A Modified Collaborative Filtering Approach for Collaborating Community

Pradnya Bhagat and Maruska Mascarenhas

Computer Engineering Department
Goa College of Engineering
Farmagudi, Ponda-Goa
pradnyabhagat91@gmail.com,
maruskha@gec.ac.in

Abstract. Web search is generally treated as a solitary service that operates in isolation servicing the requests of individual searchers. But in real world, searchers often collaborate to achieve their information need in a faster and efficient way. The paper attempts to harness the potential inherent in communities of like-minded searchers overcoming the limitations of conventional personalization methods. The community members can share their search experiences for the benefit of others while still maintaining their anonymity. The community based personalization is achieved by adding the benefits of reliability, efficiency and security to web search.

Keywords: community, personalization, collaborative filtering, collaborative web search, stemming, stopwords, lexical database.

1 Introduction

Web search is generally considered as an isolated activity. Modern search engines employ a strategy known as personalization to accommodate the differences between individuals. Several approaches have been adopted for personalization in the past; but all of these approaches have a serious limitation of treating web search as a solitary interaction between the searcher and the search engine.

In reality, web search has a distinctly collaborative flavor. Many tasks in professional and casual environment can benefit from the ability of jointly searching the web with others. Collaborative Filtering [12] is a methodology of filtering or evaluating items using the opinions of other people. The modified collaborative web search approach presented in this paper is inspired from collaborative filtering and is based on the approach followed in [3]. It tries to collaborate a community of like-minded searchers sharing similar interests to achieve personalization. The main areas focused include: the efficiency of the data structure, the reliability of the results and the security of the system from malicious uses.

M.K. Kundu et al. (eds.), *Advanced Computing, Networking and Informatics - Volume 2*,
Smart Innovation, Systems and Technologies 28,
DOI: 10.1007/978-3-319-07350-7_67, © Springer International Publishing Switzerland 2014

2 Motivation

There are many scenarios where web search takes the form of a community oriented activity. For example students seeking for information on a weekly assignment, or the employees of a company working on a common project will have similar information needs during the project span. Similarly, searches originating from the search box of a themed website, or people with similar purchase history on an e-commerce web site show the potential for collaboration. A survey conducted by M. R. Morris [9] has revealed that a large proportion of users engage in searches that include collaborative activities. The results of the survey has shown that nearly 53% of the searchers involved in sharing either the process (search terms, sites etc) of the product (useful links, facts found within sites) of web search. The respondents even showed to adapt some form of strategy like brute force, divide conquer strategy, backseat driver approach [9] to achieve their required information need faster.

These scenarios clearly show that web search is astonishingly a collaborative task but yet it is not adequately supported by the existing search engines. Hence, the main motivation behind this paper is to allow to the searchers to collaborate irrespective of the time and place of searching provided they share similar interests.

3 Literature Survey

Personalization is proving to be a vital strategy in the success of any web search engine. Two approaches have been frequently adapted to implement personalization in the past: personalization based on content analysis [1] and personalization based on hyperlink structure [6] of the web. Both of these methods have proved successful to a great extent in delivering relevant results to the searchers, but they are limited by the constraint of treating web search as a solitary activity. They fail to identify the collaboration in which users naturally engage to further refine the quality of search results.

Personalization based on user groups is a methodology that incorporates the preferences of a group of users to accomplish personalized search. An approach that is based on this ideology is knows as Collaborative Filtering.

3.1 Collaborative Filtering

Collaborative Filtering is defined as the process of filtering or evaluating items based on the opinions of other people [12]. The fundamental assumption it holds is that if two people rate on n similar items similarly then and hence will rate or act on other future items similarly. But this approach has some drawbacks including: 1) One-to-one similarity calculation and 2) Privacy Violation.

The drawbacks pose serious limitations when it comes to the use of collaborative filtering in web search personalization where the user base is very large and also the uses prefer to stay anonymous.

To overcome with these problems, a modified collaborative web search approach is proposed in [3] called community based collaborative web search.Community based collaborative Web search is based on the principle of collaborative filtering, but instead of exploiting the graded mapping between users and items, it exploits a similar relationship between queries and result pages. It can work as a meta-search working on an underlying search engine and re-rank the results returned by the underlying search engine based on the learned preferences of the community of users. The approaches adopted in literature for collaborative information retrieval can be distinguished in terms of two dimensions: Time and place. Based on these dimensions, the search can be either co-located or remote, or synchronous or asynchronous. CoSearch [10] is an example of co-located, synchronous approach. SearchTogether [8] is an example of system supporting remote search collaboration (whether synchronous or asynchronous). I-Spy [5] is another search engine that is built on the community based collaborative web search.

