ZESZYTY NAUKOWE POLITECHNIKI BIAŁOSTOCKIEJ 2008
Informatyka – Zeszyt 3

Anna Łupińska–Dubicka,[1] Marek J. Drużdżel[1,2]

ANALYZING CERTAIN TEMPORAL DEPENDENCES IN NETFLIX DATA

Abstract: Netflix (see `http://www.netflix.com/`), an American Internet-based movie rental company, uses data mining in their recommendation system. In October 2006 Netflix made a huge data base of their users and movie evaluations available to the community and announced a million dollars prize to the team that beats the accuracy of their recommendations by at least 10%. The data have since become an object of interest of the machine learning community. In this paper, we focus on one aspect of the data that, to our knowledge, has been overlooked — their temporal dependences. We have looked at the impact of the day of the week, month of the year, length of membership, month from the start of Netflix, etc., on the average evaluation.

Keywords: Data analysis, temporal dependences, Netflix

1. Introduction

Netflix (see `http://www.netflix.com/`), an American Internet-based movie rental company, offers its over 6.7 milllion subscribers access to over 85,000 DVD titles plus a growing library of over 4,000 full-length movies and television episodes that are available for instant watching on their PCs. Based on a user's viewing pattern and evaluations of previously watched movies, their user interface recommends to their clients movies that they might like as well. These recommendations, if reasonably accurate, are a valuable source of information to Netflix users.

Netflix recently announced a million dollar prize to the team that beats the accuracy of their recommendations by at least 10%. For this purpose, the company has made a huge subset of their movie evaluations database available to the community. This database contains over 500,000 users (out of roughly 2.5 million) and over 100 million evaluations. As of the start of the contest, the square root of the mean square error (RMSE) of Cinematch, the company's recommendation system, was 0.9525

[1] Faculty of Computer Science, The Bialystok Technical University, Bialystok, Poland
[2] Decision Systems Laboratory, School of Information Sciences, University of Pittsburgh, Pittsburgh, Pensylwania, USA

(see http://www.netflixprize.com/). Over 20 thousand teams have joined the competition so far and almost 700 teams have beaten Netflix, although none of them has beaten Netflix accuracy by more than 10% as of December 2007. 90 teams have improved the accuracy of Netflix predictions by more than 5%.

The Netflix data set is at the moment one of the most comprehensive and most challenging data sets available to the machine learning community. Many top machine learning and data mining teams have studied it. The best team participating in the competition as of December 2007, has beaten Netflix accuracy by 8.5%. The remaining 1.5% is fairly hard to overcome. To improve the accuracy of predictions, any source of information, even one that is weak, can play a role. We focus on one aspect of data that, to our knowledge, has been overlooked — temporal dependences among the measured variables. Every evaluation (of over 100 million evaluations available in the data) has a time stamp. Surprisingly, it turns out that this variable is quite depend on how people rate movies. We have looked at the impact of the day of the week, month of the year, length of membership, month from the start of Netflix, etc., on the average evaluation and found that incorporating this information improves the quality of predictions.

2. The data sets

Netflix has collected over 1.9 billion ratings from more than 11.7 million subscribers and over 85 thousand titles since October 1998 and has shipped over 1 billion DVDs. It receives over 2 million ratings per day. For the competition, the company provided over 100 million ratings from over 480 thousand randomly-chosen, anonymous subscribers over nearly 18 thousand movie titles. The date of each rating is also included. The data were collected between October 1998 and December 2005 and reflect the distribution of all ratings received by Netflix during this period. The ratings are on a scale from 1 to 5 and from the point of view of a user, are pictured as a collection of integral stars. To protect customer privacy, each customer id has been replaced in the data by a randomly-assigned id. Figure 1 shows the cumulative number of Netflix customers as a function of time in the Netflix data set. Assuming that the sample is representative, the figure gives an idea of the growth of the company in terms of new customers [3].

The contestants are also given the title and the year of release of each movie. The title is the Netflix movie title, year of release can range from 1890 to 2005 and typically, although not consistently, corresponds to the release year of the movie.

