A Customer Purchase Incidence Model Applied to Recommender Services *

Andreas Geyer-Schulz[1], Michael Hahsler[2], and Maximillian Jahn[2]

[1] Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany,
andreas.geyer-schulz@em.uni-karlsruhe.de,
[2] Wirtschaftsuniversität Wien, A-1090 Wien, Austria,
Michael.Hahsler@wu-wien.ac.at

Abstract. In this contribution we transfer a customer purchase incidence model for consumer products which is based on Ehrenberg's repeat-buying theory to Web-based information products. Ehrenberg's repeat-buying theory successfully describes regularities on a large number of consumer product markets. We show that these regularities exist in electronic markets for information goods, too, and that purchase incidence models provide a well founded theoretical base for recommender and alert services.
The article consists of two parts. In the first part Ehrenberg's repeat-buying theory and its assumptions are reviewed and adapted for web-based information markets. Second, we present the empirical validation of the model based on data collected from the information market of the Virtual University of the Vienna University of Economics and Business Administration from September 1999 to May 2001.

1 Introduction

In this article we concentrate on an anonymous recommender service of the correlation-type made famous by Amazon.com applied to an information broker. It is based on consumption patterns for information goods (web-sites) from market baskets (web-browser sessions) which we treat as consumer purchase histories with unobserved consumer identity. In Resnick and Varian's design space [25] this recommender service is characterized as:

1. The contents of a recommendation consists of links to web-sites.
2. It is an implicit service based on observed user behavior.
3. The service is anonymous.
4. The aggregation of recommendations is based on identifying outliers with the help of a stochastic purchase incidence model.
5. A sorted list of recommended related web sites is offered to the user of a web site (see figure 1).

* In R. Kohavi, B.M. Masand, M. Spiliopoulou, and J. Srivastava, editors, WEBKDD 2001 - Mining Log Data Across All Customer Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001, Revised Papers, Lecture Notes in Computer Science LNAI 2356, pages 25-47. Springer-Verlag, July 2002.

Fig. 1. An Anonymous Recommender Based on "Observed Purchase Behavior".

This recommender service is part of the first educational and scientific recommender system integrated into the Virtual University of the Vienna University of Economics and Business Administration (http://vu.wu-wien.ac.at) since September 1999. A full description of all recommender services of this educational and scientific recommender system can be found in Geyer-Schulz et al. [16].

For example, figure 1 shows the recommended list of web-sites for users interested in web-sites related to the `Collaborative Filtering Workshop 1996 in Berkeley`. The web-site (in figure 1 the `Collaborative Filtering Workshop 1996 in Berkeley`) in the yellow box (gray in print) is the site for which related web-sites have been requested. For every recommended web-site a Vote Box allows users to evaluate the quality of this recommendation.

In Geyer-Schulz et al. [15] we have presented the architecture of an information market and its instrumentation for collecting data on consumer behavior. We consider an information broker with a clearly defined system boundary and web-sites consisting of one or more web-pages as information products. The information broker contains only meta-data on the information products (including an external link to the home page of each information product). Clicking on an external link (which leaves the information broker and leads to the home page of a web-site) is equated as "purchasing an information product". In marketing, we assume that a consumer will only repeatedly purchase a product or a product combination if he is sufficiently content with it. The rationale that this analogy holds even for *free* information products 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 used (= purchased) together in a session repeatedly across different sessions (= buying occasions) (see Böhm et al. [8]). Data collection (logging) is restricted entirely to the information broker, only clicking on an external link is logged. This implies that on the one hand in the logs of the information broker usage behavior for information products can not be observed (and is in fact completely irrelevant as far as repeat-buying theory is concerned) and that on the other hand almost no preprocessing is necessary to obtain sessions. Without such an instrumented architecture the purchase incidence model may still be of use in combination with web-usage mining approaches as e.g. the sequence miner WUM of Spiliopoulou [28]. For a recent survey of work on this area see Srivastava et al. [29].

Such recommendations are attractive for information brokers for the following reasons:

- Observed consumer purchase behavior is the most important information for predicting consumer behavior. For offline behavior this has been known for a long time (see Ehrenberg [11]), for online behavior see Bellmann et al. [6] for a recent study.
- In traditional retail chains, basket analysis shows up to 70 percent cross-selling potential (see Blischok [7]). Such recommendations facilitate "repeat-buying" which is one of the main goals of e-commerce sites as reported in Bellmann et al. [6].
- Most important in a university environment is that such recommendations are not subject to several incentive problems found in systems based on explicit recommendations (as e.g. free-riding, bias, ...) which are analyzed by Avery and Zeckhauser [5]. The transaction cost of faking such recommendations is high, because only one co-occurrence of products is counted per user-session as usual in consumer panel analysis (see Ehrenberg [11]). Free-riding is impossible because by using the information broker each user contributes usage data for the recommendations. The user's privacy is preserved.
- And, last but not least, the transaction costs for the broker are low since high-quality recommendations can be generated without human effort. No editor, no author, no web-scout is needed.

Recently the problem of generating personalized recommendations from anonymous sessions (or market basket data) and user sessions (or consumer panels with

individual purchase histories) has attracted considerable attention. Personalization is achieved by taking the user's current navigation history into account. Mobasher has studied two variants of computing recommendations from anonymous session data, namely PACT (profile aggregation based on clustering transactions) and ARHP (association rule hypergraph partitioning) in [23] and [24]. Lawrence et al. [21] report on a hybrid recommender based on clustering association rules from purchase histories combined with the product taxanomy and the profit contribution of a product. Gaul and Schmidt-Thieme [14] construct recommender systems from frequent substructures of navigation histories. Lin et al. [22] develop an association rule mining algorithm with adaptive support for a collaborative recommender system.

However, anonymous recommendations based on consumption or usage patterns nevertheless have the following two problems which we address in this article with the help of Ehrenberg's repeat-buying theory [11]:

- Which co-occurrences of products qualify as non-random?
- How many products should be recommended?

