Behavior-Based Recommender Systems as Value-Added Services for Scientific Libraries

Andreas Geyer-Schulz[1], Michael Hahsler[2], Andreas Neumann[1], and Anke Thede[1]

[1] Schroff-Stiftungslehrstuhl Informationsdienste und elektronische Märkte, Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany
[2] Institut für Informationsverarbeitung und Informationswirtschaft, WU-Wien, Augasse 2-6, A-1090 Wien, Austria

Abstract. Amazon.com paved the way for several large-scale, behavior-based recommendation services as an important value-added expert advice service for online book shops. In this contribution we discuss the effects (and possible reductions of transaction costs) for such services and investigate how such a value-added service can be implemented in the context of scientific libraries. For this purpose we present a new, recently developed recommender system based on a stochastic purchase incidence model, present the underlying stochastic model from repeat-buying theory and analyze whether the underlying assumptions on consumer behavior hold for users of scientific libraries, too. We have analyzed the logfiles with approximately 85 million http-transactions of the web-based online public access catalog (OPAC) of the library of the Universität Karlsruhe (TH) since January 2001 and performed some diagnostic checks. A test prototype is already operational and is currently being evaluated. The recommender service will be fully operational within the library system of the Universität Karlsruhe (TH) by the end of June 2002.

Keywords

Recommender Services, Digital Libraries, Consumer Behavior, Distributed Systems

1 Introduction

Recommender systems are regarded as strategically important information systems for e-commerce. Amazon.com, one of the most profitable internet companies, aggressively and successfully uses recommender services for building and maintaining customer relationships. Recommender services are attractive for both companies and their customers because of their capability to reduce transaction costs:

For companies they

- reduce the cost of customer service and support by shifting customers to web-based self-service platforms.
- improve cross- und up-selling revenues.
- support product managers by automatically generating additional product information.
- support marketing research by continuous consumer panel analysis.

For customers they

- reduce search cost and lead to a better overview of available products.
- support the discovery of related products and product groups.
- reveal market leaders and standard products.

Several innovative and experimental digital libraries, namely ResearchIndex [NEC Research Institute, 2002] whose tools are described in [Bollacker et al., 2000] and the digital libraries of the ACM [ACM, 2002] and the IEEE [IEEE, 2002] exploit these advantages although with different types of services. Several digital library and web search engine projects implemented services and interfaces to support the user's process of search and information extraction. An example is the Stanford Digital Library Project [Group, 1995] within the scope of which the system Fab [Balabanovic, 1997], [Balabanovic and Shoham, 1997] was developed. Fab is a combination of a content-based and a collaborative recommender system that filters web pages according to content analysis and creates usage profiles for user groups with similar interests as well. Popescul et al. [Popescul et al., 2001] have experimented with estimating collaborative filter models by latent variable models represented as Bayesian networks in the context of ResearchIndex. On the ResearchIndex dataset a Bayesian network with the structure of the classical diagnostic model has been evaluated by Pennock et al. [Pennock et al., 2000]. Another example is PADDLE [Hicks et al., 2000], a system dealing with the information overload caused by the mass of documents in digital libraries by introducing customization and personalization features. Inquirus 2 is a prototype of a personalized web search engine which uses automatic query modification, a personalized result scoring function [Glover et al., 1999]. Furthermore the UC Berkeley Digital Library Project [Wilensky et al., 1999] offers users to build personalized collections of their documents of interest. Recommendation services for digital libraries and their evaluation methods are discussed by Bollen and Rocha [Bollen and Rocha, 2000]. However, traditional scientific libraries seem to be late to realize the potential of recommender services for scientists and students alike. Our main objective is to reorganize scientific libraries with the help of recommender systems to customer oriented service portals. For students, university teachers and researchers an essential advantage of recommender systems is the reduction of search and evaluation cost for information products and the reduction of information overload by customization and personalization features. In addition, recommender systems trigger customer oriented procurement processes in libraries.

