Diplomarbeit

Annehmbarkeit verschiedener Verfahren zur Wissensentdeckung auf E-Commerce Daten

Marina Neifach

Diplomarbeit
am Fachbereich Informatik
der Universität Dortmund

27. September 2001

Betreuer:

Prof. Dr. Katharina Morik
Dipl.-Inform. Ralf Klinkenberg

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

1.1 Problemstellung

Mit fortschreitender Entwicklung moderner Wirtschaft wird der Handel mit dem problem konfrontiert, welche Waren in melchen Kombinationen heutzutage angeboten werden sollen.

Für die Planung der Einführung von neuen Produkten wie auch für die Bestände von bereits angebotenen Produkten ist wichtig zu wissen, welche Artikel wie oft gekauft wurden und auch in der Zukunft gekauft werden. D.h. weiß ein Händler, was seine Kunden kaufen werden, kann er entsprechend sein Geschäft gestalten, benötigte Artikeln nachbestellen, neue Produkte einführen usw.

Dabei geht man in der Regel von der Grundannahme aus, daß ein Käufer dasjenige Produkt erwirbt, für das er die im Kaufzeitpunkt höchste Präferenz empfindet [Brockhoff, 1993].

Daher ist es sehr wichtig, Methoden zu entwickeln, mit denen der Nachfrager zielgerichtet zu genau den Produkten geführt werden kann, an denen er tatsächlich Interesse hat und die er möglicherweise kauft.

Hat der Händler die Interessen eines Benutzers herausgefunden, so kann er sie in Kundenprofilen mit geringem Aufwand für jeden Kunden dauerhaft speichern lassen. Diese Kundenprofile kann der Händler verwenden, um für jeden Kunden individuell passende Angebote zu unterbreiten.

Aufgrund einer unüberschaubaren Fülle von Angeboten und Inhalten im Internet ist es für einen Kunden schwierig, sich einen Überblik über alle Angebote zu verschaffen und das Beste für sich auszuwählen. Deshalb ist es wichtig, daß Kunden nur mit einer kleinen Auswahl, von für sie interessanten Produkten konfrontiert werden.

Sind diese für einen Kunden interessante Produkte dem Händler bekannt, so kann er einen Kunden bei der Suche behilflich sein, so daß der Kunde leichter und schneller zur den gewünschten Artikeln gelangt.

Ein Online-Geschäft hat gegenüber einem herkömmlichen den entscheidenden Vorteil, daß die Kunden sich registrieren müssen, und dadurch die Zuordnung von Verkaufsdaten einfacher ist.

Wie auch in der physischen Welt liegt in den verschiedenen Online-Shops ein wesentlicher Schlüssel zum Erfolg in der optimalen Befriedigung der Bedürfnisse der Kunden. In den Interaktiven Medien liegt die Konkurrenz buchstäblich nur einen „Mausklick“ entfernt [Runte, 2000]. Schafft ein Anbieter es nicht, den Kunden mit Angeboten zu versorgen, an denen er tatsächlich Interesse hat, so wird der Nachfrager binnen kurzer Zeit zu einem anderem Anbieter wechseln. Benötigt werden daher rechnergestützte Systeme, mit denen man erstens individuelle Präferenzen erheben und zweitens den Nutzern gemäß ihren Präferenzen „passende“ Produkte empfehlen kann [Runte, 2000].

1.2 Ziele

Ziel dieser Arbeit ist, die Einsetzbarkeit der verschiedenen Lernmethoden für die Wissensentdeckung auf E-Commerce Daten festzustellen und zu testen.

Dafür werden die Verkaufsdaten eines Online-Shops in Hinblick auf das Kundenverhalten analysiert, so das Kundengruppen gebildet werden können. Das Kriterium für das Bilden einer Gruppe soll die „Ähnlichkeit des Kaufverhaltens“, und damit auch die Ähnlichkeit der Interessen von Kunden sein. Diese wird über drei verschiedene Wege festgestellt: mit den Lernverfahren apriori, Midos und c4.5. Dann können für jede Gruppe individuelle Empfehlungen generiert werden.

Die Daten werden zuerst gesammelt. Dazu werden die Kunden gebeten, in einem Online-Shop virtuell einzukaufen. Dann werden die Kunden in Gruppen eingeteilt.

Um die Güte der Gruppeneinteileung zu überprüfen, wird ein Test durchgefürt. Den Testkunden werden bis zu fünf verschiedenen Empfehlungen vorgestellt. Jede Empfehlung besteht aus mehreren Produkten. Drei von fünf sind aufgrund der Erkenntnisse der drei Lernverfahren entstanden. Zum Vergleich werden eine Zufallsauswahl von Produkten, sowie ein Vorschlag, von dem Mittelwertverfahren geliefert, benutzt.

Das Ziel der Arbeit ist erreicht, wenn der Test zeigt, daß die Empfehlungen, die mit Hilfe der Lernverfahren generiert werden, bessere Ergebnisse liefern als die beiden Vergleichsvorschläge.

In diesem Fall sollte feststellbar sein, welches der Verfahren bei den Kunden die beste Akzeptanz genießt.

1.3 Gliederung

Die Diplomarbeit ist wie folgt aufgebaut.

Im folgenden Kapitel ist zunächst das Thema „Maschinelles Lernen“ vorgestellt. Das Maschinelle Lernen wird definiert. Dazu wird eine Definition des Lernens und der Lernaufgabe gegeben. Ein Anwendungsgebiet des Maschinellen Lernen ist Wissensentdeckung in Datenbanken. Dieses Anwendungsgebiet und drei Lernverfahren apriori, Midos und c.45 werden erläutert.

