An Effective Approach for Personalized Web Search based on Community-Cluster Analysis

T.M.Thanthriwatta[1], P.M.Karunaratne[2]

Faculty of Information Technology, University of Moratuwa, Sri Lanka
thilina.thanthriwatta@gmail.com[1] , pmkaru@itfac.mrt.ac.lk[2]

Abstract - **The concept of Personalized Web Search is commonly used for improving the quality of web search results by identifying and facilitating different users' search needs. There are several techniques such as user profiling, content analysis, hyperlink analysis and biased PageRank algorithm that are used to achieve web personalization. User Profiling is one of the widely used techniques for personalizing web search at large scale. But it contains several technical and ethical issues such as privacy violations, inefficient use of computing resources as well. Collaborative web search is also a kind of a relatively new concept which defines the way of optimizing/personalizing search results by using details of group of people and contributing the knowledge of all of them about web search. This paper presents the details of an alternative approach for personalizing web results by using user profiling technique with community cluster analysis of collaborative web search by adapting concept of reusability among web results.**

Keywords- Collaborative Web Search, Community Cluster, Personalize Web Search

I INTRODUCTION

At present, the intelligent web search has become a very common buzz word in the fields of web mining and knowledge management. The scientists are focusing to come up with solutions to integrate some form of knowledge to web searching and improve the search results quality. Despite their popularity, users' interactions with web search engines can be characterized as one size fits all [5]. The main issue of generic web search engine is that it does not have an ability to determine the users' preferences in a separate manner and facilitate different users in a distinct manner. When the same query is submitted by a different user, a typical search engine returns the same result, regardless of who submitted the query [3]. One of the user scenarios is that if there are two persons who are specialized on Zoology and Operating Systems, both of them will get the similar kind of search results for the search term 'Snow Leopard'. But most probably, the person, interested about Zoology does not need to get the web resources of Apple Mac OSX (Snow Leopard). To address this issue, the giant web search engine service providers came up with solutions to personalize the search results based on some researches. There are several standard techniques such as User Profiling and Hyperlink Analysis for personalizing the web results. Among these, User Profiling technique is widely used by the search engine service providers. In that case, the information of users such as age, gender, country and web site navigational history are tracked and stored and those parameters are used to personalize the web search results. Collaborative Web Search is a form of meta-search, relying on the search services of a set of underlying search engines, but manipulating their results in response to the learned preferences of a given community of users [8]. The unique feature of collaborative search is its ability to personalize search results for a community of users, without relying on traditional context analysis or personalization techniques [9].

In this research, an alternative system is proposed by combining important aspects of effective user profile handing and expert knowledge adapting in collaborative web search. Effective user profile handling can be elaborated as maintaining a user profile without any complexities such as simple profile with facts of most issued search queries. By using those details and clustering technologies, it is needed to group the search engine users and reusing the search results among them using some kind of enhanced algorithm. For filtering the search results further, adapting the users' knowledge about search results/links (users' feedbacks) is very important as well.

II PROBLEM OVERVIEW

Even though User Profiling is the common technique which is used for Personalized Web Search, there are several issues that are related with standard user profiling technique in Personalized Web Search. Because of those issues, most of the leading search engine service providers are unable to embed the personalization aspect for their search results in a large scale. One of the key issues is that the wastage of the computing resources likes memory, network bandwidth and storage. In a general search engine, search engine service provider maintains profiles for each user with large number of their preferences details and performs calculations in each search query to accomplish personalization on search results. Another problem of web personalization is that even though users can usually categorize to a specific domain, there may be special occurrence in that users try to access another domain. For an example, if a computer scientist who is much of 'Snow Leopard' OS may want to get information about the animal snow leopard. This dynamic behavior of users' needs cannot be fully understood by a search engine due to instability. Another scenario is that in some occasions

users may search on Internet not for their needs but others. It is very difficult to address this kind of scenario by using personalizing web results.

