International Journal of Engineering and Management Research Vol. 1, Issue-1, 2011
ISSN No.: 2250-0758
Pages: 25-29
www.ijemr.net

Re-Ranking Strategies for Ranking High Precision Information Web Search

Gary Finkelstein[1], Rob Van Hille[2]
[1]Research Guide, Department of Information Technology, Cape Town University, South Africa
[2]Sr. Scientist, Department of Information Technology, Cape Town University, South Africa
[1]garyguide@uct.ac.za, [2]robvan@uct.ac.za

ABSTRACT

A web search engine is a tool designed to search for information on the World Wide Web. The search results are usually presented in a list. The information may consist of web pages, images, information and other types of files. Some search engines also mine data available in databases. Unlike Web directories, which are maintained by human editors, search engines operate algorithmically or are a mixture of algorithmic and human input. The Search Engines often return a large volume of results with possibly a few relevant results. The notion of relevance is subjective and depends on the user and context. The Re-ranking of the results reflects the most relevant results for the user and has become very important in current scenario. In this paper, we propose a novel approach for re-ranking of the search results which would be more efficient and effective of a given user as well as the other users. Our approach is to learn the profiles of the users by making use of past browsing histories including queries posed and documents found relevant or irrelevant.

KEYWORDS: Information Retrieval, Reranking, Query, Filtering, Web Search.

I. INTRODUCTION

The tremendous growth of information available on the web made web search engines an indispensable source to find useful information. However, most often the web Search Engines return a large number of results, of which, the results relevant to a user are not often among the top few. This forces the user to scan through a long list of documents and also refine the query multiple times to find the relevant information. The relevance of the results depends on the perception of the user and the context. As the information on the web continues to grow, there is need for the current day web search engines to serve a specific user and personalize the web search to the user by adapting to his interests and needs [3].

Our work is motivated by the recent advances in collaborative research. It is based on the assumption that there exist groups of users with similar interests, requirements, expectations and motivation seeking similar information in similar contexts of the web search. In this scenario, users would benefit by sharing information, experiences and awareness among the group, typically called a community in collaborative filtering literature [3]. Collaborative filtering methods have been popular for recommending news, movies, music, research papers etc. Recommendations are typically computed using the feedback taken from the all the users in the community. It is advantageous if the users in a search system can collaborate in a similar way and share the information. This could save the laborious effort put by a user in finding the web page containing the information of interest to a great extent. Re ranking the results to contain the most relevant documents on the top by adapting to the user's interests is useful and is a well-known problem in the area of information retrieval.

A number of approaches have been proposed reranking the search using by adapting to the user's interests. We aim to improve the relevance of the result by reranking them using collaborative filtering. In doing so, the key parameters we consider are the user, query, document and the query context, which is typically the query category [1][2]. This is because we believe that a document is relevant to a given user in a given context of the query. The user's neighborhood or community is dynamically calculated in the context of the query. For example, two users might have similar interests, likes and dislikes in cooking but their interests might be totally different when it comes to sports. Hence, we calculate the user's community in context of query dynamically using the users' profiles and query category. The rank of a document is calculated using the user's profile as well as profiles of the users in the neighborhood.

II. METHODOLOGY

A critical method of successful information retrieval on the web is to identify which pages are of high quality and relevance to a user's query. There are many aspects of web Information Retrieval (IR) that differentiate it and make it somewhat more challenging than traditional problems exemplified by the TREC competition. Foremost, pages on the web contain links to other pages and by analyzing this web graph structure it is possible to determine a more global notion of page quality [5]. Early successes in this area include the Page Rank algorithm, which globally analyzes the entire web graph and provided the original basis for ranking in the Google search engine, and Kleinberg's HITS algorithm, which analyzes a local neighborhood of the web graph obtaining an initial set of

web pages matching the user's query. Since that time, several other linked-based methods for ranking web pages have been proposed including variants of both Page Rank and HITS, and this remains an active research area in which there is still much fertile research ground to be explored. Besides just looking at the link structure in web pages, it is also possible to exploit the anchor text contained in links as an indication of the content of the web page being pointed to. Especially since anchor text tends to be short, it often gives a concise human generated description of the content of a web page. By harnessing anchor text, It is possible to have index terms for a web page even if the page contains only images. Determining which terms from anchors and surrounding text should be used in indexing a page presents other interesting research venues [7] [9].

