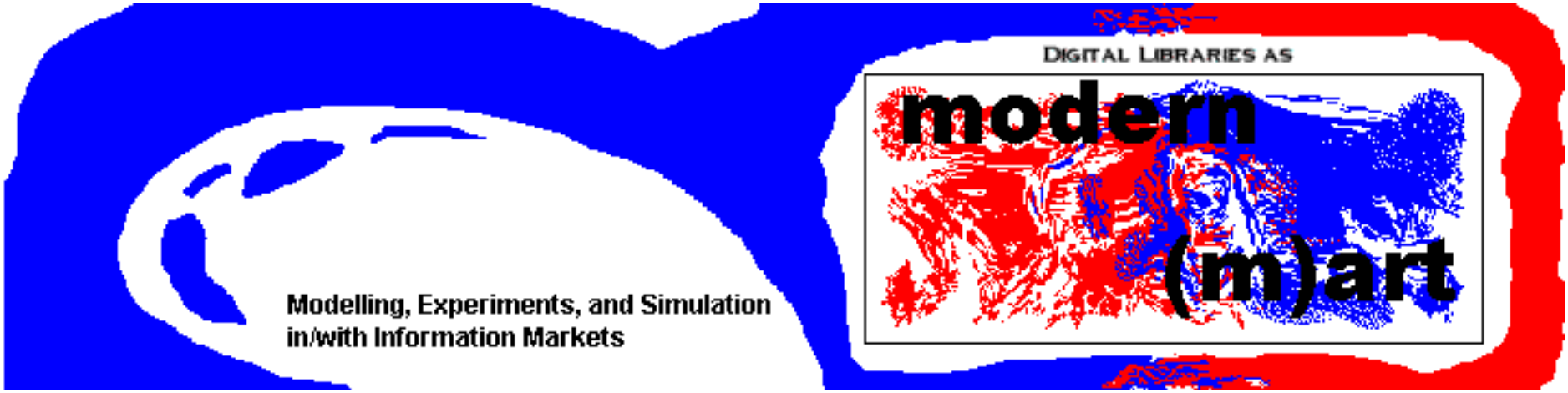
Verhaltensbasierte Empfehlungsdienste für wissenschaftliche Bibliotheken und Bibliotheksverbünde

Andreas Geyer-Schulz, Andreas Neumann, Anke Thede
Fakultät für Wirtschaftswissenschaften
Institut für Informationswirtschaft und -management

Universität Karlsruhe (TH)

7. InetBib-Tagung
12. November 2003, Frankfurt am Main



Gliederung

- Einführung und Motivation
- Web-Mining für Bibliotheksdaten
- Eine Architektur für verteilte (Legacy-)Bibliothekssysteme
- Ehrenberg's Repeat-Buying Theorie und ihre Adaption für Bibliotheken
- Das Recommendersystem der Universitätsbibliothek Karlsruhe
- Evaluation
- Ausblick

Projektteam

Institut für Informationswirtschaft und -management

Prof. Dr. Andreas Geyer-Schulz

beteiligte Mitarbeiter:

- Andreas Neumann
- Anke Thede

Universitätsbibliothek Karlsruhe

Ltd. Bibliotheksdirektor Christof-Hubert Schütte

beteiligte Mitarbeiter:

- Dr. Herbert Kristen
- Uwe Dierolf

Nutzung elektronischer wissenschaftlicher Information in der Hochschulausbildung

Klatt, R. et al. http://www.stefi.de/ (2001). BMBF-Studie

- Studierende:
  - Elektronische Recherche ist sehr wichtig: 76,4 %
  - Ihre Informationskompetenz ist durchschnittlich bis schlecht
  - Probleme:
    - Unübersichtlichkeit des Angebots: 35,9 %
    - Einschätzung der Qualität: 32,8 %
    - Informationsüberflutung: 26,4 %
  - Typische Recherchepfade: Fragen von Kommilitonen (60,2 %)
- Wissenschaftler:
  - Nutzung des Computers für Online-Recherchen:
    - Email-Austausch mit Studierenden u. Lehrenden: 91,4 %
    - Freie Suche mittels Suchmaschinen im Internet: 66,4 %

Recommendersysteme für wissenschaftliche Bibliotheken

- Wissenschaftler und Studierende sind immer mehr überfordert auf effiziente Weise relevante Literatur in konventionellen datenbankorientierten Katalogsystemen zu finden.

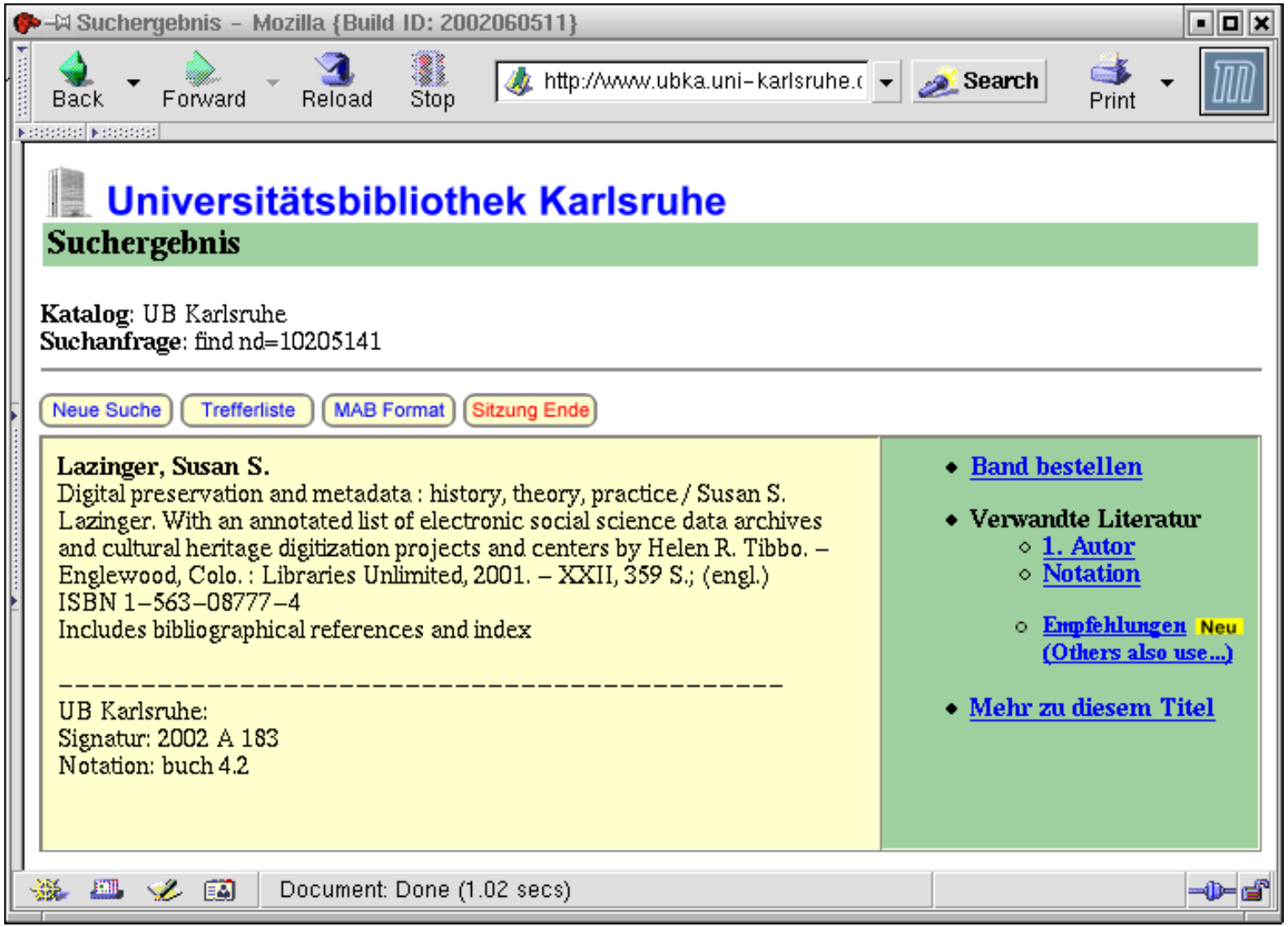
  $\rightarrow$ Weiter entwickelte Zugangswege sind nötig um Informationsüberflutung zu verhindern.

