# CS 6001 Applied Spatial and Temporal Data Analysis - Spring 2017 - Homework 4

Dalton Cole

April 3, 2017

## 1 Introduction

The purpose of this assignment is to become familiar with recommender systems. In the video that Dr.Fu asked us to watch, User-based collaborative filtering and item-based collaborative filtering was discussed. This assignment gives us first hand experience at these two filtering methods. User-based collaborative filtering is done by finding the similarity between different users based on the items, in the video's case, Netflix shows watched. Item-based collaborative filtering is essentially the transpose of user-based. Similarity between items are used instead of between users. In most cases, this provides more reliable data. This is in part because, normally, there are fewer catalog items than there are users. This reduces the dimensionality of the data.

These two methods are variations of the kNN algorithm. The basic formula is:

$$\hat{r}_{ui} = \frac{\sum\limits_{v \in N_i^k(u)} \text{sim}(u,v) \cdot r_{vi}}{\sum\limits_{v \in N_i^k(u)} \text{sim}(u,v)}$$

Three more algorithms are used: SVD, PMF, and NMF. The SVD algorithm was popularized by Simon Funk, as discussed in the KNN tutorial video. It uses a prediction value:

$$\hat{r}_{ui} = \mu + b_u + b_i + q_i^T p_u$$

The following regularized squared error should be minimized using stochastic gradient descent:

$$\sum_{r_{ui} \in R_{train}} (r_{ui} - \hat{r}_{ui})^2 + \lambda \left( b_i^2 + b_u^2 + ||q_i||^2 + ||p_u||^2 \right)$$

PMF is the unbiased version of SVD:

$$\hat{r}_{ui} = q_i^T p_u$$

NMF is very similar to SVD as well. The difference between the two is that the user and item factors are kept positive in NMF.

MAE and RMSE are two matrices that are used to evaluate performance. Mean absolute error (MAE) can be seen in Figure 1. Root-mean-square deviation (RMSD) can be found in Figure 2.

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

Figure 1: Mean Absolute Error Formula

$$\mathbf{RMSD} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}}$$

Figure 2: Root-Mean-Square Deviation Formula

## 2 Experiment

The code used for this segment can be found in *HW4.py*. The data used in this experiment is located in *restaurant_ratings.txt*, which is a text file given by Dr.Fu for this assignment. The goal of this experiment is to see which algorithm performs best given this data set. Three-fold cross validation was used. To keep each fold consistent across each algorithm and across seperate tests, the random seed generator was seeded with the value *0*.

