“How to predict preferences for new items”

AUTHORS	Volker Schlecht

ARTICLE INFO	Volker Schlecht (2008). How to predict preferences for new items. *Investment Management and Financial Innovations*, *5*(4)
RELEASED ON	Friday, 28 November 2008
JOURNAL	"Investment Management and Financial Innovations"
FOUNDER	LLC “Consulting Publishing Company “Business Perspectives”

NUMBER OF REFERENCES
0

NUMBER OF FIGURES
0

NUMBER OF TABLES
0

Volker Schlecht (Germany)

How to predict preferences for new items

Abstract

Huge amounts of data and lots of competing methods for estimating the usefulness of a certain known item to a specific user exist. However, most of these procedures only work well if the items are already well-known. Nevertheless, the users of a recommender-system might be more interested in receiving recommendations for items, which they have never heard of before than to keep getting items recommended that they have already been told or read about numerous times before. Also, from a marketer's point of view the preferences for new or even hypothetical items are more important. E.g., such information might be useful in deciding, whether a particular new item should be added to the product portfolio of an online store. A number of different techniques for estimating the preferences for new items are introduced and their performance is evaluated and compared with respect to the different purposes of preference estimation. A combination of two-mode clustering and Hierarchical Bayes regression is shown to be a good and highly interpretable estimation method. A quick heuristic procedure is developed, by which more useful recommendations with respect to new items can be generated.

Keywords: CRM, targeting, online-marketing, predictive modeling, customer insights, two-mode segmentation, strategic marketing, Bayesian statistics, hierarchical Bayes approach, applied econometrics, recommender-systems, customer centricity.
JEL Classification: C02, M21, M31, M41.

Introduction

Various online-based businesses supply their customers with recommendations based on automated collaborative filtering in order to increase customer satisfaction and customer retention. Therefore the goal of any recommender system is to provide recommendations, which are perceived as helpful by the customers. Every recommender-system is based on a quantitative procedure to estimate the utility of a certain item to a specific user for given information about his or her past (rating) behavior (Adomavicius and Tuzhilin, 2005). Approaches exist, that additionally utilize the user's demographic data and the properties of the item he or she has supplied. Since the kind of items the user is looking for may change quickly, it is necessary to include as much of the latest developments as possible. So the algorithms for the estimation of the utility have to be both quick and (reasonably) accurate.

Recently several collaborative filtering procedures were proposed which are based on two-mode clustering (Schlecht and Gaul, 2004; George and Merugu, 2005; Banerjee et al., 2005). All of those procedures were shown to outperform procedures which are based on the Bravais-Pearson correlation in terms of accuracy. It has been demonstrated, that one of those methods for two-mode clustering can also achieve results which are comparable to other competing techniques for collaborative filtering (namely a SVD-based approach and non-negative matrix factorization), but require fewer parameters, less training time and decrease the average time required for the actual estimation drastically (George and Merugu, 2005).

Nonetheless, so far two-mode clustering shares a shortcoming of all procedures for automated collaborative filtering: Until an item is rated by a substantial number of users, any automated collaborative filtering based recommender-system is unable to recommend it. Unfortunately, those recommendations are the most interesting ones for the customer and also the corresponding estimates might be the most useful ones for marketers.

Whether the user likes it or not he keeps getting recommendations for already well-known items from friends, colleagues, neighbors, short-term acquaintances, business associates – in short everyone he or she meets. Some of them might even be more helpful and better suited to the user's taste than those provided by an automated recommender-system. So any recommendation for an already well-known item might be a recommendation that actually has been given by several people before. This might still be helpful, but it would be far more interesting for the user to receive recommendations for items, which are new or less-well known, because those are the items which the user might not have heard of before or might not get any recommendation for from other sources than the recommender system.

Moreover, it is vital to managers and marketers of online-stores to approximate which and how many people might be interested in a certain product before it is introduced in the shop (or even in general). Hence, it would be very useful, if the existing data and estimation procedures could be used for the extrapolation of the utility of new items for any known user.

An already known alternative to automated collaborative filtering is a linear hierarchical Bayes

regression model, which uses additional information like certain properties of users (e.g., age and gender) and movies (for example, genre) as independent variables (Ansari et al., 2000).

The goal of this paper is to develop, apply and compare different procedures for the approximation of the utility of a certain unknown item to a known user based on the past ratings of known users for known items.

1. Two-mode clustering

Let s_{ij} be the rating of user $i \in \{1,...,I\}$ for item $j \in \{1,...,J\}$ and let I be the number of users, and J be the number of items. The corresponding matrix (s_{ij}) is a two-mode data matrix, which depicts the interaction between first mode elements (users) and second mode elements (items). There are different approaches to two-mode clustering (e.g., DeSarbo, 1982; Noma and Smith, 1985; Espejo and Gaul, 1986; DeSarbo et al., 1988). An important part of these approaches to two-mode clustering are generalizations of the ADCLUS model proposed by Shepard and Arabie (1979). Well-known generalizations of the ADCLUS model are the GENNCLUS (DeSarbo, 1982) and PENNCLUS (Both and Gaul, 1987) model. Let $k \in \{1,...,K\}$ ($l \in \{1,...,L\}$) be the index of the first (second) mode clusters and let $P = (p_{ik})$ ($Q = (q_{jl})$) be the matrix which describes the cluster-membership of the first (second) mode elements with

$$p_{ik}(q_{jl}) = \begin{cases} 1, & if\ i\ (j)\ belongs\ to\ the\ first\ (second)\ mode\ cluster\ k\ (l) \\ 0, & otherwise \end{cases}.$$

Furthermore, let V_{ij} equal one if user i has rated item j and zero if the rating of user i concerning item j is unknown. $J_i = \{j \in \{1,...,J\} \mid V_{ij} = 1\}$ is the set of items, which have been rated by user *i*. Finally let $W = (w_{kl})$ denote a matrix of weights. All ADCLUS generalizations try to find the best-fitting estimator $(\hat{S}_{ij})$ for the given two-mode data matrix (s_{ij}). There are different estimators $(\hat{S}_{ij})$ for this purpose; the most popular choice is

$$\hat{S}^1_{ij} = \sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{jl'}, \quad (\hat{S}^1 = PWQ').$$

The matrices *P*, *W* and *Q* are usually determined by minimizing the objective function

$$Z_f = \sum_{i'=1}^{I}\sum_{j' \in J_{i'}} (s_{i'j'} - \hat{s}_{i'j'})^2.$$

E.g., the alternating exchanges algorithm by Gaul and Schader (1996) could be used for this task. Because in practice most users have rated only a small part of the items in the data matrix (s_{ij}), a new version of the alternating exchanges algorithm had to be used, which is able to deal with missing values (Gaul et al., 2007). The alternating exchanges algorithm tries to improve the objective function Z_f by transferring either a row or a column element to a different cluster while recalculating (w_{kl}) accordingly:

Algorithm 1 (Alternating Exchanges Algorithm):

1. At first starting values for *P* and *Q* are chosen. *W* is calculated based on the initial values of *P* and *Q*.
2. The following steps are repeated until there are no more changes in *P* and *Q*:
   a. Try to assign each first mode element to a different first mode cluster. Recalculate *W* and the objective function Z_f based on the new matrices *P* and *W* (and on the most recent value of *Q*, which is fixed during this step). Accept the change if it has improved the objective function, otherwise reject it.
   b. Transfer each second mode element to a different second mode cluster. Account for the change in *Q* by recalculating *W*. Then determine the new value for Z_f based on this change (and the previously determined P from step 2a). Accept the changes in *Q* and *W* if the objective function Z_f decreases, else return to the previous matrices *Q* and *W*.

Each w_{kl} is just the average of all s_{ij}, whose first mode elements belong to the first mode cluster *k* and whose second mode elements are members of the *l*-th second mode cluster:

$$w_{kl} = \frac{\sum_{i'=1}^{I}\sum_{j' \in J_{i'}} p_{i'k} s_{i'j'} q_{j'l}}{\sum_{i'=1}^{I}\sum_{j' \in J_{i'}} p_{i'k} q_{j'l}}.$$

If one rearranges the two-mode data matrix so that elements from the same cluster are next to each other, the w_{kl} can be interpreted as a way to sum up a whole partition of the matrix.

The elements of one mode are clustered based on their interaction with the other mode's clusters. Thus, the result of a two-mode clustering characterizes the interaction between the rows and colums of the data matrix.

Lately it has been shown that using a more elaborate estimator $\hat{S}^{(2)}$ decreases the *AAD* significantly (Banerjee et al., 2004). The results achieved by applying $\hat{S}^{(2)}$ are comparable to the results of singular value decomposition and non-negative matrix factorization (George and Merugu, 2005). This alternative estimator $\hat{S}^{(2)}$ includes the popular estimator $\hat{S}^{(1)}$ and furthermore accounts for the differences between individual users by introducing the additional term

$$\bar{s}_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'.} = \frac{1}{|J_i|} \sum_{j' \in J_i} s_{ij'} - \sum_{k'=1}^{K} p_{ik'} \left(\sum_{l'=1}^{L} \sum_{\iota=1}^{I} \sum_{j' \in J_\iota} p_{\iota k'} q_{j'l'} \right)^{-1} \sum_{l'=1}^{L} w_{k'l'} \sum_{\iota=1}^{I} \sum_{j' \in J_\iota} p_{\iota k'} q_{j'l'}.$$

This term describes the way in which the rating behavior of user i differs from the average rating behavior of users that belong to the same first mode cluster as user i. If user i provides more (less) generous ratings than the average user from the first mode cluster that he or she belongs to, the term given above would be positive (negative). Analogously, the heterogeneity of the items is incorporated by adding

$$\bar{s}_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}_{.l'} = \frac{1}{|I_j|} \sum_{i' \in I_j} s_{i'j} - \sum_{l'=1}^{L} q_{jl'} \left(\sum_{k'=1}^{K} \sum_{\iota=1}^{I} \sum_{h \in J_\iota} p_{\iota k'} q_{hl'} \right)^{-1} \sum_{k'=1}^{K} w_{k'l'} \sum_{\iota=1}^{I} \sum_{h \in J_\iota} p_{\iota k'} q_{hl'},$$

which accounts for the fact that the item j might be liked better (worse) than the average item from the second mode cluster to which item *j* belongs. These two additional terms result in

$$\hat{S}_{ij}^{(2)} = \sum_{k'=1}^{K} \sum_{l'=1}^{L} p_{ik'} w_{k'l'} q_{jl'} + \bar{s}_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'.} + \bar{s}_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}_{.l'}.$$

Like SVD-based approaches and non-negative matrix factorization two-mode clustering estimates unknown ratings based on a low rank approximation of the original data matrix.