This article is structured into three parts:

1. In section 2 we discuss how recommender systems can be integrated with a legacy library system. We describe a loosely coupled distributed architecture which minimizes the required changes in the legacy system. It builds on the software pattern of active agents presented in [Geyer-Schulz and Hahsler, 2001] and which is a variant of a generic architecture for recommender systems shown in [Geyer-Schulz et al., 2002].

2. In section 3 we transfer a stochastic model from repeat-buying theory for recommender systems in libraries. We adapt the model for anonymous groups of users of the same information product and use this model to identify statistically significant purchase co-occurrences.
3. In section 4 we present the implementation of this system at the library of the University of Karlsruhe (TH) as a case study.

2 Recommender Services for Legacy Library Systems

The recommender services implemented at the Library of the Universität Karlsruhe (TH) are based on a generic architecture whose main idea is described by the pattern of a virtual library with active agents [Geyer-Schulz and Hahsler, 2001]. Figure 1 shows this pattern which uses Russell's and Norvig's agent analysis pattern [Russell and Norvig, 1995]. In this pattern a virtual library object, an observation agent and a interface agent interact in order to provide automated information services – in our case recommender services. The environment, which consists of the virtual library with its meta-information, the referenced information objects and the users, is perceived by the agents via their sensors. The agents gather information and influence their environment by updating information of the virtual library (observation agents) or passing results to users (interface agents).

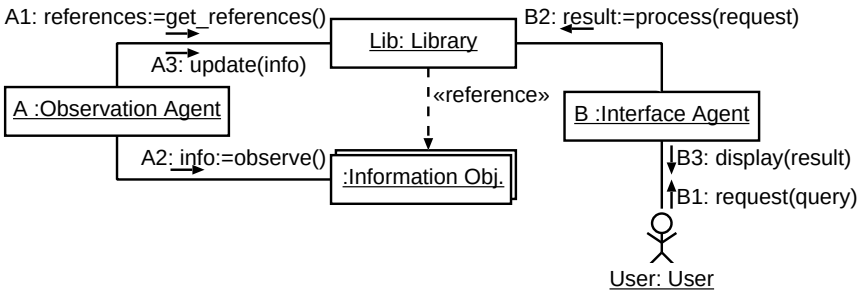

Fig. 1. Collaboration in an Agency for Virtual Libraries with Active Agents

In Figure 1 the whole task of observing a distant information object and presenting the results of the observation to a user is divided between an observation agent and an interface agent which act independently from each other. In Figure 1 this is denoted by the message sequences A1-A3 and B1-B3. This results in a weakening of consistency constraints which improves performance, reduces resource requirements (e.g. network bandwidth), and simplifies the implementation of the system. The pattern strikes a balance between consistency and performance requirements.

In Figure 2 we show an architecture for recommender services as an agency of software agents which consists of three layers, namely the Legacy Library System,

