Einsatz von Collaborative Filtering zur Datenprognose

Seminararbeit im Rahmen des Einführungsseminar Data Mining im Wintersemester 2003/2004

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

Universität Ulm
Fakultät für Mathematik und Wirtschaftswissenschaften

Ulm, im Januar 2004

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Problemstellung

Im anbrechenden Informationszeitalter und mit der rasanten Verbreitung des Internet seit Anfang der neunziger Jahre sehen wir uns immer mehr einer Informationsflut ausgesetzt, die es zu bewerten und zu nutzen gilt.

Hierfür werden Systeme benötigt, die dem Informationssuchenden Empfehlungen dafür geben, welche Informationen für die Befriedigung seiner Informationsbedürfnisse geeignet sind [Sh95]. Gleichzeitig können diese Systeme von Unternehmen dazu verwendet werden, um ihre Kunden bei der Produktsuche zu unterstützen und so ihr Angebot zu personalisieren [Ko99].

Das Internet unterscheidet sich von vielen bekannten Medien. Bei herkömmlichen Massenmedien wie Radio, Fernsehen und Printmedien ist Kommunikation vorwiegend einseitig ausgerichtet. Eine kleine Gruppe erreicht eine große Anzahl von Nutzern oder Kunden. Es findet aber kein aktiver Dialog statt. Ein rückgekoppelter geschlossener Kommunikationskreislauf kommt nicht zustande. Man spricht auch von *One-To-Many-Kommunikation*.

Die Leistungsfähigkeit interaktiver Medien wie dem Internet geht über die Möglichkeiten traditioneller Medien hinaus. So bleibt die Kommunikation nicht auf eine Richtung beschränkt, sondern eine Rückkopplung wird möglich. Man spricht von *Many-To-Many-Kommunikation*. Durch die Nutzung des expliziten oder impliziten Feedbacks des Nutzers oder Kunden an den Anbieter wird es diesem möglich, Informationen und Angebote gezielt auszurichten. Es ergibt sich die Möglichkeit einer massenhaften Individualisierung des Angebots [Ru00]. Das Spektrum beschränkt sich hierbei nicht nur auf kommerzielle Anwendungen. Für einige Beispiele sei auf Abschnitt 3.6 auf Seite 14 und Kapitel 4 auf Seite 16 verwiesen.

Gesucht sind also Techniken, die auf dem vorhandenen Informationspool aufsetzen und diese Empfehlungen aussprechen bzw. die Individualisierung herbeiführen.

1.2 Zur Bezeichnung

Der Ausdruck *Collaborative Filtering* taucht in der Literatur erstmals in einem Aufsatz über das System Tapestry auf [Go92]. Zur weiteren Historie sei auf Abschnitt 2.3 verwiesen. Der Begriff kann wörtlich als *gemeinschaftliches Filtern* übersetzt werden[1].

In der Literatur tauchen neben dieser Bezeichnung auch die Bezeichnungen *(Social) Recommender System* (gesellschaftliches Empfehlungssystem) oder *Social Filtering* (gesellschaftliches Filtern) für Systeme auf, in denen Collaborative Filtering eingesetzt wird.

[1] vgl. http://dict.leo.org

Kapitel 2

Einordnung und Abgrenzung

2.1 Einbettung in Empfehlungssysteme

Unter *Empfehlungssystemen (Recommending Systems)* werden in der Literatur Systeme verstanden, deren Aufgabe vornehmlich darin besteht, Empfehlungen oder Prognosen für den Benutzer abzugeben. Man teilt die Empfehlungssysteme in zwei Hauptklassen auf, in nicht-individualisierte und individualisierte Systeme.

In *nicht-individualisierten Systemen* kommen vergleichsweise einfache Algorithmen zur Empfehlungsabgabe zum Einsatz. Die über solche Systeme abgegebenen Empfehlungen sind für jeden Benutzer identisch und nutzen die Möglichkeiten der interaktiven Medien nur beschränkt, da keine Informationsrückkopplung stattfindet.

Interessanter sind *individualisierte Empfehlungssysteme*, welche sich in eigenschaftsbasierte Systeme und Empfehlersysteme aufteilen lassen. Die besondere Leistung dieser Systeme besteht darin, individuelle Empfehlungen oder Prognosen basierend auf den persönlichen Präferenzen des Benutzers auszusprechen. Während in *eigenschaftsbasierten Systemen* die Empfehlungen auf den Eigenschaften der prinzipiell empfehlbaren Objekten beruhen, basiert die Empfehlung in *Empfehlersystemen (Recommender Systems)* auf einer Mehrzahl anderer Benutzer, welche auch als Mentoren oder Empfehler bezeichnet werden. In eigenschaftsbasierten Systemen greift man dabei auf *eigenschaftsbasiertes Filtern (Feature Based Filtering)* zurück, wohingegen in Empfehlersystemen *kollaboratives Filtern (Collaborative Filtering)* zum Einsatz kommt [Ru00].

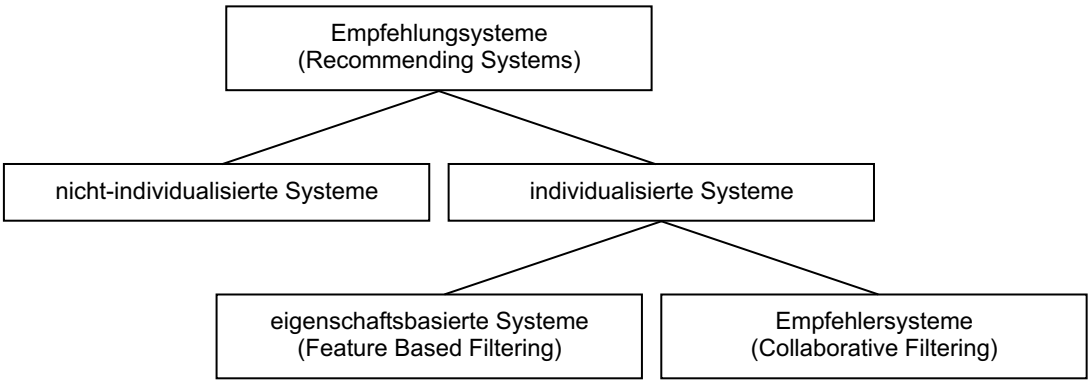

Abbildung 2.1: Arten von Empfehlungssystemen

Die Stärken des Collaborative Filtering liegen dabei oft dort, wo Feature Based Filtering nur eingeschränkt Anwendung findet. Oftmals ist es schwer oder gar unmöglich für Objekte quantifizierbare Eigenschaften zu definieren. So können eigenschaftsbasierte Techniken im Bereich der Unterhaltung (Kinofilme, Musik, Bücher) kaum angewendet werden, weil sich dort Objekte nicht anhand objektiver Eigenschaften beschreiben lassen [Sh95].

2.2 Active vs. Automated Collaborative Filtering

Man unterscheidet zwei Ansätze von Collaborative Filtering-Algorithmen. Unter dem Begriff *Active Collaborative Filtering*, den Ehrlich und Maltz 1995 eingeführt haben [Ma95], versteht man Techniken, die in Systemen zum Einsatz kommen, in denen sich Benutzer gegenseitig aktiv bestimmte Objekte empfehlen. Man spricht auch von *Push-Kommunikation*. Die wesentliche Voraussetzung ist dabei, dass der Empfehler den aktiven Benutzer persönlich kennt.

Im Gegensatz dazu fasst man Collaborative Filtering-Verfahren, die automatisiert Empfehlungen abgeben, unter dem Begriff *Automated Collaborative Filtering* zusammen. Hier holt sich ein Benutzer über eine Datenbank mit gespeicherten Benutzerprofilen eine Empfehlung (*Pull-Kommunikation*). Die Benutzer müssen sich also nicht notwendigerweise persönlich kennen. Eine manuelle Vorgabe von Filterregeln ist nicht mehr notwendig, weil die Ähnlichkeit zwischen Benutzern über ein mathematisches oder regelbasiertes Verfahren bestimmt wird [Ru00].