2. Hierarchical Bayesian model

Another successful approach is Hierarchical Bayesian Regression. We use the Hierarchical linear model by Rossi et al. (1996), which is very similar to the model by Ansari et al., (2000), that has already been used in order to model the ratings of individual users for specific items as a function of item attributes, user characteristics and expert evaluations. If the items are movies, examples for item attributes are the level of violence, suspense, action or romance and the expert evaluations are ratings which were given by professional movie critics. Examples of user characteristics are age and gender. The results of Ansari et al. suggest, that the improvements due to considering the user characteristics are very small. Also users might be deterred if they were asked too many questions about themselves. Therefore reliable information about the users might not always be available for practical purposes. For those reasons we prefer to neglect the user characteristics in our model. The relationship between (a random variable) S_{ij} and both the users taste (β_i) and the item *j*'s attributes and the expert evaluations concerning item *j* (both included in X_{ij}) can be described by the linear model

$$S_{ij} = X_{ij}' \beta_i + \varepsilon_{ij}$$

with $\varepsilon_{ij} : i.i.d.N(0, \sigma_i^2)$, for $i = 1,...,I$, $j = 1,...,\alpha_i$ and $\alpha_i = |J_i|$. The $\kappa_A + 2$ components of the vector X_{ij} include an intercept term, and all item attributes and the average expert evaluation used in the model. This way we basically get *I* different models

$$S_i = X_i \beta_i + \varepsilon_i,$$

with $S_i' = (S_{i1},...,S_{i\alpha_i})$, $X_i' = (X_{i1},...,X_{i\alpha_i})$, $\varepsilon_i' = (\varepsilon_{i1},...,\varepsilon_{i\alpha_i})$ and $\varepsilon_i : i.i.d.N(0, \sigma_i^2 I_{\alpha_i \times \alpha_i})$, $i = 1,...,I$. The different tastes of the users are accounted for by the equation

$$\beta_i = \Delta' z_i + v_i.$$

Here, $z'_{i'}$ is a row vector with d components, which describes the characteristics of the i-th user, $z'_{i'}, i = 1,...,I,$ are the rows of the matrix Z, Δ is a $d \times (\kappa_A + 2)$-matrix of regression coefficients. Δ can be used to model different types of users. In this case the type of user which is most similar to user *i* is selected by z_i. Without any prior knowledge about the users d = 1 and Z equals an I-dimensional vector of ones. It is assumed that $v_i : i.i.d.N(0, V_\beta)$.

No convenient natural prior on $\{\beta_i, \sigma_i\}, i = 1,...,I,$ is known. But given Δ, V_β and σ_i^2 the likelihood function

$$\ell(s_i \mid \beta_i, \sigma_i^2) \propto \frac{1}{|\sigma_i^2 I_{\alpha_i \times \alpha_i}|^{1/2}} \exp\left\{-\frac{(s_i - X_i\beta_i)'(s_i - X_i\beta_i)}{2\sigma_i^2}\right\},$$

is conjugate to a normal prior distribution for β_i ($\beta_i : N(\Delta' z_i, V_\beta)$), from which results the posterior distribution for β_i given σ_i^2:

$$\beta_i \mid \sigma_i^2, s_i, X_i : N\left(\overline{\beta}_i, \left(\frac{X_i' X_i}{\sigma_i^2} + V_\beta^{-1}\right)^{-1}\right),$$

with

$$\overline{\beta}_i = \left(\frac{X_i' X_i}{\sigma_i^2} + V_\beta^{-1}\right)^{-1}\left(\frac{X_i' s_i}{\sigma_i^2} + V_\beta^{-1}\Delta' z_i\right).$$

For known β_i the likelihood function given above is conjugate to the inverse Wishart prior ($\sigma_i^2 : IW(\nu, V_i)$), from which results the posterior distribution for σ_i^2 given β_i

$$\sigma_i^2 \mid \beta_i, X_i, s_i : IW\left(\nu + \alpha_i, \sqrt{\frac{\nu V_i^2 + \alpha_i \varphi_i^2}{\nu + \alpha_i}}\right),$$

with

$$\varphi_i^2 = \frac{1}{\alpha_i}(s_i - X_i\beta_i)'(s_i - X_i\beta_i).$$

As long as Δ and V_β are given, we can use these posterior distributions to design a Gibbs sampler which alternately draws σ_i^2 given β_i and then uses σ_i^2 to draw β_i given σ_i^2.

In the framework of the Hierarchical Bayes approach prior distributions for both Δ and V_β are assumed. It is well-known (see, e.g., McCulloch and Rossi, 1994) that for given $\{\beta_i\}_{i=1}^I$ the natural conjugate prior for V_β is an inverse Wishart distribution ($V_\beta : IW(\nu_0, V_0)$) with posterior distribution

$$V_\beta : IW\left(\nu_0 + I, V_0 + \sum_{i=1}^{I}(\beta_i - \overline{\beta}_i)(\beta_i - \overline{\beta}_i)'\right).$$

Also well-known is that it can be inferred from the likelihood function of the multivariate regression model $B = Z\Delta + V$ (with $B' = (\beta_1, ..., \beta_I), Z' = (z_1, ..., z_I)$, and $V' = (v_1, ..., v_I)$) that the natural conjugate prior distribution for Δ given V_β is given by $vec(\Delta) \sim N(vec(\overline{\Delta}), V_\beta \otimes A^{-1})$.

The corresponding posterior distribution is $vec(\Delta) \sim N(vec(\overline{\Delta}), V_\beta \otimes (A + Z'Z)^{-1})$. Here, vec denotes the vector operator and $\otimes$ is the Kronecker product (see e.g., Magnus und Neudecker, 1988). ν is a prior parameter, ν_0, V_0, A and $\overline{\Delta}$ are called hyperprior parameters.

One can use the posterior distributions given above to design a Gibbs Sampler for the Hierarchical Linear Model, which is given by Algorithm 2.

The convergence of the resulting Markov chain has been checked (McCulloch and Rossi, 1994). In order to achieve independence from the starting values, the first n_{BURN} of the R draws are discarded. All parameter estimates are attained by averaging over the $R - n_{BURN}$ remaining draws.

Algorithm 2 (Gibbs Sampler for the Hierarchical Linear Model):

Start with $\{\sigma^2_{i,0}\}^I_{i=1}$, Δ_0, $V_{\beta,0}$ *and* $n=0$

While $n \leq R$

1. *draw* $\beta_{i,n+1} \sim N\left(\left(\frac{X_i' X_i}{\sigma_{i,n}^2} + V_{\beta,n}^{-1}\right)^{-1}\left(\frac{X_i' s_i}{\sigma_{i,n}^2} + V_{\beta,n}^{-1}\Delta_n' z_i\right), \left(\frac{X_i' X_i}{\sigma_{i,n}^2} + V_{\beta,n}^{-1}\right)^{-1}\right)$ *for* $i = 1, ..., I$

2. *draw* $\beta^2_{i,n+1} \sim IW\left(\nu + \alpha_i, \sqrt{\frac{\nu V_i^2 + (s_i - X_i\beta_{i,n+1})'(s_i - X_i\beta_{i,n+1})}{\nu + \alpha_i}}\right)$ *for* $i = 1, ..., I$

3. *draw* $V_{\beta,n+1} \sim IW\left(\nu_0 + I, V_0 + \sum_{i=1}^{I}(\beta_{i,n+1} - \overline{\beta}_{i,n+1})(\beta_{i,n+1} - \overline{\beta}_{i,n+1})'\right)$

$$with\ \bar{\beta}_{i,n+1} = \left(\frac{X_i^{'} X_i}{\sigma_{i,n}^2} + V_{\beta,n}^{-1} \right)^{-1} \left(\frac{X_i^{'} s_i}{\sigma_{i,n}^2} + V_{\beta,n}^{-1} \Delta_n^{'} z_i \right), i = 1,...,I$$

4. *draw* $vec(\Delta_{n+1}) \sim N(vec(\bar{\Delta}), V_{\beta,n+1} \otimes (A + Z^{'} Z)^{-1})$

5. *Set n to n*+1

Information from all s_i (β_i), $i = 1,...,I$, is pooled via V_β and Δ, and then this pooled information is used to generate new draws for each individual user ($\beta_i, \sigma_i^2, i = 1,...,I$). Thus it is possible to estimate the individual preferences of users, who only supplied very few ratings by borrowing information from other users. Furthermore, these individual estimates $\beta_i, i = 1,...,I$, computed by a Hierarchical Bayes approach exhibit less variation than least squares estimates computed equation by equation (Gelman et al., 2004).

3. Hierarchical Bayes approach based on two-mode clustering

Without any prior information about the users $d = 1$ and Z equals an *I*-dimensional vector of ones. Under those circumstances Δ' is just one vector of the same dimension as every β_i, $i = 1,...,I$, and can be interpreted as the common prior expectation of all β_i, $i = 1,...,I$. The hyperprior parameter $\bar{\Delta}$ equals in this case $vec(\bar{\Delta})$ and would be the (hyperprior) expectation of Δ. Without any prior knowledge $\bar{\Delta}$ could be a vector of zeros.