One of the key issues of User Profiling in Personalized Web Search is the violation of users' privacy by making serious ethical dilemma and security vulnerability. In User Profiling, most of the important facts of users such as biological, geographical and web navigation patterns are monitored and stored. But some people argue that this process is unethical and it violates the privacy of users. In addition to that, it exposes a vulnerability point to hackers too. In some cases, the search engine service providers have to maintain much complex user profiles with large amount of details about users' preferences that are regularly updated. To avoid that, it is needed to have tighter connection to search algorithm with less information collected/less user interaction required [2]. These are the main issues related with personalizing web search results. In this research, an alternative for personalizing web results will be proposed by adapting Collaborative Web Search concept and the issue of production of less relevant search results will be addressed by reusing web results in intra community clusters in an effective manner.

III RELATED WORKS

There are several key researches and developments which are categorized under Personalized Web Search subject domain. Out of those large set of research, there are several mechanisms that are focused on the effectiveness of information retrieval by user profiling and collaborative searching. In ontological User Profiling for Personalized Web Search, ontology is used to identity topics that might be of interest to a user and it leads to conclusion that ontology is defined as a hierarchy of topics where the topics are utilized for the classification and categorization the web pages [4]. In this research, first they constructed a global dictionary of key words that were extracted from training sample of documents and those key words. After that, the system would assign a weightage value for the terms based on term frequency and inverse document frequency [7]. All the search terms are categorized as a branch of ontological user profile and if a user has an interest on a term which is in a branch, the interest scores of the search terms in upper that branch are incremented as well. One of the main drawbacks in this approach is the necessity of calculating rank score on every occasion of issuing search query by users and utilizing more computing resources and efforts.

SearchTeam.com is one of commonly used real time collaborative search engine which was developed by Zakta in July 2011 [10]. It allows to users to search web contents together as a team. The system will define a SearchSpace which is allowed to users to search web, save and edit results and put into it. Within the SearchSpace, the results are cluttered into folders and other users/collaborators can comment on those search results, add results, like and post links. iBoogie is personalized web search engine which was developed by a company name CyberTavern by embedding their former clustering engine named as Clusterizer [11]. It combines the Meta search and document clustering techniques together to facilitate users. In here Clusterizer puts the web document with similar or related concepts to a group and label it based on the common content of documents. If a user issues a search query, the system will generate a hierarchy of concepts/labels/category which is related with search term. The categories are well distributed and ordered, because of that user doesn't need to go through a long list of search results which are generated by most of search engine. Instead of that, a hierarchy of related contents of search term will be provided with small number of web documents with less complexity.

IV THE PROPOSED SYSTEM

The proposed system has mainly front end (web browser) which is enabling user login and management and back end (search engine component) with algorithms and logics related to search result production. The system overview is presented Fig 1.

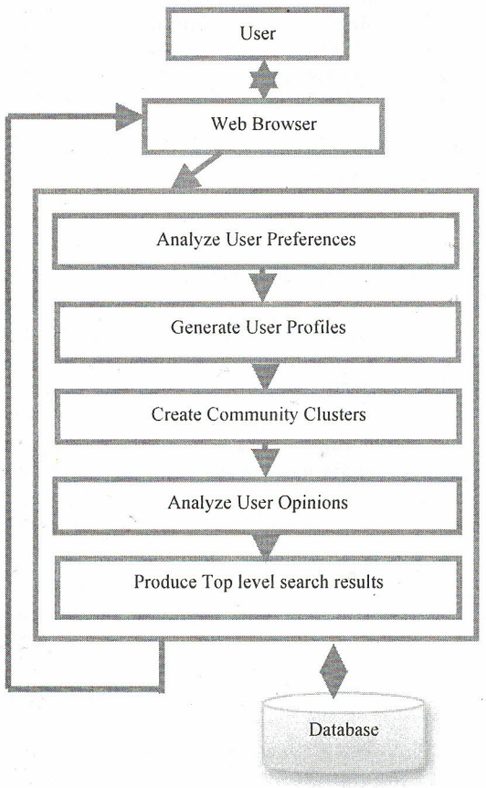

Fig 1: Overall system architecture

A Analyze users' preferences

One of the key issues in Personalized Web Search concept is that handling user profiles with full of details of users' preferences. The size of users' preferences details lead to several technical and ethical problems such as calculations cost and privacy violation. Because of that, the proposed system uses simple common factor, most issued search queries to get an idea of web navigational history of users. By that, the unnecessary complexities of user profiles are avoided.