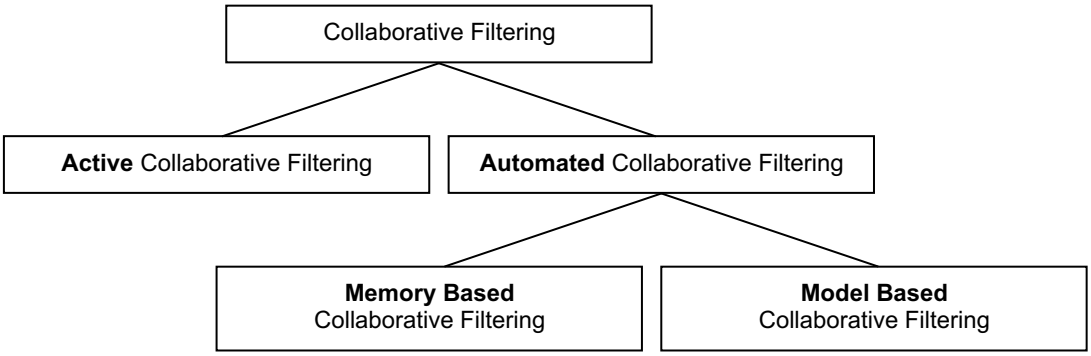

Abbildung 2.2: Active vs. Automated Collaborative Filtering

Die Verfahren, die unter dem Begriff *Automated Collaborative Filtering* gesammelt werden, lassen sich nach ihrem Prinzip in zwei Gruppen einteilen [Br98]: in die *speicherbasierten Algorithmen (Memory Based)* (vgl. Abschnitt 3.2), die bei jeder Anfrage Berechnungen über die gesamte Datenmatrix vornehmen und in die *modellbasierten Algorithmen (Model Based)* (vgl. Abschnitt 3.3), die die Datenmatrix nutzen, um offline die Parameter eines Modells zu schätzen. Auf dieses Modell wird anschließend online zurückgegriffen, um Empfehlungen oder Prognosen für einzelne Benutzer abzugeben, ohne dabei auf die gesamte Datenmatrix zugreifen zu müssen.

Der Vorteil der modellbasierten Ansätze liegt darin, dass die Schätzung der Modellparameter losgelöst von der eigentlichen Prognose einmalig vorab erfolgen kann, was in der Regel in einem deutlich geringen Rechenaufwand für die Prognose für individuelle Benutzer resultiert. Auf der anderen Seite kann durch die Informationsverdichtung im Rahmen der Modellierung ein Informationsverlust eintreten, was bei speicherbasierten Verfahren nicht der Fall ist [Ru00].

2.3 Historische Entwicklung

Die ältesten Quellen, in denen die Idee und der Algorithmus des Collaborative Filtering - obwohl nicht als solches benannt - recht detailliert beschrieben werden, sind die *US-Patente 4.870.579 und 4.999.642*. Diese wurden in den Jahren 1987 und 1989 von John Hey beantragt und 1989 bzw. 1991 bewilligt [Ru00]. Im ersten der beiden Patente werden Methoden beschrieben, die Reaktionen eines Benutzers auf Objekte vorhersagen, wobei sich diese Vorhersage auf Reaktionen von zu diesem Benutzer ähnlichen Benutzern stützt. In Patent 4.999.642 wird sogar explizit von einer Empfehlung gesprochen. Die beschriebenen Algorithmen beinhalten bereits die wesentlichen Elemente moderner Collaborative Filtering-Systeme.

Der Ausdruck *Collaborative Filtering* wird erstmals 1992 in einem Aufsatz über das System *Tapestry* verwendet [Go92]. Dieser Aufsatz wurde in der Literatur bisher als die erste und

ursprüngliche Quelle des Collaborative Filtering gewertet [Ru00]. Tapestry ist ein E-Mail-Filter-System, welches im Xerox Palo Alto Research Center eingesetzt wurde. Die Idee war, die Filterung nicht wie damals schon umgesetzt auf eigenschaftsbasierte Filterung zu beschränken, sondern durch die Einbeziehung menschlichen Urteilsvermögens effizienter zu machen. Bei diesem System ist es erforderlich, dass sich die einzelnen Benutzer kennen und einschätzen können, weil die Filterregeln individuell explizit vorgegeben werden müssen. Es gilt deshalb als ein aktives Empfehlersystem (vgl. Abschnitt 2.2).

Der nächste wichtige Artikel wird 1994 von Resnick und anderen veröffentlicht [Re94]. Er beschreibt das System GroupLens, welches mit Hilfe von Collaborative Filtering Usenet-Nachrichten empfehlen kann, um so der Informationsflut entgegenzutreten.

In den *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* erscheinen 1995 drei wichtige Veröffentlichungen über Collaborative Filtering. Im ersten Artikel von Ehrlich und Maltz werden die Begriffe *Active und Passive Collaborative Filtering* geprägt (vgl. Abschnitt 2.2) [Ma95]. Der zweite Artikel von Hill, Rosenstein und Furnas führt erstmals den Begriff *Virtual Community* im Zusammenhang mit Recommendersystemen ein [Hi95]. In diesem Aufsatz wird ein System beschrieben, dass Videofilme per E-Mail empfiehlt. Der dritte Artikel von Shardanand und Maes beschreibt das Recommendersystem Ringo, das Musik empfiehlt, und prägt den Begriff *Automating The Word Of Mouth* [Sh95].

Eine weitere wichtige Veröffentlichung stellt der Leitartikel einer Ausgabe der *Communications of the ACM* 1997 dar. In diesem geben Resnick und Varian eine Übersicht über die zu diesem Zeitpunkt existierenden Recommendersysteme [Re97]. Außerdem werden in dieser Ausgabe einige Recommendersysteme und die zu Grunde liegenden Collaborative Filtering-Techniken näher beschrieben.

Kapitel 3

(Automated) Collaborative Filtering

3.1 Grundidee des Verfahrens

Collaborative Filtering basiert auf der Annahme, dass sich Personen, die bei der Bewertung einer Anzahl von Objekten ähnlich entschieden haben, auch andere Objekte ähnlich bewerten, für die noch nicht alle Bewertungen vorliegen. So kann für eine bestimmte Person aus den Objektbewertungen der anderen Personen eine Prognose für die Präferenz dieses Objekts geschätzt werden. Da Produkteigenschaften dabei keine direkte Rolle spielen, kommt der Ansatz ohne Objektprofile aus.

Grundlegend ist es notwendig, dass man über eine große Anzahl von Personen (die Benutzer) verfügt. Diese Benutzer geben ihre individuellen Präferenzen für eine Anzahl von Objekten ab, was entweder explizit, z.B. über Bewertungsformulare, oder implizit, indem z.B. Kundenaktivitäten auf Webseiten aufgezeichnet werden, geschehen kann. Die Präferenzen werden in Benutzerprofilen gespeichert. Möchte sich nun ein bestimmter Benutzer (der aktive Benutzer) eine individuelle Empfehlung über interessante oder präferierte Objekte geben lassen, wird zunächst aus der hohen Anzahl unterschiedlicher Benutzer eine Reihe ähnlicher Benutzer ausgewählt. Die Bewertungen dieser Benutzer (im folgenden auch Mentoren oder Empfehler genannt) werden dann verwendet, um dem aktiven Benutzer eine individuell für ihn passende Empfehlung auszugeben [Ru00].

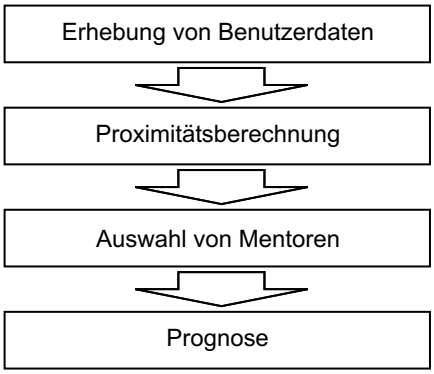

Abbildung 3.1: Typische Verfahrenselemente beim Collaborative Filtering

Um den Erfolg des Systems zu gewährleisten, sollte die Erhebung der Benutzerdaten des aktiven Benutzers nicht willkürlich erfolgen. Es bietet sich also eine explizite Befragung an, wobei folgende Ziele verfolgt werden sollten. Die Ratingvektoren sollten eine möglichst hohe Überlappung aufweisen, um eine Berechnung der Ähnlichkeiten zwischen Benutzern zu ermöglichen. Dazu sollten vorrangig Objekte mit der höchsten erwarteten Bekanntheit vorgelegt werden. Gleichzeitig sollte die Anzahl der Ratings für kontrovers bewertete Objekte maximiert werden, weil sich so Ähnlichkeiten und Unähnlichkeiten zwischen Benutzern klarer abzeichnen. Es sollten

also Objekte mit einer hohen Varianz in den Bewertungen vorgelegt werden. Drittens müssen auch unbekannte Objekte bewertet werden, damit für solche überhaupt Empfehlungen abgegeben werden können. Die Anzahl der zur Bewertung vorgelegten Objekte sollte begrenzt sein oder adaptiv erhöht werden, bis eine ausreichende Anzahl erreicht ist. Man erkennt, dass eine isolierte Verfolgung eines der Ziele nicht sinnvoll ist. In der Praxis wird oftmals eine Aufteilung zu 50% der bekanntesten Objekte und 50% zufällig ausgewählter Objekte vorgenommen [Ru00].

