A Novel Approach for Re-Ranking of Search results using Collaborative Filtering

Rohini U
Language Technologies Research Center
IIIT Hyderabad, India
rohini@research.iiit.ac.in

Vasudeva Varma
Language Technologies Research Center
IIIT Hyderabad, India
vv@iiit.ac.in

Abstract

Search Engines today 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. Re-ranking of the results to reflect the most relevant results to the user using the relevance feedback has received wide attention in information retrieval in recent years. Also, sharing of information among users having similar interests using collaborative filtering techniques has achieved wide success in recommendation systems. In this paper, we propose a novel approach for re-ranking of the search results using collaborative filtering techniques using relevance feedback of a given user as well as the other users. Our approach is to learn the profiles of the users using macine learning techniques making use of past browsing histories including queries posed and documents found relevant or irrelevant. Re-ranking of the results is done using collaborative filtering techniques. First, the context of the query is inferred from the query category. The user's community is determined dynamically in the context of the query by using the user profiles. The rank of a document is calculated using the user's profile as well profiles of the other users in the community.

1 Introduction

The tremendous growth of information available on the web made web search engines an indespensible 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 polysemny and synonymy of the words adds to the problem. The relevance of the results depends on the perception of the user and the context. (*jaguar* as cars to an automobile enthusiast and *jaguar* as cats to a zoologist). But, most search engines today still serve a generic user in a "one size fits all" fashion returning the same set of results without considering the user and his interests. 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.

Our work is motivated by the recent advances in collaborative research. It is based on the assumption that there exists 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. Collaborative filtering methods have been popular for recommending news [19], movies [1] music[3], 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. [15, 8] proposed personalized PageRank [15], Prestschner et al [16] used ontologies, Liu et. al [14] performed personalized web search by mapping a query to a set of categories using a user profile and a general profile learned from the user's search history and a category hierarchy respectively. Shen et. al [22] proposed a decision theoretic framework for implicit user modeling, Radlinski and Joachims [17] learn a ranking function using Support Vector Machines and using it to improving search results.

We aim to improve the relevance of the results by reranking them using collaborative filtering. In doing so, the key parameters we consider are the user, query, document

[1] http://movies.umn.edu

and the query context which is typically the query category. This is because we believe that a document is relevant to a given user in a given context of the query. However the same document might not be relevant for the same user for different query context. Also, the same document might not be relevant for a different user (not having similar interests as the given user) for the same query.(document on jaguar cars for a car enthusiast and a zoologist). Hence, considering all the above parameters helps us to capture the context effectively.

In our approach, we learn the profiles of the user using machine learning techniques. We make use of the past browsing history including queries posed and document found relevant or irrelevant. We use the query category as a means to infer the query context. The user's neighbourhood 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 neighbourhood.

The rest of the paper is organized as follows. Section 2 discusses the related work, Section 3 discusses the proposed re-ranking strategy in detail, Section 4 discusses the data collection, Section 5 presents our evaluation results, Section 6 presents the conclusions on our work also briefing our future work.

2 Related Work

Childovskii et al [2] 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 [26] 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 [13] used Probabilistic Latent Semantic Analysis (PLSA), a technique which stems from linear algebra. Hust [10] performed query expansion by constructing the query as a linear comination of existing old queries and their corresonding relevant documents. However, the approach does not take the user into account. In ([23], [25], [24], [1], [24]) the queries submitted and the results previously selected by a community of users are used to influence the results of searches for similar queries. [21] proposed an approach for re-ranking of search results in the context of digital libraries. Re-ranking of the results is done using the user profile 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. see ([18], [11], [5], [6], [10] etc)

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 (we used category) which we use in our work.

3 Proposed Approach

The proposed 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 SVM[27] in 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 [4] and achieved performance comparable and even better than the traditional approaches [20]. Reranking is done using the user profile and profiles of other users in the neighbourhood.

