Recommender Systems for (Scientific) Libraries.

BibTip
Ein Service der UB Karlsruhe

Version: February 19, 2008

Andreas Geyer-Schulz
Schroff-Chair
Information Services and Electronic Markets
Faculty of Business Engineering
Building 20.20 Computing Center, Level 1, Room 169
Universität Karlsruhe (TH)
D-76128 Karlsruhe

Phone: +49-721-608-8401 Fax: +49-721-608-8403
E-Mail: Andreas.Geyer-Schulz@em.uni-karlsruhe.de
Contact arranged by E-Mail

http://www.em.uni-karlsruhe.de/
# Contents

1 Menu:  & 5  

2 Motivation.  & 6  

3 Project goals.  & 7  

3.1 Example: Detailed View of Documents (UBKA)  & 8  
3.1.1 Example: List of Recommendations (UBKA)  & 9  
3.1.2 Example: List of Recommendations (UBKA)  & 10  

3.2 User Interface: Library OPAC Search (KVK)  & 11  
3.2.1 User Interface: Hit List (KVK)  & 12  
3.2.2 User Interface: Document Inspection Page (KVK)  & 13  
3.2.3 User Interface: List of Recommendations (KVK)  & 14  

4 Scientific roots and theories combined.  & 15
4.1 Paul A. Samuelson: Revealed Preference Theory
(Nobel Prize in Economics 1970) .................................................. 17
4.2 A. Michael Spence and Joseph Stiglitz:
Market Signalling and Screening by Self-Selection
(Nobel Prize in Economics 2001) .................................................. 20
4.3 Modelling as a “Gestaltproblem”:
Causal models versus noise models .............................................. 21
4.4 Ideas from ensemble methods. .................................................... 22
4.5 Ehrenbergs Repeat-Buying Theory. ............................................ 23

5 Statistical models.  ................................................................. 24
  5.1 The LSD Model: A Mixture of Poisson Processes ................. 25
     5.1.1 Stochastic Processes ......................................................... 26
     5.1.2 Generating Recommendations ........................................... 27
  5.2 Multinomial models: Small samples ........................................ 28
     5.2.1 Distributions of Co-Inspections (KVK Data) ...................... 29
     5.2.2 Data Setup .................................................................... 30
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.2.3</td>
<td>General Concept</td>
<td>31</td>
</tr>
<tr>
<td>5.2.4</td>
<td>An Ideal Decision Maker</td>
<td>32</td>
</tr>
<tr>
<td>5.2.5</td>
<td>POSICI: Probability of Single Item Co-Inspections</td>
<td>33</td>
</tr>
<tr>
<td>5.2.6</td>
<td>Inspection Probabilities for Growing $k$ and $n$ in POSICI</td>
<td>34</td>
</tr>
<tr>
<td>5.2.7</td>
<td>Interpretation of POSICI</td>
<td>35</td>
</tr>
<tr>
<td>5.2.8</td>
<td>POMICI: Probability of multiple items co-inspections</td>
<td>36</td>
</tr>
<tr>
<td>5.2.9</td>
<td>Partition Probabilities for Growing $k$ and $n$ in POMICI</td>
<td>38</td>
</tr>
<tr>
<td>5.2.10</td>
<td>Interpretation of POMICI</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>Other applications (prototypes).</td>
<td>40</td>
</tr>
<tr>
<td>6.1</td>
<td>Restricted random walk clustering for indexing</td>
<td>41</td>
</tr>
<tr>
<td>6.2</td>
<td>Adaptive recommender systems</td>
<td>42</td>
</tr>
<tr>
<td>6.3</td>
<td>Exploration of the information space</td>
<td>45</td>
</tr>
<tr>
<td>7</td>
<td>System architecture and scalability.</td>
<td>46</td>
</tr>
<tr>
<td>7.1</td>
<td>An Agency of Software Agents for Distributed Recommender Services</td>
<td>47</td>
</tr>
</tbody>
</table>
7.2 Recommender Services for Legacy Library Systems I ........ 48
7.3 Recommender Services for Legacy Library Systems II ........ 49

8 Service integration into web OPACS of scientific libraries. 50
  8.1 Integration Variant I: apache-logs, cgi-scripts .......... 51
  8.2 Integration Variants II and III: Web 2.0 Technology .... 52

9 Development of a service concept and a service organization. 53

10 Development of a global service infrastructure for (scientific) libraries. 54
  10.1 Reference Customers ............................................ 55
1 Menu:

- Project goals and motivation.
- Scientific roots and theories combined.
- Statistical models.
- Other applications (prototypes).
- System architecture and scalability.
- Service integration into web OPACS of scientific libraries.
- Development of a service concept and a service organization.
- Development of a global service infrastructure for (scientific) libraries.