Sind ausreichend viele Daten gesammelt, können im nächsten Schritt die Ähnlichkeiten zu anderen Benutzern bestimmt werden. Dazu können prinzipiell dieselben Proximitätsmaße wie z.B. für die Clusteranalyse verwendet werden. Man benötigt aber in jedem Fall ein Ähnlichkeitsmaß. Distanzmaße müssen zunächst in Ähnlichkeitsmaße transformiert werden. Problematisch ist, dass die Benutzerprofile sehr oft eine große Anzahl von Lücken aufweisen, was die Ähnlichkeitsberechnung erschwert oder gar unmöglich macht.

Im nächsten Schritt wird eine Reihe besonders ähnlicher Benutzer ausgewählt, die Mentoren. Dafür kommen im Prinzip alle Benutzer in Frage, deren Ähnlichkeit zum aktiven Benutzer berechenbar ist, die mindestens ein Objekt bewertet haben, das der aktive Benutzer noch nicht bewertet hat und die eine gewisse Mindestanzahl von Überlappungen im Benutzerprofil mit dem aktiven Benutzer aufweisen. Anschließend müssen noch die Mentoren für ein konkretes Objekt bestimmt werden. Die Menge der Objektmentoren kann prinzipiell die Teilmenge der Mentoren sein, für die eine Bewertung für das betrachtete Objekt vorliegt.

Die anschließende Prognose fehlender Bewertungen im Benutzerprofil des aktiven Benutzers erfolgt dann durch eine gewichtete Summation der Bewertungen der zugehörigen Objektmentoren. Darin liegt eine wesentliche abgrenzende Eigenschaft des Verfahrens im Gegensatz zur Clusteranalyse. Dort wird versucht, eine Menge von Objekten in eine Anzahl möglichst homogener Gruppen einzuteilen, die untereinander möglichst unterschiedlich sein sollen. Hier steht jedoch ein bestimmter Benutzer im Fokus der Betrachtung [Ru00].

Im folgenden soll vereinfachend davon ausgegangen werden, dass alle Bewertungen explizit vorgenommen wurden und ganzzahlig skaliert vorliegen (beispielsweise $\{0, \ldots, t\}$, $t \in \mathbb{N}$). Außerdem soll es sich um globale Bewertungen handeln, was bedeutet, dass pro Benutzer und Objekt nur eine Bewertung aufgezeichnet wurde. Die Benutzerprofile werden in einer Ratingmatrix

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

gespeichert. Mit $\cdot$ soll dabei im folgenden ein fehlender Eintrag in der Ratingmatrix U gekennzeichnet werden.

Die Menge der Benutzer soll mit $I = \{1, \ldots, M\}$ und die Menge der bewerteten Objekte mit $J = \{1, \ldots, N\}$ bezeichnet werden. Eine Zeile u_i der Ratingmatrix U steht also für ein Benutzerprofil, in dem für Benutzer i die individuellen Objektbewertungen für die Objekte $1, \ldots, N$ gespeichert sind.

Der aktive Benutzer soll mit a bezeichnet werden, die Menge seiner Mentoren mit M_a und die Menge der Objektmentoren für das Objekt j mit $\tilde{M}_{aj}$. Weiterhin soll ein Prognosewert für den aktiven Benutzer a und das Objekt j die Bezeichnung f_{aj} tragen.

Aus technischen Gründen soll noch die Indikatormatrix

$$V = (v_{ij})_{M,N} \in \{0,1\}^{M \times N}, v_{ij} = \begin{cases} 0 & \text{für } u_{ij} = \cdot, \\ 1 & \text{für } u_{ij} \neq \cdot \end{cases}$$

eingeführt werden, wobei ein Eintrag in V genau dann auf 1 gesetzt ist, wenn U an entsprechender Stelle eine Bewertung beinhaltet.

3.2 Speicherbasierte Algorithmen (Memory Based)

Ein naiver Ansatz, einen Prognosewert für den aktiven Benutzer a und das Objekt j zu ermitteln, wäre, das arithmetische Mittel über alle verfügbaren Bewertungen zu bilden:

$$f_{aj} = \frac{\sum_{i \in I} v_{ij} u_{ij}}{\sum_{i \in I} v_{ij}} := \bar{u}_{\cdot j} \text{ für } j \in J$$

Dieser Ansatz liefert offensichtlich für ein Objekt j für alle aktiven Benutzer identische Werte. Es handelt sich also um ein nicht-individualisiertes Verfahren.

Ein Individualisierungsverfahren macht nur dann Sinn, wenn es bessere Ergebnisse liefern kann als dieser Ansatz; schon deshalb, weil es sich bei Individualisierungsverfahren um vergleichsweise rechenintensive Verfahren handelt. Collaborative Filtering-Verfahren setzen voraus, dass die Annahme der Unabhängigkeit der Benutzerratings über eine Mehrzahl von Objekten fallen gelassen wird. Diese Verfahren versuchen, sich eben diese Abhängigkeiten zunutze zu machen und auf diesem Wege zu besseren Prognosen zu gelangen [Ru00].

Beispielhaft sei

$$M_a = \{i \in I,\, i \neq a \colon \sum_{j \in J} v_{aj} v_{ij} \geq 3\}$$

die Menge der Mentoren des aktiven Benutzers a und

$$\tilde{M}_{aj} = \{i \in M_a \colon s_{ai} v_{ij} > 0\}$$

die Menge seiner Objektmentoren für das Objekt j. Dabei steht s_{ai} für die Ähnlichkeit zwischen dem aktiven Benutzer a und dem Benutzer i. In die Menge der Mentoren sollen also alle Benutzer aufgenommen werden, die mit dem aktiven Benutzer mindestens drei Überlappungen im Ratingvektor haben. Als Objektmentoren sollen alle Mentoren gelten, die das Objekt j bewertet haben und zum aktiven Benutzer eine positive Ähnlichkeit haben.

Nun kann der Prognosewert f_{aj} als gewichteter Mittelwert nach Shardanand, Maes und Runte folgendermaßen berechnet werden [Sh95], [Ru00]:

$$f_{aj} = \frac{\sum_{i \in \tilde{M}_{aj}} s_{ai} u_{ij}}{\sum_{i \in \tilde{M}_{aj}} s_{ai}} \text{ für } \tilde{M}_{aj} \neq \emptyset \text{ und } j \in J$$

Für $\tilde{M}_{aj} = \emptyset$ empfiehlt Runte, auf den naiven Ansatz $f_{aj} = \bar{u}_{\cdot j}$ auszuweichen [Ru00].

Resnick, Breese und andere schlagen einen leicht abgewandelten Ansatz vor [Re94], [Br98]:

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

$\bar{u}_{a\cdot} = \frac{\sum_{j \in J} v_{aj} u_{aj}}{\sum_{j \in J} v_{aj}}$ ist dabei das arithmetische Mittel aller von Benutzer a abgegebenen Bewertungen. $\bar{u}_{i\cdot}$ ist analog definiert.

Interessant ist, dass in diesem Ansatz berücksichtigt wird, dass verschiedene Benutzer Bewertungsskalen unterschiedlich interpretieren können. Dies dürfte in der Praxis zu besseren Prognosewerten führen.

Nun ist noch zu klären, wie die Ähnlichkeiten s_{ai} bestimmt werden. Dazu sollen folgende zwei Ansätze vorgestellt werden:

3.2.1 Distanzbasierte Ansätze

In distanzbasierten Ansätzen geschieht die Berechnung der Ähnlichkeiten s_{ai} über ein Distanzmaß. Hier soll beispielhaft die mittlere euklidische Distanz vorgestellt werden:

$$d_{ai} = \frac{\sqrt{\sum_{j \in J} v_{aj} v_{ij} (u_{aj} - u_{ij})^2}}{\sum_{j \in J} v_{aj} v_{ij}} \text{ für } i \in M_a$$

Dieser Distanzwert muss anschließend noch in einen Ähnlichkeitswert transformiert werden, was beispielsweise über

$$s_{ai} = 1 - \frac{d_{ai}}{t} \text{ mit } t = \max_{i \in M_a} \{d_{ai}\},\ s_{ai} \in [0, 1]$$

erfolgen kann.