Ehrenberg's repeat-buying theory provides us with a reference model for testing for non-random outliers, because of the strong stationarity and independence assumptions in the theory discussed in section 2. What makes this theory a good candidate for describing the consumption behavior for information products is that it has been supported by strong empirical evidence in several hundred consumer product markets since the late 1950's. Ehrenberg's repeat-buying theory is a descriptive theory based on consumer panel data. It captures how consumers behave, but not why. Several very sophisticated and general models of the theory (e.g. the Dirichlet model by Goodhardt et al. [18]) exist and have a long tradition in marketing research. However, for our purposes, namely identifying non-random purchases of two information products, the simplest model – the logarithmic series distribution (LSD) model – will prove quite adequate. In this setting the LSD model removes all random co-purchases from the recommendation lists. It acts like a filter for noise. For a survey on stochastic consumer behavior models, see e.g. Wagner and Taudes [32].

One of the main (conceptual) innovations of this paper is that we explain how we can apply a theory for analyzing purchase histories from consumer panels to mere market baskets.

2 Ehrenberg´s Repeat-Buying Theory and Bundles of Information Products

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) [11].

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

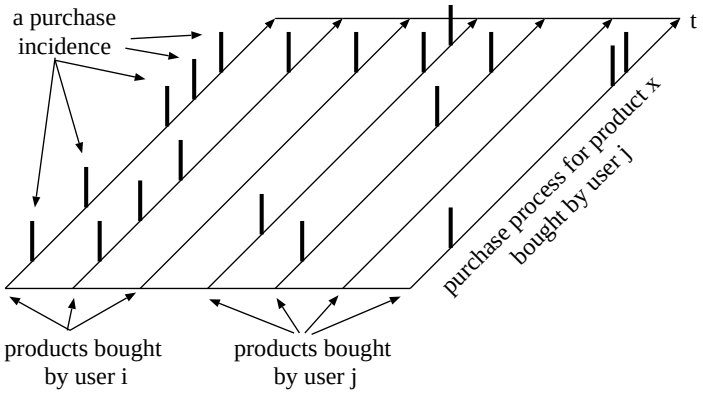

Fig. 2. Purchase Incidences as Independent Stochastic Processes.

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 by integrating these results later (see figure 2).

In a classical marketing context Ehrenberg's repeat-buying theory is based on purchase histories from consumer panels. The *purchase history* of a consumer is the sequence of the purchases in all his market baskets over an extensive periods of time (a year or more) for a specific outlet. For information products, the purchase history of a consumer corresponds e.g. to the sequence of sessions of a user in a personalized environment of a specific information broker. Note, however, a purchase history could be a sequence of sessions recorded in a cookie, in a browser cache, or in a personal persistent proxie-server, too.

A *market basket* is simply the list of items (quantity and price) bought in a specific trip to the store. In a consumer panel the identity of each user is known and an individual purchase history can be constructed from market baskets. For information products the corresponding concept is a session which contains records of all information products visited (used) by a user. In anonymous systems (e.g. most public web-sites) the identity of the user is not known. As a consequence no individual purchase history can be constructed.

Very early in the work with consumer panel data it turned out that the most useful unit of analysis is in terms of purchase occasions, not in terms of quantity or money paid. A *purchase occasion* is coded as yes, if a consumer has purchased one or more items of a product in a specific trip to a store. We ignore the number of items bought or package sizes and concentrate our attention on the frequency of purchase. For information products we define a purchase occasion as follows: a purchase occasion occurs if a consumer visits a specific information product at least once in a specific session. We ignore the number of pages browsed, repeat visits in a session, amount of time spent at a specific information product, ... Note, that this definition of counting purchases or in-

formation product usage is basic for this article and crucial for the repeat-buying theory to hold. One of the earliest uses of purchase occasions is due to L. J. Rothman [30].

Analysis is carried out in distinct time-periods (such as 1-week, 4-week, quarterly periods) which ties in nicely with other standard marketing reporting practices. A particular simplification from this time-period orientation is that most repeat-buying results for any given item can be expressed in terms of penetration and purchase frequency.

The *penetration* b is the proportion of people who buy an item at all in a given period. Penetration is easily measured in personalized recommender systems. In such systems it has the classical marketing interpretation. For this article, penetration is of less concern because in anonymous public Internet systems we simply cannot determine the proportion of users who use a specific web-site at all.

The *purchase frequency* w is the average number of times these buyers buy at least one item in the period. The mean purchase frequency w is itself the most basic measure of repeat-buying in the Ehrenberg's theory [11] and in this article.

In the following we consider anonymous market baskets as consumer panels with **unobserved consumer identity** – and as long as we work only at the aggregate level (with consumer groups) everything works out fine as provided that Ehrenberg's assumptions on consumer purchase behavior hold.

Figure 2 shows the main idea of purchase incidence models: a consumer buys a product according to a stationary Poisson process which is independent of the other buying processes. Aggregation of these buying processes over the population under the (quite general) assumption that the parameters μ of the Poisson distributions (the long-run average purchase rates) follow a truncated Γ-distribution results in a logarithmic series distribution (LSD) as Chatfield et al. [9] have shown.

The logarithmic series distribution (LSD) describes the following frequency distribution of purchases (see Ehrenberg [11]), namely the probability that a specific product is bought a total of 1, 2, 3, ..., r times without taking into account the number of non-buyers.

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

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

The variance is:

$$\sigma^2 = \frac{w}{(1-q)} - w^2 = \frac{-q\left(1 + \frac{q}{\ln(1-q)}\right)}{(1-q)^2 \ln(1-q)} \tag{3}$$

One important characteristic of the LSD is that $\sigma^2 > w$. For more details on the logarithmic series distribution, we refer the reader to Johnson and Kotz [19]. 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 of an Internet information broker with anonymous users 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 [32]).
3. The distribution of μ in the population follows a truncated Γ-distribution so that the frequency of any particular value of μ is given by $(ce^{-\mu/a}/\mu)d\mu$, 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 in the population may have the following reason (see Ehrenberg [11, p. 259]): If 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 independent of a consumer's total purchase rate of buying all the products, then it can be shown that the distribution of μ must be Γ. These independence conditions are likely to hold approximately in practice (see e.g. [4], [10], [26], [27]).
4. The market is in equilibrium (stationary). This implies that the theory does not hold for the introduction of new information products into the broker.

