Volume 7, Issue 5, May 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Content Extractor for Privacy Search

Sonali K. Shelke*
Department Computer Science & Engineering, Deogiri Institute of Engineering & Management Studies, Aurangabad, Maharashtra, India

DOI: 10.23956/ijarcsse/V7I5/0133

Abstract— The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown interesting patterns such as groups of data records, unusual records dependencies. Personalized search refers to search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided. Personalized web search (PWS) used for improving the quality of various search services on the Internet. Users might (force) experience failure when search engines return irrelevant (unrelated) results that do not meet (convene) the real intentions. In this paper PWS framework called Web Content Extractor for Privacy Protection (WCEP) that can adaptively generalize profiles by queries while respecting user specified privacy requirements. This work includes two greedy algorithms, namely GreedyDP(Greedy discriminating power) and GreedyIL (Greedy information loss), for runtime generalization.

Keywords— Privacy protection, personalized web search, WECP, GreedyIL, GrreedyDP, profile,ODP

I. INTRODUCTION

New personalization technologies are becoming increasingly widespread, raising a multitude of privacy challenges. Three trends in personalization require special attention with regard to privacy: social-based personalization, behavioral profiling, and the mobile Web. The Web had become more social, a place where people use their real identities and communicate with their family, friends,

Finally, the Web had become mobile, frequently accessed through smart phones, providing new information and possibilities that can be used for personalization. Personalization has the potential to amplify and complicate the Internet's inherent privacy risks and concerns. For example, personalized content in a social network system can reveal potentially embarrassing information directly to friends, family, and colleagues. Personalizing content according to the physical location of the user can reveal the location to unauthorized third-party entities. Examples of these types of personalization are readily apparent at many web services operating today in which users are facing a complicated privacy landscape.

A. Limitation

A major limitation of most existing information 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 completely ignored [1].

B. The main objective of WCEP implementation:

- To design an efficient application model secure user profile.
- To overcome the drawbacks faced Greedy search methods that are used for personalization techniques.

II. LITERATURE SURVEY

Personalized search gains popularity as there is the demand for more relevant information. Research has indicated low success rates among major search engines in providing relevant results; in 52% of 20,000 queries, searchers did not find any relevant results within the documents that Google returned. Personalized search can improve search quality significantly and there are mainly two ways to achieve this goal.

Long-term search history contains rich information about a user's search preferences, which can be used as search context to improve retrieval performance[7]. M. Spertta and S. Gach, User profiles, descriptions of user interests, can be used by search engines to provide PWS(Personalized web search) results. Many approaches to creating user profiles collect user information using proxy servers[11] (to capture browser history of a personal computer).

This paper presents a scalable way for users to automatically build user profiles and rich query log based on search. These profiles arrange a user's interests into a hierarchical organization according to specific interests.

A. Existing system

In our Existing System, 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 query.

Disadvantages of Existing System

- The existing profile-based PWS do not support runtime profiling.
- The existing methods do not take into account the customization of privacy requirements.
  Many personalization techniques require iterative user interactions when creating personalized search results.

B. Problem definition

To protect user privacy in profile-based PWS, researchers have to consider two contradicting effects during the search process. On the one hand, they attempt to improve the search quality with the personalization utility of the user profile. They need to hide the privacy contents existing in the user profile to place the privacy risk under control. Significant gain can be obtained by personalization at the expense of only a small (and less-sensitive) portion of the user profile, namely a generalized profile. Thus, user privacy can be protected without compromising the personalized search quality. In general, there is a tradeoff between the search quality and the level of privacy protection achieved from generalization. Unfortunately, the previous works of privacy preserving PWS are far from optimal.

C. Proposed system

This paper, proposed a personalized web search environment called WCEP (Web Content Extractor for Privacy Protection) that can adaptively generalize profiles by queries while respecting user specified privacy requirements. The key component for privacy protection is an online profiler implemented as a search proxy running on the client machine itself. We present two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization.

D. Proposed system advantages

- Increasing usage of personal and behaviour information to profile its users, which is usually gathered implicitly from query history, browsing history, click-through data bookmarks, user documents, and so forth.
- The framework allowed users to specify customized privacy requirements via the hierarchical profiles. In addition, WCEP also performed online generalization on user profiles to protect the personal privacy without compromising the search quality.

