



Applying Collaborative Filtering for Efficient Document Search

Seikyung Jung*, Juntae Kim**, Jonathan L Herlocker*,

**School of Electrical Engineering and Computer Science
Oregon State University, Corvallis OR, USA*

***Department of Computer Engineering
Donguk University, Seoul, Korea
{jung, herlock}@cs.orst.edu, jkim@dgu.edu*

Abstract

This paper presents the SERF (System for Electronic Recommendation Filtering) which is a collaborative filtering system that recommends context-sensitive, high-quality information sources for document search. Collaborative filtering systems remove the limitation of traditional content-based search by using individual's ratings to evaluate and recommend information sources. SERF uses collaborative filtering algorithms to predict the relevance and quality of each document with respect to each particular user and their specific information need. In our system, users specify their need in the form of a natural language query, and are provided with recommended documents based on ratings by other users with similar questions. Preliminary experiments show that the collaborative filtering recommendations increase the efficiency of the document search process. We also discuss some key challenges of designing a collaborative filtering system for document search.

1. Introduction

We all experience information overload. The quantity of new information available every day quickly becomes too much for a single person to deal with. Internet users often turn to search engines to find information sources (documents) relevant to a particular information need. However, traditional content-based search engines have an inherent weakness. Since they evaluate documents based on the *relevance* by comparing the query text and the document texts, it is hard to identify information sources with good *quality*.

In particular, as in Web search, if the search results contain large number of documents with almost similar keyword frequencies, the searcher must then weed through hundreds of sites to sort out the few useful resources. Because there are few barriers to publishing on the Web, individuals with no particular knowledge of a subject can publish on the Internet just as easily as recognized authorities. While some information on the Web has gone through a peer review process, there is no requirement that information on the Web meets any

quality criteria, or be reviewed by anyone other than the author. Therefore, novice searchers might find 'an' answer, but that answer may not be the appropriate or even accurate answer for their information need.

In this paper, we present a collaborative filtering system that can overcome such weakness by using the evaluations made by human users. The system is called the SERF (System for Electronic Recommendation Filtering). The SERF system recommends context-sensitive, high-quality documents by using collaborative filtering algorithms that predict the relevance and quality of each document with respect to each particular user and their specific information need. In SERF, users specify their need in the form of a natural language query, and are provided with recommended documents based on ratings by other users with similar questions.

2. Collaborative Filtering

Current search engines are based on *content-based information retrieval* approaches - based on automatic analysis of the content of documents and the content of a user's query. Search engines use content analysis to evaluate the relevance of documents with respect to the users' needs. In particular, text search engines compare the keywords in a query to the keywords of documents in a database. These content-based search and retrieval techniques are limited by the capabilities of computers to autonomously evaluate the relevance of content with respect to a user's stated need. Due to the limitations of current technology related to language processing and language understanding, content-based information retrieval systems have difficulties differentiating the quality of documents that have similar keyword frequencies.

Collaborative filtering (CF) systems eliminate this limitation of the content-based systems by placing the task of relevance evaluation in the hands of humans rather than automated software. Collaborative filtering has been a mainstream idea since 1997 when a special issue of Communications of the ACM was published on "Recommender Systems" [5].

Early studies of collaborative filtering have focused on recommending items to individuals in entertainment related domains, such as music [8], movies [3], jokes [2], and books (Amazon.com). CF-based recommender systems have often been used by e-commerce websites for presenting products that the customer is likely to buy [6], [7]. More recently, CF has been applied to the problem of recommending scientific literature in the context of the Research Index system [4], [1]. However, the article recommenders built for Research Index only supported query by example: users specified examples of scientific articles, and the citations of those articles were used to locate other related articles.

CF systems, to date, have focused on recommending items such as movies, music, news articles, and books for individuals. Current CF algorithms are based on the assumption that a user's tastes remain basically consistent over time. This assumption works well within entertainment domains, where an individual's taste in movies or music changes very slowly. However, for a search engine, the user's information needs could be substantially different each time the user initiates a search. Therefore, algorithms commonly used in CF systems cannot be applied directly to search engines.

3. Applying CF to Document Search

The goal of our work is to construct a search system that provides the strengths of collaborative filtering, while being able to support information needs that may change frequently and dramatically for each user.