Next, we present Chatfield's proof in detail because the original proof is marred by a typesetting error:

1. The probability p_r that a buyer makes r purchases is Poisson distributed:

$$\frac{e^{-\mu}\mu^r}{r!}$$

2. We integrate over all buyers in the truncated Γ-distribution:

$$\begin{aligned}
p_r &= c\int_{\delta}^{\infty}(\frac{e^{-\mu}\mu^r}{r!})(\frac{e^{-\mu/a}}{\mu})d\mu \\
&= \frac{c}{r!}\int_{\delta}^{\infty} e^{-(\mu+\mu/a)}\mu^{r-1}d\mu \\
&= \frac{c}{r!}\int_{\delta}^{\infty} e^{-(1+1/a)\mu}\frac{(1+1/a)^{r-1}}{(1+1/a)^{r-1}}\mu^{r-1}d\mu \\
&= \frac{c}{r!(1+1/a)^{r-1}}\int_{\delta}^{\infty} e^{-(1+1/a)\mu}((1+1/a)\mu)^{r-1}d\mu \\
&= \frac{c}{r!(1+1/a)^{r-1}}\int_{\delta}^{\infty} e^{-(1+1/a)\mu}((1+1/a)\mu)^{r-1}\frac{1}{(1+1/a)}d(1+1/a)\mu \\
&= \frac{c}{r!(1+1/a)^{r}}\int_{\delta}^{\infty} e^{-(1+1/a)\mu}((1+1/a)\mu)^{r-1}d(1+1/a)\mu
\end{aligned}$$

Since δ is very small, for $r \geq 1$ and setting $t = (1+1/a)\mu$ this is approximately

$$
\begin{aligned}
p_r &= \left(\frac{c}{r!(1+\frac{1}{a})^r}\right)\int_\delta^\infty e^{-t}t^{r-1}dt \\
&\approx \left(\frac{c}{r!(1+\frac{1}{a})^r}\right)\Gamma(r) \\
&= \frac{c}{(1+\frac{1}{a})^r r} \\
&= c\frac{q^r}{r} \\
&= qp_{r-1}(r-1)/r
\end{aligned}
$$

with $q = \frac{a}{1+a}$.

3. If $\sum p_r = 1$ for $r \geq 1$, by analyzing the recursion we get $p_1 = \frac{-q}{\ln(1-q)}$ and $p_r = \frac{-q^r}{r\ln(1-q)}$. (However, this is the LSD. q.e.d.)

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 $i \in X \setminus x$. The probability $p_r(x \wedge i)$ that a buyer makes r purchases of products x and i together in the same session in the observation period which follow independent Poisson processes with means μ_x and μ_i is [20]: $p_r(x \wedge i) = \frac{e^{-\mu_x}\mu_x^r}{r!}\frac{e^{-\mu_i}\mu_i^r}{r!}$. For our recommender services for product x we need the conditional probability that product i has been used under the condition that product x has been used in the same session. Because of the independence assumption it is easy to see that the conditional probability $p_r(i \mid x)$ is again Poisson distributed by

$$
\begin{aligned}
p_r(i \mid x) &= \frac{p_r(x \wedge i)}{p_r(x)} \\
&= \frac{\frac{e^{-\mu_x}\mu_x^r}{r!}\frac{e^{-\mu_i}\mu_i^r}{r!}}{\frac{e^{-\mu_x}\mu_x^r}{r!}} \\
&= \frac{e^{-\mu_i}\mu_i^r}{r!} = p_r(i)
\end{aligned}
$$

This is not the end of the story. In our data, sessions do not contain the identity of the user – it is an unobserved variable. However, we can identify the purchase histories of sets of customers (market segments) in the following way: For each information product x the purchase history for this segment contains all sessions in which x has been bought. For each pair of information products x, i the purchase history for this segment contains all sessions in which x, i has been bought. The stochastic process for the segment (x, i) – n customers which have bought product x and an other product i – is represented by the sum of n independent random Bernoulli variables which equal 1 with probability p_i, and 0 with probablity $1 - p_i$. The distribution of the sum of these variables tends to a Poisson distribution. For a proof see Feller [12, p. 292]. (And to observe this aggregate process at the segment level is the best we can do.) If we assume that the parameters μ

of the segments' Poisson distributions follow a truncated Γ-distribution, we can repeat Chatfields proof and establish that the probability of r purchases of product pairs (x,i) follow a logarithmic series distribution (LSD).

However, we expect that non-random occurrences of such pairs occur more often than predicted by the logarithmic series distribution and that we can identify non-random occurrences of such pairs and use them as recommendations. For this purpose we estimate the logarithmic series distribution for the whole market (over all consumers) from market baskets, that is from anonymous web-sessions. We compute the mean purchase frequency w and solve equation 2 for q, the parameter of the LSD. By comparing the observed repeat-buying frequencies with the theoretically expected frequencies we identify outliers and use them as recommendations.

The advantage of this approach is that the estimation of the LSD is computationally efficient and robust. The limitation is that we cannot analyze the behavior of different types of consumers (e.g. light and heavy buyers) which would be possible with a full negative binomial distribution model (see Ehrenberg [11]).

What kind of behavior is captured by the LSD-model? Because of the independence assumptions the LSD-model estimates the probability that a product pair has been used at least once together in a session by chance r-times in the relevant time period. This can be justified by the following example: Consider that a user reads – as his time allows – some Internet newspaper and that he uses an Internet-based train schedule for his travel plans. Clearly, the use of both information products follows independent stochastic processes. And because of this, we would hesitate to recommend to other users who read the same Internet newspaper the train schedule. The frequency of observing this pair of information products in one session is as expected from the prediction of the LSD-model. Ehrenberg claims that this describes a large part of consumer behavior in daily life and he surveys the empirical evidence for this claim in [11].

Next, consider complementarities between information products: Internet users usually tend to need several information products for a task. E.g. to write a paper in a foreign language the author might repeatedly need an on-line dictionary as well as some help with LaTeX, his favorite type-setting software. In this case, however, we would not hesitate to recommend a LaTeX-online documentation to the user of the on-line dictionary. And the frequency of observing these two information products in the same session is (far) higher than predicted by the LSD-model.

A *recommendation* for an information product x simply is an outlier of the LSD-model – that is an information product y that has been used more often at least once together in the same session with product x in the observation period as could have been expected from independent random choice acts. A recommendation reveals a complementarity between information products.

The main purpose of the LSD-model in this setting is to separate non-random co-occurrences of information products (outliers) from random co-occurrences (as expected from the LSD-model). We use the LSD-model as a benchmark for discovering regularities.

Table 1 shows the algorithm we use for computing recommendations. In step 1 of the algorithm repeated usage of two information products in a single session is counted once as required in repeat-buying theory. In step 2 of the algorithm we discard all fre-

Table 1. Algorithm for computing recommendations.

