ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

Preserving User's Profile Protection for Personalized Web Search

Nivi A.N [1], Vanitha S [2], Saranya K.R [3], Sivaranjani B [4], Kavitha S [5]

PG Scholars, Dept. of CSE, Dr.N.G.P. Institute of Technology, Coimbatore, India[1,3,4,5]

Assistant Professor, Dept. of CSE, Dr.N.G.P. Institute of Technology, Coimbatore, India[2]

ABSTRACT: Internet of things is in its peak in today's world. The web search is the widely performing task in the internet world. When the query is searched over web should provide the relevant information to the users. The irrelevant results may annoy the users and hence the efficiency of the query search should be improved. To improve the search, personalized web search framework is proposed to retrieve the data based on user's interest. When the information is retrieved based on user's interest, user's profile will be publicly available. In this paper, the User's profile is also protected to handle privacy threats using generalization technique with Greedy Discriminating Power and Greedy Info Loss algorithm. The efficiency of search is improved by transferring the query with user's profile to the web server to retrieve the results. If the results are not satisfied to the user then the re-ranking technique is proposed to retrieve the most relevant search results.

KEYWORDS: personalised web search, user's profile protection, re-ranking, generalization, search utility

I. INTRODUCTION

Web search is mostly performed by all the people to retrieve the useful information from the search engine. The search may return irrelevant results that are not satisfied to the users. This irrelevance is due to variation in user's perspective and also due to text ambiguity. This issue is handled with personalized web search framework [1] using profile based method. In profile based method [2], the user's profile is created based on user's search interest via queries [3],[4],[5] and browsing histories [6],[7]. It improves the quality of data retrieval and user's satisfaction along with re-ranking technique [8]. The created user profile should be protected to overcome the privacy threats using generalization algorithms which also help to improve the data utility.

A. *Security Threats:*

There are privacy threats which lead to the solution with generalization algorithms the major threat is identity disclosure of an individual as explained in [9]. The identification of individual is done with the help of sensitive values from user's profile while transferring the search query and user's profile to the web server. This problem of pseudo identity, group identity, no identity and no personal information are solved in [10], [11], [12]. The personalized privacy protection is given with generalization algorithm which is introduced in Privacy Preserving Data Publishing (PPDP) [13]. This paper further consists of Related Work in section II, Personalized web search framework in section III, Experimental Settings in section IV and Conclusion and Future Work in section V.

II. RELATED WORK

Personalized web search is mainly focused on better data utility and privacy. We proposed user's profile protection with this framework and efficiency of search results is increased with re-ranking technique [14], [15], [16]. The most of the search engines use re-ranking technique for image retrieval. The user's profile generation of this framework is in hierarchical structure with weighted graph such as ODP [2], [17], [18]. The performance of the personalized web search is estimated with metrics like average precision [19], [20], rank scoring [21] and average rank [4], [22]. Identity disclosure is described in [9]. In this paper we have analysed the threat with privacy preserving data publishing [13] algorithms.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

III. PERSONALIZED WEB SEARCH FRAMEWORK

The personalized web search is a framework where the user profile is protected during the search. The PWS is used to search the most relevant information from the web server using both the query and the user's profile. This is a profile based search in which user's profile is created with the cookies, browsing and query histories. These created profiles should be protected with the generalization techniques because it may be available on public repository and used by others as background knowledge.

A. *User's Profile Protection:*

The user's profile is generated with the details of user's browsing history, cookies and also based on the query. The profile can be generated in two phases, online and offline phases and it adopts a hierarchical structure. Consider the following figure 1 which is a general taxonomy of search from which the user profile is generated represented in figure 2 with the sensitivity of topic.