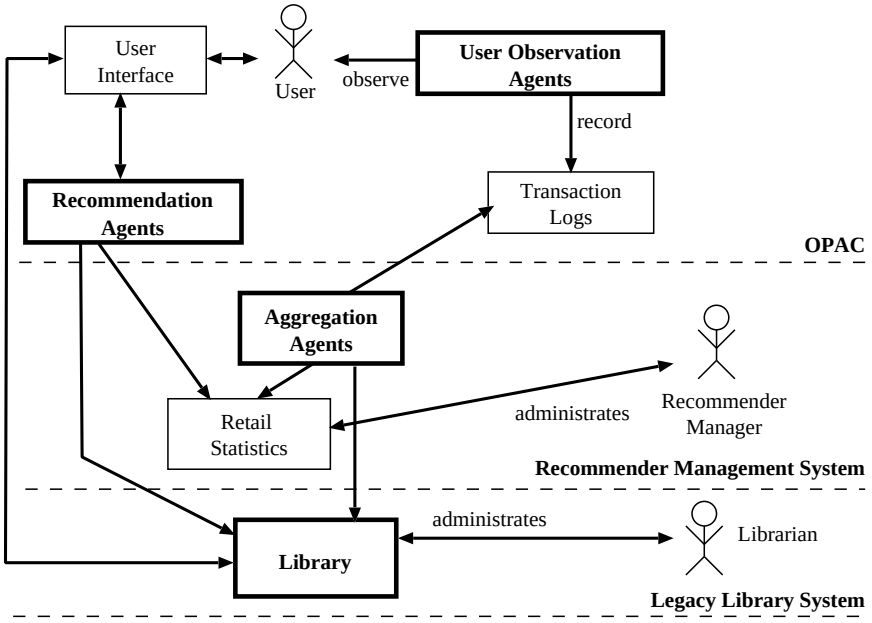

Fig. 2. An Architecture for an Information Broker with Recommender Services

the Recommender Management System, and the OPAC [Russell and Norvig, 1995]. The Legacy Library System corresponds to the meta-data management system, the recommender managment system to the broker management system, and the OPAC to the business-to-customer interface in the generic version of this architecture, see Geyer-Schulz et al. [Geyer-Schulz et al., 2002]. The task of the Observation Agent shown in Figure 1 is split in two parts and handled by the User Observation Agent and the Aggregation Agent of Figure 2. The interactions between persons, software agents and information stores is represented by arrows, where the direction indicates who starts an activity. A name near an arrow states the nature of the activity, if the arrow is unnamed, it means a simple request for information.

On the level of the legacy library system information objects are described by the library's traditional MAB format for books and journals which is the meta-data representation. However, because the interface to the other layers of this architecture is quite minimal (it requires only a method for retrieving the meta-data by a unique object key), the recommender management system and the OPAC are almost completely independent from the database technology used in this layer.

Because we use a legacy library system for meta-data management, standard interfaces for external applications are not available. The software agents we need are therefore integrated in the web interface of the OPAC. This implies that because of the legacy system the meta-data of an information object is stored distributed. Information observation agents update only meta-data stored outside the legacy library system. The distributed storage of information objects allows the integration

of agent-based information services which reduce the transaction cost of meta-data management and improve the service quality of the library system.

The recommender management system level and the OPAC are more tightly coupled. The recommender service we are describing in this article is based on observed user behavior. In an information market selecting a recommended information object (e.g. following a link) is considered as a purchase of this information object. In the library environment inspection of detailed library entries reveals interest in a certain book or journal. While lending data would have been available, for privacy reasons we have chosen to regard inspection of detailed library entries as purchase equivalent. The aggregation agent on the recommender management system level computes market-baskets, purchase histories and consumer panel statistics common in the retail industry (e.g. conditional purchase probabilities from transaction logs and customer experience profiles collected in the business-to-customer interface) which are described in detail in section 3. A consumer behavior model, a simple association rule model, and web-mining algorithms for robot elimination are integrated into the recommender management information system which supports the manager of the recommender in assessing the quality of the recommender system.

For the recommender agents in the OPAC retail statistics provide information on the preferences of users for books and journals. Analysis of http-logs with respect to additional library services with the goal of redesigning the web-site of the university library remains to be done and has the potential of increasing the convenience of retrieving scientific literature even further.

For behavior based recommendation services we need at least market baskets or purchase histories. Market baskets correspond to anonymous but session level data, purchase histories to user session data. User observation agents record the relevant transactions of a user. Several server-side recording mechanisms and their suitability for session identification have been discussed in the literature:

Http-logs. As Cooley [Cooley, 2000] has shown heuristics for session identification for pure http-logs range from 40 % to 60 % correctly identified sessions due to the combined effects of ISP proxy-servers and rotating IP addresses.

Http-logs with link embedded session IDs. Link embedded session IDs considerably improve the accuracy of session identification. However, three problem areas remain: First, robot identification [Tan and Kumar, 2002], second, public terminals which are accessed by several users sequentially, and third, the complete implementation of link embedding in the OPAC.

Http-logs with cookies. An advantage of the cookie mechanism compared to link embedded session IDs is that sessions can be identified without changes in legacy applications and that cookies are not included in bookmarks and thus do not lead to session restarts after potentially long periods of time [Geyer-Schulz et al., 2002].