1. Compute for all information products x in the market baskets the frequency distributions for repeat-purchases of the co-occurrences of x with other information products in a session, that is of the pair (x, i) with $i \in X \setminus x$. Several co-occurrences of a pair (x, i) in a single session are counted only once.
2. Discard all frequency distributions with less than l observations.
3. For each frequency distribution:
   (a) Compute the **robust** mean purchase frequency w by trimming the sample by removing x percent (e.g. 2.5%) 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.

quency distributions with a small number of observations, because no valid model can be estimated. This implies that in this case no recommendations are given. For each remaining frequency distribution, in step 3, the mean purchase frequency, the LSD parameter and the outliers are computed.

Note that high repeat-buy outliers may have a considerable impact on the mean purchase frequency and thus on the parameter of the distribution. By ignoring these high repeat-buy outliers by trimming the sample (step 3a) and thus computing a robust mean we considerably improve the chances of finding a significant LSD-model. This approach is justified by the data shown in column V of table 4 as discussed in section 4.1.

In step 3d outliers are identified by the property that they occur more often as predicted by the cumulated theoretically expected frequency of the LSD-model. Several less conservative options for determining the outliers in the tail of the distribution are discussed in the next section. These options lead to variants of the recommender service which exhibit different first and second type errors.

3 A Small Example: Java Code Engineering & Reverse Engineering

In figure 3 we show the first 16 candidates for recommendations of the list of 117 web-sites generated for the site `Java Code Engineering & Reverse Engineering` by the algorithm described in table 1 in the last section. Table 2 shows the statistics for this recommendation list. In the table as well as in the following we denote with $nf(x_{obs})$ the observed frequency distribution for r repeat-buys, by $f(x_{obs})$ the observed relative frequency distribution, by $f(x_{exp})$ the density function of the

```
Java Code Engineering & Reverse Engineering

Persons using the above web-site used the following web-sites too:

 1. Free Programming Source Code
 2. Softwareentwicklung: Java
 3. Developer.com
 4. Java-Einfuehrung
 5. The Java Tutorial
 6. JAR Files
 7. The Java Boutique
 8. Code Conventions for the Java(TM) Programming Language
 9. Working with XML: The Java(TM)/XML Tutorial
10. Java Home Page
11. Java Commerce
=== Cut ================================================
12. Collection of Java Applets
13. Experts Exchange
=== Cut ================================================
14. The GNU-Win32 Project
15. Microsoft Education: Tutorials
16. HotScripts.com
...
```

Fig. 3. List of web-sites with cuts.

estimated LSD-model, by $F(x_{obs})$ the cumulative relative frequency distribution and by $F(x_{exp})$ the distribution function of the estimated LSD-model. For a sample of n observations, the expected number of observations with r repeat buys is $nf(x_{exp}, r)$, the expected number of observations with at least r repeat buys is $nF(x_{exp}, j \geq r) = n \sum_{j=r}^{\infty} f(x_{exp}, j)$.

The observed mean purchase frequency in table 2 is 1.564. After trimming the highest 2.5 percentile (ignoring two observations with 7 and 8 repeat-buys, respectively), the robust mean purchase frequency is 1.461 and the estimated parameter q of the LSD-model is 0.511. Visual inspection of figures 4 and 5 shows that the estimated LSD-model properly describes the empirical data. This impression is supported by comparing the columns $nf(x_{obs})$ and $nf(x_{exp})$ in table 2 as well as looking at the χ^2-values in the second part of table 2. The χ^2 goodness-of-fit test for the trimmed data is highly significant with a χ^2-value of 1.099 which is considerably below 3.841, the critical value at $\alpha = 0.05$.

Table 2 also shows, that fitting a LSD-model to the original, untrimmed data results in a higher parameter q (0.567) and leads to a model with a higher χ^2-value (2.369) than the model obtained from the trimmed data. This indicates that ignoring high repeat-buy outliers improves the fit of the LSD-model. However, experimentation with several trimming percentiles in the range from 1.0 to 10 % indicated that ignoring 2.5 % of the observations contributed to an improved model fit, whereas ignoring additional obser-

vations did not lead to further improvement. In addition, in the evaluation the quality of recommendations proved rather insensitive to trimming. Our current experience suggests that 2.5 % of trimming is a robust choice for this parameter. However, additional experience with different data sets would be welcome.

All outliers whose observed repeat-purchase frequency is above the theoretically expected frequency are candidates to be selected as recommendations. In figure 5 (with a logarithmic y-axis) we explore three options of determining the cut-off point for such outliers:

Option 1. Without doubt, as long as the observed repeat-purchase frequency is above $F(x_{exp}, j \geq r)$, we have detected outliers. Look for $nf(x_{obs}, r) > n(1-F(x_{exp}, r))$ in table 2. In our example, this holds for all co-purchases with more than 3 repeat-buys which correspond to the top 11 sites shown as recommendations in figure 3. For $r = 3 \quad nf(x_{obs}, r) = 2$ is less than $n(1 - F(x_{exp}, r)) = 12.080$ in table 2, so we can not expect outliers in this class. This is the most conservative choice. Inspecting these recommendations shows that all of them are more or less directly related to Java programming, which is probably the task in which students use the example site.

Option 2. Discounting any model errors, as long as the observed repeat-purchase frequency $nf(x_{obs}, r)$ is above the theoretically expected frequency $nf(x_{exp}, r)$ is a less conservative option. For the example, we select all co-purchases with more than 3 occurrences as recommendations. Here this coincides with the option described above. See the top 11 sites in figure 3.

Option 3. If we take the cut, where both cumulative purchase frequency distributions cross, we get 13 recommendations regarding all co-purchases occurring more than twice as nonrandom – see the top 13 sites in figure 3. However, it seems, that web-sites 12 and 13, namely `Collection of Java Applets` and `Experts Exchange` seem to be not quite so related to Java programming.

The last three web-sites shown in figure 3 are not used as recommendations by any of the three explored options. And in fact they seem to be of little or no relevance for Java programming.

We have implemented the most conservative approach, namely option 1, in the recommender service based on a check of the face validity of the recommendations for a small sample of information products (25 products). We think that, at least in cases with a considerable number of candidates for the recommendation list, this is a suitable approach.

