International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-4 Issue-2, May 2015

A Novel Approach for Improved Personalized Web Search with Privacy Protection

R. Ravikumar, U. Sundhar

Abstract: ***Personalized web search is a promising way to improve search quality by customizing search results for people with individual information goals. However, users are uncomfortable with exposing private information to search engines. Thus, a balance must be struck between search quality and privacy protection. The proposed system presents a scalable way for users to automatically build rich user profiles. These profiles summarize user's interests into a hierarchical organization according to specific interests. The system proposes a Personalized Web (PWS) framework called User customizable Privacy-preserving Search (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. The system presents two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization. It also provides an online prediction mechanism for deciding whether personalizing a query is beneficial.***

Keywords: ***Privacy-preserving Search, Privacy-preserving Search, Greedy DP and Greedy IL.***

I. INTRODUCTION

WOrld Wide Web has grown explicitly. It provides access to all people at any place and at any time. By this facility any one can upload or download relevant data, so that valuable content in the website can be used in all fields. As the data in the web are unstructured and semi-structured, lots of insignificant and irrelevant document are obtained as a result after navigating several links and hence data mining cannot be applied directly. For effective retrieval of web information, web mining is used. Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as: The automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognized. The goal of data mining is to unearth relationships in data that may provide useful insights. Data mining automates the process of sifting through historical data in order to discover new information.

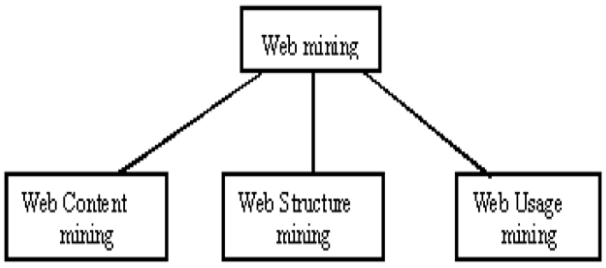
Revised Manuscript Received on 30 May 2015.
* Correspondence Author
R. Ravikumar, M.E-Scholar, Department of CSE, Thiruvalluvar College of Engineering & Technology, Tamil Nadu, India.
U. Sundhar, Asst. Prof., Department of CSE, Thiruvalluvar College of Engineering & Technology, Tamil Nadu, India.

This is one of the main differences between data mining and statistics, where a model is usually devised by a statistician to deal with a very specific analysis problem. It also distinguishes data mining from expert systems, where the model is built by a knowledge engineer from rules extracted from the experience of an expert. One of the attractions of data mining is that it makes it possible to analyses very large data sets in a reasonable time scale.
Data mining is also suitable for complex problems involving relatively small amounts of data but where there are many fields or variables to analyses. However, for small, relatively simple data analysis problems there may be simpler, cheaper and more effective solutions.

II. RELATED WORK

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Web mining can be broadly defined as discovery and analysis of useful information from the World Wide Web. There are three general classes of information that can be discovered by web mining:

- Web activity, from server logs and Web browser activity tracking.



- Web graph, from links between pages, people and other data.
- Web content, for the data found on Web
  Based on the different emphasis and different ways to obtain information, web mining can be divided into three major parts: Web Contents Mining, Web Structure Mining and Web Usage Mining.

Fig. 2.1: Web Mining Taxonomy

Web Usage Mining can be described as the discovery and analysis of user access patterns, through the mining of log files and associated data from a particular Web site. It is the type of data mining process for discovering the usage patterns from web information for the purpose of understanding and better provide the requirements of web-based applications.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

Web structure mining is a tool used to identify the relationship between web pages linked by information or direct link connection. The structured data is discoverable by the provision of web structure schema through database techniques. This connection allows a search engine to pull data relating to a search query directly to the linking web page from the web site the content rests upon. Structure mining uses minimize two main problems of the World Wide Web due to its vast amount of information.

- The first of these problems is irrelevant search results.
- The second of these problems is the inability to index the vast amount if information provided on the web.

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. Web Content Mining can be described as the automatic search and retrieval of information and resources available from millions of sites and on-line databases though search engines / web spiders. The profile-based Personalized Web Search does not support runtime profiling. A user profile is typically generalized for only once offline. Profile-based personalization may not even help to improve the search quality though exposing user profile to a server has put the user's privacy at risk. The methods do not take into account the customization of privacy requirements. They usually refine the search results with some metrics which require multiple user interactions, such as rank scoring, average rank, and so on. They need to hide the privacy contents in the user profile. To solve this problem, this project protects user privacy in profile-based PWS during the search process and also improves the search quality of the search results.