Instrumented and specialized transaction logging with cookies. With this approach preprocessing of log-files becomes obsolete at the price of instrumenting the application [Geyer-Schulz et al., 2002].

The web-server of the university library collects http-logs with link embedded session IDs which are periodically posted via http to the recommendation server after local preprocessing. Preprocessing on the library server includes extraction of http GET requests with session IDs. Preprocessing on the recommendation server implements session splitting after a break of 15 minutes to take care of public access terminals in the library building and session restarts from bookmarks.

After preprocessing the aggregation agent computes market baskets, estimates a logarithmic series distribution (LSD) for the stochastic consumer behavior model presented in section 3, identifies, and extracts outliers as recommendations. In addition, basic diagnostic statistics are provided for recommender management. The aggregation agent performs incremental updates periodically. The update algorithm has quadratic time and space complexity relative to the number of items updated. The possibility of reducing the update periods improves the scalability of the system. Because of the extreme sparseness of observations relative to the total amount of books of the library (15 million), the memory of the recommender is kept. We expect that the recommender service will profit of several years of memory. However, the effects of memory length on the quality of the recommender as well as the repeat buying behavior of library users must be investigated.

The recommendation agent resides on the recommendation server and is implemented as a CGI-script. It generates recommendation pages with the corporate identity of the university library and its associated libraries. The service is accessed via embedded links in the references of the OPAC which are only visible if recommendations are available. Fault tolerance with respect to crashes of the recommendation server is achieved by exploiting the alternate tag mechanism of the html page description language.

3 A Stochastic Model from Repeat-Buying Theory

3.1 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) [Ehrenberg, 1988].

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 by integrating these results later (see figure 3).

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

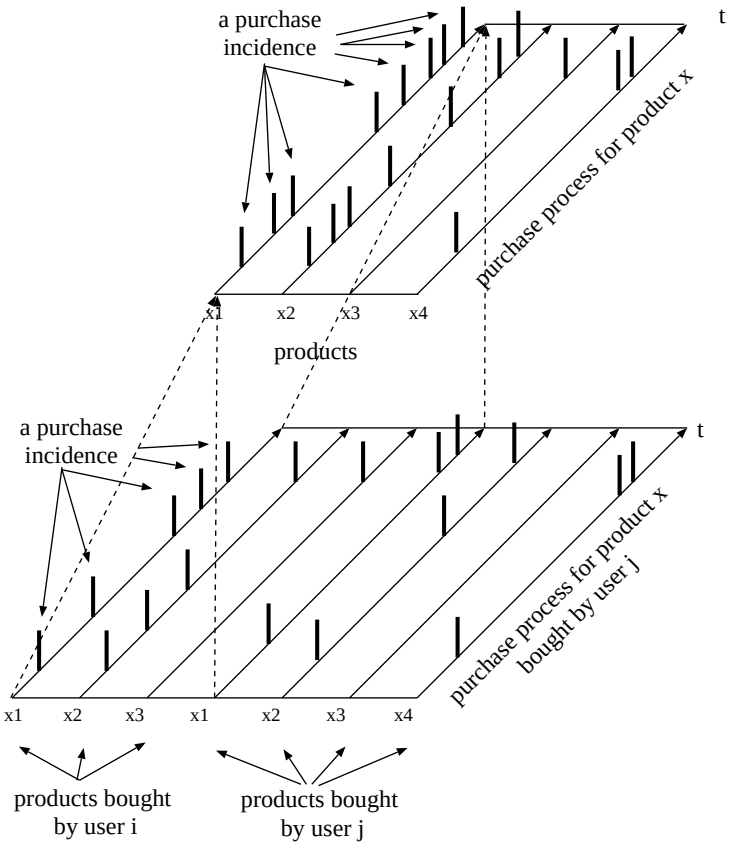

Fig. 3. Purchase Incidences as Independent Stochastic Processes.

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 information 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 [S.R.S., 1965].

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 [Ehrenberg, 1988] 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 everything works out fine as long as Ehrenberg's assumptions on consumer purchase behavior hold.

Figure 3 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. [Chatfield et al., 1966] have shown.