4 Modified Collaborative Web Search Approach

The modified collaborative web search approach is based on [3] and it attempts to harness the asynchronous search experiences of a community of like-minded remote searchers to provide improved personalized results. It is based on case-based reasoning [2], an approach which uses previous search experiences of searchers to refine future searches. It is implemented as a meta-search engine working on a background search engine like Google to further refine the results returned by the underlying search engine.

The architecture of the Collaborative Web Search (CWS) is explained in Fig. 1. Whenever a searcher submits a query, the query is sent to Google and also to the collaborative web search meta-search engine. In collaborative web search meta-search engine, the query is first passed through the pre-processing block. The output query from pre-processing block and the results of the underlying Google search form the input to the Hit data structure which keeps a record of the number of hits a page has got for a particular query. The next processing block does all the computations and presents the promoted list of results R_P to the user.

At the same time, a list of normal results returned by Google is also collected. This forms the standard list $R_{S.}$ Both promoted list and standard list and merged together and returned to the user as the final result R_{Final}. Normally the promoted results can be shown on top followed by normal Google search results. Otherwise, the promoted results can be shown in one column and standard results in another column.

4.1 Pre-processing

The query is first pre-processed to achieve efficiency in the search process. The preprocessing step consists of the following phases: 1) stopwords removal [4], 2) stemming [7] and 3) checking for synonyms [11] to avoid duplication of queries in the hit data structure.

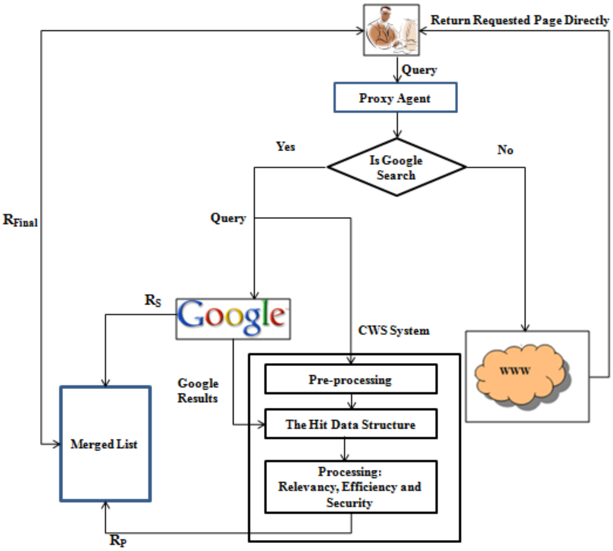

Fig. 1. Architecture of Modified Collaborative Web Search

For example, if “Pictures of Jaguar” is the target query (q_T) and “Jaguar photo” (q_i) is the one present in hit data structure then, without preprocessing, the similarity (*Sim*) computation using Jaccard correlation coefficient [3]between query q_T and q_i, equals 0.25 as given in Equation 1. The system fails to identify two exact similar queries.

$$Sim(q_T, q_i) = \frac{q_T \cap q_i}{q_T \cup q_i} = \frac{(PhotoofJaguar) \cap (JaguarPicture)}{(PhotoofJaguar) \cup (JaguarPicture)} = \frac{1}{4} = 0.25 \qquad (1)$$

The pre-processing steps are as follows. The first step is to remove the stopwords in the tagert query q_T: So the “Pictures of jaguar” will get converted to “Pictures Jaguar”. Next, using Porter Stemmer Algorithm [7] we can stem “Pictures Jaguar” to “Picture Jaguar”. Finally using a lexical database we can convert “Picture” to “Photo” so that the two queries become similar. Now, using Jaccard correlation coefficient, the similarity (*Sim*) equals:

$$Sim(q_T, q_i) = \frac{q_T \cap q_i}{q_T \cup q_i} = \frac{(PhotoJaguar) \cap (JaguarPhoto)}{(PhotoJaguar) \cup (JaguarPhoto)} = \frac{2}{2} = 1$$

4.2 The Modified Data Structure

After pre-processing, the query is given as input to the hit data structure given in Fig. 2, along with the underlying search engine (Google) results. In this specially designed modified data structure, the pages are indexed on queries with the pointer from each query leading to a linked list of pages that are associated with that query. For example in the given figure, the node consisting of query q_1 consists of two pointers. One pointer points to the node containing the next query. The other pointer points to the

