Proceedings of the IASTED International Conference
Artificial Intelligence and Soft Computing, August 9-12, 1999
Honolulu, Hawaii, USA

Thresholds for More Accurate Collaborative Filtering

Anuja Gokhale Mark Claypool
{anuja,claypool}@cs.wpi.edu Computer Science Department
Worcester Polytechnic Institute

April 14, 1999

Abstract

Today's explosive growth of information demands personalized filtering. Collaborative filtering is a technique which provides personalized predictions on the "likability" of information based on the opinions of users who think alike. We suggest improvements to current collaborative filtering algorithms to increase the accuracy of the predictions. We implement correlation and history thresholds and conduct experiments that show our improvements increase the average prediction accuracy for all users, but require a per user threshold for optimum performance. These improvements can be used to either provide users with more accurate predictions or to provide them with an indication of the confidence they can place in a prediction.

Keywords: collaborative filtering, information filtering, information retrieval

1 Introduction

Recent years have seen the explosive growth in the amount of information. The World Wide Web started in 1990 and grew to over 10,000 web sites by December 1994 [RIS+94] and to 650,000 web sites by January 1997 [Gra99]. Together these servers exported a minimum of 1.05 million documents. Studies in 1996 show that the World Wide Web is growing rapidly and is filled with transient information with an estimated 50 million pages on the web and with each page being online for only 75 days on average [Sta99]. Estimates show that in May 1993 there were over 2.6 million users of Usenet News at almost 87,000 sites and generated over 26 thousand new articles a day [Tre98]. A year later, the users on Usenet News had risen to almost 10 million. The adoption of email by corporations, governmental and educational institutes and the widespread use of newsgroups further compounds the problem of information overload by making it easy and inexpensive to send information to many recipients.

It is impossible to access even a small portion of today's information. We need automated information filters to prioritize information so that we can make more effective use of our time. Since people have different opinions about the importance/relevance of any information, there is a need for personalized filters.

Collaborative filtering is a general approach to personalized information filtering[1]. Collaborative filtering automates the process of recommending items to a user based upon the opinions of people with similar tastes. In most cases, the filtering system determines which users have similar tastes by using standard formulae for computing statistical correlations. In other systems, users can additionally specify the users whose ratings should be considered while computing the prediction.

Collaborative filtering helps a user build a correlation with every other user depending on how much that particular user agrees with them and then uses a combination of this correlation and their opinion on a particular item to help filter out or prioritize information. Take the example of a movie goer making a decision on the movie she wants to go and see. If there is a way to find all the people who have similar tastes about movies, she could make a much better decision about the movie to see by basing her decision on the opinions of all these people instead of relying on the opinions of a few friends.

Many collaborative filtering techniques use a form of weighted average to determine a prediction for a user [BP98, KMM+97, JAKR97, RIS+94, FD92, DDF+90, SKB+99, BHC98]. These techniques use the correlation (degree of similarity between two users) as the weights. Every user receives a prediction for all items and submits a rating of how well she likes an item after reading it. This feedback given by her is used along with similar feedback from other users to compute a prediction. The feedback given by the user along with the prediction received by her for a particular item is later used to update her correlation with every user.

[1] Although recently published literature on research in filtering techniques have used the term *collaborative filtering* interchangeably with the terms *social filtering* and *recommender systems*, we shall henceforth use the term collaborative filtering to refer to any three of the above.

Current collaborative filtering techniques are not tuned to the similarity between users. The effectiveness of collaborative filtering techniques relies on the confidence on the computation of the "similarity" between users. Further, the more the number of "similar" users, the more effective the prediction will be as having more users can alleviate the inconsistencies/errors that could arise if a user agreed with another on most topics except a few. For example, if user A agreed on most things with user B except that user A liked romantic comedies and B hated them, then it would lead to an inaccurate prediction if A relied only on user B's rating about romantic comedies. We implement thresholds on the correlation between users such that only the opinions of users with a correlation above the threshold are considered.

Current collaborative filtering techniques make predictions without taking the *history* between users into account. A user is given a prediction based on the ratings of users who may not have seen enough messages in common to be able to justify making a strong prediction based on their ratings. For example, if user A and B have seen only one message in common and have agreed, it is not a strong basis to predict that they will agree on the rating given to the second message too. Current techniques also do not favor those people that have a strong history and cannot give more importance to ratings of users who have agreed over a larger number of articles and so are more likely to agree in the future. We implement thresholds on the history between users such that only the opinions of users with a history in common above the threshold are considered.