The result of a procedure for two-mode clustering characterizes the interaction between the rows and columns of the data matrix. Thus the result of a two-mode clustering is informative enough to be used as prior knowledge in the framework of a Hierarchical Bayes approach.

The row clusters derived by two-mode clustering consist of users, which were grouped together because they showed similar rating behavior with respect to the same item-clusters. Thus a user's membership to a specific first mode cluster is indicative of the user's rating behavior or taste. It may be argued that this kind of cluster membersip is much more informative with respect to the user's preferences than the information about the users age and gender (which was used by Ansari et al., 2000). An easy way to exploit the results of a procedure for two-mode clustering within the described linear Hierarchical Bayes approach would be to set $d = K$ and $Z = P$. In this case Δ would be a $K \times (\kappa_A + 2)$ -matrix:

$$\Delta = \begin{pmatrix} \beta^{1'} \\ \vdots \\ \beta^{K'} \end{pmatrix},$$

where $\beta^k, k = 1,...,K$ are regression vectors which belong to one of the K different user-clusters. Analogously $\bar{\Delta}' = (\bar{\beta}^1,...,\bar{\beta}^K)$. Thus, the prior knowledge about the similarity between users dominates the prior distribution of $\beta_i, i = 1,...,I$.

By adding a vector of ones to the matrix Z both general and cluster-specific information can be combined (Rossi et al., 1996). In this case one has to add another row $\beta^{0'}$ to the matrix Δ, which corresponds to the constant column of Z. β_κ^0 measures the general effect of the κ-th attribute. If the item-cluster membership is known for all items, it can be used by defining item class membership dummies of the form

$$\delta_{jl} = \begin{cases} 1, & if\ item\ j\ belongs\ to\ the\ l-th\ item-cluster \\ 0, & otherwise. \end{cases},$$

and add all L dummy variables to the matrix X_{ij}. Thus, if this strategy is adopted, each row of X_i, X_{ij}', consists of an intercept term, the attributes of the *j*-th item, the expert evaluations of the *j*-th item and the dummy variables belonging to the *j*-th item.

If at least for most users $i \in \{1,...,I\}$ the relationship $\alpha_i > (\kappa_A + 2)L$ holds, another strategy for using the item class membership dummies in the Hierarchical Linear Model might be advisable. Here, each X_{ij} just contains the intercept term, the attributes of the *j*-th item and the expert evaluations concerning the *j*-th item. Then we redefine

$$X_i = \begin{pmatrix} \delta_{11} X_{i1}' & \cdots & \delta_{1L} X_{il}' \\ \vdots & \ddots & \vdots \\ \delta_{\alpha_i 1} X_{i\alpha_i}' & \cdots & \delta_{\alpha_i L} X_{\alpha_i 1}' \end{pmatrix},$$

and also $\beta_i = (\beta_i^{1'},...,\beta_i^{L'})'$ and use those definitions in $S_i = X_i \beta_i + \varepsilon_i, i = 1,...,I.$

Alternatively, one could also use the condensed version of the data matrix *S*, the matrix of weights *W*, to build separate regression models for each of the user-clusters $k = 1,\ldots,K$. For each user-cluster k the item-clusters which were particularly liked or disliked by the average user from this first mode cluster could be identified by comparing the elements of the k-th row of W. Those item-clusters itself could be interpreted either with respect to the common traits of the elements of this cluster or by the averages of their attributes. In most cases it should be possible to identify at least some of the characteristic attributes by comparing the item-cluster average of each attribute to the item-cluster averages of the same attribute of different item-clusters. A comparatively high value of the l-th item-cluster average of an attribute κ indicates that κ may be characteristic of l. Attributes which are as much characteristic of the high-rated item-clusters than of the low-rated item-clusters (with respect to the users of the k-th cluster) may be considered to be negligible. Thus, a number of possibly relevant attributes can be identified for each user-cluster, which should be tested for significance. Hopefully, some of those identified attributes qualify to be used as independent variables.

Since users might enjoy different attributes but not all of them with respect to the same item-cluster, it seems less advisible to infer something from the fact, that a particular attribute κ displays a low average value with respect to some item-cluster l. For example, a person might like comedies and horror movies and prefer a horror movie if it is very violent and a comedy if it is funny. Since too much humor might even spoil the thrill of the horror movie, humor might not be of relevance for the horror movie cluster for this user. Nevertheless the same user might enjoy humor in comedies very much.

Furthermore, the researcher might develop an intuition for the reason for high and low ratings by the users from cluster k if he looks closely at the clusters of particularly high-rated and low-rated items. Especially if the number of attributes, items and the number of item-clusters is high, this procedure might turn out to be helpful.

4. The new-item problem

Since no ratings exist until an item has been introduced, the ratings for new items cannot be estimated (and the new items cannot be classified) by two-mode clustering without additional information. George and Merugu (2005) recommend to use the average of each user's ratings $\bar{s}_{i.}$ as estimate for s_{ij} if item j is unknown. This is exactly the same estimate classical automated collaborative filtering approaches (Resnick et al., 1994) provide for items, that have not been rated so far. That way every new item is recommended with the same probability. Since almost always lots of known items j^* exist for which $\hat{S}_{ij^*} > \bar{s}_{i.}$ holds, new items are practically never recommended if we only use $\bar{s}_{i.}$ as an estimate. Since good estimates and recommendations for new items are highly important for business purposes, this is problematic.

In the rare case in which new items are recommended, it only happens because the user in question supplied in general very generous ratings, which has nothing to do with the new item and its properties at all. For a very generous user any unknown item would be recommended with the same probability, which could result in less satisfying recommendations for very generous users. Apparently the new item problem has to be dealt with in a different way.

5. Possible solutions to the new-item problem

If the relevant attributes of the items are known, there exist three different stragegies for solving the New-Item-Problem. The most obvious solution is to simply circumvent the problem and to apply a linear Hierarchical Bayes regression model. This procedure has provided very convincing results (Ansari et al., 2000). We call this direct estimation without two-mode clustering. A possible downside to this strategy might be that relevant (and also available) information about each individual user (which can be inferred from the data by two-mode clustering) is not included in the prior distribution. Another strategy is to perform two-mode clustering first and then incorporate information about the class-membership of the users into the prior and hyperprior distribution of a linear Hierarchical Bayes regression model, which one might term direct estimation based on two-mode clustering. The third possible solution is the indirect estimation based on two-mode clustering. Here, the second mode classes of the unknown items $j \in J^N$ are estimated based on their attributes. First a regular two-mode clustering is performed for the known items $j \in J^K$. Then the second mode class-membership and attributes of the known items are used to build a model, which relates class-membership to those attributes. Finally, this model is used to predict the class-membership of the new items $\hat{q}_{jl}, j \in J^N$. After the second mode class-membership of the new items has been estimated in the described manner, one can easily derive the $\hat{S}^1$-

estimator for the new item. By using the average rating given by a professional product tester or critic $\bar{s}_{.j}^{C} = (1/|C_j|)\sum_{i_C \in C_j} s_{i_C j}^{C}$ as a replacement for $\bar{s}_{.j}$, one is in principle also in a position to approximate $\hat{S}_{ij}^{2}$ for $j \in J^N$, if ratings s_{ij}^{C} are available for all items $j \in J^N$. (Here, $s_{i_c j}^{C}$ is the rating of a professional product tester or critic i_c concerning item j, and C_j is the set of all critics who rated item j.)

An advantage of the indirect estimation based on two-mode clustering over the direct estimation based on two-mode clustering is, that the model for the class membership uses only second mode data and can therefore be calculated much quicker. Moreover, one has to consider that if coded efficiently the algorithm for two-mode clustering is very fast (George and Merugu, 2005). Parallel versions of the algorithm for two-mode clustering already exist (George and Merugu, 2005).

5.1. Direct estimation. For the direct estimation without two-mode clustering d = 1, Z is an I-dimensional vector consisting of ones, and B' is a k-dimensional vector of zeros.

Since we focus on items which have not been rated so far, the item class membership is not available. Therefore, the direct estimation based on two-mode clustering is equivalent to the Hierarchical Bayes approach based on two-mode clustering, which only utilizes the first mode classification derived by two-mode clustering.

5.2. Indirect estimation. The indirect estimation hinges on a reliable procedure for the estimation of the second mode cluster membership of a new item. The key idea is, that every item has certain attributes, which are the reason why a specific user likes a certain item or not and which are also responsible for the classification of the corresponding item regardless of whether the item is already known or not. (Because two-mode clustering treats users and items symmetrically, the same approach could be applied to users that have not rated anything as well, provided, that any relevant information about them was known.) Since items with known ratings are also given, regular two-mode clustering can be performed with respect to the known items. Then the second mode class-membership and attributes of the known items can be used to build a model or at least to train a procedure, which relates class-membership to those attributes. Finally, this model or procedure is used to predict the class-membership of the new items. Alternatively, the attributes of the items can be used to calculate the dissimilarity between all items. Based on those dissimilarities different heuristics can be used to assign a new item to a cluster of known items. Because the choice of the model or heuristic is crucial for the success of the indirect estimation method, many different models and heuristics have to be explored.

