PERSONALIZED WEB SEARCH USING BROWSING HISTORY AND DOMAIN KNOWLEDGE

Rakesh Kumar
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, India
rakesh.kmr2509@gmail.com

Aditi Sharan
School of Computer and Systems Sciences
Jawaharlal Nehru University
New Delhi, India
aditisharan@gmail.com

***Abstract*—Generic search engines are important for retrieving relevant information from web. However these engines follow the "one size fits all" model which is not adaptable to individual users. Personalized web search is an important field for tuning the traditional IR system for focused information retrieval. This paper is an attempt to improve personalized web search. User's Profile provides an important input for performing personalized web search. This paper proposes a framework for constructing an Enhanced User Profile by using user's browsing history and enriching it using domain knowledge. This Enhanced User Profile can be used for improving the performance of personalized web search. In this paper we have used the Enhanced User Profile specifically for suggesting relevant pages to the user. The experimental results show that the suggestions provided to the user using Enhanced User Profile are better than those obtained by using a User Profile.**

Keywords—Personalized Web Search, User Modeling, Domain Knowledge, Enhanced User Profile

I. INTRODUCTION

With the development of World Wide Web, web search engines have contributed a lot in searching information from the web. They help in finding information on the web quick and easy. But there is still room for improvement. Current web search engines do not consider specific needs of user and serve each user equally. It is difficult to let the search engine know what we the user actually want. Generic search engines are following the "one size fits all" model which is not adaptable to individual users.

When different users give same query, same result will be returned by a typical search engine, no matter which user submitted the query. This might not be appropriate for users which require different information. While searching for the information from the web, users need information based on their interest. For the same keyword two users might require different piece of information. This fact can be explained as follows: a biologist and a programmer may need information on "virus" but their fields are is entirely different. Biologist is searching for the "virus" that is a microorganism and programmer is searching for the malicious software. For this type of query, a number of documents on distinct topics are returned by generic search engines. Hence it becomes difficult for the user to get the relevant content. Moreover it is also time consuming. Personalized web search is considered as a promising solution to handle these problems, since different search results can be provided depending upon the choice and information needs of users. It exploits user information and search context to learning in which sense a query refer.

In order to perform Personalized Web search it is important to model User's need/interest. Construction of user profile is an important part for personalized web search. User profiles are constructed to model user's need based on his/her web usage data.

This paper proposes an architecture for constructing user profile and enhances the user profile using background knowledge. This Enhanced User Profile will help the user to retrieve focused information. It can be used for suggesting good Web pages to the user based on his search query and background knowledge.

The paper is organized as follows: Section 2, gives the related work focusing on personalized search systems. Section 3, proposes the framework for personalized web search that satisfies each user's information need by enhancing the user's profile without user's effort. Next section, we presents the experimental results for evaluating our proposed approaches. Finally, we conclude the paper with a summary and directions for future work in Section 5.

II. RELATED WORK

Framework for Personalized search engine consists of user modeling based on user past browsing history or application he/she is using etc. And then use this context to make the web search more personalized. This section presents different approaches and the related work done in the field of Personalized Web search.

For providing personalized search results, Micro Speretta et al., [1] implemented a wrapper around the search site that collects information about user's search activity and builds user profile by classifying collected information (queries or snippets). They have used these profiles to re-rank the search results and the rank-order of the user-examined results before

978-1-4799-2900-9/14/$31.00 ©2014 IEEE

and after re-ranking were compared. They found that user profiles based on queries and user profiles based on snippets both were equally effective and re-rank gave 34% improvement in compare to rank-order.

Fang Liu et al., [2] identified that current web search engines do not consider the special needs of user or interests of user and proposes a novel technique which uses search history of user to learn user profiles. This work uses user's search history for learning of user profile and category hierarchy for learning of a general profile and then combines both profiles to categorize user's query to represent user's search intention and to disambiguate the words used in query.

Chunyan Liang [3] also identifies that different users may have need of different special information, when they use search engines and techniques of personalized web search can be used to solve the problem effectively. Three approaches Rocchio method, k-Nearest method and Support Vector Machines have been used in [3] to build user profile to present an individual user's preference and found that k-Nearest method is better than others in terms of its efficiency and robustness.