2 Algorithm Improvements

2.1 Correlation Thresholds

The basis algorithm computes a prediction by using the ratings submitted by all users irrespective of the correlation between the users. This can lead to inaccuracies if the low correlations of a large number of users overshadow the positive effect of the ratings of the possibly small number of users that have a high correlation with the user. Another potential problem with using the ratings of all the users is an increase in the computation time. To solve such problems, we implement thresholds on the correlations with users that are to be considered such that only users having a correlation higher than the threshold with the user in question are considered. We carried out experiments on data extracted from the EachMovie collaborative filtering service [EM97]. The EachMovie service was part of a research project at the Systems Research Center of Digital Equipment Corporation and contains ratings from 72,916 users on 1,628 movies. The service was started in February 1996 and lasted for 18 months. We generated the prediction for every pair of users and movies on a subset of the EachMovie data set.

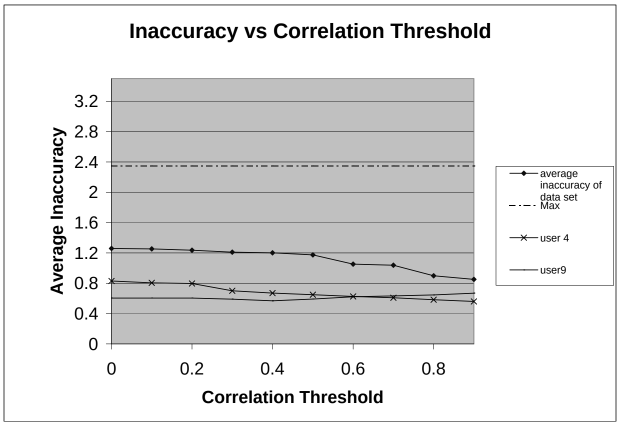

Figure 1: *Plot showing the relationship between the average inaccuracy and the correlation threshold for data from the EachMovie Data. This dataset consists of the ratings of 58 users and one 125 movies. Average inaccuracy is an average of the differences between the rating and the prediction for every pair of users and movies in the data set.*

We then compared the prediction computed as above with the rating (extracted from the EachMovie data) that the user gave for the movie. We repeated this same procedure for every user and movie (for every data element in the matrix data set). We computed the average absolute inaccuracy (average of the absolute differences between the generated rating and it's computed prediction for all pairs of users and articles) for the whole data set. This is the average inaccuracy of the data set for that particular value of the correlation threshold. We repeated the same procedure for different values of the correlation threshold.

Figure 1 illustrates the results from our experiments on a data set consisting of ratings submitted by 58 users for 125 movies. The ratings in the data set are extracted from the EachMovie data. The ratings submitted by users are in the range of one to ten (with valid ratings being 0,2,4,6,8 and 10). Data was extracted such that the extracted data had a good mix of users with varying values of correlations between them. The line *Max* shows the average of the $(rating - mean)$ of all the users in the data set. This is computed by taking the average of the inaccuracies in the prediction returned for all the users using the mean as the prediction for the user. This is the prediction returned for the user if no user in the data set has a correlation higher than the threshold with the user in question. The line labeled *average inaccuracy of data set* shows the average inaccuracy for the data set for the corresponding correlation threshold value. The lines labeled *user1* and *user9* show the average inaccuracy in the predictions for individual users. This is computed by taking the average of the inaccuracies in the predictions for all the articles for each particular user.

We can see from the graph that the mean rating of the data set (*Max*) is always higher than the average inaccuracy of the data set for all values of the threshold. As this value is above the average inaccuracy of the data set even when no thresholds are applied, using collaborative filtering techniques is more accurate than just returning the mean. The *average inaccuracy of data set* drops with increasing values of correlation thresholds. At each correlation threshold, we are considering only those users whom we agree with the most. The application of correlation thresholds leads to increased accuracy of the predictions. Lines *user1* and *user9* show the average inaccuracies for two different users in the data sets. We see that although both users benefit from the application of correlation thresholds, the inaccuracy increases when the threshold is increased beyond a certain value. In fact, there is no correlation threshold that works best for all the users in the data set. The best correlation threshold value for a user depends on the number of users who are above this correlation threshold value. In particular, the inaccuracy increases with increasing correlation thresholds if the number of users who have a correlation above the threshold value drops below a certain number. This number of users varies from 40 to 49 and is not the same for each individual user.

2.2 History Thresholds

The basis algorithm computes a prediction for the user by considering the ratings submitted by all the users of the system for the movie in question. This can lead to inaccuracies in the prediction computed by the system if the number of users with a lower history with the user in question increases or is more than the number of users with a high history with the user in question. For example, 2 movies seen in common and liked by both users does not guarantee or even strongly indicate that the two users would agree on the next movie. The prediction in such cases is uncertain and likely to have inaccuracies.

We tune our basis algorithm such that the weighted average formula for the computation of the prediction is applied only to users who have seen more messages in common than the history threshold. We implement history thresholds such that we consider the correlation and the rating given by each user only for users who have the number of messages seen in common greater than the history threshold imposed by the system. We carry out our experiments on different values of history thresholds. The advantage with this technique is that users with lower history do not unduly affect the prediction and the prediction is based only on those users who show either a proven degree of agreement or disagreement.