B Generating user profiles

For generating user profiles based on search history, it is needed to assign numeric values for each most searched query terms by user. To accomplish that task, the numeric value of a specific query term out of all search terms issued by a user in a given period of time is considered as a convenient numeric parameter. If the number of all search terms issued by a specific user (X) in a certain time period is N and number of occurrence of specific search query(S) (one of most frequently used search terms) is M, then the value of interest of that search query(v) is M/N which is resulting from equation (1).

$$v(s,x) = \frac{M(s,x)}{N} \tag{1}$$

The number of most searched query terms which are used to generate user profiles should be identical to each user. In that case, every user has a profile with frequencies of same number of most searched queries.

C Creating community clusters

For creating community clusters, it is needed to apply standard k-means algorithm with numerical data of value of interest in user profiles. As the data preprocessing step, the dimensions of data are identified by the distinct categories of most issued search terms in each user profiles. Then for obtaining a data set with identical number of dimensions, each category such as search terms like 'Java', 'Panther' etc is assigned to each user profile and users might get the dimensions which are not in their user profiles previously. After that if there are categories with NULL values, means that the assigned most search terms which are not issued by particular user previously, it is needed to assign zero for them and construct a numerical data set with multi dimensions (filtered data set) which k-means can be applied. After doing the data preprocessing, filtered dataset is applied to k-means algorithm and clustered based on value of interest.

In K-means algorithms, Euclidean distance is used to identify the similarities of data points in a set by equation (2). By adapting that scenario, it will represent the similarities of users' profiles by comparing the features (data points) of them. In this equation, d is denoted as Euclidean distance, X_i and Y_i are the feature values of X and Y data points.

$$d = \sqrt{\sum_{i=1}^{k} (X_i - Y_i)^2} \tag{2}$$

Then by analyzing the similarities (Euclidean distance values), the users who have similar kind of interest in web searching will be categorized as same cluster. Thus K-means algorithm will create community clusters by large set of users.

D Optimizing community clusters

In k-means algorithm, it is not guaranteed the optimum results for particular number of clusters, because the amount of clusters is defined by user of algorithm. It leads to a scenario that user chooses the amount of clusters in poor manner. To avoid that Rule of Thumb is used for reducing the set of candidate clusters amounts. If n is number of data points (users), then k is defined optimum number of clusters for given dataset approximately by equation (3),

$$k \cong \sqrt{\frac{n}{2}} \tag{3}$$

By applying Rule of Thumb, the approximate optimum number of clusters can be found out. This is very important because it may reduce the workload of large scale dataset. After that, a range of candidate amount of clusters are identified which is around of Rule of Thumb value. In that case, k-means is only needed to apply to that candidate list of amount of clusters which is initially reduced. After applying k-means, then it is needed to calculate silhouette coefficient of each cluster and find out the highly optimum number of clusters among set of candidate amounts of clusters.

Silhouette statistics was introduced by Peter J. Rousseeuw in his paper "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis" in 1987 [1].This technique provides a mathematical representation of how the datum lies within its cluster. In this statistics model, it is needed to find the silhouette coefficient and get the average of coefficient values of data of a clusters that describes how tightly all data grouped in that dataset. Assuming that given data set have been clustered by K-means into k number of clusters, find the average dissimilarity of each datum i with all other data within its own cluster (cohesion) and denoted it as a (i). Then it is needed to calculate the average dissimilarity of

with data of other single cluster not its own one. After repeating this calculation for each cluster but own cluster, get the lowest average dissimilarity (separation) and denote it as b(i). The cluster that is associated with lowest average dissimilarity is called as 'neigh outing cluster' that is next best cluster where the datum i fits. After that it is needed to define the silhouette statistics for datum which is given below by equation (4).

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \tag{4}$$

This definition can be represented in a much simplified way as below.

$$s(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)} & ,if\ a(i) < b(i) \\ 0 & ,if\ a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1 & ,if\ a(i) < b(i) \end{cases} \tag{5}$$

According to the above definition (5), the range of s(i) is defined between -1 and 1. If s(i) is closed to -1, it implies that datum is not well matched/grouped. If s(i) is zero, it says that datum is on the margin of two natural clusters and if s(i) is closed to 1, it means that the datum is well matched with its own cluster. After getting silhouette for all data, it is needed to calculate average s(i) of entire dataset or silhouette coefficient to measure well-clustered of entire data set. By that, the natural number of clusters within dataset can be determined without having too many or too few clusters. So the well suited number of clusters is used as community clusters which are contained the profile details of each user who has approximately similar search interest.