Instead of the 3 simple cutoff algorithms described above we consider the following error threshold procedure. We can determine for each class of products with r repeat-buys individually the probability of recommending a random web-site by dividing the theoretically expected number of occurrences by the observed number of occurrences ($f(x_{exp})/f(x_{obs})$ in table 2). If this quotient exceeds 1 the probability is 1. Consider, for example, the number of product combinations which have been bought 8 times together in table 2. Theoretically, we would expect that this is a chance event roughly in one out of ten cases (0.095). Now, we have observed 5 product combinations with 4 repeat-buys. Unfortunately, theoretically 2.789 product combinations can be expected

Table 2. Statistics for web-site Java Code Engineering & Reverse Engineering.

```
# Web-site: wu01_74 (Mon May  7 16:48:37 2001)
# Heuristic: Distr=NBD - Case 4: NBD heuristic var>mean
# Total number of observations: 117
# Sample mean=1.56410256410256 and var=1.64456233421751
# Estimate for q=0.566835385131836
#
# Robust estimation: Trimmed begin: 0 / end: 0.025 (2 observations)
# Robust estimation: Number of observations: 115
# Robust mean=1.46086956521739
# Robust estimate for q=0.511090921020508

# Plot:
# r repeat-buys  nf(x_obs)  nf(x_exp)  n(1-F(x_obs,r))  n(1-F(x_exp,r))
  1              87         83.565     117              117
  2              17         21.355      30               33.435
  3               2          7.276      13               12.080
  4               5          2.789      11                4.804
  5               3          1.140       6                2.015
  6               1          0.486       3                0.874
  7               1          0.213       2                0.389
  8               1          0.095       1                0.176

# Identifying non-random co-occurrences:
# Option 1: mixed intersection (f(x_obs) with 1-F(x_exp) at:
#     3 (leaves 11 nonrandom outliers)
# Option 2: f(x) intersection at: 3 (leaves 11 nonrandom outliers)
# Option 3: 1-F(x) intersection at: 2 (leaves 13 nonrandom outliers)

# Chi-square test (q=0.566835385131836; 117 observations):
# Class  obs     exp        chi-square
# 1      87      79.269     0.754
# 2      17      22.466     1.330
# 3      13      15.071     0.285
#                          -------
# Chi-square value:         2.369

# Chi-square test (robust) (q=0.511090921020508; 115 observations):
# Class  obs     exp        chi-square
# 1      87      82.137     0.288
# 2      17      20.990     0.758
# 3      11      11.794     0.053
#                          -------
# Chi-square value:         1.099

# Chi-square test threshold with alpha=0.01 w/1 d.f.: 10.828
# Chi-square test threshold with alpha=0.05 w/1 d.f.:  3.841
# -> LSD with alpha=0.05
```

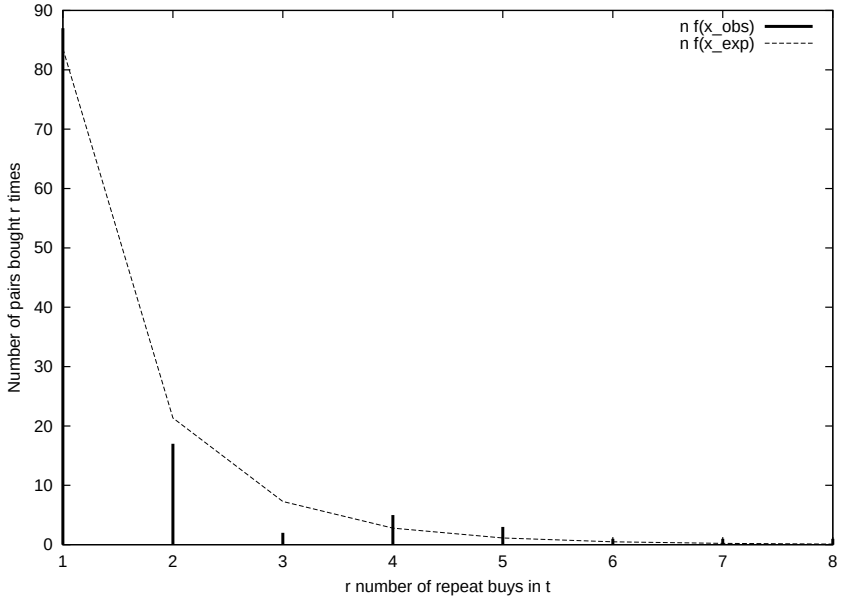

Fig. 4. Plot of frequency distribution.

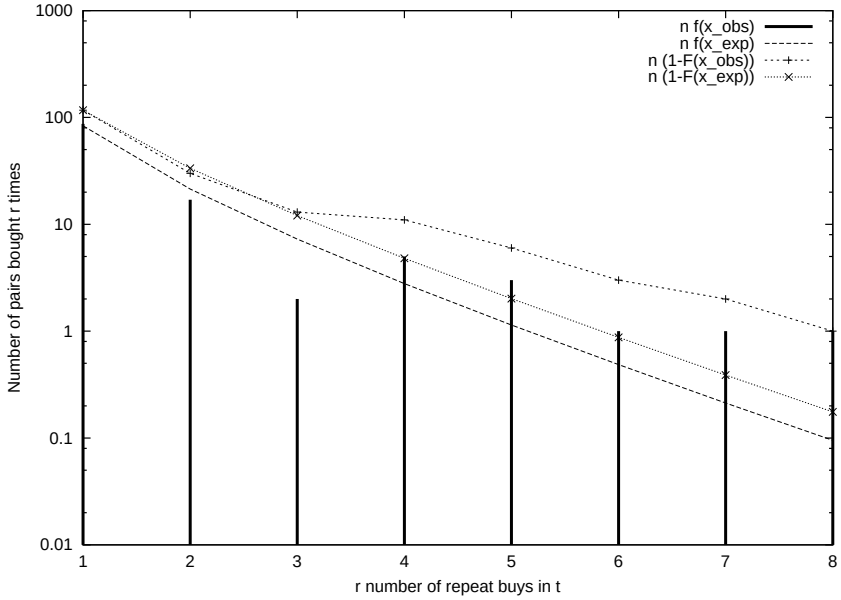

Fig. 5. Plot of log frequency distribution.