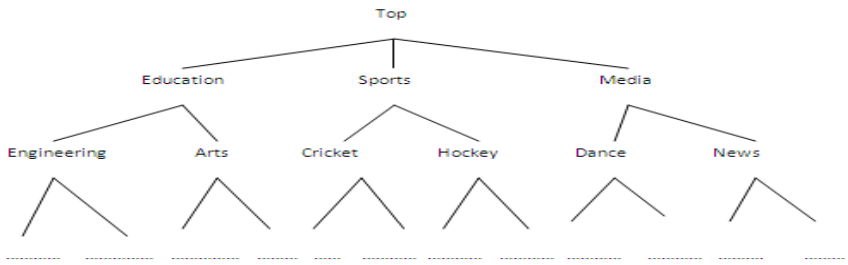

Fig 1. Taxonomy Repository

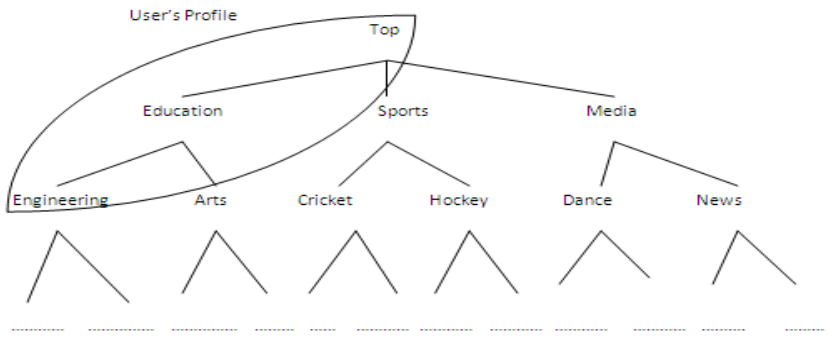

Figure 2: User's Profile creation from taxonomy

1) *Offline Phase*

The original user profile and customized privacy are constructed in the offline mode.

Profile Construction

The user's profile is constructed from the hierarchy H of the data based on user interests. It is a subset (S) of H where S∈H. to detect the respective topic in S for every document D, a naïve method is used to calculate the frequency of words (w) in topic (t) for a Naïve Bayesian Classifier (nb) using eq 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

$$nb(d, t) = \sum_{w \in t} N_{d,w} \ln \frac{N_{t,w} + \in}{\sum_{t' \in S} N_{t',w} + \in'}$$

eq. (1)

Customized Privacy Construction

There are sensitive values that are specified by users for their search to retrieve the relevant values. In this phase the user has to specify sensitive values for every topic. The cost (c) for each value should be computed for each topic as follows

$$c(t) = \sum_{t' \in c(t,H)} c(t') * p(t' \mid t)$$

eq. (2)

Once the customized profile is created with the cost then the online phase is executed

2) Online Phase

In this phase, query mapping and generalization of profile is done.

Query Mapping

For every query q, the query mapping is done with 2 steps. First, the subset S from the hierarchy H is identified relevant to q. Second, the related values between q and H are obtained. For example, consider the table 1 for the values of Computer Science.

Table.1 Query Mapping

Topics	Relevant set
Top/Education/Engineering/Computer Science	35
Top/Education/Arts/Computer Science	23

Generalization of profile

The profile generalization is implemented to preserve the privacy of users profile during the search. It is done based on the metrics of utility and privacy with generalization algorithms.

B. *Generalization:*

The critical metrics for generalization and two generalization algorithms called Greedy Discriminating Power and Greedy Info Loss are proposed in this section.

1) Metric for utility:

The quality of search is predicted with the utility metric because the quality of search depends on the user's search in the personalized web search engine. We estimate the Discriminating Power of a query for a profile instead of utility prediction when the specific topics are observed more or the topics are similar or a few topics are concentrated more. The utility metric is estimated with information content which identifies the specific topic using eq.3 and profile granularity that calculates the probability distribution using eq.4.

$$IC(t) = \log^{-1} p(t)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

eq. (3)

$$PG(q, profile) = \sum_{t \in T_{profile}(q)} p(t \mid q, profile) IC(t) - H(t \mid q, profile)$$

eq. (4)

2) Metric of Privacy:

The privacy metric is to estimate the privacy risk. The privacy risk is evaluated based on the sensitive value of a profile that can be exposed from the root node of hierarchy H. From the profile, the unnormalized risk is calculated using eq.5. The normalized risk from total sensitivity of H is calculated using eq.6

$$ur(t, profile) = \begin{cases} c(t) & \text{if t is leaf node} \\ \sum_{t' \in C(t, profile)} ur(t', profile) & \text{else} \end{cases}$$

eq. (5)

$$nr(q, profile) = \frac{ur(root, profile)}{\sum_{s \in SV} Sensitive(s)}$$

eq. (6)

Where nr is always in the interval [0,1].