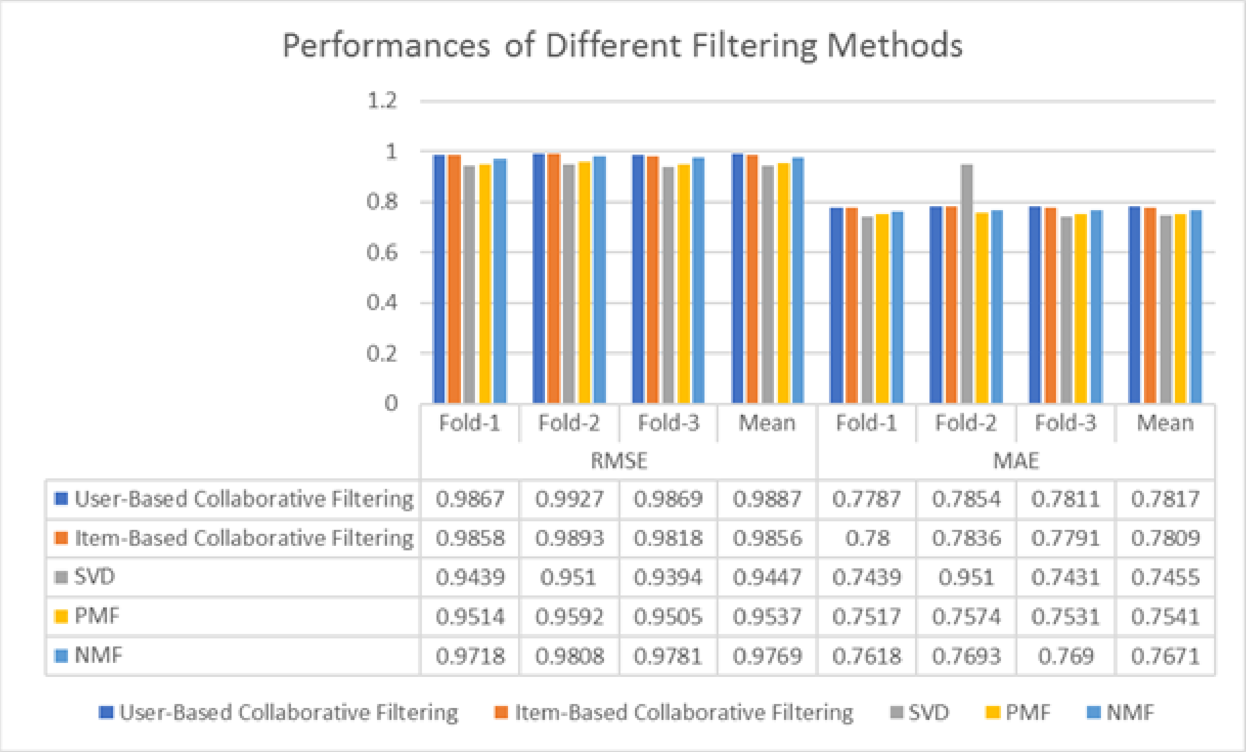
## 3 Results

The results of this experiment can be seen in Figures 3, 4, 5, and 6. As can be seen in Figure 3, SVD out performed the other methods. In this case, item-based collaborative filtering barely out performed user-based collaborative filtering. Figure 4 shows how important a similarity metric can be. Using the MSD similarity metric, item-based out performs user-based collaborative filtering. Using cosine or Pearson however gives user-based the edge.

Figure 5 shows that when k = 25, 27, and 28, RMSE was as its minimal value for user-based collaborative filtering. This implies that looking at the mid 20s to higher 20s neighbors yields the best results for. When k = 43 gave item-based collaborative filtering its best RMSE value. Item-based out performed user-based when $k \geq 30$. When $k < 30$, user-based out performs item-based.