- Viele Universitäten unterrichten eine wachsende Anzahl Studierender mit einer mehr oder weniger festen Zahl von Mitarbeitern.

  $\rightarrow$ Studierende, Hochschullehrer und Wissenschaftler können einiges ihrer wertvollen Zeit zurückgewinnen, die z. Z. benötigt wird um z. B. einander Standardliteratur ihres Arbeitsgebietes zu empfehlen. Dies kann effizient durch verhaltensbasierte Expertenempfehlungsysteme übernommen werden.

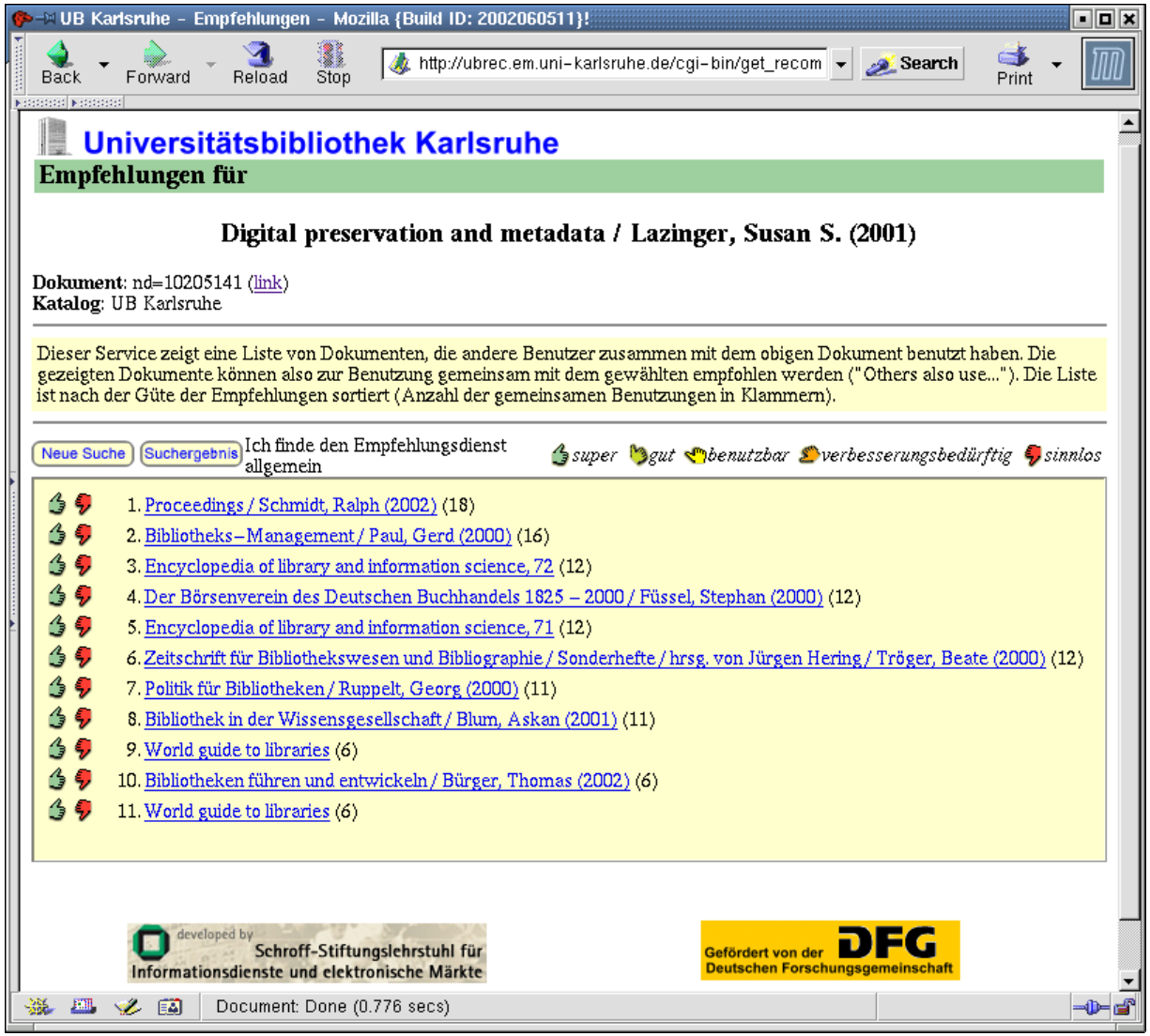
- Nutzen der Selbstselektionsbedingung um homogene Bibliotheksnutzergruppen zu identifizieren

