Repeat-Buying Theory and Its Application for Recommender Services

W. Böhm[1], A. Geyer-Schulz[2], M. Hahsler[3], and M. Jahn[3]

[1]Mathematische Methoden der Statistik, WU-Wien, A-1090 Wien, Austria

[2]Informationsdienste und Elektronische Märkte, Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

[3]Informationswirtschaft, WU-Wien, A-1090 Wien, Austria

Abstract: In the context of a virtual university's information broker we study the consumption patterns for information goods and we investigate if Ehrenberg's repeat-buying theory which successfully models regularities in a large number of consumer product markets can be applied in electronic markets for information goods too. First results indicate that Ehrenberg's repeat-buying theory succeeds in describing the consumption patterns of bundles of complementary information goods reasonably well and that this can be exploited for automatically generating anonymous recommendation services based on such information bundles. An experimental anonymous recommender service has been implemented and is currently evaluated in the Virtual University of the Vienna University of Economics and Business Administration at `http://vu.wu-wien.ac.at`.

1 Introduction

In this article we study anonymous recommender services based on consumption patterns for information goods as presented in figure 1 showing a list of recommended web-sites of courses which are recommended, because students usually use them together with M. Hahsler´s Introduction to C++. For a discussion of the design space for recommender services we refer the reader to Resnick et al. (1997).

In this setting we consider an information broker with a clearly defined system boundary. Clicking on an external link is equated as "purchasing an information product". The rationale for this stems from an analysis of the transaction costs of a user of an information broker. Even "free" information products burden the consumer with search, selection and evaluation costs. Therefore, in this article we derive recommendations from products which have been repeatedly used (= purchased) together in the same sessions (= buying occasions). The following advantages make such recommendations attractive for information brokers:

- Observed consumer purchase behavior is the most important information for predicting consumer behavior online and offline as claimed in Bellmann et al. (1999).
- Market basket analysis shows up to 70 percent cross-selling potential. See, e.g. Blischok (1995). Such recommendations facilitate "repeat-buying", as suggested in Bellmann et al. (1999).