After the second mode cluster-membership of the new items $j \in J^N$ has been estimated by one of the different models or heuristics, one can use the resulting estimates $\hat{q}_{jl}, j \in J^N$. Hence, an $\hat{S}^1$-like estimator $\hat{S}_{ij}^{*1} = \sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} \hat{q}_{jl'}$ can be calculated. In order to derive an $\hat{S}^2$-type estimator for new (unknown) items, $\bar{s}_{.j}^{C}$ might be used as a substitute for the missing $\bar{s}_{.j}$. However, not just $\bar{s}_{.j}$ but the whole difference $\bar{s}_{.j} - \sum_{l'=1}^{L} q_{jl'} \tilde{w}_{.l'}$, which is added to account for the heterogeneity of the items, needs to be replaced, since the average ratings from professional movie critics might be different from the corresponding user-averages. Thus, the term $\bar{s}_{.j} - \tilde{w}_{.l}$, which is added to account for the heterogeneity of the items from the same cluster, is replaced by $\bar{s}_{.j}^{C} - \tilde{w}_{.l}^{C}$ with

$$\tilde{w}_{.l}^{C} = \frac{\sum_{j' \in J^K} \sum_{i_C' \in C_{j'}} q_{j'l} s_{i_C' j'}^{C}}{\sum_{j' \in J^K} \sum_{i_C' \in C_{j'}} q_{j'l}}.$$

Thus, $\hat{S}_{ij}^{2}$ for $j \in J^N$ can be approximated by

$$\sum_{k'=1}^{K}\sum_{l'=1}^{L} p_{ik'} w_{k'l'} \hat{q}_{jl'} + \bar{s}_{i.} - \sum_{k'=1}^{K} p_{ik'} \tilde{w}_{k'.} + \bar{s}_{.j}^{C} - \sum_{l'=1}^{L} \hat{q}_{jl'} \tilde{w}_{.l'}^{C}.$$

Several procedures already exist, which can be used to approximate $\hat{q}_{jl}, j \in J^N$, based on the cluster-membership $q_{jl}, j \in J^K$, of the known items and the attributes of all items $j \in J = J^K \cup J^N$. Some of those procedures, logistic regression (LR), neural networks (NN) and the C4.5 decision tree algorithm, are already well-known, while other procedures like the Bayesian Multinets (BMN) and the Logistic Model Trees (LMT) do not belong to the general methodological tool kit yet. Therefore, the less well-known procedures are briefly introduced in Appendix A, while it is only sketched how logistic regression (LR) and the C4.5 algorithm can be used for the cluster-membership approximation of the new items. In addition to those methods, 3 very

simple heuristic procedures are introduced in Appendix B, which are able to perform the same task at much lower computational cost. One of those procedures, the SL -heuristic, yields very promising results if combined with the $\hat{S}_Y^1$ -estimator.

6. Data

The MovieLens data set, which is publicly available today, contains approximately 1 million ratings for 3872 movies entered by 6040 users (http://www.grouplens.org). The movies were rated on a five-point scale. A rating of 5 expresses that the user likes the movie very much, whereas a rating of 1 expresses the opposite. For each movie its title and genre are given. Additionally, movie attributes like the level of suspense of the movie were collected for a subset of these movies from another website (http://reel.com). For most of the movies at least 14 different movie attributes are given. All attributes are measured on an integer scale ranging from 0 to 10. 10 means that the corresponding attribute is characteristic of the movie, 0 indicates that the attribute is not an attribute of the movie at all. Only for a subset of the MovieLens data the movie attributes were available from http://reel.com. In addition to that movie reviews from professional movie critics like Roger Ebert (Chicago Sun Times) and James Berardinelli (ReelViews) were collected from 10 different movie critics. Because not every critic rated every movie from the MovieLens subset for which movie attributes were available from http://reel.com a subset of this subset had to be selected. This subset of the subset was selected so that each of the movies was rated by at least 4 movie critics. In addition to that all users that supplied less than 50 ratings with respect to those 418 movies were omitted. The resulting data set contains 418 movies and 1067 users. The missing value percentage is 78,9 % .

Since most of the ratings by the movie critics are not on a five-point scale, the ratings of the movie specialists were rescaled so that 1 is the lowest and 5 is the highest possible rating. Unlike the MovieLense ratings the rescaled ratings by movie critics are not whole-numbered.

7. Evaluation

A number of different evaluation metrics is used to assess the usefulness of a given set of recommendations. The most popular metric seems to be the average absolute deviation (*AAD*):

$$AAD = \left(\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} | J_{i'} | \right)^{-1} \left(\sum_{i'=1}^{I} \sum_{j' \in J_{i'}} | s_{i'j'} - \hat{S}_{i'j'} | \right).$$

Other measures tend to focus more on the usefulness of the recommendations which can be derived from the estimates $\hat{S}_{ij}, i = 1,...,I, j = 1,...,J$. In this context it is usually assumed that only items are recommended whose estimates are above a given threshold. The precision (*Prec.*) is the percentage of recommended items (presumably unknown items whose estimates are higher than the threshold) that are of interest. Recall (*Rec.*) is the percentage of interesting items which are recommended. Here, an item j is supposed to be of interest to a user i if $s_{ij} = 5$. Moreover, it is assumed that a new item j is recommended to the user i if $\hat{S}_{ij} > 4.5$. Since lots of movies exist, it is much more useful to get recommendations for some of the movies that one particularly enjoys than to get each and every movie recommended, which one might consider to be half-way decent or better and that one could enjoy – provided one had more time. Thus, the precision is considered to be more important than the recall in the context of movie recommendation. However, if the purpose of estimation is not only recommendation but also market research, it is equally important to identify as many of the prospective customers for a new product as possible. So the recall is not negligible if one focuses on estimating preferences for new products.

The Breese metric $R_{B,i}$ (Breese et al., 1998) is an estimate of the expected utility of a particular ranked list to the user i . The higher the estimate for an item is, the higher it is positioned on the ranked list $j_{list,i} = 1,...,J_{list,i}$ for the user of interest i .

The bigger the positive difference between the actual rating and the average rating (or any noncommittal rating), $\max(s_{ij} - d_i, 0)$, is, the more helpful the recommendation of item j to user i can be considered.

Each recommendation on the ranked list is less likely to be followed than the preciding one. Therefore, a recommendation for an item, which turns out to be highly enjoyable to the user *i*, is less useful if it appears at a lower position in the ranked list:

$$R_{B,i} = \sum_{j_{list,i}=1}^{J_{list,i}} \frac{\max(s_{ij_{list,i}} - d_i, 0) V_{ij_{list,i}}}{2^{(j_{list,i}-1)/(\alpha_c - 1)}}.$$

Here, α_c is the so-called halflife, which is the number of the recommendation on the list, which has a 50 % chance of being used. α_c is usually set to 5.

Let $R_{B,i}^{max}$ be the maximum achievable utility if all observed items had appeared in the order of their actual rating at the top of the recommendation list.

$$R_B = 100 \frac{\sum_{i=1}^{I} R_{B,i}}{\sum_{i=1}^{I} R_{B,i}^{max}}$$

is referred to as Breese-Score.

8. Results

The data set was divided into test and training sets. The test set consists of all ratings for 118 movies which were selected by random numbers. All ratings which deal with the remaining 300 movies were used as training set. All 118 test set movies are used as new items J^N. The training set movies correspond to the set J^K.

8.1. Indirect estimation. The two-mode cluster sizes were set to the same values which were used by Banerjee et al. (2004) and by George and Merugu (2005): K=L=10. Both the usual estimate $\hat{S}^1$ and the new estimate $\hat{S}^2$ were used to generate two-mode classifications.

All weights $\nu_h, h = 1,...,\kappa_A$ used in the weighted Euclidean distance and α were set to one. For K=L=10 the $\hat{S}^2$-based two-mode clustering procedure results in R^2=0,422 and ADD=0,660 for the training set, while the $\hat{S}^1$-based algorithm for two mode clustering leads to R^2=0,392 and AAD=0,684.

For the test set the mean value recommendation technique (*MVR*) advocated by George and Merugu (2005) is taken as baseline model. The results for the test set are shown in Table 1.

Table 1. Results from the Indirect Estimation Method ($\hat{S}^2$)

$\hat{S}^2$	MVR	SL	α	KM	C 4.5	LMT	BMN	LR	NN
R^2	0.137	0.172	0.061	0.156	0.198	0.156	0.193	0.049	0.158
			(0.175)		(0.206)	(0.196)	(0.203)	(0.116)	(0.179)
AAD	0.840	0.812	0.865	0.819	0.798	0.815	0.801	0.867	0.812
			(0.811)		(0.792)	(0.797)	(0.795)	(0.834)	(0.799)
Prec.	-	0.438	0.391	0.415	0.422	0.404	0.435	0.371	0.408
			(0.416)		(0.429)	(0.424)	(0.424)	(0.384)	(0.395)
Rec.	-	0.341	0.396	0.450	0.422	0.416	0.416	0.414	0.407
			(0.360)		(0.435)	(0.417)	(0.398)	(0.401)	(0.396)
R_B	55.99	70.24	68.54	70.75	71.67	71.04	71.66	68.44	69.66
			(70.12)		(72.43)	(71.33)	(71.44)	(69.51)	(70.15)

Note: The results of the Indirect Estimation Method ($\hat{S}^2$) for the different procedures for the estimation of the second mode cluster membership compared to the results of the mean value recommendation technique for the test data set. The numbers in brackets belong to the continuous versions.

The *SL*- and *KM*-heuristics as well as C4.5, logistic model trees and Bayesian Multinets (BMN) clearly outperform the mean value technique in every way with respect to the test set. Interestingly the Breese Score is very high for all estimates derived by two-mode clustering based on $\hat{S}^2$ even if the fit is less than statisfactory. By integrating $\bar{s}_{i.}$ and $\bar{s}_{.j}^{C}$ directly into the $\hat{S}^2$-estimates, those estimates do not only depend on the correct classification. Even if the classification is wrong those terms still contribute to the estimation, which means, that the estimates for items, which are likely to be high-rated, and the estimates for generous users still tend to be above average. Thus, the higher the $\hat{S}^2$-estimate for a given rating is, the more plausible it becomes, that this rating is indeed quite high, so that the highest estimates are most likely to belong to high ratings. For every user i the items which were generally higher-rated than the average item from the same item-cluster are more likely to be recommended to the user i than others. Indeed, the higher an item is generally rated, the more it will tend to be at the top of the recommendation list, which explains why the Breese Score is high even if not only the fit but also recall and precision leave much to be desired. The strategy of recommending items which are generally preferred is combined with the strategy of personal recommendation. This property is not shared by the two-mode clustering estimates based on $\hat{S}^1$.