III. APPROACHES

The main objective of the system is to protect user privacy in profile-based Personalized Web Search and to improve the search quality with the personalization utility of the user profile. This system attempts to hide the privacy contents existing in the user profile to control the privacy risk and provides runtime profiling, which in effect optimizes the personalization utility while respecting user's privacy requirements.

IV. MOTIVATING EXAMPLE

A novel framework is proposed for web image re-ranking. Instead of manually defining a universal concept dictionary, it learns different semantic spaces for different query keywords individually and automatically. The semantic space related to the images to be re-ranked can be significantly narrowed down by the query keyword provided by the user. The visual and textual features of images are then projected into their related semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the semantic space of the query keyword. The semantic correlation between

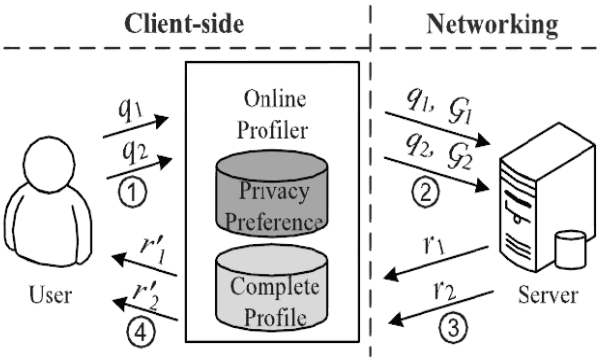


Fig. 4.1: The Circumstances Screening System Architecture

concepts is explored and incorporated when computing the similarity of semantic signatures. 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 reported 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. It probably makes some user privacy to be overprotected while others insufficiently protected. All the sensitive topics are detected using an absolute metric called surprises based on the information theory, assuming that the interests with less user document support are more sensitive. It usually refines 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, there is a need for predictive metrics to measure the search quality and breach risk after personalization, without incurring iterative user interaction.

V. PROBLEM ANALYSIS

5.1: Profile Construction:

Each user profile in UPS adopts a hierarchical structure. Our profile is constructed based on the availability of a public accessible taxonomy,

This satisfies the following assumption.

1. The repository is a huge topic hierarchy covering the entire topic domain of human knowledge. That is, given any human recognizable topic t, a corresponding node can be found in Repository
2. Given a taxonomy repository, the repository support is provided by itself for each leaf topic.

The first step is to build the original user profile in a topic hierarchy H that reveals user interests. The system assumes that the user's preferences are represented in a set of plain text documents.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

To construct the profile, take the following steps:
1. Detect the respective topic in Repository for every document d € D. Thus, the preference document set D is transformed into a topic set T.
2. Construct the profile H as a topic-path prefix tree with T, i.e., H = tire (T).
3. Initialize the user support $sup_h(t)$ for each topic t€T with its document support from D, then compute $sup_h(t)$ of other nodes of H with (4).

5.2: Privacy Requirements Customization:

Customized privacy requirements can be specified with a number of sensitive-nodes in the user profile, whose disclosure introduces privacy risk to the user. The sensitive nodes are a set of user specified sensitive topics. User's privacy concern differs from one sensitive topic to another. To address the difference in privacy concerns, allow the user to specify sensitivity for each node. Sensitivity is a positive value that quantifies the severity of the privacy leakage caused by disclosing the node. Considering the sensitivity of each sensitive topic as the cost of recovering it, the privacy risk can be defined as the total sensitivity of the sensitive nodes. The approach to privacy protection of PWS has to keep this privacy risk under control. Privacy requirement customization can be done in 2 steps as follows:
1. Requests the user to specify a sensitive-node set and the respective sensitivity value for each topic.
2. Generates the cost layer of the profile by computing the cost value of each node as follows:

- For each sensitive-node, cost(t)= sen(t)
- For each nonsensitive leaf node, cost(t)= 0
- For each nonsensitive internal node, cost (t) is recursively computed.

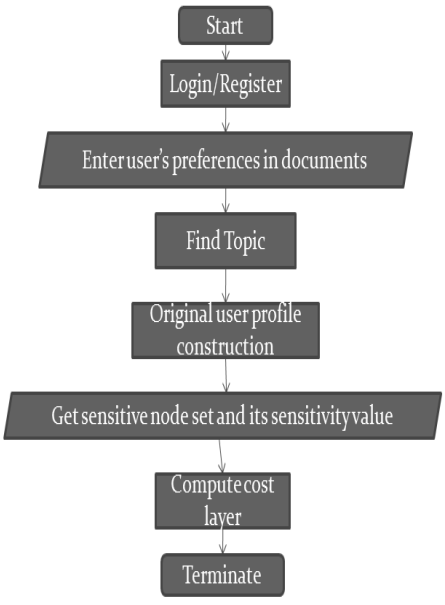


Fig. 5.1: Profile Construction and Privacy Requirements Customization

Thus, the customized profile is obtained with its cost layer available.

