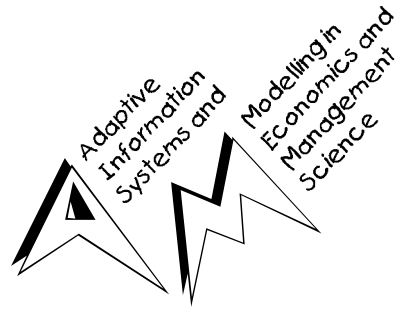


Working Paper Series

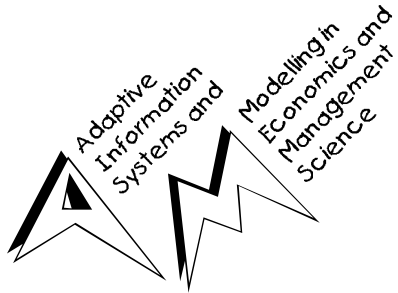


A Critical View on Recommendation Systems

Andreas Mild
Martin Natter

Working Paper No. 82
July 2001

Working Paper Series



July 2001

SFB
'Adaptive Information Systems
and Modelling in Economics and Management Science'

Vienna University of Economics and Business Administration
Augasse 2-6
1090 Vienna
Austria

<http://www.wu-wien.ac.at/am>

This piece of research was supported by the
Austrian Science Foundation (FWF) under grant SFB # 010
(‘Adaptive Information Systems and Modelling in Economics and
Management Science’).

Mild, Andreas, Natter, Martin

A Critical View on 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. In contrast to previous work on this data set, here a significantly higher subset is used. The effects caused by the 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.

1. 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 customer search process. Recommendation systems (Negroponte 1970) 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 (Resnick and Varian 1997). 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 (Ansari et al. 2000). Practical implementations of such systems can be found at Amazon.com (books, CDs) or MovieCritic.com (movies).

In the literature, different approaches to recommendation systems have been studied. Sarwar et al. (2000) compare collaborative filtering systems based on similarities between users to methods which consider similarities between items. They show that the item-based approach is preferable in terms of recommendation quality and computational effort. Breese et al. (1998) find that Bayesian networks with decision trees at each node and correlation methods outperform Bayesian clustering and vector-similarity methods. Chen and George (2000) compare several Bayesian models to the original collaborative filtering approach proposed by Shardanand and Maes (1995) and find that their approach performs better. Runte (2000) 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 unpersonalized recommendations.

In contrast to collaborative filtering systems, Ansari et al. (2000) propose a hierarchical Bayesian methodology which makes use of additional demographic data and external expert ratings. They find that for their specific (small) data set, simple linear regression performs almost as well as their proposed hierarchical Bayesian methodology. However, they argue that linear regression forecasts meet the average rating but do not explain any variance. Good et al. (1999) 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.

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 subsets of the original data sets. 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.

Authors	Customers	Movies	Percentage of ratings used
Ansari et al. (2000)	2000	340	2,0 %
Breese et al. (1998)	4119	1623	6,8 %
Chen & George (1999)	1373	41	0,05%
Runte (2000)	1995	683	3,7 %
Present study	61007	419	75,2 %

Table 1: Design of previous studies on recommender systems

Table 1 shows that previous studies use only 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 in practice most relevant case where a recommendation system is used to predict ratings for new users of a given set of films.

If these effects prove to be significant, interesting methodological and managerial implications can be derived. 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. However, the usage of the entire data set is not advisable as it contains users who have rated hardly any films (<3) and movies which were seldom (<50) rated. 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.

2. 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, some 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. Consequently, we restrict our study to movies which have more than 50 ratings and to users who rated more than 3 movies. The remaining data set consists of 61007 users and 419 movies. We split the set of available customers into three groups, i.e.:

- a training sample consisting of 50,000 randomly selected customers, this data set is used for model estimation
- a validation sample consisting of 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

Collaborative Filtering:

For the calculation of the similarities between items we use two different methods, i.e., 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:

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

where $R_{u,i}$ is the voting of user u on movie i and \bar{R}_i is the average voting 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 (Sarwar et al. 2000). 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):

$$p_{u,i} = \frac{\sum_n sim(i,n) R_{u,n}}{\sum_n |sim(i,n)|}$$

The predictions are then transformed into discrete ratings on a 6-point scale. The optimal number of neighbors is determined on the basis of the validation sample.