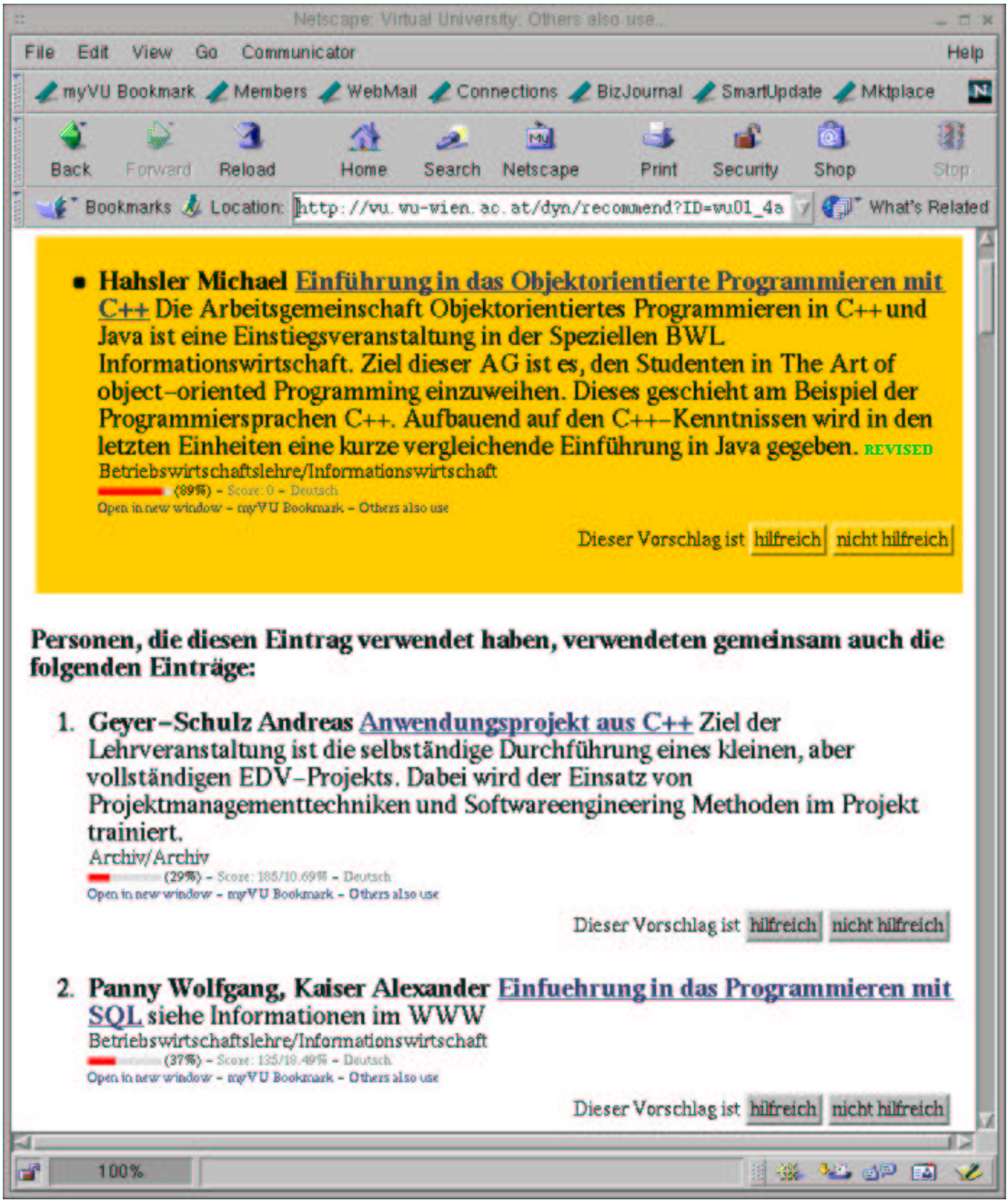

Figure 1: Example: An Anonymous Recommender Based on “Observed Purchase Behavior”

- Such recommendations are not subject to several incentive problems found in systems based on explicit recommendations as free-riding, bias, ... See Avery and Zeckhauser (1997). Faking such recommendations leads to high transaction costs, because only a single co-occurrence of products is counted per user-session. Free-riding is impossible, because each user automatically contributes usage data for the recommendations. The user's privacy is preserved.
- The transaction costs for the broker are low, because such recommendations can be generated without editor, author, or web-scout.

Anonymous recommendations based on consumption patterns have been made famous by Amazon.com and e.g. the first phase of the a-priori algorithm for association rules of Agrawal and Srikant (1994) can compute anonymous recommendations – even without a model and its underlying behavioral assumptions – quite well. The reason for this is that the first few entries of such frequency sorted lists of recommendations usually are good candidates for recommendations. Provided usage is counted in the same way and the thresholds for the support and confidence parameters of the a-priori algorithm are set to 0 the raw frequency-sorted recommendation lists produced by phase 1 of the a-priori algorithm will be the same as in this article. However, all such approaches still have the following two problems which we address in this article with Ehrenberg's repeat-buying theory (see Ehrenberg (1988)): Which co-occurences of products are regarded as non-random? And how many products should we recommend? Showing random co-occurences of products runs a high risk of giving bad recommendations (type I error), whereas suppressing non-random co-occurences of products implies that possible useful recommendations are not given (type II error).

Ehrenberg's repeat-buying theory is a descriptive theory based on consumer panel data well suited for this task, because of the strong stationarity and independence assumptions in the theory discussed in section 2, and because it has been supported by strong empirical evidence in consumer product markets since the late 1950's. Although quite sophisticated and general models of the theory (e.g. the Dirichlet model (Goodhardt et al. (1984)) exist, for our purposes the simplest model – the logarithmic series distribution (LSD) model – will be sufficient. For a survey see e.g. Wagner and Taudes (1987).

The careful and experienced reader certainly will ask at this point, how we can apply a theory for analyzing purchase histories from consumer panels to mere market baskets. A market basket contains all information products which a user has visited (= purchased) in a session (= purchase occasion), but not the identity of the consumer. In a consumer panel, in addition, the identity of each user is known. The purchase history of a consumer is just the sequence of the purchases in his market baskets. Well, the answer is simple: We consider anonymous market baskets as consumer panels with **unobserved consumer identity** – and as long as we work only at the aggregate level, everything works out fine. And keep in mind, that for repeat-buying analysis we count all occurences of an information product in a market basket just once.

