Volume 6, Issue 1, January 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Personalized Web Content Extractor for Privacy Protection

Sonali K. Shelke*, Padmapani P. Tribhuvan
Dept .Computer Science & Engineering, Deogiri Institute of Engineering & Management Studies,
Aurangabad, Maharashtra, India

Abstract— 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. This irrelevance is largely due to the enormous variety of user contexts and backgrounds, as well as the ambiguious texts. However, research studies show that user's private information during search has become known to publicly due to proliferation of PWS. We propose a PWS framework called WCEP that can adaptively generalize profiles by queries without violating user specified privacy requirements. We present 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

I. INTRODUCTION

The overall goal of the web mining process is to extract web content from a data set and transform it into an understandable format for further use. The real task of data mining 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[5],[6]. Personalized search is to search experiences that are tailored specifically to an individual's intensions/interests by incorporating information about the individual beyond specific query provided. It introduces potential privacy problems in which a user may not be aware that their search results are personalized for them [1], and surprised why the things that they are interested in have become so relevant. An interesting point about personalization that often gets overlooked is the privacy versus personalization in terms of battle. Google provides a host of services to people, and many of these services do not require personal information to be collected about a person to be customizable. Since there is no threat of privacy invasion with these services.

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]. In this paper, we study how to exploit implicit feedback information, including previous queries and Click through information, to improve retrieval accuracy in an interactive information retrieval setting.

II. LITERATURE SURVEY

Personalized web search is a promising way to improve search quality by customizing search results for people with individual information goals. However, users are not comfortable with exposing private preference information to search engines. On the other hand, privacy is not absolute, and often can be compromised if there is a gain in service or profitability to the user[10]. Thus, there should be a balance between search quality and privacy protection[2].

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

The existing methods do not take into account the customization of privacy requirements. This probably makes some user privacy to be overprotected while others insufficiently protected[8],[9].

For example, in, all the sensitive topics are detected using an absolute metric called based on the information theory, assuming that the interests with less user document support are more sensitive[3],[6],[7]. However, if user search through personalized web search, his/her details are tracked with browser history. Sometimes userget failure with false-clik methods in run time environment, and he/she completely loss then content of privacy.

B. Proposed system

Web search engines (e.g. Google, Yahoo, Microsoft Live Search, etc.) are widely used to find certain data among a huge amount of information in a minimal amount of time. However, these useful tools also pose a privacy threat to the users: web search engines profile their users by storing and analyzing past searches submitted by them [4]. We propose a PWS framework called WCEP (Web content extraction for privacy) 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. Privacy to a user account is distinguish between normal search and personalized search. To accomplish this we present two greedy algorithms, namely Greedy DP and Greedy IL, for runtime generalization.

The framework works in two different phases of security.

1) User contents are secured with their account
2) Enhanced privacy search with personal and general key for secure search.

C. Proposed system advantages

- Works on different types of queries from user.
- Customization of privacy requirements.
- More scalable in terms of computation complexity
- It achieves better accuracy when compared with the Existing system

III. SYSTEM DESIGN

A. System architecture

User profiles are generalized using greedy IL. The finding motivates us to maintain a priority queue of candidate prune-leaf operators in descending order of the information loss caused by the operator. This queue, enables fast retrieval of the best so- far candidate operator. Filtering results based on WCEP and results are shown to user. Figure 1shows the detail system architecture of WCEP framework.