3.1. The Concept

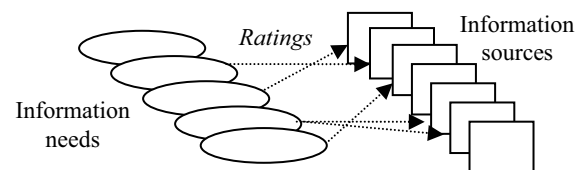
As we discussed previously, we cannot simply apply existing CF algorithms to the problem of document search. Rather we must adapt those algorithms to incorporate some mechanism for specifying the user's immediate information need.

To adapt the idea of CF to a document search, we adjusted the traditional CF data model to integrate the concept of an *information need*. An information need is some representation about what the user wants to find at the time the request is submitted to the system. It can be a set of keywords, a natural language sentence, a sample document, or a pair of (user information, question). If we include information about the user in the representation of the information need, same question can formulate different information need when it is issued by different users. For example, asking about AI by a college student majoring in CS and by an elementary school student are two totally different information needs.

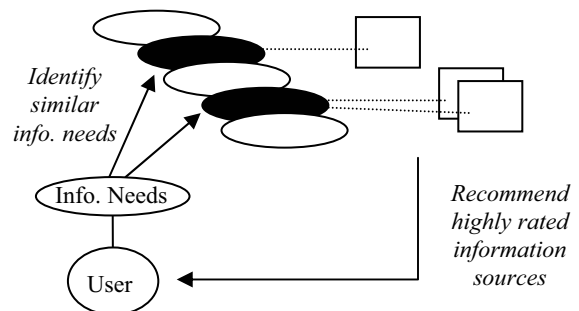
Rather than modeling users rating items, as it is done in traditional CF, we instead model the information needs as the entities that rate documents. Figure 1 illustrates this

concept of our model. Rather than trying to find users with similar preferences, we are instead taking the information need submitted by a user, and trying to match it up with similar past information needs that were submitted by similar users with similar interests.

The problem is at one level similar to traditional CF. We have information needs, (user, query) entities for example, that provide ratings for documents. Some ratings are observed, but most are not. For a new (user, query), we must predict the unobserved ratings of that (user, query). However, there are some substantial differences, particularly in the characteristics of the data.



(a) Ratings of information sources for each information need



(b) Recommending high quality information sources by finding similar information needs

Figure1 The concept of CF-based search

3.2. Identifying Similar Information Needs

The primary difference is that the space of all possible information needs and all possible information sources are not practically enumerable for any search engine with reasonable breadth of information. This is in contrast to traditional CF, where there are always a discrete and manageable number of users and items. Related to this difference is the sparsity of data that will be associated with most queries. In other words, we expect that only rarely will more than one person ask the exact same query.

Therefore, instead of comparing the rating vectors as in traditional CF, we compare information needs directly by defining a distance between them. For the information needs represented as a set of keywords, a cosine similarity can be used as a distance function

Once the system has identified previously requested similar information needs, it displays a set of recommendations. Instead of just displaying the recommended individual documents, the system displays the recommended documents *and* the original information need for which they were rated highly. The user who issued the search can then examine the information needs for which the documents were rated high and determine for themselves if they have any relation to their current information need. This enables us to indirectly use the intellect of the searching user to help identify if past information needs are indeed similar to the user's current need. Furthermore, if the user insufficiently specifies their information need to the system, the user should be able to recognize if past information needs are similar, even if the system does not have enough information to do so.

4. Current Implementation: SERF

We have designed and developed a web-based search system according to the model described above. We call our system the System for Electronic Recommendation Filtering or SERF for short. Figure 2 shows the architecture of SERF. The goal of SERF is to explore the viability of integrating collaborative filtering into a document search engine.