III. SYSTEM DESIGN

1. System architecture

- Online Profiler

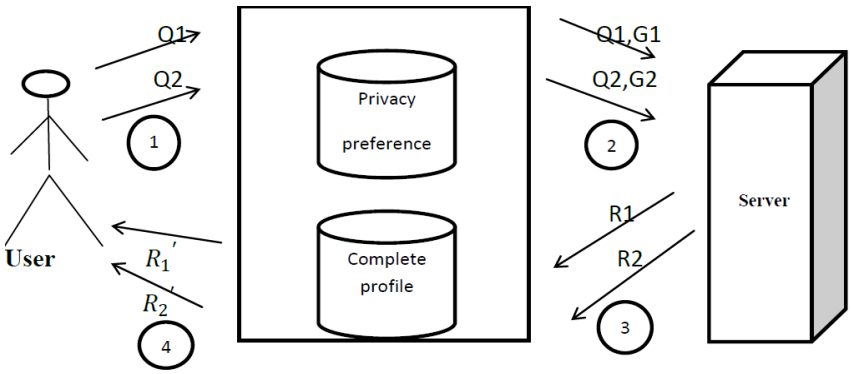

Figure 1 : Block Diagram of online profiler

As shown in figure online profiler comprises of number of customers/clients and a server for sufficient customers demand. In customer's machine, the online profiler is performed as search intermediary who keeps up clients profile in advanced system of hubs additionally keep up the client resolute security essential as an arrangement of sensitive hubs. There are two stage, specifically Offline and Online stage for the structure. Among Offline, a various leveled client profile is made and client determined security condition is stamped on it. The inquiry let go by client is taken care of in the online stage as: At the point when client fires a question on to the customer, intermediate produces client profile in run time.

2. Algorithm

A greedy algorithm is a mathematical process that looks for simple, easy-to-implement solutions to multipart, multi-step problems by deciding which next step will provide the most obvious benefit. This algorithm is applied for pruning process. But it leads to computational cost increasingly. Greedy DP algorithm is a bottom up manner during the interaction to maintain best profile .

- **GreedyDP (Discriminating Power) Algorithm**

This algorithm works in a bottom up manner. Starting with the leaf node, for every repetition, it chooses leaf topic for clipping thus trying to maximize convenience of output. During repetition a best profile-so- far is maintained satisfying the Risk constriction. The repetition stops when the root topic is reached. The best profile-so-far is the final result. GreedyDp algorithms require recompilation of profiles which adds up to computational cost and memory requirement.

- **Algorithm:**
  1. Prune the leaf node from profile
  2. Add Leaf to Personalized Search
  3. If Leaf=0 goto step 1

- **Greedy-IL (Information Loss) Algorithm**

GreedyIL algorithm is advances simplification productivity. GreedyIL continues importance queue for candidate clip leaf operator in descending order. This decreases the computational cost. GreedyIL states to dismiss the repetition when Risk is satisfied or when there is a single leaf left. Since, there is less computational cost compared to GreedyDP, GreedyIL out performs GreedyDP.

IV. PERFORMANCE ANALYSIS

Web Content Extractor for privacy is implemented with different metrics for analysis. Different methods are used to figure out various performance metrics such as scalability of generalization of algorithms, average response time and various effective analysis based on selected queries.

This framework produced the best average response time for test queries. On average, this work outperformed privacy in improved manner. This work also evaluated by comparing discriminative naive bays classifier with previous work. Proposed system proved significant improvement over existing system.

A. Experimental Setup

The WCEP framework is implemented on a PC (Localhost) with a PENTIUM IV 2.6 GHz, Intel Core 2 Duo and 512 MB DD RAM, running Microsoft Windows 7. All the algorithms are implemented in Java. This paper discuss the results for following metric of utility for performance analysis.

1) Scalability of Generalization Algorithms
2) Effective Analysis of Personalization

i) Scalability of Generalization Algorithms

Experiment 1:

This metric is studied using the scalability of the proposed algorithms by varying the data set size (i.e., number of queries). It takes randomly 100 queries from the real query log. The count of GL and GP is evaluated as in big-oh computation time complexity. For 100 queries GP is performed with 6.49 second which lowers the average overhead of re-computation.

Table 1 shows the average response time in improved scalability and Figure 2 illustrates the reduction of re-computation.

Parameter	Existing system	Proposed System
Average Response Time(sec)	8(Sec)	6.49(Sec)

Experiment 2:

In this experiment, plotted the result of enhanced efficiency with following results

- Resultof GreedyDP
- Resultof GreedyIL

For fair comparison with existing system, system achived the better result of GreedyDP than the existing system. The detail illustration of highest rank of 1-100 queries from real query log with respect to response time is performed in Figure 4.13(a and b),y-axis for total time and x-axis for set of keyword at each iteration.

