Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Journal of Retailing and Consumer Services 10 (2003) 123–133
JOURNAL OF RETAILING AND CONSUMER SERVICES
www.elsevier.com/locate/jretconser

An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data

Andreas Mild[a,*], Thomas Reutterer[b]

[a] Department of Production Management, Vienna University of Economics & BA, Pappenheimgasse 35, A-1200 Vienna, Austria
[b] Department of Retailing and Marketing, Vienna University of Economics & BA, Augasse 2-6, A-1090 Vienna, Austria

Abstract

Retail managers have been interested in learning about cross-category purchase behavior of their customers for a fairly long time. More recently, the task of inferring cross-category relationship patterns among retail assortments is gaining attraction due to its promotional potential within recommender systems used in online environments. Collaborative filtering algorithms are frequently used in such settings for the prediction of choices, preferences and/or ratings of online users. This paper investigates the suitability of such methods for situations when only binary pick-any customer information (i.e., choice/non-choice of items, such as shopping basket data) is available. We present an extension of collaborative filtering algorithms for such data situations and apply it to a real-world retail transaction dataset. The new method is benchmarked against more conventional algorithms and can be shown to deliver superior results in terms of predictive accuracy.
© 2003 Elsevier Science Ltd. All rights reserved.

Keywords: Collaborative filtering; Recommender systems; Market basket analysis

1. Introduction

Consumers are permanently involved in multi-category decision making, such as grocery shopping trips, mail-order purchasing, or financial portfolio choice. In a retailing context, such multi-category decision processes result in the formation of shopping or market baskets which comprise the set of categories (or items) that individual consumers purchase on one and the same purchase occasion. Both on- and off-line retailers are traditionally interested in under-standing the composition of their customers' market baskets, since valuable insights for designing micro-marketing and/or targeted cross-selling programs can be derived (cf. Russell and Kamakura, 1997).

More recently, the extensive diffusion of in-store scanning technologies, and especially the availability of customers' navigation and clicking information through online environments like the web, makes mass-customization of both content (i.e., information, products, or product categories offered to web-site visitors at specific prices and conditions) and design elements via interactive electronic media eminently possible. Another consequence of the high flexibility and the virtually unlimited 'shelf space' of online retail environments is the capability of online retailers to provide consumers with a very large number of products available at greatly reduced search costs (cf. Bakos, 1997). On the other (the customer) side, increasingly complex online retail assortments very rapidly risk to bust the constraints imposed by the cognitive limitations of human information processing. Consequently, assistance of automated recommendation agents is called for reducing the complexity at the customer end of the market (Alba et al., 1997).

Online firms, such as amazon.com, cdnow.com, or barnesandnoble.com, are therefore using personalized recommendation systems that suggest to customers lists of items on the basis of the preferences of their other customers. As a consequence, the analysis of (online) market basket data for fine-tuning a company's offerings at the individual level is (re-)attracting the interest of marketing researchers. The introduction of such recommendation systems in computer-mediated shopping environments is often posited to have an impact on

*Corresponding author. Tel.: +43-1-31336-5628; fax.: +43-1-31336-5610.
E-mail addresses: andreas.mild@wu-wien.ac.at (A. Mild), thomas.reutterer@wu-wien.ac.at (T. Reutterer).

0969-6989/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0969-6989(03)00003-1

the composition of consumers' consideration sets both in terms of size and quality (see, e.g., Alba et al., 1997; Winer et al., 1997). Using content filtering on self-explicated attribute importance weights to generate recommended personalized item lists, Haeubl and Trifts (2000) were among the first who provided empirical evidence that online shoppers can improve both the quality and the efficiency of purchase decisions (see also West et al., 1999). Hence, it may be expected that increasing the predictive accuracy of recommendation systems is beneficial to customer satisfaction and loyalty, at the same time enhancing customer profitability and product return.

Most of the customized recommendation systems currently operating in real-world online shopping environments are based on so-called collaborative filtering (CF) methods. These methods are mimicking word-of-mouth recommendations by using data from users with similar preferences in order to determine an active customer's preferences and the recommended item list derived from these. The majority of the CF algorithms presented so far, however, are predominantly designed for the analysis of customers' preference rating data (such as the GroupLens research project; see, e.g., Konstan et al., 1997) which makes them less appropriate for the binary world of retail transaction data (i.e., choice/non-choice of customers among product categories offered in retail assortments).

In the remainder of this paper, we first provide a brief overview of related research on market basket data analysis and recommendation systems. Encouraged by pertinent critics and limitations included in the relevant literature (see, e.g., Breese et al., 1998; Ansari et al., 2000) as well as recent empirical findings reported in an experimental study using conventional CF algorithms conducted by Mild and Reutterer (2001), we next propose an improved CF approach designed for predictions of customers' product category choices. The performance of this modified CF method is illustrated using market basket data across 54 product categories from a typical grocery retail assortment. Using this dataset, we compare the predictive accuracy of our approach to results obtained from standard CF methodology and a probabilistic approach for predicting cross-category dependencies. Finally, we draw conclusions and discuss the practical significance of our results.

2. Background and related literature

We divide the following discussion into two subsections: First, an overview of existing approaches to choice-based modelling of cross-category preferences (market basket analysis) is provided. The CF methodology is then integrated into that framework, and a research agenda for proposing an extension of the basic CF procedure that is more qualified for the analysis of category choice data is briefly outlined.

2.1. Modelling multiple-category preferences using market basket data

A market basket consists of a set of items (or categories) purchased by a customer during one single shopping occasion. The methodological toolbox enabling researchers to study the composition of such baskets (or bundles) of products is usually referred to as market basket analysis. Two papers by Russell et al. (1997, 1999) provide reviews of state-of-the-art methods for market basket data analysis. According to the terminology adopted by marketing researchers from data theory (cf. Coombs, 1964), market basket data are qualified as 'pick-any/J', whereas the choice set (here: the basket size) can be constrained by a maximum number of products (categories) J or unconstrained, i.e. 'pick-any'.