5.3: Query-Topic Mapping:

Given a query q, the purposes of query-topic mapping are to compute a rooted sub tree of H, which is called a seed profile, so that all topics relevant to q are contained in it and to obtain the preference values between q and all topics in H.
This procedure is performed in the following steps:
1. Find the topics in R that are relevant to q. Then develop an efficient method to compute the relevance of all topics in R with q. These values can be used to obtain a set of no overlapping relevant topics denoted by T(q), namely the relevant set. These topics are required to be no overlapping so that T(q), together with all their ancestor nodes in R, comprise a query-relevant tire denoted as R(q).Apparently, T(q) are the leaf nodes of R(q).Note that R(q) is usually a small part of R.
2. Overlap R (q) with H to obtain the seed profile G0, which is also a rooted sub tree of H. The seed profile's size is significantly reduced compared to the original profile.

5.4: Profile Generalization

This procedure generalizes the seed profile G0 in a cost-based iterative manner relying on the privacy and utility metrics. In addition, this procedure computes the discriminating power for online decision on whether personalization should be employed.

VI. EXPRIMENT RESULTS AND ANALYSIS

Search Quality refers to the relevant search results on issuing the query and the generalized profile as per user's interests constructed in their profile gives the comparison of the existing system of Greedy DP(Red color) and the proposed system of GreedyIL(Blue color) based on the Quality of search. The numbers of query sets (set of 20 queries Q1-Distinct Q2- Medium, Q3- Ambiguous Q4-Very ambiguous) are plotted in X-axis and the numbers of relevant URLs are plotted in Y-axis. When the query set varies, the number of relevant URLs will also vary based on the user's profile. The GreedyIL achieves 13% of improvement in search quality than the GreedyDP.
The response time comparison of the existing system of GreedyDP (Green color) and the proposed system of GreedyIL (Blue color) based on the response time taken by the query sets.

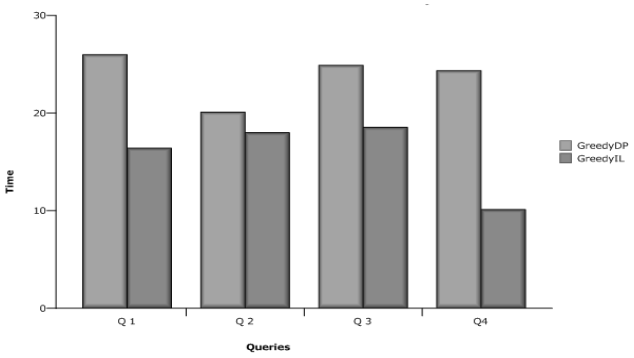


Fig. 6.1: Performance Comparison Based On Response Time

The numbers of query sets (set of 20 queries Q1-Ambiguous Q2-Medium Q3-Distinct Q4-Very ambiguous) are plotted in X-axis and the average time (in sec) are plotted in Y-axis. Based on the query set taken, the response time varies with respect to the user's profile.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The GreedyIL achieves 12% of improvement in response time than the GreedyDP. The precision is the quality of being accurate and also called positive predictive value. The comparison of the existing system of GreedyDP (Green color) and the proposed system of GreedyIL (Blue color) based on the effectiveness of personalization.

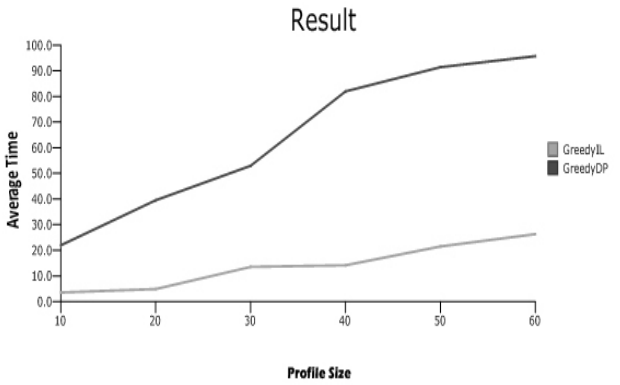


Fig. 6.2: Performance Comparison Based On Profile Size

Scalability refers the system's ability to handle the growing profile size in a capable manner or its ability to be enlarged to accommodate that growth. The comparison of the existing system of GreedyDP (Blue color) and the proposed system of GreedyIL (Green color) based on the scalability of varying profile size. The Profile Size (number of nodes) is plotted in X-axis and the average time (in sec) is plotted in Y-axis. Based on the number of nodes, the average precision varies with respect to generalization. The GreedyIL achieves 11% of improvement in scalability than the GreedyDP.

