Generalization Screen

The performance of proposed system is evaluated based on present results and proposed results. Proposed and present solution results are represented by using two colors namely blue and red respectively.

Figure 5: Comparison Graph

Present solutions contain results which are retrieved from the repository when a user enters a keyword in UPS search bar. When the user clicks on links in present solution, a rank of the clicked link gets increased, and this link appears at the top in proposed solutions. Proposed solutions contain present solution results as well as the ranked links on the top which are clicked by user frequently. Proposed solutions recommend user interested links. Thus, proposed solution results are better than the present.

5. Conclusion and Future Scope

The proposed system presented a client-side privacy protection framework called UPS for personalized web search. UPS framework could potentially be adopted by any PWS that gathers user profiles in a hierarchical taxonomy. The framework allowed users to specify their customized privacy requirements via the hierarchical profiles. In the existing system, they didn't consider user personal profile information for PWS, like age, postal code. Here, the proposed system takes user personal information for personalized web search (PWS), like their interests and for the security of the personal profile information, the system anonymized the data like to k-diversity. Also, UPS performed runtime generalization on user profiles to protect their privacy without compromising the search quality.

In future work, user security can be improved by using different parameters and techniques except the parameters and technique used in proposed work.

6. Acknowledgment

The authors would like to thank the publishers, researchers for making their resources available and teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Finally, we would like to extend heartfelt gratitude to friends and family members.

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, 2014.

[2] 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.

[3] 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.

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

[5] Xu, Yabo, et al. "Online anonymity for personalized web services," Proceedings of the 18th ACM conference on Information and knowledge management, ACM, 2009.

[6] 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.

[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] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, "Analysis of geographic queries in a search engine log," Proc. of the International Workshop on Location and the Web, 2008.

[9] P. Palleti, H. Karnick and P. Mitra, "Personalized Web Search Using Probabilistic Query Expansion," International Conferences on Web Intelligence and Intelligent Agent technology Workshops (IEEE/WIC/ACM), Pp. 83 – 86, 2007.

[10] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc.ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
[11] P.A. Chirita, C. Firan, and W. Nejdl, "Summarizing Local Context to Personalize Global Web Search," Proc. ACM Int'l Conf. Information and Knowledge Management (CIKM), 2006.
[12] P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 178-185, 2005.
[13] A. J. Ratnakumar, "An Implementation of Web Personalization Using Web Mining Techniques," Journal of Theoretical and Applied information technology, 2005.
[14] 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 '05), pp. 449-456, 2005.
[15] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive Web Search Based on User Profile Constructed without Any Effort from Users," Proc. 13th Int'l World Wide Web Conf. (WWW '04), pp. 675-684, 2004.