The logarithmic series distribution (LSD) describes the following frequency distribution of purchases (see Ehrenberg [Ehrenberg, 1988]), 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 [Johnson et al., 1993]. 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 [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 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 [Ehrenberg, 1988, p. 259]): If for different products $P, Q, R, S, \ldots$ the average purchase rate of P is independent of the purchase rates of the other products, and $\frac{P}{(P+Q+R+S+\ldots)}$ 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. [Research, 1975], [Charlton and Ehrenberg, 1976], [Powell and Westwood, 1978], [Sichel, 1982]).
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:

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

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

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

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

3. If $\sum p_r = 1$ for $r \geq 1$, 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 at the same buying occasion which follow independent Poisson processes with means μ_x and μ_i is [Johnson et al., 1997]: $p_r(x \wedge i) = \frac{e^{-\mu_x}\mu_x^r}{r!}\frac{e^{-\mu_i}\mu_i^r}{r!}$. 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

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

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 [Feller, 1971, 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 [Ehrenberg, 1988]).

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 by chance r-times together in a session. 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 [Ehrenberg, 1988].

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 in the same

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.

session 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 frequency 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 **??** as discussed in section **??**.

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.

4 A Recommender System for the Library of the University of Karlsruhe

Fig. 4. Anonymous Recommender of UB Karlsruhe

```
   Rechnungslegung nach IAS, US-GAAP und HGB im Vergleich / Born,
Karl (1999)
Empfehlungen:
1. Internationale Rechnungslegung / Buchholz, Rainer (2001)
2. Internationale Rechnungslegung / Kremin-Buch, Beate (2000)
3. IAS / US-GAAP / HGB im Vergleich / Hayn, Sven (2000)
4. Rechnungslegung nach IAS, US-GAAP und HGB im Vergleich / Born,
Karl (2000)
```

```
5. Internationales Rechnungswesen / Müller, Werner (2001)
6. Der Konzernabschluss nach HGB, IAS und US-GAAP / Schildbach,
T.(2001)
7. Analyse von Jahresabschlüssen nach US-GAAP und IAS / Dangel,
P.(2001)
8. Rechnungslegung international / Born, Karl (1999)
9. IAS / US-GAAP / HGB im Vergleich / Hayn, Sven (2000)
10. Bilanzanalyse international / Born, Karl (2001)
11. International accounting standards / Lüdenbach, Norbert (2001)
12. Rechnungslegung kompakt / Dusemond, Michael (2001)
====================== C U T ==========================
13. Rückstellungen nach HGB, US-GAAP und IAS / Daub, Sebastian
(2000)
14. Konzernabschluss international / Prangenberg, Arno (2000)
15. Internationale Rechnungslegung / Selchert, Friedrich W. (1998)
...
```

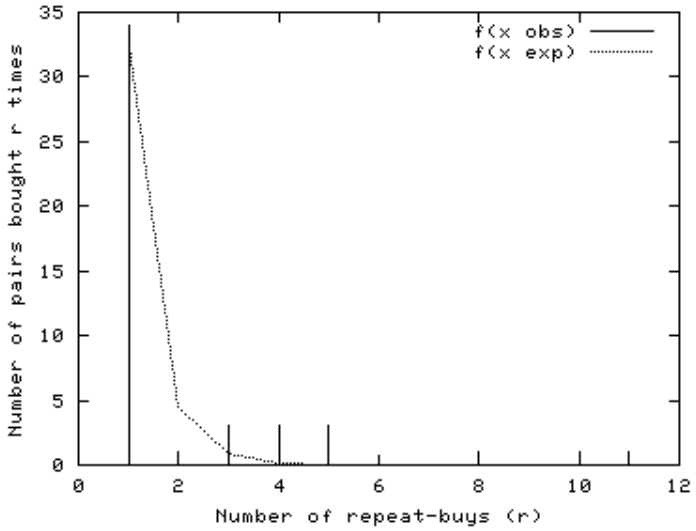

Fig. 5. Plot with linear y-axis scale

5 Conclusion

6 Acknowledgement

We gratefully acknowledge the funding of the project “Scientific Libraries in Information Markets” (“Wissenschaftliche Bibliotheken in Informationsmärkten”) by