2 Ehrenberg´s Repeat-Buying Theory for Bundles of Information Goods

Of the thousand and one variables which might affect buyer behavior, it is found that nine hundred and ninety-nine usually do not matter. Many aspects of buyer behavior can be predicted simply from the penetration and the average purchase frequency of an item, and even these two variables are interrelated. A.S.C. Ehrenberg (1988).

The key result of Ehrenberg's repeat-buying theory which we exploit in this paper for anonymous recommender services is that the logarithmic series distribution (LSD) describes the following frequency distribution of purchases (Ehrenberg (1988)), namely how many buyers buy a specific product 1, 2, 3, ... times (without taking into account the number of non-buyers)?

$$P(r \quad \text{purchases}) = \frac{-q^r}{r \ln(1-q)}, \quad r \geq 1 \tag{1}$$

$$\text{Mean purchase frequency} \quad w = \frac{-q}{(1-q)\ln(1-q)} \tag{2}$$

In purchasing a product a consumer basically makes two decisions: when does he buy a product of a certain product class (purchase incidence) and which brand does he buy (brand choice). Ehrenberg claims that almost all aspects of repeat-buying behavior can be adequately described by formalizing the purchase incidence process for a single brand and to integrate these results later. The logarithmic series distribution results from the following assumptions about the consumers' purchase incidence distributions:

1. The share of never-buyers in the population is not specified. In our setting this definitely holds.
2. The purchases of a consumer in successive periods follow a Poisson distribution with a certain long-run average μ.
   The purchases of a consumer follow a Poisson distribution in subsequent periods, if a purchase tends to be independent of previous purchases (as is often observed) and a purchase occurs in such an irregular manner that it can be regarded as if random (see Wagner and Taudes (1987)).
3. The distribution of μ in the population follows a truncated Γ-distribution, so that the frequency of any particular value of μ is $(ce^{-\mu/a}/\mu)d\mu$, for $\delta \leq \mu \leq \infty$, where δ is a very small number, a a parameter of the distribution, and c a constant, so that $\int_\delta^\infty (ce^{-\mu/a}/\mu)d\mu = 1$.
   A Γ-distribution of the μ_i in the population may result from the following independence conditions (see Ehrenberg (1988)): For different products $P, Q, R, S, ...$ the average purchase rate of P is independent of the purchase rates of the other products, and $\frac{P}{(P+Q+R+S+...)}$ is inpendent of a consumer´s total purchase rate of buying all the products. These independence conditions are likely to hold approximately in practice.

4. The market is in equilibrium (stationary).

For the sake of completeness (and to correct a persistent typesetting error in the original proof), we include the following short proof which is due to Chatfield (see Chatfield et al. (1966) and Ehrenberg (1988)):

1. The probability p_r of r purchases is Poisson distributed: $p_r = \frac{e^{-\mu}\mu^r}{r!}$
2. We integrate over all buyers in the truncated Γ-distribution: $p_r = c\int_\delta^\infty (\frac{e^{-\mu}\mu^r}{r!})\,(\frac{e^{-\mu/a}}{\mu})d\mu = \frac{c}{r!(1+\frac{1}{a})^r}\int_\delta^\infty e^{-(1+\frac{1}{a})\mu}\left((1+\frac{1}{a})\mu\right)^{r-1} d\left((1+\frac{1}{a})\mu\right)$. Since δ is very small, for $r \geq 1$ this is approximately $\left(\frac{c}{r!(1+\frac{1}{a})^r}\right)\Gamma(r) = \frac{c}{(1+\frac{1}{a})^r r} = c\frac{q^r}{r} = qp_{r-1}(r-1)/r$, with $q = \frac{a}{1+a}$.
3. If $\sum p_r = 1$ for $r \geq 1$, we get $p_1 = \frac{-q}{\ln(1-q)}$ and $p_r = \frac{-q^r}{r\ln(1-q)}$.