Regression methods

We compare the performance of 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 ratings over all available customers in the training sample serve as the dependent variable for each movie. The ratings for all movies except the one under consideration serve as independent variables. After estimation of the model parameters, recommendations from the regression models can be received by transforming the predictions into discrete ratings on a 6 point scale.

As in our analysis most settings are characterized by ratios of customers per movie which typically leads to over-fitting, model selection is expected to play an important role in getting good recommendations. In the model selection phase, we try to find those movies which optimize the performance on the validation set. Since in view of the large number of settings a classical backward model selection is computationally prohibitive, 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:

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

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, we investigate only 3 other model sizes:

- a) $J - \min(0.5 * J, 0.05 * C)$
- b) $J - \min(J - 10, 0.2 * C)$
- c) $\text{Round}(0.5 * (a + b))$

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 a ridge regression (RidgeReg) and a logistic regression (Logistic Reg), performing a model specific calculation of the importance weights.

3. 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). 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.

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

Table 2: Design of the study

For the evaluation of the results, we use four different performance measures calculated from the generalization data set, i.e.:

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 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. (2000), we use the perfect predictions and their nearest neighbors (± 1) to calculate the hit-rate.

4. Results

In a first step, we separately analyze the two classes of methodologies described, i.e. regression based-models and collaborative filtering. Table 3 and Table 4 present the results in terms of our four performance measures for the two categories of methods. From Table 3 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 and Figure 2 show 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. (2000) 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 (lower) number of customers (movies) leads to a lower optimal number of neighbors.

	MAE	MAE	RMSE	RMSE	r-sq	r-sq	Hitrate	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

Table 3: Mean and standard deviations over all designs for the performance measures of the collaborative filtering methods

	MAE	MAE	RMSE	RMSE	r-sq	r-sq	Hirate	hitrate
	mean	std	Mean	std	mean	std	Mean	std
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 4: Mean and standard deviations over all designs for the performance measures of the regression methods

Table 4 contains the results for the regression-based methods. Surprisingly, LinReg (B) significantly ($\alpha=0.01$) outperforms all other methods in terms of MAE, RMSE, and r-square. Only the hit-rate is highest for ridge regression.

For the main comparison between regression and collaborative filtering, we restrict our analysis to CF (A) and LinReg(B). Furthermore, due to high correlation between MAE, hit-rate and RMSE (see Table 5), we restrict the evaluation of our analyses to r-square and MAE.

	r-square	RMSE	Hirate	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

Table 5: Correlations between performance measures over all designs. All correlations are significant ($\alpha=0.01$).

Figure 3 and Figure 4 depict 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. For 2000 (6000) customers or more, linear regression should be preferred to collaborative filtering in terms of r-square (MAE). Figure 3 and Figure 4 indicate 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 from a recommendation systems perspective. However, the marginal benefits of increased customer databases significantly depend 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. R-square 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. Table 6 reflects the results of this analysis, confirming our graphical analysis. Most interestingly, from an increasing number of customers

collaborative filtering does not significantly gain in performance. 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 6 is positive.

Figure 5 and Figure 6 plot the r-square and MAE as a function of the number of movies used as independent variables in the designs. Both figures illustrate the benefit of a higher number of products, i.e., collaborative filtering and linear regression are able to improve their recommendations for larger assortments. Our model (Table 6) 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.

	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**

Table 6: Model explaining r-square as dependent variable. ** (*) denotes parameters significant at $\alpha=0.01$ ($\alpha=0.05$).

5. 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 higher subset was used here. 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 customers and 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 (2000), 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.

As our 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, we feel that also the performance of other methods depends on the amount of data used. It would certainly be interesting to see studies that analyze the trade-off between the disadvantage of having fewer customers and the advantage arising from segment-specific recommendations.