3) Greedy Discriminating Power Algorithm:

The Greedy Discriminating Power Algorithm is used for generalizing the data. It is a greedy algorithm which generalizes the data in bottom up manner in the hierarchical structure of the user's profile. It starts from the leaf node and the pruning of the profile's data is done with repeated iteration for better data utility and privacy of the user's profile. For all user profiles it requires re-computation for pruning from leaf node to root node. The disadvantage of this algorithm is it requires more memory space and computations.

4) Greedy Info Loss Algorithm:

The Greedy Info Loss Algorithm is used for more efficient generalization which uses heuristics for pruning. Here the priority of user's profile is maintained in the queue for pruning till the end of all iterations for generalizing the user's profile. It also identifies the specific topic of the user's profile search from the pruned leaf. The computation is simplified than Greedy Discriminating Power algorithm.

C. *Re-ranking Technique:*

Re-ranking is an efficient method used to retrieve the results from the search engines. It is mainly used in image retrieval where a query yields a set of images as result and if the result is irrelevant further search takes image from already retrieved results as a query for refined search and more relevant search results. The comparison of input query image with others in the database is done with the semantic features of the image that avoids irrelevant results to be retrieved.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

IV. EXPERIMENTAL SETTINGS

In this section, the experimental setup and experimental results are explained in detail.

A. *Experimental Setup:*

An online shopping website is created for personalized web search using Java and HieldSQL with JDBC connectivity is shown in fig 4. When the registered user login and search for the desired shopping his/her user profile is generated using their search query and previous searching histories. Once the user profile is created, it is generalized using GreedyDP and Greedy IL algorithms. The generalized user profile along with the query is send to the web server. The obtained results can be ranked again if it is not satisfied to the user using re-ranking technique. The ranked query will be send to the sever and the most relevant results are retrieved without losing the user's privacy.

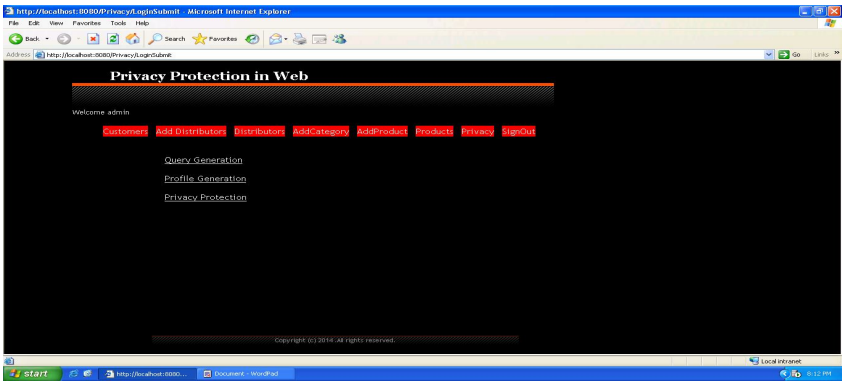

Fig. 3 personalized web search GUI

B. *Experimental Results:*

From this experiment we observed that Greedy Info Loss outperforms Greedy Discriminating Power algorithm and hence we used Greedy Info Loss Algorithm for user's profile protection is shown in fig. 4.

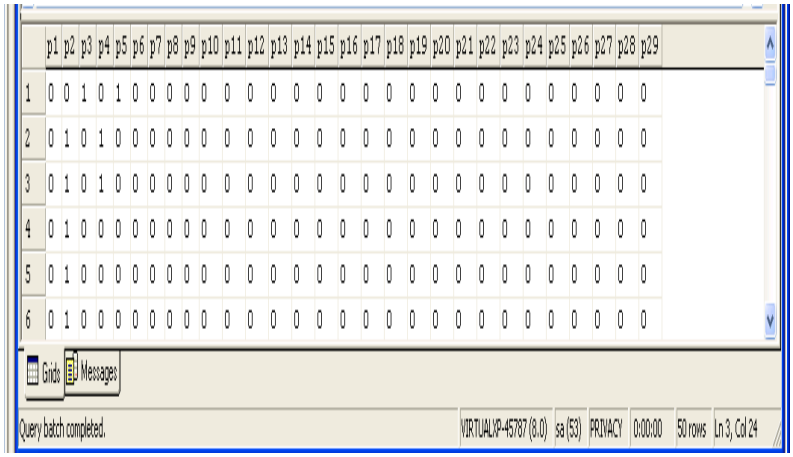

Fig. 4 User's Profile Protection

The efficient search result is obtained with the re-ranking technique is shown in fig. 5.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