Table 2. Statistics for “Rechnungslegung nach IAS...”

```
# Web-site: 6870526§BLB_OPAC
# Total number of observations: 47, Max repeat-buys: 11
# Sample mean=2.14893617021277 and var=5.44590312358533
# Case: E var>mean, Estimate for q=0.74712006072998

# Robust estimation: Trimmed begin 0: 0 / end 0.2: 9 (9 observations)
# Robust estimation: Number of observations: 38
# Robust mean=1.18421052631579 and var=0.308171745152355
# Robust estimate for q=0.280454684448242

# Plot: Observed repeat-buys and robust estimated LSD (q=0.280454684448242)

# r repeat-buys  nf(x_obs)  nf(x_exp)  f(x_exp)/f(x_obs)  show
        1           34        32.380          0.952         0
        2            1         4.541          4.541         0
        3            3         0.849          0.283         1
        4            3         0.179          0.060         1
        5            3         0.040          0.013         1
        6            0         0.009              -         0
        7            0         0.002              -         0
        8            1         0.001          0.001         1
        9            0         0.000              -         0
       10            1         0.000          0.000         1
       11            1         0.000          0.000         1

# Recommendations found with threshold=0.5: 12
# Chi-square test for q=0.74712006072998 and 47 observations

# Class  nf(x_obs)  nf(x_exp)  chi-square
# 1        34        25.541     2.802
# 2         1         9.541     7.646
# 3        12        11.918     0.001
#                             ------
#                             10.449

# Chi-square test for q=0.280454684448242 and 38 observations invalid.
# Less than 3 classes.
# Robust estimate performes better with chi-square value: 0, Col: II
```

the Deutsche Forschungsgemeinschaft (DFG) within the scope of the research initiative “Distributed Processing and Delivery of Digital Documents” (DFG-SPP 1041 “V^3D^2: Verteilte Vermittlung und Verarbeitung Digitaler Dokumente”).

Table 3. Finding classes with less than 0.50 random observations

r repeat-buys	$nf(x_{obs})$	$nf(x_{exp})$	$f(x_{exp})/f(x_{obs})$	Class shown
1	34	32.380	0.952	0
2	1	4.541	4.541	0
3	3	0.849	0.283	1
4	3	0.179	0.060	1
5	3	0.040	0.013	1
6	0	0.009	-	0
7	0	0.002	-	0
8	1	0.001	0.001	1
9	0	0.000	-	0
10	1	0.000	0.000	1
11	1	0.000	0.000	1

Table 4. Detailed results for observation period 2001-01-01 to 2002-06-09

	I q undef.	II no χ^2 (< 3 classes)	III Sign. $\alpha = 0.05$	IV Sign. $\alpha = 0.01$	V Not sign.	$\sum$
A Obs. < 10	1,638,782 (0)	83,057 (15,053)	0 (0)	0 (0)	0 (0)	1,721,839 (15,053)
B $\overline{x} = 1$	9,942 (0)	0 (0)	0 (0)	0 (0)	0 (0)	9,942 (0)
C $\overline{x} > \sigma^2$ $r \leq 3$	0 (0)	18,740 (6,451)	0 (0)	17 (12)	0 (0)	18,757 (6,463)
D $\overline{x} > \sigma^2$ $r > 3$	0 (0)	11,684 (11,684)	0 (0)	160 (160)	136 (136)	11,980 (11,980)
E $\sigma^2 > \overline{x}$	0 (0)	8,652 (8,652)	11 (11)	68 (68)	764 (764)	9,495 (9,495)
$\sum$	1,648,724 (0)	122,133 (41,840)	11 (11)	245 (240)	900 (900)	1,772,013 (42,991)

