Einsatz von Collaborative Filtering zur Datenprognose

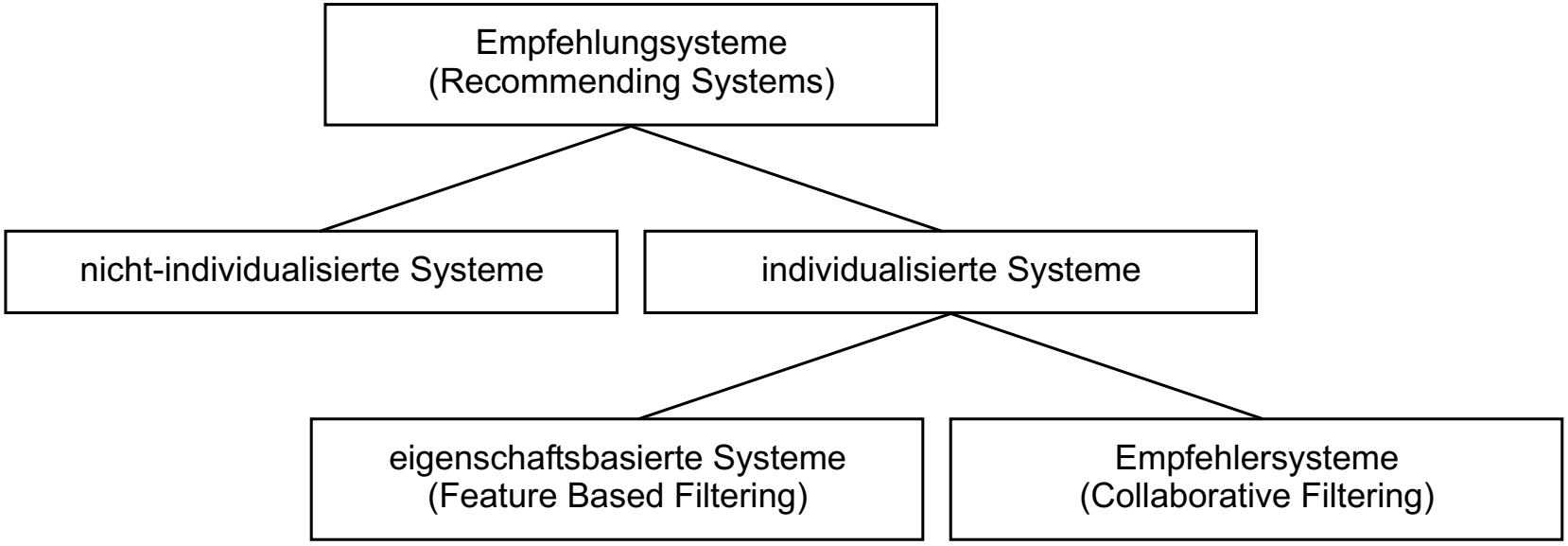
Fabian Bohnert
fabian.bohnert@mathematik.uni-ulm.de