A. Dealing with Spam on the Web

One particularly intriguing problem in web IR arises from the attempt by some commercial interests to unduly heighten the ranking of their web pages by engaging in various forms of spamming. One common method of spamming involves placing additional keywords in invisible text on a web page so that the page potentially matches many more user queries, even if the page is really irrelevant to these queries [4]. Such methods can be effective against traditional IR ranking schemes that do not make use of link structure, but have more limited utility in the context of global link analysis.

B. Evaluating Search Results

Even when advances are made in the ranking of search results, proper evaluation of these improvements is a non-trivial task. In contrast to traditional IR evaluation methods using manually classified corpora such as the TREC collections, evaluating the efficacy of web search engines remains an open problem and has been the subject of various workshops. Recent efforts in this area have examined interleaving the results of two different ranking schemes and using statistical tests based on the results users clicked on to determine which ranking scheme is "better". There has also been work along the lines of using decision theoretic analysis as a means for determining the "goodness" of a ranking scheme. Commercial search engines often make use of various manual and statistical evaluation criteria in evaluating their ranking functions [6][9].

III. PRIOR APPROACH

Childovskii et al perform collaborative reranking of results using user and community profiles built from the documents marked as relevant by the user or community respectively. The search process and the ranking of relevant documents are accomplished within the context of a particular user or community point of view. Sugiyama et. al performed personalization by constructing a user term weights matrix analogous to user-item matrix in memory based collaborative filtering algorithms and then applied traditional collaborative filtering predictive algorithms to predict a term weight in each user profile. Liu et. Al used Probabilistic Latent Semantic Analysis (PLSA), a technique which stems from linear algebra. Hust performed query expansion by constructing the query as a linear combination of existing old queries and their corresponding relevant documents. However, the approach does not take the user into account. In the queries submitted and the results previously selected by a community of users are used to influence the results of searches for similar queries. They proposed an approach for re-ranking of search results in the context of digital libraries [2]. Re-ranking of the results is done using the user pro- file and profile of others users in the community as selected by the user. Several other works have made use of past queries mined from the query logs to help the current searcher.

The focus of this work aims at providing customized search results to a user in response to a query by re-ranking them using collaborative filtering. As it can bee seen, in most of the earlier works, at least one of user, query, document or category has not been used. But, we believe that using all of these would enable us to capture the context appropriately. Also, earlier approaches assumed a static community or group of users and used them for personalizing. But, we believe that the community would depend on the context of the query which we use in our work [5][8].

IV. OUR APPROACH

Our approach constitutes two main steps. The first is learning the user profile and second is re-ranking the search results using the user profile. Learning of the user profile is done using machine learning approaches. Among the machine learning techniques, we have investigated the use of this work. It has been applied with great success in various text applications like text classification, web pages classification and others. Recently, they have been used for text retrieval and achieved performance comparable and even better than the traditional approaches. Reranking is done using the user profile and profiles of other users in the neighborhood.

A. Learning User Profile

A user's profile is a representation of his interests. A user could be interested in different categories. Therefore, we consider a user profile as a collection of discrete sub profiles, one for each category that he is interested in. Each sub profile corresponds to the user's interest in a particular category. Each sub profile is learnt from the queries that belong to a particular category and the corresponding relevant and irrelevant documents for the queries which is learnt by training the text of the queries and the content of the relevant and irrelevant documents. Each sub profile is thus a vector consisting of weights of different terms and is represented. Similarly we

obtain sub profiles for all the categories in which the user has posed at least one query earlier.

B. Re–Ranking of search results

The given query is submitted to a search engine and all the results returned by the search engine are collected. Then, Re-ranking of the search results is done in 2 steps.

- In the first step, the dynamic neighborhood of the user for this query is identified.
- In the second step, rank of each document is computed using the user profile and the profile of all the users in the dynamic neighborhood and the results are sorted and presented in decreasing order of computed rank.

B.1. First Step

Here the user is computed with respect to the query category. In this work, we assume that the category of the query is given.

The user's dynamic neighborhood is computed as follows. At first, all the users having sub profiles in the query category are retrieved. For each of these users having sub profiles in the category, we compute his similarity with the active user. This is done using a simple cosine of the sub profiles of the two users in the query category.

