Collaborative Filtering or Regression Models for Internet Recommendation Systems?

Article *in* Journal of Targeting Measurement and Analysis for Marketing · March 2002

DOI: 10.1057/palgrave.jt.5740055

2 authors:

Andreas Mild
Wirtschaftsuniversität Wien

Martin Natter
University of Zurich

COLLABORATIVE FILTERING OR REGRESSION MODELS FOR INTERNET RECOMMENDATION SYSTEMS?

Andreas Mild and Martin Natter

Department of Production Management

Vienna University of Economics and Business Administration,

Pappenheimgasse 35/3/5

A-1200 Vienna, Austria

e-mail: andreas.mild@wu-wien.ac.at, martin.natter@wu-wien.ac.at

Published in: Journal of Targeting, Measurement and Analysis for Marketing, **2002,Vol. 10, Nr 4, pp. 304-313**

COLLABORATIVE FILTERING OR REGRESSION MODELS FOR INTERNET RECOMMENDATION SYSTEMS?

ABSTRACT

The literature on recommendation systems indicates that the choice of the methodology significantly influences the quality of recommendations. The impact of the amount of available data on the performance of recommendation systems has not been systematically investigated. We study different approaches to recommendation systems using the publicly available EachMovie data set containing ratings for movies and videos. In contrast to previous work on this data set, here a significantly larger subset is used. The effects caused by the available number of customers and movies as well as their interaction with different methods are investigated. We compare two commonly used collaborative filtering approaches to several regression models using an experimental full factorial design. According to our findings, the number of customers significantly influences the performance of all approaches under study. For a large number of customers and movies, we show that simple linear regression with model selection can provide significantly better recommendations than collaborative filtering. From a managerial perspective, this gives suggestions about the selection of the model to be used depending on the amount of data available. Furthermore, the impact of an enlargement of the customer database on the quality of recommendations is shown.

Keywords: Internet Recommendation Systems, Collaborative Filtering, Linear Regression, Ridge Regression, Logistic Regression

INTRODUCTION

E-commerce applications typically provide customers with larger product assortments than brick-and-mortar stores. In contrast to physical stores where products are nicely arranged around the shop, computer interfaces have a limited space of representation. For customers who already know which products they are looking for, simple search functions can help. However, for many product categories such as books, compact discs or movies, variety seeking plays an important role in choice decisions; i.e., simple search functions are not sufficient for supporting the customers' search process. Recommendation systems[1] endeavor to bridge the gap between the customer's demand for search assistance and her/his inability to express preference structures. In analogy to successful real-world sellers, recommendation systems use their customers' purchase history to determine the preference structure and identify products that a customer is likely to buy. In most applications, these systems use no actual product content but are based on choice or preference patterns of other users. Implicitly, one assumes that a good way to predict the products of interest to a customer is to look at other people who show similar behavior[2]. Besides the reduction of the search effort for customers, recommendation systems promise greater customer loyalty, higher sales, more advertising revenues, and the benefit of targeted promotions[3]. Practical implementations of such systems can be found at Amazon.com (books, CDs) or www.cdnow.com (CDs).

In the literature, different approaches to recommendation systems have been studied. Sarwar et al.[4] compare collaborative filtering systems based on similarities between users to methods which consider similarities between products (items). They show that the item-based approach is preferable in terms of recommendation quality and computational effort. Breese et al.[5] find that Bayesian networks with decision trees at each node and correlation methods outperform Bayesian clustering and vector-similarity methods. Chen and George[6] compare several Baysian models to the original collaborative filtering approach proposed by Shardanand and Maes[7] and find that their approach performs better. Runte[8] investigates the performance of correlation-based and distance-based collaborative filtering approaches and compares them to unpersonalized recommendations (item-specific averages). He finds that distance-based methods outperform correlation-based predictions which, in turn, perform better than non-personalized recommendations. The literature mentioned shows that various approaches have been proposed and compared. Several contributions use the mean absolute error as a performance measure. However, in our opinion the different results cannot be compared since they all use different sizes of subsets of the original data set. Although the literature considered indicates that the choice of the methodology adopted significantly influences the quality of recommendations, we suppose that some of the results maintained in the above-mentioned studies might be a result of the specific design (data selection) chosen.