(x) indicates x lists with recommendations

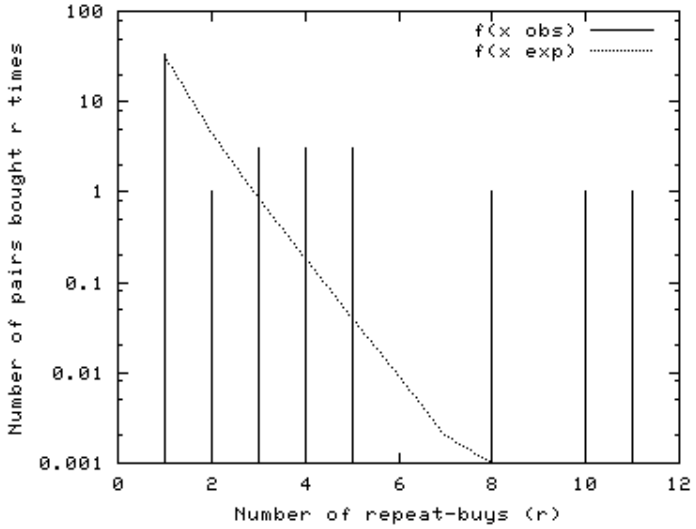

Fig. 6. Plot with logarithmic y-axis scale

References

[ACM, 2002]ACM (2002). ACM digital library. `http://www.acm.org/dl/`.

[Balabanovic, 1997]Balabanovic, M. (1997). An adaptive web page recommendation service. In *Proceedings of the 1st International Conference on Autonomous Agents*, Marina del Rey, California.

[Balabanovic and Shoham, 1997]Balabanovic, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

[Bollacker et al., 2000]Bollacker, K., Lawrence, S., and Giles, C. L. (2000). Discovering relevant scientific literature on the web. *IEEE Intelligent Systems*, 15(2):42–47.

[Bollen and Rocha, 2000]Bollen, J. and Rocha, L. M. (2000). An adaptive systems approach to the implementation and evaluation of digital library recommendation systems. In Borbinha, J. and Baker, T., editors, *Proceedings of the 4th European Conference on Digital Libraries*, volume 1923 of *LNCS*, pages 356–359. Springer.

[Charlton and Ehrenberg, 1976]Charlton, P. and Ehrenberg, A. S. C. (1976). Customers of the lep. *Applied Statistics*, 25:26–30.

[Chatfield et al., 1966]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*, 129:317–367.

[Cooley, 2000]Cooley, R. W. (2000). *Web Usage Mining: Discovery and Application of Interesting Patterns and Web Data*. PhD thesis. Faculty of the Graduate School the University Minnesota. Advisor: Jaideep Srivastava.

[Ehrenberg, 1988]Ehrenberg (1988). *Repeat-Buying: Facts, Theory and Applications*. Charles Griffin & Company Ltd, London, 2 edition.

[Feller, 1971]Feller, W. (1971). *An Introduction to Probability Theory and Its Application*, volume 2. John Wiley, New York, 2 edition.

Fig. 7. List of documents selected by class

```
Rechnungslegung nach IAS, US-GAAP und HGB im Vergleich /
Born, Karl (1999)

Empfehlungen:

1. Internationale Rechnungslegung / Buchholz, Rainer (2001)
2. Internationale Rechnungslegung / Kremin-Buch, Beate (2000)
3. IAS / US-GAAP / HGB im Vergleich / Hayn, Sven (2000)
4. Rechnungslegung nach IAS, US-GAAP und HGB im Vergleich /
Born, Karl (2000)
5. Internationales Rechnungswesen / Müller, Werner (2001)
6. Der Konzernabschluss nach HGB, IAS und US-GAAP /
Schildbach, T.(2001)
7. Analyse von Jahresabschlüssen nach US-GAAP und IAS /
Dangel, P.(2001)
8. Rechnungslegung international / Born, Karl (1999)
9. IAS / US-GAAP / HGB im Vergleich / Hayn, Sven (2000)
10. Bilanzanalyse international / Born, Karl (2001)
11. International accounting standards / Lüdenbach, Norbert
(2001)
12. Rechnungslegung kompakt / Dusemond, Michael (2001)
```


[Gaul and Ritter, 2002]Gaul, W. and Ritter, G., editors (2002). *Classification, Automation, and New Media*, volume 20 of *Studies in Classification, Data Analysis, and Knowledge Organization*, Heidelberg. Gesellschaft für Klassifikation e.V. (German Classification Society) http://www.gfkl.de, Springer-Verlag. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Passau, March 15-17, 2000. 535 pp.

