International Journal of Research
Available at https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

Implementation of Personalized Web Search Using Learned User Profiles

M.Vanitha[1] & P.V Kishan Rao[2]
[1]P.G-Scholar Dept. of CSE TKR College of Engineering andTechnology, TS, Hyderabad.
[2]Assoc.professorDept. Of CSETKR College ofEngineering and Technology, TS, Hyderabad

Abstract:

With increasing number of websites the Web users are increased with the massive amount of data available in the internet which is provided by the Web Search Engine (WSE). Personalized web search (pws) refers to search experiences that are tailored specifically to an individual's interests by incorporating information about the individual beyond specific query provided. Which is involving modifying the user's query and the other re-ranking search results.[1] Generally WSE is to provide the relevant search result to the user with the behavior of the user click were they performed. WSE provide the relevant result on behalf of the user frequent click based method. From this method no assurance to the user privacy and also no securities were providing to their data. Hence users were afraid for their private information during search has become a major barrier. They were many techniques were proposed by researchers most of that based on the server side, it has provide less security. For minimizing the privacy risk here propose the client side based technique with the combination of Greedy method to prevent the user data that we applied in Knowledge mining area. Proposed framework called UPS that can adaptively generalize profiles by queries while respecting user's privacy requirements. Proposed work consists two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization.

Index Terms— Privacy Protection; profile; personalized web search; risk; UPS

1. INTRODUCTION

The web search engine has gained a lot of popularity and importance for users seeking information on the web. Since the contents available in web is very vast and ambiguous, users at times experience failure when an irrelevant result of user query is returned from the search engine. Therefore, in order to provide better search result a general category of search technique Personalized Web search is used. In personalized web search, user information is collected and analyzed in order to find intention behind issued query fired by user. The explosive growth of documents in the Web makes it difficult to determine which are the most relevant documents for a particular user, given a general query. Recent search engines rank pages by combining traditional information retrieval techniques based on page content, such as the word vector space [4, 6], with link analysis techniques based on the hypertext structure of the Web [7, 8]. Personalized search has gained great popularity to improve search effectiveness in recent years [10, 12, 2]. The objective of personalized search is to provide users with information tailored to their individual contexts. We propose to personalize Web search based on features extracted from hyperlinks, such as anchor terms or URL tokens. Our methodology personalizes PageRank vectors by weighting links based on the match between hyperlinks and user profiles. In particular, here we describe a profile representation using Internet domain features extracted from URLs.

We identify two aspects of link analysis. One is the global importance of pages as estimated from analyzing the Web link graph structure. There is a major body of research exploring retrieval techniques based on link popularity such as PageRank [5] and HITS [3]. Another aspect of link analysis is the structure of the hyperlinks themselves. For example, anchor text has been shown to be a very good predictor of content of the linked page. One can expect that keywords in the anchor text of a link might be highly related with the content of that page. The accuracy and quality of a page can also be estimated by looking at its URL. Web pages published under an educational institution Web site might be deemed to have higher prestige compared to those published under free Web hosting sites. In this research, we combine these two aspects of link analysis: PageRank and hyperlink structure to improve search effectiveness through personalized search.

Although our formalization is general, in this paper we specifically consider its application to the task of personalization using topic-based profiles. We have one discrete variable for each document whose states specify the topic of the document. The state space that we use corresponds to the top two levels of the human-generated ontology provided by the Open Directory Project (ODP, dmoz.org). Some example categories are 'Sports', 'Arts/Movies', and 'Shopping'. In a pre-processing step, we use a text-based classifier, trained with logistic regression, to obtain the distribution over topics for each document in the index. This allows the personalized ranking to be computed extremely quickly at query time.

2. RELATED WORK

In Existing approaches mainly focused on users interests. There is a growing interest in the information retrieval and machine learning communities in moving beyond context free search experiences, and toward examining how knowledge of a searcher's interests and search context can be used to improve various aspects of search (e.g., ranking, query suggestion, query classification). For example, there has been work on using session context, such as the previous few searches or result clicks, to personalize search results and improve retrieval performance. Short-term session profiles have also been used for other tasks such as predicting future interests [11], query categorization [9], query suggestion, and URL recommendation. We focus on personalizing using user profiles constructed from logs comprising long-term interaction behaviors, potentially providing a richer view of searcher interests over time. Another line of prior research uses long-term histories to directly improve retrieval effectiveness. Teevan et al. [15] constructed user profiles from indexed desktop documents and showed that this information could be used to re-rank search results and improve relevance for individuals. Matthijs and Radlinski [18] constructed user profiles using users' browsing history, and evaluated their approach using an interleaving methodology. Rather than using all of the previous search history, Tan et al. [16] focused only on the most relevant prior queries and constructed language models for this task. Personalization is not equally effective on all queries. Teevan et al. [17] introduced a framework to identify the potential-for-personalization for different queries. In particular, the implicit measure click entropy (the number of different results that different people clicked) was highly correlated with explicit judgments of relevance by individuals. All of these approaches to personalization use word-based profiles, and ranking is done by reweighting terms using an existing scoring method such as BM25 or TFIDF. In contrast, our approach uses a higher-level representation.