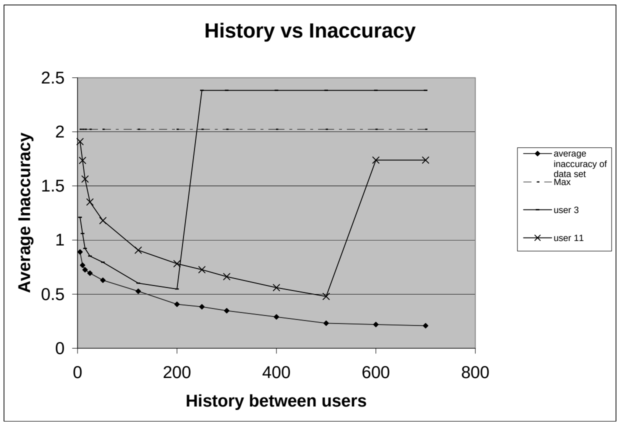

Figure 2: *Average inaccuracy and the history in common between any two users. The number of users is kept constant at 125. Average inaccuracy is an average of the differences between the rating and the prediction for every pair of users and movies in the data set.*

We conducted experiments similar to those on correlation thresholds except that we considered only those users who had a history in common above the threshold with the user in question. The graph shown below illustrates the relationship between the inaccuracy of the prediction and the number of users in the data set extracted from the EachMovie database [EM97].

The average inaccuracy plotted in the graph is the average inaccuracy for the data set for the corresponding number of users in the data set. The average inaccuracy was calculated by taking the average of inaccuracies in all the predictions computed by the system for every pair of users and ratings.

The ratings in the data set are extracted from the EachMovie data. The line labeled *Max* shows the average of the $(rating - mean)$ of the data set. This is the prediction returned for the user if no user in the data set has a history in common higher than threshold with the user in question. The line labeled *average inaccuracy of data set* shows the average inaccuracy for the data set for the corresponding history threshold value. Lines labeled *user 3* and *user 11* show the average inaccuracy in the predictions for individual users. This is computed by taking the average of the inaccuracies in the predictions for all the articles for that particular user. We can see from the graph that the mean rating of the data set (*Max*) is always higher than the average inaccuracy of the data set for all values of the threshold. The line *average inaccuracy of data set* shows the average inaccuracy of the data set for different history thresholds. We can see that the average inaccuracy of the data set drops with increasing values of history thresholds. This is because at each history threshold, we are increasingly considering only those users with whom we have seen the most articles in common. This implies that the application of history thresholds leads to increased accuracy of the predictions.

Lines labeled *user 3* and *user 11* show the average inaccuracies for two different users in the data sets. We see that although both users benefit from the application of history thresholds, the inaccuracy increases when the threshold is increased beyond a certain value. In fact, there is no history threshold that works best for all the users in the data set. The best history threshold value for a user depends on the number of users who are above this history threshold value. In particular, the inaccuracy increases with increasing history thresholds if the number of users who have a history above the threshold value drops below a certain number. The history thresholds, therefore, should be different for different users to be able to compute the best predictions.

We also observed that this best history threshold value depends on the number of users who are above the history threshold value. In particular, the inaccuracy increases with increasing history thresholds if the number of users who have a history in common above the threshold value drops below a certain number. This number of users, however, is not the same for all users and varies between 35 to 50.

We believe that the "strength" of predictions computed by considering users who have a low history in common with the user in question would be low. For example, consider a situation where user A and user B have seen just two articles in common and have agreed on both. There is no justification in assuming that they will think alike on the next article seen by them (i.e. the prediction is inherently weak). We have also discussed earlier that predictions computed by using the ratings of users who have a low history also leads to predictions that are not consistently accurate. The history between users is not high enough to guarantee that the correlation computed between these users is a true indication of their degree of similarity or dissimilarity. We would thus assume that such predictions would have a high standard deviation.

We perform analysis on the data from the above experiments to see the effect of the history on the strength of the prediction. We compute the standard deviation of the inaccuracies in the predictions. We then plot the average inaccuracy of the data set and also plot graphs which are at one standard deviation distance from the average. For normally distributed data approximately 60 percent of the total predictions would lie in the range within one standard deviation of the mean.

The graph shown below illustrates the relationship of the inaccuracy as well as the standard deviation of the inaccuracy of the prediction and the history in common between the users in the data set for data extracted from the EachMovie database.