Benutzerschnittstelle im OPAC I



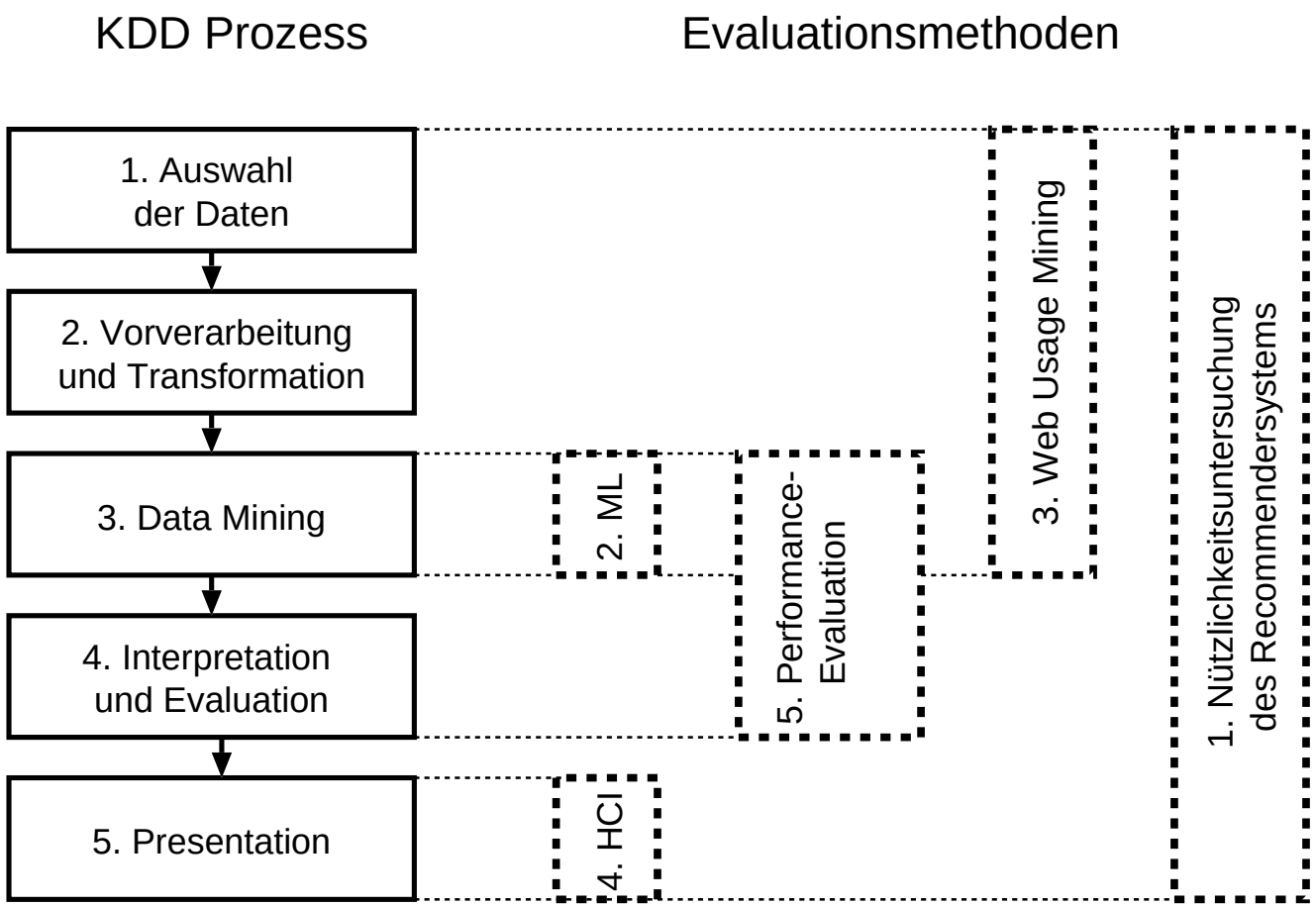
Detailansicht von Dokumenten (Bücher, Journale, Multi-Media, ...)

Benutzerschnittstelle im OPAC II



Empfehlungsliste

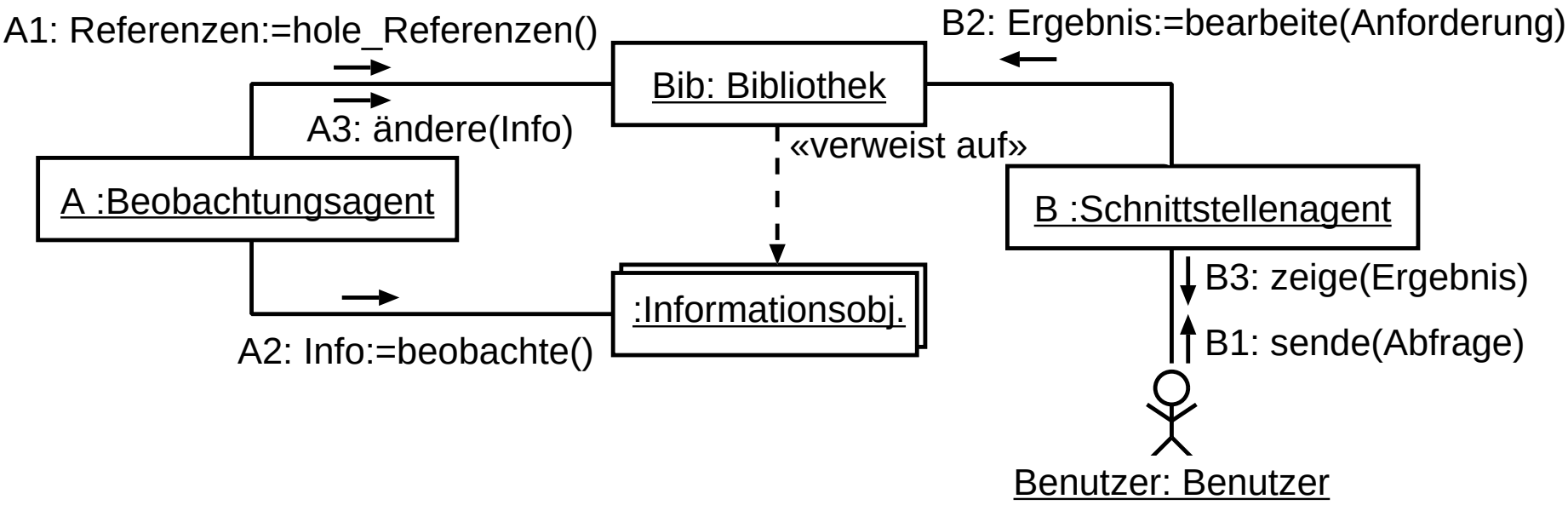
Web-Mining für Bibliotheksdaten



ML ... Maschinelles Lernen
HCI ... Human Computer Interface Design
KDD ... Knowledge Discovery und Data Mining

