International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume2, Issue 8 (August '2016)

SUSTAINING PRIVACY PROTECTION IN PERSONALIZED WEB SEARCH OVER RE-RANKING WITH CACHE SERVICES MODEL IN WEB DATABASE

P.Vinothini,
M.Phil Research Scholar,
Department of Computer Science,
Vivekanandha College of Arts and Sciences for Women
(Autonomous), Tiruchengode.

M.Jothilakshmi,
Assistant Professor,
Department of Computer Science,
Vivekanandha College of Arts and Sciences for Women
(Autonomous), Tiruchengode.

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. This paper analysis a privacy protection in PWS applications that model user preferences as hierarchical user profiles. This paper proposes a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user-specified privacy requirements. The proposed 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 paper presents a greedy algorithm, namely Enhanced GreedyIL, for runtime generalization. It also provides an online prediction mechanism for deciding whether personalizing a query is beneficial.

Keywords: Personalized web search, profile, Catch base Search Tree, Fusion Re-ranking Algorithm, Security privacy.

I. INTRODUCTION

Data mining is about finding new information in a lot of data. Data mining, the extraction of hidden predictive information from large databases, it is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Security and privacy are not very new concepts in data mining, but there is too much that can be done in this area with data mining. Analysis of social networks and group dynamics from electronic communication give a thorough analysis of impact of social networks and group dynamics. Specifying the need to understand cognitive networks, it also models knowledge network using the Enron E-mail corpus. Recording of electronic communication like email logs, and web logs have captured human process. Analysis of this can present an opportunity to understand sociological and psychological process. Privacy preserving data analysis on graph and social networks provides various types of privacy breach and present an analysis using k-candidate anonymity, K-degree anonymity and k-neighborhood anonymity.

Previous works on profile-based PWS mainly focus on improving the search utility. The basic idea of these works is to tailor the search results by referring to, often implicitly, a user profile that reveals an individual information goal. In the remainder of this section, we review the previous solutions to PWS on two aspects, namely the representation of profiles, and the measure of the effectiveness of personalization. Many profile representations are available in the literature to facilitate different personalization strategies. 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. The hierarchical profile automatically via term-frequency analysis on the user data and proposed UPS framework, do not focus on the implementation of the user profiles.

Actually, our framework can potentially adopt any hierarchical representation based on taxonomy of knowledge. As for the performance measures of PWS in the literature, Normalized Discounted Cumulative Gain (nDCG) 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), Rank Scoring, and Average Rank .

II. RELATED WORK

Ruihua Song [1] describe this problem and provides some preliminary conclusions. They 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 MSN query logs. By analyzing the results, they reveal that personalized search has significant improvement over common web search on some queries but it has little effect on other queries (e.g., queries with small click entropy). It even harms search accuracy under some situations. Furthermore, they show that straight- forward click-based personalization strategies perform consistently and considerably well, while profile-based ones are unstable in their experiments. They also reveal that both long- term and short-term contexts are very important in improving search performance for profile-based personalized search strategies.

Xuehua Shen et al [2] describes the 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, they 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 click through data of past searches that are related to the current query

Kenji Hatano et al [3] Web search engines help users find useful information on the World Wide Web WWW). However, when the same query is submitted by different users, typical search engines return the same result regardless of who submitted the query. Generally, each user has different information needs for his/her query. Therefore, the search results should be adapted to users with different information needs. In this paper, authors first propose several approaches to adapting search results according to each user's need for relevant information with-out any user effort, and then verify the effectiveness of their proposed approaches. Experimental results show that search systems that adapt to each user's preferences can be achieved by constructing user profiles based on modified collaborative filtering with detailed analysis of user's browsing history in one day.