Except for the procedure which uses the single linkage heuristic every procedure for indirect

estimation based on the usual estimate $\hat{S}^1$ is clearly outperformed by the mean value technique. A possible reason might be that the $\hat{S}^1$-estimates react more sensitively to misclassification. The more elaborate procedures like logistic regression, C4.5, Bayesian Multinet and logistic model trees are usually used for classification tasks with fewer than 10 classes, which might lead to more misclassification.

Surprisingly, the indirect estimation results based on the *SL*-heuristic are rather good as can be seen from Table 2.

Table 2. Results from the Indirect Estimation Method ($\hat{S}^1$)

R^2	AAD	Prec.	Rec.	R_B
0.270 (0.392)	0.765 (0.684)	0.597	0.149	70.77

Note: The results for the Indirect Estimation Method based on $\hat{S}^1$ and the single linkage heuristic (without brackets). The results in brackets belong to the training set.

R^2, *AAD*, recall and precision are much better than for the indirect estimation results based on $\hat{S}^2$. (The same pattern was reproduced for a different test and training set, which were also selected by random numbers.)

Since the *SL* results produced by the $\hat{S}^2$-based two-mode clustering are not as good as those results, it seems that the $\hat{S}^1$-based method for two-mode clustering generates more homogeneous item-clusters. Nevertheless the Breese Score R_B is lower than for the indirect estimation procedures based on $\hat{S}^2$ and C4.5, logistic model trees and Bayesian Multinets.

8.2. Direct estimation. The attributes and the average rating by the professional movie critics of each movie were used as independent variables for the linear Hierarchical Bayes regression model. The effects of most of the 14 attributes are negligible if they are used for all users. However, 3 attributes could be identified that seem to influence the rating more strongly than others: the levels of action, suspense and character development (char.). A diffuse but proper prior distribution was chosen (A=0.01I, ν=3, ν_0=8, $V_i = \hat{\text{var}}(s_i)$, V_0=*diag*(4,0.01,0.01,0.01,0.1), every entry of $\overline{\Delta}$ was set to zero). This parametrization can be referred to as conservative. Since the prior expectation of each attribute's effect is set to zero, the small prior variances prevent the three attributes to contribute significantly to the estimator – unless this is demanded by the data. Thereby, overfitting can be avoided and more accurate predictions are possible. Both is very important for the desired purposes. However, it is certainly true, that this might lead to an underestimation of the resulting posterior variances. Therefore, the significance tests concerning the parameters of those attributes might be considered to be problematic. However, those tests still provide some guidance with respect to the question, whether it makes sense to introduce a certain variable – at a specified level of the prior variance of this variable. (For purposes, which require the exact estimation of the effect of certain variable, the used parametrization would be undesirable.) Here, this parametrization yields the best results.

The estimation was performed for $Z = Z_1, Z_2$ and Z_3 (and properly adjusted $\Delta, \overline{\Delta}$). Z_1 is simply a vector of ones, $Z_2 = P$ and Z_3 is the combined matrix $(Z_1 Z_2)$. By using Z_1 the estimation is performed without any information from the two-mode clustering analysis. For Z_2 only information from the users which belong to the same first mode cluster is pooled. The choice Z_3 combines both approaches: General effects (β^0) are estimated using pooled information from all users and cluster-specific effects $\beta^k, k = 1,...,K$ are calculated based exclusively on people which belong to the same cluster. Then general and cluster-specific information is combined via the prior distribution to estimate the effects for individual users. The results are given in Table 3.

Except with respect to recall and Breese score the results for Direct Estimation clearly outperform all results of the direct estimation approach. The results for all procedures for direct estimation listed in Table 3 differ from each other only slightly.

Table 3. Results from the Direct Estimation Method

	R^2	AAD	R^2	AAD	Prec.	Rec	R_B
Data set	training set	training set	test set	test set	test set	test set	test set
Z_1	0.332	0.715	0.281	0.762	0.613	0.043	70.02
Z_2	0.332	0.715	0.284	0.760	0.644	0.051	70.52
Z_3	0.332	0.715	0.284	0.760	0.641	0.050	70.53

Note: The results for the Direct Estimation Method based on the attributes action, suspense and character development and the average rating by professional movie critics. Here, Z_2 and Z_3 use the results from an $\hat{S}^2$ two-mode clustering.

Table 4 provides some information about the posterior distribution of Δ based on $Z=Z_3$. Both the posterior means of the Δ-coefficients and their posterior probability of being positive or negative are presented. Moreover, the unobserved heterogeneity and the ρ_η^2-measure for each independent variable are given. The unobserved heterogeneity of an independent variable η (η=1,…K_M) is the square root of the posterior mean of the corresponding diagonal element of V_β: $\sqrt{(V_\beta)_{\eta\eta}}$. The ρ^2-measure $\rho_\eta^2 = 1 - var(\varepsilon_\eta)/var(\beta_\eta) = 1 - (V_\beta)_{\eta\eta}/var(\beta_\eta), \eta = 1,...,\kappa_M$. Here, $var(\beta_\eta)$ is the total variation of β_η, and $(V_\beta)_{\eta\eta}$ is the conditional variance of β_η. (κ_M equals 5.) ρ_η^2 is an R^2-like quantity, which measures how much of the variation of each coefficient $\beta_\eta, \eta = 1,...,\kappa_M$ is due to random effects. This measure was introduced by Rossi et al., (1996).

In general, the results suggest, that movies are more likely to be high-rated, if they are liked by the critics, focus strongly on character development and also manage to build up a high level of suspense. High levels of action seem to have a slightly negative effect on the rating.

Table 4. Posterior distribution of Δ-coefficients

	Gen.*	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	u.h.**	ρ_η^{2*}
Intercept	**1.34**	**1.29**	**-0.71**	**-0.58**	**0.31**	**0.24**	**0.62**	**-0.76**	**0.30**	0.04	**0.38**	0.712	0.34
$\eta = 1$	(0.96)	(0.99)	[0.87]	(0.68)	(0.69)	(0.87)	(0.97)	[0.96]	(0.68)	(0.48)	(0.74)		
Action	**-0.014**	**0.045**	**-0.062**	**-0.020**	**-0.008**	**0.004**	**-0.045**	**-0.011**	**-0.005**	**0.039**	**0.039**	0.024	0.51
$\eta = 2$	[0.68]	(1.00)	[1.00]	[0.95]	[0.77]	(0.68)	[1.00]	[0.79]	[0.69]	(0.99)	(0.97)		
Suspense	**0.028**	**-0.017**	**0.018**	**0.006**	**0.004**	**-0.005**	**0.011**	0.001	**0.010**	**0.010**	0.002	0.022	0.87
$\eta = 3$	(0.99)	[0.99]	(0.98)	(0.77)	(0.72)	[0.76]	(0.92)	(0.53)	(0.90)	(0.94)	(0.65)		
Char.**	**0.057**	**-0.030**	**0.043**	**0.019**	-0.001	**-0.012**	0.001	**0.037**	**0.030**	**-0.023**	**-0.005**	0.024	0.84
$\eta = 4$	(1.00)	[1.00]	(1.00)	(0.97)	[0.55]	[0.87]	(0.58)	(0.99)	(0.98)	[0.99]	[0.67]		
$\overline{s}_{.j}^{C***}$	**0.371**	**-0.181**	**0.206**	**0.234**	**0.045**	**0.066**	-0.005	**0.230**	**-0.033**	**0.041**	0.005	0.121	0.91
$\eta = 5 = \kappa_M$	(1.00)	[1.00]	(1.00)	(1.00)	(0.73)	(0.79)	[0.55]	[1.00]	[0.65]	(0.75)	(0.54)		

Notes: * The term general model refers to the first row of Δ ($\beta^{0'}$); ** unobservable heterogeneity (u.h.) as measured by the square roots of the posterior mean of the diagonal elements of V_β; $^*\rho_\eta^2 = 1 - var(\varepsilon_\eta)/var(\beta_\eta) = 1 - (V_\beta)_{\eta\eta}/var(\beta_\eta), \eta = 1,...,\kappa_M$. Here, $var(\beta_\eta)$ is the total variation of β_η, and $(V_\beta)_{\eta\eta}$ is the conditional variance of β_η, and $var(\beta_\eta)$ is the total variation of β_η (for details see Rossi et al., 1996); ** character development; *** average rating by professional movie critic; ( ) indicates probability that coefficient is positive; [ ] indicates probability that coefficient is negative; Bold indicates probability exceeds 0.95.

Table 5. The matrix of weights W for $\hat{S}^2$

	l=1	l=2	l=3	l=4	l=5	l=6	l=7	l=8	l=9	l=10
k=1	3.72	3.65	3.58	3.59	3.61	3.30	3.65	3.40	3.98	3.65
k=2	2.60	2.14	3.84	3.94	3.33	4.21	3.13	2.70	3.26	3.83
k=3	3.31	2.43	4.18	3.67	3.73	4.10	3.13	2.78	3.87	3.70
k=4	3.42	3.19	3.81	3.71	3.94	4.11	3.64	2.63	3.78	3.58
k=5	3.26	2.97	3.98	3.97	3.32	3.62	3.57	2.95	3.95	3.11
k=6	3.33	3.09	3.98	4.02	3.85	3.98	3.64	3.51	3.42	4.20
k=7	2.92	2.58	4.02	3.90	3.27	4.21	3.38	2.29	3.89	3.30
k=8	3.31	2.97	3.83	3.80	3.80	4.34	3.37	3.34	3.75	2.90
k=9	3.33	3.25	3.69	2.97	3.23	3.71	3.36	2.85	4.13	2.58
k=10	3.48	3.37	3.97	3.25	3.68	4.05	3.30	3.29	4.13	3.94

The remaining Δ-coefficients can be used to learn something about the user-clusters. The first user-cluster (k = 1) has the highest intercept term of all clusters. Also all remaining coefficients $\Delta_{2\eta}, \eta = 2,...,5,$ have the exact opposite sign than the attributes for the general model $\Delta_{1\eta}, \eta = 2,...,5,$ which makes the actual movie characteristics less important. This cluster is the one, that seems to favor action movies most strongly. Also, those users seem to prefer higher levels of character development. Nevertheless, they also have the smallest coefficient for character development, which means that character development is comparatively unimportant to them. The level of suspense seems to be irrelevant to them. From the first row of the matrix of weights W (as given in Table 5) it can be seen that those users provided quite high and also very similar average ratings for all of the movie clusters. The only cluster which those users seem to favor slightly is cluster l = 9. One might think that either this group of users does not have any particular preferences, or they hide them well by their tendency to high-rate movies. However, one might get a hint about their particular tastes by the observation that cluster k = 1 provided the lowest average rating for the movies from item-cluster l = 6, which is a cluster of items that are strongly high-rated by all other user-clusters except for cluster k = 5. Also their average rating for item-cluster l = 6 is the lowest average rating in the whole first row of the matrix W. This hint will be dealt with later.