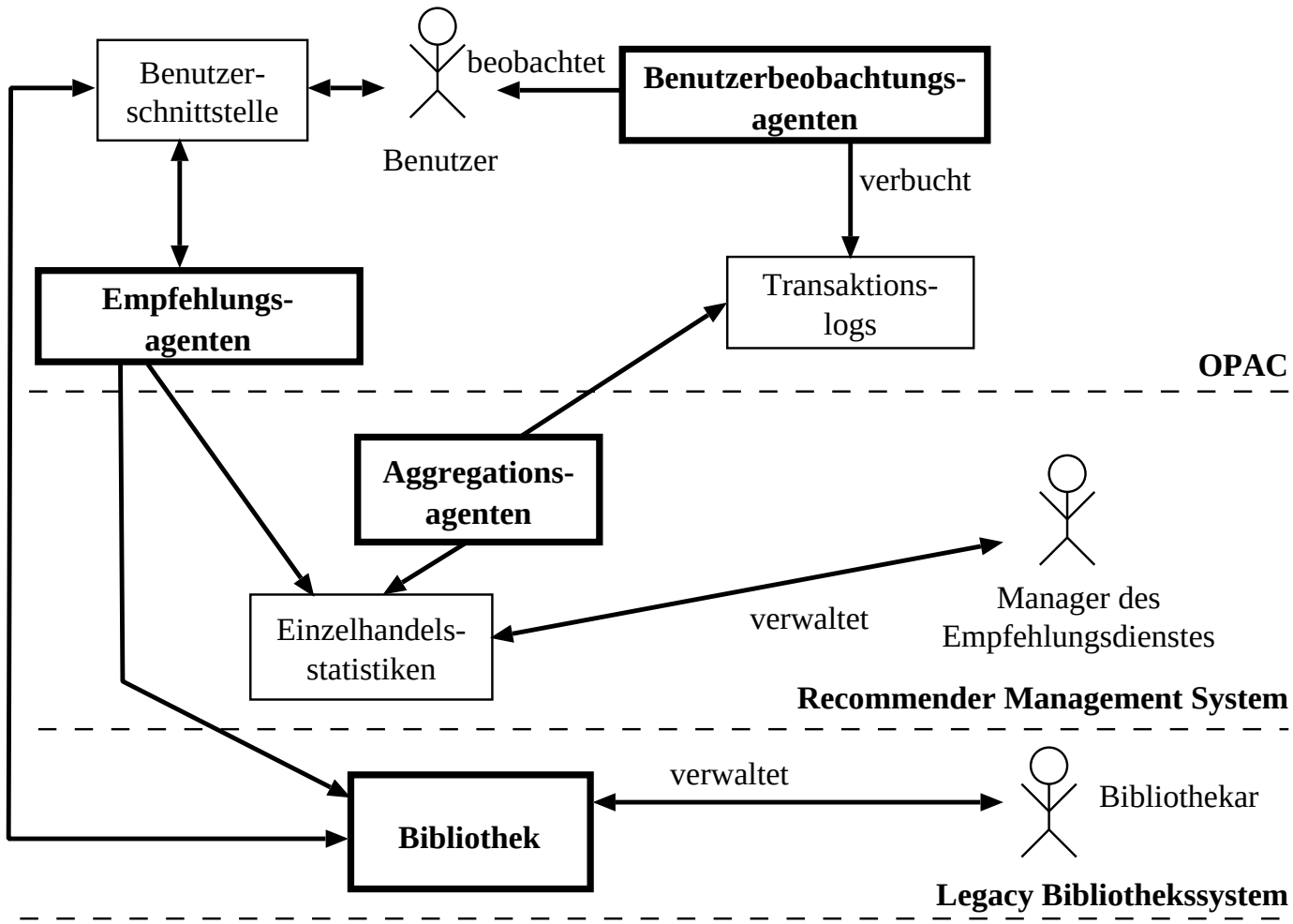
Der Web-Mining-Prozess für Bibliotheksdaten