Universität Ulm
Fakultät für Mathematik und Wirtschaftswissenschaften

Inhalt

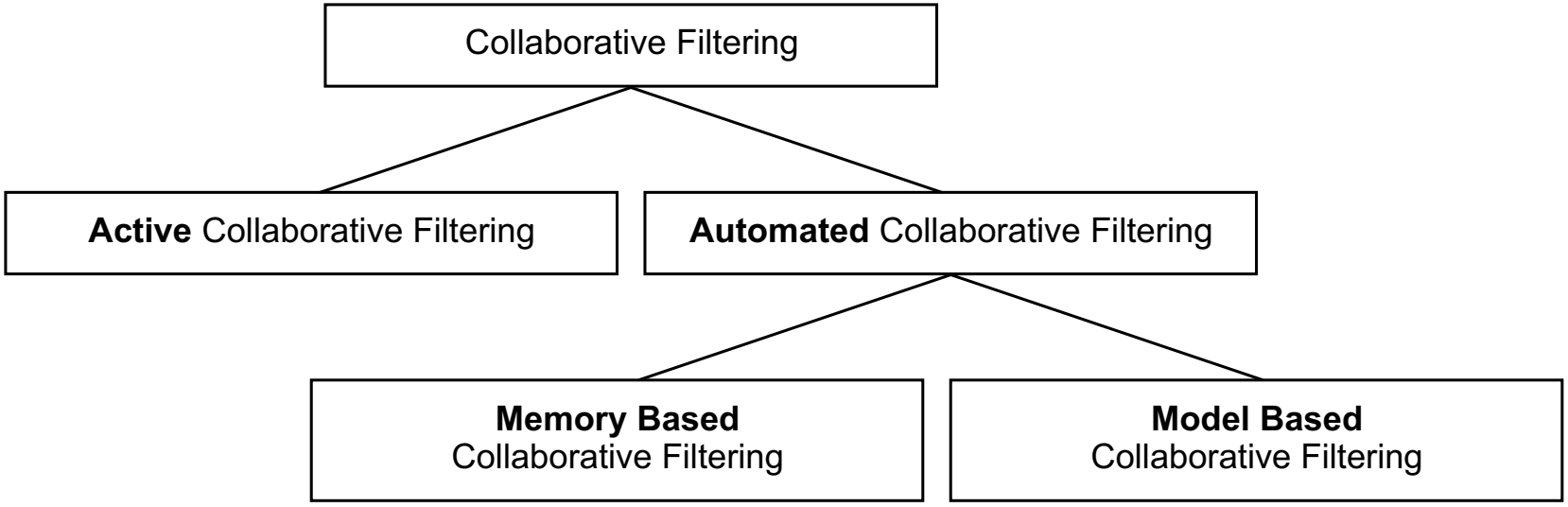
1. Einführung
2. Historie
3. **(Automated) Collaborative Filtering**
   - Grundidee
   - Speicherbasierte Algorithmen
   - Modellbasierte Algorithmen
   - Gütemaße
   - Vor- und Nachteile
   - Anwendungsgebiete
4. Beispiele
5. Datenschutz

Arten von Empfehlungssystemen



- **Empfehlungssysteme**
  Hauptaufgabe: Empfehlungen oder Prognosen für den Benutzer
- **individualisierte Empfehlungssysteme**
  individuelle Empfehlungen oder Prognosen basierend auf den persönlichen Präferenzen des Benutzers
- **eigenschaftsbasierte Systeme**
  Empfehlungen basierend auf Objekteigenschaften
- **Empfehlersysteme**
  Empfehlungen basierend auf ähnlichen Benutzern (Mentoren) durch CF

Active vs. Automated CF



- **Active Collaborative Filtering**
  aktive Empfehlungen durch Push-Kommunikation
- **Automated Collaborative Filtering**
  Pull-Kommunikation,
  Empfehlungen mit Hilfe eines mathematischen oder regelbasierten Verfahrens
- **speicherbasierten Algorithmen (Memory Based)**
  bei jeder Anfrage Berechnungen über die gesamte Datenmatrix
- **modellbasierten Algorithmen (Model Based)**
  offline Schätzung von Modellparametern auf Basis der Datenmatrix,
  online Empfehlungsabgabe auf Basis des Modells

Historie

- **US-Patente 4.870.579 und 4.999.642**
  Hey 1987, Hey 1989
- **Tapestry**, E-Mail-Filter-System
  Goldberg 1992
- **GroupLens**, Usenet-News-Empfehlersystem
  Resnick 1994
- **Musik-Empfehlersystem Ringo**
  Shardanand und Maes 1995
- **Bellcore's Video Recommender**
  Hill 1995
- Leitartikel einer Ausgabe der **Communications of the ACM**
  Resnick und Varian 1997