Next, consider for some fixed information product x in the set X of information products in the broker, the purchase frequency of pairs of (x, i) with $i \in X - \{x\}$. The reason for considering pairs of information products is that we expect complementarities between information products, because Internet users usually tend to use several information products for a task. Because of the independence assumptions outlined above, the frequency distribution that such pairs occur 1, 2, 3, .., -times, follows a logarithmic series distribution by the same line of reasoning as above. And we expect that non-random occurences of such pairs occur more often than predicted by the logarithmic series distribution. A *recommendation* in this setting simply implies that co-occurences occur more often than expected from independent random choice acts and that a recommendation reveals a complementarity between information products. We use the stochastic model as a benchmark for discovering regularities.
Finally, we present a short overview of the algorithm for computing recommendations:

1. *Compute for all information products x in the market baskets the frequency distributions for repeat-purchases of the co-occurences of x with other information products in a session, that is of the pair (x, i) with $i \in X - \{x\}$.*
2. *Discard all frequency distributions with less than l observations. (We set $l < 10$ in order to prune frequency distributions which are unlikely to lead to a significant LSD model. More than 80 % of these frequency distributions contain no repeat-buys in our data. For the rest, a χ^2-goodness-of-fit test should not be used.)*
3. *For each frequency distribution, we compute:*
   (a) *Compute the* **robust** *mean purchase frequency w by trimming e.g. the 2,5 percentil of the high repeat-buy pairs.*

(b) *Estimate the parameter q for the LSD-model from $w = \frac{-q}{(1-q)(\ln(1-q))}$ with either a bisection or Newton method.*
(c) *Apply a χ^2-goodness-of-fit test with a suitable α (e.g. 0.01 or 0.05) between the observed and the expected LSD distribution with a suitable partitioning.*
(d) *Determine the outliers in the tail. (We suggest to be quite conservative here: Outliers at r are above $\sum_r^\infty p_r$.)*
(e) *Finally, we prepare the list of recommendations for information product x, if we have a significant LSD-model with outliers.*

3 Data Set and Results

The data set used for the example shown in figures 2 and 3 is from the anonymous recommender services of the Virtual University of the Vienna University of Economics and Business Administration (http://vu.wu-wien.ac.at) for the observation period from 1999-09-01 to 2001-03-05. The agent architecture as well as the data collection techniques have been described in Geyer-Schulz et al. (2001a). Personalized recommendations based on self-selection combined with interactive evolutionary algorithms can be found in Geyer-Schulz et al. (2000). The potential of recommender systems in education and scientific research are discussed in Geyer-Schulz et al. (2001b). A revised and expanded version of this contribution was presented at the WEBKDD2001 conference and is under review (Geyer-Schulz et al. (2001c)). This (later) version presents the model in more detail. Special emphasis in Geyer-Schulz et al. (2001c) is on the identification of non-random outliers, on the discussion of detailed results, and on the validation of recommendations.

In figure 2 we show the first 17 recommendations generated for the research site `Intelligent Software Agents (CMU)` by the method described in the previous section. 101 other information products have been found in market baskets together with this research site. The (robust) mean purchase frequency is 1.556, the parameter q of the LSD-model is 0.562. A χ^2 goodness-of-fit test is highly significant ($\chi^2 = 10.763$ which is considerably below 30.144, the critical value at $\alpha = 0.05$).

We regard outliers whose observed repeat-purchase frequency is above the theoretically expected frequency as recommendations. In figure 3 we explore three options of determining the cut-off point for such outliers:

1. Without doubt, as long as the observed repeat-purchase frequency is above the cumulated theoretically expected frequency, we have detected outliers. In our example, this holds for all observations of more than 5 co-purchases which correspond to the top 5 sites shown as recommendations in figure 2. (This is the most conservative choice.)
2. Discounting any model errors, as long as the observed repeat-purchase frequency is above the theoretically expected frequency is a less conservative option. For the example, we select all co-purchases with more than 3 occurences as recommendations. See the top 11 sites in figure 2.

```
Intelligent Software Agents (CMU)

Persons using the above entry
used the following entries too:
  1.Intelligent Software Agents (Sverker Janson)
  2.agent (Definition and Links from webopedia) META
  3.Intelligent Software Agents on the Internet
    (Björn Hermans) META
  4.Mobasher - List of Publications Personalization
    and Adaptive Web Sites from Web-Usage Patterns.
  5.Intelligent Software Agents and New Media
=== Cut ============================================
  6.Geyer-Schulz Intelligente Internet Agenten
  7.Agent Technology Projects
    in the Stanford Digital Library
  8.vista's virtual friends
  9.Books on Software Agents
 10.AVALANCHE - Agent-Based Value Chain Experiment
 11.The Zeus Agent Building Toolkit
=== Cut ============================================
 12.Let's Browse: A Collaborative Web Browsing Agent
 13.German Agent Pages
 14.Mobile Service Agents
=== Cut ============================================
 15.Foundation for Intelligent Physical Agents
 16.Auctions and Bargaining in Electronic Commerce
 ...
```