VII. CONCLUSION

A client-side privacy protection method called UPS for personalized web search is created. UPS could potentially be adopted by any PWS that captures user profiles in a hierarchical taxonomy. The method allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, UPS performed online generalization on user profiles to protect the personal privacy without compromising the search quality. Two greedy algorithms, namely Greedy DP and Greedy IL, are proposed for the online generalization. The UPS could achieve quality search results while preserving user's customized privacy requirements. The future work attempts to resist adversaries with broader background knowledge, such as richer relationship among topics (e.g., exclusiveness, sequentially, and so on), or capability to capture a series of queries from the victim. A more sophisticated method to build the user profile and metrics on privacy to predict the performance (especially the utility) of UPS can be developed.

REFERENCES

1. Baeza-Yates R and Ribeiro-Neto B (1999), Modern Information Retrieval Addison Wesley Longman.
2. Breese J.S, Heckerman D, and Kadie C.M (1998), "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52.
3. Chirita P.A, Nejdl W, Paiu R, and Kohlschutter C (2005), "Using ODP Metadata to Personalize Search," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR).
4. Dou. Z, Song R, and Wen J.-R (2007), "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590.
5. Gabrilovich E and Markovich S (2006), "Overcoming the Brittleness Bottleneck Using Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge," Proc. 21st Nat'l Conf. Artificial Intelligence (AAAI).
6. Hafner K (2006), Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times.
7. Ja¨rvelin K and Keka¨la¨inen J(2000), "IR Evaluation Methods forRetrieving Highly Relevant Documents," Proc. 23rd Ann. Int'lACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), pp. 41-48.
8. Krause A and Horvitz E (2010), "A Utility-Theoretic Approach to Privacy in Online Services," J. Artificial Intelligence Research, vol. 39, pp. 633-662.
9. Pitkow J, Schu H ¨tze, Cass T, Cooley R, Turnbull V, Edmonds, A, Adar E, and Breuel T(2002), "Personalized Search," Comm. ACM, vol. 45, no. 9, pp. 50-55.
10. Pretschner A and Gauch S (1999), "Ontology-Based Personalized Search and Browsing," Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99).
11. Ramanathan K, Giraudi J, and Gupta A (2008), "Creating Hierarchical User Profiles Using Wikipedia," HP Labs.
12. Shen X, Tan B, and Zhai C (2005), "Implicit User Modeling for Personalized Search," Proc. 14th ACM Int'l Conf. Information and Knowledge Management (CIKM).
13. Shen X,Tan B, and Zhai C(2005), "Context-Sensitive Information Retrieval Using Implicit Feedback," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR).
14. Shen X, Tan B, and Zhai C (2007), "Privacy Protection in Personalized Search," SIGIR Forum, vol. 41, no. 1, pp. 4-17.
15. Spertta M. and Gach S (2005), "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI).
16. Sugiyama K, Hatano K, and Yoshikawa M (2004), "Adaptive Web Search Based on User Profile Constructed without any Effort from Users," Proc. 13th Int'l Conf. World Wide Web (WWW).
17. Tan B, Shen X, and Zhai C (2006), "Mining Long-Term Search History to Improve Search Accuracy," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).
18. Teevan. J, Dumais S.T, and Horvitz E (2005), "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.
19. Qiu F and Cho J (2006), "Automatic Identification of User Interest for Personalized Search," Proc. 15th Int'l Conf. World Wide Web (WWW), pp. 727-736.
20. Xu Y, Wang K, Zhang B, and Chen Z (2007), "Privacy-Enhancing Personalized Web Search," Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600.

AUTHORS PROFILE

R. Ravikumar M.E.-Scholar, Department of CSE, Thiruvalluvar College of Engineering & Technology, Tamil Nadu, India.

U. Sundhar Asst.Professor, Department of CSE, Thiruvalluvar College of Engineering & Technology, TamilNadu, India.