The complete data set consists of two separated subsets: the training set and the qualifying set. The ratings from the training set were released to contestants while

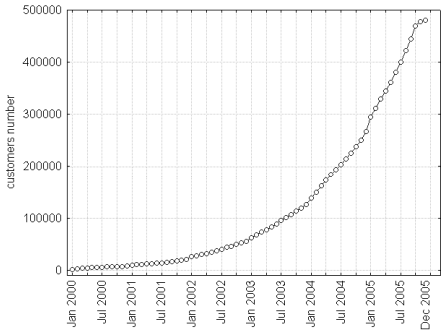


Fig. 1. Cumulative number of customers as a function of time in the Netflix data

the qualifying ratings were withheld and form the basis of the contest scoring system. Contestants are required to make prediction for all 3 million withheld ratings in the qualifying data set. Testing contestants' results on the most recent ratings reflects Netflix's business goal of predicting future ratings based on past ratings. The RMSE is computed automatically and reported to the contestants [4].

3. Differences between the probe and the training data sets

It is critical for analyzing the Netflix data to realize that they consist of two data sets with quite different statistical properties. We will review briefly the method by which the two data sets were obtained.

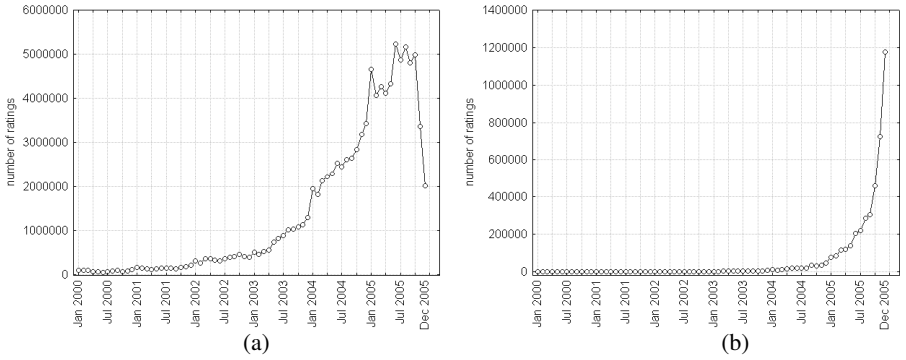
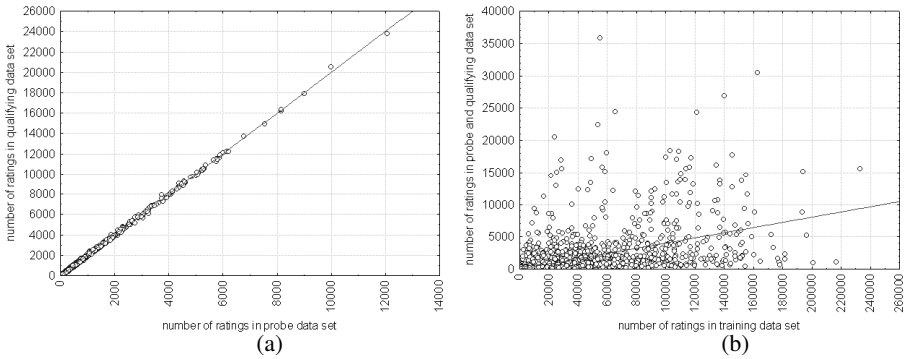


Fig. 2. Number of ratings as a function of time in the training data set without probe subset (a) and in the probe and qualifying data sets (b)

The complete data set consists of two separated subsets: the training data set and the qualifying data set. These sets were created by randomly selecting a subset of all users who provided at least 20 ratings between October 1998 and December 2005. Then the most recent ratings of these users were randomly assigned, with equal probability, to three subsets: quiz, test, and probe. The qualifying data set comprises the quiz and test subsets. The training data set was created from all remaining ratings and the probe subset [1].

Figure 2 shows the distribution of the number of ratings in the training data set (excluding the probe subset) and in the qualifying and probe data sets combined. It is clear that most of the ratings in the qualifying and the probe data sets are very recent.



(a) (b)

Fig. 3. The relations between the number of ratings per movie in the probe and qualifying data sets (a) and in the training data set and in the probe and qualifying data sets (b)

Please note that the number of ratings per movie in Figure 3 indicates that the ratings in the probe and qualifying subsets are not randomly selected from the training data set. In Figure 3-a, the scatterplot shows the relation between the number of ratings per movie in the probe and qualifying data sets. In Figure 3-b, it is noticeable that the relationship between the number of ratings per movie in the training data set and in the probe and qualifying data sets is much weaker, if at all.