(Context: Recommender services.)
2 Motivation.

• Scientists and students are more and more incapable of efficiently finding relevant literature in conventional database oriented catalog systems.
  – Problems: Supply complexity, evaluation of the quality, information overload
  – Typical literature research path: Asking peers
  → More sophisticated access paths to prevent information overload are needed.

• Many universities are teaching a growing number of students with a more or less fixed staff size.
  → Students, university teachers and researchers could free some of their valuable time that is needed now to recommend each other e.g. standard literature of their fields which could be done easily by behavior-based expert advice services.
3 Project goals.

- Recommender systems aggregate knowledge from many peer groups to the level of expert advice services.
- Recommender systems bear the potential to significantly reduce transaction costs for literature searches by means of their aggregation capabilities.

Goal: To convert library OPACs into customer oriented service portals supporting the interaction of the customers.
3.1 Example: Detailed View of Documents (UBKA)

Universitätsbibliothek Karlsruhe

Suchergebnis

Katalog: UB Karlsruhe
Suchanfrage: find nd=10205141

Lazinger, Susan S.
Includes bibliographical references and index

UB Karlsruhe:
Signatur: 2002 A 183
Notation: buch 4.2

- Band bestellen
- Verwandte Literatur
  - 1. Autor
  - Notation
  - Empfehlungen Neu
    (Others also use...)
- Mehr zu diesem Titel
3.1.1 Example: List of Recommendations (UBKA)

![Example Image]

**University of Karlsruhe Library**

**Empfehlungen für**

**Digital preservation and metadata / Lazinger, Susan S. (2001)**

**Dokument: ad=2020141 (link)**
**Katalog: UB Karlsruhe**

Dieser Service zeigt eine Liste von Dokumenten, die andere Benutzer zusammen mit dem aktuellen Dokument benutzt haben. Die gezeigten Dokumente können also zur Benutzung gemeinsam mit dem gewählten empfohlen werden ("Others also use..."). Die Liste ist nach der Güte der Empfehlungen sortiert (Anzahl der gemeinsamen Benutzungen in Klammern).

3. Encyclopedia of Library and Information science, 72 (13)
5. Encyclopedia of Library and Information science, 71 (12)
9. World guide to libraries (6)
11. World guide to libraries (6)
3.1.2 Example: List of Recommendations (UBKA)
3.2 User Interface: Library OPAC Search (KVK)
3.2.1 User Interface: Hit List (KVK)
3.2.2 User Interface: Document Inspection Page (KVK)
3.2.3 User Interface: List of Recommendations (KVK)

**Empfehlungen für: Die Phönizier / Sommer, Michael, 2005**

4 Scientific roots and theories combined.
• Paul A. Samuelson: Revealed Preference Theory (Nobel Prize in Economics 1970)

• A. Michael Spence and Joseph Stiglitz: Market Signalling and Screening by Self-Selection (Nobel Prize in Economics 2001)

• Modelling as a “Gestaltproblem”. Causal models versus noise models.

• Ideas from ensemble methods: e.g. Albert Einstein’s work about the molecular Brownian motion. (Nobel prize in physics 1921).

• Ehrenbergs repeat-buying theory.
4.1 Paul A. Samuelson: Revealed Preference Theory
(Nobel Prize in Economics 1970)
Purchases reveal preferences.
4.2 A. Michael Spence and Joseph Stiglitz: Market Signalling and Screening by Self-Selection (Nobel Prize in Economics 2001)

Signalling (M. Spence):
- The choice of John von Neumann’s Quantenmechanik (quantum mechanics) signals experience and expert knowledge.
- Incentive compatibility: voluntary and truthful.

Screening (J. Stiglitz):
- A professor offers a menu of course books each of which requires a certain level of experience and expert knowledge.
- Selection of a book by a student reveals the student’s level.
- Incentive compatibility: voluntary and truthful.

Self-selection processes build homogenous clusters of readers.
4.3 Modelling as a “Gestaltproblem”:
Causal models versus noise models

Economics and management science versus natural sciences.
4.4 Ideas from ensemble methods.

Ensemble methods are methods which measure the behavior and the interactions of a large set of particles only by observing of the states of the aggregate (e.g. pressure, temperature and volume of a gas).

