Academic Papers

Collaborative filtering or regression models for Internet recommendation systems?

Received (in revised form): 21st January, 2002

Andreas Mild
is Assistant Professor at the Wirtschaftsuniversität Wien, Austria. He holds a PhD in Business Administration. His previous work has been published in *Lecture Notes in Artifical Intelligence*, *MIS Quarterly* and *Management Science*. His major research focuses on knowledge discovery techniques, agent-based simulation and new product development.

Martin Natter
is Assistant Professor at the Wirtschaftsuniversität Wien, Austria. His previous research has appeared in *Marketing Letters*, *Management Science*, *Journal of Retailing and Consumer Services*, *International Journal of Production Economics*, *European Journal of Operational Research* and others. His research interest lies in the field of basket analysis and new product development.

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. The authors 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. Two commonly used collaborative filtering approaches are compared with several regression models using an experimental full factorial design. According to the findings, the number of customers significantly influences the performance of all approaches under study. For a large number of customers and movies, it is shown 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.

Andreas Mild
Department of Production
Management, Vienna
University of Economics
and Business
Administration,
Pappenheimgasse 35/3/5,
A-1200 Vienna, Austria.

Tel: +43 1 31336/5628;
Fax: +43 1 31336/905613;
e-mail: andreas.mild@
wu-wien.ac.at

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. For many product categories such as books, compact discs or movies, however, variety seeking plays an important role in choice decisions; ie simple search functions are not sufficient to support a customer's search process.

Table 1 Design of previous studies on recommendation systems

Study	Customers	Movies	Percentage of ratings used %
Ansari *et al.*[3]	2,000	340	2.0
Breese *et al.*[5]	4,119	1,623	6.8
Chen and George[6]	1,373	41	0.05
Runte[8]	1,995	683	3.7
Present study	61,007	419	75.2

Recommendation systems[1] endeavour 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 the customer's purchase history to determine the preference structure and identify products that the 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, it is assumed that a good way to predict the products of interest to a customer is to look at other people who show similar behaviour.[2] Besides reducing 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 Bayesian 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 unpersonalised recommendations (item-specific averages). He finds that distance-based methods outperform correlation-based predictions which, in turn, perform better than non-personalised recommendations. The literature mentioned shows that various approaches have been proposed and compared. Several contributions use the mean absolute error as a performance measure. In the authors' opinion, however, 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, they 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 shows that previous studies only use a small fraction of the data available. It is hypothesised 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. Collaborative filtering approaches are benchmarked against several variants of multivariate regression analysis. The 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, it is necessary to consider larger portions of the EachMovie database than those dealt with in the mentioned studies. The data analysed in this work represents more than 75 per cent 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, 72,916 users entered a total of 2,811,983 numeric ratings for 1,628 different movies (films and videos). This data set was made available to researchers for testing new algorithms. The movies are rated on a six-point scale. From the 1,628 movies many have very few ratings. Due to computational restrictions, the full data set could not be used. From the original database, users who rated more than three movies were selected. This is equivalent to real-world systems which refuse recommendations before a minimum number of ratings is delivered. From this selection, the most relevant movies in terms of the number of ratings were picked. By selecting movies with more than 50 ratings in each of the samples, a manageable data set size was finally arrived at (199.07 MB in MATLAB format). Although the reduced data set contains about 75 per cent of all ratings and 84 per cent of all users, use is only made of about 26 per cent of all movies. Due, however, to the design of the study (see Table 2), situations where only a very limited number of ratings for a specific movie or customer is available, are also analysed. The remaining data set consists of 61,007 users and 419 movies. The set of available customers was split 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 neighbours (collaborative filtering) or stepwise parameter selection (regression models)
- a generalisation sample consisting of 6,007 randomly selected customers, this data set serves for performance measurement.

In the following subsections, the models used for generating movie recommendations are described. The task of a recommendation system is to predict a movie's rating for a specific customer (dependent variable) based on their weighted ratings on other movies (independent variables). The models differ in the way the weights are calculated. All models, however, 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 six-point scale.