Similarly, Figure 4 shows that the probe and the training data sets have different average ratings for both movies and users. It seems that the most recent movies (i.e. those in the qualifying and probe data sets) tend to get higher ratings. The average rating per user in the training data set is a little bit smaller and has smaller variance, which is the result of a small number of data

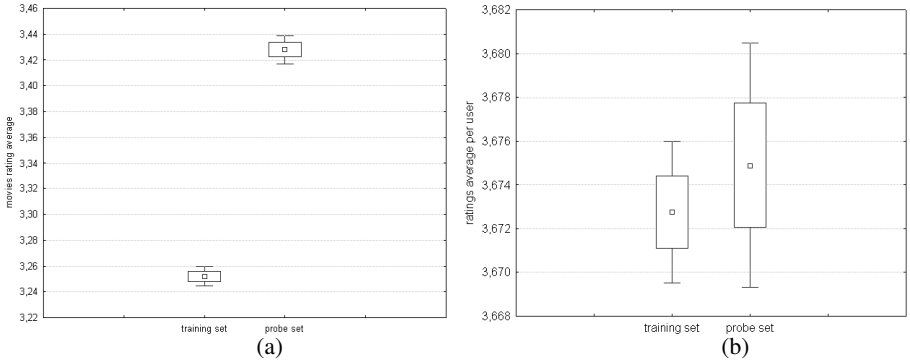


Fig. 4. Average ratings per movie (a) and per user (b) in the trainings and probe data sets

4. Movie age

We found that older movies tend to get better user ratings than new movies (Figure 5-a). This could be explained by the fact that most movies rented by Netflix are in DVD format. While there are fewer older than there are new movies, those that have been published in DVD may simply belong to masterpieces and, hence, effectively get high ratings. Virtually, all new movies are published in the DVD format and, as is usually the case, only a fraction of them are masterpieces. Hence, the average rating is lower.

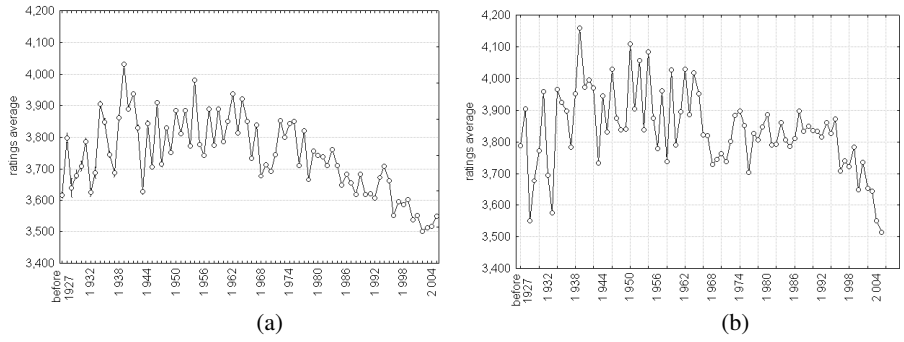


Fig. 5. Average rating as a function of movie's production year in training (a) and probe (b) data set

Comparing Figure 5-a and Figure 5-b, it is not difficult to see that this trend is still observable in the probe subset. The movies from '40s, '50s, and '60s receive higher ratings than these produced nowadays.

5. Netflix Age

Movies tend to get higher ratings as time goes by (see Figure 6). This may have to do with a varying population of Netflix customers. For example, it is well known that technology-based companies attract in the beginning so called "early adopters". Mainstream customers, typically more conservative than the "early adopters," get interested only later. The difference in average evaluation can possibly be explained by the difference between these two groups of customers.

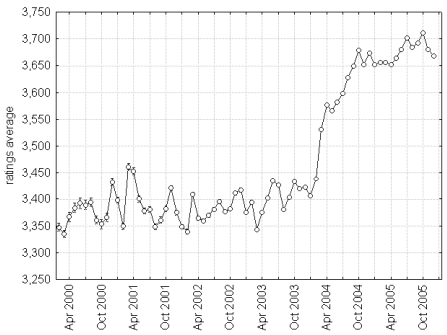


Fig. 6. Average rating as a function of Netflix's age

It is quite interesting to observe a sudden increase (about 0.15 point) in average ratings between January and April, 2004. To our knowledge, this increase was associated neither with a significant increase in the number of Netflix customers (see Figure 1), nor with the number of ratings (see Figure 2).

