VOL. 10, NO. 10, JUNE 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PROVIDING PRIVACY AND PERSONALIZATION IN SEARCH

T. Mercy Priya and R. M. Suresh
Department of Computer Science and Engineering, Sri Muthukumaran Institute of Technology, Chikkarayapuram, Chennai, Tamil Nadu, India
E-Mail: mercypriya.fm@gmail.com

ABSTRACT

The aim of this project is to construct a PWS framework called UPS that can generate a profile for the given user queries. The framework works in two phases, online phase and offline phase. The PWS is a search technique which provides a better search result which will satisfy the user needs. Here a personalized web search is provided without affecting their privacy. This work has supported to expose the user profile at run time with the given user query and to personalize it. For the PWS the semi structured data is indexed with RDF. Here we use two greedy algorithms, namely GreedyDP (Discriminating Power) and GreedyIL (Information Loss). These algorithms are used in the run time generalization.

Keyword: privacy, personalize, user profile, user query.

1. INTRODUCTION

It has become increasingly difficult for the users to find information on the internet which would satisfy the real intension of the user. The user might experience failure when the search engine returns the irrelevant answer for the user query. This occurs due to the ambiguity of the text and the enormous variety of user content. The PWS (Personalized Web Search) aiming at providing the better search result, which satisfy the individual user needs. The user information is collected and stored in the RDF (Resource Description Framework) manner. In general, each user has different information needs for their query. For example, for the query "java" [14], some user maybe interested in the document dealing with the programming language, "java", while other users may want documents related to "coffee" or a place located in Indonesia. Therefore, Web search result should adapt to user with different information needs.

In order to predict such information needs, the PWS is used here. The PWS is generally categorized into two types, namely click-log-based method and profile-based ones [1]. The click-log-based method imposes the clicked page in the user's query history. But this can work only on the repeated queries from the same user which act as a strong limitation. In contrast, Profile-based method improves with complicated user interest which can act effective iterative user for almost all sort of queries but unstable at some situations. The UPS provides a runtime profiling to optimize the personalized utility of the user's privacy requirement, allows the privacy needs to be customized and it doesn't require any user interaction [2]. Even though there are pros and cons in both the techniques, the profile-based one are found to be more effective. The information to the profile is gathered implicitly from query history [8] [9], browsing history [2], click-through data, book marks and so forth. The collected personal data can easily reveal the user's privacy lives which have been a major barrier for the wide proliferation of PWS services.

2. RELATED WORK

2.1 Personalized web sites

The personalized web site is constructed with the help of the link topology and the structure and contents of the web pages. The frameworks of these systems are "Link Personalization" [14] and "Content Personalization" [14].

2.1.1 Link personalization

This involves in selecting the links that are more relevant the user and changing the original navigation. The user who gives similar rating to the similar object is presumed to have similar preference and the site suggests those recommendations that are most popular among the class. At the E-commerce site for Amazon.com, this was taken by constructing a "New for you" home page and presenting it to each user with the new products that the user may be interested in it. This implicit recommendation is done by the purchase history.

2.1.2 Content personalization

The content personalization is done when pages present different information to different users. The approach in this application is that the users should be able to "construct" their own pages and even the layout may be customized. However, the users have to input their preference or the information based on the previous questionnaires.

2.2 Search history mining

In this information retrieval, after the query is submitted the system will return a set of documents with titles and summaries displayed. The user can then select to view the full texts of some results. Thus the search history generally includes three components: past queries, their search result and the information on which results were clicked/ viewed. The goal of the search history mining is to estimate the best history model. There are several challengers in this task: First, the past search contains different components (query, result and click through). Second, not all the past queries are equally important.

Third, the search history has hundreds or thousands of entities.

2.3 Semantic web service discovery

The user given content will pass through the refinement steps to extract the annotations (noun) that are used to discover the services. The refinement process model includes three steps: Removing noise from the query, Tokenizing the query and Filtering "stop" words. The user query is a sentence which may contains several special characters like comma (,), dot (.), plus (+) and etc. These characters are considered as noise and so they are removed from the content. Then the tokenizer will separate the each word as token. The result of this tokenizer will contain the "stop" words. The "stop" words are nothing but the verbs in natural languages, such as "want", "which", "is", "what", etc. Thus these "stop" words are removed from the tokenized content and thus a original content is obtained. Thus the original content obtained through the above steps in the refining process is widely used in many places to book tickets and a lot more.

2.4 Profile-based personalization

The Profile based PWS mainly focus on the search utility. The idea of this search results in creating a user profile that reveals an individual information goal. The previous solution of the PWS [4] is based on the two aspects, namely the profile representations and effective measure of the personalization. The recent work builds profiles in the hierarchical structure due to their strong descriptive ability, higher efficiency and better scalability.