Table 3. Finding classes with less than 0.40 random observations.

r repeat-buys	$nf(x_{obs})$	$nf(x_{exp})$	$f(x_{exp})/f(x_{obs})$	Class shown
1	87	83.565	0.961	0
2	17	21.355	1.256	0
3	2	7.276	3.638	0
4	5	2.789	0.558	0
5	3	1.140	0.380	1
6	1	0.486	0.486	0
7	1	0.213	0.213	1
8	1	0.095	0.095	1

```
Java Code Engineering & Reverse Engineering

Persons using the above web-site used the following web-sites too:

 1. Free Programming Source Code
 2. Softwareentwicklung: Java
 4. Java-Einfuehrung
 5. The Java Tutorial
 6. JAR Files
```

Fig. 6. List of web-sites selected by class.

from pure chance. In this class we observe now a mixture of random product combinations and non-random product combinations, but we are not able to distinguish them. However, we can specify a threshold for the chance of falsely presenting a random co-occurrence, e.g. below 0.40. Table 3 summarizes this procedure. That is we pick only those classes with r observed repeat-buys where the probability of falsely presenting an outlier is below $\theta = 0.40$.

In the example, we would then present the web-sites in classes 5, 7, and 8, but not the web-site in class 6. That is, we would present web-sites 1, 2, 4, 5, and 6 as shown in figure 6. Web-site 3 (`Developer.com`, a site definitely not exclusively devoted to Java programming) and others from figure 3 are not shown because of the high probability of presenting random web-sites.

Considering the theoretically expected number of occurrences $f(x_{exp})$ of a class as the error to give a recommendation for a random co-occurrence (the type II error) leads to a thresholding strategy based on the specification of acceptable expected type II error β: We add to the recommender products from the tail of the distribution as long as the expected type II error of the recommender is smaller than β. If there is more than one product in a class r, the type II error of this class is divided by the number of products in the class and products in a class may be picked in arbitrary sequence. This strategy has been implemented and used for the evaluation of the recommender in section 4.3.

Table 4. Detailed results for observation period 1999-09-01 – 2001-05-07.

	I q undef.	**II** no χ^2 (< 3 classes)	**III** Sign. $\alpha = 0.05$	**IV** Sign. $\alpha = 0.01$	**V** Sign. (trim)	**VI** Not sign.	$\sum$
A Obs.< 10	1128 (0)	66 (63)	0 (0)	0 (0)	0 (0)	0 (0)	1194 (63)
B $\overline{x} = 1$	1374 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1374 (0)
C $\overline{x} > \sigma^2; r \leq 3$	2375 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	2375 (0)
D $\overline{x} > \sigma^2; r > 3$	201 (0)	617 (605)	105 (105)	46 (46)	15 (15)	372 (145)	1356 (916)
E $\sigma^2 > \overline{x}$	3 (0)	86 (86)	222 (222)	194 (194)	93 (93)	253 (253)	851 (848)
$\sum$	5081 (0)	769 (754)	327 (327)	240 (240)	108 (108)	625 (398)	7150 (1827)

(x) indicates x lists with at least 1 outlier.

In the analysis of outliers there is still room for improvement as e.g. by developing statistical tests for identifying outliers. For example, the choice of the threshold value should be analyzed in terms of the tradeoff between errors of type I and II for a sample of co-occurrence lists all of whose entries have been completely evaluated by experts with regard to their usefulness as recommendations.

4 First Empirical Results

To establish that a recommender service based on Ehrenberg's repeat buying-theory is supported by empirical evidence, we proceed as follows:

1. In section 4.1 we investigate how well the LSD-model explains the actual data for 7150 information products.
2. However, that the LSD-model fits the data well does not yet mean that the outliers we have identified are suitable recommendations for a user. In section 4.2 we present the results of a first small face evaluation experiment whose result suggests that these outliers are indeed valuable recommendations.

The data set used for the example given in section 3 and in this section 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. Co-occurrences have been observed for

8596 information products. After elimination of web-sites which ceased to exist in the observation period co-occurrences for 7150 information products remain available for analysis.

4.1 Fit of Data to LSD-Models

Table 4 summarizes the results of applying the algorithm for computing recommendations presented in table 1. If the sample variance is larger than the sample mean this may indicate that a negative binomial distribution (NBD)-model (and thus its LSD-approximation) is appropriate (see Johnson and Kotz [19, p.138]). This heuristic suggests 851 candidates for an LSD-model (see table 4, row E).

The rows of the table represent the following cases:

A The number of observations is less than 10. In this row we find co-occurrence lists either for very young or for very rarely used web-sites. These are not included into the further analysis. Cell (A/II) in table 4 contains lists which have repeated co-occurrences despite the low number of observations. In this cell good recommendation lists may be present (4 out of 5).

B No repeat-buys, just one co-occurrence. These are discarded from further analyses.

C Less than 4 repeat-buys and trimmed sample mean larger than variance. Trimming outliers may lead to the case that only the observations of class 1 (no repeat-buys) remain in the sample. These are discarded from further analyses.

D More than 3 repeat-buys and trimmed sample mean larger than variance. In cell (D/I) after trimming only class 1 web-sites remain in the trimmed sample (no repeat-buys). As a future improvement, the analysis should be repeated without trimming. In cell (D/II) the χ^2-test is not applicable, because less than 3 classes remain.

E (Trimmed) sample variance larger than sample mean. For cell (E/I) we recommend the same as for cell (D/I). For cell (E/II) we observe the same as for cell (D/II).

The columns I – VI of table 4 have the following meaning:

I The parameter q of the LSD model could not be estimated. For example, only a single co-occurrence has been observed for some product pairs.

II The χ^2 goodness-of-fit test could not be computed, because of lack of observations.

III, IV The χ^2 goodness-of-fit test is significant at $\alpha = 0.05$ or $\alpha = 0.01$, respectively.

V The χ^2 goodness-of-fit test is significant at $\alpha = 0.01$ using the trimmed data. All high repeat-buy pairs in the 2.5 percentile have been excluded from the model estimation.

VI The χ^2-test is not significant.

As summarized in table 5 we tested the fitted LSD-model for the frequency distributions of co-occurrences 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 $\alpha = 0.01$.