3.2.2 Korrelationsbasierte Ansätze (Pearson)

Bei korrelationsbasierten Ansätzen werden die Ähnlichkeiten s_{ai} direkt über ein Ähnlichkeitsmaß bestimmt. Oftmals wird dazu der Pearsonsche Korrelationskoeffizient herangezogen. Er findet unter anderem in den Systemen GroupLens [Re94] und Ringo [Sh95] Verwendung:

$$q_{ai} = \frac{\sum_{j \in J} v_{aj} v_{ij} (u_{aj} - \bar{u}_{ai})(u_{ij} - \bar{u}_{ia})}{\sqrt{\sum_{j \in J} v_{aj} v_{ij} (u_{aj} - \bar{u}_{ai})^2 \sum_{j \in J} v_{aj} v_{ij} (u_{ij} - \bar{u}_{ia})^2}} \text{ für } i \in M_a \text{ und } Nenner > 0,$$

wobei

$$\bar{u}_{ai} = \frac{\sum_{j \in J} v_{aj} v_{ij} u_{aj}}{\sum_{j \in J} v_{aj} v_{ij}} \text{ für } i \in M_a$$

des arithmetische Mittel aller Bewertungen u_{aj} des Benutzers a ist, für die auch bei Benutzer i entsprechende Bewertungen vorliegen und $\bar{u}_{ia}$ analog definiert ist. Es gilt: $q_{ai} \in [-1, 1]$.

Über

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

stellt man sicher, dass nur positive Korrelationswerte in einem Ähnlichkeitswert größer Null resultieren und andere Korrelationswerte unberücksichtigt bleiben.

Es sollte erwähnt werden, dass eine Korrelationsberechnung lediglich die lineare Abhängigkeit der Bewertungen zweier Benutzer ausdrückt.

Zu bemerken ist noch, dass in obigen Ansätzen nur Ratingwerte einfließen, die sowohl beim aktivem Benutzer a wie auch beim Benutzer i vorliegen. Dies wird über die Verwendung der Indikatormatrix V erreicht. Es ist möglich, die Verfahren so zu modifizieren, dass auch Ratingwerte berücksichtigt werden können, die nur einseitig vorliegen (beispielsweise über *Default Voting* [Br98]). Damit kann künstlich eine höhere Ratingdichte erzeugt werden.

In [Br98] untersuchen Breese, Heckerman und Kadie weitere Methoden zur Bestimmung der Ähnlichkeiten s_{ai}: einen weiteren Basisalgorithmus (*Vector Similarity*) und zwei weitere Modifikationen der Standardalgorithmen (*Inverse User Frequency* und *Case Amplification*).

3.3 Modellbasierte Algorithmen (Model Based)

Die Grundidee modellbasierter Algorithmen ist, dass auf Basis der Bewertungen in der Datenmatrix die Parameter eines Modells offline geschätzt werden, so dass die online berechnete Prognose deutlich schneller erfolgen kann. Zur Schätzung der Parameter beschränkt man sich aus Effizienzgründen oftmals auf ein sogenanntes *Trainingsset*, eine Teilmatrix der Datenmatrix. Durch diese Informationsverdichtung im Rahmen der Modellbildung kann allerdings ein erheblicher Informationsverlust eintreten.

Vor einem probabilistischen Hintergrund kann der Prognosewert f_{aj} als Erwartungswert angesehen werden [Br98]:

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

wobei $J_a \subset J$ die Menge der Objekte sein soll, für die Benutzer a eine Bewertung abgegeben hat. $\mathbb{P}(U_{aj} = i \mid u_{ak},\, k \in J_a)$ ist also die Wahrscheinlichkeit, dass der aktive Benutzer a die Bewertung i abgibt, unter der Voraussetzung, dass seine Bewertungen für die Objekte aus J_a gegeben sind. Vereinfachend wird auch hier angenommen, dass die vorliegenden Bewertungen aus $\{0, \ldots, t\}$, $t \in \mathbb{N}$ sind.

3.3.1 Cluster-Modelle

Die Cluster-Modellen zu Grunde liegende Idee ist, dass bestimmte Gruppen oder Typen von Benutzern nahezu gleiche Bewertungen abgeben und man diese deshalb zu Clustern zusammenfassen kann. Das eigentliche Collaborative Filtering beschränkt sich dann auf eine probabilistische Zuordnung der Benutzer zu Clustern. Auf dieser Basis werden Wahrscheinlichkeiten dafür bestimmt, dass ein Benutzer ein Objekt präferiert.

Unter der Annahme, dass die Wahrscheinlichkeiten der Bewertungen bei gegebener Klassenzugehörigkeit unabhängig sind, kann man die gemeinsame Wahrscheinlichkeit, dass Klasse c und die Bewertungen $u_{a1}, \ldots, u_{an}$ vorliegen, nach Bayes folgendermaßen ausdrücken [Bo03]:

$$\mathbb{P}(C_a = c, u_{a1}, \ldots, u_{an}) = \mathbb{P}(C_a = c) \prod_{i=1}^{n} \mathbb{P}(u_{ai} \mid C_a = c)$$

Die Parameter des Modells, die Wahrscheinlichkeiten der Klassenzugehörigkeiten $\mathbb{P}(C_a = c)$ und die bedingten Wahrscheinlichkeiten $\mathbb{P}(u_{ai} \mid C_a = c)$, werden offline geschätzt. Online werden dann nur noch die Wahrscheinlichkeiten $\mathbb{P}(C_a = c \mid u_{ak},\, k \in J_a)$ für den aktiven Benutzer a bestimmt und anschließend genutzt, um die Wahrscheinlichkeiten $\mathbb{P}(U_{aj} = i \mid u_{ak},\, k \in J_a)$ zu schätzen, die zur Berechnung des Erwartungswertes $\mathbb{E}(U_{aj})$ benötigt werden [Bo03]:

$$\mathbb{P}(U_{aj} = i \mid u_{ak},\, k \in J_a) = \sum_{c} \mathbb{P}(U_{aj} = i \mid C_a = c) \cdot \mathbb{P}(C_a = c \mid u_{ak},\, k \in J_a)$$

Man sollte bemerken, dass nicht-beobachtete Bewertungen als *missing-at-random* angesehen werden. Dies trifft in vielen Fällen aber nicht zu, da die Tatsache, dass eine Bewertung fehlt, oftmals Information enthält [Bo03].

Da die Klassenzugehörigkeiten nicht explizit ermittelt werden, muss auf Techniken zurückgegriffen werden, bei denen die Klassenzugehörigkeiten versteckte Parameter sind. In der Literatur werden Algorithmen wie der sogenannte *EM-Algorithmus* (z.B. *K-Means-Clustering* und *Wiederholtes K-Means-Clustering*) oder *Gibbs Sampling* beschrieben [Un98]. Gibbs Sampling stellt dabei nach Ansicht von Ungar und Foster die bessere Alternative dar.

In [Fi99] schlagen Fisher und andere einen sogenannten *Clustered Pearson-Algorithmus* vor, der sich vom entsprechenden speicherbasierten Ansatz (vgl. Abschnitt 3.2.2) dadurch unterscheidet, dass offline Clusterverfahren wie *K-Means-Clustering* angewendet werden und online nicht mehr auf die eigentlichen Bewertungen in der Datenmatrix, sondern nur noch auf die ähnlichsten Cluster zurückgegriffen wird.

3.3.2 Bayessche Netze

Einen weiteren Ansatz, die benötigten bedingten Wahrscheinlichkeiten $\mathbb{P}(U_{aj} = i \mid u_{ak}, k \in J_a)$ zu bestimmen, stellen Bayessche Netze dar [Br98]. Über geeignete Lernalgorithmen werden Entscheidungsbäume für jedes Objekt generiert, mit denen man diese Wahrscheinlichkeiten berechnen kann. Die Knoten repräsentieren hierbei die Objekte, die Kanten zu untergeordneten Knoten Bewertungen für diese Objekte. Als Bayessches Netz bezeichnet man die Gesamtheit der Entscheidungsbäume für alle Objekte. Ist das Bayessche Netz erst einmal berechnet, so sind individuelle Prognosen mit sehr geringem Zeitaufwand durchführbar.

Zu bemerken ist, dass beim Aufbau der Entscheidungsbäume, Wahrscheinlichkeiten für bestimmte Ausprägungen auf einer Bewertungsskala ermittelt werden und deshalb keine stetigen Skalen verwendet werden können, sondern stets mit diskreten Skalen gearbeitet werden muss. Im Bedarfsfall kann eine stetige Skala auch auf diskrete Ausprägungen heruntertransformiert werden. Es sollte erwähnt werden, dass für Collaborative Filtering-Anwendungen, in denen statt einer binären eine mehrwertige Skala vorliegt, die Komplexität der Entscheidungsbäume und Lernalgorithmen stark ansteigt, was sich nachteilig auf die Rechenzeit des Lernalgorithmus auswirkt [Ch97]. Zu beachten ist dabei, dass in den meisten Fällen auch noch Ausprägungen für fehlende Ratings zu berücksichtigen sind.