Ferner ist in diesem Kapitel eine Einführung in das Thema „Personalisierung“ gegeben. Ihre Arten werden beschrieben, die Möglichkeiten des praktischen Einsetzens und die damit verbundenen Probleme dargestellt. Personalisierung wird im Empfehlungssystemen benutzt, deshalb werden auch diese definiert, und die verschiedenen Arten der Empfehlungssystemen werden vorgestellt. Eines der Verfahren, das bei Empfehlungssystemen eingesetzt wird, heißt „Kollaboratives Filtern“ und wird umfassend erläutert.

Im nächsten Kapitel wird nach der Beschreibung des Ansatzes die Datenanalyse beschrieben. Die Datensammlung mit dem Lebensmittelshop und das Einkaufen wird vorgestellt. Dann wird für jedes der drei Lernverfahren die Suche nach geeigneter Repräsentation der Daten und das Einsetzen von jedem Lernverfahren beschrieben. Danach wird kurz das System Kepler vorgestellt. Auf das zum Vergleich benutzte Mittelwertverfahren wird eingegangen.

Die gewonnenen Ergebnissen werden dann getestet. Im fünften Kapitel wird der Test und die Testergebnisse beschrieben. Es folgen Fazit und Ausblick.

Kapitel 2

Grundlagen

2.1 Maschinelles Lernen

Möchte man maschinelles Lernen definieren, so ist zuerst das Lernen als solches zu definieren. Diese Definition ist aber nicht trivial. In der Literatur werden verschiedene Definitionen verwendet. Herbert Simon definiert Lernen so: [Simon, 1983]:

> *Lernen ist jede Veränderung eines Systems, die es ihm erlaubt, eine Aufgabe bei der Wiederholung derselben Aufgabe oder einer Aufgabe derselben Art besser zu lösen.*

Diese Definition ist an verschiedenen Stellen kritisiert worden. Zum einen kann man die Leistung eines Systems auf verschiedene Arten messen, d.h. eine vernünftige Leistungsmessung setzt voraus, daß man das Ziel der Handlung kennt. Der zweite Kritikpunkt ist der, daß ein System auch lernen kann, ohne das gelernte Wissen direkt anzuwenden. Als alternative Definition hat Michalski daher vorgeschlagen [Michalski, 1986]:

> *Lernen ist das Konstruieren oder Verändern von Repräsentationen von Erfahrungen.*

Mitchell (und auch Simon) definiert maschinelles Lernen so, daß es anwendbar wird [Mitchell, 1997]:

> *Ein Programm lernt aus Erfahrung E in Bezug auf Aufgabe T und einem Leistungsmaß P, wenn es seine Leistung P in Bezug auf Aufgabe T durch die Erfahrung E verbessert.*

Beispielsweise verbessert ein Datenbanksystem die Antwortzeit P auf Anfragen T auf Grund der Erfahrungen E, die es aus den Aktualisierungen der Benutzer hat.

Alternativ kann man versuchen das maschinelle Lernen durch die einzelnen Typen von Lernaufgabe zu definieren. Der Vorteil von diesen Vorgehen liegt daran, daß die einzelnen Lernaufgaben präzise definiert werden können [Morik et al., 2000].

> *Eine Lernaufgabe wird definiert durch eine Beschreibung der dem lernenden System zur Verfügung stehenden Eingaben (ihrer Art, Verteilung, Eingabezeitpunkte, Darstellung und sonstigen Eigenschaften), der vom lernenden System*

erwarteten Ausgaben (ihrer Art, Funktion, Ausgabezeitpunkte, Darstellung und sonstigen Eigenschaften) und den Randbedingungen des Lernsystems selbst (z.B. maximale Laufzeiten oder Speicherverbrauch).

Eine häufig untersuchte Lernaufgabe ist das Funktionslernen aus Beispielen [Morik et al., 2000]:

Es sei X eine Menge möglicher Instanzenbeschreibungen, D eine Wahrscheinlichkeitsverteilung auf X, und Y eine Menge möglicher Zielwerte. Es sei weiterhin H eine Menge zulässiger Funktionen (auch als Hypothesensprache L_H bezeichnet). Eine Lernaufgabe vom Typ Funktionslernen aus Beispielen sieht dann wie folgt aus.

Gegeben:

Eine Menge E von Beispielen der Form $(x, y) \in X \times Y$, für die gilt: $y = f(x)$ für eine unbekannte Funktion f.

Finde:

Eine Funktion $h \in H$ so daß der Fehler $error_D(h, f)$ von h im Vergleich zu f bei gemäß der Verteilung D gezogenen Instanzen aus X möglichst gering ist.

Diese Lernaufgabe gehört zu der Gruppe der **prädiktiv** orientierten Aufgaben. Das bedeutet, daß das Ziel dieser Lernaufgaben ist, eine unbekannte Funktion möglichst gut zu approximieren, also ein globales Modell zu finden [Morik et al., 2000].

Im Gegensatz dazu versucht man beim **deskriptiven** Lernen, die durch Hypothesen beschriebene Teilbereiche des Instanzenraums zu identifizieren, über die lokal interessante Aussagen gemacht werden können [Morik et al., 2000].

Später wird ein Entscheidungsbaumverfahren erläutert, das einen Lösungsansatz für das Funktionslernen aus Beispielen darstellt, und zwei Verfahren zum deskriptiven Lernen vorgestellt.