3. PERSONALIZED WEB SEARCH

Today's search engines usually cannot distinguish different users' needs well. For example, a computer scientist may use the search query "leopard" to locate information on Apple OS X Leopard and a biologist may use the same query for the animal leopard; however, a search engine usually treats the two queries the same way. Alternatively, personalized search provides customized results.

Based on literature work we introduced a scoring function for personalizing search results. The function uses four characteristics to score a term that matches the user profile (called UIH), which is learned from the user's interests. 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 issued query. The main contributions of this paper are:

1. When a user issues a query, the proxy generates a user profile in runtime in the light of query terms.

2. Subsequently, the query and the generalized user profile are sent together to the PWS server for personalized search.

3. The search results are personalized with the profile and delivered back to the query proxy.

4. Finally, the proxy either presents the raw results to the user, or reranks them with the complete user profile

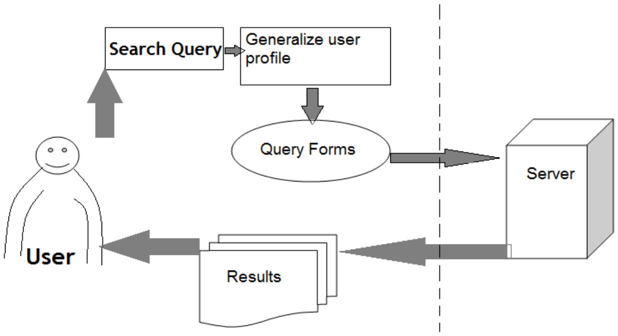
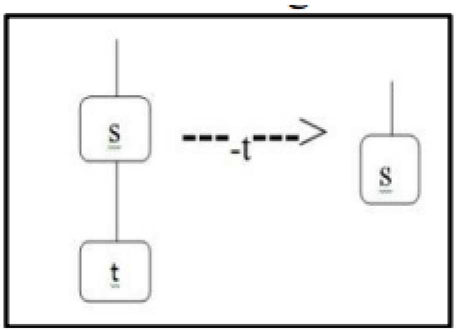


Fig. 1.System architecture of UPS.

The project uses two algorithms for

1) Greedy DP: Greedy Discriminating power [7].This algorithm gives optimal solution hencecalled a Near Optimal Greedy Algorithm. Forremoval of leaf topic from profile we will introducean operator ----t--> This is called Prune leaf . Wemay have 2 cases for removal of leaf.

Case 1: When t has no siblings t has no siblings



Once a leaf topic t is pruned, only the candidateoperators pruning t's sibling topics need to be updated in Q. Inother words, we only need to recompute the IL values foroperators attempting to prune t's sibling topics.

2) GreedyIL: To increase the efficiency GreedyIL algorithm is used [7].

Following terminologies are used in GreedyIL algorithm.

G0: Seed profile

q:query

δ : Privacy Threshold.

G*: Generalized profile satisfying δ- Risk.

```
        Q: IL-priority queue of prune-leaf decision.
i: Iteration index initialized to 0.
input is G0, q, δ.
    Output: G*.
Following steps will be carried out for online decision
whether to personalize q or not
If DP(q,R) < μ then do following:
     Obtain the seed profile G0 from Online-1,
Insert(t,IL(t)) into Q for all to ε T(q)
 While risk(q,Gi) > δ do
     Pop a prune-leaf operation on t from Q
     Set s ←part(t,$G_i$)
      Process prune leaf Gi If t has no siblings then
//case 1
      Insert(s,IL(s)) to Q Else if t has siblings then
//case2
    Merge t into shadow-sibling
 If No operation on t's siblings in Q then
   Insert(s,IL(s)) to Q
Else Update IL- value for all operations on t's sibling
    Update i <=i+1
    Return Gi as G* return root(R) as G*
```

Based on literature reviews proposes a privacy-preserving personalized web search framework called UPS i.e User customizable Privacy-preserving Search, that generalize profile for every query as per user privacy specification. Based on personalization and privacy risk metric, this paper formulate Risk Profile Generation, with its NP- hardness proved. It develops two simple but effective generalization algorithms, GreedyDP and GreedyIL, to support runtime profiling. GreedyDP maximize the discriminating power (DP) while GreedyIL minimize the information loss (IL). This paper also provide a mechanism for the client to decide whether or not to personalize a query in UPS. This decision is made before each runtime profiling to enhance the stability of the search results.