The information we get about the users from cluster k = 2 at the first glance is much more informative. Those users strongly dislike action and clearly prefer movies which are high-rated by the professional movie critics and that focus on character development. Suspense is also important to the members of cluster k = 2.

Another possibility for using the results of a two-mode clustering as starting point for a Bayesian regression model is to identify characteristic attributes for each user-cluster as discussed at the end of section 4. This procedure will be illustrated by two practical examples. First, a model for user-cluster k = 2 will be developed.

By looking at the second row of the matrix of weights W the item-clusters l = 4 and l = 6 to be the highest-rated movie clusters by the users from cluster k = 2. Furthermore, the movies from cluster l = 2 seem to be particularly disliked by this user-cluster.

The movies from cluster l = 4 have the highest average value for character development of all item-clusters. Both for l = 2 and l = 6 character development does not seem to be an important attribute. Thus, character development could turn out to be useful as independent variable. Also the movies from the 4-th cluster seem to display very few action. Neither movie cluster l = 2 nor l = 6 have remarkable average values with respect to this attribute. So action might prove to be a good independent variable.

Movie cluster l = 6 has a quite high average but not the highest value for the attribute hollywood style (6.84). Cluster l = 4 has quite a comparatively low value for hollywood style (5.84) and cluster l = 2 has a hollywood style degree which is very similar to that of cluster l = 6. Since movies from the second movie cluster are strongly disliked and movies from cluster l = 6 are liked, one might infer, that hollywood style is immaterial to the users from the second user-cluster. Therefore, there is no need to put hollywood style on the list of possibly useful variables.

However, it should not be argued that hollywood style should not be taken into further consideration because the average value for hollywood style was low in cluster l = 4, since users might enjoy different attributes – but not all of them with respect to the same item-cluster. For example, a person might like comedies and horror movies and prefer a horror movie if it is very violent and a comedy if it is funny. Since too much humor might even spoil the thrill of the horror movie, humor might not be of relevance for the horror movie cluster for this user. Thus, one should always focus on the attributes with high item-cluster averages.

Finally, we have to deal with cluster l = 2. Cluster l = 2 has the highest average value for cinematography of all item-clusters and a comparatively high average value for offbeat energy.

With respect to cinematography the 4-th item-cluster also has a quite high average value. However, there are three movie clusters, which actually have even higher average values concerning this attribute. Thus, cinematography should not be considered to be a characteristic of cluster l = 4, which means that this does count as argument against the acceptance of cinematography as possibly useful variable. Cluster l = 6 also has a quite high average value for cinematography, which is at the same time the second lowest average user-cluster level for cinematography. Thus one should not infer anything about the importance of cinematography from the cluster average of item-cluster l = 6. Thus, cinematography is a possibly useful variable.

Offbeat energy has a comparatively low value for movie cluster l = 6 and carries not much weight in cluster l = 4. Since offbeat energy is comparatively low in cluster l = 6 (which is liked by users from the second user-cluster) and has a high average value

for the movie cluster $l = 2$ (which is detested by the second user-cluster) it should be considered as a possibly helpful addition to the set of independent variables.

These considerations yield a set of possibly useful independent variables (with respect to the second user-cluster) which consist of action, character development, offbeat energy and cinematography. All of those variables except for cinematography turn out to be significant. The model which uses the intercept, the levels of action, character development and offbeat energy describes the data from user-cluster $k = 2$ as well.

Table 6. Separate regression models

Model	Relevant attributes*
k=1	Hollywood style, action, character development
k=2	Action, character development, offbeat energy
k=3	Action, suspense, character development, offbeat energy
k=4	Humor, suspense, character development, offbeat energy
k=5	Violence, character development, offbeat energy
k=6	Action, suspense, character development, offbeat energy
k=7	Character development, offbeat energy
k=8	Character development, offbeat energy
k=9	Family appeal, hollywood style, suspense, character development, cinematography
k=10	Hollywood style, suspense, character development, cinematography

Note: * The term relevant attributes refer to all attributes which are useful choices for independent variables.

Since all item-clusters were given similar average ratings by the persons from the fist user-cluster, the task of developing a model for this group of users is a bit less straightforward. One can learn from the matrix W that these users provided the lowest average rating for the movies from item-cluster $l = 6$, which is a cluster of items that are strongly high-rated by all other user-clusters with the exception of the 5-th user-cluster. Also their average rating for item-cluster $l = 6$ is the lowest average rating they provided for any of the movie clusters. These findings suggest, that the first user-cluster might dislike attributes of the 6-th item-cluster even though their average rating for these movies is not particularly low. The characteristic attribute for the item-cluster $l = 6$ is hollywood style. The only other movie cluster which could be of relevance in this context is the 9-th one, which seems to be slightly preferred by this group of users. This item-cluster has no significantly high values for any of the attributes. However, it has also the lowest average value for hollywood style of all item-clusters. Thus it should be tried as regressor.

Since one possibly relevant independent variable might not be sufficient, information from other sources should also be taken into account. Because the results of the previous Bayesian regressions suggest that action and character development also might be useful as independent variables, it might be a good idea to try these attributes in addition to hollywood style and the average rating given by a professional movie critic. Finally all three attributes turn out to be a significant contribution to the regression model for user-cluster $k = 1$.

Of course one should not only check the significance but also the consistency of the sign of the resulting Δ-coefficients with the argument for its introduction. If the variable is significant but has an unexpected sign, this means that we may have by accident discovered a useful additional independent variable but have not succeeded in discovering the most important influences.

In the case of user-cluster $k = 1$ the assumption was, that since the users of this cluster seem to dislike movies from movie cluster $l = 6$ and hollywood style is the only attribute which has a high average cluster value for this item-cluster, the sign of hollywood style should be negative. If hollywood style proves to be a significant contribution to the model but has a positive sign, we have proven that the high level of hollywood style is not the reason for the relative unpopularity of movie cluster $l = 6$ with respect to the first user-cluster. In this case further investigations should be undertaken, since a reason for the relative unpopularity has not been discovered yet and might turn out to improve the regression model. Since in this example no other attribute has a significantly high average cluster value, we could only try to identify attributes with particularly low average levels for $l = 6$. However, it has to be emphasized that people might enjoy different attributes but not all of them with respect to the same group of items. Thus an exceptionally low cluster average is a weaker argument for the relevance of the respective attribute than a considerably high cluster average value.

Following this procedure one is in a position to exploit the two-mode classification in order to derive 10 different regression models for ten different groups of users. An overview of the attributes which are used as independent variables in each of these models is given in Table 6. For all models the average rating of a professional movie specialist is also used as exogenous variable. Also an intercept term is estimated for each of the models.

Since the user-clusters are disjoint sets it is possible to combine those ten different models to a model for the whole data set. This yields the results presented in Table 7.

Table 7. Results of the combined model

R^2 (training set)	AAD (training set)	R^2 (test set)	AAD (test set)	Prec. (test set)	Rec. (test set)	R_B (test set)
0.364	0.698	0.305	0.748	0.638	0.089	71.44

Note: The results of the model that combines the estimates derived by the models outlined in Table 6.

The results from the combined models outperform all other methods for indirect estimation. They also compare favorably to all of the clustering models with respect to R^2 (test set) and *AAD* (test set). However the recall is still remarkably low. At the same time the precision is much higher for the regression models than for the cluster approaches. This suggests, that the threshold of 4.5 might be too high for regression models. The Breese Score is high but also slightly smaller than for some of the $\hat{S}^2$-clustering models.

9. The importance of variable selection

Due to the applied conservative parametrization of the linear hierarchical Bayesian regression, even different attribute-variable selections yield similar and reliable results with respect to the hierarchical approach.

In contrast, the results of the proposed combination of the *SL*-heuristic with $\hat{S}^1$ two-mode clustering, depend very strongly on the set of variables, which is used to classify the new items. In order to select a promising set of variables, a forward-backward selection technique based on the Breese-Score can be recommended. Before this procedure starts, the given (300) items are divided into two disjoint sets. With respect to the first item-set (and all users) the ($\hat{S}^1$) two-mode clustering procedure is applied. In each step, the *SL*-heuristic is utilized to determine the cluster-membership of the items, which belong to the other set. All ratings concerning those items are used to calculate the resulting Breese-Score. At the beginning, all variables, which could possibly be relevant are included in the set of variables. The forward-backward procedure alternates between two procedure-types (referred to as forward- and backward-step), until no further improvements are possible. Each of those procedure-types consists of a number of different steps.

During the so-called backward-selection, variables from the set are omitted from the set, as long as the resulting Breese-Score is thereby increased. In every backward-selection step, the (single) variable is omitted from the set, without which the maximum Breese-Score can be obtained, unless the omission does not yield any improvement of the resulting measure of utility. The backward-selection is finished as soon as the set contains no variable, that can be left out in order to obtain a higher Breese-Score.

Then, the forward-selection begins, which tries to boost the approximated utility of the resulting recommendation list by adding further variables to the set, as long as the Breese-Score can be increased thereby. During each forward-selection step, the (single) variable is added to the (resulting) set, which leads to the highest Breese-Score, unless no further improvements in terms of estimated utility are possible by the addition of variables to the set.