Xuwei Pan et al., [4] suggested a context based personalized web search model. In this paper the authors have given a personalized web search outcome which is in accordance with the need of user in various situations. The analysis of model has resulted in three concepts to implement the model, which is semantic indexing for web resources, modeling and acquiring user context and semantic similarity matching between web resources and user context. The author has defined it as context based adaptive personalized web search

K. W. T. Leung et al., [5] have proposed a Personalized Web search model with location preferences. In this paper the location and content concept has been separated and is organized into different ontology to make an ontology-based, multi-facet (OMF) profile which is captured by web history and location interest. This model actually gives results by outlining the concepts in accordance with the preference of user. By keeping the diverse interest of the users in mind, location entropy is introduced for finding the degree of interest and information related to location and query. The personalized entropies actually estabilize the relevant output content and location content. At last, an SVM based on the ontology is derived which can be used for future purpose for ranking or re-ranking. The experiments shows that the results produced by OMF profiles are more accurate in comparison with the ones which use baseline method.

O. Shafiq et al., [6] have proposed a personalized web search model that combines community based and content based evidences based on novel ranking technique. Nowadays, uploading data on internet has become a daily activity. A massive amount of data is uploaded in the form of web pages, news, and blogs etc. on a regular basis. So, it becomes very difficult for the user to search for relevant content. Not only for users but also for search engines like Google and Yahoo it becomes difficult. Information overload is the only reason behind this difficult situation. Other than this user's preference is the second problem, which is not taken into consideration while producing the results. The author tried to solve this problem through this model which produce results on the basis of preference and interest of the user. In this paper, authors proposed a unique approach to find out the interest and preference of the user. It's a two way approach, first it will find out the activities of user through his/her profile in social networking sites. Secondly, it will find out information from what the social networking sites provide to the user through friends and community. Based on the results, user's interest and preference will be prioritized by the web search or it is personalized.

III. FRAMEWORK FOR PROPOSED SYSTEM

We propose a framework for personalized web search which considers individual's interest into mind and enhances the traditional web search by suggesting the relevant pages of his/her interest. We have proposed a simple and efficient model which ensures good suggestions as well as promises for effective and relevant information retrieval. In addition to this, we have implemented the proposed framework for suggesting relevant web pages to the user.

General Architecture of Proposed Framework

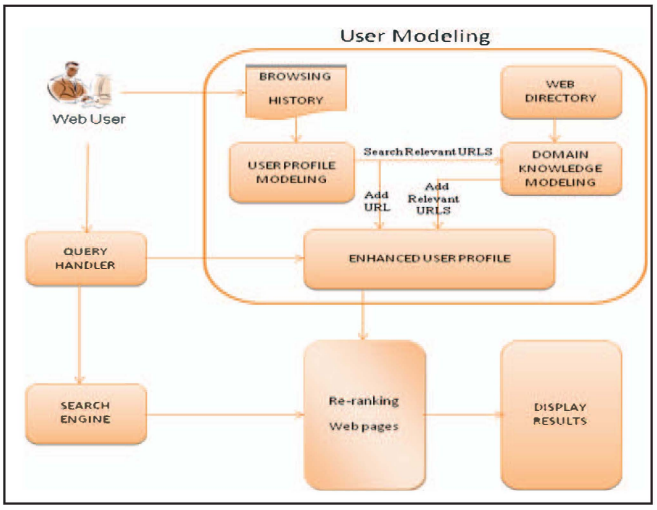

Figure 3.1 General Architecture of Proposed Framework

Our system considers user's profile (based on user's weblog/navigation browsing history) and Domain Knowledge in order to perform personalized web search. Using a Domain Knowledge, the system stores information about different domain/categories. Information obtained from User Profile is classified into these specified categories. The learning agent learns user's choice automatically through the analysis of user navigation/browsing history, and creates/updates enhanced User Profile conditioning to the user's most recent choice. Once the user inputs query, the system provides good suggestions for personalized web search based on enhanced user profile. Further our model makes good use of the advantages of popular search engines, as it can re-rank the results obtained by the search engine based on the enhanced user profile.

A. Domain Knowledge Modeling

Domain knowledge is the background knowledge that we used to enhance the user profile. The source which we have used for preparing Domain Knowledge is DMOZ directory. For preparing Domain Knowledge, first we have crawled the web pages from DMOZ directory for some specified categories, where each category is represented by collection of URL's present in that category.