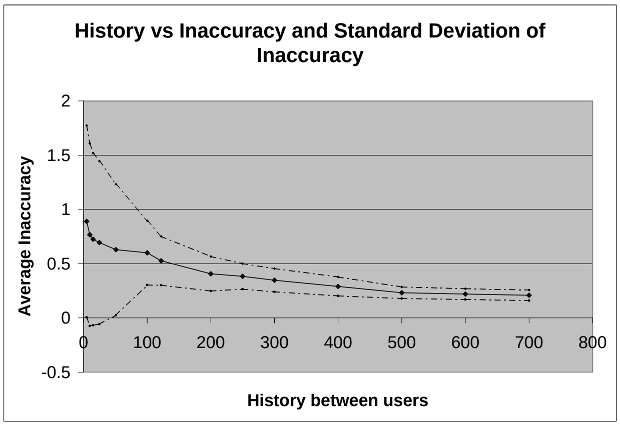

Figure 3: *Plot showing the relationship between the average inaccuracy, the standard deviation of the average inaccuracy and the history in common between any two users in the data set for a data set consisting of ratings extracted from the EachMovie database. Average inaccuracy is an average of the differences between the rating and the prediction for every pair of user and article in the data set.*

We observe in figure 3 that the standard deviation is large for a history in common of 5 articles and then rapidly drops with an increase in the history in common between users. This supports our earlier conclusion that an increase in the history in common between any 2 users will lead to predictions that are not just more accurate but also more consistently accurate. These can also be used to generate a confidence level in the prediction.

3 Conclusions

The use of correlation and history thresholds leads to more accurate predictions. Both correlation and history threshold values are specific to a user. There is no threshold value that works best for all users. In fact, the application of a single threshold value for all users can increase the inaccuracy of the predictions for some users beyond certain threshold values. We also observe that the implementation of history thresholds decreases the standard deviation of error in the prediction and thus leads to predictions that are not just more accurate but also more *consistently* accurate.

The improvements discussed above in subsections 2.1 and 2.2 can be implemented to increase the accuracy of predictions in a collaborative filtering system as well as to decrease computation. Collaborative filtering is inherently computationally expensive, and in really large collaborative filtering systems, this becomes an important consideration. Our standard deviation analysis on history thresholds can also be used to return an indication of the confidence a user can place in the prediction.

In summary, in contributing with this work, we:

- Designed and developed an algorithm that implements correlation thresholds to compute more accurate predictions.
- Designed and developed an algorithm that implements history thresholds to compute more accurate predictions.
- Implemented history thresholds to compute predictions that are more *consistently accurate*.
- Presented a way to measure confidence in predictions.

There are many areas for future work. Correlation and history thresholds do not perform very well if the number of users considered is low. A combination of the above techniques along with user thresholds may be more effective. Correlations between two users, A and B, will not be captured correctly with the Pearson's correlation coefficient if they have a low correlation when B likes an item but have a high correlation whenever B hates the item (they dislike similar items but may not like similar items). Implementation of a non-linear correlation coefficient will be able to reflect such relationships between users more accurately. Collaborative filtering systems cannot give a collaborative filtering prediction for a new article that no users have rated. Such situations can be dealt with if the system also maintains correlations between articles. If two articles are similar, the user is likely to have similar opinions about the articles.

References

[BHC98] Chumki Basu, Haym Hirsh, and William Cohen. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. *American Association for Artificial Intelligence*, pages 714–720, 1998.

[BP98] Daniel Billsus and Michael Pazzani. Learning Collaborative Information Filters. *Machine Learning: Proceedings of the Fifteenth International Conference*, 1998.

[DDF+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, pages 391–407, 1990.

[EM97] EachMovie Website, http://www.EachMovie.com. *web document*, 1997.

[FD92] P.W. Foltz and S.T. Dumais. Personalized Information Delivery: An Analysis of Information Filtering Methods. *Communications of the ACM*, 35(12):51–60, 1992.

[Gra99] Matthew Gray. Web Growth Summary. *http://www.mit.edu:8001/people/mkgray /net/web-growth-summary.html*, 1999.

[JAKR97] Bradley N. Miller Joseph A. Konstan and John Riedl. Experiences with Grouplens: Making Usenet News useful again. *Proceedings of the USENIX 1997 Annual Technical Conference*, pages 219–231, 1997.

[KMM+97] Joseph A. Konstan, Bradley N. Miller, David Maltz, Jonathan L. Herlocker, Lee R. Gordon, and John Riedl. GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, 40(3):77–87, March 1997.

[RIS+94] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. An Open Architecture for Collaborative Filtering of Netnews. pages 175–186. ACM Conference on Computer Supported Co-operative Work, 1994.

[SKB+99] Badrul M. Sarwar, Joseph A. Konstan, Al Borchers, Jon Herlocker, Brad Miller, and John Riedl. Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. *Computer Supported Cooperative Works 1998 Conference*, 1999. Seattle.

[Sta99] Internet Archive. *http://www.archive.org*, 1999.

[Tre98] Win Treese. The Internet Index. *http://www.openmarket.com/intindex/98-01.htm*, (21), Jan 1998.