Details des Verfahren werden beispielsweise in [Ch97] beschrieben.

3.3.3 Hauptkomponentenanalyse

Ein weiteres interessantes Verfahren basiert auf der sogenannten Hauptkomponentenanalyse. Es wird beispielsweise im Recommendersystem Jester 2.0 verwendet (vgl. Abschnitt 4.3) [Gu99]. Mithilfe der Hauptkomponentenanalyse werden hier die Präferenzen der Benutzer offline von zehn Globalurteilen auf zwei Eigenvektoren (Hauptachsen) verdichtet. Anhand der Eigenvektoren findet dann online die individuelle Präferenzprognose statt. Bei diesem Verfahren steht eine schnelle Empfehlungsabgabe im Vordergrund. Im Gegenteil zu anderen Verfahren, die eine diskrete Bewertungsskala voraussetzen, kann hier mit einer stetigen Skala gearbeitet werden.

Wesentlicher Bestandteil der Hauptkomponentenanalyse ist die *Hauptachsentransformation*. Zur Bestimmung der Hauptachsen benötigt man zunächst ein Trainingsset, in dem für alle Benutzer und alle Objekte die Bewertungen vorliegen und betrachtet im folgenden die k Objektprofile in diesem Trainingsset. Als erstes wird der Ursprung des Koordinatensystems in den Schwerpunkt des Trainingssets gesetzt. Dazu werden die Objektprofile normalisiert. Im zweiten Schritt wird das Koordinatensystem dann so gedreht, dass die erste Koordinate in Richtung der größten Varianz des Trainingssets zeigt. Damit ist die erste Hauptachse festgelegt und die Varianz in dieser Richtung ist die erste Hauptkomponente. Die nächste Drehung wird dann um diese Koordinatenachse durchgeführt, und zwar so, dass die zweite Hauptachse (die orthogonal zur ersten stehen muss) in Richtung der größten verbleibenden Varianz zeigt. Dieser Vorgang wird k-mal wiederholt, bis eine neue Basis geschaffen ist. Nach dieser Transformation ist die Varianz des Trainingssets so auf neue Koordinaten verteilt, dass die Varianz mit zunehmender Achsennummer abnimmt. Bei der weiteren Betrachtung beschränkt man sich dann auf die Hauptachsen, welche ausreichend Varianz ausdrücken.

Genaueres zum Verfahren kann in [Go00] nachgelesen werden.

3.3.4 Neuronale Netze

Möglicherweise sind im Kontext von Collaborative Filtering auch Modelle einsetzbar, die aus dem Bereich der neuronalen Netze stammen. Mit deren Hilfe kann man nicht-lineare funktionale Beziehungen modellieren, indem man versucht, die Funktionsweise des menschlichen Gehirns nachzubilden.

Ein neuronales Netz besteht aus sogenannten Eingangs- und Ausgangsneuronen, welche insgesamt die Werte eines Eingangs- bzw. Ausgangsvektor annehmen. Diese Neuronen werden über weitere Neuronen so verknüpft, dass über spezielle funktionale Zusammenhänge der Eingangsvektor in einen Ausgangsvektor transformiert wird. Neben der Wahl einer geeigneten Netzwerktopologie, also dem Aufbau des Netzes, ist vor allem der Lernalgorithmus von Bedeutung, mit dem die Beziehungen zwischen den Neuronen über eine geschickte Wahl von Verknüpfungsgewichten hergestellt werden. Man verwendet zur Bestimmung ein Trainingsset aus Input- und Outputvektoren. Abweichungen zwischen dem vom Netz ausgegebenen Outputvektor und dem tatsächlichen Outputvektor des Trainingssets resultieren in bestimmten Änderungen der Gewichte. Der Lernalgorithmus läuft so lange, bis das Netz mit hinreichender Genauigkeit den funktionalen Zusammenhang zwischen Ein- und Ausgangsvektoren wiedergibt. Danach ist das Netz im Idealfall in der Lage, für alle möglichen Eingangsvektoren die passenden Ausgangsvektoren zu ermitteln. Ob neuronale Netze gute Ergebnisse im Bereich des Collaborative Filtering erbringen, ist bislang noch nicht erforscht [Ru00].

3.4 Prognosegüte der Algorithmen

Zur Untersuchung der Leistungsfähigkeit der betrachteten Algorithmen kann auf eine Reihe von Gütemaßen zurückgegriffen werden. Einige sollen im folgenden vorgestellt werden.

3.4.1 Prognosefehler

Zur Messung der Prognosegenauigkeit eignen sich insbesondere Maße, in die die Differenz zwischen vorhergesagten Bewertungen und tatsächlichen Bewertungen einfließt.

Hierzu bieten sich der *mittlere absolute Prognosefehler (Mean Absolute Error, MAE)* und der *mittlere quadrierte Prognosefehler (Mean Square Error, MSE)* an. Der quadrierte Prognosefehler unterscheidet sich dabei vom absoluten Prognosefehler dadurch, dass durch die Quadrierung größere Prognoseabweichungen stärker gewichtet werden:

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

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

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

Wie auch schon in Abschnitt 3.2 angedeutet, macht ein Individualisierungsverfahren nur dann Sinn, wenn es bessere Ergebnisse liefern kann als der naive Ansatz, einen Prognosewert f_{ij} für den Benutzer i und das Objekt j als das arithmetische Mittel über alle für das Objekt j verfügbaren Bewertungen zu ermitteln [Sh95], [Fi99]. Dies ist dann der Fall, wenn der mittlere Prognosefehler kleiner ausfällt als der analog bestimme Prognosefehler für den naiven Ansatz.

3.4.2 Anteil berechenbarer Prognosen (Coverage)

Ein weiteres Gütemaß ist die *Coverage*, die angibt, für welchen Anteil der Objekte überhaupt Prognosen durch den Algorithmus erstellt werden können. Die Coverage liegt bei Verfahren, die vor allem in der Startphase auf eine Berechnung der Prognosen über den naiven Ansatz ausweichen, nahe bei 1.

Alternativ könnte beispielsweise auch der *Anteil besserer Prognosen* gemessen werden, der den Anteil an der Gesamtzahl der Prognosen angibt, für den der Algorithmus bessere Prognosen liefert als der naive Ansatz.

3.4.3 Klassifikationsgüte (ROC-Curve)

Eine weitere Möglichkeit, eine Aussage über die Güte eines Collaborative Filtering-Verfahrens zu treffen, ist, die *Klassifikationsgüte oder Trennschärfe* des Algorithmus zu bestimmen. Sie ist ein Maß dafür, wie oft durch das Verfahren für den Benutzer tatsächlich relevante Objekte präsentiert und für den Benutzer tatsächlich irrelevante Objekte nicht präsentiert werden.

Dazu bestimmt man zunächst *Sensitivität* und *Spezifizität* des Filters. Die Sensitivität gibt die Wahrscheinlichkeit an, mit der ein relevantes Objekt präsentiert wird. Die Spezifizität misst die Wahrscheinlichkeit, mit der ein irrelevantes Objekt gefiltert (also nicht präsentiert) wird. Die Einteilung in für den Benutzer relevante und irrelevante Objekte muss manuell vorgenommen werden.

Trägt man nun den funktionalen Zusammenhang von Sensitivität und Spezifizität in einem Diagramm auf, indem man den Schwellenwert des Filters stückweise erhöht, erhält man die sogenannte *Receiver Operational Characteristic Curve (ROC-Curve)*. Diese könnte beispielsweise so aussehen:

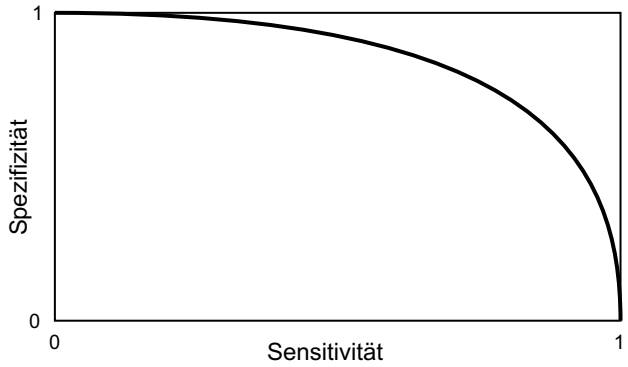

Abbildung 3.2: Beispiel für eine ROC-Kurve

Eine Spezifizität von 1 und eine Sensitivität von 0 bedeutet, dass alle irrelevanten Objekte herausgefiltert werden, aber auch kein einziges relevantes Objekt akzeptiert wird. Andersherum bedeutet eine Sensitivität von 1 und eine Spezifizität von 0, dass alle relevanten Objekte akzeptiert werden, aber gleichzeitig auch alle irrelevanten präsentiert werden.