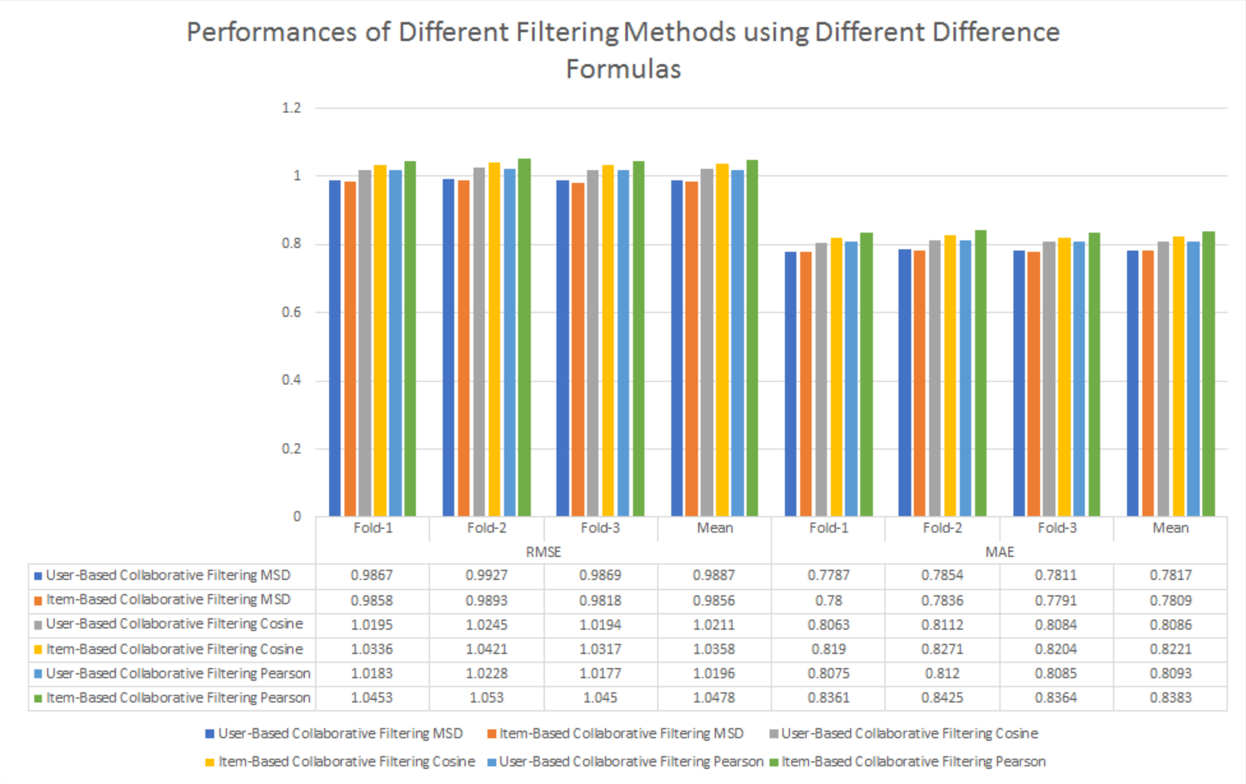
## 4 Conclusion

Different machine learning algorithms produce different results. Different error measuring formulas also produce different results. It is important to keep in mind both of these facts when selecting algorithms. It is important to be consistent with the error measurements. In this report, RMSE was primarily used to compare algorithms. It has been shown, that with the given data set, and with the given random folds, SVD is the



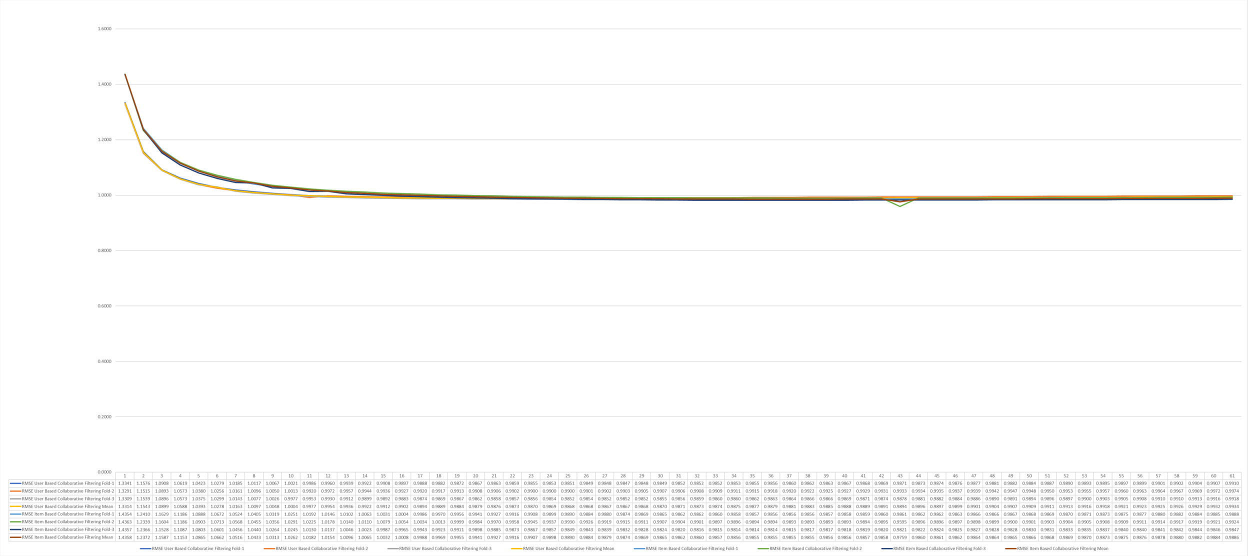
| | RMSE | | | | MAE | | | |
|---|---|---|---|---|---|---|---|---|
| | Fold-1 | Fold-2 | Fold-3 | Mean | Fold-1 | Fold-2 | Fold-3 | Mean |
| User-Based Collaborative Filtering | 0.9867 | 0.9927 | 0.9869 | 0.9887 | 0.7787 | 0.7854 | 0.7811 | 0.7817 |
| Item-Based Collaborative Filtering | 0.9858 | 0.9893 | 0.9818 | 0.9856 | 0.78 | 0.7836 | 0.7791 | 0.7809 |
| SVD | 0.9439 | 0.951 | 0.9394 | 0.9447 | 0.7439 | 0.951 | 0.7431 | 0.7455 |
| PMF | 0.9514 | 0.9592 | 0.9505 | 0.9537 | 0.7517 | 0.7574 | 0.7531 | 0.7541 |
| NMF | 0.9718 | 0.9808 | 0.9781 | 0.9769 | 0.7618 | 0.7693 | 0.769 | 0.7671 |