2.2 Wissensentdekung in Datenbanken

Ein Anwendungsgebiet des maschinellen Lernens ist die Analyse von vorher gesammelten Datenbeständen mit Hilfe von Lernverfahren, die sogenannte Wissensentdeckung in Datenbanken (Knowledge Discovery in Databases).

Im wissenschaftlichen Bereich wird die folgende Definition des Begriffs KDD oft zitiert [Fayyad et al., 1996]:

Wissensentdeckung in Datenbanken ist der nichttriviale Prozeß der Identifikation gültiger, neuer, potentiell nützlicher und schlußendlich verständlicher Muster in (großen) Datenbeständen.

Data Mining wird dabei als Bezeichnung für den eigentlichen Analyseschritt, in dem Hypothesen gesucht und bewertet werden, verwendet, d.h. Data Mining ist ein Teilschritt des KDD-Prozesses.

Brockhausen und Morik [Brockhausen und Morik, 1998] definieren Wissensentdeckung so:

Sei eine Datenbank E und eine Repräsentationssprache L_H gegeben. Die Aufgabe der Wissensentdeckung besteht darin, eine interessante und charakteristische Beschreibung H der Daten mit $H \in L_H$ zu finden, wobei die Interessantheit über ein Prädikat p bestimmt wird, sodaß gilt:

$$H(E,p) = \{h \in L_H | p(E, h(E)) \text{ ist wahr}\}$$

Die folgenden Abschnitten behandeln das Lernen von Assoziationsregeln und die Subgruppenentdeckung — zwei in Data Mining derzeit bekannten deskriptiven Lernaufgaben.

2.2.1 Lernen von Assoziationsregeln

Eine bekannte Anwendung von Assoziationsregelverfahren ist die sogenannte Warenkorbanalyse. Bei diesem Verfahren werden die Einkaufskörbe aller Kunden genau erfaßt, d.h. in einer Datenbank wird abgelegt, aus welchen Artikeln der jeweilige Einkauf (Transaktion) bestanden hat. Die Datenbank wird dann zur Optimierung der Geschäftsprozesse genutzt. Dabei hoffen die Anwender dieser Methode Antworten auf die Frage zu erhalten, welche Artikel unter welchen Bedingungen bei einer Transaktion gemeinsam gekauft werden. Wurden solche Artikel gefunden, so wäre eine mögliche Konsequenz, diese Artikel räumlich nah bei einander zu platzieren. Angebracht wäre auch, den Kunden günstige Angebote für einen Teil der Artikel zu machen (Lockvogelangebote).

Als eine Spezialisierung der deskriptiven Lernaufgaben wird das Entdecken von Assoziationsregeln wie folgt präzisiert [Morik et al., 2000]:

*Sei I eine Menge von Objekten („items") und sei T eine Menge von Transaktionen, wobei $\forall t \in T : t \subseteq I$. Sei weiterhin $s_{min} \in [0;1]$ eine benutzergegebene Minimalhäufigkeit (**„minimal support"**) und $c_{min} \in [0;1]$ eine benutzergegebene Minimalkonfidenz. Bei der Lernaufgabe Finden von Assoziationsregeln sind, gegeben I, T, s_{min} und c_{min}, alle Regeln der folgenden Form zu finden:*

$$X \rightarrow Y$$

wobei $X \subseteq I$ und $Y \subseteq I$ und $X \cap Y = \emptyset$, und es gilt:

$$s(r) := \frac{|\{t \in T | X \cup Y \in t|\}}{|T|} \geq s_{min}$$

sowie

$$c(r) := \frac{|\{t \in T | X \cup Y \in t\}|}{|\{t \in T | X \in t\}|} \geq c_{min}$$

Aus dieser Definition wird ersichtlich, daß eine Assoziationsregel dann als Lösung zulässig ist, wenn sie erstens eine gewisse Minimalhäufigkeit s_{min} für die in einer Regel vorkommenden Artikel aufweist, und zweitens wird verlangt, daß von den Transaktionen, die die Prämisse X beinhalten, mindestens ein Anteil c_{min} auch die Konklusion Y beinhalten soll.

procedure Apriori(I, T, s_{min}, c_{min})

$L := HÄUFIGE - MENGEN(I, T, s_{min})$
$R := REGELN(L, c_{min})$
$return\ R$

procedure $HÄUFIGE - MENGEN(I, T, s_{min})$
$C_1 := \cup_{i \in I}\{i\},\ k := 1,$
$L_1 := PRUNE(C_1)$
$while\ L_k \neq 0$
$\quad C_{k+1} := ERZEUGE - KANDIDATEN(L_k)$
$\quad L_{k+1} := PRUNE(C_{k+1}, T)$
$\quad k := k + 1$
$return\ \cup_{j=2}^{k} L_j$

procedure $REGELN(L, c_{min})$

$R := 0$
$forall\ l \in L, k := |l| \geq 2$
$\quad H_1 := \cup_{i \in l}\{i\}, m := 1$
$\quad loop$
$\quad\quad forall\ h \in H_m$
$\quad\quad\quad if \frac{s(l_k)}{s(l_k/h)} \geq c_{min}$
$\quad\quad\quad then\ add\ l_k/h \rightarrow h\ to\ R$
$\quad\quad\quad else\ H_m := H_m/\{h\}$
$\quad\quad while\ m \leq k - 2$
$\quad\quad\quad H_{m+1} := ERZEUGE - KANDIDATEN(H_m)$
$\quad\quad\quad m := m + 1$
$return\ R$

