International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Security Approach for Personalized Web Search Framework

Farha Naaz[1], **Asma Parveen**[2]

[1]P G Student, Department of Computer Science and Engg, Khaja Banda Nawaz College of Engineering, Gulbarga, Karnataka, India

[2]H.O.D, Department of Computer Science and Engg, Khaja Banda Nawaz College of Engineering, Gulbarga, Karnataka, India

Abstract: ***Web Search engines (e.g. Google,Yahoo,Microsoft,Live Search, etc) are widely used to find certain data among a huge amount of information in a minimal amount of time.These useful tools also pose a privacy threat to the users. Web search engine profile their users on the basis of past searches submitted by them. 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. 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: Mining, Query, Personalized, Web, Search, Greedy, Profile, Risk

1. Introduction

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. Personalized web search (PWS) is a general category of search techniques aiming at providing better search results, which are tailored for individual user needs. As the expense, user information has to be collected and analyzed to figure out the user intention behind the issued query.

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, 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 potentially effective for almost all sorts of queries, but are reported to be unstable under some circumstances.

Although there are pros and cons for both types of PWS techniques, the profile-based PWS has demonstrated more effectiveness in improving the quality of web search recently, with increasing usage of personal and behavior information to profile its users, which is usually gathered implicitly from query history , browsing history , click-through data , bookmarks, user documents, and so forth. Unfortunately, such implicitly collected personal data can easily reveal a gamut of user's private life. Privacy issues rising from the lack of protection for such data, for instance the AOL query logs scandal, not only raise panic among individual users, but also dampen the data-publisher's enthusiasm in offering personalized service. In fact, privacy concerns have become the major barrier for wide proliferation of PWS services.

2. Literature Survey

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.
for Personalized Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM), 2005.

1. A Large-Scale Evaluation and Analysis of Personalized Search Strategies

Although personalized search has been proposed for many years and many personalization strategies have been investigated, it is still unclear whether personalization is consistently effective on different queries for different users, and under different search contexts. In this paper, we study this problem and provide some preliminary conclusions. We present a large-scale evaluation framework for personalized search based on query logs, and then evaluate five personalized search strategies (including two click-based and three profile-based ones) using 12-day MSN query logs. By analyzing the results, we reveal that personalized search has significant improvement over common web search on some queries but it has little effect on other queries . It even harms search accuracy under some situations. Furthermore, we show that straightforward click-based personalization strategies perform consistently and considerably well, while profile-based ones are unstable in our experiments. We also reveal that both longterm and short-term contexts are very important in improving search performance for profile-based personalized search strategies.

2. Personalized search based on user search histories

User profiles, descriptions of user interests, can be used by search engines to provide personalized search results. Many approaches to creating user profiles collect user information through proxy servers (to capture browsing histories) or desktop bots (to capture activities on a personal computer). Both these techniques require participation of the user to install the proxy server or the bot. In this study, we explore the use of a less-invasive means of gathering user information for personalized search. In particular, we build user profiles based on activity at the search site itself and study the use of these profiles to provide personalized search results. By implementing a wrapper around the Google search engine, we were able to collect information about individual user search activities. In particular, we collected the queries for which at least one search result was examined, and the snippets (titles and summaries) for each examined result. User profiles were created by classifying the collected information (queries or snippets) into concepts in a reference concept hierarchy. These profiles were then used to re-rank the search results and the rank-order of the user-examined results before and after re-ranking were compared. Our study found that user profiles based on queries were as effective as those based on snippets. We also found that our personalized re-ranking resulted in a 34% improvement in the rank order of the user-selected results.

3. Mining Long-Term Search History to Improve Search Accuracy

Long-term search history contains rich information about a user's search preferences, which can be used as search context to improve retrieval performance. In this paper, we study statistical language modeling based methods to mine contextual information from long-term search history and exploit it for a more accurate estimate of the query language model. Experiments on real web search data show that the algorithms are effective in improving search accuracy for both fresh and recurring queries. The best performance is achieved when using clickthrough data of past searches that are related to the current query.