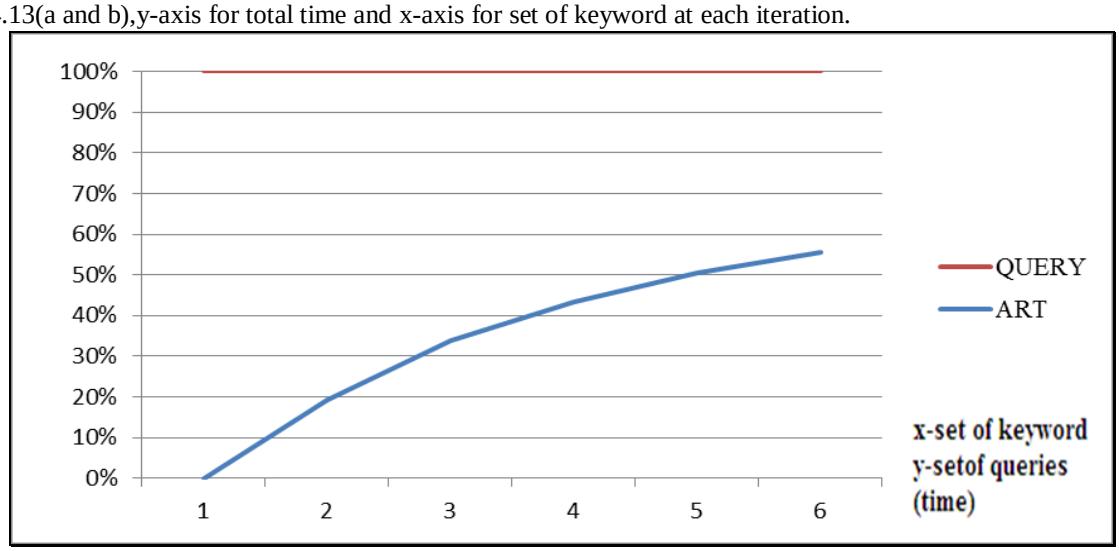

Figure 2 Result of GreedyDP

- **Effective Analysis of Personalization**

In this experiment, mentioned three different types of queries as "Wikipedia" for distinct queries, "Freestyle" for medium queries, and "Program" for ambiguous queries. Results are obtained as the real search quality over commercial search engines using WCEP framework. The search results are combined with the generalized profile output by GreedyIL over 50 target uploaded files.

During Offline-1 procedure, the relevance is obtained. A naive method [1] is to compute for each pair of *d* and *t 2 R* their relevance with a discriminative naive Bayesian classifier are defined in following equation.

$$dnb(d,t)=\sum_{w\in t}^{N} Nd,w * ln * \frac{Nt,w+\epsilon}{\sum t' \epsilon RNt',w+\epsilon,} \quad \text{.....(1)}$$

Results obtained in table 4.2 illustrates that our framework computes more effective relevance for the given topic in the document than the existing system. In given Figure x-axis represent relevance count and y-axis represent type of query.

Table 2: Evaluation of relevance for queries

Type of Query	ODP	WCEP
Distinct	1	1.324
Medium	0.44	1.118
Ambiguous	0.82	1.069

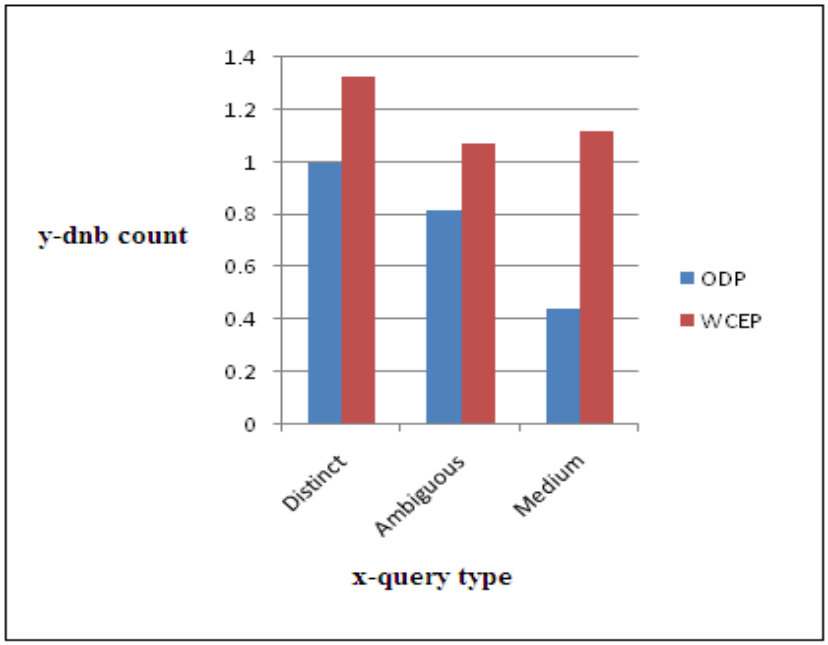

Figure 3: Comparison between ODP and WCEP framework for Relevance

V. CONCLUSION

This paper presented the experimental results of WCEP, shows significant improvements in user search results. WCEP 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. WCEP also performed online generalization on user profiles to protect the personal privacy without compromising the search quality. User can experience better search quality with effective privacy protected web content.

REFERENCES

[1] L.Shou, H.Chen, and G. Chen, "Supporting Privacy Protection in Personalized Web Search," IEEE transactions on KDD, VOL. 26, NO. 2, pp. 453-467, 2014.

[2] 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.

[3] 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.

[4] A. Ukande, N. Shivale, "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.

[5] 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.

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

[8] 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.

[9] J. Pitkow, H. Schu¨ tze, T. Cass, R. Cooley, D. Turnbull, A. Edmonds, E. Adar, and T. Breuel, "Personalized Search," Comm. ACM, vol. 45, no. 9, pp. 50-55,2002.

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

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