3.1 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 (also called user-category 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. It is learnt by training an SVM using the text of the queries and the content of the relevant and irrelevant documents. SVM is trained on 2 classes, relevant and irrelevant. The query and relevant documents form the examples for relevant class and the irrelvant documents form the examples for the irrelevant class. The words in documents(after removal of stopwords) attached a numerical identifier are used as the features for training SVM. We used SVM light [12] to perform the learning. After training, we obtain a weight vector which is converted to a standard

vector space model representation. Each sub profile is thus a vector consisting of weights of different terms and is represented by $W_{u,c}$. Similarly we obtain sub profiles for all the categories in which the user has posed atleast one query earlier. In some category, if the user has not posed a query, then the sub profile for that category does not exist for the user.

3.2 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 neighbourhood 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 neighbourhood and the results are sorted and presented in decreasing order of computed rank.

We assume that the sub profile of the user for this category exists. In the rest of the section we refer to the user who has posed the query as the active user.

3.2.1 Computing Dynamic Neighbourhood

The dynamic neighbourhood of 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 neighbourhood is computed as follows. At first, all the users having sub profiles in the query category are retrieved. For each of this 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,c_j} and the active user a whose sub profile in this category is W_{a,c_j}

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

where $\cdot$ 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 neighbourhood.

3.2.2 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 neighbourhood weighted by the similarity between the given user and the user in the neighbourhood computed using f described above.

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

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

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

4 Data Collection

One of the common and important problem 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 more earlier proposed approaches. Alltheweb.com has recently released query log data which has been collection in 2001. 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 (a day), it is difficult to observe the behaviour and interests of the user which is important in our paper. Also, it is difficult to observe repitition in the needs of the users (which our approach exploits) in such a short time period. In this regard, we use an simulation process to simulate such an evironment.

Osmot [2],[17] is an open source search engine which simulates user behaviour on the web. The tool uses some randomization processes and simulate the user behaviour on the web studies in [7]. The user first poses a query, 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. Due to space constraints, the details of the simulation process in Osmot are skipped here. They can be found in [17].

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

[2] http://www.cs.cornell.edu/ filip/osmot/

Method	Avg. Min Accuracy(%)	P@10
baseline	80.00	0.25
Approach 1	91.09	0.327
Approach 2	**92.68**	**0.362**

Table 1. Results

simulated user behaviour of a large number of users. We used queries available from KDDCup 2005 data [3] as the query set. The query set also contained categories labelled. We obtained part of ODP data by crawling it and used it as the document collection.

The data collection thus created consists of the tuples user, query, query category, documents clicked, documents seen but not clicked.

5 Evaluation

To test our re ranking approach, we use the simulated test collection described in Section 4. 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. This is a reasonable assumption considering the clicked documents as relevant and it has been used in earlier works (like [22] etc). The baseline is simple TFIDF scoring of terms and cosine similarity based ranking of documents. 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 Appraoch which we called Apprpach 2.

We report and compare the minimum accuracy, precision @ 10, the most widely used metric for evaluating personalized search.

- Minimum Accuracy
  Minimum Accuracy (see [25]) 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 1.

[3] http://kdd05.lac.uic.edu/kddcup.html

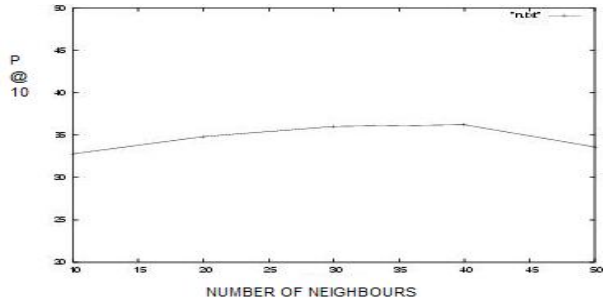

Figure 1. Effect of neighbourhood on P@10

- 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. Tbe results are shown in Table 1 averaged over all users and queries.

- Effect of Neighbourhood on Precision @10
  Finally, we discuss an interesting experiment on the effect of the neighbourhood size on the performance of our proposed approach ie 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(ie top K) 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 neighbourhood has to be done.[9].

In summary, our approach of using user neighbourhood 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 neighbourhood chosen. [9].