Figure 3: Performance of Different Filtering Methods



| | RMSE | | | | MAE | | | |
|---|---|---|---|---|---|---|---|---|
| | Fold-1 | Fold-2 | Fold-3 | Mean | Fold-1 | Fold-2 | Fold-3 | Mean |
| User-Based Collaborative Filtering MSD | 0.9867 | 0.9927 | 0.9869 | 0.9887 | 0.7787 | 0.7854 | 0.7811 | 0.7817 |
| Item-Based Collaborative Filtering MSD | 0.9858 | 0.9893 | 0.9818 | 0.9856 | 0.78 | 0.7836 | 0.7791 | 0.7809 |
| User-Based Collaborative Filtering Cosine | 1.0195 | 1.0245 | 1.0194 | 1.0211 | 0.8063 | 0.8112 | 0.8084 | 0.8086 |
| Item-Based Collaborative Filtering Cosine | 1.0336 | 1.0421 | 1.0317 | 1.0358 | 0.819 | 0.8271 | 0.8204 | 0.8221 |
| User-Based Collaborative Filtering Pearson | 1.0183 | 1.0228 | 1.0177 | 1.0196 | 0.8075 | 0.812 | 0.8085 | 0.8093 |
| Item-Based Collaborative Filtering Pearson | 1.0453 | 1.053 | 1.045 | 1.0478 | 0.8361 | 0.8425 | 0.8364 | 0.8383 |

Figure 4: Performance of Different Filtering Methods using Different Difference Formulas



| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RMSE User Based Collaborative Filtering Fold-1 | 1.3341 | 1.1576 | 1.0908 | 1.0619 | 1.0423 | 1.0279 | 1.0185 | 1.0117 | 1.0067 | 1.0021 | 0.9986 | 0.9960 | 0.9939 | 0.9922 | 0.9908 | 0.9897 | 0.9888 | 0.9882 | 0.9872 | 0.9867 |
| RMSE User Based Collaborative Filtering Fold-2 | 1.3291 | 1.1515 | 1.0893 | 1.0573 | 1.0380 | 1.0256 | 1.0161 | 1.0096 | 1.0050 | 1.0013 | 0.9920 | 0.9972 | 0.9957 | 0.9944 | 0.9936 | 0.9927 | 0.9920 | 0.9917 | 0.9913 | 0.9908 |
| RMSE User Based Collaborative Filtering Fold-3 | 1.3309 | 1.1539 | 1.0896 | 1.0573 | 1.0375 | 1.0299 | 1.0143 | 1.0077 | 1.0026 | 0.9977 | 0.9953 | 0.9930 | 0.9912 | 0.9899 | 0.9892 | 0.9883 | 0.9874 | 0.9869 | 0.9867 | 0.9862 |
| RMSE User Based Collaborative Filtering Mean | 1.3314 | 1.1543 | 1.0899 | 1.0588 | 1.0393 | 1.0278 | 1.0163 | 1.0097 | 1.0048 | 1.0004 | 0.9977 | 0.9954 | 0.9936 | 0.9922 | 0.9912 | 0.9902 | 0.9894 | 0.9889 | 0.9884 | 0.9879 |
| RMSE Item Based Collaborative Filtering Fold-1 | 1.4354 | 1.2410 | 1.1629 | 1.1186 | 1.0888 | 1.0672 | 1.0524 | 1.0405 | 1.0319 | 1.0251 | 1.0192 | 1.0146 | 1.0102 | 1.0063 | 1.0031 | 1.0004 | 0.9986 | 0.9970 | 0.9956 | 0.9941 |
| RMSE Item Based Collaborative Filtering Fold-2 | 1.4363 | 1.2339 | 1.1604 | 1.1186 | 1.0903 | 1.0713 | 1.0568 | 1.0455 | 1.0356 | 1.0291 | 1.0225 | 1.0178 | 1.0140 | 1.0110 | 1.0079 | 1.0054 | 1.0034 | 1.0013 | 0.9999 | 0.9984 |
| RMSE Item Based Collaborative Filtering Fold-3 | 1.4357 | 1.2366 | 1.1528 | 1.1087 | 1.0803 | 1.0601 | 1.0456 | 1.0440 | 1.0264 | 1.0245 | 1.0130 | 1.0137 | 1.0046 | 1.0023 | 0.9987 | 0.9965 | 0.9943 | 0.9923 | 0.9911 | 0.9898 |
| RMSE Item Based Collaborative Filtering Mean | 1.4358 | 1.2372 | 1.1587 | 1.1153 | 1.0865 | 1.0662 | 1.0516 | 1.0433 | 1.0313 | 1.0262 | 1.0182 | 1.0154 | 1.0096 | 1.0065 | 1.0032 | 1.0008 | 0.9988 | 0.9969 | 0.9955 | 0.9941 |

| | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RMSE User Based Collaborative Filtering Fold-1 | 0.9863 | 0.9859 | 0.9855 | 0.9853 | 0.9851 | 0.9849 | 0.9848 | 0.9847 | 0.9848 | 0.9849 | 0.9852 | 0.9852 | 0.9852 | 0.9853 | 0.9855 | 0.9856 | 0.9860 | 0.9862 | 0.9863 | 0.9867 |
| RMSE User Based Collaborative Filtering Fold-2 | 0.9906 | 0.9902 | 0.9900 | 0.9900 | 0.9900 | 0.9901 | 0.9902 | 0.9903 | 0.9905 | 0.9907 | 0.9906 | 0.9908 | 0.9909 | 0.9911 | 0.9915 | 0.9918 | 0.9920 | 0.9922 | 0.9925 | 0.9927 |
| RMSE User Based Collaborative Filtering Fold-3 | 0.9858 | 0.9857 | 0.9856 | 0.9854 | 0.9852 | 0.9854 | 0.9852 | 0.9852 | 0.9852 | 0.9855 | 0.9856 | 0.9859 | 0.9860 | 0.9860 | 0.9862 | 0.9863 | 0.9864 | 0.9866 | 0.9866 | 0.9869 |
| RMSE User Based Collaborative Filtering Mean | 0.9876 | 0.9873 | 0.9870 | 0.9869 | 0.9868 | 0.9868 | 0.9867 | 0.9867 | 0.9868 | 0.9870 | 0.9871 | 0.9873 | 0.9874 | 0.9875 | 0.9877 | 0.9879 | 0.9881 | 0.9883 | 0.9885 | 0.9888 |
| RMSE Item Based Collaborative Filtering Fold-1 | 0.9927 | 0.9916 | 0.9908 | 0.9899 | 0.9890 | 0.9884 | 0.9880 | 0.9874 | 0.9869 | 0.9865 | 0.9862 | 0.9862 | 0.9860 | 0.9858 | 0.9857 | 0.9856 | 0.9856 | 0.9856 | 0.9857 | 0.9858 |
| RMSE Item Based Collaborative Filtering Fold-2 | 0.9970 | 0.9958 | 0.9945 | 0.9937 | 0.9930 | 0.9926 | 0.9919 | 0.9915 | 0.9911 | 0.9907 | 0.9904 | 0.9901 | 0.9897 | 0.9896 | 0.9894 | 0.9894 | 0.9893 | 0.9893 | 0.9893 | 0.9893 |
| RMSE Item Based Collaborative Filtering Fold-3 | 0.9885 | 0.9873 | 0.9867 | 0.9857 | 0.9849 | 0.9843 | 0.9839 | 0.9832 | 0.9828 | 0.9824 | 0.9820 | 0.9816 | 0.9815 | 0.9814 | 0.9814 | 0.9814 | 0.9815 | 0.9817 | 0.9817 | 0.9818 |
| RMSE Item Based Collaborative Filtering Mean | 0.9927 | 0.9916 | 0.9907 | 0.9898 | 0.9890 | 0.9884 | 0.9879 | 0.9874 | 0.9869 | 0.9865 | 0.9862 | 0.9860 | 0.9857 | 0.9856 | 0.9855 | 0.9855 | 0.9855 | 0.9855 | 0.9856 | 0.9856 |

| | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RMSE User Based Collaborative Filtering Fold-1 | 0.9868 | 0.9869 | 0.9871 | 0.9873 | 0.9874 | 0.9876 | 0.9877 | 0.9881 | 0.9882 | 0.9884 | 0.9887 | 0.9890 | 0.9893 | 0.9895 | 0.9897 | 0.9899 | 0.9901 | 0.9902 | 0.9904 | 0.9907 | 0.9910 |
| RMSE User Based Collaborative Filtering Fold-2 | 0.9929 | 0.9931 | 0.9933 | 0.9934 | 0.9935 | 0.9937 | 0.9939 | 0.9942 | 0.9947 | 0.9948 | 0.9950 | 0.9953 | 0.9955 | 0.9957 | 0.9960 | 0.9963 | 0.9964 | 0.9967 | 0.9969 | 0.9972 | 0.9974 |
| RMSE User Based Collaborative Filtering Fold-3 | 0.9871 | 0.9874 | 0.9878 | 0.9881 | 0.9882 | 0.9884 | 0.9886 | 0.9890 | 0.9891 | 0.9894 | 0.9896 | 0.9897 | 0.9900 | 0.9903 | 0.9905 | 0.9908 | 0.9910 | 0.9910 | 0.9913 | 0.9916 | 0.9918 |
| RMSE User Based Collaborative Filtering Mean | 0.9889 | 0.9891 | 0.9894 | 0.9896 | 0.9897 | 0.9899 | 0.9901 | 0.9904 | 0.9907 | 0.9909 | 0.9911 | 0.9913 | 0.9916 | 0.9918 | 0.9921 | 0.9923 | 0.9925 | 0.9926 | 0.9929 | 0.9932 | 0.9934 |
| RMSE Item Based Collaborative Filtering Fold-1 | 0.9859 | 0.9860 | 0.9861 | 0.9862 | 0.9862 | 0.9863 | 0.9866 | 0.9866 | 0.9867 | 0.9868 | 0.9869 | 0.9870 | 0.9871 | 0.9873 | 0.9875 | 0.9877 | 0.9880 | 0.9882 | 0.9884 | 0.9885 | 0.9888 |
| RMSE Item Based Collaborative Filtering Fold-2 | 0.9894 | 0.9895 | 0.9595 | 0.9896 | 0.9896 | 0.9897 | 0.9898 | 0.9899 | 0.9900 | 0.9901 | 0.9903 | 0.9904 | 0.9905 | 0.9908 | 0.9909 | 0.9911 | 0.9914 | 0.9917 | 0.9919 | 0.9921 | 0.9924 |
| RMSE Item Based Collaborative Filtering Fold-3 | 0.9819 | 0.9820 | 0.9821 | 0.9822 | 0.9824 | 0.9825 | 0.9827 | 0.9828 | 0.9828 | 0.9830 | 0.9831 | 0.9833 | 0.9835 | 0.9837 | 0.9840 | 0.9840 | 0.9841 | 0.9842 | 0.9844 | 0.9846 | 0.9847 |
| RMSE Item Based Collaborative Filtering Mean | 0.9857 | 0.9858 | 0.9759 | 0.9860 | 0.9861 | 0.9862 | 0.9864 | 0.9864 | 0.9865 | 0.9866 | 0.9868 | 0.9869 | 0.9870 | 0.9873 | 0.9875 | 0.9876 | 0.9878 | 0.9880 | 0.9882 | 0.9884 | 0.9886 |