Examples:

- Ludwig Boltzmann’s Thermodynamics.
- Albert Einstein’s work about the molecular Brownian motion. (Nobel prize in physics 1921).
4.5 Ehrenbergs Repeat-Buying Theory.

Independent Poisson processes act as models of noise and filters: Outliers are recommendations.
5 Statistical models.
5.1 The LSD Model: A Mixture of Poisson Processes

The logarithmic series distribution describes, how many consumers have bought a specific product $1, 2, 3, \ldots$ times in the observation period (without taking non-buyers into account):

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

Mean Purchase Frequency:

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

Variance:

$$\sigma^2 = \frac{-q}{(1 - q) \ln(1 - q)} \frac{1+q}{(1 - q)^2 \ln(1 - q)}$$ \hspace{1cm} (3)
5.1.1 Stochastic Processes

From Consumer Panels to Anonymous Library Users
5.1.2 Generating Recommendations

- Products which are bought together more frequently than expected by the stochastic model
- violate independence assumptions of the model

![Graph showing Observed Distribution Function with f(x_obs) and f(x_exp)]
5.2 Multinomial models: Small samples

<table>
<thead>
<tr>
<th></th>
<th>UBKA</th>
<th>KVK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of total documents in catalog</td>
<td>1,000,000</td>
<td>&gt; 10,000,000</td>
</tr>
<tr>
<td>Number of total co-inspected documents</td>
<td>527,363</td>
<td>255,248</td>
</tr>
<tr>
<td>Average market baskets size</td>
<td>4.9</td>
<td>2.9</td>
</tr>
<tr>
<td>Av. aggregated co-inspections per document</td>
<td>117.4</td>
<td>5.4</td>
</tr>
</tbody>
</table>

- When starting to monitor new catalogs (stores) no information about the usage age of the documents (products) in the catalog is available.
- With smaller market baskets and a greater number of total documents it takes much more time for many co-inspections per document to occur.
- Due to sample size constraints methods using statistical tests on distributions (like LSD) are only reliably applicable with many co-inspections.
- Special small sample statistics are needed to compute recommendations out of samples of few co-inspections.
5.2.1 Distributions of Co-Inspections (KVK Data)
5.2.2 Data Setup

- The number of total documents $n + 1$ in the catalog is finite but unknown.
- Recommendations are computed separately for each document $D$.
- Each user session (market basket) contains all documents that the user inspected within that session, multiple inspections of the same document are counted as one.
- All user session are aggregated.
- The set $C(D)$ contains all documents, that at least one user has inspected together with $D$.
- The number of co-inspections with $D$ of all elements of $C(D)$ is known, this histogram is called $H(D)$.
- $H(D)$ can be interpreted as an integer partition with the number of co-inspections of each co-inspected document as the summands. (For partition theory see Euler, MacMahon, Hardy, Ramanujan, Rademacher.)
5.2.3  General Concept

Assumption: All documents in the catalog have the same probability of being co-inspected.

- In real systems generally this assumption does not hold.
- Especially when starting to observe new catalogs no information about the underlying distribution of the inspection processes of documents is known.
- We do not develop a causal model of a decision maker, we construct an ideal decision maker without preferences.
5.2.4 An Ideal Decision Maker

- Is a prototype for a class of homogenous decision-makers
- Makes choices randomly. But from what choice set?
- The conceptual model of the choice set:
  - Total set $\supseteq$ awareness set $\supseteq$ consideration set $\supseteq$ choice set (Kotler 1980)
- Evidence for the model:
  - Awareness set from 3 – 11 products (Narayana and Markin (1975)
  - Interaction of emotional and rational brain activity (Bechara et al. 1997): Emotional activity reduces size of choice set of normal persons, whereas for patients with prefrontal damage this could not be observed.
  - In a brand choice experiment successful branding reduces the choice set to 1 (the winning brand) and the regions of the brain responsible for emotions show high activity when measured by fMRI (Deppe et al. 2005).

Consequence: We consider a sequence of increasing event spaces.
5.2.5 POSICI: Probability of Single Item Co-Inspections

Question: What is the probability $p_j$ that at least one other document has been co-inspected exactly $j$-times with document $D$?