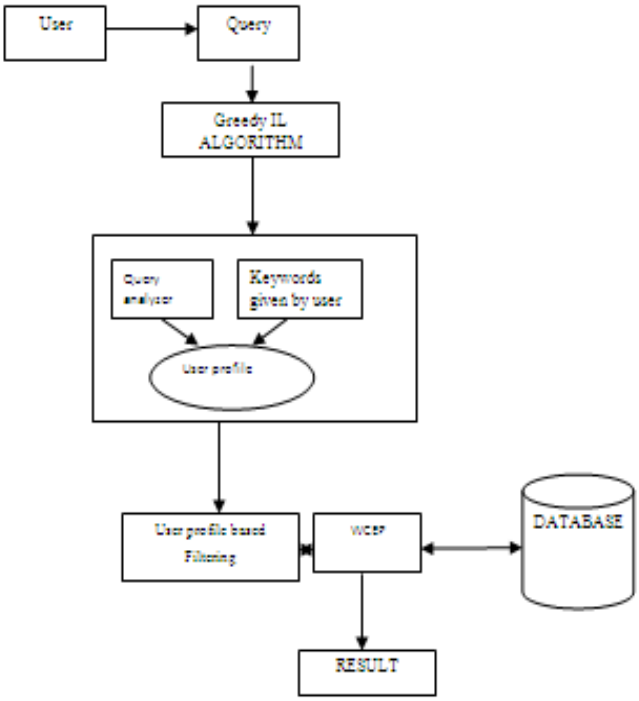

Figure 1: System architecture of WCEP

B. Methodologies

Following modules involves

- User interface.
- Query processing.
- Combining User profile and query
- Online Generalization
- Search personalization

- USER INTERFACE DESIGN:
  - To connect with server user must give their username and password then only they can able to connect the server. If the user already exits directly can login into the server else user must register their details such as username, password and Email id, into the server. Server will create the account for the entire user to maintain upload and download rate.

- QUERY PROCESSING:
  - In this module, the data is given by customer requests goes to server , When a user issues a query 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 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 were administrator maintains all files and responsible for storing that files into cloud.

- COMBINING USER PROFILE AND QUERY:
  - In this model, user given query and the generalized user profile are sent together to the PWS server for personalized search. Query with related user preferences stored in a user profile with the aim of providing better search results.

- ONLINE GENERALIZATION:
  - In this model, user given query based on privacy requirements and cost of profiling search results are checked whether to personalize or not.

- SEARCH PERSONALIZATION:
  - In this model user given query search results are personalized according to user profile and delivered back to the query proxy. After results are shown to user.

IV. PERFORMANCE ANALYSIS

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. In this paper we discuss results for following metric of utility for performance analysis.

- Scalability of Generalization Algorithms
- Effective Analysis of Personalization

i) Scalability of Generalization Algorithms

EX 1:

We study the scalability of the proposed algorithms by varying the data set size (i.e., number of queries). we randomly choose 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 graph 1 illustrates the reduction of re-computation.

Table 1: Comparison between Existing System and Proposed System

Parameter	Existing system	Proposed System
Avg. Response Time(sec)	8	6.487833333

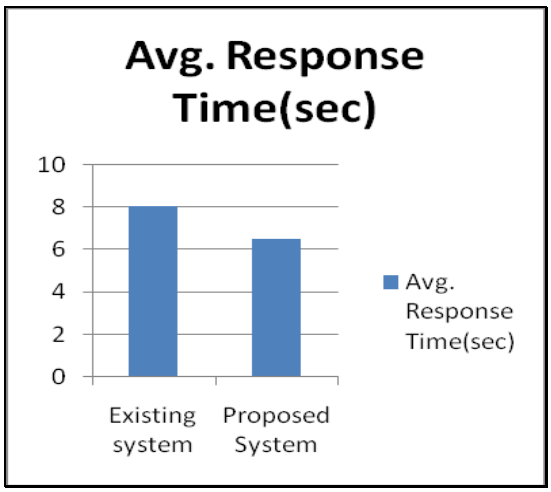

Figure 2: Graphical Analysis of Average Response Time

EX2: In this experiment, we plotted the result of enhanced efficiency with following results

- Resultof GreedyDP
- Resultof GreedyIL

For fair comparison with existing system, we 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 3(a and b).