Here, the outlined procedure yields a set of variables that consist of action, suspense, character development and the average rating of the professional movie critics. If ($\hat{S}^1$) two-mode clustering is applied to all ratings concerning the usual 300 movies and the *SL* heuristic based on the (by forward-backward selection determined) variable-set is utilized to classify the remaining 118 items, the results presented in Table 8 can be obtained.

Table 8. Results after variable selection

R^2 (training set)	AAD (training set)	R^2 (test set)	AAD (test set)	Prec. (test set)	Rec. (test set)	R_B (test set)
0.382	0.677	0.294	0.754	0.647	0.345	73.75

Note: The results of the combination of the $\hat{S}^1$ two-mode clustering method with the *SL*-method based on the variable-set, which was determined by forward-backward selection.

Conclusion

If the quality of the resulting recommendations are of primary importance, the procedure, which is a combination of the *SL*-heuristic and the $\hat{S}^1$-method for two-mode clustering, clearly outperforms all other procedures, as long as it is based on the results of a previously performed procedure for variable selection. Even though the direct estimation method could lead to more accurate predictions, the difference in terms of accuracy (*AAD* and R^2) seems almost negligible. Therefore, the much quicker heuristic can be recommended.

It has been shown that the method for direct estimation which uses the two-mode classification to derive different models for each of the user-clusters yields convincing results. Also it has been demonstrated by practical examples that one can learn a great deal more about the data by combining two-mode clustering and Bayesian regression. Thereby one can learn a great deal more about the data and develop a deeper understanding with respect to the taste of each user-cluster.

Since online shop data matrices are in general huge it might be helpful to use two-mode clustering to divide the huge amount of users into smaller clusters which can be handled by a Hierarchical Bayes regression approach.

Quick updating procedures for two-mode clustering already exist (George and Merugu, 2005). Once the regression models have been build and the decision has been made, for which groups of users regression estimates should be used instead of two-mode clustering estimates, updating should be straightforward. New Bayesian modelling efforts should be undertaken as soon as either a substantial number of users has been switched to different clusters or the structure of the matrix of weights W is changed significantly by the two-mode clustering updates. All changes of W that make the previous interpretation of W seem unsatisfactory should be considered significant.

The resulting recommendation framework has the advantage that even ratings (preferences) of individual users for purely hypothetical items could be estimated. Thus, it is in principle even possible to use the outlined methods in order to decide, whether a certain item should be offered by a particular online store in the future.

References

1. Adomavicius, Gediminas and Tuzhilin, Alexander (2005), *"Toward the Next Generation of recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", IEEE Transactions on Knowledge and Data Engineering*, 17, 734-749.
2. Asim, Essegaier, Skander, and Kohli, Rajeev (2000), *"Internet Recommendation Systems"*, *Journal of Marketing Research,* 37, 363-375.
3. Banerjee, Arindam, Dhillon, Inderjit S., Ghosh, Joydeep, Merugu, Srujana, and Modha, Dharmendra S. (2004), *"A Generalized Maximum Entropy Approach to Bregman Co-Clustering and Matrix Approximation"*, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 509-514.
4. Both and Gaul, Wolfgang (1987), *"Ein Vergleich zweimodaler Clusteranalyseverfahren", Methods of Operations Research*, 57, 593-605.
5. Brand, Matthew (2003), "*Fast online SVD revisions for lightweight recommender systems", Proceedings of the SIAM 3rd International Conference on Data Mining*, 37-48.
6. Breese, John S., Heckerman, David, and Kadie, Carl (1998), *"Empirical Analysis of Predictive Algorithms for Collaborative Filtering", Proceedings of the 14-th Annual Conference on Uncertainty in Artificial Intelligence*, 43-52.
7. Breiman, Leo, Friedman, H.,Olshen, J.A., and Stone, C.J. (1984), *"Classification and Regression Trees"*, *Chapman & Hall*.
8. DeSarbo, Wayne S. (1982), *"Gennclus: New Models for General Nonhierarchical Clustering Analysis", Psychometrika*, 47, 449-475.
9. DeSarbo, Wayne S., DeSoete, Geert, Carroll, J. Douglas, Ramaswamy, Venkatram (1988), *"A New Stochastic Ultrametric Tree Methodology for Assessing Competitive Market Structure"*, *Applied Stochastic Models and Data Analysis,* 4, 185-204.
10. Dhillon, Inderjit S., Mallela, Subramanyam, Modha, Dharmendra S. (2003), *"Information-Theoretic Co-clustering", Proceedings of the 9-th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 89-98.
11. Gaul, Wolfgang and Schader, Martin (1996), *"A New Algorithm for Two-Mode Clustering", in Data Analysis and Information Systems, Hans-Hermann Bock and Wolfgang Polasek, eds.*, 15-23.
12. Espejo, E. and Gaul, Wolfgang (1986), *"Two-Mode Hierarchical Clustering as an Instrument for Marketing Research", in Classification as a Tool of Research, Wolfgang Gaul and Martin Schader, eds.*, 121-128.
13. Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert (2000), *"A Statistical View of Boosting"*, *The Annals of Statistics*, 28, 337-374.
14. George, Thomas and Merugu, Srujana (2005), *"A Scalable Collaborative Filtering Framework based on Co-Clustering", Proceedings of the Fifth IEEE Conference on Data Mining (ICDM)*, 625-628.
15. Gelman, Andrew, Carlin, John B., Stern, Hal S., Rubin, Donald B. (2004), *"Bayesian Data Analysis"*, *Chapman & Hall*.
16. Kim, Jin H., Pearl, Judea (1983), *"A Computational Model for Causal and Diagnostic Reasoning in Inference Engines", Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 190-193.
17. Lauritzen, Steffen L., Spiegelhalter, David (1988), *"Local Computations with Probabilities on Graphical Structures and their Applications to Expert Systems", Journal of the Royal Statistical Society B*, 50, 157-224.
18. Landwehr, Niels, Hall, Mark, and Frank, Eibe (2005), *"Logistic Model Trees"*, *Machine Learning,* 59, 161-205.
19. Magnus, Jan R., Neudecker, Heinz (1988), *"Matrix Differential Calculus with Applications in Statistics and Econometrics", Wiley*.

20. McCulloch, Robert, Rossi, Peter E. (1994), *"An Exact Likelihood Analysis of the Multinomial Probit Model", Journal of Econometrics*, 64, 207-240.
21. McFadden, Daniel (1974), *"Conditional Logit Analysis of Qualitative Choice Behavior", in: Zarembka, Paul (Ed.): Frontiers of Econometrics, Academic Press*, 105-142.
22. Noma, Elliot and Smith, D. Randall (1985), *"Benchmark for the Blocking of Sociometric Data", Psychlogical Bulletin,* 97 (3), 583-591.
23. Pearl, Judea (1982), *"Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach", Proceedings of the Second National Conference on Artificial Intelligence*, 133-136.
24. Pearl, Judea (1986), *"Fusion, Propagation, and the Structuring in Belief Networks", Artificial Intelligence*, 29, 241-288.
25. Quinlan, J. Ross (1993), *"C4.5: Programs for Machine Learning", Morgan Kaufman*.
26. Resnick, Paul, Neophytos, Iacovon, Mitesh, Suchak, Bergstrom, Peter, and Riedl, John (1994), *"Grouplens: An Open Architecture for Collaborative Filtering of Netnews"*, *Proceedings of the Conference on Computer Supported Cooperative Work,* 175-186.
27. Rossi, Peter E., McCulloch, Robert E., and Allenby, Greg M. (1996), *The Value of Purchase History Data in Target Marketing*, *Marketing Science,* 15, 321-340.
28. Sarwar, Badrul M., Krypis, George, Konstan, Joseph A., and Riedl, John (2002), *"Incremental SVD-based algorithms for highly scaleable recommender systems", Proceedings of the 5th International conference on Computer and Information Technology*, 173-181.
29. Schlecht, Volker und Gaul, Wolfgang (2004), *"Fuzzy Two-mode Clustering vs. Collaborative Filtering", Classification – the Ubiquitous Challenge, Weihs, Gaul, eds,* 410-417.
30. Shachter, Ross (1986), *"Evaluating Influence Diagrams", Operations Research*, 34, 871-882.
31. Shen, Bin, Su, Xiaoyuan, Greiner, Russell, Musilek, Petr, and Cheng, Corrine (2003), *"Discriminative Parameter Learning of General Bayesian Network Classifiers"*, *Proceedings of the Fifteenth IEEE International Conference on Tools with Artificial Intelligence*.

Appendix A. Estimation of item-cluster membership based on known procedures

The following procedures can be used to approximate the item-cluster membership of new items.

Logistic Regression (LR). The regression parameters of the multinomial logit model (McFadden, 1974) can be calculated based on all items from the set J^K with the item classes

$$q_j^c = \arg\max_{l' \in \{1,...,L\}} q_{jl'},$$

as realization of the (nominal) dependent variable and the components of the vector a_j as independent variables for all $j \in J^K$. (Since every item j belongs to exactly one cluster, $|\{l \in \{1,...,L\} \mid q_{jl} = 1\}| = 1$, which implies, that q_j^c is a well-defined dependent variable.) Let $C_2(l)$ be the set of elements, that belong to the l-th item-cluster. Then the parameters of this model $\hat{\beta}_{A,0}^l, \hat{\beta}_A^l, l \in \{1,...,L\},$ can be used to estimate each new item's ($j \in J^N$) probability of belonging to a particular item class based on its attributes a_j :

$$P(j \in C_2(l) \mid a_j) = \frac{\exp(\hat{\beta}_{A,0}^l + a_{j'} \hat{\beta}_A^l)}{\sum_{l'=1}^{L} \exp(\hat{\beta}_{A,0}^{l'} + a_{j'} \hat{\beta}_A^{l'})}.$$

Thus, the probabilities of cluster-membership are estimated by $\hat{P}(j \in C_2(l) \mid a_j)$, from which discrete degrees of cluster-membership can be inferred by using

$$\lambda_{LR}(j_1) = \arg\max_{l \in \{1,...,L\}} \hat{P}(j_1 \in C_2(l) \mid a_j) \quad \rightarrow \quad \hat{q}_{j_1 l}^{LR,d} = \begin{cases} 1, & if\ l = \lambda_{LR}(j_1) \\ 0, & otherwise \end{cases}.$$

Decision Tree (C4.5). In the same way a decision tree like the C4.5 algorithm by Quinlan (1993) may be built based on J^K, which is called the training-set. Item-classes for the new items can be determined by applying the induced decision tree to their attributes $a_j, j \in J^N$. Here, J^N is referred to as test-set. Again, continuous degrees of cluster-membersip are calculated and discrete cluster-membership can be inferred.