Table 5. Summary of results for observation period 1999-09-01 – 2001-05-07.

	n	%
Information products	9498	100.00
Products bought together with other products	7150	75.28
Parameter q defined	2069	21.78
Enough classes for 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
$n < 10$ and no χ^2-test	66	0.69

4.2 Face Validation of Recommendations

In order to establish the plausibility of the recommendations identified by the recommender service previously described we performed a small scale face validation experiment. The numbers in parenthesis in table 4 indicate the number of lists for which outliers were detected. From these lists 100 lists of recommendations were randomly selected. Each of the 1259 recommendations in these lists was inspected for plausibility. Plausible recommendations were counted as good recommendations by pressing the affirmative symbol (a hook) in the Vote Box shown in figure 1 in the introduction of this paper.

This small scale face validation experiment of inspecting recommendations for plausibility led to a quite satisfactory result:

- For the 31 lists for which a significant LSD-model could be fitted, 87.71 % of the recommendations were judged as good recommendations.
- 25 lists for which an LSD model was not significant contained $89,45$ % good recommendations.
- Only 75.74 % good recommendations were found in the 44 lists for those LSD-models where no χ^2 test could be computed which is a significantly lower percentage.

Surprisingly, the class of models where the LSD model was not significant contains a slightly higher number of recommendations evaluated as good. However, a number of (different) reasons may explain this:

- First, we might argue that even if the LSD-model is insignificant, it still serves its purpose, namely to identify non-random outliers as recommendations.
- 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 e.g. by a shift in user behavior (non-stationarity) or too regular behavior as e.g. for cigarettes in consumer markets (see Ehrenberg [11]). If too regular behavior is the reason that the LSD-model is insignificant, again, we still identified the non-random outliers.

– Another factor which might contribute to this problem is that several web-sites in this group belong to lists integrated in the web-sites of other organizational units. These lists, at least some of them, contain web-sites which have been carefully selected by the web-masters of these organizational units for their students. For example, the list of web-sites for student jobs is integrated within the main web-site of the university. The recommendations for such lists seem to reflect mainly the search behavior of the users. A similar effect is known in classic consumer panel analysis, if the points of sale of different purchases are not cleanly separated. This implies that e.g. purchases in a supermarket are not distinguished from purchases from a salesman. In our analysis, the purchase occasions are in different web-sites, namely the broker system and the organizational web-site with the embedded list. Ehrenberg's recommendation is to analyse the data separately for each purchase occasion.

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.

4.3 A First Comparison with a Simple Association Rule Based Recommender System

Compared to several other recently published recommender systems (e.g. [21], [24], [23]) which combine several data mining techniques, most noteably association-rule mining algorithms and various clustering techniques, the recommender system discussed here is very simple. Computing the frequency distribution corresponds exactly to the identification of frequent itemsets in association-rule mining algorithms whose most famous representative is Agrawal's a-priori algorithm ([2], [3]) with support and confidence of 0. Recent improvements of these algorithms include TITANIC of Stumme et al. [31] for highly correlated data sets, the association rule mining algorithm with adaptive support of Lin et al. [22], a graph-based approach for association rule mining by Yen and Chen [33], and a new approach for the online generation of association rules by Aggrawal and Yu [1]. TITANIC efficiently exploits the concept lattice and is based on concept analysis (see Ganter and Wille [13]).

The difference between our system and the association rule framework is in the way, "significant" item pairs are identified. In our model "significant" item pairs are outliers with regard to the LSD model, in association rule approaches "significant pairs" have more than a specified amount of support and more than a specified amount of confidence.

We have evaluated these two simple recommender systems on the VU data set for the observation period of 2001-01-01 to 2001-06-30. For the purpose of comparing our recommender system with a simple association rule based recommender system we have randomly drawn association lists for 300 information products with a total of

Table 6. Comparison of LSD-model with no trim and an acceptable expected type II error rate of 0.1 and with association rules with a support of 0.00015 and a confidence of 0.01.

No. of	LSD-model	Association Rules
produced recommendations	145	154
correct recommendations	96	100
unknown recommendations	35	38
false recommendations (type II error)	14	16
missed recommendations (type I error)	465	461

1966 pairs of information products. For each pair (x, i) a member of our research group answered with yes or no to the following question "Is i a good recommendation for x?". For 561 pairs the answer was yes, for 1100 pairs the answer was no. Unfortunately, however, for 305 pairs no expert evaluation could be done, because the web-site for i ceased to exist.

This evaluation approach differs from the usual machine learning methodology of splitting the data set in a training and a testing data set for testing the capability of the algorithm to extract patterns. In addition we test whether frequent co-purchase (or co-usage) is a suitable indicator for recommendations.

Table 6 shows a first result for the two suitably parametrized models. A rough comparison indicates that the LSD-model is at least as good as the association rule approach. However, a complete analysis of the trade-off of the type I and II error over the parameter range of the two models and a sensitivity analysis with regard to misspecification of parameters is still on the todo-list. With regard to parametrization, the type II error threshold is the only parameter of the LSD-model based algorithm. This parameter is independent of the size of the data set. The support and confidence parameters of association rule algorithms seem to depend on the data set size.

5 Further Research

The main contribution of this paper is that Ehrenberg's classical repeat-buying models can be applied to market baskets and describe – despite their strong independence and stationarity assumptions – the consumption patterns of information products – at least for the data set analyzed – surprisingly well. For e-commerce sites this implies that a large part of the theory developed for consumer panels may be applied to data from web-sites, too, as long as the analysis remains on the aggregate level.

For anonymous recommender services they seem to do a remarkable job of identifying non-random repeated-choice acts of consumers of information products as we have demonstrated in section 3. The use of the LSD-model for identifying non-random co-occurrences of information products constitutes a major improvement which is not yet present in other correlation-type recommender services.

However, establishing an empirical base for the validity of repeat-buying models in information markets as suggested in this article still requires a lot of additional evidence

and a careful investigation of additional data sets. We expect that such an empirical research program would have a good chance to succeed because to establish Ehrenberg's repeat-buying theory a similar research program has been conducted by Aske Research Ldt., London, in several consumer product markets (e.g. dentifrice, ready-to-eat cereals, detergents, refrigerated dough, cigarettes, petrol, tooth-pastes, biscuits, color cosmetics, ...) from 1969 to 1981 (see Ehrenberg [11]).