Table 1: Design of previous studies on recommendation systems

Study	**Customers**	**Movies**	**Percentage of ratings used**
Ansari et al.	2000	340	2,0 %
Breese et al.	4119	1623	6,8 %
Chen & George	1373	41	0,05%
Runte	1995	683	3,7 %
Present study	61007	419	75,2 %

Table 1 shows that previous studies only use a small fraction of the data available. We hypothesize that both the amount of data and the interaction between the amount of data and the method used have a significant impact on the quality of recommendations, too. We benchmark collaborative filtering approaches against several variants of multivariate regression analysis. Our analysis is focused on the most relevant case where a recommendation system is used to predict ratings for (new) users for a given set of films. An analysis of these effects could yield interesting methodological and managerial implications. On the one hand, such results provide suggestions about the model to be selected depending on the amount of data available. On the other hand, the impact of an enlargement of the customer database on the quality of recommendations can be shown. As a higher quality of recommendations is expected to enhance customer loyalty which, in turn, increases the customer lifetime value, this research topic is of high practical relevance. To be able to study the effects of different customer database sizes, we need to consider larger portions of the EachMovie database than those dealt with in the mentioned studies. The data we analyze in this work represents more than 75% of all ratings in the EachMovie data set (Table 1). This is a significantly higher percentage than that investigated in all other studies here considered.

DATA AND RECOMMENDATION MODELS

Data

To experiment with a collaborative filtering algorithm, the Compaq Systems Research Center ran the EachMovie recommendation service for 18 months. During that time, 72916 users entered a total of 2811983 numeric ratings for 1628 different movies (films and videos). This data set was made available to researchers for testing new algorithms. The movies are rated on a 6-point scale. From the 1628 movies many have very few ratings. Due to computational restrictions, we could not use the full data set. From the original data base, we selected users who have rated more than 3 movies. This is equivalent to real world system which refuse recommendations before a minimum number of ratings is delivered. From this selection, we picked the most relevant movies in terms of the number of ratings. By selecting movies with more than 50 ratings in each of the samples, we finally arrived at a manageable data set size (199.07 MB in MATLAB format). Although the reduced data set contains about 75% of all ratings and 84% of all users, we only make use of about 26% of all movies. However, due to our design of the study (see Table 2), also situations where only a very limited number of ratings for a specific movie or customer is available, are analyzed. The remaining data set consists of 61007 users and 419 movies. We split the set of available customers into the following three groups:

- a training sample consisting of 50,000 randomly selected customers, this data set is used for model estimation
- a validation sample containing 5,000 randomly selected customers, this data set is used for tuning model parameters such as the number of neighbors (collaborative filtering) or stepwise parameter selection (regression models)

- a generalization sample consisting of 6007 randomly selected customers, this data set serves for performance measurement.

In the following subsections, we describe the models used for generating movie recommendations. The task of a recommendation system is to predict a movie's rating for a specific customer (dependent variable) based on her weighted ratings on other movies (independent variables). The models differ in the way the weights are calculated. However, all models use the ratings of other customers for weight estimation. After estimation of the model parameters, recommendations from the models can be received by transforming the predictions into discrete ratings on a 6 point scale.

Collaborative Filtering

We consider two variants of collaborative filtering regarding the calculation of the similarities between movies (items), namely, a correlation-based similarity measure and a distance based one. The correlation-based method simply calculates the Pearson-r correlation on the basis of co-rated movies. Let the set of users who rated the movies *i* and *j* be denoted by *U.* Then, the similarity is defined as:

$$sim(i,j) = \frac{\sum_{u \in U} (R_{u,i} - \overline{R}_i)(R_{u,j} - \overline{R}_j)}{\sqrt{\sum_{u \in U} (R_{u,i} - \overline{R}_i)^2} \sqrt{\sum_{u \in U} (R_{u,j} - \overline{R}_j)^2}}$$

where $R_{u,\mathrm{i}}$ is the rating of user *u* on movie *i* and $\overline{R}_i$ is the average rating for movie *i*. For the distance-based method, the squared distance between two movies is calculated as follows:

$$dist(i,j) = \sum_{u \in U} (R_{u,i} - R_{u,j})^2$$

The distance is then transformed to a similarity measure, which lies in the range of [0;1]:

$$sim(i,j) = \frac{1}{1 + dist(i,j)}$$

For the calculation of predictions the weighted sum algorithm is used[9]. This method computes the prediction $p_{u,i}$ of a rating on an item *i* for a user *u* by computing the sum of the ratings given by the user on the items similar to *i*. Each rating is weighted by the corresponding similarity *sim(i,j)*. We adapt this method by restricting the number of similar movies to a sorted list of the N most similar movies (sorted by the absolute similarity):

$$p_{u,i} = \frac{\sum_{n=1}^{N} sim(i,j) R_{u,j}}{\sum_{j=1}^{N} \left| sim(i,j) \right|}$$