Bin Tan and ChengXiang Zhai [4] describe a major limitation of most existing retrieval models and systems is that the retrieval decision is made based solely on the query and document collection; information about the actual user and search context is largely ignored. In this paper, authors study how to exploit implicit feedback information, including previous queries and clickthrough information, to improve retrieval accuracy in an interactive information retrieval setting. They propose several context- sensitive retrieval algorithms based on statistical language models to combine the preceding queries and clicked document summaries with the current query for better ranking of documents. They use the TREC AP data to create a test collection with search context information, and quantitatively evaluate their models using this test set. Experiment results show that using implicit feedback, especially the clicked document summaries, can improve retrieval performance substantially.

III. SYSTEM METHODOLOGY

A. EXISTING SYSTEM METHOLOGY

The existing system works in two phases, namely the offline and online phase, for each user. During the offline phase, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The online phase handles queries as follows:

- When a user issues a query qi 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 Gi satisfying the privacy requirements. The generalization process is guided by considering two conflicting metrics, namely the personalization utility and the privacy risk, both defined for user profiles.
- Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.
- The search results are personalized with the profile and delivered back to the query proxy.
- Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile.
    - No capability to capture a series of queries.
    - User profile is categorized into single node in the tree structure only.
    - Past query based suggestion is not given to user.

B. PROPOSED METHODOLOGY

In addition with all the existing system activities, the proposed system also includes the new following processes. Users may be grouped in more than one profile so that the result includes both types of search outputs. In addition, previous query based suggestions are also provided.

- Capability to capture a series of queries.
- User profile is categorized into multiple nodes in the tree structure.
- Past query based suggestion is given to user.

Specific Objective

1. The average numbers of the relevant shots at different depths of the result list are used as the evaluation criteria. It is not hard to reach a conclusion that relevant shots are scarce in the top ranked results. However in real world application scenarios users merely restrict their attention to the top rank shots. In the proposed system this problem

is eliminated by fusing the clustering produced during the search results so that the results are more effective and relevant to end user's requirement.
2. The new search mechanism satisfies the user's information needs.
3. The fusion is carried out to the intermediate clusters only and so the calculation time and overhead is reduced much.
4. Unwanted Web page links are eliminated to better extent.
5. Time Reduction in searching and result Efficiency is improved
6. Accurate Re-ranking of result and Multiple feature space based search is possible

IV. PRIVACY POLICY CONSTRUCTION

A. PROFILE CONSTRUCTION

In this section, a hierarchical user profile is constructed and customized with the user-specified privacy requirements. The first step of the offline processing is to build the original user profile in a topic hierarchy H that reveals user interests. Assume that the user's preferences are represented in a set of plain text documents, denoted by D. The profile construction in which the new node will be additionally added to the in this module in the user menu that contains addition of the user and the View to the previous existing user.

The next one is the category in these categories that are added already will be shown that list will be gathered in this view category. The new category can also be added in this add category and each node that is created already and newly will be shown in this view. The profile construction in which the different set of the user will be created that is each set of the user will have the different set of the restrictions for the created user that kind of settings is called profile settings in the creation of profile.

B. SEARCH RESULTS

In this module, the search results are personalized with the profile and delivered back to the query proxy that is all the user will not have the permission to work in that specific user login the signed user can work in that login that is the user already have the permission to work in that will be able to work in that login options. The signed user can view the results that they got as the output. The search result will be the user query search the output gathered will be in the form of the user requirements it may be the webpage link or the reference link that is been provided by the web server to the searched query. The encryption category that is from which part of the user the encryption is been performed that kind of the data is been retrieved form the data base that is given according to the user level encryption. The encryption is based on the user group according to that the data will be encrypted and added into the database.

C. RERANKING

Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile. The displayed results will be reranked in this module the particular result will be displayed exactly ranked in this module the output will be based on the ranked output. The query that is given from the user side that will be for the searching of data the searched data is found when the user attains the query word. The lot of results will be displayed for the single query word according to the result displayed that result will not be in the form that user view that is the user searched page will be in last page of the displayed result are in the middle page so that will be sorted and the related data will be shown in the first page of the results that type of the sorting is called as the re ranking. The search result will be gathered form the server and then according to the reranked output will be given to the user. The reranking will be carried out using the cross reranking algorithm.