Abb. 2.1: Apriori-Algorithmus [Morik et al., 2000]

Ein Beispiel:

Es stellt sich heraus, daß mindestens 1% aller Kunden Windeln, Babynahrung und Bier kaufen, $s_{min} = 0,01$. Außerdem kaufen 50% aller Windeln- und Babynahrungkäufer auch Bier ($c_{min} = 0,5$):

$$I = \{\text{Windeln; Babynahrung; Bier; ...}\}$$

Also haben wir folgende Assoziationsregel gefunden:

$$\{\text{Windeln; Babynahrung}\} \rightarrow \{\text{Bier}\},$$

Übrigens diese, auf den ersten Blick verwirrende Regel, läßt sich leicht erklären: Viele Familienväter werden von ihren Gattinnen zum Einkaufen geschickt. Dabei kaufen sie Windeln und Babynahrung, holen aber gleichzeitig auch eine kleine „Belohnung“ für sich — ein paar Bierflaschen.

2.2.1.1 Der Apriori-Algorithmus

Das ursprüngliche Verfahren AIS ist vom von Agrawal [Agrawal et al., 1996] entwickelt worden, und wurde dann nach ca. zwei Jahren so verbessert, daß es nach wie vor als das Standardverfahren in diesem Bereich gilt. Der Algorithmus ist freiverfüglich z.B. unter http://fuzzy.cs.uni-magdeburg.de/ borgelt/software.html.

Die Idee des Apriori-Algorithmus ist, Kandidaten für häufige Mengen durch Induktion über die Teilmengenbeziehung zu generieren, d.h. mit 1-elementigen Kandidaten anzufangen und schrittweise zu 2-elementigen, 3-elementigen etc. Kandidatenmengen aufzusteigen. Dies ist möglich, da sich die Eigenschaft „häufig“ auf alle Teilmengen vererbt, d.h. umgekehrt jede häufige Menge als Vereinigung häufiger Mengen geringerer Kardinalität gebildet werden kann.

Der Algorithmus ist in der Abbildung 2.1 dargestellt. Dabei ist I die Artikelmende, T – die Transaktionsmenge. Er enthält zwei Teilprozeduren (siehe Abbildung 2.2). Die Teilprozedur $PRUNE$ entfernt diejenigen Mengen die die geforderte Minimalhäufigkeit nicht erreichen. Die zweite Teilprozedur $ERZEUGE - KANDIDATEN$ erzeugt aus den bereits gefundenen k-elementigen Mengen alle $k+1$-elementigen.

2.2.2 Subgruppenentdeckung

Das Ziel der Subgruppenentdeckung ist das automatische Finden derjenigen Teilgruppen, die hinsichtlich eines bestimmten Merkmals die interessantesten Auffälligkeiten zeigen. Dies kann z.B. zur Auswertung von Marktstudien oder Kundendatenbanken genutzt werden. Betrachten wir z.B. eine Versicherungsgesellschaft, die ihren Kunden Hausrat-, Haftpflicht-, und Lebensversicherungen anbietet und eine Kundendatenbank besitzt, in der für jeden Kunden verzeichnet ist, welche Produkte des Unternehmens bereits genutzt werden [Morik et al., 2000].

Ein deskriptives Lernverfahren könnte folgende lokale Beobachtungen entdecken:

> *„Unter den alleinstehenden jungen Männern in ländlichen Regionen ist der Anteil der Lebensversicherungskunden signifikant niedriger als im gesamten Kundenbestand.“*

Oder

> *„Verheiratete Männer mit Pkws der Luxusklasse machen nur zwei Prozent der Kunden aus, erzeugen aber vierzehn Prozent der Lebensversicherungsabschlusssumme.“*

Beobachtungen dieser Art geben zwar keine Antwort auf die Frage, welcher Kunde nun tatsächlich eine Lebensversicherung kaufen wird, sie stellen aber evtl. eine wichtige Basis dar, z.B. für die Planung von Geschäftsstrategien.

procedure $ERZEUGE - KANDIDATEN(L_k)$

$L_{k+1} := 0$
  $forall\ l_1, l_2 \in L_k so\ daß$
    $l_1 = \{i_1, ..., i_{k-1}, i_k\}$
    $l_2 = \{i_1, ..., i_{k-1}, i_k^*\}$
    $i_k^* \leq i_k (lexikografische\ Ordnung)$
  $let\ l := \{i_1, ..., i_{k-1}, i_k, i_k^*\}$
  $if\ alle\ k - elementigen\ Teilmengen\ von\ l\ sind\ in\ L_k$
  $then L_{k+1} := L_{k+1} \cup \{l\}$
$return\ L_{k+1}$

procedure $PRUNE(C_{k+1}, T)$

$forall\ c \in C_{k+1} : s(c) := 0$
  $forall\ t \in T$
    $forall\ c \in C_{k+1}, c \subseteq t : s(c) := s(c) + 1$
$return\ \{c \in C_{k+1} | s(c) \geq s_{min} * |T|\}$

Abb. 2.2: Apriori-Teilprozeduren [Morik et al., 2000]

Die den gerade beschriebenen Beobachtungen zugrunde liegende Lernaufgabe wird üblicherweise als Subgruppenentdeckung bezeichnet [Morik et al., 2000]:

Sei X ein Instanzenraum mit einer Wahrscheinlichkeitsverteilung D und L_H ein Hypothesenraum, in dem jede Hypothese als Extension eine Teilmenge von X hat:

$$ext(h) \subseteq X \text{ für alle } h \in L_H.$$

Sei weiterhin

$$S \subseteq X$$

eine gegebene, gemäß D gezogene Stichprobe der Gesamtpopulation. Es sei schließlich q eine Funktion

$$q := L_H \to \mathbf{R} \text{ (reele Zahl).}$$

Die Lernaufgabe Subgruppenentdeckung kann dann auf zwei Arten definiert werden:

1.Gegeben X, S, L_H, q und eine Zahl $q_{min} \in \mathbf{R}$, finde alle $h \in L_H$, für die $q(h) \geq q_{min}$.

oder/und

2.Gegeben X, S, L_H, q und eine natürliche Zahl $k \geq 1$, finde eine Menge $H \subseteq L_H, IHI = k$ und es gibt keine $h \in H$, $h' \in L_H$ H : $q(h') \geq q(h)$.

Ein Kernpunkt ist in dieser Definition die Qualitätsfunktion q. Für eine gefundene Subgruppen bewertet sie, wie „interessant“ diese ist. Gegeben q, verlangen wir bei der Subgruppenentdeckung entweder alle Subgruppen oberhalb einer bestimmten Mindestqualität q_{min}, oder/und die k gemäß q besten Hypothesen aus unserem Hypothesenraum [Morik et al., 2000].

2.2.3 Entscheidungsbaumverfahren C4.5

C4.5 wurde von J. Ross Quinlan 1993 [Quinlan, 1993] aus dem Entscheidungsbaumlerner ID3, ebenfalls von Quinlan, entwickelt. Entscheidungsbaumverfahren sind einfach zu bedienen, haben nur relativ kurze Laufzeiten, und die produzierten Entscheidungsbäume sind für die Benutzer relativ einfach zu verstehen. Wahrscheinlich deswegen sind diese Verfahren populär und häufig benutzt [Schneider, 2000]

Morik et al. [Morik et al., 2000] definieren Entscheidungsbaumverfahren so:

> *Entscheidungsbaumverfahren lösen die Lernaufgabe Funktionslernen aus Beispielen, wobei X eine Menge von durch n numerische oder diskrete Attribute beschriebenen Instanzen ist, Y eine kleine Menge von diskreten Klassenwerten ist und L_H die Menge der aus diesen Attributen und ihren Werten konstruierbaren Entscheidungsbäumen ist.*

Ein Entscheidungsbaum wird top-down, von der Wurzel zum Blatt gebildet. Zuerst wird für die Wurzel die gesamte Menge der Instanzen durch ein Attribut so zerlegt, daß der Informationsgewinn (Information Gain, siehe unten) am größten ist. Dadurch entsteht ein Baum der Höhe 2. In der Wurzel steht die Vergleichsbedingung, in den Blättern die Teilmengen. Dann wird die gleiche Prozedur rekursiv jeweils mit den Teilmengen in den neuen Knoten durchgeführt. Der Algorithmus stoppt, wenn in einem Knoten nur die Instanzen einer Klasse enthalten sind, oder falls alle möglichen Vergleichsbedingungen (Knotentests) bereits angewandt wurden. In den Blätter des Entscheidungsbaumes steht dann die jeweils zu vorhesagende Klasse, diese Knoten nennt man Blattknoten.

Möchte man eine unbekannte Instanz identifizieren, so fängt man bei der Wurzel an, und prüft den Wert des in der Wurzel angegebenen Attributs. Dann wird der Pfad zum nächsten Knoten, der dem in der zu klassifizierenden Instanz vorgefundenen Wert des Attributes entspricht, und so weiter bis zu einem Blattknoten.

Abbildung 2.3 zeigt ein Beispiel für einen Entscheidungsbaum, der verschiedene Früchte klassifiziert.

Gibt man dem Algorithmus beispielsweise eine unbekannte, runde, gelbe, pelzige Frucht, wird zunächst überprüft, ob diese Frucht rund ist. Trifft die Behauptung zu, wird geguckt, ob die runde Frucht orange ist. Das stimmt für unsere Frucht genauso wenig, wie die Annahme, daß sie grün ist. Sie ist aber pelzig, also handelt es sich um einen Pfirsich. Eine entsprechende Regel würde lauten: Wenn eine Frucht rund und nicht orange und nicht grün und pelzig ist, dann ist diese Frucht ein Pfirsich.

Der Grundalgorithmus für Entscheidungsbaumlernverfahren ist in der Tabelle 2.4 dargestellt (TDIDT ist eine Abkürzung für Top-Down Induction of Decision Trees).

In diesem Algorithmus ist $Qualität(T, E)$ ein Qualitätsmaß, das auch Informationsgewinn oder Information Gain genannt wird [Morik et al., 2000]:

Sei T ein Test mit k Ausgangen, der die Beispielmenge E in Teilmengen $E_1, ..., E_k$ gemäß dem Testausgang (je Ausgang eine Testmenge)zerlegt. Es seien $p_{i,1}, ..., p_{i,m}$ die relativen Häufigkeiten der m Klassen in der Teilmenge E_i. Dann definiere:

$$Qualit\ddot{a}t(T, E) := IG(T, E) := -\sum_{i=1}^{k} \frac{|E_i|}{|E|} I(p_1, ..., p_m),$$

wo $p_1, ..., p_m$ Auftretenswahrscheinlichkeit eines Symbols aus einer m-Elementigen Menge ist, und $I(p_1, ..., p_m)$ — Informationsgehalt (Entropie). In den Informationstheorie nennt man Informationsgehalt einer bestimmten Menge von Symbolen die durchschnittliche in Bit gemessene Anzahl der für die Identifizierung eines Symbols aus dieser Menge notwendigen Fragen:

$$I(p_1, ..., p_m) := \sum_{i=1}^{m} -p_i log p_i$$

Für den Test, der den durchschnittlichen Informationsgehalt der neuen Teilmengen am stärksten reduziert, ist die Informationsgewinn am größten.

