WPI-CS-TR-99-17 June 1999

Correlation Thresholds for More Accurate Collaborative Filtering

by

Anuja Gokhale and Mark Claypool

Computer Science
Technical Report
Series

WORCESTER POLYTECHNIC INSTITUTE

Computer Science Department
100 Institute Road, Worcester, Massachusetts 01609-2280

Correlation Thresholds for More Accurate Collaborative Filtering

Anuja Gokhale
Mark Claypool

{anuja,claypool}@cs.wpi.edu

Computer Science Department
Worcester Polytechnic Institute

August 4, 1999

Keywords: collaborative filtering, information filtering, information retrieval

Introduction We are in the age of information overload. The explosive growth of computers and networks have provided us with too much information and too few tools to deal with it. People feel overwhelmed by the amount of online information available through the Internet, via the Web, Usenet News or email. There is a clear demand for filters to sort information with respect to users' individual preferences.

We commonly use social recommendations to filter information in everyday life. For example, we may choose a book that a friend has recommended or go see a movie that the critics gave a good rating. Instead of only using recommendations from the people we know or a few public sources, the opinions of users at large can be solicited to let us make better-informed descisions on what information we see and what information we do not. This form of information filtering is called *collaborative filtering*. As a detailed example of how collaborative filtering works, consider two users, John and Mary. John and Mary often see the same movies. In the past, when John has liked a movie, Mary has also liked the movie. So, if John sees a movie that Mary has not yet seen, there is a good chance Mary will like it, too. If we extend this community to many users, we have the potential for extremely accurate predictions.

Many well-known collaborative filtering systems, such as *Firefly* (`www.firefly.com`) and *GroupLens* [KMM+97], base recommendations on a weighted average of the opinions of others, with the weights being influenced by pairwise correlations among users. Other systems like *Fab* [BS97] match a user's profile with content analysis and build correlations between users based on their profiles. Other research has applied machine learning to collaborative filtering by classifying recommendations [BHC98] and extracting features from a group of ratings [BP98]. *Ringo* [SM95] uses correlations between user profiles to perform a similarity assessment of users.

All the above systems, irrespective of whether they use pure collaborative filtering or also perform additional content analysis, rely upon the opinions of all users while computing predictions regardless of the strength of their agreement to the users in question. This may result in predictions that are not as accurate as recommendations based only on the opinions of like-minded users.

We propose *correlation thresholds* that can improve the accuracy of collaborative filtering methods by removing the opinions of users with whom we do not readily agree and concentrating predictions on those with whom we have had strong agreement. Applying correlation thresholds to the example above, assume Mary has seen the same movies as Ann, but has not always liked the same movies. In this case, recommendations for Mary would only be based on John's finding and not on Ann's. In addition, our thresholds can reduce the computation time for generating predictions by decreasing the number of users on which the prediction is based.

Experiments We designed several experiments to evaluate the benefits of correlation thresholds. We built random data sets representing the ratings given to articles by different users. These randomly generated ratings were then used to compute correlation between users. We predicted ratings for each user by first assuming that the particular rating was not present in the data set and then applying a standard prediction algorithm. The accuracy of the prediction was measured by taking the absolute difference between the original rating and the prediction. We varied the correlation threshold and observed the effects it has on average accuracy.

Figure 1 illustrates some preliminary results from our experiments. Point A on the curve shows the the average inaccuracy when there is no correlation threshold applied (correlation threshold is 0.0). Point B on the graph shows the average inaccuracy when the correlation threshold is 0.1. As we can see there is a marked increase in the accuracy. Point C shows a rise in the inaccuracy when the correlation threshold is increased to 0.3, due to a low number of users with a correlation value this high.

In addition, we vary additional experimental parameters such as the number of users, number of articles, sparsity of the matrices and the level of agreement among users, in order to determine the effects of correlation thresholds for a variety of collaborative filtering domains. Lastly, we validate the results from our above experiments by applying our analysis methods to a publicly available database of movie ratings [McJ97] and a collaborative filtering system for an online newspaper that we are developing.

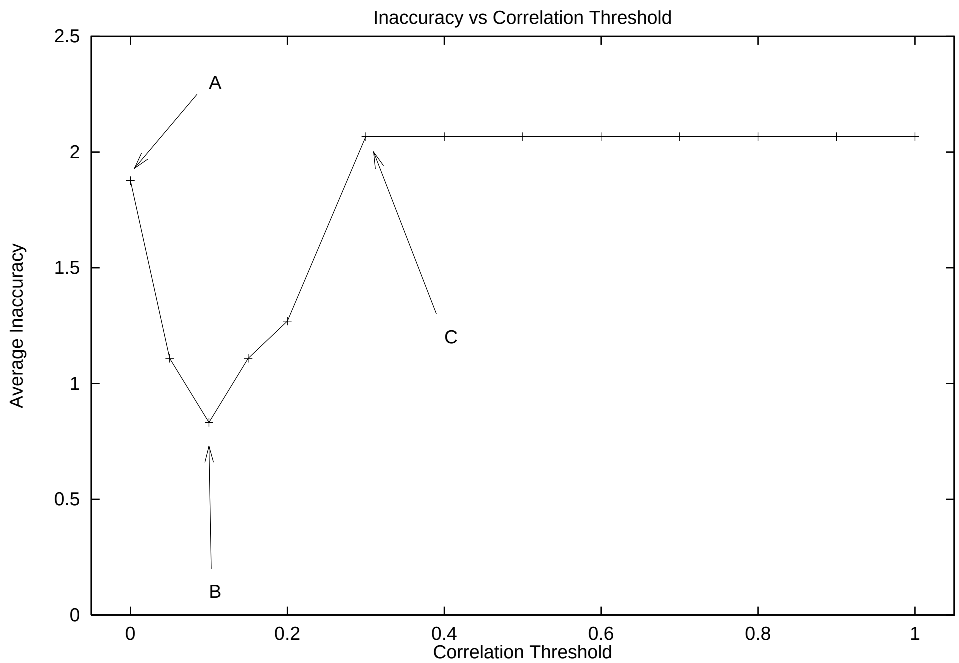

Figure 1: Inaccuracy vs Correlation Threshold

Summary Our preliminary conclusions from the above are that properly chosen correlation thresholds can dramatically improve the accuracy of collaborative filtering systems. However, poorly chosen thresholds can, in fact, hurt the accuracy of the predictions. Our complete paper clearly demonstrates which collaborative filtering domains will benefit from correlation thresholds and provides detailed information on how to choose optimal thresholds.

In summary:

- We present experimental evidence that demonstrates the potential improvement to correlation-based collaborative filtering through the use of correlation thresholds.
- To the best of our knowledge, we are the first to provide a careful analysis of collaborative filtering accuracy as the number of articles and users vary.

References

[BHC98] Chumki Basu, Haym Hirsh, and William Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI)*, Madison, WI, July 1998.

[BP98] Daniel Billsus and Michael Pazzani. Learning collaborative information filters. In *Proceedings of the International Conference on Machine Learning.*, Madison, WI, 1998.

[BS97] Marko Balabanovic and Yoav Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3), March 1997.

[KMM+97] Joseph Konstan, Bradley Miller, David Maltz, Jonathan Herlocker, Lee Gordon, and John Riedl. Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, 40(3):77 – 87, 1997.

[McJ97] P. McJones. Eachmovie collaborative filtering data set. *DEC Systems Research Center*, 1997.

[SM95] U. Shardananad and P. Maes. Algorithms for automating 'word of mouth'. *Conference on Human Factors in Computing Systems (CHI)*, 1995.