The optimal number of neighbors is determined on the basis of the validation sample.

Regression methods

We compare three different regression models, i.e. linear regression, logistic regression and ridge regression. As a benchmark, we use for each movie a simple linear regression model without parameter selection (denoted as LinReg (A)). The application of regression models with a large number of parameters (movies) will only yield reliable results when the number of observations (customers) is sufficient. In our analysis, most settings are characterized by problematic ratios of parameters to customers which typically leads to over-fitting. Therefore, the elimination of irrelevant parameters (model selection) is expected to play an important role in getting reliable recommendations. In the model selection phase, we determine those parameters which optimize the performance on the validation set. A classical backward model selection is computationally prohibitive due to the large number of settings. Therefore, we decided to calculate importance weights, $w_{i,j}$, for all dependent variables i and independent variables j (movies) on the basis of the following heuristic:

$$w_{i,j} = \left| r_{i,j} \right| * s_j * \left| b_{i,j} \right|$$

where $r_{i,j}$ denotes the correlation between ratings of movies i and j. s_j represents the standard deviation of the ratings of movie j over all customers who have rated i and j. $b_{i,j}$ is the initial parameter estimate obtained by LinReg (A) for movie j. According to this heuristic, movies get higher importance values with higher (absolute) correlation between the dependent variable and the independent variable, with a higher standard deviation and with higher (absolute) initial parameter estimates. Movies with lower importance weights are potential candidates for parameter elimination. Besides the full model (i.e., the model where no parameters were eliminated), we investigate only 3 other model sizes: *(a) J-*

*min(0.5*J, 0.05*C); (b) J-min(J-10, 0.2*C) and (c) Round(0.5*(a+b))* where *J* denotes the number of movies in the design and *C* the number of customers who have rated movie *i*. a) is a relatively large model, b) is rather sparse and c) lies in between. The choice of the final model size is based on the performance on the validation set. Our performance measures are then calculated on the generalization data set. The linear regression model with model selection is denoted by LinReg (B). In addition, we applied the same selection procedure with ridge regression (RidgeReg) and logistic regression (Logistic Reg), calculating model specific importance weights.

DESIGN OF THE STUDY

To analyze the effects of the number of customers and movies used for model estimation, we implement a full factorial design as shown in Table 2. We vary the number of customers between 1,000 and 50,000 and the number of movies in the range between 25 and 419. For all these combinations we estimate the users' ratings applying the 6 different methodologies (see Table 2).

Table 2: Design of the study

Factor	**Levels**
Customers	1000, 2000, 5000, 10000, 25000, 50000
Movies	25, 50, 150, 250, 350, 419
Methodology	Collaborative Filtering (A)
	Collaborative Filtering (B)
	Linear Regression (A)
	Linear Regression (B)
	Ridge Regression
	Logistic Regression

The movies used for each design are sampled randomly. Furthermore, we replicate each design, the number of replications depending on the number of movies employed. Since the standard deviations of the performance measures increase with a lower number of movies, we chose a higher number of replications for such settings. In total, we calculated the performance measures for 1224 different scenarios. For the evaluation of the results, we use four different performance measures calculated from the generalization data set:

1. MAE: mean absolute error between actual and predicted ratings. This measure is the most commonly used performance measure in this field of research.
2. RMSE: root mean squared error between actual and real ratings. This measure is more sensitive than the MAE to larger deviations from the actual ratings. Such deviations are problematic in Internet recommendation systems, since the customers may be disappointed and no longer make use of the recommendation engine.
3. R-square: squared correlation between model forecasts and real ratings. R-square is a frequently used measure for model comparison. As this measure has not been used in previous studies, we think that it may give some additional insights into the performance of recommendation systems.
4. Hit-rate: We calculate a matrix of actual versus predicted ratings (6x6) where one cell contains the probability that a person giving a specific rating gets exactly the same rating as a recommendation. As proposed by Ansari et al.[10], we use the perfect predictions and their nearest neighbors (±1) to calculate the hit-rate.

