Survey of Collaborative Filtering Techniques for a Movie Recommendation Engine

Nirmal Krishnan (nkrishn9, nkrishn9@jhu.edu), Emily Brahma (ebrahma1, ebrahma1@jhu.edu)

4 October 2016

1 Abstract

In 2009, Netflix offered a one million dollar prize to an individual that could create a recommendation engine better than its own. Many solutions were attempted, but the most successful algorithms were those that used a technique known as *collaborative filtering*, in which a user's interest in new items are computed based on the recommendations of users with similar interests. Almost all implementations of collaborative filtering use a neighbor-based prediction algorithm similar to k-nearest neighbors (k-NN) (Herlocker et al. 2002). However, there is no consensus on how parameters are weighted, and what kind of similarity metrics are used. In this study, we will be surveying the effects of different parameter weighting techniques on collaborative filtering in the context of the movie recommendation problem.

2 Methods

We will perform collaborative filtering using multiple similarity weighting schemes. The basic process for collaborative filtering is as follows:

1. Pick an active user for which we would like to predict that user's rating for an arbitrary movie.
2. Use a weighting technique to weight all of the users in the data-set for similarity to the active user.
3. Pick a set of users based on closest similarity to active user.
4. Make a prediction based on the subset selected in part 3.

The weighting techniques we plan on surveying include the Pearson correlation coefficient, the Spearman rank correlation coefficient, and the mean-squared difference (used the Ringo recommendation pattern).

3 Resources

The data that will be used for the study is publicly available through an online database called MovieLens. This database contains 100,004 ratings created by 671 users, and covers 9,125 movies. Each user was chosen at random and has rated at least 20 movies, ensuring that the data-set is not too sparse for analysis.
In regards to implementation details, we will be writing our project in python, using num.py for speed and stability. We plan on implementing our machine learning algorithm from scratch, in order to meet the weighting goals.

4 Milestones

- November 18th: Establish data pipeline
- November 23rd: Complete Pearson correlation coefficient
- November 26th: Complete Spearman rank correlation coefficient
- December 2nd: Complete mean-squared difference
- December 7th: Rough draft of final write-up
- December 12th: Finish write-up

4.1 Must achieve

We must achieve an implementation of collaborative filtering that uses all three of the weighting similarity techniques listed above, such that we can make a recommendation on the best one.

4.2 Expected to achieve

We are expected to achieve an implementation of a non-statistical technique for collaborative filtering. This approach, known as cosine similarity, is based on a linear algebra design where users are treated as $|N|$-dimensional vectors and "similarity is measured by the cosine distance between two rating vectors." We would like see how this technique compares with the traditional statistical approaches.

4.3 Would like to achieve

If possible, we would like to survey other weighting similarity techniques that could potentially yield better results. These techniques may include Constrained Pearson correlation and Pearson Correlation with threshold-based damping. If time permits, we would also like to brainstorm and test ideas for our own similarity metric.

5 Final Writeup

In the final writeup, we will be discussing our findings extensively. We plan on including sections that describe our methods, results, and a brief explanation of these results. In regards to specifics, in our methods section, we plan on including details for how we setup our environment, the prediction rules used, and the weighting similarity algorithms. In our results section, we will show the accuracies of each weighting technique, and in our discussion section, we will explain which technique was optimal and why we believe this is the case.

6 Bibliography

- Herlocker, Konstan, and Riedl. An Empirical Analysis of Design Choices in Neighborhood-Based Collaborative Filtering Algorithms. Information Retrieval, 5, 2002.
- Melville, Mooney, and Nagarajan. Content-Boosted Collaborative Filtering for Improved Recommendations. Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-2002), Edmonton, Canada, July 2002.
- Breese, Heckerman, and Kadie. Empirical analysis of predictive algorithms for collaborative filtering. 1998.
- Cooper and Moral. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98). San Francisco.