In this respect, the logic of numerous approaches to market structuring is applicable for such a pick-any type of dominance data and, thus, suitable for the task of deriving representations of inter-category relationship patterns inherent in market basket data. The contributions of DeSarbo et al. (1993, 1994) present an excellent overview of contemporary tree structure and multidimensional scaling methods as used in the market structure literature. Models of multiple-category demand, however, have to consider interrelationships between brands or categories that are unrestricted, i.e. substitutional, complementary or independent (cf. Russell and Kamakura, 1997).

Conventional methods for market basket analysis can be classified into exploratory and explanatory approaches to measuring such cross-category relationships. This differentiation criterion is also included in an extensive but not exhaustive summary of existing approaches compiled in Table 1. Exploratory approaches try to uncover and condense the complex interdependency structures typically observed among multiple categories in a managerially meaningful manner. A few applications are preprocessing pick-any choice data using simple association measures that indicate coincidence or affinity of (pairs of) items in market baskets to identify product category relationships (see references under [1] in Table 1). Configurations of relationship patterns hidden in such a priori aggregated matrices of product category associations can also be derived via decomposition using methods introduced in the market structuring literature. Another prominent stream of research in the field of exploratory market basket analysis (see [2] in Table 1) utilizes data mining techniques to generate association rules among subsets of product categories for a given database of

Table 1
Overview of various existing approaches to analyzing market basket information

Method and selected references	Type of analysis (characteristics)	Primary task of analysis	Level of aggregation	Marketing mix
[1] Pairwise associations: Boecker (1978), Hruschka (1985), Dickinson et al. (1992), and Julander (1992)	Exploratory (affinity analysis)	Representation of relationships	Aggregate	No
[2] Association rules: Agrawal and Srikant (1994), Buechter and Wirth (1998), and Hilderman et al. (1998)	Exploratory (mining of large datasets)	Discover rules for symptomatic category purchase associations	Aggregate	No
[3] Vector quantization: Schnedlitz et al. (2001)	Exploratory (compression of basket data)	Respresentation of segment-level associations	Disaggregate (segment)	No
[4] Finite mixture model: Russell and Kamakura (1997), and Andrews and Currim (2002)	Exploratory or explanatory (latent class analysis)	Identification of segments with homogeneous basket composition	Disaggregate (segment)	Possible via inclusion in RUT framework
[5] Multivariate logistic model: Hruschka (1991), and Hruschka et al. (1999), Russell and Petersen (2000)	Exploratory or explanatory (measurement of category inter-dependencies)	Estimation and prediction of cross-category effects	Aggregate	Possible via inclusion in RUT framework
[6] Regression analysis: Walters (1991), and Mulhern and Leone (1991)	Explanatory (ridge or seemingly unrelated regression	Assessing impact of price on product and category choices	Aggregate	Yes (price and others)
[7] Intercategory choice dynamics: Harlam and Lodish (1995), Chintagunta and Haldar (1998)	Explanatory (hazard and sequential choice models)	Choice sequencies and purchase timing across categories	Aggregate	Yes (price and others)
[8] Logit/probit model: Ainslie and Rossi (1998), Kim et al. (1999), Seetharaman et al. (1999), and Manchanda et al. (1999)	Explanatory (empirical Bayesian analysis of variance components)	Modelling of intercategory choice decisions in a RUT framework	Disaggregate (individual level)	Yes (price and others)

shopping baskets. Quantization of similarities among pick-any/J vectors of item choices is postponing the aggregation step performed by the above-mentioned exploratory approaches prior to analysis. The use of advanced vector quantization techniques allows a disaggregated (segment-level) representation of relationship patterns between product categories (for an application, see [3] in Table 1).

While the primary task of exploratory approaches is to uncover and represent hidden category relationships in shopping baskets, explanatory models focus on the identification and quantification of complementary cross-category choice effects of some marketing variables under managerial control, such as price, promotions, or in-store marketing features. Traditional contributions in this direction as cited under Section [6] in Table 1 are using variants of regression analysis. Sequential choice models or sophisticated bivariate hazard and probit models are developed to further include interpurchase timing effects on cross-category choices (for applications, see [7] in Table 1). Especially with the adoption of random utility theory (RUT) in a finite mixture modelling framework as demonstrated by authors under Section [4] in Table 1, the inclusion of (cross-category) marketing mix effects in a model for identification of consumers with homogeneous choice behaviors across product categories becomes possible.

A particulary interesting approach suitable for predicting cross-category effects under various marketing mix conditions is presented by Hruschka (1991). The author proposes a probabilistic model for estimation of conditional purchase probabilities within product categories. Using a set of logit equations, this model in principle allows for the incorporation of both direct effects on category choice from other categories and cross-category dependencies by specification of interaction effects. Due to parameter restrictions, for applications using real-world market basket data the approach

is usually restricted to first-order (i.e., pairwise category) effects only. The choice model for all market baskets can be expressed as a multivariate logistic (MVL) distribution and estimated using model selection criteria for a typically large number of parameters. In an extension of this approach, Hruschka et al. (1999) introduce cross-category sales promotion effects influencing the category choice probabilities. Consistent with the work of Hruschka and colleagues, a similar model is presented by Russell and Petersen (2000). In their MVL model, the authors allow for variations of household characteristics, marketing mix variables, and any type of demand relationships (complementary, independence, or substitution) across product categories to effect the conditional category choice probabilities in a random utility framework.

The MVL approach to modelling the pick-any/J choice task, however, is confronted with estimation problems as J becomes larger and reaches real-world shopping basket sizes. In this regard, with the continuing improvement of Markov chain Monte Carlo (MCMC) simulation methodologies, individual-level estimation of the full conditional models can be expected to be possible. Such empirical Bayesian approaches for the estimation of logit- and probit-type analyses of market basket data have already been applied, as mentioned under Section [8] of Table 1.