RESULTS

In a first step, we analyze the two classes of methodologies described, i.e. regression based-models and collaborative filtering.

Table 3: Mean and standard deviations over all designs for the performance measures

	MAE		RMSE		R-square		Hitrate	
	Mean	std	Mean	Std	mean	std	mean	std
CF (A)	0,92	0,03	1,238	0,03	0,13	0,03	80,1	1,7
CF (B)	0,93	0,03	1,245	0,03	0,11	0,04	79,8	1,8
LinReg (A)	1,04	0,21	1,41	0,28	0,11	0,06	77,0	5,5
LinReg (B)	0,94	0,06	1,27	0,08	0,13	0,05	79,5	2,4
Logistic Reg	1,13	0,13	1,56	0,16	0,11	0,05	72,6	3,3
Ridge Reg	1,04	0,29	1,45	0,40	0,10	0,05	80,6	2,2

Table 3 presents the results in terms of our four performance measures for the two categories of methods. It can be seen that the correlation-based approach (CF (A)) outperforms the distance-based approach (CF (B)) in terms of MAE, RMSE, R-square and hit-rate. All differences are significant at the 5% error level.

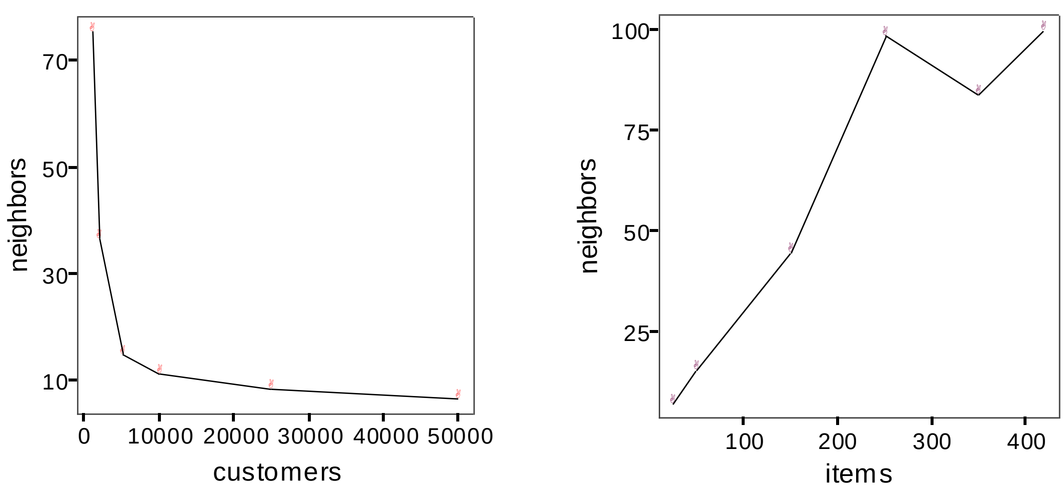

Figure 1 The optimal number of neighbors as a function of the number of customers (left hand side) and the number of items (right hand side) for CF (A).

Figure 1 shows the optimal number of neighbors for the CF (A) method as a function of the number of customers and movies in the design. Sarwar et al.[11] propose an optimal number of 80-120 neighbors for the MovieLens data set. Our study confirms this finding for the specific number of customers used in their work. However, Figure 1 shows that this only holds for this particular number of customers. Ceteris paribus, a higher number of customers or a lower number of movies leads to a lower optimal number of neighbors. Surprisingly, LinReg (B) significantly (α=0.01) outperforms all other regression methods in terms of MAE, RMSE, and R-square (Table 3). Only the hit-rate is highest for ridge regression. As a consequence of the above analysis, we restrict our presentation to CF (A) and LinReg (B). Furthermore, due to high correlation between MAE, hit-rate and RMSE (see Table 4), we restrict the evaluation of our analyses to R-square and MAE.

Table 4: Correlations between performance measures over all designs. All correlation coefficients are significant (α=0.01).