6. Weekly variations

We were trying to investigate why there is a difference in average rating among various days of the week. One might think that it may have to do with varying characteristics of the populations that rate on the weekends and those that rate on weekdays

– weekend customers, for example, may be more relaxed and less critical, which results in higher ratings. And in this case, confronting comeback to daily business may cause a more critical attitude and a decrease in ratings.

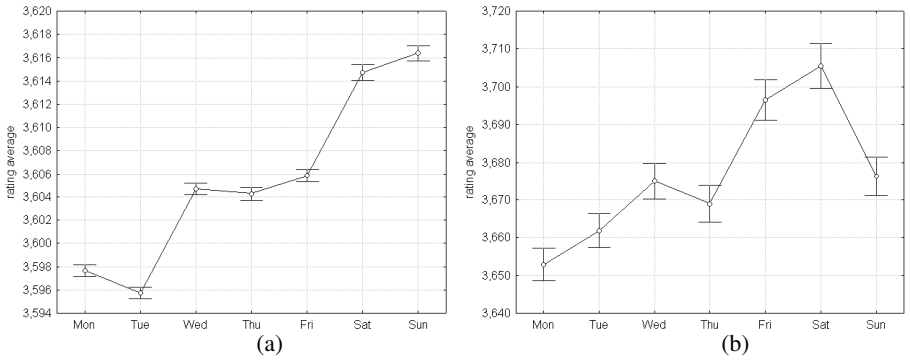


Fig. 7. Average rating as a function of day of the week in training (a) and probe (b) data set

We also looked at the number of ratings on each day of the week (see Figure 8). Almost twice as many customers rate on Mondays and Tuesdays than on weekends. This may have to do with the viewing attern – quite likely more movies are watched on weekends. Watched movies are evaluated and new movies are ordered after the weekend viewing.

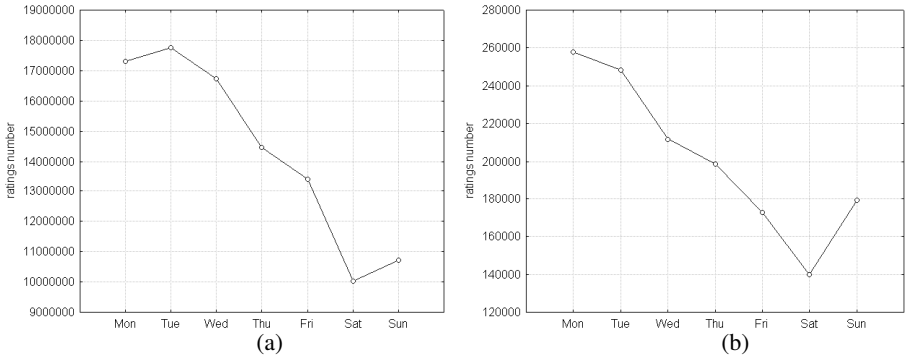


Fig. 8. Number of ratings as a function of day of the week in training (a) and probe (b) data set

We believe that while there could be a correlation between the day of the week, the customer's mood and the ratings he gives, the observed differences are rather due to individual differences among customers. Some customers typically vote on weekdays, other on weekend – probably they are the majority, as shown in Figure 8. It is also possible that those who rate on weekends, rate higher than those who rate on weekdays.

To check if there is a relationship between ratings of particular user and day of week we chose randomly a few customer with different number of ratings. Figure 9 (custmers who rated more than 16 thousand times), Figure 10 (custmers who rated about three thousand times), and Figure 11 (custmers who rated about 500 times) show their average rating and the number of ratings as a function of the day of the week. We observed that both average rating and the number of ratings falling on the day of the week is very a individual feature.

For example customer 305344 (Figure 9a, c) rates the most his movies on Fridays but gives in this day the lowest evaluations. The highest ratings he gives on Sundays and Mondays. In his case the average rating is the lower the more he votes. Customer 2439493 (Figure 9b, d) is active the most in the middle of the week, his ratings are going up and down. Customer 981753 (Figure 10a, c) rates a lot on Tuesdays but similarly to customer 305344, the more ratings the lower they are. Customer 1092521 (Figure 10b, d) is an example of a user who votes mainly once a week – on Saturday, giving at that time the highest ratings. Again, customer 912242 (Figure 11a, c) rates for the most part on Tuesdays and his ratings are then the lowest. Customer 1187552 (Figure 11b, d) rates the movies almost at the same level, independently of the day of the week. None of his ratings was given on Saturdays; Sunday, Tuesday and Wednesday are also days with almost zero activity.