The current version of the anonymous recommender services (and the analysis in this article) still suffers from several deficiencies. The first is that new information products are daily added to the information broker's database 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, e.g. by testing the behavioral assumptions with data from the personalized part of the VU, as well as validation either by studying user acceptance or by controlled experiments, still has to be done. Third, for performance reasons the co-occurrence lists for each information product do not contain time-stamps. Therefore, the development of time-dependent e.g. alert systems has not been tried, although Ehrenberg's theory is in principle suitable for this task.

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, ...). Embedded in a personalized environment Ehrenberg's repeat-buying models may serve as the base of continuous marketing research services for managerial decision support which provide forecasts and classical consumer panel analysis in a cost efficient way.

6 Acknowledgment

We acknowledge the financial support of the Jubiläumsfonds of the Austrian National Bank under Grant No. 7925 without which this project would not have been possible. For the evaluation of the system we acknowledge support of the DFG for the project "Scientific Libraries in Information Markets" of the DFG program SPP 1041 "V3D2: Verteilte Verarbeitung und Vermittlung digitaler Dokumente". Thanks to Anke Thede and Andreas Neumann for correcting the final versions of this contribution and to one of the anonymous reviewers whose constructive critic helped us to improve this contribution considerably.

References

1. Aggrawall, C.C., Yu, P.S.: A New Approach to Online Generation of Association Rules. IEEE Trans. on Knowledge Eng. **13(4)** (2001) 527-540
2. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. Proc. of the 1993 ACM SIGMOD Int'l Conf. on Management of Data (1994) 207-216
3. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. 20th Int'l Conf. on Very Large Databases (1994) 478-499
4. Aske Research: The Structure of the Tooth-Paste Market. Aske Research Ltd., London (1975)

5. Avery, C., Zeckhauser, R.: Recommender Systems for Evaluating Computer Messages. CACM **40(3)** (1997) 88–89
6. Bellmann, S., Lohse, G.L., Johnson, E.J.: Predictors of Online Buying Behavior. CACM **42(12)** (1999) 32–38
7. Blischok, T.J.: Every Transaction Tells a Story. Chain Store Age Executive **71(3)** (1995) 50–62
8. Böhm, W., Geyer-Schulz, A., Hahsler, M., Jahn, M.: Repeat Buying Theory and its Application for Recommender Services. In: Opitz, O. (Ed.): Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg (to appear)
9. Chatfield, C., Ehrenberg, A. S. C., Goodhardt, G. J.: Progress on a Simplified Model of Stationary Purchasing Behavior. J. of the Royal Stat. Society A **129** (1966) 317–367
10. Charlton, P., Ehrenberg, A. S. C.: Customers of the LEP, Appl. Statist. **25** (1976) 26-30.
11. Ehrenberg, A. S. C.: Repeat-Buying: Facts, Theory and Applications. Charles Griffin & Company Limited, London (1988)
12. Feller, W.: An Introduction to Probability Theory and Its Applications. Vol. 2. John Wiley & Sons, New York (1971)
13. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg (1999)
14. Gaul, W., Schmidt-Thieme L.: Recommender Systems Based on User Navigational Behavior in the Internet. Behaviormetrika **29(1)** 2002 to appear.
15. Geyer-Schulz, A., Hahsler, M., Jahn, M.: myVU: A Next Generation Recommender System Based on Observed Consumer Behavior and Interactive Evolutionary Algorithms. In: Gaul, W., Opitz, O., Schader, M. (Eds.): Data Analysis – Scientific Modeling and Practical Applications, Studies in Classification, Data Analysis, and Knowledge Organization, Vol. 18, Springer, Heidelberg (2000) 447-457
16. Geyer-Schulz, A., Hahsler, M., Jahn, M.: Educational and Scientific Recommender Systems: Designing the Information Channels of the Virtual University. Int. J. of Engineering Education **17(2)** (2001) 153-163
17. Geyer-Schulz, A., Hahsler, M., Jahn, M.: Recommendations for Virtual Universities from Observed User Behavior. In: Gaul, W., Ritter, G., Schader, M. (Eds.): Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Heidelberg (to appear)
18. Goodhardt, G.J., Ehrenberg, A.S.C., Collins, M.A.: The Dirichlet: A Comprehensive Model of Buying Behaviour. J. of the Royal Stat. Society A **147** (1984) 621-655
19. Johnson, N.L., Kotz, S.: Discrete Distributions. Houghton Mifflin, Boston (1969)
20. Johnson, N.L., Kotz, S., Balakrishnan, N.: Discrete Multivariate Distributions. John Wiley & Sons, New York (1997)
21. Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., Duri, S.S.: Personalization of Supermarket Product Recommendations. Data Mining and Knowledge Discovery **5** (2001) 11–32
22. Lin, W., Alvarez, S.A., Ruiz, C.: Efficient Adaptive-Support Association Rule Mining for Recommender Systems. Data Mining and Knowledge Discovery **6(1)** (2002) 83-105
23. Mobasher, B., Cooley, R., Srivastava J.: Automatic Personalization Based on Web Usage Mining. CACM **43(8)** (2000) 142–151
24. Mobasher, B., Dai, H., Luo, T., Nakagawa, M.: Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. Data Mining and Knowledge Discovery **6** (2002) 61–82
25. Resnick, P., Varian, H.R. (1997): Recommender Systems. CACM **40(3)** (1997) 56–58
26. Powell, N., Westwood, J.: Buyer-Behaviour in Management Education. Appl. Statist. **27** (1978) 69-72
27. Sichel, H. S.: Repeat-Buying and the Poisson-Generalised Inverse Gaussion Distributions. Appl. Statist. **31** (1982) 193-204

28. Spiliopoulou, M.: Web Usage Mining for Web Site Evaluation. CACM **43(8)** (2000) 127–134
29. Srivastava, J., Cooley, R., Deshpande, M., Tan P.-N.: Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations **1(2)** (2000) 1–12
30. S.R.S.: The S.R.S. Motorists Panel. S.R.S. Sales Research Service, London (1965)
31. Stumme, G., Taouil, R., Bastide Y., Pasquier N., Lakhal L.: Computing Iceberg Concept Lattices with TITANIC. J. on Knowledge and Data Engineering to appear.
32. Wagner, U., Taudes, A.: Stochastic Models of Consumer Behaviour. Europ. J. of Op. Res. **29(1)** (1987) 1–23
33. Yen, S.J., Chen, A.L.P.: A Graph-Based Approach for Discovering Various Types of Association Rules. IEEE Trans. on Knowledge Eng. **13(5)** (2001) 839-845