- Let $n+1$ be the number of total documents, $k$ the number of non-aggregated co-inspections (multiple inspections in different sessions are counted separately)
- Let $(N_1, \ldots, N_n)$ be the vector of the number of times document $i$ ($1 \leq i \leq n$) was co-inspected with $D$
- Then $(N_1, \ldots, N_n) \sim \mathcal{M}(k; p_1, \ldots, p_n), \quad p_i = \frac{1}{n}, \quad 1 \leq i \leq n$
- Let $A_i = \{N_i = j\}$

Answer: (inclusion-exclusion principle)

$$p_j = P \left( \bigcup_{i=1}^{n} A_i \right) = \sum_{\nu=1}^{n} (-1)^{\nu-1} \sum_{1 \leq i_1 < \ldots < i_\nu \leq n} P(A_{i_1} \cap \ldots \cap A_{i_\nu})$$
5.2.6 Inspection Probabilities for Growing $k$ and $n$ in POSICI

Co-inspection probabilities for $k = 5$ ($n = 5$ to 50)

- P(1)
- P(2)
- P(3)
- P(4)
- P(5)
5.2.7 Interpretation of POSICI

Example for $k = 9$:

- Observed partitions:
  - (1) 5+2+2
  - (2) 5+2+1+1

- The 5-times co-inspected document has the same probability to be a recommendation in both cases.

POSICI is built on the theory, that co-inspections other than $j$-times add more noise than information about the incentive to co-inspect the current document $j$-times. (The form of the histogram does not reveal additional information)
5.2.8 **POMICI: Probability of multiple items co-inspections**

**Question:** What is the probability $p_{part}$ that the partition corresponding to the complete histogram $H(D)$ of all co-inspections with $D$ occurs?

- Let $n+1$ be the number of total documents, $k$ the number of non-aggregated co-inspections (multiple inspections in different sessions are counted separately)

- Let $X$ be the set of words of length $k$ from an alphabet of $n$ letters, and $l_i$ the number of letters (i.e. documents), that occur exactly $i$-times in $x \in X$ (i.e. in $H(D)$)

- First examine the actions of the group $G = S_n \times S_k$ on the set $X$, and then the actions of the stabilizer subgroup $G_x$ on the set $S_n$ for the identity $id \in S_n$
**Answer:** (applying two times the orbit-stabilizer theorem together with Lagrange's theorem from group theory and some counting arguments)

\[
p_{\text{part}} = \frac{|G_x|}{|X|} = \frac{|G|}{|G_x|} = \frac{|G|}{|G_{x \text{id}}| |G_{x \text{id}}|} = \frac{n! \ k!}{(n - \sum_{i=1}^{k} l_i)! \ \prod_{j=1}^{k} l_j! (j!)^{l_j}}
\]
### 5.2.9 Partition Probabilities for Growing $k$ and $n$ in POMICI

#### Partition probabilities for $k = 7$ ($n = 7$ to 50)

![Partition probabilities graph for $k = 7$ ($n = 7$ to 50)]
5.2.10 Interpretation of POMICI

Example for $k = 9$:

- Observed partitions:
  - (1) 5+2+2
  - (2) 5+2+1+1

- The 5-times co-inspected document has not the same probability to be a recommendation in both cases.

- Partition (2) is more likely, so here the 5-times co-inspected document has a very slightly higher chance of being recommended.

POMICI is built on the theory, that the distribution of co-inspections other than $j$-times reveals more information than noise about the incentive to co-inspect the current document $j$-times.
6 Other applications (prototypes).
6.1 Restricted random walk clustering for indexing
6.2 Adaptive recommender systems

Recommendations for

Finding groups in data - an introduction to cluster analysis / Leonard Kaufman ; Peter J. Rousseeuw (1990)

- Few precise hits
- Many, but less precise hits

(75) Robust regression and outlier detection / Peter J. Rousseeuw ; Annick M. Leroy (1987)
### Recommendations for


<table>
<thead>
<tr>
<th>Few precise hits</th>
<th>Many, but less precise hits</th>
</tr>
</thead>
</table>

- (83) Clustering algorithms / John A. Hartigan (1975)
- (83) Clusteranalyse - anwendungsorientierte Einführung / Johann Bacher (1996)
- (80) Forecasting economic time series / Michael P. Clements and David F. Hendry (1998)
- (77) Data analysis - scientific modeling and practical application; with 45 tables / Wolfgang Gaul ... (eds.) (2000)
- (70) Cluster analysis for applications / Michael R. Anderberg (1973)
Finding groups in data - an introduction to cluster analysis / Leonard Kaufman ; Peter J. Rousseeuw (1990)