2.2. Collaborative filters as recommendation agents for binary choice data

The current customized recommender systems can be classified into two main classes, namely content-based and CF-methods (for a detailed overview of contemporary CF approaches, cf. Runte, 2000). In content-based filtering approaches, recommendations are made on the basis of consumer preferences for product attributes in order to retrieve items, such as relevant textual documents (see, e.g., Salton and Buckley, 1988; Maes, 1994), with a content most similar to a specific customer's interests. The task of CF is to predict preferences of an active user given a data-base of preferences of other users, where preferences are typically represented as dominance data. The latter are typically recorded as explicit preference ratings obtained from users on a subset of available items or implicit behavioral reactions (such as purchase frequencies, click-stream, or choice data) regarding a given item set.

Memory-based CF algorithms, as outlined in more detail in the next section, are deterministic by nature. They rely on a database of previous users' preferences and perform certain calculations (similarity matching) on the database each time a new prediction is required (see, e.g., Breese et al., 1998). The most common representatives are neighbor-based algorithms where a subset of users most similar to an active user is chosen and a weighted average of their preference ratings is used to estimate preferences of an active user on other items (see, e.g., Konstan et al., 1997; Sarwar et al., 2000). Model-based algorithms, including Bayesian clustering, Bayesian networks and other classification-based algorithms (see, e.g., Breese et al., 1998; Ungar and Foster, 1998), first develop a descriptive model of the database and use it to make predictions for an active user.

According to the criteria used in the synopsis as outlined in Table 1, when applied to study market basket data CF methods can be qualified as an exploratory analysis for making predictions of multiple item choices at the individual customer level. As discussed below in more detail, the predictions derived from CF procedures are crucially dependent on the way how similarity matching between the category choice profiles of various users is accomplished. Hence, there is no explicit behavioral model behind the derived item recommendations. Furthermore, since CF algorithms neither account for any marketing mix variations across purchase occasions nor for differences in individual household characteristics, they rely on an implicit stability assumption of customers' preference structures (see also Ansari et al., 2000).

In a comparative study using alternative CF methods, Breese et al. (1998) report neighbor-based CF algorithms to be superior to model-based approaches in terms of predictive accuracy. However, there are several other data-related factors, such as sparsity of item lists included in the database or the amount of information available per person, that are frequently reported to critically affect the performance of CF-based recommender systems as well. To circumvent potential problems caused by data sparsity, Ansari et al. (2000) propose a hierarchical Bayesian recommendation system that makes use of additional demographic customer data and external expert ratings. As an empirical Bayesian model, the system incorporates 'learning' and is suitable to exploit various types of information gathered from users included in the database in order to derive more accurate recommendations. Unfortunately, most of the published work on CF methods including the model developed by Ansari et al. (2000) is elaborated on preference ratings, and there are only limited indications available in the literature of how these methods perform when using binary pick-any choice data derived by customers' shopping baskets. Mild and Reutterer (2001) are experimenting with various settings of standard memory-based CF algorithms and attest only poor performance in terms of predictive accuracy of actual product category choices. In order to alleviate these limitations, we next outline a modified CF approach designed for making recommendations based on such data.

3. Model description

The formal structure of our model for personalized recommendations of product categories based on an individual customer's shopping basket composition is derived from memory-based CF algorithms which are typically proceeding in a two-step framework. For a given number of I customers and J product categories included in the database (in the relevant literature on recommender systems, customers are frequently denoted as 'users' and product categories as 'items'), the conventional CF process involves completion of the following two tasks (Karypis, 2000; Sarwar et al., 2000):

1. Calculation of numerical predictive values $p_{a,j}$, expressing the purchase or choice likeliness of item $j \in J$ for the 'active' (or online) user a with the observed current shopping basket $c_{a,j}$,
2. Building of a recommended list of N items that the active user is expected to prefer (and choose) most likely based on the item-specific predictive values (this task of CF is also known as '*Top-N* recommendation').

Notice that the shopping basket for the active customer $c_{a,j}$ as well as each basket of all other customers $c_{i,j}$ $(i \in I)$ included in the database are represented as J-dimensional binary (pick-any) vectors with elements coded as 0 indicating 'no choice' and 1 indicating 'choice' of an item or product category. Naturally, the number of items to be included in a user-specific recommendation list is substantially smaller than the total number of available items ($N \ll J$), with a size of $N=1$ representing the extreme of recommending to the active user the most likely preferred item j^*. In the latter case, the predictive value of the recommended item is required to exceed the predictors of all other items by meeting the condition $p_{a,j^*} > p_{a,j\neq j^*}, \forall j \in J$. For practical applications, however, suitable sizes of recommendation lists could be easily determined by deducting the number of items already included in the active user's basket (i.e., the number of one-coded values in the vector $c_{a,j}$) from the average basket size observed in the customer database.

In step (1) of the CF process, the predictive value $p_{a,j}$ for the active user a and a specific item j is computed on the basis of a weighted sum of the 'votes' of other similar users. The term 'vote' is frequently used in the CF literature as an expression of the rating scores provided by users; in the context of shopping basket data as in our subsequent empirical application, 'vote' is corresponding to the binary choice vectors $c_{i,j}$. We use a modified version of the function proposed by Breese et al. (1998):