4. Implicit User Modeling for Personalized Search

Information retrieval systems (e.g., web search engines) are critical for overcoming information overload. A major deficiency of existing retrieval systems is that they generally lack user modeling and are not adaptive to individual users, resulting in inherently non-optimal retrieval performance. For example, a tourist and a programmer may use the same word "java" to search for different information, but the current search systems would return the same results. In this paper, we study how to infer a user's interest from the user's search context and use the inferred implicit user model for personalized search. We present a decision theoretic framework and develop techniques for implicit user modeling in information retrieval. We develop an intelligent client-side web search agent (UCAIR) that can perform eager implicit feedback, e.g., query expansion based on previous queries and immediate result reranking based on clickthrough information. Experiments on web search show that our search agent can improve search accuracy over the popular Google search engine.

5. Automatic Identification of User Interest for Personalized Search

One hundred users, one hundred needs. As more and more topics are being discussed on the web and our vocabulary remains relatively stable, it is increasingly difficult to let the search engine know what we want. Coping with ambiguous queries has long been an important part in the research of Information Retrieval, but still remains to be a challenging task. Personalized search has recently got significant attention to address this challenge in the web search community, based on the premise that a user's general preference may help the search engine disambiguate the true intention of a query. However, studies have shown that users are reluctant to provide any explicit input on their personal preference. In this paper, we study how a search engine can learn a user's preference automatically based on her past click history and how it can use the user preference to personalize search results. Our experiments show that users' preferences can be learned accurately even from small click-history data and personalized search based on user preference yields significant improvements over the best existing ranking mechanism in the literature.

3. Existing System

The existing profile-based Personalized Web Search do not support runtime profiling. A user profile is typically generalized for only once offline, and used to personalize all queries from a same user indiscriminatingly. Such "one profile fits all" strategy certainly has drawbacks given the variety of queries. One evidence reported in is that profile-based personalization may not even help to improve the search quality for some ad hoc queries, though exposing user profile to a server has put the user's privacy at risk.

The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected. For example, in, all the sensitive topics are detected using an absolute metric called surprisal based on the information theory, assuming that the interests with less user document support 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.

Many personalization techniques require iterative user interactions when creating personalized search results. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on. This paradigm is, however, infeasible for runtime profiling, as it will not only pose too much risk of privacy breach, but also demand prohibitive processing time for profiling. Thus, we need predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

Disadvantages

- All the sensitive topics are detected using an absolute metric called surprisal based on the information theory.
- The existing methods do not take into account the customization of privacy requirements.
- Privacy issues rising from the lack of protection for such data.
- The existing profile-based PWS do not support runtime profiling.

4. Proposed System

We propose 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 itsNP-hardness proved.

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 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

Advantages:

1. It enhances the stability of the search quality.
2. It avoids the unnecessary exposure of the user profile.

5. GreedyAlgorithm

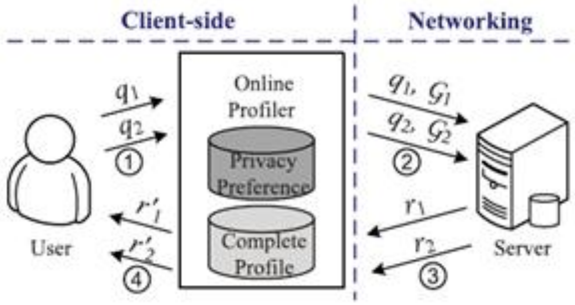
A **greedy algorithm** is a mathematical process that recursively constructs a set of objects from the smallest possible constituent parts. Recursion is an approach to problem solving in which the solution to a particular problem depends on solutions to smaller instances of the same problem.