E Produce the top level search results

Then the weightage of each URL (Uniform Resource Locator) should be calculated in the each set of URLs which are grouped by category and cluster. For the normalization purpose, first it is needed to assign 1 for all the records of attribute 'weightage'. After that it is needed to normalize the value of 'interest' field between -1 and 1 and define the rank of each URL. For an example, if there are eight URLs in cluster 'c1' and category 'a', then assume that URL 'x' is in cluster 'c1' an category 'a'. Then the users' interests that is calculated by subtracting plus votes from minus which are given by users as feedbacks. After that all the URLs of specific cluster and category can be normalized based on interest values by using below mentioned normalization formula.

$$N(x) = 2 \times \left(\frac{\text{X} - \text{Min}}{\text{Max} - \text{Min}}\right) - 1 \tag{6}$$

In equation (6), interest of URL (x), minimum interest value of URLs of specific category by a particular community cluster and maximum interest value of URLs of specific category by a particular community cluster and normalized value of web document are denoted as X, Min, Max and N(X) respectively. So it will give a search result set which is most updated according to the feedbacks of users in own community cluster by ranking the URLs by descending order of ranking value which is calculated by adding normalized value of each URL to one. Since the normalization process, there is no need to calculate all ranking scores from the beginning or do the clustering process but simply reusing the most updated results set when a users in the same community cluster issues the same query for web searching. If the search term is never used by the users in own cluster, it will check whether it has been used by users in other community clusters. If it is used, then produce the most updated search results set used by that other community cluster for this specific search term. By adapting the reusability concept, it will save more computing resources and efforts which are utilized for clustering and calculating ranking processes for each and every user's search requests.

V EVALUATION

The evaluation of search quality is also considered as highly researchable area in information retrieval domain. To identify the quality of search results, average precision calculation is used by comparing the proposed system with existing related applications. Average precision is a kind of measure which gives an overview of relevancy of ranked search results according to users' needs. Because the proposed system already has a function component to get the users' feedbacks/opinions about web sites, it is relatively easy to identify average precision based on their feedbacks. The average precision for a single topic is the mean of the precision obtained after each relevant document is retrieved [6]. According to equation (7), k defines as the rank of particular web site among retrieved web sites while n is the number of all retrieved web sites. rel (k) is considered as an indicator function which is 1 if rank k is relevant document or zero other way. P (k) defines as the precision of cut-off k in the list which is the result of dividing number of all relevant documents by 1st to kth document by k.

$$\text{AveP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}} \tag{7}$$

MAP (Mean Average precision) is calculated by averaging the average precisions of set of queries (Q) as equation (8).

$$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q} \tag{8}$$

In this experiment, there are three personalized and collaborative search engines such as 'SearchTeam', 'iBoogie' and 'PWS_PLUS' (proposed system itself) used.

After producing search results sets of 20 search queries by 10 and analyzing users' feedback (relevant/irrelevant), the mean average precisions are calculated. The output of the experiment is shown below in table 1.

TABLE 1: Result set of evaluation process

Users	Search engines		
	PWS_PLUS	iBoogie	SearchTeam
01	0.7844	0.453	0.756
02	0.8976	0.875	0.654
03	0.789	0.654	0.7
04	0.79	0.612	0.234
05	0.657	0.421	0.278
06	0.435	0.675	0.412
07	0.76	0.34	0.45
08	0.54	0.12	0.34
09	0.41	0.5763	0.3123
10	0.7432	0.589	0.421

By analyzing the statistics of the output of evaluation, the quality of the search results set is relatively in a higher position than other alternatives. Throughout whole evaluation process, average precision of PWS_PLUS's search content marked the top position with 80%. Fig 2 shows the graphical representation of the result set of relevancy evaluation of in proposed system and other alternatives.