The function **f** denotes the similarity between a user **u** whose sub profile in this category (c_j) is $W_{u,cj}$ and the active user **a** whose sub profile in this category is $W_{a,cj}$

$$f(a, u, c_j) = W_{u,cj} \cdot W_{a,cj}$$

Where · denotes the vector dot product. Then we sorted down the users based on the f value and picked the top K users.

If the sub profile of the user does not exist, then all the users having sub profiles in the given query category consist of the neighborhood.

B.2. Calculating rank of a document

The rank of a document is computed as a linear combination of the rank for a document computed with respect to the active user and the community rank for the document. The rank for a document with respect to the active user is computed as the cosine similarity between the document and sub profile of the active user in the query category. The community rank of a document is the average of the rank computed with respect to all the users in the computed neighborhood weighted by the similarity between the given user and the user in the neighborhood computed using **f** described above.

The rank of a document d for the query q with respect to the user a is calculated as

$$R_{a,d,q} = \alpha (W_{a,cj} \cdot D\, W_{a,cj} \cdot Q) + \beta \sum_{\text{top K Users}} f(q)\, (W_{u,cj} \cdot D\, W_{u,cj} \cdot Q)$$

Where **D** is vector representation of the content of the document **d** and $W_{a,cj}$ and $W_{u,cj}$ are the sub profiles of the active user **a** and a user **u** in the computed neighborhood respectively. This kind of weighted combination helps us perform ranking of the document using the feedback given by the community users. The parameters α, β can be adjusted in order to reflective the relative weights of importance given to information from user and community.

V. DATA COLLECTION

One of the common and important problems in personalized search and related research is the unavailability of large scale datasets for evaluation of the approaches. The unavailability of common test beds poses a serious problem when one has to compare one or earlier proposed approaches. However when tried to download the clicked urls from the data, almost 50% of the clicked urls are unavailable now. Also, since the data was collected in a short time period, it is difficult to observe the behavior and interests of the user which is important in our paper. Also, it is difficult to observe repetition in the needs of the users in such a short time period. In this regard, we use a simulation process to simulate such an environment.

Osmot is an open source search engine which simulates user behavior on the web. The tool uses some randomization processes and simulates the user behavior on the web. The user first poses a query, and then the search engine returns a list of results, the user then looks at the results from top to bottom and possibly clicks one or more results.

Osmot uses synethic queries and documents picked randomly from texts using zipf's law. We slightly modified Osmot to incorporate real queries and real document collections and used the simulation process in Osmot to create simulated user behavior of a large number of users. We used queries that contained categories labeled. We obtained part of ODP data by crawling it and used it as the document collection.

VI. EVALUATION

To test our re ranking approach, we use the simulated test collection. The data consists of 50 users and a total of 31089 queries (600 queries on an average per user). It consists of 4.94 clicked documents and 15.7 seen but not clicked documents per query on an average. The data is divided into 2 sets training set consisting of about 20,000 queries and their corresponding clicked and unclicked results and testing set consisting of around 11,089 queries and their corresponding clicked and unclicked results. User profile learning is done on the training data and the re ranking approach is evaluated on the testing data.

We evaluate the performance of our approach by comparing with the clicked documents in the data. We compare two different methods over the baseline in this section. Firstly the re-ranking done using only the user's profile which we called Approach 1 and our proposed collaborative re ranking Approach which we called Approach 2.
We report and compare the minimum accuracy, precision @ 10, the most widely used metric for evaluating personalized search.

• *Minimum Accuracy*

Minimum Accuracy measures the ability of a search engine to return at least a single relevant result in returned results. The percentage of the queries for which at lest one relevant result is returned is computed. We compare the top 30 results returned by our ranking approaches in calculating the minimum accuracy as shown in Table below:

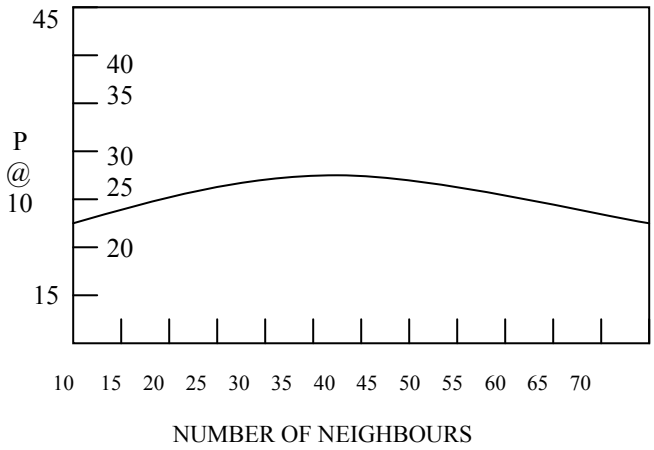
Method	Avg Min Accuracy(%)	P @ 10
Baseline	80.00	0.25
Approach1	91.09	0.327
Approach2	92.62	0.362

• *Precision @10*

We used precision @10 (P@10), the most widely used metrics for evaluating approaches performing re ranking of results. It measures the number of relevant documents found in the top 10 results. The results are shown in above Table averaged over all users and queries.

• *Effect of Neighborhood on Precision @10*

Finally, we discuss an interesting experiment on the effect of the neighborhood size on the performance of our proposed approach i.e. Approach 2. This experiment is done only on Approach 2 because baseline and Approach 1 are unaltered by it. We observed that as the number of users in consideration for re ranking increases, the precision@10 increased. But as the number of users increased beyond 40, the precision dropped. With more number of users in our experiments we expect that the precision might drop even further. We believe this is because the noised added. This is common in approaches using collaborative filtering approaches and careful selection of user neighborhood has to be done.
In summary, our approach of using user neighborhood for re ranking of the results (Approach 2) showed improvement over Approach 1 and baseline in terms of minimum Accuracy, Precision @10. We have seen that the results were dependent on the size of the neighborhood chosen.



VII. CONCLUSION

In this paper, we have proposed an approach for re-ranking the search results reflecting the user's interests. The user's profile is learnt from the query, query category and the clicked documents. Then for re-ranking of the search results for a given query, we first inferred the query's context by mapping the query to one of the pre-defined categories. Then we computed the user's neighborhood in the current context using the query category. We proposed a novel interesting approach for evaluating collaborative web search approaches with minimal manual effort from the users for data collection. Test collections are automatically created with user click through data by simulating user behavior on the web. We evaluated our approach the simulated data. Our evaluation has shown an improvement of performance by using the neighbors profile over using only user's profile. We plan to investigate the use of other alternative functions for computing user-user similarities, which is a major component in our approach.

REFERENCES

[1]. Laura A. Granka, Matthew Feusner, Lori Lorigo, "Eye Monitoring in Online Search,, Passive Eye Monitoring", pp. 283-304, 2008.
[2]. Kazuhiro Seki, Hiroyuki Hattori, Kuniaki Uehara, "Generating Diverse Katakana Variants Based on Phenomic Mapping", Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 793-794, 2008.
[3]. Carrie Grimes, Diane Tang, Daniel Russell, "Query logs alone are not enough", Workshop on Query Log Analysis: Social and Technological Changes, 2007.
[4]. Andrei Z. Broder, Nadav Eiron, Marcus Fontoura, Michael Herscovici, Ronny Lempel, John McPherson, Runping Qi, Eugene J. Shekita, "Indexing Shared Content in Information Retrieval Systems", EDBT, pp. 313-330, 2006.
[5]. E. Balfe and B. Smyth. "An analysis of query similarity in collaborative web search. In Proceedings of

the European Conference on Information Retrieval", pages 330–344. Springer-Verlag, 2005.
[6]. X. Shen., B. Tan., and C. Zhai. "Context-sensitive information retrieval using implicit feedback". In Proceedings of SIGIR 2005.
[7]. B. Smyth, Balfe, O. Boydell, K. Bradley, P. Briggs, M. Coyle, and J. Freyne. "A live-user evaluation of collaborative web search". In Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, 2005.
[8]. H. Lin., G.-R. Xue., H.-J. Zeng., and Y. Yu. "Using probabilistic latent semantic analysis for personalized web search". In Proceedings of APWEB'05, 2005.
[9]. U. Rohini and A. Vamshi. "A collaborative filtering based re-ranking strategy for search in digital libraries". In proceedings of 8th ICADL – 2005.