Eine Agentur von Softwareagenten für verteilte Recommenderdienste



Eine Agentur von Softwareagenten als Analyse Pattern

Recommenderdienste für Legacy-Bibliothekssysteme I

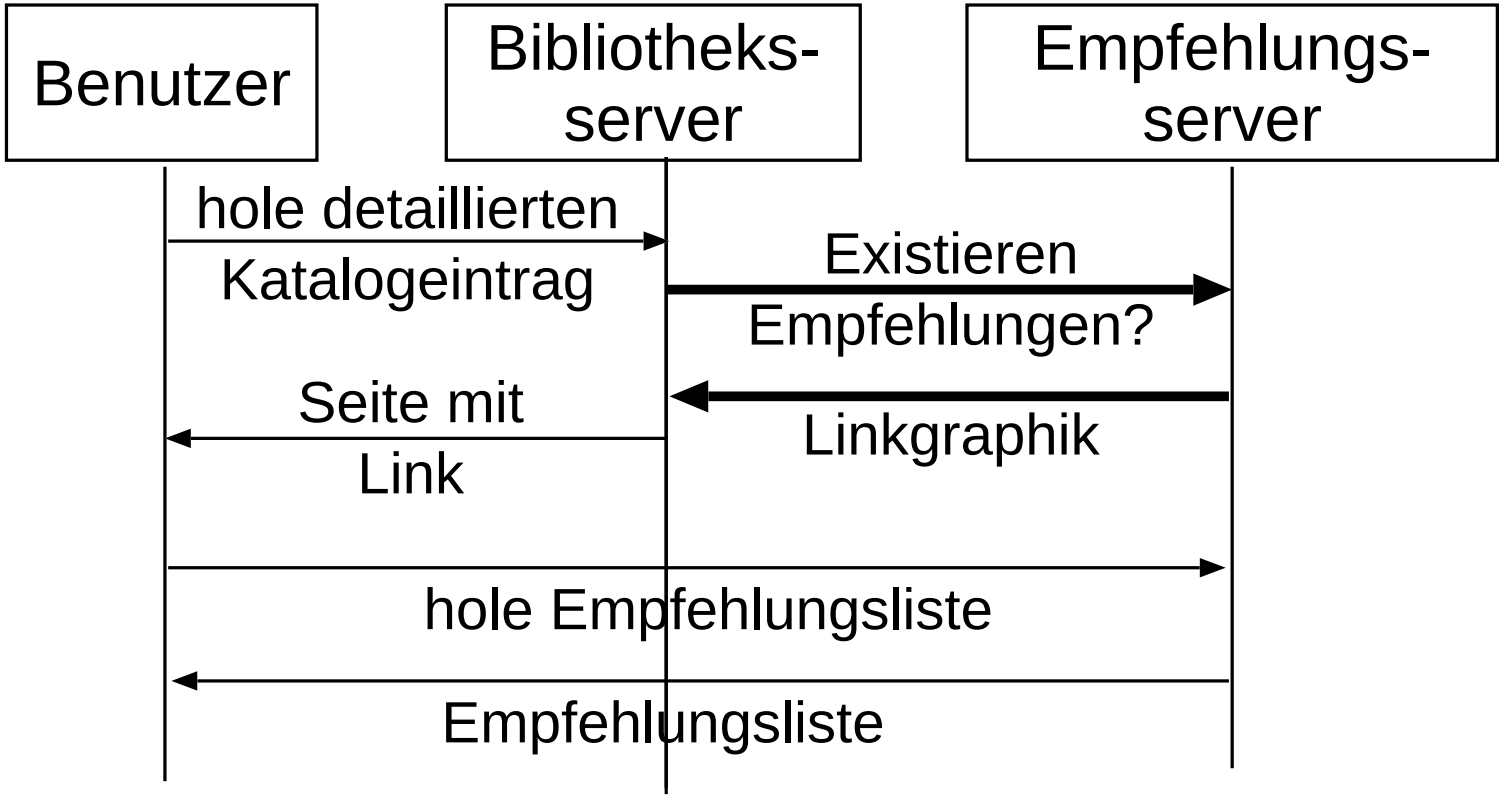


Architektur eines Bibliothekssystems mit Recommenderdiensten

Recommenderdienste für Legacy-Bibliothekssysteme II

- Benutzerbeobachtungsagent:
  - Informationsmarkt: Informationsauswahl = Kauf
  - http-logs mit link embedded session IDs
  - Preprocessing:
    - Extrahieren der GET requests
    - Session-Splitting zur Berücksichtigung von öffentlichen Terminals
- Aggregationsagent:
  - Erzeugt Warenkörbe
  - Schätzt eine Logarithmic Series Distribution (LSD-Verteilung)
  - Identifiziert and extrahiert Ausreißer als Empfehlungen
  - Führt periodisch inkrementelle Updates durch
- Empfehlungsagent:
  - CGI-script auf dem Recommendation Server
  - Generiert Empfehlungsseiten
  - Aufruf mittels Links aus dem OPAC

Recommenderdienste für Legacy-Bibliothekssysteme III



Sequenzdiagramm und Message Trace

Ein Stochastisches Kaufverhaltensmodell

- Basierend auf Ehrenbergs Repeat-Buying Theorie (1988), beschreibende Theorie zum Konsumentenverhalten
- Anonyme Kunden
- Analyse von Warenkörben
- Kaufverteilung von Produktpaaren ist eine LSD-Verteilung
- Berechnung der beobachteten und erwarteten LSD-Verteilung
- Anwenden eines χ^2-Tests zwischen beobachteter und erwarteter Verteilung
- Herausfinden der Ausreißer im Tail (Paare mit hoher Wiederkaufsrate) $\rightarrow$ Empfehlungen

Ehrenbergs Repeat-Buying Theorie I

- Deskriptive Theorie für Konsumverhalten
- Verallgemeinerung von regulärem Verhalten
- Starke empirische Evidenz für mehrere hundert Märkte für Konsumgüter seit den 1950ern
- Analyse von Haushaltspanels

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's Repeat-Buying Theorie II

Konsumentscheidungen:

1. Ob/Wann kauft ein Konsument ein bestimmtes Produkt? (Kaufentscheidung)

2. Wenn er kauft, welche Marke kauft er? (Markenwahl)

→ **Formalisierung des Kaufprozesses:**

- Das Konzept der Kaufgelegenheit

- Analysevariablen

  1. Länge der Analyseperiode
  2. Penetration - Anteil der Kunden, die Produkt gekauft haben.
  3. Durchschnittliche Kaufhäufigkeit - durchschnittliche Anzahl der Käufe für ein Produkt

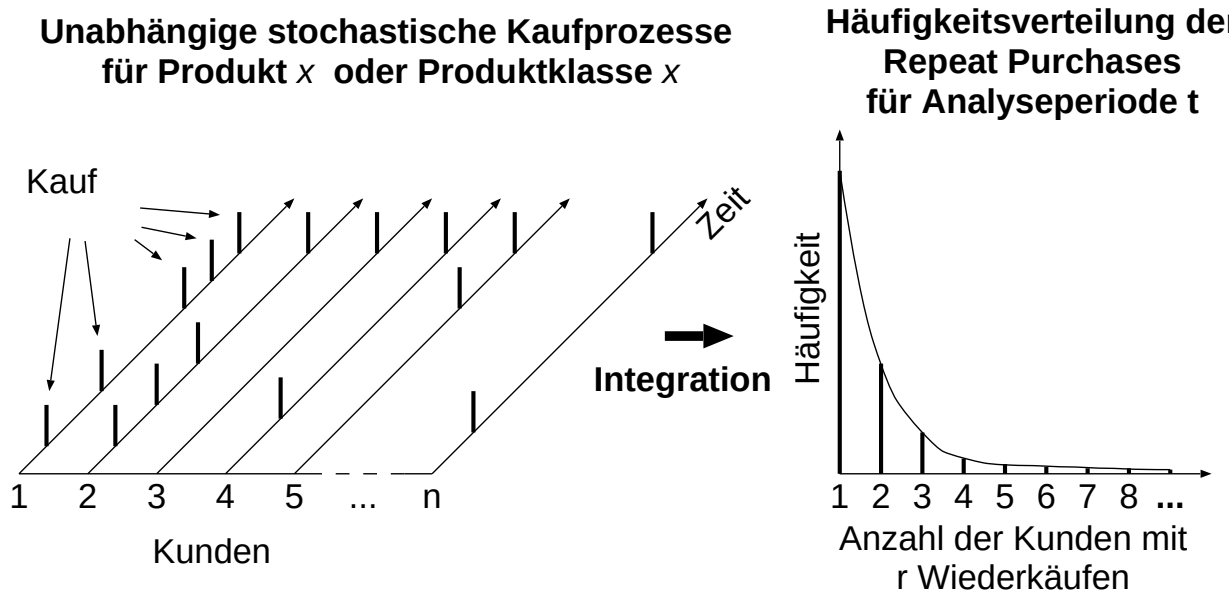
- Der Markt ist im Gleichgewicht (stationär)

Ehrenberg's Repeat-Buying Theorie III

Modelle und Parameter:

- Negatives Binomialverteilungs (NBD) Modell
  - m - Durchschnittliche Anzahl der Käufe pro Haushalt
  - k - Negativer binomial Exponent (geschätzt durch die Penetration b)
- **Logarithmische Reihenverteilungsmodell (LSD)**: Einfacheres Modell, wenn die Penetration unter 20% ist.
  - q - geschätzt aus w (durchschnittliche Kaufhäufigkeit)
- Dirichlet Modell: Ein allgemeineres Modell, mit dem auch Markenwahl und Marketing Mix berücksichtigt werden können.