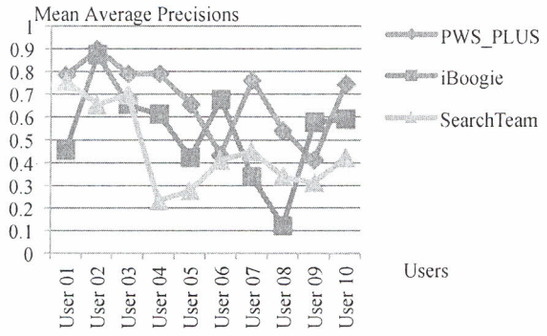

Fig 2: Evaluation of ranked results quality

As a summary, average of mean average precisions of proposed system is calculated as 0.6920 while the other systems have the corresponding value of 0.5306 and 0.4558. The statistics provides a measurement of the quality of search results of proposed system in relative to the other alternatives.

VI CONCLUSION

Under the theoretical domain of web search engine architecture, there are several important factors that can be concerned about the implementation of an alternative approach for traditional personalized web search by minimizing its in-built drawbacks. With the rapid development in computer science and application development, users are always expecting the user friendliness of a system to a great extent. Because of that, people need to come out of the traditional generic web search paradigm and experience the aliveness of search results by obtaining search results that are nearly matched with personal preferences. Even though, existing systems fulfill users' needs to a large extent, the issue is always with the developers to implement the system with minimizing searching time, effort and increase the effectiveness of the system. To address this issue, this application has been developed through utilizing the benefit of reusing search results and cluster technologies.

There are several limitations can be identified as the factors of defining the boundaries of this proposed system in technical and ethical aspects. Inability to do the real time clustering due to lack of infrastructure such as speedy high end machines. Because of that, the clustering should be done once a given period of time and regulated clusters dynamically. Lack of storage area is also another limitation that is coming up with the growth of amount of users in search engine exponentially. Rapid changes of users' preferences are needed to be concerned in this type of a system. This may lead to some errors in this system because the community cluster generating is totally depends on user profiles that consisted with users' preferences details. To minimize this, it is needed to perform the clustering process regularly without having long time duration. Inability to identify the users' need correctly, because there are some occasions where users search for others' needs not themselves. There is a high probability to fail in this type of scenario due to the dynamic nature of human thinking and behavior. The zero violation of privacy cannot be obtained this kind of system because it is related with user profiling. But with reducing complexities of user profiles, the privacy violations of users are reduced as much as possible. So making user profile generation is much more simplistic as this proposed system will be a solution for reducing privacy violations for query issuers.

ACKNOWLEDGMENT

The authors would like to thank all staff members of faculty of Information Technology, University of Moratuwa who have given support towards success of this research.

REFERENCES

[1] P.J.Rousseeuw, "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, pp. 53-56, 1987.

[2] P.A.Chiritha, "Current Approaches to Personalize Web Search," in Workshop on the future of Web Search, Barcelona, Spain, 2006.

[3] C.Yu, W.Meng, F.Liu, "Personalized Web Search For Improving Retrieval," Knowledge and Data Engineering, IEEE Transactions on, pp. 28-40, Feb. 2004.

[4] A.Sieg, B.Mobasher, R.Burke, "Ontological User Profiles for Personalized Web Search," 2007.

[5] J.Allan; et. al., "Challenges in information retrieval and language modeling," in ACM SIGIR Forum, vol. 37 (1), New York, NY, USA, 2003, p. 31–47.

[6] "Annual Review of Information Science and Technology, Volume 40," The American Society for Information Science and Technology ISSN: 0066-4200, 2006.

[7] G.Salton, M.J.Macgill, Introduction to Modern Information Retrieval. New York, NY, USA: McGraw- Hill, 1983.

[8] O.Boydell, B.Smyth, C.Gurrin, A.F.Smeaton, "A Study of Selection Noise in Collaborative Web Search," in IJCAI'05 Proceedings of the 19th international joint conference on Artificial intelligence, San Francisco, CA, USA, 2005, pp. 1595-1597.

[9] B.Smyth, E.Balfe, P.Briggs, M.Coyle, J. Freyne, "Collaborative Web Search," in IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, 2003, pp. 1417-1419.

[10] (2011)TheSearchTeamBlog.[Online].Available:http://blog.searchteam.com/2011/01/13/a-tour-of-searchteam-real-time-collaborative-search-engine/. [Accessed: December 24, 2012].

[11] iBoogie. [Online]. Available:http://iboogie.com/Text/about.asp. [Accessed: December 28, 2012].