4. PROPOSED WORK

In Addition of proposed system in Personalized web search (PWS), we are taking user personal information for PWS, like their interests, for instance user interested MYSQL in SQL hierarchy, when user search for MYSQL the system will retrieve results like SQL/DATABASE/MYSQL, that means based public hierarchy *P*, the results will retrieving n number of ways related to his interests. In proposed they didn't consider user personal profile information for PWS, like age, postal code. In proposed work also considering where they located, what is the age group of user? Like youtube we have option like 'Popular in INDIA'. That means they are doing pws like which videos are famous in India.

If consider users age, postal code (Address) we can retrieve results based on the users age category, like middle age group people what are they willing to search, so on so Hyderabad people what they willing to search like that. But for security of the personal profile information, For security of users personal information we anonymized the data like to k- anonymity with data Suppression. Age we are doing k-anonymity for data individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' 11-20', the value '29' by '21-30' ,etc

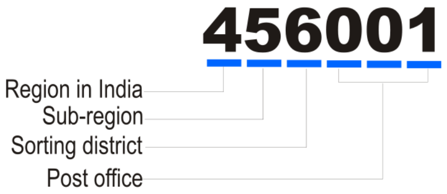
For example

Age	Postal Code
19	500031
25	504231
39	500016

Anonymization table,

Age	Postal Code
11-20	500***
21-30	504***
30-40	500***

And postal code we make data suppression, first 3 digits represents city name and exact location. So If suppress last 3 digits, we will get information of only city, not exact area.

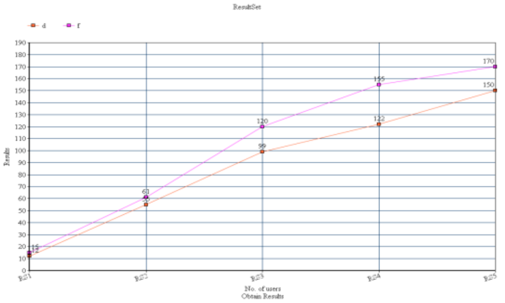


So we can achieve best results as well as security in profile.

5. EXPERIMENTAL RESULTS

In this section, we present the experimental results of UPS. The UPS framework is implemented on a PC with aPentium Dual-Core 2.50-GHz CPU and 2-GB main memory,running Microsoft Windows XP. All the algorithms areimplemented in Java.

In this experiment, we analyze and compare the effect of the generalization on queries with different discriminating power, and study the tradeoff between existing and proposed in the GreedyDP/GreedyIL algorithm.



6. CONCLUSION

This paper improves our previous work on personalized ranking by enhancing the accuracy of scoring function. A client side privacy protection framework called UPS i.e User customizable Privacy preserving Search is presented in the paper. Any PWS can adapt UPS for creating user profile in hierarchical taxonomy. UPS allows user to specify the privacy requirement and thus the personal information of user profile is kept private without compromising the search quality. UPS framework implements two proposed greedy algorithms for this purpose, namely GreedyDP and GreedyIL. Our 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.

7. REFERENCES

[1] https://en.wikipedia.org/wiki/Personalized_search

[2]. Fang Liu, Clement Yu, WeiyiMeng: Personalized Web Search For Improving Retrieval Effectiveness. IEEE Transactions on Knowledge and Data Engineering, January 2004

[3]. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (1999) 604– 632

[4]. Salton, G., McGill, M.: An Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY (1983)

[5]. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks 30 (1998) 107–117

[6]. van Rijsbergen, C.: Information Retrieval. Butterworths, London (1979) Second edition.

[7]. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks 30 (1998) 107–117

[8]. Kleinberg, J.: Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (1999) 604–632

[9] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li. Context-aware query suggestion by mining click-through and session data. In KDD '08, pages 875–883, 2008.

[10]. Haveliwala, T.: Topic-sensitive PageRank. In Lassner, D., De Roure, D., Iyengar, A., eds.: Proc. 11th International World Wide Web Conference, ACM Press (2002)

[11] R. W. White, P. N. Bennett, and S. T. Dumais. Predicting short-term interests using activity-based search context.In CIKM '10, pages 1009–1018, 2010.

[12]. Jeh, G., Widom, J.: Scaling personalized Web search. In: Proc. 12th International World Wide Web Conference. (2003)

[13] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford University Database Group (1998)

[14]. Brin, S., Motwani, R., Page, L., Winograd, T.: What can you do with a Web in your pocket. IEEE Data Engineering Bulletin 21 (1998) 37–47

[15] J. Teevan, S. Dumais, and E. Horvitz.Personalizing search via automated analysis of interests and activities. In SIGIR '05, pages 449–456, 2005.

[16] B. Tan, X. Shen, and C. Zhai. Mining long-term search history to improve search accuracy. In SIGKDD '06, pages 718–723, 2006.

[17] J. Teevan, S. Dumais, and E. Horvitz.Potential for personalization. ACM TOCHI, 17(1), 2010

[18] N. Matthijs and F. Radlinski.Personalizing web search using long term browsing history. In Proceedings of the fourth ACM international conference on Web Search and Data Mining, WSDM '11, pages 25–34, New York, NY, USA, 2011. ACM.