It looks like there are several types of customers. One of them prefer to vote on weekdays giving then lower ratings, the second like rating on weekends and their evaluations are higher, the others give the more critical ratings the more frequently they rate, regardless of the day of the week. In case of the whole Netflix customers' population, the singular influence of these different groups can cancel each other out. Finding these groups of users, similar to each other with respect to the number of ratings and the rating average, could help to benefit from this kind temporal dependency.

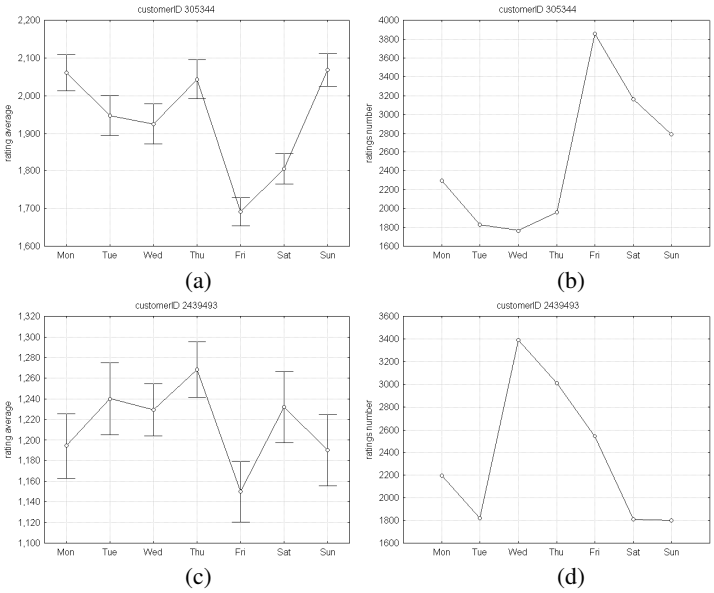


Fig. 9. Average rating (a, c) and the number of ratings (b, d) as a function of the day of the week for two customers who rated more than 16 thousand times

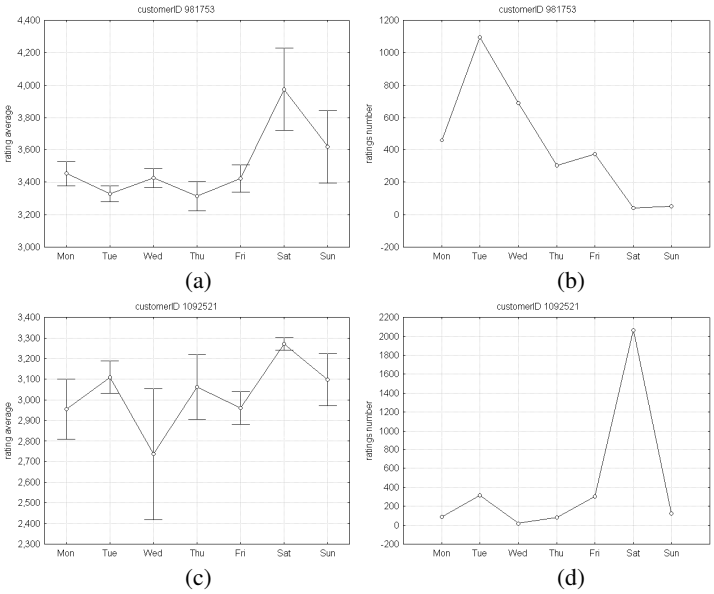


Fig. 10. Average rating (a, c) and the number of ratings (b, d) as a function of the day of the week for two customers who rated about three thousand times