After crawling, we have extracted the keywords from the crawled web pages. The collections of keywords form the vocabulary for the crawled pages. Now we form a term-category matrix, which specifies weight of each term in each category. The weight may be represented by frequency of the term in that category. Here w_{ij} represents number of times the term t_i is present in Category $Cate_j$. The matrix may be represented as follows:

Table 3.1 Terms – Category Matrix (TCM)

Term / Category	**$Cate_1$**	**$Cate_2$**	**$Cate_3$**	**……..**	**$Cate_n$**
t_1	w_{11}	w_{12}	w_{13}		w_{1n}
t_2	w_{21}	w_{22}	w_{23}		w_{2n}
t_3	w_{31}	w_{32}	w_{33}		w_{3n}
.					
t_m	w_{m1}	w_{m2}	w_{m3}		w_{mn}

B. User Profile Modeling

User profile is used to reflect user's interest and predict their intentions for new queries. User Profile also helps to deal with ambiguous queries. To create the user profile, we need to classify the web pages accessed by a user into particular category. AlchemyAPI has been used for classifying web pages. AlchemyAPI classifies a web page by giving it a particular category along with confidence (numerical value) which shows its probability of belonging to that particular category. If the web page is classified with confidence above the specified threshold level then only we have consider that page to contribute for that category.

As we are using DMOZ for background knowledge, we have to map these Alchemy categories to DMOZ categories.

Thus in our model, a User Profile is represented as a category preference vector, where weight of each category represents user's interest in that category. As shown in the Figure 3.1, users browsing history is used to build user profile. When the number of web pages browsed by the user grows above the specified threshold, the learning agent updates user profile. User interest will thus be represented by fix number of categories weights. It can be denoted by

$$U= \{cw_1, cw_2, cw_3, ….., cw_m\}$$

Where, cw_i will be the number of web pages of category i visited by that user, normalized by maximum number of page visits among all categories.

Table 3.2 Alchemy API to DMOZ Category Mapping

Alchemy Categories	**DMOZ Categories**
Arts & Entertainment	Arts
Business	Business
Computers & Internet	Computers
Culture & Politics	Regional
Gaming	Games
Health	Health
Law & Crime	Society
Religion	
Recreation	Recreation
Science & Technology	Science
Sports	Sports
Weather	News

For modeling User Profile we have used Vector Space Model (VSM). We consider all the webpages present in browsing history of particular user. Each web page corresponds to a specific document. The outcome of vector space model is a term document matrix (TDM) which represents each webpage/document as a feature vector of terms. Here we consider each document as a URL.

Table 3.3 Terms – Document Matrix (TDM)

	d_1	d_2	d_3	……..	d_n
t_1	w_{11}	w_{12}	w_{13}		w_{1n}
t_2	w_{21}	w_{22}	w_{23}		w_{2n}
t_3	w_{31}	w_{32}	w_{33}		w_{3n}
.					
t_m	w_{m1}	w_{m2}	w_{m3}		w_{mn}

C. Enhanced User Profile

Enhanced User Profile is an important part in our framework. An Enhanced User Profile improves the User Profile by using the Domain Knowledge. For preparing the

Enhanced User Profile we have considered each URL of the User Profile, match it with Domain Knowledge URLs and add most relevant URLs to the Enhanced User Profile.

Following steps explain the process of preparing the Enhanced User Profile. Perform the following steps for each document (URL) in user profile :

- Select the URL from the User Profile.
- Add the URL to the Enhanced User Profile.
- Find the cosine similarity of this URL with the URLs present in user specific categories from the Domain Knowledgebase.
- Rank the URLs on descending order of cosine similarity.
- Retrieve top 20 URLs.
- Calculate the average of the cosine similarity of these top 20 URLs.
- From the top 20 URLs add only those URLs to the enhanced user profile whose similarity value is above the average value.

To summarize the process, for each URL (form user profile) most relevant URLs from the user specific Domain Knowledge category are added to prepare enhanced user profile.

The cosine formula used for the similarity of the URL u in User Profile to each web pages d_j in Domain Knowledge is as follows:

$$cosine(d_j, u) = \frac{< d_j * u >}{\|d_j\| \times \|u\|}$$

A cosine similarity measure is the angle between the web page in User Profile u and the document vector d_j.