Figure 2: List of entries with cuts

3. If we take the cut, where both cumulative purchase frequency distributions cross, we get 14 recommendations regarding all co-purchases occuring more than twice as nonrandom – the top 14 sites in figure 2.

However, we recommend that the most conservative approach should be implemented. This recommendation is based on a check of the face validity of the recommendations for a small sample of information products (25 products).
As summarized in table 1 we fitted a LSD-model for the frequency distributions of co-occurencies for 1300 information products. For 675 information products, that is more than 50 percent, the estimated LSD-models pass a χ^2 goodness-of-fit test at $\alpha = 0.01$. A small scale face validation experiment of inspecting every entry in 100 randomly selected recommendation lists for plausibility led to a quite satisfactory result: 87.71 % good recommendations (LSD significant, 31 lists), $89,45$ % good recommendations (LSD not signif-

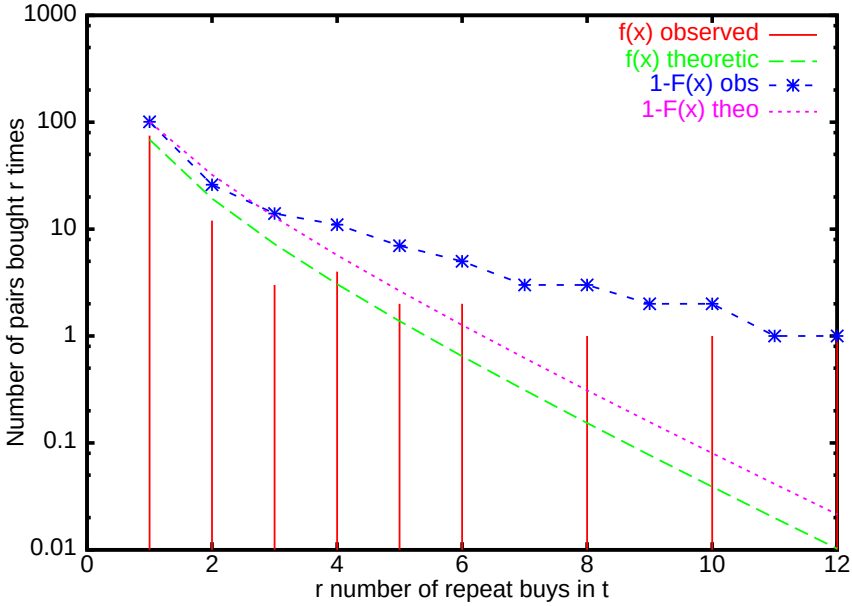

Figure 3: Plot of log distributions

icant, 42 lists). Only 75.74 % good recommendations were found for those LSD-models where no χ^2 test could be computed (25 lists), which is significantly lower.

Surprisingly, the class of models where the LSD model was not significant contains a slightly higher number of recommendations evaluated as good. A close inspection of frequency distributions for these lists revealed the quite unexpected fact that several of these frequency distributions were for information products which belong to the oldest in the data set and which account for many observations. The reasons for this may be explained by a shift in user behavior (non-stationarity) or too regular behavior (e.g., for cigarettes in consumer markets).

Also, the fact that the data set contains information products with different age may explain some of these difficulties. However, to settle this issue further investigations are required.

Finally, table 1 indicates that identification of non-random outliers is important for the perceived quality of recommender systems because of the high risk of recommending random co-occurences of products. The fact that recommendations for LSD-models for which no χ^2 test could be computed were considered as containing significantly more bad recommendations supports this conclusion.