Das LSD Modell



1. Kaufprozesse von Konsumenten folgen unabhängigen, stationären Poissonprozessen mit Mittelwert μ

2. Die Parameter μ folgen einer abgeschnittenen Γ-Verteilung

$\rightarrow$ Die Häufigkeitsverteilung folgt einer logarithmischen Reihenverteilung.

Das LSD Modell: Verwendung

Die logarithmische Reihenverteilung (LSD) beschreibt, wieviele Käufer ein spezisches Produkt 1, 2, 3, ... mal (ohne die Anzahl der Nichtkäufer zu berücksichtigen) kaufen:

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

Durchschnittliche Kaufhäufigkeit:

$$w = \frac{-q}{(1-q)\ln(1-q)} \tag{2}$$

Varianz:

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

Das LSD Modell: Annahmen

- **Der Anteil der Nichtkäufer in der Population ist unbekannt.**

  $\rightarrow$ Wahr für die meisten Dienste im Web (im Internet).

- **Die Käufe eines Konsumenten in aufeinanderfolgenden Perioden folgen einer Poissonverteilung mit einem längerfristig stabilen Mittelwert μ.**

  $\rightarrow$ Käufe erfolgen unabhängig von vergangenen Käufen und sie erfolgen so unregelmäßig, dass sie als zufällig betrachtet werden können.

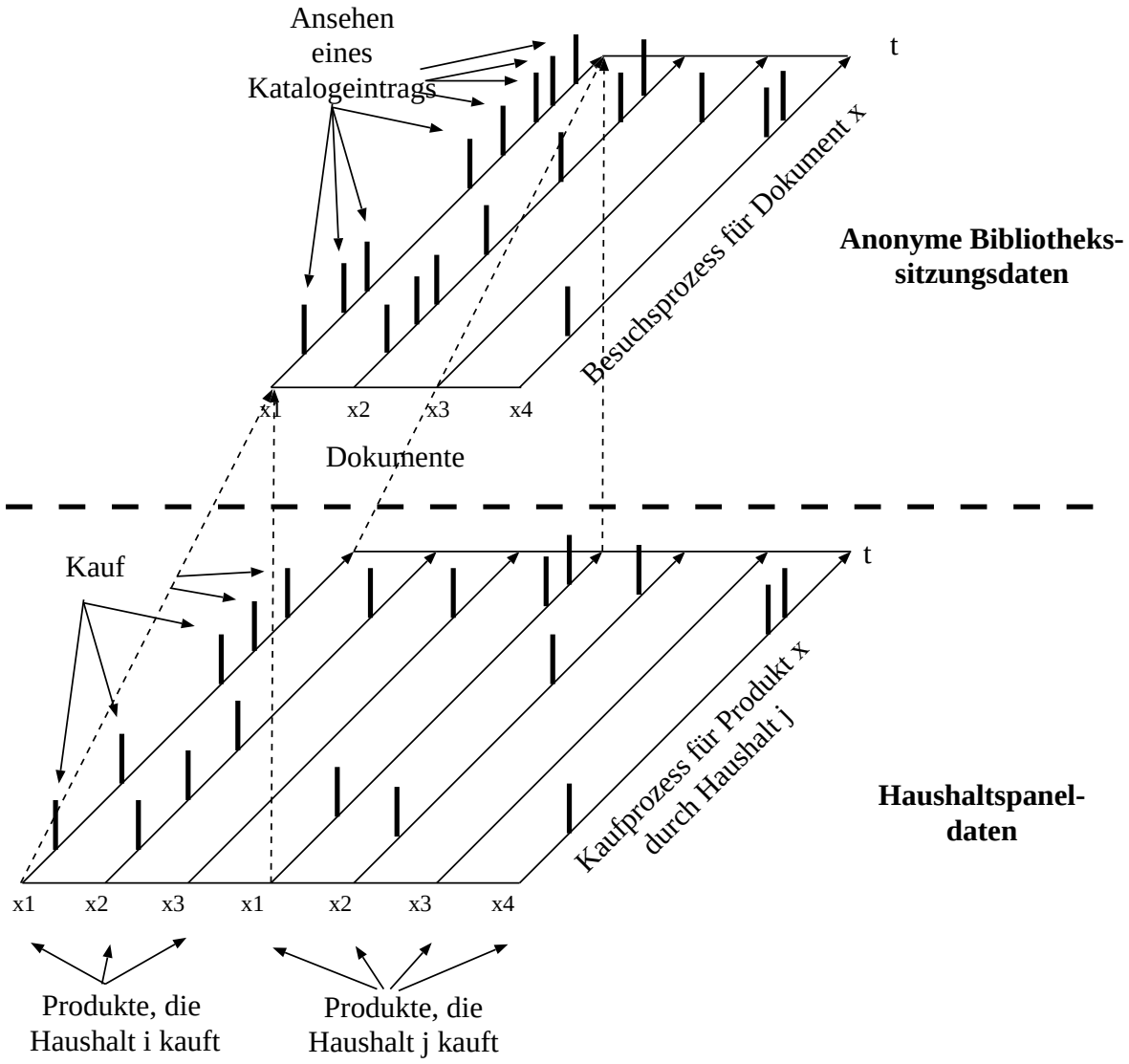
- **Die Verteilung von μ in der Population folgt einer abgeschnittenen Γ-Verteilung.**

  $\rightarrow$ folgt aus den Unabhängigkeitsannahmen und Γ-Verteilung sehr flexibel.

- **Der Markt ist im Gleichgewicht (stationär).**

  $\rightarrow$ Empirische Studien zeigen, dass die meisten Märkte für Konsumprodukte die meiste Zeit in oder nahe beim Gleichgewicht sind.