3. PROPOSED METHOD

We proposed a Privacy-preserving personalized web search framework called UPS (User customized Privacy preserving Search). This UPS will generalize a profile for each and every user query according to the user preserved privacy requirement. The framework assumes that it does not provide any sensitive information and it also aims at protecting the privacy of the individual [5] user profile. The UPS [1] consists of many numbers of client and non trusty search engines. The client access the search services by them self at their own risk. The privacy of this search service is given by the online profiler who is implemented as a search proxy running on the client machine itself. The proxy maintains the user profile and the user-specified privacy requirements.

There are two types of phases in which the framework works [13] [15], namely the offline phase and the online phase. During the offline phase the user profile is constructed and it is customized with the user-specified privacy requirement. The online phase is performed by the following steps, at first as soon as the user issues a query the proxy generates a generalized profile that satisfy the user's privacy requirement. The generated generalized profile with the issued query sent together to the PWS [4] server for the personalized search. The search results are queried back to the proxy which will now re-rank it and send it to the client to obtain the in tented result.

This technique relies on the two conflict metrics, namely personalization utility and privacy risk. We develop two simple but effective generalized algorithms namely, GreedyDP [1] and GreedyIL which is used to the runtime profiling. At first it tries to maximize the DP and then try to minimize the IL. We also decide to provide a mechanism in an inexpensive way to whether personalize a query in UPS. The decision is made before each run time profiling to enhance the search engine and avoid unnecessary exposure of the profile.

4. TECHNIQUES IN UPS

The UPS [14] is distinguished from the conventional PWS by the following ways: It provides a runtime profiling, to optimize the personalized utility of the user's privacy requirement, allows the privacy needs to be customized and it does not require any iterative user interaction. These are the three methods which is not presented in the existing system. The architecture of the UPS is shown in Figure-1.

In this section, we explain the procedures carried out during the each phase of execution, namely online phase and offline phase. The offline phase constructs the original user profile and the topic sensitivity specified by the user to perform the privacy requirement in a customized manner. The online generalization procedure is guided by the global risk and utility metrics. The metric is computation relies on two intermediate data structure, namely cost layer and preference layer define on user profile.

The user has to undertake the procedure solutions as follows: Offline profile construction, offline privacy requirement customization and online generalization. The first step of the offline processing is to build the original user profile that reveals the user interest. In the offline privacy requirement, they first request the user to specify the sensitivity-node value. The sensitive-node value represents the data to be sensitive or non-sensitive by the obtained value. The sensitive-node value is taken here as cost. Now we can obtain the customized profile with its cost layer available.

5. ALGORITHM

We propose two Greedy algorithms namely, Greedy DP (Discriminating Power) and Greedy IL (Information Loss).

5.1 The Greedy DP algorithm

To reduce the complexity of our problem we introduce the operation called prune-leaf, r indicates the removal of a leaf topic *t* from a profile. The first greedy algorithm Greedy DP works in bottom-up matter. Starting from bottom node for every *i*th iteration, Greedy DP chooses a leaf topic for pruning and trying to maximize the utility of the output of the current iteration. During the iterative, we maintain the best profile so far that have the highest discriminating power while satisfying the risk constrains. The iterative process terminates when the profile is generalized to a root-topic. The main problem of

GreedyDP is that it requires recomputation of all candidate profiles generated from the attempts of prune leaf. This causes significant memory requirement and computational cost.

5.2 The Greedy IL algorithm

The Greedy IL algorithm improves the efficiency of the generalization using heuristics based on several findings. The finding motivates us to maintain a priority queue of candidate prune-leaf operations in descending order of the information loss caused by the operator. This lead to the following heuristic, which reduces the total computational cost significantly. The iterative process can terminate whenever risk is satisfied. Once a leaf topic is pruned, only the candidate operations pruning sibling topics need to be updated. In other word, we only need to recompute the IL values for operators attempting to prune sibling's topics.

6. RESULT AND ANALYSIS

Consistent with many previous works in personalized web services, each user profile in UPS [14] adopts a hierarchical structure. Here, our profile (as shown in Figure-2 and Figure-3) is constructed based on the availability of the public accessible taxonomy. The user profile is constructed based on the sample taxonomy repository.

The RDF is constructed for semantic data (as shown in Figure-4) on a relational database containing structured as well as unstructured data. The RDF is also generated by mining the text content uploaded by the users in blogs (as shown in Figure-5) and contents of the file are analyzed and the Meta contents are manipulated. The Meta contents are the key for search process so that the file can rendered on demand. The words are analyzed in WordNet API so that the related terms can be found for use in the Meta content in generation of RDF.

Similar data's are grouped together that relate to the same resource. The data level processes are subjected to structure level processing by indexing the semantic data elements. Multiple RDFs are grouped and structures together to form a master RDF data that holds all the semantic information of a server that support reasoning in any formats of query processing.