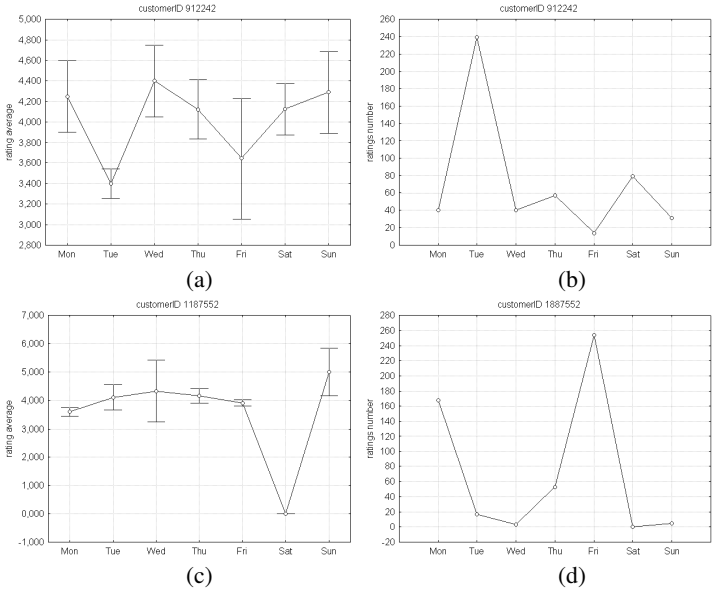


Fig. 11. Average rating (a, c) and the number of ratings (b, d) as a function of the day of the week for two customers who rated about 500 times

7. Customer behavior as a function of the length of membership

Customers tend to give higher ratings in the beginning of their membership. This goes down within the first year but picks up again somewhat after a few years (see Figure 12).

A probable cause for this trend is that in the beginning of their membership customers choose movies they have seen in movie theaters, liked, and desired to see again. As time goes by, they run out of movies that they want to watch again or that were recommended by friends or their families. Effectively, their rental may become more accidental, based on less strong recommendations and desire to watch, which may entail a decrease in average rating.

The last 2–3 years in Figure 12 show a large variation in ratings. It is quite likely the effect of a relatively small sample size – the number of customers who have been Netflix's members for more than 4 years is rather small.

To verify this, we changed the scale on the time axis and plotted the average rating as a function of lenght membership in reverse order, i.e. point 0 shows the most recent ratings, point 1 ratings from month before, etc. It seems like the average rating is going up with time (i.e., going down in reverse time scale).

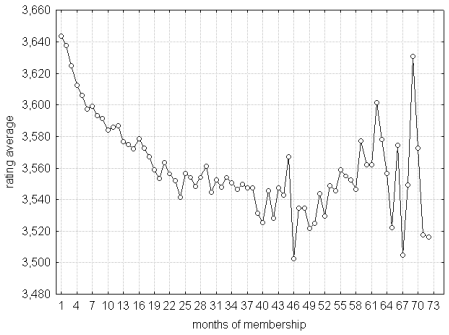
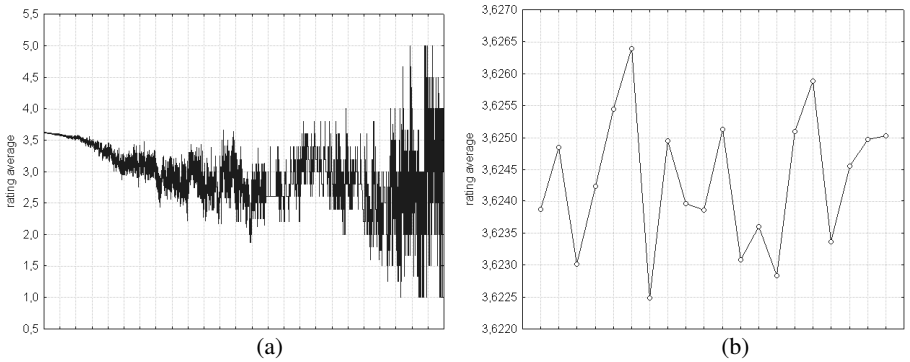


Fig. 12. Average rating as a function of a customer's membership age



(a) (b)

Fig. 13. Average rating as a function of rating number in order from the last rating to the first – all ratings (a) and the last 20 (b)

8. Experimental study

As mentioned before, Netflix provided two separated data sets: the training set and the qualifying set for the competition. While movie ratings for the qualifying set are not known, it is possible to test rating algorithms on the "probe" subset of the training data, identified by Netflix and statistically similar to the qualifying set. In our previous work [2], we removed the probe subset from all available ratings and were trying to guess the ratings from the "probe" subset. In this paper, we concentrated on predicting the ratings from the qualifying data set. First we made a prediction on the basis of all data in the trainig data set, then we checked if leaving only most recent data (provided in the probe subset) could improve the results.

There are many ways in which temporal dependences can be used in guessing the rating. In this paper we use a few simple techniques, such as movie average and user average, just to illustrate our point that knowledge of temporal dependences can reduce the guessing error.