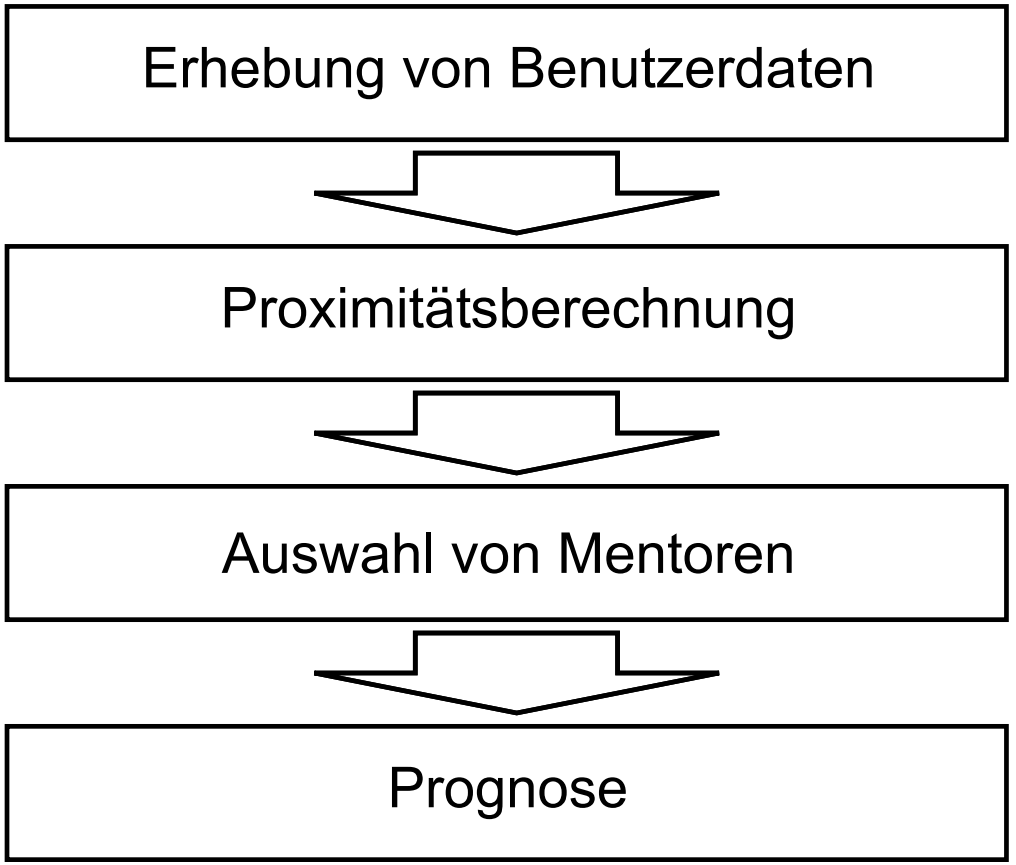
ACF – Grundidee (1/2)

- Ratingmatrix

$$U = (u_{ij})_{M,N} = \begin{bmatrix} u_{11} & \cdots & u_{1N} \\ \vdots & & \vdots \\ u_{i1} & \cdots & u_{iN} \\ \vdots & & \vdots \\ u_{M1} & \cdots & u_{MN} \end{bmatrix} \in \{0, \ldots, t, \cdot\}^{M \times N}$$

- $I = \{1, \ldots, M\}$ Menge der Benutzer
- $J = \{1, \ldots, N\}$ Menge der bewerteten Objekte
- eine Zeile u_i der Ratingmatrix U steht für ein Benutzerprofil

ACF – Grundidee (2/2)



Grundannahme:
Personen mit ähnlichem Profil treffen ähnliche Entscheidungen

ACF – speicherbasierte Ansätze (1/2)

- Prognosewert f_{aj} für den aktiven Benutzer a und das Objekt j (im Prinzip gewichteter Mittelwert der Bewertungen der zugehörigen Objektmentoren):

$$f_{aj} = \bar{u}_{a\cdot} + \frac{\sum_{i \in \tilde{M}_{aj}} s_{ai}\left(u_{ij} - \bar{u}_{i\cdot}\right)}{\sum_{i \in \tilde{M}_{aj}} s_{ai}} \text{ für } \tilde{M}_{aj} \neq \emptyset \text{ und } j \in J$$

- s_{ai} Ähnlichkeit der Benutzer a und i
- $\bar{u}_{a\cdot}$ arithmetisches Mittel aller von Benutzer a abgegebenen Bewertungen ($\bar{u}_{i\cdot}$ analog)
- M_a Menge der Mentoren
- $\tilde{M}_{aj} \subset M_a$ Menge der Objektmentoren für Objekt j

ACF – speicherbasierte Ansätze (2/2)

- **Pearsonscher Korrelationskoeffizient** (Ähnlichkeitsmaß für lineare Abhängigkeit):

$$q_{ai} = \frac{\sum(u_{aj} - \bar{u}_{ai})(u_{ij} - \bar{u}_{ia})}{\sqrt{\sum(u_{aj} - \bar{u}_{ai})^2 \sum(u_{ij} - \bar{u}_{ia})^2}} \text{ für } i \in M_a$$

  Summation über alle Objekte, für die sowohl u_{aj} als auch u_{ij} vorliegen

- $$s_{ai} = \begin{cases} q_{ai} & \text{für } q_{ai} > 0, \\ 0 & \text{für } q_{ai} \leq 0 \end{cases}$$

- **Alternative**: Berechnung der s_{ai} über Distanzmaß
- **Erweiterungen**: z.B. *Default Voting*, um Ratingdichte zu erhöhen

ACF – modellbasierte Ansätze (1/3)

- Prognosewert f_{aj} für den aktiven Benutzer a und das Objekt j (vor probabilistischem Hintergrund als Erwartungswert):

$$f_{aj} = \mathbb{E}(U_{aj}) = \sum_{i=0}^{t} i \cdot \mathbb{P}(U_{aj} = i \mid u_{ak},\, k \in J_a)$$

  $J_a \subset J$ Menge der Objekte, für die Benutzer a eine Bewertung abgegeben hat.