(75) Robust regression and outlier detection / Peter J. Rousseeuw ; Annick M. Leroy (1987)
(50) Clusteralyse - anwendungsorientierte Einführung / Johann Becker (1986)
(50) Bayeslösungen des Ausreißerproblems / Friedrich Gebhardt (1961)
(50) Ausreisser bei ein- und mehrdimensionalen Wahrscheinlichkeitsverteilungen / Rudolf Mathar (1981)
(33) Clustering algorithms / John A. Hartigan (1975)
(33) Data analysis - scientific modeling and practical application; with 46 tables / Wolfgang Gaul, ... (eds.) (2000)
(33) Modern regression methods / Thomas F. Ryan (1997)
(33) Untersuchung zur zeitlich-räumlichen Ähnlichkeit von phänologischen und klimatologischen Parametern in Westdeutschland u. v. Xiaobu Chen (1984)
(33) Cluster analysis / Brian Everitt (1974)
(33) Social Science Research Council / Social Science Research Council ()
(25) New approaches in classification and data analysis / E. Diday ... (eds.) (1984)
(25) Fallstudien Cluster-Analyse / Helmut Späth (1977)
(11) Mastering data mining - the art and science of customer relationship management / Michael J. A. Berry ; Gordon Linoff (2000)
(11) Kundenwert - Grundlagen - quantitative Konzepte - praktische Umsetzungen / Bernd Günter ... (Hrsg.) (2003)
(11) Customer Lifetime Value - Management - Kundenwert schaffen und erhöhen: Konzepte, Strategien, Praxisbeispiele / Markus Hofmann ... (Hrsg.) (2000)
(8) Scheduling theory / Tanaev, Vjaceslav S. (1994)
6.3 Exploration of the information space
7 System architecture and scalability.
7.1 An Agency of Software Agents for Distributed Recommender Services

A1: references:=get_references()

A2: info:=observe()

A3: update(info)

Lib: Library

B2: result:=process(request)

B1: request(query)

B3: display(result)

User: User

An Agency of Software Agents as Analysis Pattern
### 7.2 Recommender Services for Legacy Library Systems I

**The Architecture of a Library System with Recommender Services**
7.3 Recommender Services for Legacy Library Systems II

Sequence Diagram and Message Trace for a Detailed Document View
8 Service integration into web OPACS of scientific libraries.
8.1 Integration Variant I: apache-logs, cgi-scripts

• User Observation Agent:
  – http-logs with link embedded session IDs
  – preprocessing:
    ∗ extraction of GET requests and session splitting to take care of public access terminals

• Aggregation Agent:
  – information market: selecting information = purchase
  – computes market-baskets, estimates a logarithmic series distribution (LSD), and identifies and extracts outliers as recommendations
  – performs incremental updates periodically

• Recommendation Agent:
  – CGI-script on the recommendation server generates recommendation pages and is accessed via embedded links in the OPAC

Problems with maintenance and operations.
8.2 Integration Variants II and III: Web 2.0 Technology

• As a web-service:
  – The recommender server is an (internal) web-service. (Interface: SOAP, XML, WSDL)
  – Connection server may provide a web-service interface.

  **Advantage:** Customer has a full control on layout and layout change processes.

• With JavaScript:
  – A few lines of JavaScript are integrated at the customer’s OPAC.
  – The script is dynamically rewritten and controls the process shown in the sequence diagram.

  **Advantage:** Minimal integration overhead.
9 Development of a service concept and a service organization.

- Personal needs
- Personal values
- Previous experiences
- Company/local image
- Marketing communication
- Word of mouth
- Absence of communication
10 Development of a global service infrastructure for (scientific) libraries.

- Partners/alliances.
- Development of infrastructure.
- Migration to a distributed system.
10.1 Reference Customers

- Universität Karlsruhe (TH) (OPAC und KVK), operational since 2001, 2005.
- University of Freiburg, observed.
- Badische Landesbibliothek (a State library of Baden-Württemberg), observed.
- Technische Universität Berlin, observed.
- Universität Stuttgart, in preparation.
- Universität Hohenheim, in preparation.
Thanks for your attention!

**Project Links:**
http://reckvk.em.uni-karlsruhe.de/
http://www.ubka.uni-karlsruhe.de/catalog/
http://www.bibtip.org

**Contact:**
Andreas.Geyer-Schulz@kit.edu
Andreas.Neumann@iism.uni-karlsruhe.de

**Acknowledgement.**
We thank the DFG for the continued support and funding of the project Recommender Systems for Meta Library Catalogues.
References


Annual Conference of the German Classification Society. Springer. To appear.