	R-square	RMSE	Hitrate	MAE
R-square	1,00	-0,39	0,39	-0,39
RMSE		1,00	-0,90	0,98
Hitrate			1,00	-0,93
MAE				1,00

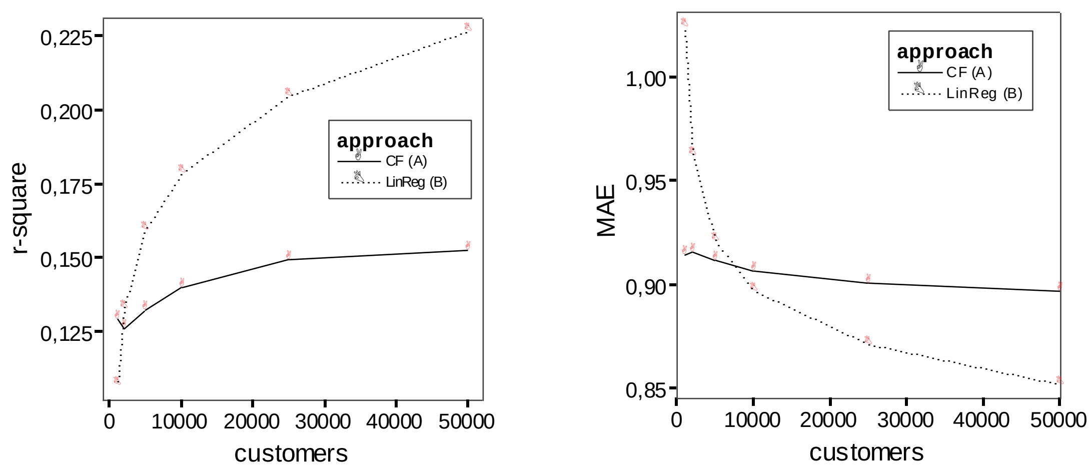

Figure 2 R-square (left hand side) and MAE (right hand side) as a function of customers for the case of 419 movies. The dashed line shows the mean R-square values for LinReg (B), the straight line represents mean R-square values for CF (A).

Figure 2 depicts the results for the maximum number of movies (419) in our design as a function of the number of customers in terms of R-square and MAE, respectively. For a low number of customers, collaborative filtering clearly performs better than linear regression. CF (A) shows a relatively stable performance for the entire range considered. In contrast to CF (A), recommendations generated by linear regression significantly improve as the number of customers increases. In terms of R-square (MAE), linear regression should be preferred to collaborative filtering when more than 2000 (6000) customers are in the database. Figure 2 indicates that for the regression model the performance of both measures could even be improved with a higher number of customers (>50000) than used in our study. From a managerial perspective, our findings justify the constant effort of enlarging customer databases. However, the marginal benefits of an increased customer database significantly depends on the methodology used. To estimate the effects of the method used, the number of customers and movies as well as their interactions, we formulated a simple linear model. The R-square between actual and predicted ratings acts as the dependent variable, whereas the method and the interaction between the method and the number of customers (log-transformed) and the interaction between the method, the number of customers and the number of items serve as independent variables (factors). Table 5 reflects the results of this analysis, confirming our graphical analysis. Positive (negative) coefficients indicate a higher (lower) accuracy of the predictions for a given factor. High absolute t-values for a factor indicate a significant impact on the dependent variable. Most interestingly, collaborative filtering does not significantly gain in performance with an increasing number of customers. Linear Regression, in contrast, significantly increases the performance with a higher number of customers. Since we chose more replications for designs with lower numbers of movies where collaborative filtering performs better, the coefficient for CF(A) in Table 5 is positive.