6. References

Ansari, A., Essegai, S., Kohli, R., (2000), Internet Recommendations Systems, *Journal of Marketing Research*, (August) 363-375.

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*, 43-52.

Chen, Y., George, E., (2000), A Bayesian Model for Collaborative Filtering, Technical Report, Statistics Department, University of Texas at Austin.

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*.

Maes, P. (1995), Agents that Reduce Work and Information Overload, Readings in Human-Computer Interaction, Toward the Year 2000, Baecker, Grudin and Buxton, editors, Morgan Kaufman.

Negroponte, N. , (1970), The Architecture Machine, Boston, MIT Press.

Resnick, P., Varian, H., (1997), Recommender Systems, Communications of the ACM, 40 (3), 56-58.

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

Sarwar, B. M., Karypis, G., Konstan, J. A., Riedl, J., (2001), Item-based Collaborative Filtering Recommender Algorithms, Proceedings of The WWW10 Conference. May, 2001.

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).

Shardanand, U., Maes, P., (1995), Social information filtering: algorithms for automating "word of mouth", Proceedings of the Conference on Human Factors in Computing Systems (CHI95), Denver, CO, ACM, 210-217.

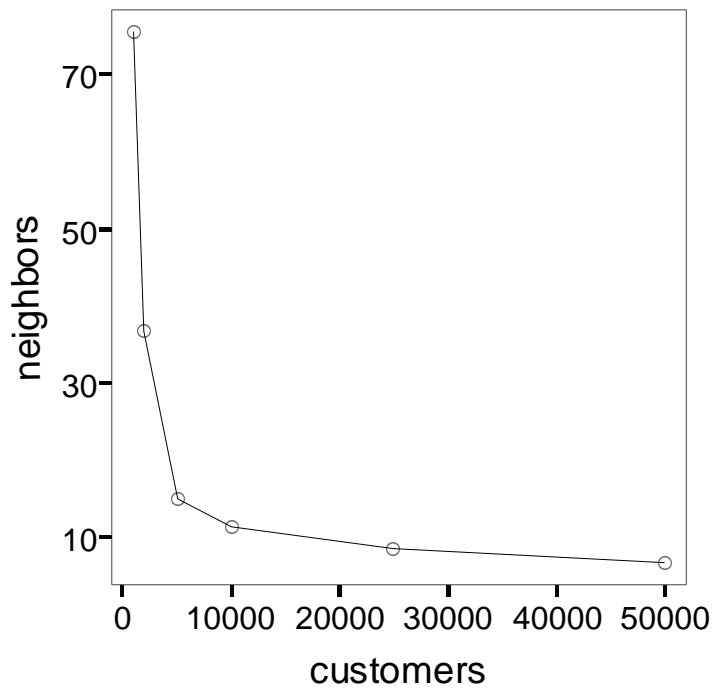


Figure 1: The optimal number of neighbors as a function of the number of customers for CF (A).

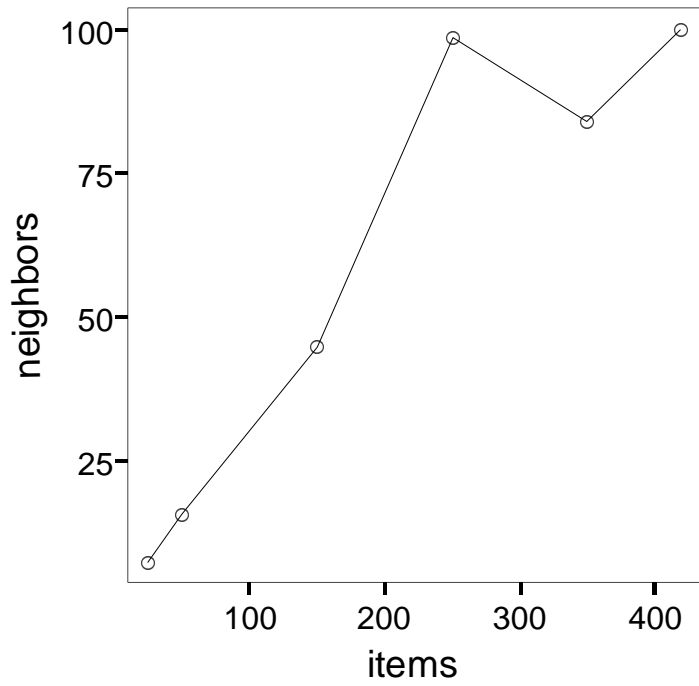


Figure 2: The optimal number of neighbors as a function of the number of movies (items) for CF (A).