$$p_{a,j} = \kappa \sum_{i=1}^{I} w(a,i) c_{i,j}. \tag{1}$$

The propensity of a user a to purchase item j thus depends on the similarity weights $w(a, i)$ between the active user a and each individual other user i from the available database and the actual shopping behavior $c_{i,j}$ of the respective user i. κ is a normalizing factor to ensure that the absolute values of the weights sum to unity. Most studies on CF methods make use of correlation coefficients (see, e.g., Resnick et al., 1994) or a measure of vector similarity based on the cosine of the angle between two vectors (Sarwar et al., 2000) for the calculation of the similarity between users. Since there is only very limited variance to be expected in similarities (using, e.g., correlational measures) constructed for very sparse binary datasets, we propose the usage of the well-known Jaccard or Tanimoto coefficient (for a brief description of the properties of this proximity measure, see, e.g., Anderberg, 1973, or Kaufman and Rousseeuw, 1990). The Tanimoto similarity between two users a and i is defined as

$$w(a,i) = \frac{n(c_a \cap c_i)}{n(c_a \cup c_i)} = \frac{n(c_a \cap c_i)}{n(c_a) + n(c_i) - n(c_a \cap c_i)}, \tag{2}$$

where $n(X)$ represents the number of elements in the customer basket (or item-set) X. As is obvious from the above description, the Tanimoto coefficient ignores the number of coinciding non-chosen elements (i.e., zeros). In an experimental study testing the impact of various proximity measures and data-related conditions (such as sparsity of item lists and available amount of information available per customer) on the predictive accuracy of conventional CF algorithms, this property is shown to be clearly advantageous in the case of extremely asymmetric distributed or sparse data vectors like shopping basket data (cf. Mild and Reutterer, 2001).

So far, the CF algorithm arrives at user-item specific predictive values, which normally are used directly for the subsequent personalized item recommendations according to stage two of the procedure. It can be shown, however, that the use of the raw predictive values $p_{a,j}$ for constructing (typically 'Top-*N*') recommended item lists proves rather problematic when this method is applied to real-world data: First, item-specific mean predictive values $\bar{p}_j$ depend on the overall purchase frequency of each single item. Therefore, a decision rule using a general threshold value for recommending items (such as, e.g., $\bar{p}_j = 0.5$) is biased by the unequal purchase frequency distribution across items typically observable in real-world shopping baskets. Secondly, there exists no unique measure that tells us for which items the predictor delivers accurate recommendation results.

Using the same database as employed in the subsequent empirical application, Fig. 1 illustrates this for the case of two extreme product categories by opposing mean predicted values with their dispersion ranges (indicated as 0.95 percentile error bars) for customers who have actually chosen and such who have not chosen an item from the respective categories. Notice, that in both cases the application of a naive but

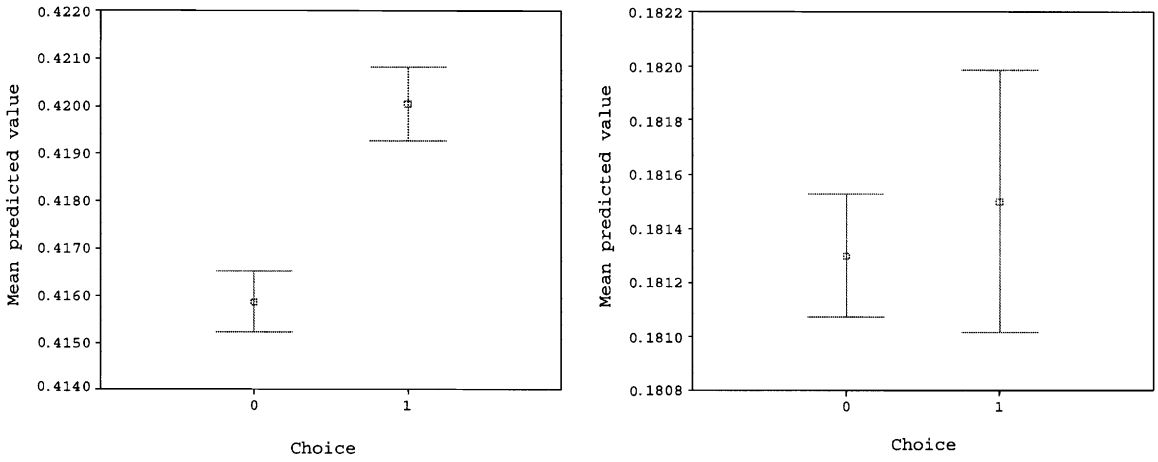

Fig. 1. Predictable (left-handed) items are characterized by predictors that clearly discriminate between actual choices and non-choices. If the dispersions of predictive values for the choice/non-choice options are highly overlapping (right-handed), the item group is omitted from the candidate list of recommended items.

nevertheless 'plausible' candidate threshold, such as $\bar{p}_j = 0.5$, would result in a rejection of the categories from a item recommendation list. Irrespective of absoulte values, however, mean predictive values for the left-hand item are capable of clearly discriminating between observable category choices and non-choices in the customer database. For the right-hand side situation, the opposite applies. Due to the high variation, and more precisely the overlapping dispersion of predictive values for the choice and non-choice option, raw predicted values turn out to be poor predictors of category choice, thus leading to inaccurate recommendations.

For recommendation purposes, we therefore propose an extension of the basic CF algorithm by formulating a decision rule with respect to the predictability of single items. As shown in the above illustration, the accuracy of predictive values as derived by Eq (1) for correctly discriminating between chosen and non-chosen items crucially depends on the separability of the dispersions around respective mean values. To account for this, in the recommendation stage of the CF procedure we derive item-specific thresholds based on a percentile of the dispersion around two empirical means before inclusion in the list of recommended items. The rationale is as follows: Given a customer database of choices among categories of a retail assortment, we compute the confidence intervals for the predictors of each item in the subsamples for both actual choice and non-choice realizations. Using the confidence interval for the mean and an unknown variance, the upper confidence bound for the non-choice option of item j is defined as follows:

$$c^u_{0,j} = \bar{p}_{0,j} + z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\bar{p}_{0,j}}, \tag{3}$$

where $\bar{p}_{0,j}$ represents the mean predicted value for actual non-choice item realizations with sample standard deviation $\hat{\sigma}$ and $z_{(1-(\alpha/2)}$ as the (one sided) normal inverse distribution function with error level α.[1] Analogously, the lower confidence bound for an actual purchase of item j is:

$$c^l_{1,j} = \bar{p}_{1,j} - z_{(1-\frac{\alpha}{2})}\hat{\sigma}_{\bar{p}_{1,j}}. \tag{4}$$