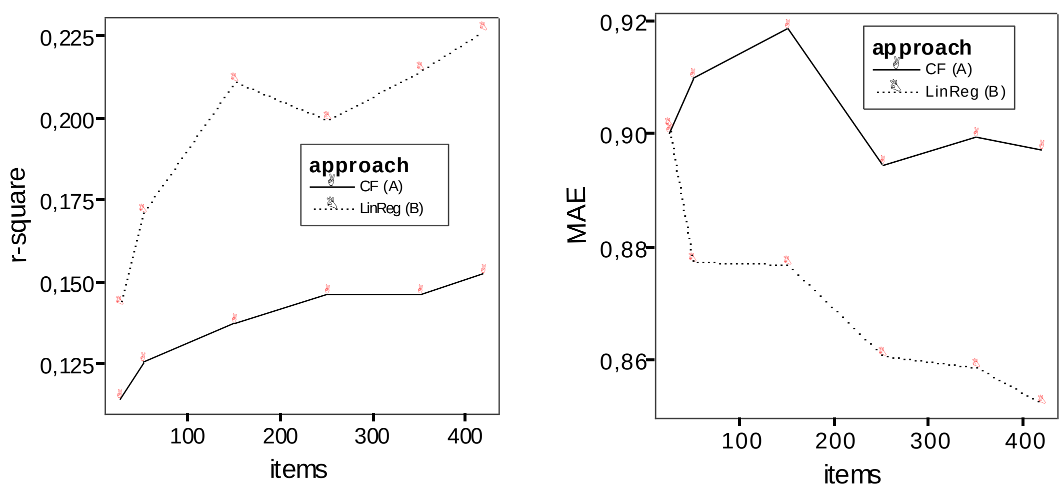

Figure 3 R-square (left hand side) and MAE (right hand side) as a function of the number of movies for the case of 50000 customers. The dashed line shows the mean R-square values for LinReg (B), the straight line represents mean R-square values for CF (A).

Figure 3 plots the R-square and MAE as a function of the number of movies used as independent variables in the designs. This figure illustrates the benefit of a higher number of movies, i.e., collaborative filtering and linear regression are able to improve their recommendations for larger assortments. Our model (Table 5) shows that this effect only arises when the number of customers considered for model estimation is high whereas for a lower number of customers it becomes insignificant.

For the case of very limited data on a specific movie, the use of additional demographic data and external expert ratings such as proposed by Ansari et al.[12] can help to provide users at least with some basic recommendations. Ansari et al. find that for their specific data set, simple linear regression performs almost as well as their proposed hierarchical Bayesian methodology. They argue that linear regression forecasts meet the average rating but do not explain any variance. Our results support this finding only for small data sets such as the ones used by the authors. Similarily, Good et al.[13] analyze the predictive ability of collaborative filtering and information filtering. Information filtering focuses on the analysis of item content and the development of a personal user interest profile. They find that the combination of both methods leads to the most useful recommendations.

Table 5: Model explaining R-square between actual and predicted ratings as dependent variable. ** (*) denotes parameters significant at α=0.01 (α=0.05).

Factor	Coefficient	t-value
Constant	-0,0175775	-0,87
CF(A)	0,1686730	5,92**
CF(A) * ln(customer)	-0,0032639	-1,45
LinReg (B) * LOG_C	0,0156319	6,96**
CF (A) * [customer=1000] * items	-0,0000085	-0,17
CF (A) * [customer=2000] * items	0,0000053	0,11
CF (A) * [customer=5000] * items	0,0000493	1,11
CF (A) * [customer=10000] * items	0,0000763	1,71
CF (A) * [customer=25000] * items	0,0000968	2,09*
CF (A) * [customer=50000] * items	0,0001025	2,10*
LinReg (B) * [customer=1000] * items	-0,0000239	-0,49
LinReg (B) * [customer=2000] * items	0,0000287	0,62
LinReg (B) * [customer=5000] * items	0,0000997	2,24*
LinReg (B) * [customer=10000] * items	0,0001443	3,24**
LinReg (B) * [customer=25000] * items	0,0001685	3,64**
LinReg (B) * [customer=50000] * items	0,0002039	4,18**

SUMMARY AND CONCLUSION

In this study, we investigate different approaches to recommendation systems using the publicly available EachMovie data set. In contrast to previous work on this data set, here a significantly larger subset was used. This allows us to investigate implications that were not identified before. In particular, we analyze the effects of the number of customers and movies as well as their interaction with different methods. We compare two commonly used collaborative filtering approaches to several regression models (linear regression, logistic regression, ridge regression). In an experimental full factorial design with replications (in total 1224 settings), we evaluate the quality of the recommendations in terms of the mean absolute error, the root mean squared error, R-square, and the hit-rate.
Among the collaborative filtering approaches, the correlation-based outperforms the distance-based one. Of the regression-based approaches, the linear regression is superior to its alternatives. However, model selection is a crucial factor of success, especially if the ratio between the number of observations (customers) and parameters (movies) is low. Collaborative filtering shows a satisfying performance if the number of customers available for model estimation is low.