IV. Experimental Results and Analysis

In the absence of standard benchmark datasets which is suitable for our problem, we have designed our own dataset. In our Experiment, we have used the browsing history of 10 different users from our university, 6 from Computer department and 4 from Life Science department. Our Experiment is conducted for 50 queries of which 35 queries from Computer domain and 15 queries from Life Sciences domain.

In order to collect the domain knowledge, we have crawled the datasets from DMOZ for the selected topics using Apache Nutch, while Apache Solr has been used for indexing crawled pages. By setting crawling parameters of Nutch we have restricted the crawling to specific DMOZ topic.

Using the information of user browsing history and domain knowledge, we create an Enhanced User Profile. Once the Enhanced User Profile is created, we take the user query and suggest the relevant web pages with respect the query. In our Experiment, we have used User Profile as a base case for suggesting the relevant pages and compared the results with the pages suggested from Enhanced User Profile. For each query, we suggest top 20 relevant documents from User Profile and for the same query we also suggest top 20 relevant documents from Enhanced User Profile. In order to compare the efficiency of the result, we compared the similarity of suggested documents with the user query.

In order to represent the result graphically, we have used the bar graph. The analysis of the result is done in 2 different ways. First one is the individual cosine similarities of suggested pages and the second one is the average cosine similarity obtained for top 20 suggested pages.

For each query, we draw a bar graph of the cosine similarity measure for each suggested web page. The Figure 4.1 to Figure 4.4 shows the graph for Query1 to Query4.

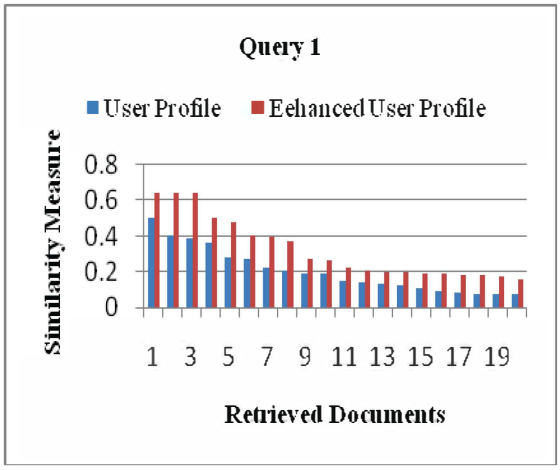

Figure 4.1 Bar graph of Cosine Similarities for Query1.

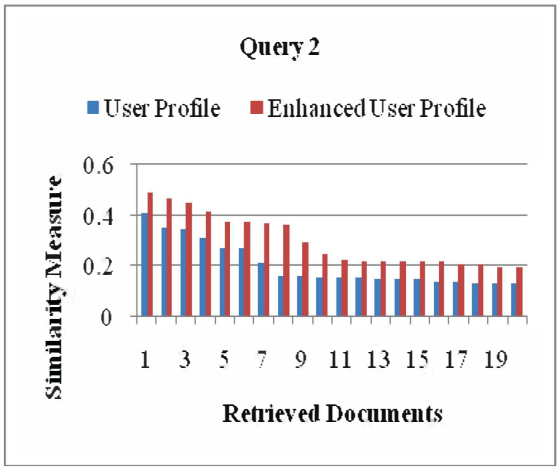

Figure 4.2 Bar graph of Cosine Similarities for Query2.

In this section, we have analyzed the results for different Queries. For each query, we have retrieved top 20 relevant web pages with User Profile and Enhanced User Profile. As we can see clearly from the above figures (Figure 4.1 to Figure 4.4) that all the queries show the improved result for Enhanced User Profile as compared to those suggested using User Profile.

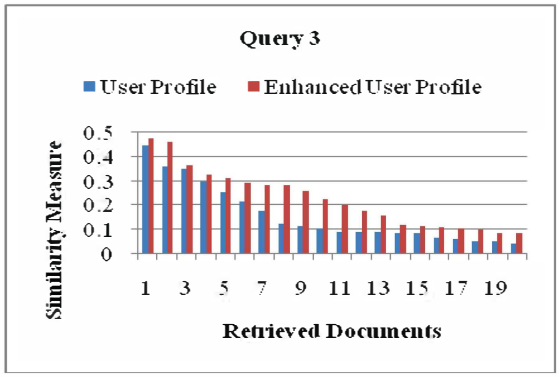

Figure 4.3 Bar graph of Cosine Similarities for Query 3.