Based on these simple computations on the collection of raw predictive values we now can select items for inclusion in recommendation lists by identifying items which satisfy the following condition:

$$c^l_{1,j} > c^u_{0,j}. \tag{5}$$

Items with raw predictive values that do not meet this condition are skipped from the candidate list of recommended items due to their desiderative discriminatory power between observed choices and non-choices of the respective item from shopping baskets in the existing customer database. For the remaining items we can use the values $c^l_{1,j}$, as item-specific thresholds for p_j; i.e., all items, whose predictions for an active user $p_{a,j}$ are exceeding $c^l_{1,j}$, should be recommended (or included in a recommendation list of a priori, say 'Top-N', fixed size in descending order of their, e.g., product returns) and can be expected to be purchased with a probability of 1—α. In the next Section, we will demonstrate the performance of the proposed modified CF algorithm using real-world retail transaction data.

4. Data and design of the empirical demonstration study

For empirical validation, we are using market basket data across 54 product categories. The data are representing 2241 grocery retail transactions (i.e.,

[1] Note that for small sample sizes (<30), this distribution should be replaced by the student t inverse distribution function.

customers' purchases of multiple product categories), each basket containing a minimum of five items or categories. For validation purposes, we split the available data into a training and a hold-out sample. The training sample consists of 1000 shopping baskets for model estimation while the remaining 1241 baskets are used for generating recommendations in order to evaluate the predictive accuracy of the modified CF method with item-level recommendation thresholds proposed above, subsequently denoted as CF_{mod}, relative to the performance of alternative approaches.

Both the prediction and the recommendation steps of CF are performed for each single item using purchase information (i.e., choice/non-choice) of the user database regarding the remaining items. In other words, one complete permutation of CF sequences through the total dataset was performed, with the predictive values $p_{a,j}$ as well as thresholds $c^{l}_{1,j}$ and $c^{u}_{0,j}$ being calculated using data from the training sample and the item recommendation step being applied for the hold-out data (based on the set of predictions for the training data).

For the evaluation of the predictive performance, we calculate the following hit rate h_j for each product category: h_j is defined as the fraction of the number of correctly predicted or recommended item choices (i.e., the number of occasions item j was recommended and actually purchased) and the total number of recommendations made for item j. In the information retrieval literature, this widely used metric is also denoted as the *precision* of an algorithm (see, e.g., Kowalski, 1997). Compared to competing definitions of hit rates like, for example, the ratio of correctly recommended to all actual—i.e., predicted and unpredicted—item choices (denoted as the *recall* rate), this measure takes into account the problem with so-called *false negatives*. False negatives are recommended items that do not meet the interests or preferences of users and might involve the risk of deterring customers from further using the recommendation engine (see also Sarwar et al., 2000).

The results obtained from the CF_{mod} procedure are compared to the recommendations derived from alternative methods which are subject to the same permutations through both the training and hold-out dataset. We benchmark the CF_{mod} results against recommendations based on (a) a sequence of binary logit models (*BLM*) similar to the modeling approach as proposed by Hruschka (1991),[2] (b) recommendations based on simple a-priori item choice probabilities (*APPROB*) as well as (c) recommendations based on raw predictive values (CF_{raw}) as derived by the basic CF approach.

[2] We do not model first-order interaction effects and skip the model selection sub-routine due to its computationally prohibitive effort in practical applications as well as the relatively small number of available transaction data in the present data set.

When applying *APPROB*, the predictive values $p_{a,j}$ are set to item j's observed choice probability $\bar{c}_j$ in the training sample $T \subset I$, and a given number of items with descending purchase probabilities are recommended to all users in the hold-out sample. Since the item-specific choice probabilities are identical to the average number of shopping baskets containing the item across all baskets available ($\bar{c}_j = \Sigma^{T}_{i=1} c_{i,j}/T$), they represent expected values of observing an item's j choice and thus provide a natural performance benchmark for individualized recommendation procedures like CF methods. Notice that the predictive values delivered by *APPROB* are unique for all users and, consequently, do not account for heterogeneity in customer preferences. Thus, predictions based on *APPROB* may be expected as a lower performance benchmark which should be outperformed by CF and comparable methods like the *BLM* approach.

The raw predictive values used by procedure CF_{raw} for further recommendation are calculated for all users in the training sample using Eq (1). Interpreting these values as conditional probabilities of observing an item j's choice for the active user a (with implicit similarity weights $w(a,i)$ regarding other users), we use a threshold decision rule of $\bar{p}_j = 0.5$ for recommending an item in the hold-out sample. In order to allow for derivation of CF_{mod} based recommendations, upper and lower confidence bounds have to be computed for the set of raw CF predicted values $p_{a,j}$ as derived from the training sample according to expressions (3) and (4) for each item. Next, decision rule (5) is applied for the construction of a candidate list for item recommendation. For the available training dataset, 29 out of 54 product categories were included in that subset of items. For the remaining items the predictors failed to discriminate clearly between chosen and non-chosen items.

When applying *APPROB* and the *BLM* procedure, for the hold-out sample the number of recommended items per customer is set equal to the average basket size (8 items). Contrarily, the length of the recommendation list is flexible for both the CF_{mod} and the CF_{raw} based methods. Here, all items with $p_{a,j}$ exceeding the respective threshold are recommended to customer a.

5. Results and discussion

As a result of the estimation and item recommendation procedures as described above, Table 2 shows the 29 item-specific hold-out (test sample) results for the competing methods investigated.