All previous studies on collaborative filtering methods base their investigations on such small data sets. Runte[14], for instance, finds that for collaborative filtering methods a higher number of ratings does not lead to better recommendations. This is consistent with our findings. Our analysis indicates an insignificant impact of a higher number of customers on the performance of collaborative filtering methods. In contrast, we find that the number of ratings (customers) strongly influences the performance of regression-based methods. For a larger number of customers, we show that simple linear regression with model selection can provide significantly better recommendations in terms of all our measures.

Both collaborative filtering and linear regression are able to improve their recommendations in case of larger product assortments. However, this effect only arises when the number of customers considered for model estimation is high enough. From a managerial viewpoint, our findings justify the constant effort of enlarging customer databases for recommendation systems. However, the marginal benefits of increased customer databases significantly depend on the method used. Our analysis suggests that in the early phase of the life-cycle of a recommendation system - when there are relatively few customers - collaborative filtering can be used. In later stages, when the customer database has grown, linear regression is the method to be preferred.

This study was carried out on the basis of movie ratings and the question arises whether the conclusions can be generalized to other applications like book or CD recommendation systems. We expect that the identified characteristics remain the same. However, since the number of available books or CDs typically is much higher as compared to movies, the data will probably be even more sparse. Therefore, we assume that a higher number of customers will be necessary for preferring regression models over collaborative filtering. As our study is limited, several ideas for future research can be suggested: Given that this works is based on a small subset of methods applicable for recommendation systems, we feel that also the performance of other methods depends on the amount of data used. If one considers segment specific models, for instance, it would be interesting to study the trade-off between the disadvantage of having fewer customers per segment and the advantage arising from segment-specific recommendations.

Acknowledgement

We would like to express our gratitude to Kerry Barner and two anonymous referees for the valuable comments and suggestions on a previous version of this paper. The data set for this paper was generously provided by Compaq Equipment Corporation.

References:

1 Negroponte, N. (1970) 'The Architecture Machine', Boston, MIT Press.
2 Resnick, P., Varian, H. (1997) 'Recommender Systems', Communications of the ACM, 40 (3), pp. 56-58.
3 Ansari, A., Essegaier, S., Kohli, R. (2000) 'Internet Recommendations Systems', Journal of Marketing Research, (August) pp. 363-375.
4 Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J. (2001) 'Item-based Collaborative Filtering Recommender Algorithms', Proceedings of The WWW10 Conference.
5 Breese, J. S., Heckerman, D., Kadie, C. (1998) 'Empirical Analysis of Predictive Algorithms for Collaborative Filtering', Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence, pp. 43-52.
6 Chen, Y., George, E. (2000) 'A Bayesian Model for Collaborative Filtering', Technical Report, Statistics Department, University of Texas at Austin.
7 Shardanand, U., Maes, P. (1995) 'Social information filtering: algorithms for automating "word of mouth"', Proceedings of the Conference on Human Factors in Computing Systems (CHI'95), Denver, CO, ACM, pp. 210-217.
8 Runte, M. (2000) 'Personalisierung im Internet - Individualisierte Angebote mit Collaborative Filtering', DUV, Wiesbaden.
9 Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J. (2000) 'Analysis of Recommender Algorithms for E-Commerce', Proceedings of the 2nd ACM E-Commerce Conference (EC'00).
10 Ansari, A., Essegaier, S., Kohli, R. (2000) 'Internet Recommendations Systems', Journal of Marketing Research, (August) pp. 363-375.
11 Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J. (2000) 'Analysis of Recommender Algorithms for E-Commerce', Proceedings of the 2nd ACM E-Commerce Conference (EC'00).
12 Ansari, A., Essegaier, S., Kohli, R. (2000) 'Internet Recommendations Systems', Journal of Marketing Research, (August) pp. 363-375.
13 Good, N., Schafer, J., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., Riedl, J. (1999) 'Combining collaborative filtering with personal agents for better recommendations', Proceedings of the Sixteenth National Conference on Artificial Intelligence.

14 Runte, M. (2000) ‘Personalisierung im Internet - Individualisierte Angebote mit Collaborative Filtering’, DUV, Wiesbaden.