8.1 Prediction based on customer's average rating from the training data set

RMSE = 1.0655

$$\text{prediction} = \text{cAVG}\ , \tag{1}$$

where the variable cAVG means average of all customer's ratings.

8.2 Prediction based on movie's average rating from the training data set

RMSE = 1.0536

$$\begin{aligned} &\text{if } (\text{cStdDev} < 0.01)\ \text{prediction} = \text{cAVG} \\ &\text{else prediction} = \text{mAVG}\ , \end{aligned} \tag{2}$$

where variable cStdDev stands for standard deviation of all customer's ratings, variable cAVG means average of all customer's ratings, variable mAVG is equal to average of all movie's ratings.

8.3 Prediction based on weekly variations of customer's average from the training data set

RMSE = 1.1266

$$\text{prediction} = 0.9\ \text{cAVG} + 0.1\ \text{cDayOfWeekAVG}\ , \tag{3}$$

where variable cDayOfWeekAVG denotes average of all customer's ratings on a given day of the week.

8.4 Prediction based on weekly variations of movie's average from the training data set

RMSE = 1.0525

$$\begin{aligned} &\text{if } (\text{cStdDev} < 0.01)\text{prediction} = \text{cAVG} \\ &\text{else prediction} = 0.9\ \text{mAVG} + 0.1\ \text{mDayOfWeekAVG}\ , \end{aligned} \tag{4}$$

where variable mDayOfWeekAVG stands for average of all ratings that this movie has been given on a specific day of the week.

8.5 Prediction based on customer's average rating from the probe data set

RMSE = 1.0672

$$\begin{aligned} &\text{if } (\text{cNoRatingsProbe} > 5) \text{ prediction} = \text{cAVGfromProbe} \\ &\text{else prediction} = \text{cAVG} , \end{aligned} \tag{5}$$

where the variable cNoRatingsProbe means the number of ratings this customer contained in the probe subset, variable cAVGfromProbe denotes average of customer's ratings from the probe subset, variable cAVG stands for average of all customer's ratings (from the training and probe data sets).

8.6 Prediction based on movie's average rating from the probe data set

RMSE = 1.0490

$$\begin{aligned} &\text{if } (\text{mNoRatingsProbe} > 7) \text{ prediction} = \text{mAVGfromProbe} \\ &\text{else if } (\text{cStdDev} < 0.01) \text{ prediction} = \text{cAVG} \\ &\text{else prediction} = \text{mAVG} , \end{aligned} \tag{6}$$

where the variable mNoRatingsProbe denotes the number of ratings this movie from the probe subset, variable mAVGfromProbe stands for average of movie's ratings contained in the probe subset, variable cStdDev is equal to standard deviation of all customer's ratings, variable mAVG stands foraverage of all movie's ratings (from the training and probe data sets).

8.7 Prediction based on weekly variations of customer's average from the probe data set

RMSE = 1.1418

$$\begin{aligned} &\text{if } (\text{cDayOfWeekNoRatingsProbe} >= 5) \\ &\text{prediction} = 0.9\ \text{cAVG} + 0.1\ \text{cDayOfWeekAVGfromProbe} \\ &\text{else if } (\text{cDayOfWeekNoRatings} > 0) \\ &\text{prediction} = 0.9\ \text{cAVG} + 0.1\ \text{cDayOfWeekAVG} \\ &\text{else prediction} = \text{cAVG} , \end{aligned} \tag{7}$$

where variables cDayOfWeekNoRatingsProbe and cDayOfWeekNoRatings denote the number of customer's ratings on a given day of the week in the probe data and

training (including probe subset) data sets respectively, variables cDayOfWeekAVGfromProbe and cDayOfWeekAVG stand for the average of all customer's ratings on a given day of the week in the probe data and training (including probe subset) data sets respectively, variable cAVG means average of all customer's ratings, independently of the day of the week.