As expected, *APPROB* recommendations are clearly outperformed by the CF_{mod} approach across all product categories. A comparison of average hit rates over all items (see row 'average precision' in Table 2), reflects a convincing improvement in terms of overall hold-out

Table 2
Test sample hit rates (precision) for different product categories and methods used

Product category	CF_{mod}	*BLM*	*APPROB*	CF_{raw}	Frequency
Sausage	0.472	0.215	0.420	0.000	521
Poultry	0.174	0.125	0.000	0.000	166
Pork	0.167	0.308	0.000	0.000	173
Beef	0.249	0.210	0.000	0.000	262
Fruit	0.552	0.287	0.529	0.530	657
Vegetables	0.693	0.327	0.600	0.600	744
Dairy products	0.840	0.433	0.781	0.781	969
Non-perishable products	0.100	0.263	0.000	0.000	131
Cheese	0.350	0.361	0.315	0.000	391
Frozen foods	0.276	0.266	0.259	0.000	321
Bread	0.666	0.331	0.609	0.609	756
Staple food	0.209	0.305	0.000	0.000	186
Vinegar/oil	0.251	0.303	0.000	0.000	272
Sweetener	0.160	0.186	0.000	0.000	117
Herbs	0.054	0.118	0.000	0.000	71
Soup	0.092	0.000	0.000	0.000	85
Diet products	0.018	0.143	0.000	0.000	24
Baking products	0.066	0.114	0.000	0.000	57
Pet food	0.112	0.105	0.000	0.000	118
Non-alcoholic drinks	0.594	0.285	0.524	0.523	650
Beer	0.249	0.419	0.000	0.000	222
Spirituous beverages	0.101	0.111	0.000	0.000	57
Wine	0.115	0.100	0.000	0.000	81
Chips	0.120	0.075	0.000	0.000	135
Long life bakery products	0.291	0.275	0.000	0.000	261
Chocolate	0.303	0.322	0.000	0.000	279
Candies	0.086	0.083	0.000	0.000	81
Personal hygiene products	0.000	0.000	0.000	0.000	3
Books/music	0.208	0.039	0.000	0.000	223
Average precision	0.261	0.211	0.139	0.105	
Frequency correlation	0.994	0.673	0.955	0.895	

precision at the favor of Cf_{mod}. Naturally, application *of APPROB* renders recommendations and, as a consequence, noteworthy hit rates for the most frequently purchased items only. Notice that even for these categories CF_{mod} (but not *BLM*) delivers superior results in terms of precision of recommendation. The same group of items is obviously also favored by CF_{raw} recommendations.

From a marketing managerial perspective, however, the benefit of recommendation lists restricted to high frequency product categories (such as dairy products, bread, vegetables, non-alcoholic drinks, etc.) that are included in the majority of standard grocery shopping baskets remains more than questionable. Moreover, items of these fast-moving consumer good categories are almost permanently subject to aggressive price promotions (loss-leader pricing) and, therefore, earn low or even negative profit margins which do not justify the effort of personalized recommendation. In particular, this is a striking argument against the implementation efforts caused by the CF_{raw} procedure, since average hit rates are even lower (according to Table 2, 0.105 against 0.139) and the precision of individual item recommendations is not improved compared to an a priori offer of constant recommendations as provided by *APPROB*.

As the results suggest, however, the poor performance of the basic CF procedure for binary market basket data can be noticeably improved when adopting the proposed modified CF approach. There is another interesting point to be noticed when comparing the various results: Just as *APPROB* and CF_{raw}, also the CF_{mod} procedure results in a high forecasting precision for frequently purchased product categories. In fact, item-specific hit rates of CF_{mod} recommendations are almost perfectly correlated with the (holdout sample) purchase frequency distribution across items (see the column of item-specific purchase frequencies and the correlation coefficients in the last row of Table 2).[3] This high correlation is only partially due to the fact that good predictions of choices in categories with high purchase frequencies are possible (which in fact is the case for the correlations achieved by *APPROB* and CF_{raw} predictions). It is rather how the

[3] The correlations are only computed for the 29 categories which turned out to be predictable by the CF_{mod} selection procedure. Of course, correlations would be proportionally lower if all 54 categories were considered.

weighting procedure for shopping baskets obtained from other users in CF_{mod} is customizing a set of item recommendations to the individual user and as an outcome is mimicking the choice frequency distribution across categories at the aggregate level. Quite obviously, a conventional CF process as represented by the CF_{raw} results fails to do so.

Category-specific comparisons between CF_{mod} and *BLM* recommendation results yield no straightforward conclusion regarding excellence in terms of predictive accuracy. While CF_{mod} is predominant in 16 categories, *BLM*-based recommendations are superior in 13 cases. For a final overall comparison of the precision of recommendations derived from the methods applied in this study, we calculate the average differences of category-specific hit rates achieved by each of the competing methods ($h(CF_{mod})_j$, $h(BLM)_j$, $h(APPROB)_j$, and $h(CF_{raw})_j$) for each pair of methods. Before doing so, in order to account for differences in the magnitude of category purchase frequencies, the hit rates were weighted by the relative purchase frequencies $f(j)$. Hence, the overall weighted difference in precision between, for example, methods CF_{mod} and *BLM* is derived as $\Sigma_{j=1}^{J} f(j)(h(CF_{mod})_j - h(BLM)_j)/J$. The resulting pairwise differences are represented in the rows of Table 3, with an asterix indicating significance at the 95% confidence level.

For each row in Table 3, positive differences are indicative for superior overall performance when comparing one method against the others. It can be seen that CF_{mod} significantly outperforms all other methods under investigation in terms of weighted hit rates. Again, CF_{raw} is not able to excel the much more easy to derive *APPROB* recommendations. Interestingly, *BLM* is significantly outperformed by all other methods due to its poor performance in the most frequently purchased categories (in the unweighted case, CF_{mod} and *BLM* deliver similar results outperforming the remaining methods). This effect is also signified by the lower correlation coefficient for *BLM* recommendations with purchase frequencies as given in Table 2.