Beschneiden von Entscheidungsbäumen

Entscheidungsbäume können relativ groß werden, wenn sie aus großen Datenmengen erstellt wurden. Dabei kann es vorkommen, daß einige Äste nur durch wenige Datensätze entstanden sind und nicht sehr hilfreich sind, wenn sie zum Klassifizieren einer Testmenge eingesetzt werden. Wird bei der Überprüfung eines Entscheidungsbaumes mit einer Testmenge festgestellt, daß ein Ast eine gleiche oder geringere Fehlerrate hätte, wenn er durch ein Blatt mit einem festen Wert ersetzt würde, so wird der Entscheidungsbaum an dieser Stelle beschnitten (Pruning Decisiontrees). C4.5 schätzt dabei die Fehlerrate, die der Entscheidungsbaum auf neuen Datensätze hat. Dazu wird einfach die Zahl der Fehler, die der Entscheidungsbaum bei dem Trainingsset hat, durch die Zahl der Datensätze im Trainingsset geteilt.

Regeln

Regeln entstehen aus Entscheidungsbäumen, indem man den Pfad von der Wurzel bis zum Blatt als Bedingung sieht (linke Seite) und dann den Wert des Blattes daraus schließen kann (rechte Seite). C4.5 kann Regeln aber noch vereinfachen. Wenn eine Regel aus dem Pfad eines Entscheidungsbaumes gefunden wurde, wird versucht, eine Bedingung wegzulassen, und anhand der Trainingsmenge überprüft, ob diese Verallgemeinerung die Fehlerrate übermäßig erhöht. Ist dies nicht der Fall, wird mit der verallgemeinerten Regel weitergearbeitet. So könnten alle Bedingungen überprüft werden. Dadurch entsteht ein Regellernverfahren.

2.3 Personalisierung

2.3.1 Definition

Häufig, wenn die Rede von Anpassung von Produkten an Kundenwünsche (auch Individualisierung genannt), Electronic Commerce, Software–Konfiguration (Benutzereinstellungen), feste News-channel, Anpassung von Webseiten an Kunden oder Besucher usw. ist, fällt das Stichwort „Personalisierung ".

Da dieser Begriff in so vielen verschiedenen Kontexten verwendet wird, gibt es keine allgemein anerkannte Definition. In dieser Arbeit soll unter der Aufgabe der Personalisierung Folgendes verstanden werden [Sonntag, 1998]:

> *Kunden/Interessenten werden mit genau den Informationen versorgt oder ihnen werden genau diejenigen Produkte angeboten, von denen aufgrund der durch das Sammeln, Speichern und Analysieren von Daten über die Benutzergewohnheiten ermittelten, individuellen Nutzerprofile anzunehmen ist, daß sie für den Benutzer von gesteigerten Interesse sind.*

2.4 Sinn und Zweck der Personalisierung

Personalisierung hat sowohl für den Anbieter als auch für den Kunden sowohl Vorteile, als auch Nachteile. Es liegt also im beiderseitigen Interesse, vorhandene Vorteile zu nutzen und Nachteile zu überwinden oder in Kauf zu nehmen. In folgenden Abschnitten werden verschiedene Einsatzmöglichkeiten der Personalisierung mit den dazugehörenden Vor- bzw. Nachteilen aufgelistet.

2.4.1 Zielgerichtete Werbung

Für den Anbieter von Webseiten hat Personalisierung unter anderen den Vorteil, daß er mehr Informationen über seine Benutzer erhält. Damit kann der Anbieter die Werbung auf seiner Webseite zielgerichteter platzieren. Da die Informationen damit auch für die Benutzer wertvoller werden, werden sie öfters diese Werbung anklicken — höhere Click-Through-Raten[1] werden dadurch erreicht. Je höher diese Zahl ist, desto mehr Benutzer haben die Seite der Werbetreibenden besucht. Die Click-Through-Rate ist demnach ein wichtiger Indikator für den Erfolg der Werbekampagne und die Qualität der Banner.

Diese Methode wird bereits in einer ähnlichen Form bei Suchmaschinen eingesetzt, wo die Werbung oft nach den eingegebenen Suchwörtern ausgewählt wird.

Außerdem hat man gemerkt, daß Werbeanzeigen um so erfolgreicher sind, je besser sie in die eingebettete Umgebung passen: Eine Anzeige für schnelle Sportwagen wird auf der Seite mit Informationen über die Formel 1 erfolgreicher

[1] Ein „Click-Through" ist das Anklicken eines Banners oder Werbebuttons durch den Besucher einer Internetseite. Die Rate der so genannten „Click-Throughs" oder „Ad-Clicks" wird berechnet, indem die Anzahl der Clicks-Through durch die Anzahl aller Seitenabrufe dividiert wird, mal 100. Die Click-Through-Rate ist damit eine wichtige Maßzahl für die Wirksamkeit einer Bannerwerbung. Sie verdeutlicht, wie viele Website-Besucher sich den Werbeinhalt eines Banners tatsächlich angesehen haben [PNP-Omline, 2000].

sein, als auf der Seite mit den Wirtschaftsnachrichten, auch wenn sie von derselben Person gelesen wird, die sich grundsätzlich dafür interessiert [Sonntag, 1998]. Damit ermöglicht die Personalisierung dem Anbieter der Webseite eine schnellere Reaktion auf das Interesse der Benutzer an einem Produkt: Klickt der Benutzer auf eine Anzeige, kann der Anbieter ihm ein passendes Angebot machen oder eine Verkaufsabwicklung vereinfachen.