The framework works in two phases, namely the offline and online phase. During the offline phase, a hierarchical user profile is constructed and customized with user-specified privacy requirements. In the online phase, the user will send a query to the online profiler and that will create a generalized profile and send it with the query. The server will receive it and produce a raw data. The raw data will be obtained as it is or it will be re-ranked and provided to the user.

The created web page will appear as shown in Figure-6. As the sensitivity values explicitly indicates the user's privacy concerns, the most straight forward privacy preserving method is to remove sub trees rooted at all sensitive-nodes whose sensitivity value is greater than threshold. This method is referred to as forbidding. This is performed by two methods, Query-topic mapping and profile generation.

7. CONCLUSIONS

This paper present a client side privacy protection framework called UPS for personalized web search. UPS could potentially be adopted by any PWS that captures users to specify customer privacy requirement via the hierarchical profiles. In addition, the privacy is given without compromising the search quality. We propose two greedy algorithms, namely GreedyDP and GreedyIL. The RDF is indexed here to create a semi-structured data. Our experiment result also confirmed the effectiveness and efficiency of our solution.

For future work, we will try to resists adversaries with broader background knowledge, such as richer relationship among topics. We will also seek more sophisticated method to build the user profile.

ACKNOWLEDGMENT

I am not lucky! But still I am blessed. First of all I would like to thank the Almighty for his grace towards me. I would also like to thank my parents and my brothers for whom I am right now. I am Thankful towards my guide for his guidelines towards my project work. At last, I would like to thank my friends and my family.

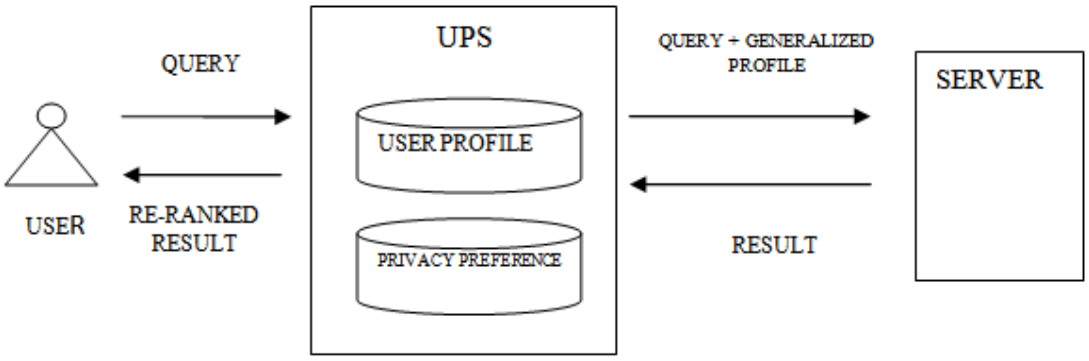

Figure-1. System architecture of UPS.

Figure-2. Register the user profile.

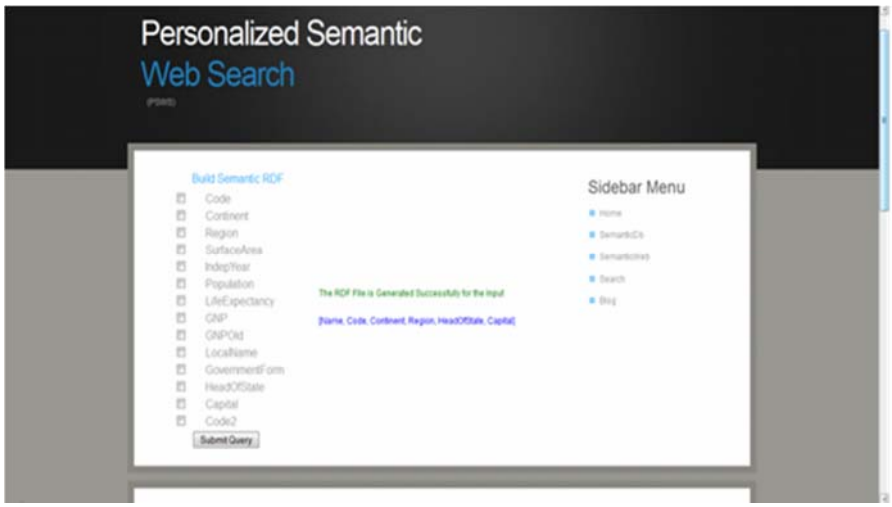

Figure-3. User Profile constructed with the available public taxonomy.

Figure-4. Created RDF for the semantic user profile.