However, this finding should be taken with caveat. The dataset is relatively sparse and contains rather few retail transactions as compared to the number of independent variables in a regression-type analysis. In this regard, Mild and Natter (2002) presented a study on the influence of the available data on the performance of CF-based recommender systems as compared to regression-based methods. They report a positive impact of a higher number of available transactions on all methods, while regression-based methods are expected to gain most in terms of predictive accuracy. Although the authors are dealing with (pseudo-metric) rating data, their findings are very likely to apply to the series of binary logit models as employed by the *BLM* approach used in the present study.

Table 3
Pairwise differences in overall weighted hit rates for the competing recommendation methods investigated

Difference	CF_{mod}	*BLM*	*APROB*	CF_{raw}
CF_{mod}	—	0.171*	0.108*	0.161*
BLM	−0.171*	—	−0.063*	−0.010*
APROB	−0.108*	0.063*	—	0.053*
CF_{raw}	−0.161*	0.010*	−0.053*	—

*Indicates significance at the 5% error level.

6. Summary and conclusion

The advent of computer-mediated shopping environments makes improvements of the accuracy of personalized recommendation systems for predicting multiple-product category or information item choices potentially beneficial to customer satisfaction and loyalty. In the relevant marketing literature briefly reviewed in this paper, methodologies to study customers' cross-category preferences are summarized under the term 'market basket analysis'.

The majority of practical applications of recommender systems rely on the use of collaborative filtering methods, most of them originally designed for (pseudo-metric) rating data. The present paper investigates the suitability and limitations of these methods for situations when only binary customer information (i.e., choice/non-choice of items, such as shopping basket data) is available. Based on these considerations an extension of collaborative filtering algorithms suitable for pick-any/J choice data is presented. The proposed modification includes an automatizable and managerially adjustable (by means of α-values for confidence intervals) criterion-based selection procedure of item candidates for inferring recommendations. This new approach is demonstrated using a real-world retail transaction dataset consisting of customers' multiple category purchases across 54 product categories.

We have benchmarked the prediction of the proposed procedure against three alternatives, namely (a) a sequence of binary logit models, (b) simple a priori choice probability-based recommendations as well as (c) recommendations based on the predictive values resulting from a more conventional collaborative filtering approach. In terms of a hit rate criterion measuring the precision of item recommendations, the modified algorithm presented clearly outperforms both (b) and (c) across all product categories. Using a weighted difference between the hit rates accounting for purchase frequency variations across categories, the new algorithm also significantly excels (a).

Although the proposed method exhibits promising predictive accuracy, it would be interesting to see further

research investigating the influence of factors like the sparsity of the dataset, the amount of available transactions and the ratio between available transactions and the number of items on the relative performance.

Acknowledgements

We would like to express our gratitude to the guest editors for this special issue and two anonymous reviewers for providing valuable comments on a prior version of this paper.

References

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. Proceedings of the 20th VLDB Conference, Santiago, Chile.

Ainslie, A., Rossi, P.E., 1998. Similarities in choice behavior across product categories. Marketing Science 17 (2), 91–106.

Alba, J., Lynch, J., Weitz, B., Janiszewski, Ch., Lutz, R., Sawyer, A., Wood, S., 1997. Interactive home shopping: consumer, retailer, and manufacturer incentives to participate in electronic marketplaces. Journal of Marketing 61 (July), 38–53.

Anderberg, M.R., 1973. Cluster Analysis for Applications. Academic Press, New York.

Andrews, R.L., Currim, I.S., 2002. Identifying segments with identical choice behaviors across product categories: an intercategory logit mixture model. International Journal of Research in Marketing 19, 65–79.

Ansari, A., Essegaier, S., Kohli, R., 2000. Internet recommendation systems. Journal of Marketing Research 37 (August), 363–375.

Bakos, J.Y., 1997. Reducing buyer search costs: implications for electronic marketplaces. Management Science 43 (12), 1676–1708.

Boecker, F., 1978. Die Bestimmung der Kaufverbundenheit von Produkten. Duncker & Humblot, Berlin.

Breese, J.S., Heckerman, D., Kadie, C., 1998. Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, Madison, WI.

Buechter, O., Wirth, R., 1998. Discovery of association rules over ordinal data: a new and faster algorithm and its application to basket analysis. In: Wu, X., Kotagiri, R., Korb, K.B. (Eds.), Research and Development in Knowledge Discovery and Data Mining. Second Pacific-Asia Conference, PAKDD-98, Melbourne, Australia. Springer, Berlin, pp. 36–47.

Chintagunta, P.K., Haldar, S., 1998. Investigating purchase timing behavior in two related product categories. Journal of Marketing Research 35 (February), 43–53.

Coombs, C.H., 1964. A Theory of Data. Wiley, New York.

DeSarbo, W.S., Manrai, A.K., Manrai, L.A., 1993. Non-spatial tree models for the assessment of competitive market structure: an integrated review of the marketing and psychometric literature. In: Eliashberg, J., Lilien, G.L. (Eds.), Handbooks in Operations Research and Management Science, Vol. 5. Marketing. North-Holland, Amsterdam, pp. 193–257.

DeSarbo, W.S., Manrai, A.K., Manrai, L.A., 1994. Latent class multidimensional scaling: a review of recent developments in the marketing and psychometric literature. In: Bagozzi, R.P. (Ed.), Advanced Methods of Marketing Research. Blackwell, Cambridge, MA, pp. 190–222.

Dickinson, R., Harris, F., Sircar, S., 1992. Merchandise compatibility: an exploratory study of its measurement and effect on department store performance. International Review of Retail, Distribution and Consumer Research 2 (4), 351–379.