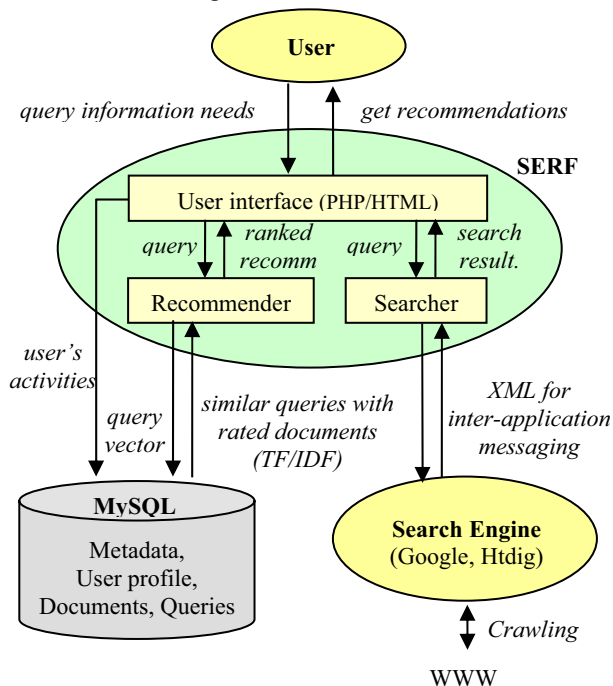


Figure 2 The architecture of SERF

Users initiate a session with SERF by specifying their immediate information need in the form of a text query. Users are asked to enter a question or statement of need and are encouraged to enter their query in human-readable complete sentences, although the system will

function adequately if the user specifies keywords without any context. This is done so that when the queries can be redisplayed to later users, they are more likely to present enough information to be meaningful.

SERF locates past queries that are similar to the current query using the same technique that many full-text search engines use. Each query is treated as a document which is converted into a vector of keywords and their frequencies. Similarity between vectors is computed using a cosine similarity. To speed this process, past queries are indexed by the keywords they contain. Keywords are stemmed before indexing and searching to account for variances in suffixes where there is little variance in meaning (e.g. pick vs. picking vs. picked vs. picker). After past similar queries have been identified, they are ranked based on the cosine similarity (TF/IDF) and listed with the documents rated by each query.

We have designed SERF as a framework on which different topic domains can be easily implemented. SERF has been in development in two domains, first as a search portal to the OSU Library and second as a search portal to Tsunami information on the web (the Tsunami Digital Library).

5. Experiment and Results

We have some promising initial results to report based on controlled testing that we did with a working prototype of the Tsunami information. We investigated the SERF-enhanced Tsunami information in a series of experiments. The objective of these experiments was to investigate the question: "Will the SERF enable more efficient and convenient information search than the traditional key word search?" We tested this question with volunteer student users from our university. We divided 33 students into two experimental groups of 15 and 18, and gave both groups access to our current set of content for the Tsunami information. The first group (Non-CF group) had only a traditional full-text search engine and the second group (CF group) had access to the full-text search plus the SERF recommendations. The two groups were given the same 8 tasks, and after searching they filled out a survey indicating their answer to the tsunami question they had just attempted.

The data collected in our experiments shows that the subjects with SERF recommendations found their answers more efficiently and effectively.

Figure 3 demonstrates the effectiveness of the SERF system. For all the 8 tasks, the percentage of finding correct answer was higher in CF group. Especially, for task 8, 71% of the people in CF group found the correct answer, while only 25% of the people in non-CF group found it. This is because there was no keyword in the sentence describing the task that can be directly used to find the answer. It clearly demonstrates the effect of CF.

The average percentage of finding correct answer was 91% for the CF group and 69% for the non-CF group. An analysis of variance (ANOVA, $p < 0.001$) indicates that the CF group is significantly likely to find a correct answer than those that did not.