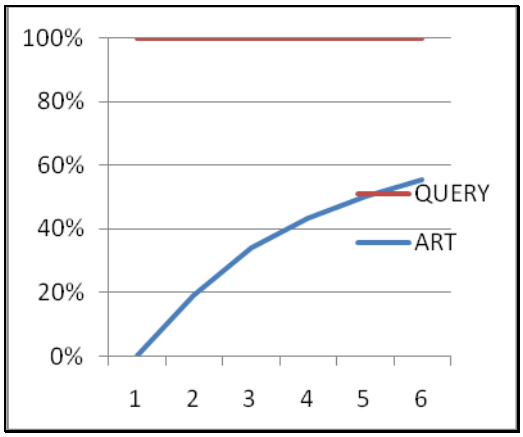

Figure 3(a) Result of GreedyDP

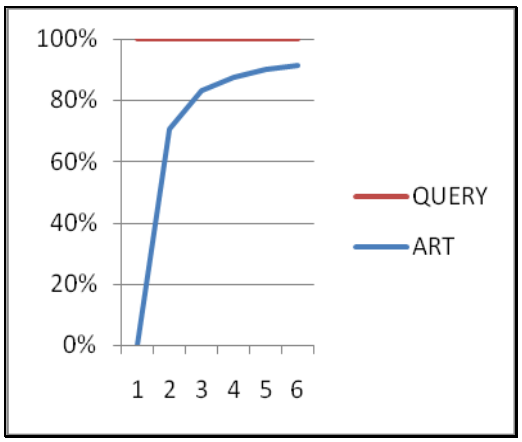

Figure3 (b) Result of GreedyIL

ii. Effective Analysis of Personalization

In this experiment, we mentioned three different types of queries as "Wikipedia" for distinct queries, "Freestyle" for medium queries, and "Program" for ambiguous queries. We performed the real search quality over commercial search engines using WCEP framework. The search results is combined with the generalized profile output by GreedyIL over 50 target uploaded files. The final search quality is evaluated using the Average Precision of the click records of the users, which is defined as

$$\mathcal{AP} = \sum_{i=1}^{n} \frac{i}{l_i.rank} / n,$$

where li is the ith relevant link identified for a query, and n is the number of relevant links[1]. The details of results compared with Google search engine is given in following table 2 and chart 1.

Table 2: Evaluation of Three Representative Queries

Type of Query	Google	WCEP
Distinct	0.005	0.025
Ambiguous	0.0042	0.01
Medium	0.02	0.1

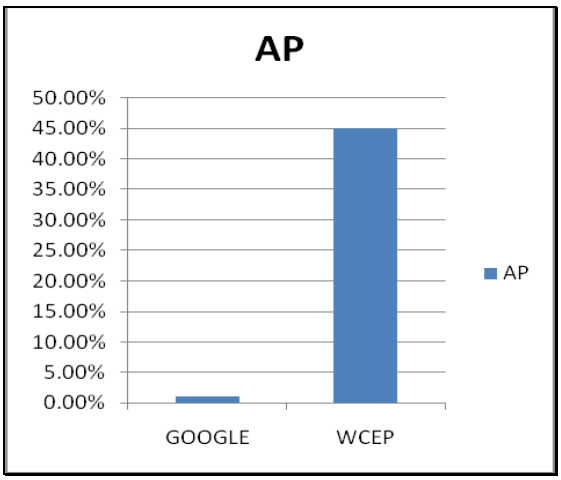

Chart 1: Comparison between Google search and WCEP framework w.r.t Average Precision

During Offline-1 procedure, we need to calculate the relevance. A naive method is to compute for each pair of d and t 2 R their relevance with a discriminative naive Bayesian classifier as defined as

$$dnb(d,t) = \sum_{w \in d} N_{d,w} \ln \frac{N_{t,w} + \epsilon}{\sum_{t' \in \mathcal{R}} N_{t',w} + \epsilon},$$

Results obtained in table 3 illustrates that our framework computes more effective relevance for the given topic in the document than the existing system.

Table 3: Evaluation of Relevance For Queries

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

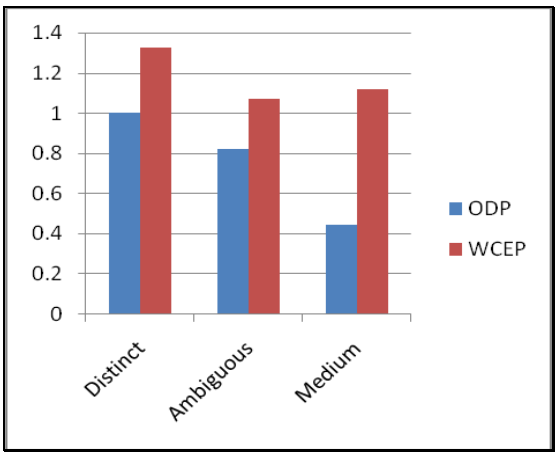

Chart 2: Comparison between ODP and WCEP framework for Relevance

V. CONCLUSION

This paper presented the experimental results of UPS, 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.

In future work we will try to enhance the keyword search through broader relationship among the data over the real time environment. We will also seek to evaluate better utility performance metrics.

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.