	Number	%
Number of information products	9498	100.00
Number of products bought together with other products	7150	75.28
Not a uniform distribution and $n > 9$	4582	48.24
Enough repeat buys to compute LSD parameter and χ^2 test	1300	13.69
LSD with $\alpha = 0.01$ (robust)	675	7.11
LSD not significant	625	6.58
LSD fitted, no χ^2 test	703	7.40

Table 1: Summary of results. (Observation period: 1999-09-01 – 2001-05-07)

4 Further Research

The main contribution of this paper is that Ehrenberg's classical repeat-buying models describe – despite their strong independence and stationarity assumptions – the consumption patterns of information products surprisingly well. For anonymous recommender services they do a remarkable job of identifying non-random repeated-choice acts of consumers of information products which serve as the base of automatically generated recommendations of high-quality. However, the current version of the anonymous recommender services (and the analysis in this article) still suffers from two deficiencies. The first is that new information products are daily added to the information broker's data base, so that the stationarity assumptions for the market are violated and the information products in the data set are of non-homogenous age. The second drawback is that testing the behavioral assumptions of the model, as well as validation either by studying user acceptance or by controlled experiments still has to be done.

In addition, we expect Ehrenberg's repeat-buying models to be of considerable help to create anonymous recommender services for recognizing emerging shifts in consumer behavior patterns (fashion, emerging trends, moods, new subcultures, ...) and imbedded marketing research services which provide forecasts and classical consumer panel analysis in a cost efficient way.

5 Acknowledgement

The financial support of the Jubiläumsfonds of the Austrian National Bank under Grant No. 7925 is gratefully acknowledged.

References

AGRAWAL, R. and SRIKANT, R. (1994): Fast Algorithms for Mining Association Rules, Proceedings of the 20th VLDB Conference, Santiago, 487–499.

AVERY, C. and ZECKHAUSER, R. (1997): Recommender Systems for Evaluating Computer Messages. *Communications of the ACM, 40(3),* 88–89.

BELLMANN, S., LOHSE, G.L., and JOHNSON, E.J. (1999): Predictors of Online Buying Behavior. *Communications of the ACM, 42(12),* 32–38.

BLISCHOK, T.J. (1995): Every Transaction Tells a Story. *Chain Store Age Executive, 71(3),* 50–62.

CHATFIELD, C., EHRENBERG, A.S.C., and GOODHARDT, G.J. (1966): Progress on a Simplified Model of Stationary Purchasing Behavior. *Journal of the Royal Statistical Society A*, Vol. 129, 317-367.

EHRENBERG, A. S. C. (1988): *Repeat-Buying: Facts, Theory and Applications.* Charles Griffin & Company Limited, London.

GEYER-SCHULZ, A., HAHSLER, M., and JAHN, M. (2000): myVU: A Next Generation Recommender System Based on Observed Consumer Behavior and Interactive Evolutionary Algorithms. In: W. Gaul, O. Opitz, M. Schader (Eds.): *Data Analysis – Scientific Modeling and Practical Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, Vol. 18, Springer, Heidelberg, 447-457.

GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001a): Recommendations for Virtual Universities from Observed User Behavior. In: W. Gaul, Ritter, M. Schader (Eds.): Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg, to appear.

GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001b): Educational and Scientific Recommender Systems: Designing the Information Channels of the Virtual University. *International Journal of Engineering Education.* Vol. 17, N. 2, 153-163.

GEYER-SCHULZ, A., HAHSLER, M., JAHN, M. (2001c): A Customer Purchase Incidence Model Applied to Recommender Services. Submitted Procs. WEBKDD2001, LNCS, Springer, 20 pages.

GOODHARDT, G.J., EHRENBERG, A.S.C., COLLINS, M.A. (1984): The Dirichlet: A Comprehensive Model of Buying Behaviour. *Journal of the Royal Statistical Society*, A, 147, 621-655.

RESNICK, P. and VARIAN, H.R. (1997): Recommender Systems. *Communications of the ACM, 40(3),* 56–58.

WAGNER, U. and TAUDES, A. (1987): Stochastic Models of Consumer Behaviour. *European Journal of Operations Research, 29(1),* 1–23.