- $\mathbb{P}(U_{aj} = i \mid u_{ak},\, k \in J_a)$ Wahrscheinlichkeit, dass der aktive Benutzer a die Bewertung i abgibt, bei gegebenen Bewertungen für die Objekte aus J_a
- **Vorgehen**: offline Schätzung dieser Wahrscheinlichkeiten, damit die eigentliche Prognose online schnell erfolgen kann

ACF – modellbasierte Ansätze (2/3)

Cluster-Modelle

- **Idee**: bestimmte Gruppen oder Typen von Benutzern (Cluster) geben nahezu gleiche Bewertungen ab
- Annahme: Wahrscheinlichkeiten der Bewertungen bei gegebener Klassenzugehörigkeit sind unabh.

  $\mathbb{P}(C_a = c, u_{a1}, \ldots, u_{an}) = \mathbb{P}(C_a = c) \prod_{i=1}^{n} \mathbb{P}(u_{ai} \mid C_a = c)$
- $\mathbb{P}(U_{aj} = i \mid u_{ak}, k \in J_a) =$
  $\sum_c \mathbb{P}(U_{aj} = i \mid C_a = c) \cdot \mathbb{P}(C_a = c \mid u_{ak}, k \in J_a)$
- **Offline**: Schätzung von $\mathbb{P}(C_a = c)$ und $\mathbb{P}(u_{ai} \mid C_a = c)$
- **Online**: Schätzung von $\mathbb{P}(C_a = c \mid u_{ak}, k \in J_a)$ und Bestimmung des Erwartungswertes f_{aj}

ACF – modellbasierte Ansätze (3/3)

- Klassenzugehörigkeiten werden nicht explizit ermittelt $\rightarrow$ z.B. *K-Means-Clustering*, da die Klassenzugehörigkeiten hier versteckte Parameter sind

Andere Ansätze

- Bayessche Netze
- Hauptkomponentenanalyse
- Neuronale Netze

ACF – Gütemaße

- **Prognosegüte**
  mittlere Differenz zwischen vorhergesagten und tatsächlichen Bewertungen:

$$MAE = \frac{1}{|P|} \sum_{(i,j) \in P} |f_{ij} - u_{ij}|$$

$$MSE = \frac{1}{|P|} \sum_{(i,j) \in P} (f_{ij} - u_{ij})^2$$

  $P \subset I \times J$ Menge der Indexpaare (i, j), für die sowohl eine Prognose f_{ij} berechenbar ist als auch eine tatsächliche Bewertung u_{ij} existiert

- **Coverage**
  gibt an, für welchen Anteil der Objekte überhaupt Prognosen durch den Algorithmus erstellt werden können

ACF – Vorteile

- Aufdeckung von Beziehungen zwischen Benutzern und Objekten, die nicht mit objektiven Eigenschaften beschreibbar sind
- Erfahrungsaustausch zwischen einer hohen Anzahl von Benutzern, die sich nicht zwingend kennen müssen (*Virtual Community*)
- aufwendige Ermittlung (wenn überhaupt möglich) von Objekteigenschaften und das Führen einer zugehörigen Datenbank entfällt
- Empfehlung von Objekten, auch wenn nicht nach ihnen gesucht wurde
- Empfehlung von Objekten, die sich von bisherigen Präferenzen unterscheiden (ständige Neubildung der Mentorenmenge)

ACF – Nachteile

- gewisse Mindestanzahl von Benutzerprofilen ist notwendig
  (*Kaltstart- oder Bootstrapping-Problematik*)
- Objekteigenschaften auch wenn verfügbar nicht in die Prognose mit einbezogen
- Risiko schlechter Empfehlungen durch zufällige Zusammenhänge
- Prognosen haben *Black Box-Charakter*

Ausweg: Kombination mit Feature Based Filtering
(**Feature Guided Collaborative Filtering**)