Die Trennschärfe des Filters misst man, indem man die Fläche unterhalb der Kurve berechnet. Für einen Zufallsfilter beträgt die Trennschärfe 0,5. Falls der betrachtete Algorithmus einen größeren Wert für die Trennschärfe liefert als der naive Ansatz, so liefert er eine bessere Filterleistung. An der Objektivität der Trennschärfenmessung kann gezweifelt werden. Zu berücksichtigen ist, dass die Trennschärfe über eine subjektive Komponente verfügt, was sich aus der Einteilung in relevante und irrelevante Objekte ergibt [Ru00].

3.5 Vor- und Nachteile

Collaborative Filtering-Techniken setzen grundlegend eine große Anzahl von Benutzern und ein vielfältiges Angebot an Objekten voraus und können dann eingesetzt werden, wenn die Präferenzen der Benutzer subjektiv geprägt sind.

Es können Beziehungen zwischen Benutzern und Objekten aufgedeckt werden, die nicht mit objektiven Eigenschaften der betrachteten Objekte beschreibbar sind. Die Ansätze ermöglichen einen Erfahrungsaustausch zwischen einer hohen Anzahl von Benutzern, die sich dabei nicht zwingend persönlich kennen müssen (*Virtual Community*). Eine aufwendige Ermittlung (wenn überhaupt möglich) von Objekteigenschaften und das Führen einer zugehörigen Datenbank entfällt. Es werden keine Eigenschaftsprofile der Objekte benötigt, stattdessen werden Präferenzen in Benutzerprofilen gespeichert. Auch Prognosen in produktspartenübergreifenden Gebieten sind so möglich. Eine wesentliche Stärke der Techniken liegt darin, dass Objekte empfohlen werden, die nicht anhand von Eigenschaften gefunden worden wären und die sich von bisherigen Präferenzen unterscheiden, da in den Empfehlungsprozess durch eine ständige Neubildung der Mentorenmenge veränderte Interessen einfließen. Objekte werden empfohlen, auch wenn nicht nach ihnen gesucht wurde. Je länger das System besteht und je größer die Menge der Benutzer ist, desto treffsicherer werden Empfehlungen, da die Wahrscheinlichkeit ähnliche Benutzer zu finden steigt.

Dies erweist sich gleichzeitig aber als problematisch. Es ist nämlich grundlegend notwendig, über ein Minimum an Benutzerprofilen zu verfügen, um sinnvolle Empfehlungen abgeben zu können. Man spricht auch von der *Kaltstart- oder Bootstrapping-Problematik*. Ferner ist für neue Objekte keine Prognose möglich, da für sie zunächst ein Mindestanzahl von Bewertungen abgegeben werden muss. Außerdem werden Objekteigenschaften auch dann nicht in die Prognose mit einbezogen, wenn sie verfügbar oder gar relevant sind. Nicht zu unterschätzen ist auch das Risiko schlechter Empfehlungen durch zufällige Zusammenhänge, da bestimmte übereinstimmende Präferenzen nicht zwingend auch eine Übereinstimmung hinsichtlich anderer Kriterien bedeuten. Die erzeugten Prognosen haben *Black Box-Charakter* [Ru00], was bedeutet, dass der Empfehlungsprozess nicht besonders transparent ist.

Man erkennt, dass die Stärken des Collaborative Filtering oftmals dort liegen, wo Feature Based Filtering Schwächen hat und umgekehrt. Durch Kombination von Collaborative Filtering und Feature Based Filtering versucht man, die jeweiligen Nachteile auszugleichen. Man spricht dann von *Feature Guided Collaborative Filtering* oder auch von *Content Based Collaborative Filtering*.

Auf Basis der Objekteigenschaften nimmt man hierbei eine Vorauswahl der in Frage kommenden Objekte vor und ermittelt die persönlichen Präferenzen anschließend über Collaborative Filtering. Dabei können neben den vorliegenden Bewertungen für Objekte auch die individuellen Bedeutungsgewichte der Objekteigenschaften (man spricht auch von Partialpräferenzen) in die Prognose mit einfließen.

3.6 Typische Einsatzgebiete

Aus den betrachteten Vor- und Nachteilen lassen sich einige typische Anwendungsgebiete ableiten. Überall dort, wo auf eine große Benutzerzahl zurückgegriffen werden kann, wo Objekte nicht anhand objektiver Eigenschaften beschreibbar sind oder deren Erhebung zu aufwendig ist und wo Objekte besser über subjektive Präferenzen beschrieben werden, kann Collaborative Filtering zum Einsatz kommen.

Collaborative Filtering bietet sich beispielsweise zur Empfehlung von Literatur, Musik, Videos, Filmen, Webseiten oder Restaurants an. In diesen Bereichen des One-To-One-Marketing liegen Anwendungsgebiete des Collaborative Filtering im Rahmen des E-Commerce. Effizientes

Marketing wird möglich, weil individuell auf den Benutzer zugeschnittene Kaufempfehlungen abgegeben werden können und sich so die Wahrscheinlichkeit des Kundeninteresses am angebotenen Produkt steigert. Außerdem wird die Kundenbeziehung vertieft und die Kundenloyalität erhöht.

Anwendungsgebiete liegen aber durchaus auch in anderen Bereichen. Stefanie Maute hat beispielsweise im Rahmen einer Diplomarbeit versucht, Collaborative Filtering-Techniken für eine Prognose von Blutzuckerwerten, genauer SDS-Werten, im Bereich der Diabetologie, Medizin heranzuziehen. Idee ist dabei, dass der Blutzuckerwert in hohem Maße genetisch bedingt ist und nicht von biologischen Informationen wie Gewicht, Ernährung oder Lebensgewohnheiten abhängt [Ma02].

Für einige Beispiele sei auf Kapitel 4 verwiesen.

Kapitel 4

Beispiele für aktuelle Anwendungen

4.1 Amazon.de - http://www.amazon.de/

Mit seiner Webseite gilt der Online-Buchhändler Amazon.de als Vorreiter für den Einsatz des Collaborative Filtering zur Generierung personalisierter Angebote. Schon beim Betreten der Homepage wird der Kunde nach seiner Identifikation (per Cookie oder Login) persönlich begrüßt. Voraussetzung dafür ist, dass er sich vorab einmalig explizit identifiziert hat. Der Kunde erhält dann Buchvorschläge, die auf Basis des Collaborative Filtering aus bisherigen Informationen wie Einkaufsverhalten oder expliziten Produktbewertungen abgeleitet werden. Darüberhinaus besteht für den Benutzer die Möglichkeit, sein Profil selbst zu konfigurieren, indem er explizite Produkbewertungen abgibt. Das ist beispielsweise dann sinnvoll, wenn keine passenden Vorschläge generiert werden können. Auch bei der Suche nach einem bestimmten Buch werden dem Kunden weitere Empfehlungen unterbreitet, indem das Buch mit Käufen und Bewertungen anderer Kunden abgeglichen wird. Die Bestellung und der Warenausgang werden durch eine personalisierte E-Mail bestätigt. Zusätzlich kann der Kunde auf der Webseite den Fortschritt seiner Bestellung nachvollziehen. Empfehlungen kann er auch per Newsletter anfordern.

Die Collaborative Filtering-Techniken, die bei Amazon.de zum Einsatz kommen, werden von Net Perceptions (vgl. Abschnitt 4.4) bereitgestellt und basieren auf den Verfahren, die im Rahmen des GroupLens-Projektes [Re94] entwickelt wurden.

Abbildung 4.1: Amazon.de über vorgelegte persönliche Empfehlungen

4.2 MovieLens - http://movielens.umn.edu

movielens
helping you find the *right* movies

MovieLens ist ein webbasiertes Empfehlersystem für Kinofilme. Es ist Teil des GroupLens-Forschungsprojektes der University of Minnesota und setzt auf den Collaborative Filtering-Techniken auf, die im Rahmen des GroupLens-Projektes [Re94] 1994 entwickelt wurden.