Weiterhin kann auch die Präsentation der Produkte entsprechend den Anforderungen oder Interessen der Kunden angepaßt werden.

Die Effektivität der Werbung und damit die Treffsicherheit wird dadurch noch weiter erhöht. Mit der Erhöhung der Treffsicherheit steigt auch die Zahl der Bestellungen und letztendlich auch die Einnahmen der Anbieter.

2.4.2 Erhöhte Kundenbindung bei Electronic Commerce

Es gibt mehrere Gründe, warum der Einsatz von personalisierten Webseiten zu höherer Bindung der Kunden führen kann [Sonntag, 1998]:

- Der Kunde wird persönlich angesprochen und ist kein anonymer Besucher mehr. Dies erzeugt ein gewisses Gefühl der Verbundenheit und kann ihn zur Wiederkehr animieren. Es wird ihm sozusagen mitgeteilt, daß er als Person willkommen ist, und nicht einfach irgend jemand ist, dem etwas verkauft werden soll.

- Dem Kunden können speziell abgestimmte Angebote gemacht werden: Kunden, die diejenigen Artikel angeboten bekommen, die sie brauchen bzw. für die sie sich interessieren, haben eine viel stärkere Bindung als solche, denen viele nicht relevante Produkte angeboten werden.

- Da die Kundeninteressen und die von ihm bisher gekauften Waren bekannt sind, ist eine weitere Betreuung möglich, um zusätzliche Produkte gezielt zu bewerben (Siehe auch Punkt 2.4.1).

2.4.3 Speziell abgestimmte Informationen

Wie bereits früher erwähnt, bringt Personalisierung nicht nur für den Betreiber der Seite Vorteile, sondern auch für den Benutzer: Im günstigsten Fall werden ihm gezielt genau die für ihn interessanten Informationen zu Verfgung gestellt, ohne daß er sich selbst darum kümmern muß (Push-Technologie). Unter Umständen bedeutet es eine enorme Arbeitsersparnis. Für den Benutzer sind wenige, dafür ausgewählte und relevante, Informationen von größerem Wert, als eine enorme Menge an Daten, die er zwar zur Verfügung hat, aber nicht weiß, ob und wenn ja wo, sich die wichtigen Teile befinden [Sonntag, 1998]. Einen zusätzlichen Wert bringt auch die Zusammenstellung von Informationen aus verschiedenen Quellen, da dies eine eindeutige Zeit- und Arbeitsersparnis bedeutet.

Kein System kann aber sicherstellen, daß es alle und nur relevante Informationen liefert. Es kann also passieren, daß dem Benutzer einige Informationen verborgen bleiben, die er sonst gesehen haben könnte, falls das System zu wenige Informationen liefert, oder aber auch, falls das Benutzermodell fehlerhaft oder nur zu einfach ist, liefert das System dennoch viele für den Benutzer nicht interessante Informationen, was wiederum eine Mehrarbeit für ihn bedeutet.

2.4.4 Reduktion der Datenflut

Je mehr Informationen über das Internet übertragen werden müssen, desto stärker ist die Gefahr, daß die zur Verfügung stehende Bandbreite nicht ausreicht. Sollte die anfallende Datenmenge die Leistungsgrenzen überschreiten, wird die Kommunikation entweder sehr langsam oder bricht gänzlich ab. Diese Situation bezeichnet man als Bandbreitenproblem.

Da aber nun mehr an relevanten Informationen übertragen wird, kann das Bandbreitenproblem zumindest teilweise gelöst werden. Davon können sowohl die einzelnen Benutzer wie auch die Anbieter von Webseiten profitieren. Ohne Personalisierung werden große Datenmengen übertragen, weil wichtige Informationen in ihnen vermutet werden. Z.B. bei der Internet–Recherche liefern die meisten Suchmaschinen eine große Anzahl von verschiedenen Links. Doch in sehr vielen Fällen stellt sich dann heraus, daß die Information doch nicht die gewünschte ist, und die Daten werden einfach nicht weiter benutzt. Bis jedoch diese Entscheidung getroffen wird, wurde bereits eine erhebliche Bandbreite belegt und stand für wirklich wichtige Informationen nicht zur Verfügung.

Dadurch benötigt der Betreiber der Seite einen viel schmaleren Zugang zum Internet. Das kann unter Umständen vorteilhaft für ihn sein. Er hat dann zwar absolut weniger Besucher auf seiner Seite, doch sind diese für ihn wertvoller, da sie wirklich am Inhalt interessiert sind [Sonntag, 1998]. Selbstverständlich kann dadurch das Bandbreitenproblem im Großen und Ganzen nicht komplett gelöst werden, aber zumindest wird es verringert.