Collaborative filtering

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 are considered. 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:

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

where $R_{u,i}$ 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:

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

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 $sim(i,j)$. This method is adapted by restricting the number of similar movies to a sorted list of the N most similar movies (sorted by the absolute similarity):

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

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

Regression methods

Three different regression models are compared, ie linear regression, logistic regression and ridge regression. As a benchmark, a simple linear regression model without parameter selection (denoted as LinReg (A)) is used for each movie. 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 this analysis, most settings are characterised 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, those parameters which optimise the performance on the validation set are determined. A classical backward model selection is computationally prohibitive due to the large number of settings. Therefore, it was 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:

$$w_{i,j} = |r_{i,j}| \star s_j \star |b_{i,j}|$$

Table 2: Design of the study

Factor	Levels
Customers	1,000, 2,000, 5,000, 10,000, 25,000, 50,000
Movies	25, 50, 150, 250, 350, 419
Methodology	Collaborative filtering (A) Collaborative filtering (B) Linear regression (A) Linear regression (B) Ridge regression Logistic regression

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 (ie the model where no parameters were eliminated), only three other model sizes are investigaged: (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. The performance measures are then calculated on the generalisation data set. The linear regression model with model selection is denoted by LinReg (B). In addition, the same selection procedure was applied to ridge regression (RidgeReg) and logistic regression (Logistic Reg), calculating model specific importance weights.

DESIGN OF THE STUDY

To analyse the effects of the number of customers and movies used for model estimation, a full factorial design as shown in Table 2 was implemented. The number of customers was varied between 1,000 and 50,000 and the number of movies in the range between 25 and 419. For all these combinations the users' ratings were estimated applying the six different methodologies (see Table 2).

The movies used for each design were sampled randomly. Furthermore, each design was replicated, 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, a higher number of replications for such settings was chosen. In total, the performance measures for 1,224 different scenarios were calculated. For the evaluation of the results, four different performance measures calculated from the generalisation data set were used:

— MAE: mean absolute error between actual and predicted ratings. This measure is the most commonly used performance measure in this field of research

— 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

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

	MAE mean	std	RMSE mean	std	R-square mean	std	Hit-rate 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

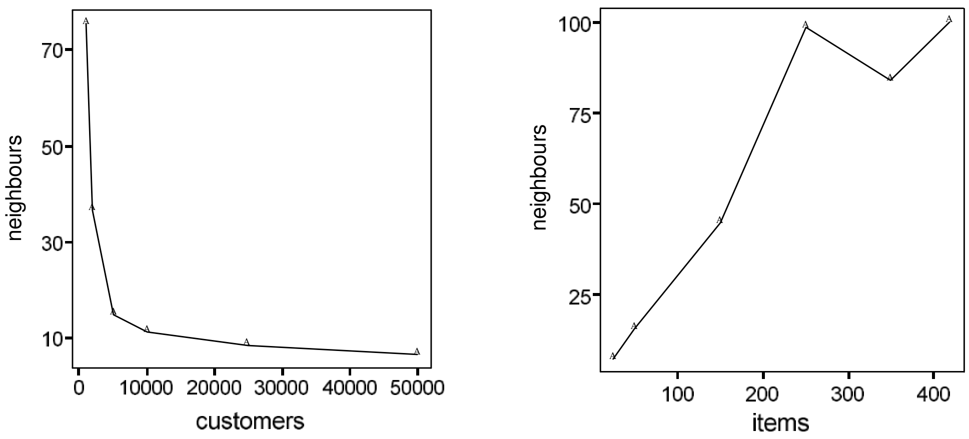

Figure 1 The optimal number of neighbours as a function of the number of customers (left-hand side) and the number of items (right-hand side) for CF (A)

problematic in Internet recommendation systems, since the customers may be disappointed and no longer make use of the recommendation engine

— 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, it may give some additional insights into the performance of recommendation systems

— hit-rate: a matrix of actual versus predicted ratings (6 × 6), where one cell contains the probability that a person giving a specific rating gets exactly the same rating as a recommendation, is calculated. As proposed by Ansari *et al.*,[10] the perfect predictions and their nearest neighbours (± 1) are used to calculate the hit-rate.

RESULTS

In a first step, the two classes of methodologies described are analysed, ie regression based-models and collaborative filtering.