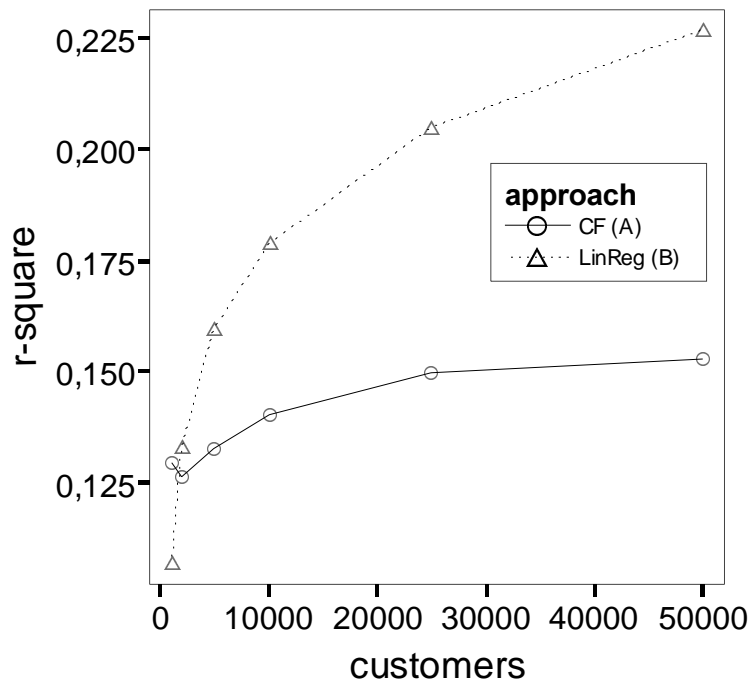


Figure 3: r-square as a function of customers for the case of 419 movies. The dashed line shows the mean r-square values for LinReg (B), the straight line represents mean r-square values for CF (A).

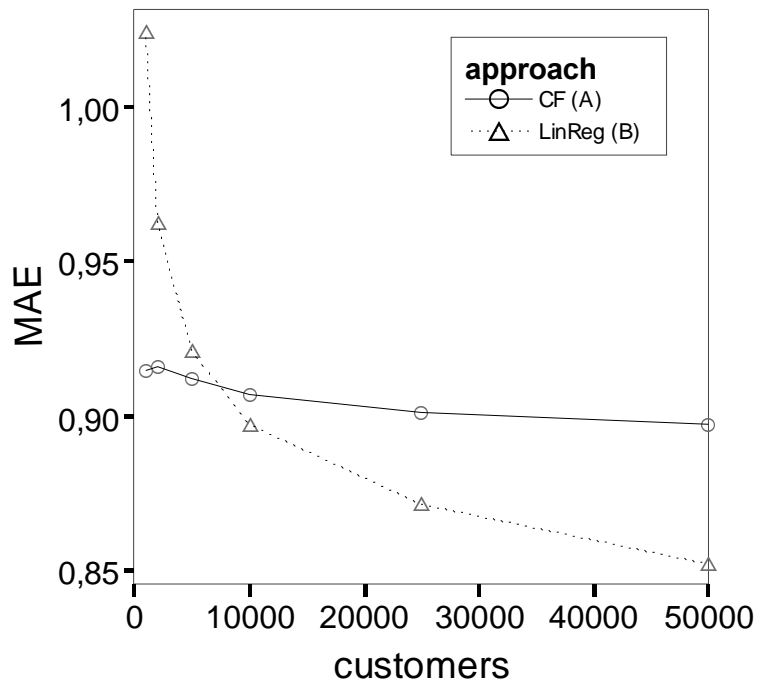


Figure 4: Mean absolute error (MAE) as a function of customers for the case of 419 movies. The dashed line shows the mean MAE values for LinReg (B), the straight line represents mean MAE values for CF (A).