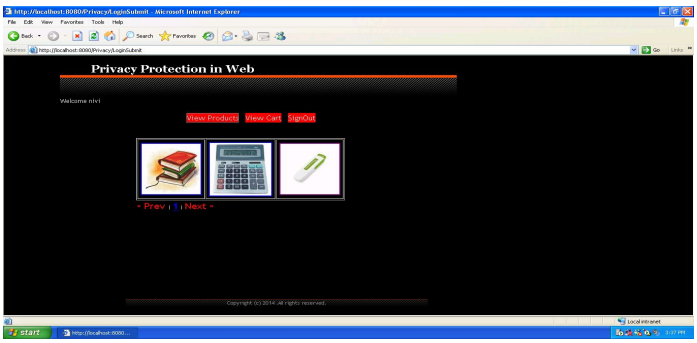

Fig. 5 Search Results

V. CONCLUSION AND FUTURE WORK

The web search is a major task in the internet. The search history and queries of the user are saved by the search engines. The saved data can be accessed by the users as a profile for analysis or to provide other relevant data for users. The browsing histories and search queries of a user creates a user's profile and it should be protected to avoid privacy threats. From our experiment, it is observed that Greedy Info Loss algorithm works better for generalization of user's profile that protects the identity disclosure of a user. The data retrieval is made efficient with re-ranking technique which retrieves more relevant results for the user's query.

In future, the other privacy threats can be handled with efficient algorithms and the scalability of data can be analyzed with personalized web search.

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, February 2014.
2. 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.
3. J. Teevan, S.T. Dumais, and E. Horvitz, '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, 2005.
4. M. Spertta and S. Gach, 'Personalizing Search Based on User Search Histories', Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI), 2005.
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. K. Sugiyama, K. Hatano, and M. Yoshikawa, 'Adaptive Web Search Based on User Profile Constructed without any Effort from Users', Proc. 13th Int'l Conf. World Wide Web (WWW), 2004.
7. X. Shen, B. Tan, and C. Zhai, 'Context-Sensitive Information Retrieval Using Implicit Feedback', Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.
8. Xiaogang Wang, Ke Liu, Xiaoou Tang, 'Web Images Re-Ranking Using Query- Specific Semantic Signatures,' IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume:36 , Issue: 4) April 2014, DOI:10.1109/TPAMI.201.
9. X. Shen, B. Tan, and C. Zhai, 'Privacy Protection in Personalized Search', SIGIR Forum, vol. 41, no. 1, pp. 4-17, 2007.
10. K. Hafner, Researchers Yearn to Use AOL Logs, but They Hesitate, New York Times, Aug. 2006.
11. 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.
12. 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.
13. X. Xiao and Y. Tao, 'Personalized Privacy Preservation', Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2006.
14. J. Cui, F. Wen, and X. Tang, 'Real time google and live image search re-ranking', In Proc. ACM Multimedia, 2008.
15. J. Cui, F. Wen, and X. Tang, 'Intentsearch: Interactive on-line image search re-ranking', In Proc. ACM Multimedia. ACM, 2008.
16. B. Luo, X. Wang, and X. Tang, 'A world wide web based image search engine using text and image content features', In Proceedings of the SPIE Electronic Imaging, 2003.
17. P.A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschu¨ tter, 'Using ODP Metadata to Personalize Search', Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development Information Retrieval (SIGIR), 2005.
18. Pretschner and S. Gauch, 'Ontology-Based Personalized Search and Browsing', Proc. IEEE 11th Int'l Conf. Tools with Artificial Intelligence (ICTAI '99), 1999.
19. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley Longman, 1999.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 4, April 2015

20. Y. Xu, K. Wang, B. Zhang, and Z. Chen, 'Privacy-Enhancing Personalized Web Search', Proc. 16th Int'l Conf. World Wide Web (WWW), pp. 591-600, 2007.
21. J.S. Breese, D. Heckerman, and C.M. Kadie, 'Empirical Analysis of Predictive Algorithms for Collaborative Filtering', Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), pp. 43-52, 1998.
22. F. Qiu and J. Cho, 'Automatic Identification of User Interest for Personalized Search', Proc. 15th Int'l Conf. World Wide Web (WWW), pp. 727-736, 2006.