Haeubl, G., Trifts, V., 2000. Consumer decision making in online shopping environments: the effects of interactive decision aids. Marketing Science 1 (19), 4–21.

Harlam, B.A., Lodish, L.M., 1995. Modeling consumers' choices of multiple items. Journal of Marketing Research 32 (November), 404–418.

Hilderman, R.J., Carter, C.L., Hamilton, H.J., Cercone, N., 1998. Mining market basket data using share measures and characterized item-sets. In: Wu, X., Kotagiri, R., Korb, K.B. (Eds.), Research and Development in Knowledge Discovery and Data Mining. Second Pacific-Asia Conference, PAKDD-98, Melbourne, Australia. Springer, Berlin, pp. 159–173.

Hruschka, H., 1985. Der Zusammenhang zwischen paarweisen Verbundbeziehungen und Kaufakt- bzw. Kaeuferstrukturmerkmalen. Zeitschrift fuer betriebswirtschaftliche Forschung 37 (3), 218–231.

Hruschka, H., 1991. Bestimmung der Kaufverbundenheit mit Hilfe eines probabilistischen Messmodells. Zeitschrift fuer betriebswirtschaftliche Forschung 43 (5), 418–434.

Hruschka, H., Lukanowicz, M., Buchta, Ch., 1999. Cross-category sales promotion effects. Journal of Retailing and Consumer Services 6 (2), 99–105.

Julander, C.-R, 1992. Basket analysis. A new way of analyzing scanner data. International Journal of Retail and Distribution Management 20 (7), 10–18.

Karypis, G., 2000. Evaluation of item-based Top-N recommendation algorithms. Technical Report CS-TR-00-46, Computer Science Department, University of Minnesota.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, New York.

Kim, B.-D., Srinivasan, K., Wilcox, R.T., 1999. Identifying price sensitive consumers: the relative merits of demographic vs. purchase information. Journal of Retailing 75 (2), 173–193.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., Riedl, J., 1997. GroupLens: applying collaborative filtering to usenet news. Communications of the Association for Computing Machinery 40 (3), 77–87.

Kowalski, G., 1997. Information Retrieval Systems: theory and implementation. Kluwer Academic Publishers, Norwell, MA.

Maes, P., 1994. Agents that reduce work and information overload. Communications of the Association for Computing Machinery 37 (7), 30–40.

Manchanda, P., Ansari, A., Gupta, S., 1999. The shopping basket: a model for multi-category purchase incidence decisions. Marketing Science 18 (2), 95–114.

Mild, A., Natter, M., 2002. Collaborative filtering or regression models for internet recommendation systems. Journal of Targeting, Measurement and Analysis for Marketing 10 (4), 304–313.

Mild, A., Reutterer, T., 2001. Collaborative filtering methods for binary market basket data analysis. In: Liu, J., Yuen, P.C., Li, C., Ng, J., Ishida, T. (Eds.), Active Media Technology, Lecture Notes in Computer Science, Vol. 2252. Springer, Berlin, pp. 302–313.

Mulhern, F.J., Leone, R.P., 1991. Implicit price bundling of retail products: a multi-product approach to maximizing store profitability. Journal of Marketing 55 (October), 63–76.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J., 1994. Grouplens: an open architecture for collaborative filtering of netnews. Proceedings of the Association for Computing Machinery 1994 Conference on Computer Supported Cooperative Work. Anaheim, CA, pp. 219-231.

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

Russell, G.J., Bell, D., Bodapati, A., Brown, C., Chiang, J., Gaeth, G., Gupta, S., Manchanda, P., 1997. Perspectives on multiple category choice. Marketing Letters 8 (3), 297–305.

Russell, G.J., Kamakura, W.A., 1997. Modeling multiple category brand preference with household basket data. Journal of Retailing 73 (4), 439–461.

Russell, G.J., Petersen, A., 2000. Analysis of cross-category dependence in market basket selection. Journal of Retailing 76 (3), 367–392.

Russell, G.J., Ratneshwar, S., Shocker, A.D., Bell, D., Bodapati, A., Degeratu, A., Hildebrandt, L., Kim, N., Ramaswami, S., Shankar, V.H., 1999. Multiple-category decision-making: review and synthesis. Marketing Letters 10 (3), 319–332.

Salton, G., Buckley, Ch., 1988. Term-weighting approaches in automatic text retrieval. Information Processing and Management 24 (5), 513–523.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J., 2000. Analysis of recommendation algorithms for E-commerce. Proceedings of the ECOO, Minneapolis, Minnesota.

Schnedlitz, P., Reutterer, T., Joos, W., 2001. Data-Mining und Sortimentsverbundanalyse im Einzelhandel. In: Hippner, H., Kuesters, U., Meyer, M., Wilde, K. (Eds.), Handbuch Data Mining im Marketing. Knowledge Discovery in Marketing Databases, pp. 951-970.

Seetharaman, P.B., Ainslie, A., Chintagunta, P.K., 1999. Investigating household state dependence effects across categories. Journal of Marketing Research 36 (November), 488–500.

Ungar, H., Madison, W.I., Foster, D.P., 1998. Clustering methods for collaborative filtering. Proceedings of Workshop on Recommendation Systems at the 15th National Conference on Artificial Intelligence.

Walters, R.G., 1991. Assessing the impact of retail promotions on product substitution, complementary purchase, and inter-store sales displacement. Journal of Marketing 55 (April), 17–28.

West, P.M., Ariely, D., Bellman, S., Bradlow, E., Huber, J., Johnson, E., Kahn, B., Little, J., Schkade, D., 1999. Agents to the rescue? Marketing Letters 10 (3), 285–300.

Winer, R.S., Deighton, J., Gupta, S., Johnson, E.J., Mellers, B., Morwith, V.G., O'Guinn, T., Rangaswamy, A., Sawyer, A.G., 1997. Choice in computer-mediated environments. Marketing Letters 8 (3), 287–296.