8.8 Prediction based on weekly variations of movie's average from the probe data set

RMSE = 1.0525

$$\begin{aligned} &\text{if } (\text{cStdDev} < 0.01) \text{prediction} = \text{cAVG} \\ &\text{else if } (\text{mDayOfWeekNoRatingsProbe} >= 5) \\ &\text{prediction} = 0.9\ \text{mAVG} + 0.1\ \text{mDayOfWeekAVGfromProbe} \\ &\text{else if } (\text{mDayOfWeekNoRatings} > 0) \\ &\text{prediction} = 0.9\ \text{mAVG} + 0.1\ \text{mDayOfWeekAVG} \\ &\text{else prediction} = \text{mAVG}\ , \end{aligned} \tag{8}$$

where variable cStdDev is equal to standard deviation of all customer's ratings, variable mDayOfWeekNoRatingsProbe and mDayOfWeekNoRatings stand for the number of movie's ratings that this movie has been given on a specific day of week in the probe data and training (including probe subset) data sets respectively, variable mDayOfWeekAVGfromProbe and mDayOfWeekAVG denote the average of ratings that this movie has been given on a specific day of week in the probe data and training (including probe subset) data sets respectively, variable mAVG is equal to the average of all movie's ratings, independently of the day of the week.

9. Concluding remarks

Recommender systems are programs and algorithms that measure user interest in given items or products to provide personalized recommendations for items that will suit the user's taste. More broadly, recommender systems attempt to profile user preferences and model the interaction betwen users and products. One possible strategy for building such systems relies only on past user behaviour, without requiring creation of explicit profiles.

Netflix, an on-line movie subscription rental service, allows people to rent movies for a fixed monthly fee, maintaining a prioritized list of movies they wish to view.

The company encourages subscribers to rate the movies that they watch, expressing an opinion about how much they liked (or disliked) each movie. To predict the customer's rating of the movie, the Netflix recommendation system, Cinematch analyses the past ratings using a variant of Pearson's correlation. In October 2006, the company released a large data set of movie-ratings and challenged the data mining, machine learning, and statistical communities to develop systems that could improve the accuracy of Cinematch. At the start of the contest Cinematech's RMSE was 0.9525. One year later, in December 2007, the best team has achieved a 8.5% improvement over this score. It seems that the remaining 1.5% is fairly hard to overcome.

Given the current impasse and lack of significant progress, it is clear that every, even a weak source of information about customers' preferences can play an important role in improving the predictions. In previous paper [2], we presented certain observations that focused on temporal dependencies in Netflix' data. We have shown that there is a strong dependence of the average rating on the the day of week, Netflix' age and length of a customer's membership.

Our next step was to focus on the differences between the probe subset and the rest of the training data set. Both, the probe and the qualifying subsets were created from the most recent ratings and are very different from the training data set and, as it was shown, they have similar properties. It seems that our idea that the most recent ratings can provide much more information about future votes was good direction. When predictions are based on the movies' average ratings only from the probe subset, the RMSE is going down. Some problems appeared when we tried to guess with basis on the customers' average ratings. The RMSE was growing up when we confined to the probe subset. The possible explanation can be relatively small size of the probe subset in comparison with the number of rating contained in the training data set. When we are taking account of movies' average, the number of ratings per movie is enough to cause the decrease of RMSE. But considering customers' average one might notice that the proportion of 1.5 million of ratings in the probe subset and to 480 thousand of customers is too small to receive reliable results. We believe that enlargement the probe subset with maintenance its properties, such as number of ratings per movie and per user, should bring improvement of the accuracy of predictions based on the customer's preferences.

Bibliography

[1] Bennett, J. Lanning, S.: The Netflix Prize, Proceedings of KDD Cup and Workshop 2007, Aug. 12, 2007.

[2] Łupińska–Dubicka, A. Drużdżel, M.J. Temporal Aspects of Netflix Data, 2007.

[3] http://www.netflix.com
[4] http://www.netflixprize.com
[5] http://www.netflixprize.com/community

ANALIZA WYBRANYCH ZALEŻNOŚCI CZASOWYCH W DANYCH NETFLIX

Streszczenie: Działający w Stanach Zjednoczonych Ameryki Netflix (http://www.netflix.com/) jest jedną z największych na świecie internetowych wypożyczalni filmów. W celu uzyskania wyższej jakości proponowanych przez system ocen filmów, w październiku 2006 roku Netflix udostępnił bazę danych użytkowników oraz ich ocen i ogłosił nagrodę dla tego, kto uzyska co najmniej 10-cio procentową poprawę w stosunku do wyników Cinematch (RMSE=0.9525). W tym artykule postawiliśmy sobie za cel zbadanie, czy zależności czasowe, takie jak dzień tygodnia lub długość członkostwa, są w stanie zwiększyć jakość prognozowania.

Słowa kluczowe: Zależnosci czasowe, Netflix