D. PREVIOUS QUERIES

In this module, if the query contains the phrase and the user is already registered (not guest), then the previous queries are also suggested during the query input, so that the search result is related with previous searches. The previous query will not be updated in the existing system that results in taking more time for the search for reducing the search time by establishing the previous query to store and it can taken as the query when the search is performed. The previous query that is given by the user to search for the particular data the particular data search will be performed with that query that query is called as the previous query it may be the traveled path by the user regarding the particular search that search will be in the form of web address or any other reference link of the previous query this may be of any other type. The user searched with the different query that is stored in the data base will be retrieved whenever the user need to work on it that is the previous query.

Proposed system conceptually simple technique protects user privacy to a certain degree, but at the cost of the semantic loss incurred by suppressing tags. Other approaches based on data perturbation include the submission of false tags. For example, a user wishing to tag the webpage www.mentalhelp.net with "depression" could use the tag "sports" instead, to conceal their interest for this resource. In doing so, the user distorts their actual profile, although at the expense of a far greater impact on semantic functionality than suppression does resources are assigned tags that do not describe, in principle, the actual content of such resources.

E. FUSION-RERANKING ALGORITHM

In this thesis, they introduce a re-ranking method, called Fusion-Re ranking, which combines query modal features in the manner of cross reference. The fundamental idea of Fusion-Re ranking lies in the fact that the semantic understanding of query content from different modalities can reach an agreement. Actually, this idea is derived from the multi-view learning strategy, a semi supervised method in machine learning. Multi view learning first partitions available

attributes into disjointed subsets (or views), and then cooperatively uses the information from various views to learn the target model. Its theoretical foundation depends on the assumption that different views are compatible and uncorrelated. In this context, the assumption means that various modalities should be comparable in effectiveness and independent of each other. Multi view strategy has been successfully applied to various research fields, such as concept detection. However, this strategy, here, is utilized for inferring the most relevant shots in the initial search results, which is different from its original role. CR-Reranking method contains three main stages: clustering the initial search results separately in diverse feature spaces, ranking the clusters by their relevance to the query, and hierarchically fusing all the ranked clusters using a cross-reference strategy.

Input: file and Query Label
Output: RERanking Encrypted files
Step 1: Select file form local Database
Step 2: Select Detector base file in to database.
Step 3: Search file
Step 4: The user implicitly relates the retrieved (down-loaded) file
Step 5: Assuming category level searching order transitions in the order of the keywords the aim of the proposed approach is to quantify logical connections between keywords.
Step 6: The new probability (based on M + m key-words) is calculated by the recurrent formula

```
/* Fusion Re Ranking Algorithm */
Input: Query Q
Output: Cross Ranking Results CR
Q ← Null
for i=1;<=1000;i++
Q ← i
Search (Qi, Qi2....Qn)
Query files ← (Qi++ ∩ Qn)
CR ← Q
Return CR
```

This procedure constructs a category level searching order where each keyword corresponds to a state. Each time a keyword appears in a query, its state counter is advanced; if another keyword follows in the same query, their interstate link counter is also advanced

Step 7: Compare directly the probability vectors πi and πj calculated in the previous step for two file, one faces the zero-frequency problem. By itself, the fact that a user puts certain keywords together in a query implicitly renders the keywords relative to each other regardless of the file that may or may not be picked by this user.
Step 8: Optimization step. The AMC will be used to cluster the keyword space and define explicit relevance links between the keywords by means of this clustering.

