ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Supporting Privacy Preservation technique for Personalized Web Search Based on User History

Ms.V.Manga
Final M.Tech Sudent,
Dept of Computer Science and Engineering,
Avanthi Institute of Engineering & Technology,
Narsipatnam, Andhra Pradesh.

Ms.V.Sunitha
Assistant Professor,
Dept of Computer Science and Engineering,
Avanthi Institute of Engineering & Technology,
Narsipatnam, Andhra Pradesh.

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 user-specified 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.

Introduction:

Huge amount of information gets added to the Web every day. Publicly visible text creation is of the order of 10 GB per day and private text creation (including user email, IM messages, tags, reviews etc) is of the order of 3 terabytes per day. This rapidly increasing scale of the web is in many ways limiting the utility of the web. There is a high level of noise beginning from spam and ending with a lot of uninteresting, irrelevant and duplicated content. Search engines and other forms of ranking are unable to keep up with this. Recently, search engines have started showing Wikipedia links as the top search result because ranking has become very hard. Personalized search is a promising way to improve the accuracy of web search, and has been attracting much attention recently.

However, effective personalized search requires collecting and aggregating user information, which often raises serious concerns of privacy infringement for many users. Indeed, these concerns have become one of the main barriers for deploying personalized search applications, and how to do privacy-preserving personalization is a great challenge. The web search engine has long become the most important portal for ordinary people looking for useful information on the web. However, users might experience failure when search engines return irrelevant results that do not meet their real intentions. Such irrelevance is largely due to the enormous variety of users' contexts and backgrounds, as well as the ambiguity of texts.

A. Why Privacy Protection needed?

During the search process it considers two contradicting effects in order to provide privacy protection in user profile based PWS. Considering personalization utility of the user profile which attempt to improve the search quality. On the other hand, they need to hide the privacy contents that exist in the user profile to control the privacy risk.[4]Personalized Web Search (PWS) is a general category of search techniques which aims to provide better search results, according to the individual user needs. So, for this user information has to be collected and analyzed so that the perfect search results required for the user behind the issued query is to be given to the user. The solution to this is Personalized Web Search (PWS). Personalized search is a promising way to improve search quality by customizing search results for people with different information goals. Many recent research efforts have focused on this area.

Most of them could be categorized into two general approaches: Re-ranking query results returned by search engines locally using personal information; or sending personal information and queries together to the search engine. A good personalization algorithm relies on rich user profiles and web corpus. However, as the web corpus is on the server, re-ranking on the client side is bandwidth intensive because it requires a large number of search results transmitted to the client before re-ranking. Alternatively, if the amount of information transmitted is limited through filtering on the server side, it pins high hope on the existence of desired information among filtered results, which is not always the case. Therefore, most of personalized search services online like Google Personalized Search and Yahoo! My Web adopt the second approach to tailor results on the server by analyzing collected personal information, e.g. personal interests, and search histories.

II. Literature review:

2.1 Profile-Based Personalization:

Earlier techniques utilize term lists/vectors or bag of words to represent their profile. However, most recent works build profiles in hierarchical structures due to their stronger descriptive ability, better scalability, and higher access efficiency. The majority of the hierarchical representations are constructed with existing weighted topic hierarchy/graph, such as ODP , [1], [2], [3], [5], and so on. Another work in [10] builds the hierarchical profile automatically via term-frequency analysis on the user data. As for the performance measures of PWS in the literature, Normalized Discounted Cumulative Gain (nDCG) [8] is a common measure of the effectiveness of an information retrieval system. It is based on a human graded relevance scale of item-positions in the result list, and is, therefore, known for its high cost in explicit feedback collection. To reduce the human involvement in performance measuring, researchers also propose other metrics of personalized web search that rely on clicking decisions, including Average Precision (AP) [9], Rank Scoring, and Average Rank [3], [4].

2.2 Privacy Protection in PWS System:

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 in [5]. Typical works in the literature of protecting user identifications (class one) try to solve the privacy problem on different levels, including the pseudo-identity, the group identity, no identity, and no personal information. Solution to the first level is proved to fragile. The third and fourth levels are impractical due to high cost in communication and cryptography. Therefore, the existing efforts focus on the second level. Both and 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[6].

In , 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. Viejo and Castell_a-Roca [2] use legacy social networks instead of the third party to provide a distorted user profile to the web search engine. In the scheme, every user acts as a search agency of his or her neighbours. They can decide to submit the query on behalf of who issued it, or forward it to other neighbours. The shortcomings of current solutions in class one is the high cost introduced due to the collaboration and communication.

III. Problem Definition:

The existing personalized web search on profile based was concentrated on server-side as the results of the search engines are common to all the users and it provides a less security to the user. In many studies the click based method proved they simply impose bias to clicked pages inthe user’s query history.

It can only work on repeated queries from the same user, which is a strong limitation, but the privacy protection was poor as the contents as been lost. The profile based method also provided better personalized relevant searches but the drawback is that it does not support runtime profiling and used to personalize all queries from a same user. [3]

IV. Proposed System:

- 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, we formulate the problem of privacy-preserving personalized search as #-Risk Profile Generalization, with its N P-hardness proved.