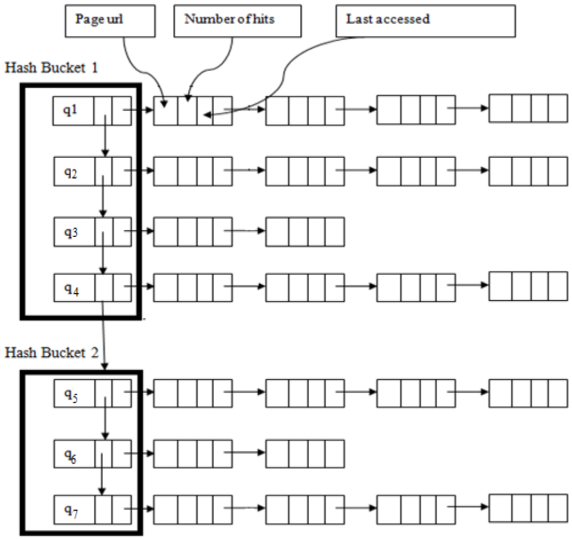

Fig. 2. Data Structure used in Modified Collaborative Web Search

corresponding linked list of pages associated with that query. The nodes in the linked list consist of the following four fields: 1) The URL of the Page, 2) The number of hits for the page, 3) Last Accessed Date and 4) Pointer to the Next Node.

Further, the queries are hashed into several buckets to increase the insert and retrieval efficiency. The pages are ordered in decreasing order whenever the load on the system is reduced, based on the number of hits so that pages having most number of hits are located in beginning and search time will be reduced to a significant extent.

4.3 Achieving Collaboration of Community

The relevance (*Rel*) of a page with some target query is calculated as given in Equation 2 where q_i refers to the target query which is already present in the data structure and p_j is the page whose relevance we are calculating:

$$Rel(p_j, q_i) = \frac{H_j \times \frac{1}{n_j}}{\sum_{\forall j} H_{ij} \times \frac{i}{n_{ij}}} \tag{2}$$

H_j refers to the number of hits that page has got for query and n_j refers to the number of days passed since its last access. This creates a bias towards never pages. Now, the Weighted Relevance (*WRel*) [3] of page p_j to some new target query is a combination of Rel(p_j, q_i) values for all cases q_1, …, q_n that are deemed to be similar to q_T and can be calculated as given in Equation 3:

$$WRel(p_j, q_1, \dots, q_n) = \frac{\sum_{i=1\dots n} Rel(p_j, q_i).Sim(q_T, q_i)}{\sum_{i=1\dots n} Sim(q_T, q_i)} \tag{3}$$

where, $Exists(p_j, q_i) = 1\; if H_{ij} \neq 0$ and $0\; otherwise$

The weighted relevance metric rank orders the search results from the community case base and presents the promotion candidates to users for the target query. Further, since in this approach it is not possible to identify individual users, malicious users may simply click irrelevant pages to increase their hit counts. To deal with this, instead of users, the check is kept on the pages accessed. If any page is getting accessed far more number of times compared to a threshold, a bias towards that page can be detected which can be an activity of malicious users and a check can be kept on that page.

The approach has been implemented on Java platform on test bases. The current implementation is limited to a single community. The system has proved to deliver better performance compared to the underlying search engine and the original approach [3].

5 Results

The dataset is selected from an online bookmarking service delicious.com [13]. Each of the bookmarks can be considered as the result selection and each tag as the query term. To find users having common information need, we tried to discover users having interest in a similar field. These users have the potential to show a significant overlap in searches giving evidence of collaboration.

The data set consisted of 30 users having interest in computer technologies from the delicious dataset. The total size of the dataset consists of nearly 4000 tagged keyword-url pairs with each user on an average having about 135 bookmarks. These 30 uses have proved to show surprising collaboration in their information need when identified properly. With the dataset consisting of about 30 users, if we take any random user, we found that at least 70 percent of the queries typed, are already searched by other community members, while only about 30 percent of the information need varies.

The graph in Fig. 3 shows the evidence of collaboration that can be harnessed to deliver better personalized results to users, hence saving their significant amount of search time. This huge amount of query overlap hints that there can be overlap in the solution need also. That is, given that the query typed is the same, the links selected also can be same. To study this behavior, we used our modified collaborative web search system. So, next we use these about 60 percent of the repeated queries only and try to find out how much percent of the solution overlap we can find. That is, given the query of the new user which query was already placed by the other users we need to find out how many percent of the times even the same link was presented for the same query.