Step 9: All the clusters are spitted into three sub groups as High, Medium and Low for each cluster. For example, Cluster A is spitted in A_{high}, A_{medium} and A_{Low} and Cluster B is spitted in B_{high}, B_{medium} and $B_{Low.}$ Then the sub groups are intersected like

1. Unwanted Web page links are eliminated to better extent.
2. Time Reduction in searching and result Efficiency is improved
3. Accurate Re-ranking of result and Multiple feature space based search is possible.

$$p_i (K1, K2) = M\ p_i (K1, K2) + m / M + m$$

V. PERFORMANCES ANALYSIS

The table 5.1 describes an experimental result for greedy Information loss algorithm and Cross References algorithm. The table contains total number of profile, number of profile created and search count details for greedy information loss (GL) algorithm and number of profile created and search count details for Fusion Re-Ranking (FU-RR) algorithm. The comparison between existing and proposed algorithm is reduces the noisy profile created by search given login process.

S.NO	NUMBER OF PROFILE QUERY [Count]	GL ALGORITHM [Count]	FURR ALGORITHM [Count]
1	100	56	43
2	200	143	112
3	300	227	203
4	400	345	325
5	500	413	397
6	600	505	492

Table 6.1 Comparison between GLL and FURR Algorithm-Performances Analysis

The Figure 5.1 describes an experimental result for greedy Information loss algorithm and Cross References algorithm. The figure contains total number of profile, number of profile created and search count details for greedy information loss (GL) algorithm and number of profile created and search count details for Fusion Re-Ranking (FU-RR)algorithm. The comparison between existing and proposed algorithm is reduces the noisy profile created by search given login process.

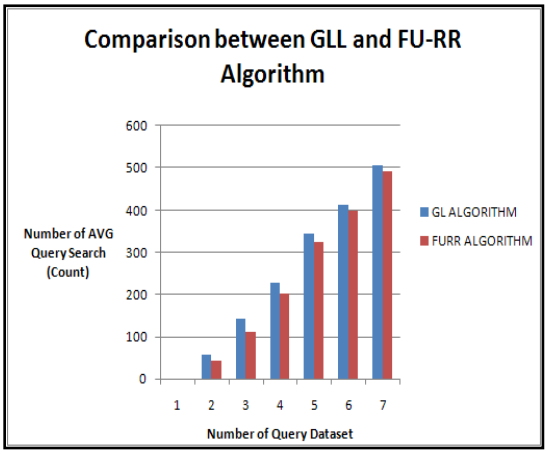

Figure 5.1 Comparison between GLL and FU-RR Algorithm

VI.CONCLUSION

The proposed system is 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. It proposed a greedy algorithm, namely GreedyIL, for the online generalization. The 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. The main benefits are capability to capture a series of queries, User profile is categorized into multiple nodes in the tree structure and past query based suggestion is given to user.

VII.FUTURE ENHANCEMENT

At present, the project presented a client-side privacy protection framework called UPS for personalized web search. For future work, the thesis will try 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 (relaxing the second constraint of the adversary) from the victim. It will also seek more sophisticated method to build the user profile, and better metrics to predict the performance (especially the utility) of UPS.

VIII.REFERENCE

[1]. Z. Dou, R. Song, and J.-R. Wen, "A Large-Scale Evaluation and Analysis of Personalized Search Strategies," Proc. Int'l Conf. World Wide Web (WWW), pp. 581-590, 2007.

[2]. R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. Information Systems, 10(2):115–141, 1992.

[3]. X. Shen, B. Tan, and C. Zhai. Implicit user modeling for personalized search. In Proceedings of CIKM '05, pages 824–831, 2005.

[4]. J. Teevan, S. T. Dumais, and E. Horvitz. Beyond the commons: Investigating the value of personalizing web search. In Proceedings of PIA '05, 2005.

[5]. X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In Proceedings of SIGIR '05, pages 43–50, 2005.

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

[7]. OnStat. Most people use 2 word phrases in search engines according to onestat.com, February 2004. http://www.onstat.com.

[8]. Seruku. Seruku toolbar, 2004. http://www.seruku.com/index.html.

[9]. Mike Giles. Furl, 2003. http://www.furl.net.

[10]. 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.