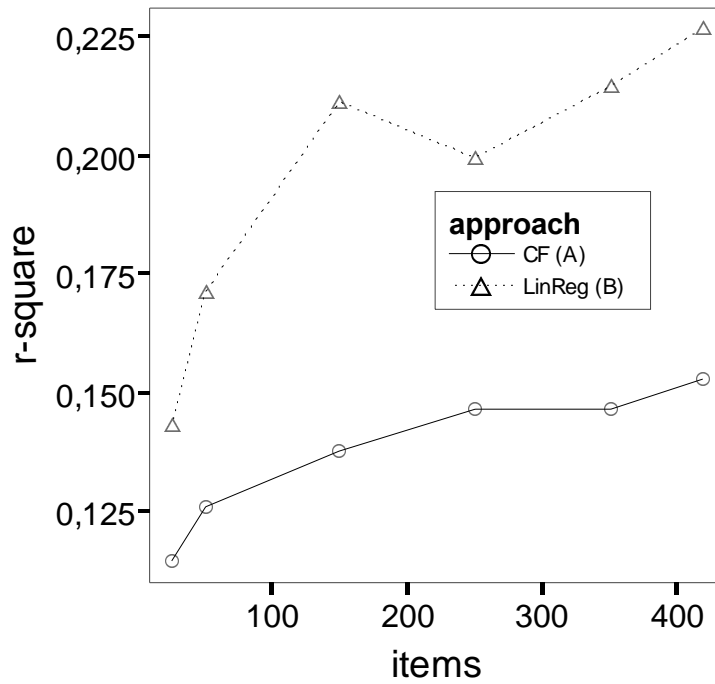


Figure 5: r-square as a function of the number of movies for the case of 50000 customers. The dashed line shows the mean r-square values for LinReg (B), the straight line represents mean r-square values for CF (A).

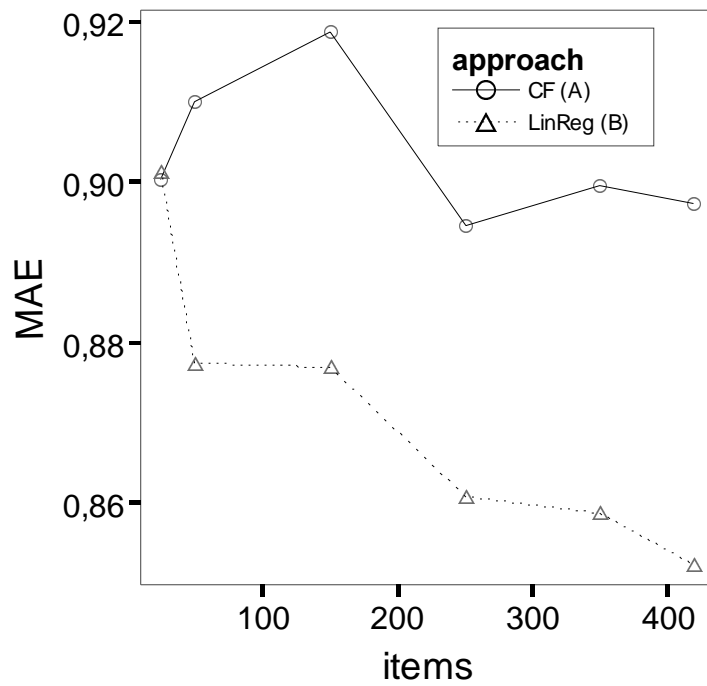


Figure 6: Mean absolute error (MAE) as a function of the number of movies (items) for the case of 50000 customers. The dashed line shows the mean MAE values for LinReg (B), the straight line represents mean MAE values for CF (A).