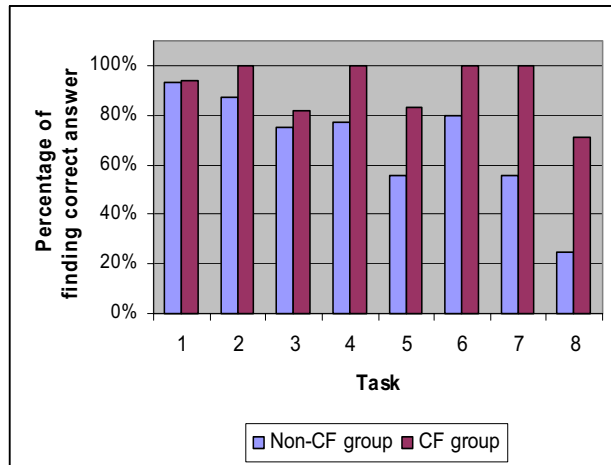


Figure 3 Percentage of finding correct answers

Figure 4 shows the efficiency of search using SERF. The average number of query revisions of the CF group is less than that of the non-CF group for all the 8 tasks. This means that the people in CF group were able to find the answers with fewer rephrasing of search queries. The average number of query revisions for task 8 was 0.44 for the CF group, while it was 1.75 for the non-CF group. Also note that the average number of query revisions was high in task 3. This is because the 3th task does not include the word “tsunami”. The CF group showed less number of revisions in this case too, which means that the SERF approach can help find information even without the necessary content word. An analysis of variance (ANOVA, $p < 0.001$) indicates that the mean number of reformulations of search queries for non-CF group is significantly higher than the mean number of

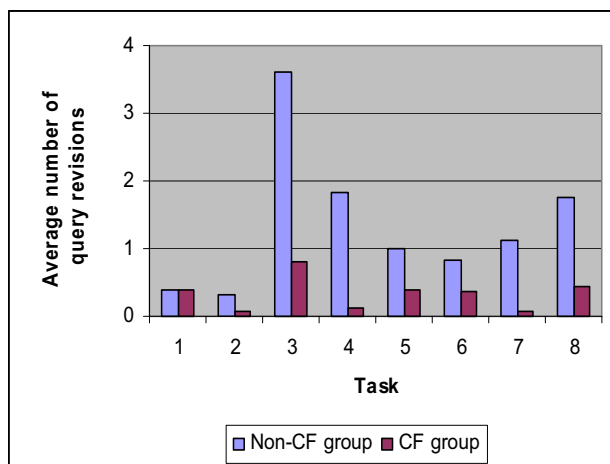


Figure 4 Average number of query revisions

reformulations of search queries for CF group.

Another thing to notice was that subjects’ perception of the quality of aesthetic issues orthogonal to the recommendations was higher among those who received the recommendations than among those that did not, in spite of the fact that visual layouts of the two systems are almost identical. For example, subjects receiving the SERF recommendations felt that the vocabulary used in messages was easier to understand.

6. Conclusion

The SERF application that we have described here applies collaborative filtering to create an efficient and effective document search engine. Since the document recommendations are based on human evaluations, rather than the automated content analysis, it enables search results to incorporate ranking based on hard-to-measure characteristics such as quality, readability, or usability as well as relevancy.

Acknowledgements

We would like to acknowledge Janet Webster, Kevin Harris, Kami Vaniea, Kristina Pence, and Michael Tichenor for their hard work in making the SERF happen.

References

- [1] Cosley, D. Lawrence, S. and Pennock, D.M. 2002. REFEREE: An open framework for practical testing of recommender systems using ResearchIndex. *Proceedings of the 28th VLDB Conference*.
- [2] Goldberg, K., Roeder, T., Guptra, D. and Perkins, C. 2001. Eigentaste: A Constant-Time Collaborative Filtering Algorithm. *Information Retrieval*, 4(2): 133-151.
- [3] Hill, W., Stead, L., Rosenstein, M. and Furnas, G. 1995. Recommending and evaluating choices in a virtual community of use. *Proceedings of SIGCHI '95*, 194-201.
- [4] McNee, S.M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S.K., Rashid, A.M., Konstan, J.A. and Riedl, J. 2002. On the Recommending of Citations for Research Papers. *Proceedings of ACM CSCW '02*.
- [5] Resnick, P., and Varian H. 1997. Recommender Systems. *Communication of the ACM*, 40(3):56-58.
- [6] Schafer, J.B., Konstan, J. and Riedl, J. 1999. Recommender Systems in E-Commerce. *Proceedings of the ACM 1999 Conference on Electronic Commerce*.
- [7] Schafer, J.B., Konstan, J. A. and Riedl, J. 2001. E-Commerce Recommendation Applications. *Journal of Data Mining and Knowledge Discovery*.
- [8] Shardanand, U. and Maes, P. 1995. Social Information filtering: Algorithms for automating “word of mouth”, In *Proceedings of ACM CHI '95 Conference on Human Factors in Computing Systems*.