[Geyer-Schulz and Hahsler, 2001]Geyer-Schulz, A. and Hahsler, M. (2001). Pinboards and virtual libraries - analysis patterns for collaboration. Technical Report 1, Institut für Informationsverarbeitung und -wirtschaft, Wirtschaftsuniverität Wien, Augasse 2-6, A-1090 Wien.

[Geyer-Schulz et al., 2002]Geyer-Schulz, A., Hahsler, M., and Jahn, M. (2002). Recommendations for virtual universities from observed user behavior. In [Gaul and Ritter, 2002], pages 273–280. Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e.V., University of Passau, March 15-17, 2000. 535 pp.

[Glover et al., 1999]Glover, E., Lawrence, S., Gordon, M. D., Birmingham, W., and Giles, C. L. (1999). Recommending web documents based on user preferences. In *SIGIR 99 Workshop on Recommender Systems*, Berkeley, CA.

[Group, 1995]Group, T. S. D. L. (1995). The stanford digital library project. *Communications of the ACM*, 38(4):59–60.

[Hicks et al., 2000]Hicks, D., Tochtermann, K., and Kussmaul, A. (2000). Augmenting digital catalogue functionality with support for customization. In *Proceedings of 3rd International Conference on Asian Digital Libraries*.

[IEEE, 2002]IEEE (2002). IEEE digital library. `http://www.ieee.org/products/onlinepubs/`.

[Johnson et al., 1993]Johnson, N. L., Kemp, A. W., and Kotz, S. (1993). *Univariate Discrete Distributions*. Wiley Series in Probability and Mathematical Statistics. John Wiley, 2nd edition.

[Johnson et al., 1997]Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. John Wiley, New York, 1 edition.

[NEC Research Institute, 2002]NEC Research Institute (2002). Researchindex. `http://citeseer.nj.nec.com/`.

[Pennock et al., 2000]Pennock, D., Horvitz, E., Lawrence, S., and Giles, C. L. (2000). Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence, UAI 2000*, pages 473–480, Stanford, CA.

[Popescul et al., 2001]Popescul, A., Ungar, L., Pennock, D., and Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *17th Conference on Uncertainty in Artificial Intelligence*, pages 437–444, Seattle, Washington.

[Powell and Westwood, 1978]Powell, N. and Westwood, J. (1978). Buyer-behaviour in management education. *Applied Statistics*, 27:69–72.

[Research, 1975]Research, A. (1975). The structure of the tooth-paste market. Technical report, Aske Research Ltd., London.

[Russell and Norvig, 1995]Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach - The Intelligent Agent Book*. Prentice-Hall, Upper Saddle River. Introduction and survey of AI.

[Sichel, 1982]Sichel, H. S. (1982). Repeat-buying and a poisson-generalised inverse gaussian distributions. *Applied Statistics*, 31:193–204.

[S.R.S., 1965]S.R.S. (1965). The S.R.S. motorists panel. Technical report, Sales Research Service, London.

[Tan and Kumar, 2002]Tan, P.-N. and Kumar, V. (2002). Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6:9–35.

[Wagner and Taudes, 1987]Wagner, U. and Taudes, A. (1987). Stochastic models of consumer behaviour. *European Journal of Operational Research*, 29(1):1–23.

[Wilensky et al., 1999]Wilensky, R., Forsyth, D., Fateman, R., Hearst, M., Hellerstein, J., Landay, J., Larson, R., Malik, J., Stark, P., Twiss, R., Tygar, D., House, N. V., Varian, H., Baird, H., Hurley, B., Kopec, G., Hirata, K., Li, W.-S., and Amir, A. (1999). Reinventing scholarly information dissemination and use. Technical report, University of California, Berkeley. `http://elib.cs.berkeley.edu/pl/about.html`.