Table 3 presents the results in terms of the 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 per cent error level.

Figure 1 shows the optimal number of neighbours for the CF (A) method as a

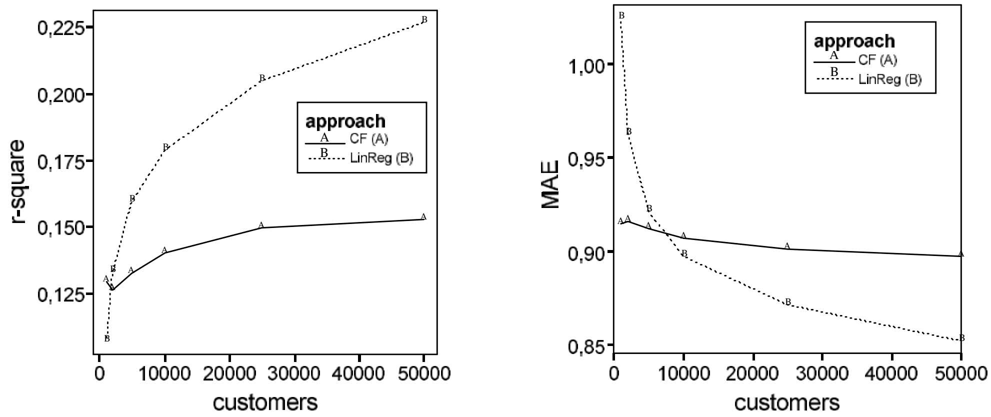

Figure 2 R-square (left-hand side) and MAE (right-hand side) as a function of customers for the case of 419 movies

Note: The dashed line shows the mean R-square values for LinReg (B), the straight line represents mean R-square values for CF (A).

function of the number of customers and movies in the design. Sarwar *et al.*[11] propose an optimal number of 80–120 neighbours for the MovieLens data set. This study confirms this finding for the specific number of customers used in their work. Figure 1 shows, however, that this only holds for this particular number of customers. Other things being equal, a higher number of customers or a lower number of movies leads to a lower optimal number of neighbours. Surprisingly, LinReg (B) significantly ($\alpha = 0.01$) 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, presentation is restricted to CF (A) and LinReg (B). Furthermore, due to high correlation between MAE, hit-rate and RMSE (see Table 4), the evaluation of the analyses is restricted to R-square and MAE.

Table 4: Correlations between performance measures over all designs

	R-square	RMSE	Hit-rate	MAE
R-square	1.00	−0.39	0.39	−0.39
RMSE			−0.90	0.98
Hit-rate			1.00	−0.93
MAE				1.00

All correlation coefficients are significant ($\alpha = 0.01$)

Figure 2 depicts the results for the maximum number of movies (419) in the 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 2,000 (6,000) 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 (> 50,000) than used in the study. From a managerial perspective, these findings justify the constant effort of enlarging customer databases. The marginal benefits of an increased customer database, however, significantly depend on the methodology used. To estimate the effects of the

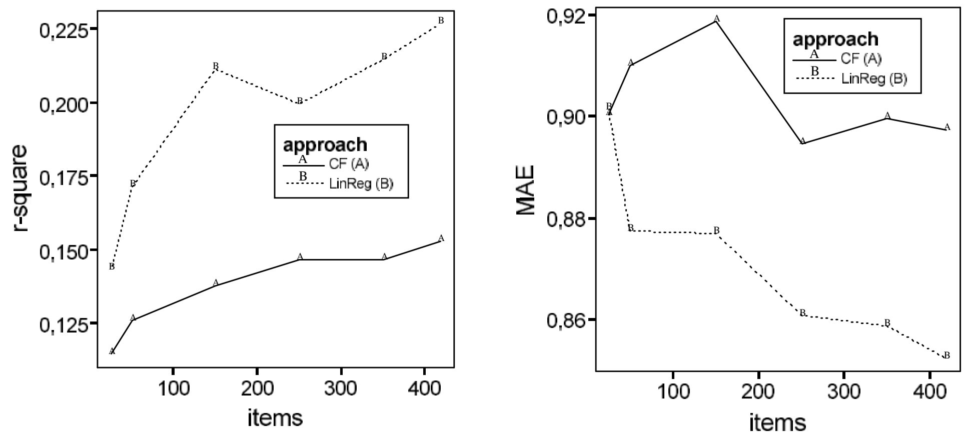

Figure 3 R-square (left-hand side) and MAE (right-hand side) as a function of the number of movies for the case of 50,000 customers