Figure 5: RMSE Filtering with each Fold and Mean of the Three Folds

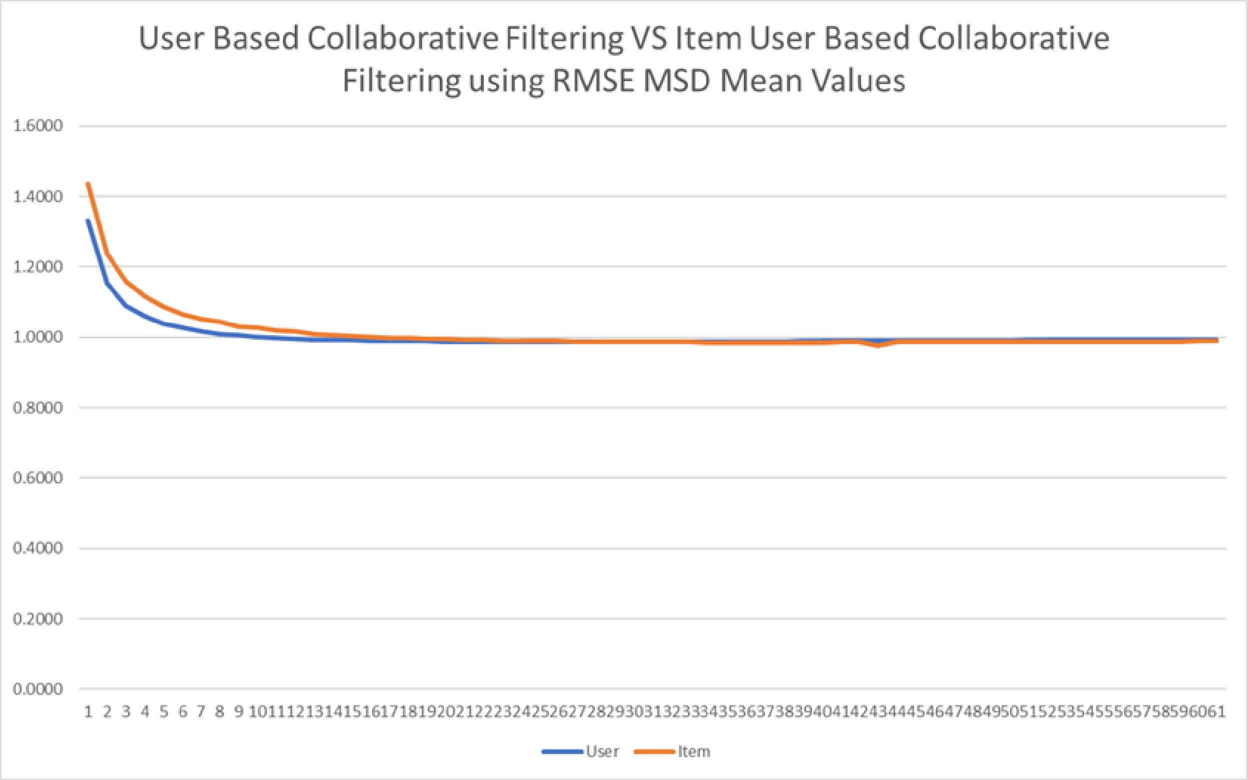


Figure 6: User-Based Collaborative Filtering VS Item-Based Collaborative Filtering using RMSE MSD Mean Values

best filtering algorithm out of the five in the experiment. Between user-based and item-based collaborative filtering, item-based generally out performed user-based. Since these two methods are variations of the kNN method, it is important to keep in mind the value of k. When k has a value less than 30, user-based collaborative filtering out performed item-based. This will vary depending on the data set, but the general trend that user-base will out perform item-based up until a certain threshold is expected.

Different similarity measurements matter highly as well. As shown in Figure 4, item-based Pearson method performed the worst, however, out of the user-based methods, using the cosine similarity performed the worst. Different similarity measurements vary on different algorithms and data sets.