Stochastische Prozesse



Vom Haushaltspanel zur anonymen Bibliotheksnutzung

Kombination von Produkten

- Zwei Produkte x, i
- Zwei unabhängige Kaufprozesse (Poissonprozesse mit μ_x und μ_i)

$$p_r(x \wedge i) = \frac{e^{-\mu_x}\mu_x^r}{r!}\ \frac{e^{-\mu_i}\mu_i^r}{r!} \tag{4}$$

$$p_r(i \mid x) = \frac{p_r(x \wedge i)}{p_r(x)} = \frac{\frac{e^{-\mu_x}\mu_x^r}{r!}\ \frac{e^{-\mu_i}\mu_i^r}{r!}}{\frac{e^{-\mu_x}\mu_x^r}{r!}} = \frac{e^{-\mu_i}\mu_i^r}{r!} \tag{5}$$

$\rightarrow$ Die Häufigkeitsverteilung folgt wieder einer LSD

Selbstselektion I

- Durch die Wahl eines Produkts enthüllen Konsumenten ihre Zugehörigkeit zu bestimmten Gruppen mit lokal homogenen Präferenzen.
- Die Selbstselektion von Konsumenten erlaubt die Identifikation von Kaufgeschichten für Konsumenten mit homogenen (lokalen) Präferenzen aus Warenkörben.
- Dadurch kann die Repeat-Buying Theorie auf Warenkörbe (Kaufgeschichten mit latenter Identität von Haushalten) übertragen werden.

$\rightarrow$ Ehrenbergs Repeat-Buying Theorie ist auf Bibliotheken und Bibliotheksverbünde anwendbar.

In der angewendeten Form ist der Datenschutz (die Privatsphäre) für Bibliotheksnutzer gewährleistet.

Selbstselektion II

- Die wiederholte Wahl eines Produktes stellt ein glaubwürdiges Signal über die wahren Präferenzen eines Konsumenten dar, da die Selbstselektionsbeschränkungen halten:
  - Die Anzahl der Wiederkäufe eines Produktes muß so gewählt sein, dass Konsumenten nicht gewillt sind, ihre wahren Präferenzen durch Käufe anderer Produkte zu verschleiern, auch wenn ihnen dieses Verhalten Vorteile brächte.
  - Wenn dieses Wiederkaufsniveau für ein Produkt nicht erreicht wird, soll damit auch korrekt signalisiert werden, dass dieses Produkt nicht den Präferenzen der Käufer entspricht.

Algorithmus

1. Compute for all information products x in the browser sessions the frequency distributions for repeat-purchases of the co-occurrences of x with other information products in a session.

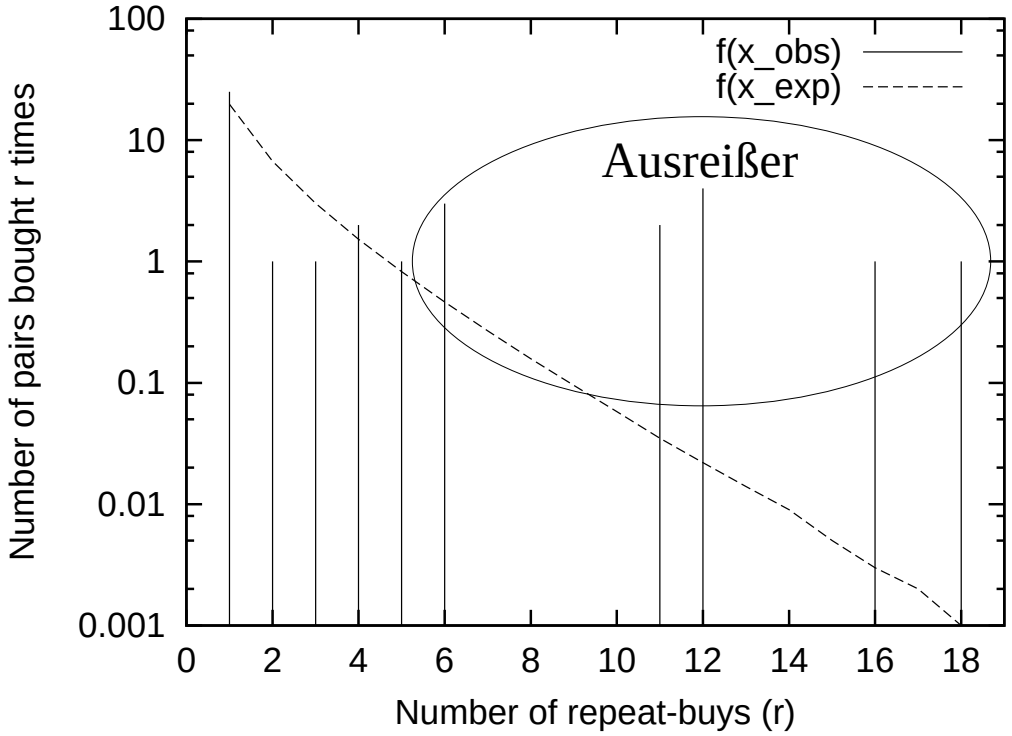
2. Discard all frequency distributions with less than l observations.

3. For each frequency distribution:

   (a) Compute the **robust** mean purchase frequency w by trimming the t percentile of the high repeat-buy pairs.
   (b) Approximate the parameter q for the LSD-model from $w = \frac{-q}{(1-q)(\ln(1-q))}$ (bisection/Newton method).
   (c) Apply a χ^2-goodness-of-fit test with a suitable α between the observed and the expected LSD distribution with a suitable partitioning.
   (d) Determine the outliers in the tail.
   (e) Finally, prepare the list of recommendations for information product x, if the LSD-model is significant and outliers exist.

Generierung der Empfehlungen

- Empfehlungen:
  - Produkte, die häufiger zusammen gekauft wurden als vom stochastischen Model erwartet,
  - widersprechen den Unabhängigkeitsannahmen des Modells.



Verteilungsfunktionen

Eigenschaften des Verfahrens

- Inkrementelle Updates
  - $O(n^2)$ in Zeit und Speicherplatz, n Anzahl der zu aktualisierenden Dokumente
  - Verbessert Skalierbarkeit
  - Gedächtnis des Recommenders bleibt bestehen

- Vergleich mit Assoziationsregeln (AR)
  - Niedrigere Update-Komplexität als AR-Algorithmus
  - Bei AR ist die Wahl des Minimum Support und Confidence kritisch und statistisch nicht hinreichend gesichert
  - AR-Verhalten ist nicht notwendigerweise stabil über die Zeit (gewählter Support und Confidence muß regelmäßig überprüft werden)
  - Berechnung der LSD-Verteilung erfolgt lokal für jedes Dokument, bei AR hängen Support und Confidence vom gesamten Datensatz ab $\Rightarrow$ Effizienter lokaler inkrementeller Update-Mechanismus möglich

Implementierung

- Universitätsbibliothek Karlsruhe
  - Internet/Catalog Service Provider für den südwestdeutschen Bibliotheksverbund
  - 23 Bibliotheken im Verbund
  - 15 Millionen Dokumente im Verbundkatalog
  - Empfehlungen über Kataloggrenzen hinweg
- Recommendation Server
  - 1,2 GHz AMD Athlon Processor
  - 1,5 GB Main Memory
  - Mandrake Linux (Kernel Version 2.4.17)
  - Recommender Software: Perl
  - MySQL Datenbank für Autor, Titel, etc.