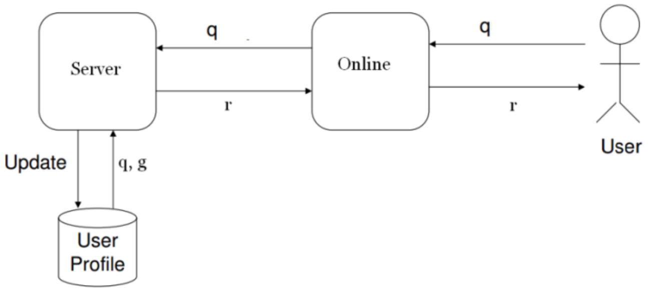
Greedy algorithms look for simple, easy-to-implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit. Such algorithms are called greedy because while the optimal solution to each smaller instance will provide an immediate output, the algorithm 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 that solutions 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 hoc mobile networking to efficiently route packets with the fewest number of hops and the shortest delay possible. They are also used in machine learning, business intelligence (BI), artificial intelligence (AI) and programming.

6. System Architecture





7. Modules

1) Profile-Based Personalization.
2) Privacy Protection in PWS System.
3) Generalizing User Profile.
4) Online Decision.

Modules Description

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.

2)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.

3) Generalizing User Profile
The generalization process has to meet specific prerequisites to handle the user profile. This is achieved by preprocessing the user profile. At first, the process initializes the user profile by taking the indicated parent user profile into account. The process adds the inherited properties to the properties of the local user 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.

4) 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.

5) Online Decision
The profile-based personalization contributes little or even reduces the search quality, while exposing the profile to a server would for sure risk the user's privacy. To address this problem, we develop an online mechanism to decide whether to personalize a query. The basic idea is straightforward. if a distinct query is identified during generalization, the entire runtime profiling will be aborted and the query will be sent to the server without a user profile.

8. 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.

References

[1] Z.Dou, R. Song, and J.-R.Wen,"ALarge-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 Searchvia Automated Analysis of Interests and Activities,"Proc.28thAnn.Int'lACMSIGIRConf.Research andDevelopmentinInformationRetrieval (SIGIR),pp. 449-456, 2005.

[3] M.SperttaandS.Gach,"PersonalizingSearchBasedonUser SearchHistories,"Proc.IEEE/WIC/ACMInt'lConf.WebIntelligence (WI), 2005.

[4] B.Tan,X. Shen,andC.Zhai,"Mining Long Term Search History to Improve Search Accuracy, "Proc. ACMSIGKDD Int' lConf. Knowledge Discovery and Data Mining (KDD),2006.

[5] [5]K.Sugiyama,K.Hatano,andM.Yoshikawa,"AdaptiveWebSearchBasedonUserProfileConstructedwithoutanyEffortfromUsers,"Proc.13thInt'l Conf. World Wide Web (WWW),2004.

[6] X.Shen, B.Tan, and C.Zhai,"Implicit User Modeling for Personalized Search,"Proc.14thACMInt'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'lACMSIGIR Conf. Research and Development Information Retrieval (SIGIR),2005.

[8] F.QiuandJ.Cho,"Automatic Identification of User Interest for Personalized Search, "Proc.15thInt'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,andT.Breuel,"PersonalizedSearch,"Comm.ACM,vol.45, no. 9, pp. 50-55, 2002.

[10] Y.Xu,K.Wang,B.Zhang,andZ.Chen,"Privacy Enhancing Personalized Web Search, "Proc.16th Int'l Conf. World Wide Web (WWW), pp. 591-600, 2007.

[11] K. Hafner, Researchers Yearn toUse AOLLogs, but They Hesitate, NewYorkTimes,Aug.2006.

[12] A. Krause and E.Horvitz,"A Utility-Theoretic Approach to Privacy in OnlineServices,"J. Artificial IntelligenceResearch,vol.39, pp. 633-662, 2010.

[13] J. S. Breese, D. Heckerman, and C. M.Kadie, "Empirical Analysis of redictive 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 Metadatato Personalize Search, "Proc. 28th Ann.Int'l ACMSIGIR Conf. Research and Development Information Retrieval (SIGIR),2005.

[15] A. Pretschner and S.Gauch, "Ontology-Based Personalized Search and Browsing, "Proc.IEEE11th Int'l Conf. Tools with Artificial Intelligence (ICTAI'99),1999