ACF – Anwendungsgebiete

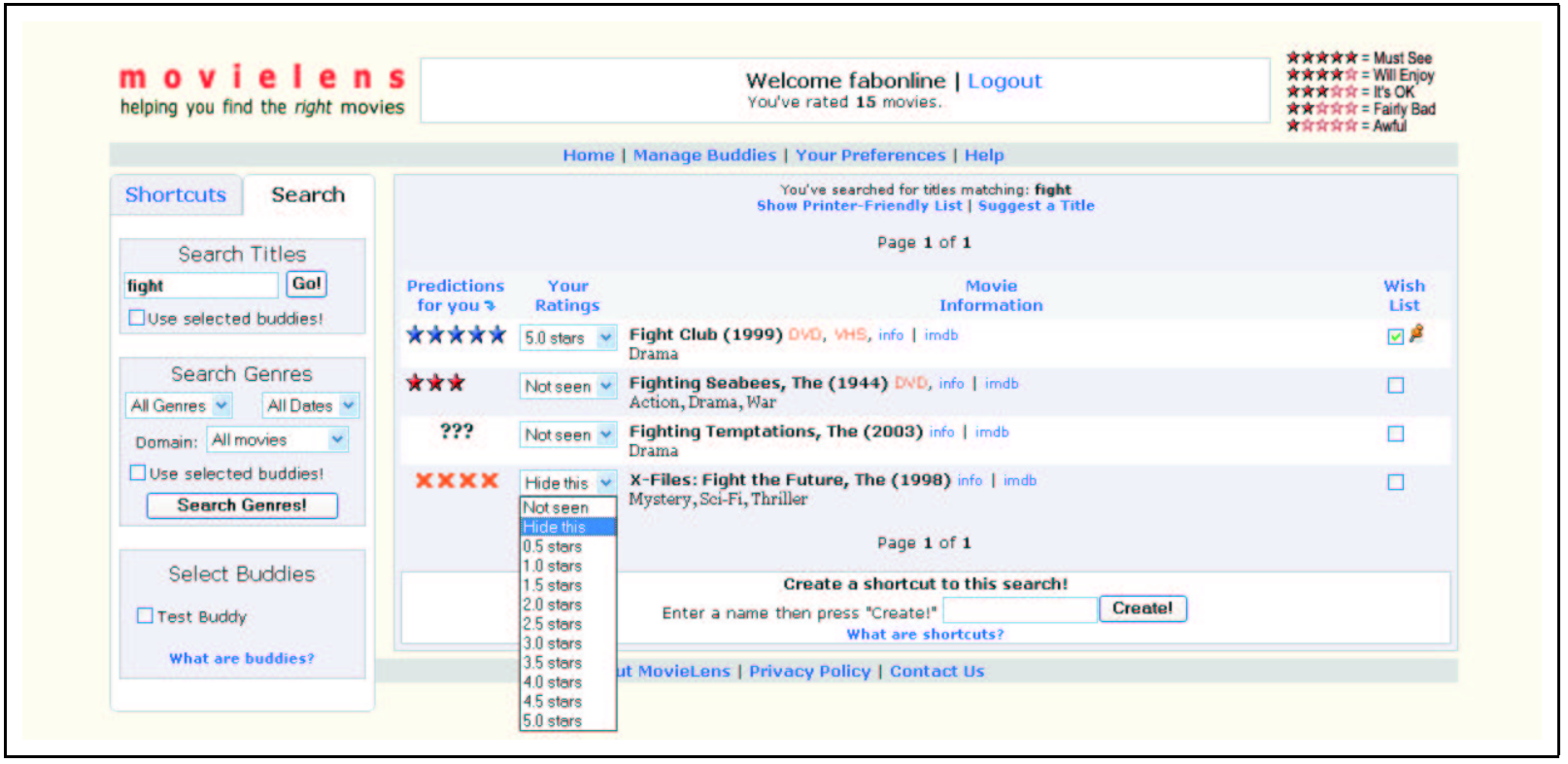
Voraussetzungen

- große Benutzer- und Objektzahl
- Objekte nicht anhand objektiver Eigenschaften oder besser über subjektive Präferenzen beschreibbar

Empfehlung von

- Literatur
- Musik
- Videos
- Filmen
- Webseiten
- Restaurants

Beispiel – MovieLens



- webbasiertes Empfehlersystem für Kinofilme
- Teil des GroupLens-Forschungsprojektes der University of Minnesota
- Filmempfehlungen für ganze Benutzergruppen möglich
- http://movielens.umn.edu