Nach der Registrierung werden dem neuen Benutzer Filme vorgelegt, bis er 15 Filme bewertet hat. Anschließend ist das System in der Lage, individualisierte Empfehlungen auszusprechen. Insgesamt kann dabei auf mehrere Millionen abgegebener Bewertungen zurückgegriffen werden. Der Benutzer kann die Anzahl seiner bewerteten Filme erhöhen und damit die Prognosegenauigkeit erhöhen, indem er neue Ratings abgibt. Gleichzeitig kann er schon abgegebene Ratings nachträglich ändern. In einer Wunschliste kann der Benutzer Filme sammeln, die er besonders gerne mag. Über eine umfangreiche Suchfunktion kann der Benutzer auch direkt nach bestimmten Filmen suchen und das für ihn prognostizierte Rating einsehen. Fehlt ein Film in der Datenbank, ist es möglich, diesen zur Aufnahme vorzuschlagen. Außerdem bietet das System die Möglichkeit, eine Buddy-Liste zu führen und kann Filmempfehlungen für ganze Benutzergruppen abgeben, beispielsweise für einen gemeinsamen Videoabend.

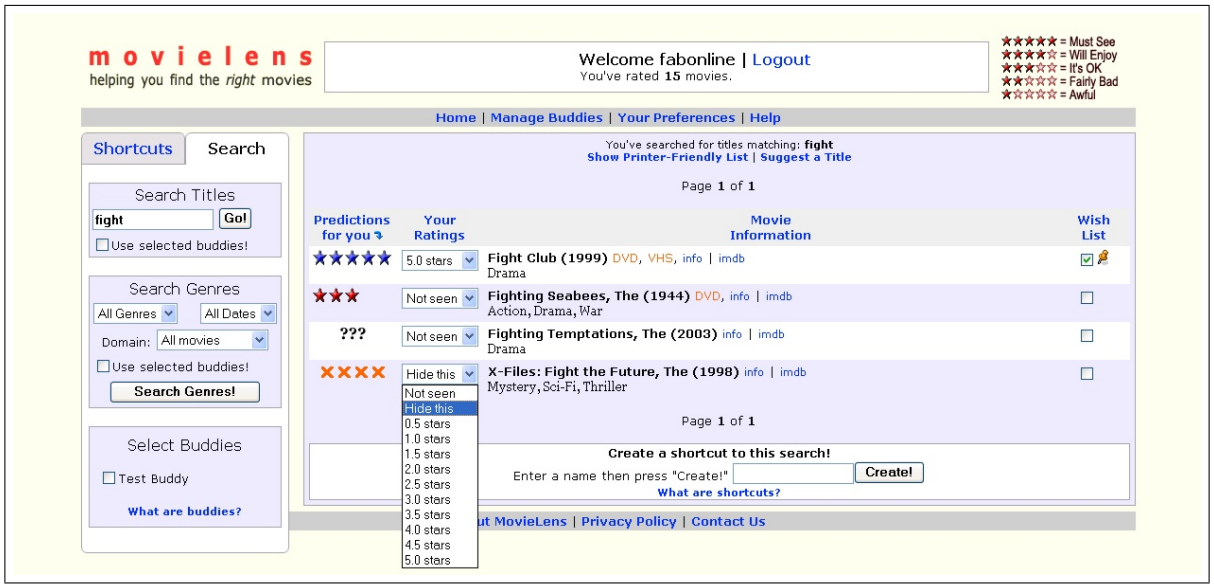

Abbildung 4.2: MovieLens - Suchergebnis für “fight” (Ratings in blau, Prognosewerte in rot)

4.3 Jester 2.0 - http://eigentaste.berkeley.edu/

Auch das Recommendersystem Jester 2.0 ist im Rahmen von Forschungsarbeiten [Gu99], [Go00] entstanden. Es wurde an der University of California Berkeley entwickelt und ist in der Lage Witze vorzulegen, die dem persönlichen Geschmack entsprechen. Zuvor muss sich der Benutzer registriert und ein Trainingsset von zehn festgelegten Witzen und fünf zufällig ausgewählten Witzen auf einer stetigen Skala von -10 bis $+10$ bewertet haben. Anschließend legt das System personalisierte Witze vor und sammelt gleichzeitig weitere Ratings. Das System kann derzeit dabei auf mehr als vier Millionen abgegebene Ratings von ungefähr 75000 Benutzern für 100 Witze zurückgreifen.

Der zur Empfehlungsabgabe eingesetzte Algorithmus, den die Autoren mit *Eigentaste* bezeichnen, ist den modellbasierten Verfahren des Collaborative Filtering zuzu-

ordnen (vgl. Abschnitt 3.3). Idee ist, dass eine auf dem gesamten zur Verfügung stehenden Datenmaterial online berechnete Prognose nicht effizient genug ist. Deshalb setzen die Entwickler von Jester auf eine der Prognose offline vorgeschaltete Hauptkomponentenanalyse und Clustering. So kann erreicht werden, dass die eigentliche Empfehlung mit konstantem Aufwand erfolgen kann [Go00].

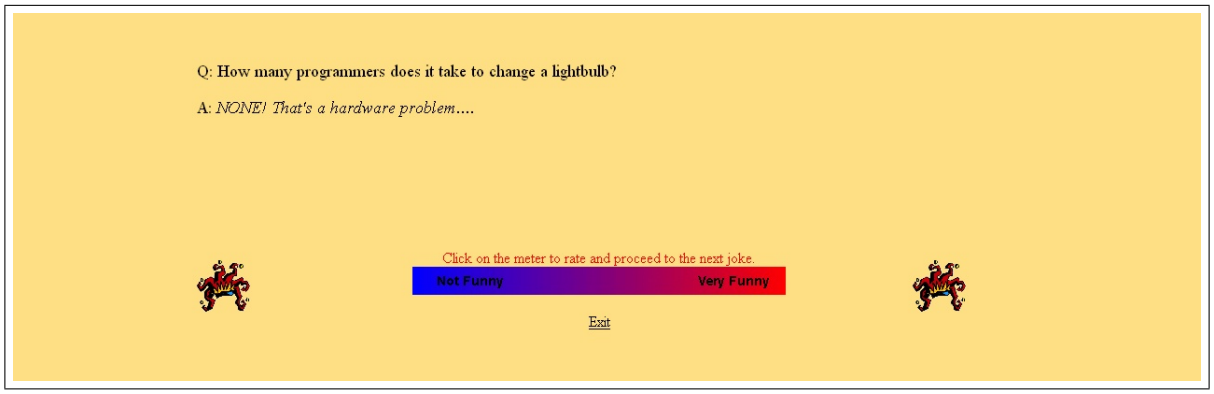

Abbildung 4.3: Jester 2.0 - Screenshot

4.4 Kommerzielle Anbieter

Hier sollen kurz stichwortartig einige kommerzielle Anbieter von Recommendersystemen genannt werden:

- **Net Perceptions (http://www.netperceptions.com/)**
  Die eingesetzte Collaborative Filtering-Techniken basieren auf der im Rahmen des GroupLens-Projektes [Re94] entwickelten Technologie. Beispielsweise kommt die Software derzeit bei Amazon.de, E!Online im MovieFinder und auf Half.com zum Einsatz.

- **Macromedia (http://www.macromedia.com/)**
  Der *Macromedia LikeMinds Preference Server* bedient sich Collaborative Filtering, um auf Webseiten Kundenverhalten zu analysieren und individualisierte Empfehlungen auszusprechen. IBM hat LikeMinds in seine *WebSphere-Produktreihe* integriert[1]. LikeMinds findet beispielsweise auf CineMax.com und Levis.com Anwendung.

- **Firefly Network**
  Ein der größten Unternehmen, die auf Collaborative Filtering setzten, war Firefly Network. Es wurde im März 1995 gegründet. Schon wenige Monate nach der Gründung startete es den *Online-Service Firefly*. Dabei handelte es sich um ein Verkauf- oder Vertriebssystem, das auf Basis intelligenter Agent-Software arbeitete, und im Media Lab des Massachusetts Institute of Technology entwickelt wurde. Idee war, Benutzerdaten zentral abzulegen und diese bei Erlaubnis durch den Benutzer Unternehmen zur Individualisierung ihres Angebots bereitzustellen. Das Unternehmen wurde 1998 von Microsoft übernommen[2].

- **WiseWire**
  Auch das 1995 gegründete Unternehmen WiseWire war einer der Marktführer auf dem Gebiet der Webseitenpersonalisierung durch Collaborative Filtering. WiseWire wurde 1998 von Lycos übernommen[3].