Note: The dashed line shows the mean R-square values for LinReg (B), the straight line represents mean R-square values for CF (A).

method used, the number of customers and movies as well as their interactions, a simple linear model was formulated. 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 the 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 more replications for designs with lower numbers of movies where collaborative filtering performs better were chosen, the coefficient for CF(A) in Table 5 is positive.

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, ie collaborative filtering and linear regression are able to improve their recommendations for larger assortments. The 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. The present results support this finding only for small data sets such as the ones used by the authors. Similarly,

Table 5: Model explaining R-square between actual and predicted ratings as dependent variable

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**

**(*) denotes parameters significant at $\alpha = 0.01$ ($\alpha = 0.05$)

Good *et al.*[13] analyse 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.

SUMMARY AND CONCLUSION

In this study, the authors 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 the authors to investigate implications that were not identified before. In particular, they analyse the effects of the number of customers and movies as well as their interaction with different methods. Two commonly used collaborative filtering approaches are compared to several regression models (linear regression, logistic regression, ridge regression). In an experimental full factorial design with replications (in total 1,224 settings), the authors 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 approach outperforms the distance-based one. Of the regression-based approaches, the linear regression one is superior to its alternatives. Model selection, however, 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 the findings here. The present analysis indicates an insignificant impact of a higher number of customers on the performance of collaborative filtering methods. In contrast, it is found that the number of ratings (customers) strongly influences the performance of regression-based methods. For a larger number of customers, it is shown that simple linear regression with model selection can provide significantly better recommendations in terms of all the measures.

Both collaborative filtering and linear regression are able to improve their recommendations in case of larger product assortments. This effect, however, only arises when the number of customers considered for model estimation is high enough. From a managerial viewpoint, the findings justify the constant effort of enlarging customer databases for recommendation systems. The marginal benefits of increased customer databases, however, significantly depend on the method used. The 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 generalised to other applications like book or CD recommendation systems. The authors expect that the identified characteristics remain the same. Since, however, the number of available books or CDs typically is much higher as compared to movies, the data will probably be even more sparse. Therefore, it is assumed that a higher number of customers will be necessary for preferring regression models over collaborative filtering. As this study is limited, several ideas for future research can be suggested. Given that this work is based on a small subset of methods applicable for recommendation systems, the performance of other methods may also depend on the amount of data used. If segment specific models are considered, 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.

Acknowledgment

The authors would like to express their 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', MIT Press, Boston.

2 Resnick, P. and Varian, H. (1997) 'Recommender systems', *Communications of the ACM*, Vol. 40, No. 3, pp. 56–58.

3 Ansari, A., Essegaier, S. and Kohli, R. (2000) 'Internet recommendations systems', *Journal of Marketing Research*, August, pp. 363–375.

4 Sarwar, B. M., Karypis, G., Konstan, J. A. and Riedl, J. (2001) 'Item-based collaborative filtering recommender algorithms', Proceedings of The WWW10 Conference.

5 Breese, J. S., Heckerman, D. and 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. and George, E. (2000) 'A Bayesian model for collaborative filtering', Technical Report, Statistics Department, University of Texas at Austin.

7 Shardanand, U. and 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. and Riedl, J. (2000) 'Analysis of recommender algorithms for e-commerce', Proceedings of the 2nd ACM E-Commerce Conference (EC'00).

10 Ansari, *et al.* (2000) *op cit.*

11 Sarwar *et al.* (2000) *op. cit.*

12 Ansari *et al.* (2000) *op. cit.*

13 Good, N., Schafer, J., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J. and Riedl, J. (1999) 'Combining collaborative filtering with personal agents for better recommendations', Proceedings of the Sixteenth National Conference on Artificial Intelligence.

14 Runte (2000) *op. cit.*