6 Conclusions and Future Work

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 neighbourhood in the current

context using the query category. We proposed a novel interesting approch 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 behaviour on the web. We evaluated our approach the simulated data. Our evaluation has shown an improvement of performance by using the neighbours 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 compoenent in our approach. Our final experiment,studying the effect of the size of the neighbourhood on P@10 warns that careful selection of user neighbourhood has to be done. Typically best N neighbours are to be chosen where N might range from application to application. Careful examination, perhaps on a separate validation set is needed to appropriately chose the value of N.

References

[1] E. Balfe and B. Smyth. An analysis of query similarity in collaborative web search. In *In Proceedings of the European Conference on Information Retrieval*, pages 330–344. Springer-Verlag, 2005.

[2] B. Chidlovskii, N. Glance, and A. Grasso. Collaborative reranking of search results. In *Proc. AAAI-2000 Workshop on AI for Web Search.*, 2000.

[3] F. W. Cohen W. Web-collaborative filltering: Recommending music by crawling the web. In *In Proceedings of WWW-2000*, 2000.

[4] S. B. Drucker H and G. D. Relevance feedback using support vector machines. In *Proceedings of the 18th International Conference on Machine Learning*, pages 122–129, 2000.

[5] L. Fitzpatrick and M. Dent. Automatic feedback using past queries: Social searching? In *In Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–313. ACM Press, 1997.

[6] N. S. Glance. Community search assistant. In *In Proceedings of the International Conference on Intelligent User Interfaces*, pages 91–96. ACM Press, 2001.

[7] L. Granka, T. Joachims, and G. Gay. Eyetracking analysis of user behaviour in www search. In *Poster Abstract, Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), 2004*, 2004.

[8] T. H. Haveliwala. Topic-sensitive pangerank. In *Proceedings of the 11th International World Wide Web Conference (WWW2002)*, pages 517–526, 2002.

[9] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. In *SIGIR*, pages 230–237, 1999.

[10] A. Hust. Query expansion methods for collaborative information retrieval. *Inform., Forsch. Entwickl.*, 19(4):224–238, 2005.

[11] J.-Y. Ji-Rong Wen and H.-J. Zhang. Query clustering using user logs. *ACM Transactions on Information Systems (TOIS)*, 20(1):59–81, 2002.

[12] T. Joachims. Making large-scale svm learning practical. *Advances in Kernel Methods - Support Vector Learning*, 1999.

[13] 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.

[14] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM '02), ACM Press*, pages 558–565, 2002.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] A. Pretschner and S. Gauch. Ontology based personalized search. In *ICTAI.*, pages 391–398, 1999.

[17] F. Radlinski and T. Joachims. Evaluating the robustness of learning from implicit feedback. In *ICML Workshop on Learning In Web Search*, 2005.

[18] V. V. Raghavan and H. Sever. On the reuse of past optimal queries. *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 344–350, 1995.

[19] P. Resnick, N. Iacovou, M. Suchak, and J. R. P. Bergstorm. Grouplens: An open architecture for collaborative filtering of netnews. In *Proc. of the ACM 1994 Conference on Computer Supported Cooperative Work (CSCW '94)*, pages 175–186, 1994.

[20] J. J. Rocchio. Relevance feedback in information retrieval, the smart retrieval system. *Experiments in Automatic Document Processing*, pages 313–323, 1971.

[21] U. Rohini and A. Vamshi. A collaborative filtering based re-ranking strategy for search in digital libraries. In *proceedings of 8th ICADL - 2005*, 2005.

[22] X. Shen., B. Tan., and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proceedings of SIGIR 2005*, page 4350, 2005.

[23] 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.

[24] B. Smyth, E. Balfe, P. Briggs, M. Coyle, and J. Freyne. Collaborative web search. In *In Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI-03*, pages 1417–1419. Morgan Kaufmann, 2003.

[25] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. Exploiting query repetition & regularity in an adaptive community-based web search engine. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, pages 383–423, 2004.

[26] K. Sugiyama, K. Hatano, and M. Yoshikawa. Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of WWW 2004*, pages 675 – 684, 2004.

[27] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.