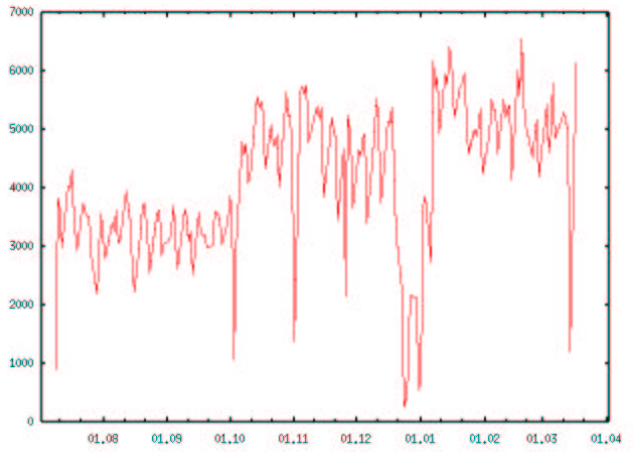
Auswertungsergebnisse 01.01.2001 – 08.11.2003

	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	637.185 (0)	105.143 (20.348)	0 (0)	0 (0)	0 (0)	742.328 (20.348)
B $\overline{x} = 1$	236.281 (0)	0 (0)	0 (0)	0 (0)	0 (0)	236.281 (0)
C $\overline{x} > \sigma^2$ $r \leq 3$	2.453 (0)	196.519 (49.003)	1.065 (893)	6.656 (3.359)	5.283 (2.220)	211.976 (55.475)
D $\overline{x} > \sigma^2$ $r > 3$	0 (0)	65.352 (65.215)	7.507 (7.507)	10.822 (10.793)	16.784 (15.693)	100.465 (99.208)
E $\sigma^2 > \overline{x}$	0 (0)	39.322 (39.322)	19.980 (19.980)	13.846 (13.846)	19.735 (19.735)	92.883 (92.883)
$\sum$	875.919 (0)	406.336 (173.888)	28.552 (28.380)	31.324 (27.998)	41.802 (37.648)	1.383.933 (267.914)

(n) bedeutet n Empfehlungslisten, Anzahl Empfehlungen: 2.646.166

Nutzungsstatistiken I

- Durchschnittliche Anzahl Aufrufe Juli 2002 – März 2003 (Montag – Freitag):
  - exist_recommendations: 4148,17 täglich
  - get_recommendations : 118,83 täglich
- WS 2002/2003: 122 Empfehlungslisten täglich genutzt.
- SS 2003: 168 Empfehlungslisten täglich genutzt.



Aufrufe von get_recommendation

Nutzungsstatistiken II

- Coverage:
  - 14.5 Millionen Dokumente im Verbund
  - 214 980 Empfehlungslisten mit 2 204 980 Empfehlungen (per 18/06/2003)
  - Coverage der Dokumente im Verbund: 1.48 %
  - Coverage der tatsächlich verwendeten Dokumente: 46.21 %

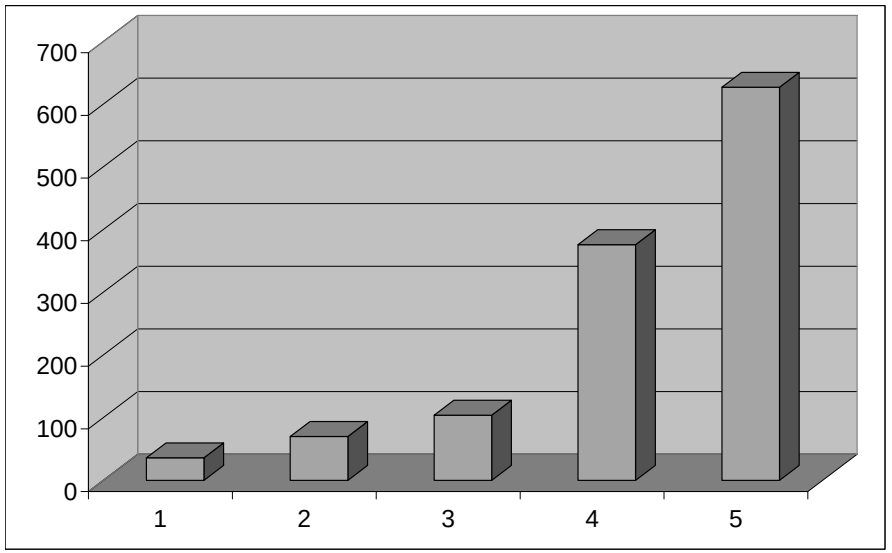
- Verwendung der Empfehlungen:
  - 2.6 % der gezeigten Links auf Empfehlungsseiten werden tatsächlich verfolgt
    (Vergleich: im Supermarkt zwischen 1.8% (IBM) und 2% (Netperceptions))
  - 18.6 % der Empfehlungslisten führten zum Besuch weiterer Dokumente durch Benutzer

Evaluation des Dienstes durch die Benutzer

- Frage:
  - Ich finde den Empfehlungsdienst allgemein:
  - (1) sinnlos – (2) verbesserungsbedürftig – (3) benutzbar – (4) gut – (5) super

- Ergebnis (Befragung von 07/02/2003 - 16/09/2003):
  - 1213 Beurteilungen
  - Mittelwert: 4,23
  - Standardabweichung: 1,02

- Mittlere Antwortrate: 4,9%



Ausblick

- Benutzerevaluation des Dienstes (laufend)
- Expertenevaluation einzelner Empfehlungen (in Arbeit)
- Berücksichtigung von Zeitinformationen zur Analyse von Moden, neuen Trends,. . .
- Diffusionsmodelle
- Repeat-Buying Information fließt in die Modellierung des Benutzerverhaltens für Marktmodelle ein
- Nutzen der Repeat-Buying Information zum Bestandsmanagement

Danksagung

Die Autoren danken der Deutschen Forschungsgemeinschaft (DFG), die das Projekt „Wissenschaftliche Bibliotheken in Informationsmärkten“ im Rahmen des Schwerpunktprogramms V^3D^2: Verteilte Vermittlung und Verarbeitung Digitaler Dokumente (DFG-SPP 1041) finanziert hat.

Der Empfehlungsdienst ist verfügbar unter:
www.ubka.uni-karlsruhe.de

Fragen?

Andreas Geyer-Schulz, Andreas Neumann, Anke Thede

Informationsdienste und elektronische Märkte,
Institut für Informationswirtschaft und -management,
Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany
{geyer-schulz, neumann, thede}@em.uni-karlsruhe.de