Bayesian Multinets (BMN). Bayesian Multinets (BMN) are another promising but less well-known Machine Learning procedure for using J^K as training-set in order to estimate the class-membership for the test-set J^N. Since Bayesian Multinets consist of local Bayesian Networks, it is necessary to explain Bayesian Networks first.

Every Bayesian Network (X, E, θ) consists of a directed acyclic graph (X, E), whose nodes $X_1,...,X_M$, ($X = \{X_1,...,X_M\}$) can represent random variables and whose arcs E express the conditional dependencies between those random variables. Another important part of every Bayesian Network is $\theta = \{\theta_1,...,\theta_M\}$. For every node $X_\mu, \mu = 1,...,M$ a conditional probability $\theta_\mu = P(X_\mu \mid \pi(X_\mu))$ exists. Here, $\pi(X_\mu)$ denotes the parents of node X_μ. The Bayesian Network encodes a joint probability distribution over a set of random variables $X_\mu, \mu = 1,...,M$.

If the discrete random variable X_1 describes to which class or cluster a certain item $j \in J^K$ belongs and all the remaining random variables $X_2,...,X_M$ are the elements of the item attribute vector a_j ($M = 1 + \kappa_A$), a Bayesian Network can be used for classification purposes.

Regardless of the value X_1 takes, a Bayesian Network for a specific set of training data will express the same conditional dependencies among the random variables – even if for some given values of X_1 some of those conditional dependencies do not exist in the training data set (Geiger and Heckerman, 1996). E.g., if the items to be classified are movies it could be, that for a group (class) consisting mainly of horror, action and adventure movies dependencies exist, which are different from the dependencies which can be discovered for classes which essentially contain comedies, romances, dramas and documentaries.

Bayesian Multinets (Geiger and Heckerman, 1996) allow for those different dependencies. Let X_1 denote the cluster membership variable. To the latter the set of possible realizations $\{1,...,L\}$ belongs. A Bayesian Multinet for the classification of objects into L groups (classes) consists of L Bayesian Networks $(\tilde{X}, E^l, \theta^l)$ with (set of) nodes $\tilde{X} = \{X_2,...,X_M\}$, conditional dependencies E^l and conditional probabilities θ^l, which depend on the value l, which is taken by X_1 and the probabilities of class-membership $P(X_1 = l), l = 1,...,L.$ Each of those Bayesian Networks $(\tilde{X}, E^l, \theta^l)$ is determined based on items, which belong to (the respective) class $l, l = 1,...,L$. The conditional probabilities are $\theta^l_\mu = P(X_\mu \mid \pi(X_\mu), X_1 = l)$. Given empirical approximations of θ^l_μ and $P(X_1 = l)$, $\hat{P}(X_\mu \mid \pi(X_\mu, X_1 = l)$ and $\hat{P}(X_1 = l)$, the probabilities of cluster-membership can be estimated by

$$\hat{P}(X_1 = l \mid X_2 = a_{j1},..., X_M = a_{j\kappa_A}) = \frac{\prod_{\mu'=2}^{1+\kappa_A} \hat{P}(X_{\mu'} \mid \pi(X_{\mu'}), X_1 = l)\hat{P}(X_1 = l)}{\sum_{l'=1}^{L}\prod_{\mu'=2}^{1+\kappa_A} \hat{P}(X_{\mu'} \mid \pi(X_{\mu'}), X_1 = l')\hat{P}(X_1 = l')}.$$

Friedman et al. (1997) have shown that Bayesian Multinets outperform Bayesian Networks and Naive Bayes estimation and also compare favorably to C4.5. First, the training set J^K is used to estimate both the structure of the directed acyclic graph and the conditional probabilities of the resulting Bayesian Multinet. Then, the test set can be classified via the resulting estimators $\hat{P}(X_1 = l \mid X_2 = a_{j1},..., X_M = a_{j\kappa_A}),\ l = 1,...,L, j \in J^N$. Like logistic regression the Bayesian Multinets estimate probabilities for class-membership. Those probabilities can be used to derive the discrete cluster-membership via the usual procedure.

Logistic Model Trees (LMT). Logistic Model Trees (LMT) were recently introduced by Landwehr et al. (2005). The idea behind this technique is to combine logistic regression with tree induction.

First a standard classification tree has to be built, which is accomplished by the C4.5 algorithm. At each node of the resulting tree the data are further partitioned into disjoint data subsets with respect to a specific attribute. Starting from the root node at each node of the tree a logistic model is fitted to the subset associated with the current node by the LogitBoost algorithm (Friedman, Hastie, and Tibshirani, 2000). Then the tree is pruned with the help of error complexity pruning, which is the pruning method that is also used by the CART algorithm (Breiman, Friedman, Olshen, and Stone, 1984).

For the sake of computational efficiency the regression results from each parent node are to be used as starting values for the estimation of the logistic regression model with respect to the subsets of data associated with its child nodes (Landwehr et al., 2005). Hence, the logistic model for the subset connected with the parent node has to be estimated before the logistic regressions, which are based on the subsets of data, that are associated with its children, can be carried out.

Finally, each specific item is matched to the end node, which corresponds to its attributes. The logit model, which belongs to this end node is used to calculate the probability of the item's cluster-membership. Again, the cluster-membership is approximated according to the cluster-membership probabilities.

Appendix B. Approximation of item-cluster membership based on simple heuristics

In addition to the already known methods, which were described in Appendix A, three simple heuristics can be introduced in order to approximate item-cluster membership. Those new procedures are all based on dissimilarity measures between items.

Let $a_{j\kappa}$ be the κ-th attribute of the j-th item and $\kappa \in \{1,...,\kappa_A\}$. κ_A is the number of item attributes used in the model. Furthermore, let $a_j = (a_{j1},...,a_{j\kappa_A})'$ be the vector of attributes which describes the *j*-th item. Heuristically one can define the dissimilarity $d(j_1, j_2)$ between two different items j_1 and j_2 as the weighted Euclidean distance between a_{j_1} and a_{j_2}:

$$d(a_{j_1}, a_{j_2}) = \sqrt{\sum_{\kappa'=1}^{\kappa_A} \nu_{\kappa'}(a_{j_1\kappa'} - a_{j_2\kappa'})^2}.$$

The dissimilarities $d(a_{j_1}, a_{j_2})$ of objects $j_1 \in J^N$ and $j_2 \in J^K$ are the basis of 3 different heuristics. Here, J^K is the set of all known items and J^N is the set which contains all new items.

***SL*-Heuristic.** Let $j_{1'} = \text{argmin}_{j_2 \in J^K} d(a_{j_1}, a_{j_2})$ be the known second mode element, which is most similar to the new item j_1. The *SL*-heuristic then simply assigns j_1 to the second mode cluster to which $j_{1'}$ belongs:

$$\lambda_{SL}(j_1) = \arg\max_{l' \in \{1,...,L\}} q_{j_1 l'} \quad \rightarrow \quad \hat{q}^{SL}_{j_1 l} = \begin{cases} 1, & \text{if } l = \lambda_{SL}(j_1) \\ 0, & \text{otherwise} \end{cases}.$$

***KM*-Heuristic.** The *KM*-heuristic first calculates for each cluster of second mode elements the vector of average item-cluster attributes:

$$\bar{a}^l = \frac{\sum_{j \in J^K} q_{jl} a_j}{\sum_{j \in J^K} q_{jl}}.$$

Then each new item j_1 is assigned to the cluster, to whose vector average item-cluster attributes $\bar{a}^l$ it has the smallest Euclidean distance:

$$\lambda_k(j_1) = \arg\max_{l' \in \{1,...,L\}} d(a_{j_1}, \bar{a}^{l'}) \quad \rightarrow \quad \hat{q}^{k}_{j_1 l} = \begin{cases} 1, & \text{if } l = \lambda_k(j_1) \\ 0, & \text{otherwise} \end{cases}.$$

***a*-Heuristic (continuous version).** Unlike the first two heuristics the *a*-heuristic has a discrete and a continuous version like the procedures introduced in Appendix A. The *a*-Heuristic calculates a degree of membership for every $j_1 \in J^N$ and $l \in \{1,...,L\}$

$$\tilde{q}_{j_1 l} = \frac{\sum_{j_2 \in J^K} \left(\frac{1}{d(a_{j_1}, a_{j_2})} \right)^{\alpha} q_{j_2 l}}{\sum_{l'=1}^{L} \sum_{j_2 \in J^K} \left(\frac{1}{d(a_{j_1}, a_{j_2})} \right)^{\alpha} q_{j_2 l'}},$$

and its continuous version uses the (continuous) $\tilde{q}_{j_1 l}$ instead of $q_{j_1 l}$. Here α is a parameter which is usually set to 1 or 2.

***a*-Heuristic (discrete version).** The discrete version of the *a*-heuristic assigns each $j_1 \in J^N$ to the cluster with the biggest degree of membership:

$$\lambda_a(j_1) = \arg\max_{l \in \{1,...,L\}} \tilde{q}_{j_1 l} \quad \rightarrow \quad \hat{q}^{a}_{j_1 l} = \begin{cases} 1, & \text{if } l = \lambda_a(j_1) \\ 0, & \text{otherwise} \end{cases}.$$