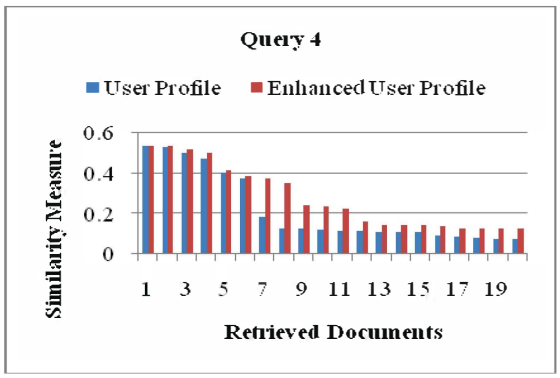

Figure 4.4 Bar graph of Cosine Similarities for Query4.

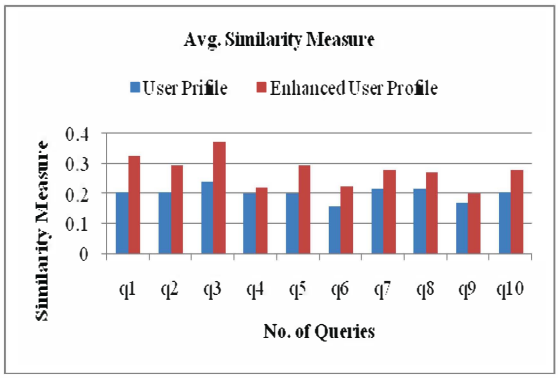

Figure 4.5 Bar graph of Avg. Cosine Similarities along with queries.

Figure 4.5 represents the bar graph of average cosine similarity obtained for top 10 queries. For all 10 queries, the average cosine similarity measure of User Profile and Enhanced User Profile has been calculated.

Figure 4.5 shows that the average improvement in Enhanced User Profile for all 10 queries as compared to User Profile. Experimental results show that our proposed model for personalize web search is effective for focused information retrieval and suggests good Web pages

V. Conclusion and Future Work

In this paper, we have proposed a framework for personalized web search using User Profile and Domain Knowledge. Based on the User Profile and the Domain Knowledge, the system keeps on updating the user profile and thus builds an enhanced user profile. This Enhanced user profile is then used for suggesting relevant web pages to the user. The proposed framework has been implemented by performing some experiments. These experiments shows that the performance of the system using enhanced user profile is better than those which are obtained through the simple user profile. Our work is significant as it improves the overall search efficiency, catering to the personal interest of the user's. Thus, it may be a small step in the field of personalized web search. In future this framework may be applied for re-ranking the web pages retrieved by search engines on the basis of user priorities. We may also apply collaborative filtering for personalized web search in our framework.

References

[1] M Speretta and S Gauch, "Personalized Search Based on User Search Histories", Proceeding Of International Conference on Web Intelligence, pp. 622-628, 2005.

[2] F Liu, C Yu and W Meng, "Personalized Web Search for Improving Retrieval Effectiveness", IEEE Transactions On Knowledge And Data Engineering, pp. 28-40, Volume 16, 2004.

[3] C Liang, "User Profile for Personalized Web Search", International Conference on Fuzzy Systems And Knowledge Discovery, pp. 1847–1850, 2011.

[4] X Pan, Z Wang and X Gu, "Context-Based Adaptive Personalized Web Search for Improving Information Retrieval Effectiveness", International Conference on Wireless Communications, Networking and Mobile Computing, pp. 5427 – 5430, 2007.

[5] K.W.T. Leung, D.L. Lee and Wang-Chien Lee, "Personalized Web search with location preferences", IEEE 26th International Conference on Data Engineering, pp. 701 – 712, 2010.

[6] O. Shafiq, R. Alhajj and J. G. Rokne, "Community Aware Personalized Web search", International Conference on Advances in Social Networks Analysis and Mining, pp. 3351 – 355, 2010.

Appendix

Query 1	How merge sort works
Query 2	How to find the inverse of a matrix
Query 3	What is k-mean clustering technique
Query 4	How to insert data in my sql