[1] vgl. http://www-5.ibm.com/de/versicherungen/personal_web.html, http://www-306.ibm.com/software/genservers/commerce/community/partners/macromedia.html
[2] vgl. http://www.microsoft.com/presspass/press/1998/apr98/freflypr.asp
[3] vgl. http://www.lycos.com/press/wirebuss.html

Kapitel 5

Datenschutzproblematik: Der gläserne Kunde

Kunden individuell ansprechen und informieren, auf ihre Wünsche und Bedürfnisse eingehen und ein entsprechendes Leistungsangebot bereitstellen – für immer mehr E-Commerce-Betreiber ist Personalisierung ein unverzichtbarer Baustein eines erfolgreichen E-Business. Die meisten der führenden Online-Händler haben damit begonnen, ihr Angebot nutzerspezifisch auszurichten [Ma01].

Die Gründe liegen auf der Hand: Vorlieben und Abneigungen der Konsumenten zu erkennen und umzusetzen, ist ein wichtiger Bestandteil jeder Unternehmensstrategie. Bei einer personalisierten E-Business-Seite besteht eine höhere Wahrscheinlichkeit, dass Besucher aufmerksam werden, ihr Interesse aufrechterhalten und dadurch letztlich mehr Umsatz erzielt wird. Der Anbieter kann zielgerichtet Produkte vermarkten, personalisierte Nachrichten bereitstellen, Dokumente empfehlen oder Ratschläge erteilen. Auch die dargestellte Werbung kann individualisiert werden. Dies verbessert nicht nur die Kaufwahrscheinlichkeit, sondern die Werbung wird als weniger lästig empfunden, da sie fast schon Informationscharakter hat. Die Kunden sind zufriedener, wenn sie Informationen schneller finden und der Service an ihre Anforderungen angepasst ist.

Aus Unternehmersicht scheint die Richtung vorgegeben zu sein. Doch wie sieht es auf der anderen Seite aus? Ist der Kunde tatsächlich bereit, so viel an Information über ihn preiszugeben? Hat er überhaupt die Möglichkeit, sich dem zu entziehen?

Laut [Ka02] kommt Personalisierung an. Mehrere Studien haben nachgewiesen, dass Nutzer von Online-Shops lieber auf Webseiten einkaufen, die auf ihre persönlichen Bedürfnisse zugeschnitten sind (56%). Dagegen würden 87% der Nutzer genervt reagieren, wenn ihnen die gleiche Information mehrmals gegeben wird. 90% der Online-Kunden sind bereit, ihre E-Mail-Adresse anzugeben, wenn diese genutzt wird, um personalisierte Informationen zu erstellen. Eine Bereitschaft, die in anderen Bereichen längst nicht so hoch ist.

Doch wie sieht es aus, wenn der Kunde das nicht möchte? In Deutschland sind die Erhebung und Verarbeitung personenbezogener Daten nur zulässig, wenn gesetzliche Vorschriften sie ausdrücklich erlauben oder der Betroffene ausdrücklich eingewilligt hat. Den rechtlichen Rahmen für den Schutz personenbezogener Daten stellt das Bundesdatenschutzgesetz (BDSG) dar. Speziell für Teledienste finden sich weitere Regelungen im Informations- und Kommunikationsdienste-Gesetz (IuKDG) und hier insbesondere im Teledienstdatenschutzgesetz (TDDSG). Personenbezogene Daten sind hier als "Einzelangaben über persönliche [...] Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person" definiert und Nutzungsprofile sind demnach nur dann erlaubt, wenn sie anonymisiert sind. Sobald die Daten aber auf ausländischen Servern zwischengelagert werden, gelten leider andere Regeln. Findige Anbieter lösen das Problem durch nebulöse Formulierungen im Kleingedruckten ihrer Geschäftsbedingungen [Br00].

Der Schutz personengebundener Daten ist ein wesentliches Kriterium für die Fortentwicklung von individualisierten Angeboten. Unternehmen können das Vertrauen der Kunden nur gewinnen,

wenn sie sorgfältig und transparent mit personenbezogenen Daten umgehen. Das Angebot sollte bespielsweise nur dann personalisiert werden, wenn der Kunde das ausdrücklich wünscht und bereit ist, Informationen über sich preiszugeben.

Ein Schritt in die richtige Richtung ist der vom WWW-Konsortium (W3C) entwickelte Vorschlag einer *Platform for Privacy Preferences (P3P)*[1]. P3P wurde mit dem Ziel entworfen, dem Benutzer beim Besuch von Webseiten mehr Kontrolle über die Nutzung seiner persönlichen Informationen zu geben. Idee des Konsortiums ist, dass Betreiber einer Webseite, Informationen über die Verwendung von Kundendaten in maschinenlesbarer Form (XML) bereitstellen. Der Browser des Besuchers kann dann diese Angaben mit dem vom Benutzer angelegten Set eigener Privacy-Regeln abgleichen. Auf diese Weise kann der Nutzer entscheiden, ob und unter welchen Bedingungen persönliche Daten übertragen werden.

Es ist klar. Durch Individualisierung verliert der Kunde an Anonymität. Ob er bereit ist, diesen Preis zu zahlen, muss er letztendlich selbst entscheiden.

[1] vgl. http://www.w3.org/P3P/

Abbildungsverzeichnis

Literaturverzeichnis

[Br00] Detlev Brechtel und Jochen Mai, *Gläserner König*. In Wirtschaftswoche Nr. 07, S. 103, 2000

[Br98] 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
http://research.microsoft.com/users/breese/cfalgs.html

[Bo03] Craig Boutilier und Richard S. Zemel, *Online Queries For Collaborative Filtering*. In Proceedings of the 9th International Workshop on Artificial Intelligence and Statistics, 2003
http://research.microsoft.com/conferences/aistats2003/proceedings/171.pdf

[Ch97] 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
ftp://ftp.research.microsoft.com/pub/tr/tr-97-07.ps

[Fi99] Danyel Fisher et al., *SWAMI: A Framework For Collaborative Filtering Algorithm Development And Evaluation*, 1999
http://guir.cs.berkeley.edu/projects/swami/swami-paper/paper.html

[Go92] 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
http://doi.acm.org/10.1145/138859.138867

[Go00] Ken Goldberg et al., *Eigentaste: A Constant Time Collaborative Filtering Algorithm*, 2000
http://www.ieor.berkeley.edu/~goldberg/pubs/eigentaste.pdf

[Gu99] Dhruv Gupta et al., *Jester 2.0: Evaluation Of A New Linear Time Collaborative Filtering Algorithm*. In Proceedings of the SIGIR, ACM, 1999
http://portal.acm.org/citation.cfm?doid=312624.312718

[Hi95] 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
http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/wch_bdy.htm

[Jo97] Patrick Joseph, *On-Line Advertising Goes One-On-One*. In Scientific American - Cyber View, 1997

[Ka02] Sven Kauffelt, *Personalisierung im Web - Schlüssel zum Stammkunden*. In absatzwirtschaft Nr. 03, S.98, 2002

[Ko99] 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
http://doi.acm.org/10.1145/336992.337035

[Ma95] 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
http://www-2.cs.cmu.edu/~dmaltz/ACF95-draft8.txt

[Ma01] Klaus Manhart, *Wege aus der Anonymität*. In Market Nr. 08, S. 58, 2001

[Ma02] Stefanie Maute, *Konzeption und Realisierung eines Data Mining-Verfahrens für das Qualitätsmanagement in der Diabetologie*, Diplomarbeit an der Universität Ulm, 2002

[Ni97] David Nichols, *Implicit Rating And Filtering*. In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering, S. 31-36, 1997
http://www.comp.lancs.ac.uk/computing/research/cseg/projects/ariadne/docs/delos5.html

[Pe00] David Pescovitz, *Accounting For Taste*. In Scientific American - Cyber View, 2000

[Re94] 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
http://doi.acm.org/10.1145/192844.192905

[Re97] Paul Resnick und H. R. Varian, *Recommender Systems*. In Communications of the ACM, Vol. 40, No. 3, S. 56-58, 1997
http://doi.acm.org/10.1145/245108.245121

[Ru00] Matthias Runte, *Personalisierung im Internet - Individualisierte Angebote mit Collaborative Filtering*, Deutscher Universitätsverlag, 2000

[Sh95] 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
http://www.acm.org/sigchi/chi95/Electronic/documnts/papers/us_bdy.htm

[Un98] Lyle H. Ungar und Dean P. Foster, *Clustering Methods For Collaborative Filtering*, AAAI Workshop on Recommendation Systems, 1998
http://www.cis.upenn.edu/datamining/Publications/clust.pdf

[Link1] Übersicht über wichtige Papers zu Collaborative Filtering
http://www.theether.org/collab/papers.html

[Link2] Collaborative Filtering - Eine Übersicht
http://www.sims.berkeley.edu/resources/collab/