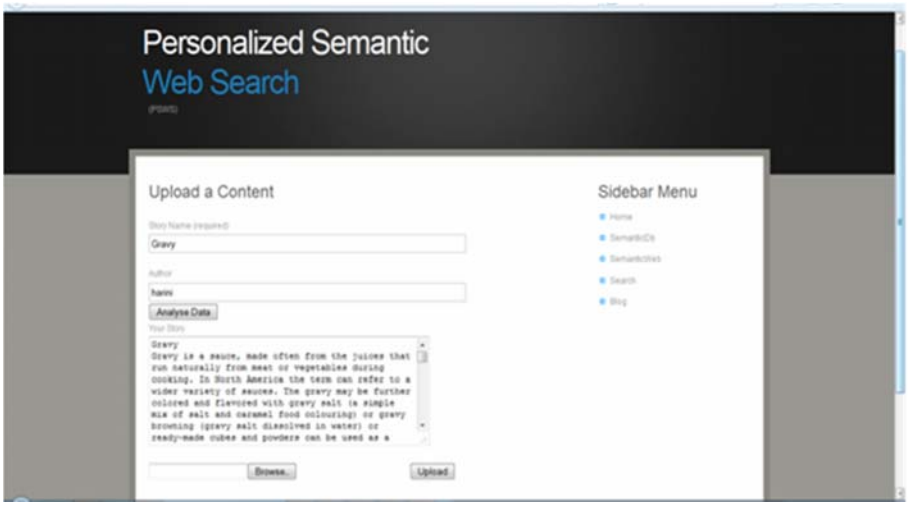

Figure-5. User upload their text content.

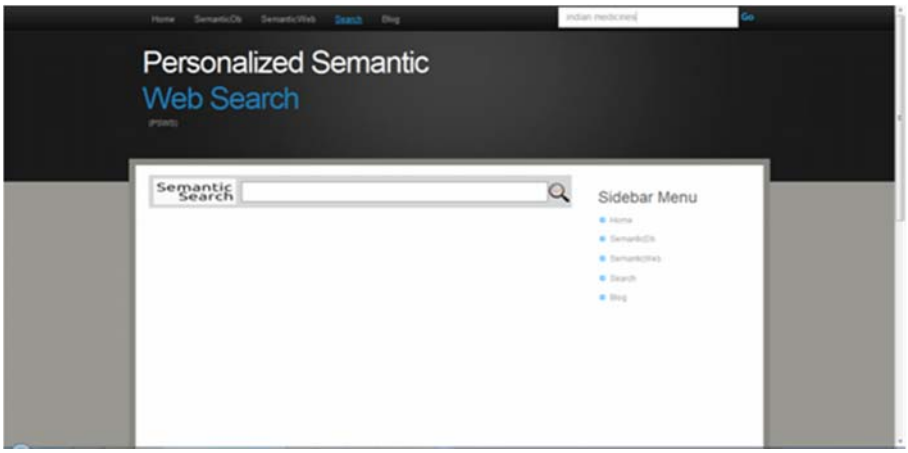

Figure-6. Created PWS page.

REFERENCES

[1] Chen. G, Bai. H, Shou. L, Chen. K and Gao. Y. 2011. Ups: Efficient Privacy Protection in Personalized Web Search," Proc. 34th Int'l ACM SIGIR Conf. Research and Development in Information, pp. 615- 624.

[2] 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.

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

[4] Lidan shou, He Bai, Ke Chen, and Gang Chen. 2014. "Supporting Privacy Protection in Personalized Web Search", Issue No.02 – February 2014, vol.26, pp. 453-467

[5] Ramanathan. K, Giraudi. J and Gupta. A. 2008. Creating Hierarchical User Profiles Using Wikipedia," HP Labs.

[6] 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.

[7] 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).

[8] 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).

[9] Spertta.M and Gach.S. 2005. "Personalizing Search Based on User Search Histories," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI).

[10] 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).

[11] Teevan.J, Dumais.S.T and Liebling. D.J. 2008. "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.

[12] Viejo. A and Castell_a Roca.J. 2010. "Using Social Networks to Distort Users' Profiles Generated by Web Search Engines," Computer Networks, vol. 54, no. 9, pp. 1343-1357. [12]

[13] Xu. Y, Wang. K, Yang. G, and Fu. A.W.C. 2009. "Online Anonymity for Personalized Web Services," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM), pp. 1497-1500.

[14] Xing.D, Xue.G.R, Yang.Q and Yu.Y. 2008. "Deep Classifier: Automatically Categorizing Search Results into Large-Scale Hierarchies," Proc. Int'l Conf. Web Search and Data Mining (WSDM), pp. 139-148.

[15] Zhu.Y, Xiong.L and Verdery.C. 2010. "Anonymizing User Profiles for Personalized Web Search," Proc. 19th Int'l Conf. World Wide Web (WWW), pp. 1225-1226.