Weitere Beispiele

Amazon.de

- personalisierte Buchempfehlungen
- eingesetzte CF-Techniken von Net Perceptions bereitgestellt
- basiert auf Technologie des GroupLens-Projektes

Jester 2.0

- Witze-Empfehlersystem
- modellbasierte CF-Verfahren basierend auf Hauptkomponentenanalyse und Clustering
- stetige Ratingskala

Datenschutz

- **gesetzliche Vorschriften**
  Bundesdatenschutzgesetz (BDSG), Informations- und Kommunikationsdienste-Gesetz (IuKDG), hier insb. Teledienstdatenschutzgesetz (TDDSG)
- **Platform for Privacy Preferences (P3P)**
  vom WWW-Konsortium (W3C) entwickelter Vorschlag, um dem Benutzer von Webseiten mehr Kontrolle über die Nutzung seiner persönlichen Informationen zu geben

Vielen Dank für Ihre Aufmerksamkeit.

Literatur (1/3)

- Detlev Brechtel und Jochen Mai, *Gläserner König*. In Wirtschaftswoche Nr. 07, S. 103, 2000
- Jack Breese, David Heckerman und Carl Kadie, *Empirical Analysis Of Predictive Algorithms For Collaborative Filtering*. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 1998
- Craig Boutilier und Richard S. Zemel, *Online Queries For Collaborative Filtering*. In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, 2003
- David Chickering, David Heckerman und Christopher Meek, *A Bayesian Approach To Learning Bayesian Networks With Local Structure*. In Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence, 1997
- Danyel Fisher et al., *SWAMI: A Framework For Collaborative Filtering Algorithm Development And Evaluation*, 1999
- David Goldberg et al., *Using Collaborative Filtering To Weave An Information Tapestry*. In Communications of the ACM, Vol. 35, No. 12, S. 61-70, 1992
- Ken Goldberg et al., *Eigentaste: A Constant Time Collaborative Filtering Algorithm*, 2000
- Dhruv Gupta et al., *Jester 2.0: Evaluation Of A New Linear Time Collaborative Filtering Algorithm*. In Proceedings of the SIGIR, ACM, 1999

Literatur (2/3)

- Will Hill et al., *Recommending And Evaluating Choices In A Virtual Community Of Use*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, S. 194-201, 1995
- Patrick Joseph, *On-Line Advertising Goes One-On-One*. In Scientific American - Cyber View, 1997
- Sven Kauffelt, *Personalisierung im Web - Schlüssel zum Stammkunden*. In absatzwirtschaft Nr. 03, S.98, 2002
- Joseph Konstan, John Riedi und J. Ben Schafer, *Recommender Systems In E-Commerce*. In Proceedings of the 1st ACM Conference on Electronic Commerce, S. 158-166, 1999
- David Maltz und Kate Ehrlich, *Pointing The Way: Active Collaborative Filtering*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, S. 202-209, 1995
- Klaus Manhart, *Wege aus der Anonymität*. In Market Nr. 08, S. 58, 2001
- Stefanie Maute, *Konzeption und Realisierung eines Data Mining-Verfahrens für das Qualitätsmanagement in der Diabetologie*, Diplomarbeit an der Universität Ulm, 2002
- David Nichols, *Implicit Rating And Filtering*. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, S. 31-36, 1997

Literatur (3/3)

- David Pescovitz, *Accounting For Taste*. In Scientific American - Cyber View, 2000
- Paul Resnick et al., *GroupLens: An Open Architecture For Collaborative Filtering Of Netnews*. In Proceedings of the 1994 Computer Supported Collaborative Work Conference, S. 175-186, 1994
- Paul Resnick und H. R. Varian, *Recommender Systems*. In Communications of the ACM, Vol. 40, No. 3, S. 56-58, 1997
- Matthias Runte, *Personalisierung im Internet - Individualisierte Angebote mit Collaborative Filtering*, Deutscher Universitätsverlag, 2000
- Upendra Shardanand und Pattie Maes, *Social Information Filtering: Algorithms For Automating "Word Of Mouth"*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, S. 210-217, 1995
- Lyle H. Ungar und Dean P. Foster, *Clustering Methods For Collaborative Filtering*, AAAI Workshop on Recommendation Systems, 1998
- Übersicht über wichtige Papers zu Collaborative Filtering http://www.theether.org/collab/papers.html
- Collaborative Filtering - Eine Übersicht http://www.sims.berkeley.edu/resources/collab/