- Two simple but effective generalization algorithms are introduced to support runtime profiling: Greedy DP and Greedy IL. 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, Greedy IL out performs Greedy DP significantly.

- Using this we get an inexpensive mechanism 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.

- Our extensive experiments demonstrate the efficiency and effectiveness of our UPS framework.

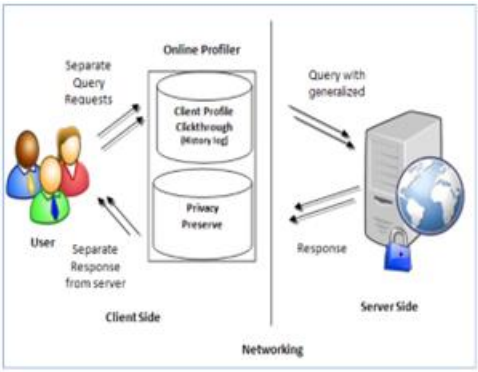

Fig: System architecture of UPS

4.1. The GreedyDP Algorithm:

The first greedy formula GreedyDP [11]works in an exceedingly bottom up way .The unvarying technique terminates once the profile is generalized to a root-topic. The best-profile-so-far square measure the final results of the rule. the most disadvantage of GreedyDP is that it desires recompilation of all candidate profiles (together with their discriminating power and privacy risk) generated from makes an attempt of prune-leaf. This causes important memory wants and machine worth. GreedyDP formula works on 2 key ingredients they are:

4.1.1. Optimal sub-structure:

A best answer to the complete downside contains inside it best solutions to sub issues (this is additionally true of dynamic programming)

4.1.2. Greedy Choice Property:

Greedy choice + Optimal sub-structure gives the correctness of the greedy algorithm

4.2. The GreedyIL Algorithm:

The GreedyIL rule [11] improves the efficiency of the generalization practice heuristics supported several findings. One necessary finding is that any prune-leaf operation reduces the discriminating power of the profile. In alternative words, the exile displays monotonicity by prune-leaf. GreedyIL algorithmic rule selection Properties are:

4.2.1. Locally Optimal Choice:

– Make optimistic choice available at a given moment.

4.2.2. Locally Optimal Choice Globally Optimal Solution:

– In other words, the selection of greedy is always safe.

–To prove this algorithm Exchange Argument are used.

4.2.3. Contrast With Dynamic Programming:

– Choice at a given step may be depend on solutions to sub problems (bottom-up)

B. Implementation Issues:

There exist some open problems in the UPS process, this can be solved using a mechanism called an inverted-indexing mechanism for computing the query topic relevance. The publicly available repositories permit the editing as well as manual tagging on each topic. These topics contain textual data which consist of a document repository, which allows each leaf topic to identify its associated document set. Each document in document repository is assigned to one leaf topic only. Thus, it is possible to generate an inverted-index for each leaf topic, which contains entries such as term; doc id; topic id for all the documents. At the end, a hierarchy of inverted indices is obtained, where all the documents within the taxonomy will be contained in the inverted index file. Thus this structure enables each user to resourcefully process keyword search and retrieval. Specifically, the root index files are able to maintain the entire document set that can sustain term-based topic searching in repository. During the Offline-1 procedure, it is needed to detect for each document the respective topic in repository. For this a naive method is to compute the relevance for each pair of document and their topic to repository with a discriminative naive Bayesian classifier (dnb). The topic that exhibits with the largest dnb value is considered the result of the search. But, if many of the topics in repository are not relevant to the documents then the naive method is inefficient. Exploiting the user's click log to be the set of document will be a more efficient way (and the one used in this implementation). The click log contains entries such as query in the log and document clicked by the user after issuing a query. Thus, this allows reducing the necessity of computing the topics that are retrieved by the query from the topmost inverted index and then all documents relevant to the query are retrieved from the inverted index and their associated topics are obtained from the topic id. Then, the dnb value for each topic is computed

V. Conclusion:

This paper presented a client-side privacy protection framework called UPS for personalized web search. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. 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. We proposed two greedy algorithms, namely Greedy-DP and Greedy-IL, for the online generalization. Our experimental results revealed that UPS could achieve quality search results while preserving user's customized privacy requirements. The results also confirmed the effectiveness and efficiency 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 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.

V. 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. Privacy Preservation in Personalized Web Search International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163 Issue 10, Volume 2 (October 2015).

3. PRIVACY PRESERVING WEB SEARCH BASED ON USER HISTORY International Journal For Technological Research In Engineering Volume 3, Issue 1, September-2015 ISSN (Online): 2347 – 4718.

4. Privacy Protection in Personalized Web Search Using UPS International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869, Volume-3, Issue-3, March 2015.

[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] 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.

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

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

[9] 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.

[10] 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.

[11]. Nivi.A.N1, Vanitha.S2, Saranya.K.R3 and Yamini.S4 Privacy Protection using Various Algorithms in Personalized Web Search. INTERNATIONAL JOURNAL FOR RESEARCH IN EMERGING SCIENCE AND TECHNOLOGY, VOLUME-1, ISSUE-6, NOVEMBER-2014.