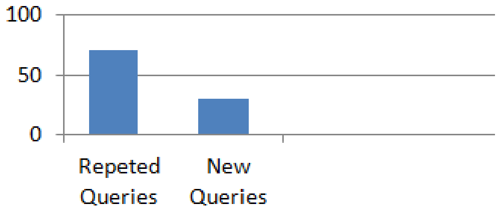

Fig. 3. Repetition in queries found

The CWS system rank orders the results in decreasing order of the weighted relevance. The top ten results are presented to the searcher as the promoted candidates. A search session is marked as successful if at least one result was selected by the users from the top 10 promoted results. The first case taken to find the results is, a user is taken as the test user and its queries are not included in the dataset consisting of result selections. So the promotions that user will be getting will be solely based on the promotions presented by other community member to which that user belongs. Finally, the list which is clicked by the test user for that query is checked if it is present in the top 10 promoted results. If it is present, the search session is considered to be a success.

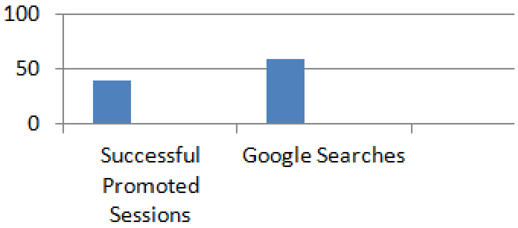

Fig. 4. Successful v/s Unsuccessful sessions when the queries of test user are excluded from the dataset

As can be seen from Fig. 4, about 40% percent of the search sessions were successful in promoting the clicked result of the test user in the top 10 results. Sometimes the promotions might come from the user himself. It refers to searching something which we have already searched before. If these are considered in the dataset, the success rate goes to noticeably higher showing that more than 50% of the search sessions were successful (Fig.5).

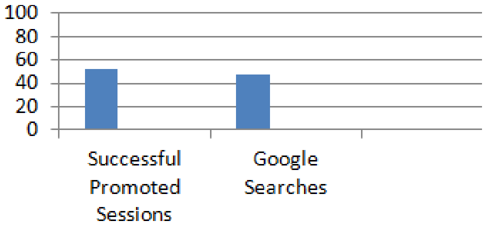

Fig. 5. Successful v/s Unsuccessful sessions when the queries of test user are included in the dataset

6 Conclusion

The motivating insight on this research is that there are important features missing from mainstream search engines like Google. These search engines offer no solution for sharing of the search results between users despite of the fact that there is tremendous potential that can be explored to further refine the quality of search results returned. The system of collaborative web search approach inspired from collaborative filtering allows members of a community of like-minded searchers to share their search experiences for the benefit of other community members. The members of the community can asynchronously collaborate irrespective of the distance between them to improve the search experience. The approach is proved to deliver better performance with respect to precision and recall in comparison to the other search engines.

References

1. Pretschner, A., Gauch, S.: Ontology based personalized search. In: Proceedings of 11th IEEE International Conferenceon Tools with Artificial Intelligence, pp. 391–398 (1999)
2. Aamodth, A., Plaza, E.: Case-based reasoning: foundational issues, methodological variations and system approaches. AI Communication 7(1), 39–59 (1994)
3. Smyth, B., Coyle, M., Briggs, P.: The altrustic seacher. In: Proceedings of 12th IEEE International Conference on Computational Science and Engineering (2009)
4. Buckley, C., Salton, G.: Stop Word List. SMARTInformation Retrieval System, Cornell University
5. Freyne, J., Smyth, B.: Cooperating search communities. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) AH 2006. LNCS, vol. 4018, pp. 101–110. Springer, Heidelberg (2006)
6. Wen, J.-R., Dou, Z., Song, R.: Personalized web search. Encyclopedia of Database Systems, pp. 2099–2103 (2009)
7. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
8. Morris, M.R., Horwitz, E.: Searchtogether: an interface for collaborative web search. In: Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology, UIST 2007 (2007)
9. Morris, M.R.: A survey of collaborative web search practices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1657–1660 (2008)
10. Amershi, S., Morris, M.R.: Cosearch: a system for co-located collaborative web search. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1647–1656 (2008)
11. Peredson, T., Patwardhan, S., Michelizzi, J.: WordNet:Similarity - Measuring the Relatedness of Concepts. In: American Association for Artificial Intelligence, pp. 38–41 (2004)
12. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence 2009 (2009)
13. Wetzker, R., Zimmermann, C., Bauckhage, C.: Analyzing social bookmarking systems: A delicious cookbook. In: Mining Social Data (MSoDa) Workshop Proceedings, pp. 26–30 (2008)