2.4.5 Verteilung von Informationen in Betrieben

Auch einigen Betrieben kann Personalisierung Vorteile bringen. Sie kann helfen, Informationen an die diese benötigenden Personen schneller und zielsicherer zu bringen. Existiert in dem Betrieb kein Personalisierungssystem, so müssen bei Bedarf an einer bestimmten Information die firmeninternen Wissensdatenbanken jedesmal durchsucht werden. Ist hingegen das Personalisierungssystem über die Arbeit unterrichtet, kann es automatisch verwandte Informationen suchen und neue Daten direkt an die Betroffenen weiterleiten. Dies bedeutet einen entscheidenden Fortschritt, weil der Ersteller der Informationen nicht wissen muß, wen diese betreffen [Sonntag, 1998].

2.4.6 Genaue Benutzerüberwachung möglich

Natürlich gibt es auch Nachteile. Ein Personalisierungssystem hat viel mehr an genaueren Informationen über seine Benutzer als es sonst der Fall ist. Deshalb ist es ohne weiteres möglich, daß die systeminternen Informationen auf eine dafür nicht vorgesehene Weise verwendet werden, die nicht den Wünschen der Benutzer entspricht. Dies kann von verstärkter Werbung (nur lästig) über Vertreterbesuche (unangenehm) bis hin zur Verwertung durch staatliche Behörden (u.U. sehr unangenehm) gehen. Leider gibt es keine erfolgversprechenden Ansätze, um das Entstehen dieses Problems zu verhindern [Sonntag, 1998].

Das Datenschutzgesetz Nordrhein-Westfalen (DSG NRW) schreibt vor, daß das Erheben personenbezogener Daten nur dann zulässig ist, wenn ihre Kenntnis zur rechtmäßigen Erfüllung der Aufgaben der erhebenden Stelle erforderlich ist. Durch die Art und Weise der Erhebung darf das allgemeine Persönlichkeitsrecht der betroffenen Person nicht beeinträchtigt werden. Personenbezogene Daten sind bei der betroffenen Person mit ihrer Kenntnis zu erheben; bei anderen Stellen oder Personen dürfen sie ohne ihre Kenntnis nur unter den Voraussetzungen des 13 Abs. 2 Satz 1 Buchstabe a und c bis g oder i des DSG NRW erhoben werden (siehe Anhang A).

Die betroffene Person muß über den Verwendungszweck aufgeklärt und über eine eventuelle Freiwilligkeit der Angaben unterrichtet werden.

Eine Verarbeitung für andere Zwecke ist mit wenigen Ausnahmen nur dann zulässig, wenn eine Rechtsvorschrift dies erlaubt und die betroffene Person eingewilligt hat.

Die Daten dürfen nur für Zwecke weiterverarbeitet werden, für die sie erhoben worden sind. Daten, von denen die Stelle ohne Erhebung Kenntnis erlangt hat, dürfen nur für Zwecke genutzt werden, für die sie erstmals gespeichert worden sind.

Die Übermittlung personenbezogener Daten an Personen oder Stellen ist zulässig, wenn eine Rechtsvorschrift dies erlaubt und sie zur rechtmäßigen Erfüllung der in der Zuständigkeit der übermittelnden Stelle liegenden Aufgaben erforderlich ist und der Auskunftsbegehrende ein rechtliches Interesse an der Kenntnis der zu übermittelnden Daten glaubhaft macht und kein Grund zu der Annahme besteht, daß das Geheimhaltungsinteresse der betroffenen Person überwiegt.

Die betroffene Person muß darüber in Kentnis gesetzt werden, daß die Übermittlung durch Gesetz oder eine andere Rechtsvorschrift ausdrücklich vorgesehen ist oder die Daten für Zwecke von Statistiken, die durch Gesetz oder eine andere Rechtsvorschrift vorgeschrieben sind, verarbeitet werden.

Der Empfänger darf die übermittelten Daten nur für die Zwecke verarbeiten, zu denen sie ihm übermittelt wurden. Hierauf ist er bei der Übermittlung hinzuweisen (siehe Anhang A).

Die meisten der Suchmaschinen z.B. Lycos, My Jahoo, die auch personalisierte Nachrichten anbieten, bieten Informationen über die Verwendung der persönlichen Daten eher verschämt an (winzige Links auf nur wenigen Seiten, etwa auch nicht auf der Startseite). Allgemein ist der Standard im Bezug auf die Vertraulichkeit eher gering. Z.B. werden die persönlichen Informationen auch dann weitergegeben, wenn die Firma nur annimmt, daß sie gesetzlich dazu verpflichtet sei [Sonntag, 1998].

Dennoch hat der Benutzer fast immer die Möglichkeit, den Empfang von unangenehmen oder nicht gewünschten Informationen zu unterbinden. Ebenfalls werden die Informationen meistens nur in aggregierter Form weitergeleitet (z. B. 65% der Besucher sind männlich).

2.5 Datenerhebung

Jedes Personalisierungssystem braucht Daten, die zuerst erfaßt werden müssen. Bei der Erfassung von Daten über den Benutzer gibt es drei Möglichkeiten:

- Reaktive Verfahren: Fragebogen, Formulare. (Beispiele in Abschnitten 2.6.1, 2.6.2)

  Bei der expliziten Messung kann das System selbst bestimmen, welche Objekte zur Bewertung vorzulegen sind. Für den Seitenbetreiber ist diese Methode sehr einfach, effizient und mit wenig Aufwand verbunden. Da der Benutzer alle Eingaben freiwillig macht, entstehen keine datenschutzrechtlichen Probleme bezüglich der Datenerhebung (siehe auch Abschnitt 2.4.6). Die weitere Verarbeitung von Daten muß selbstverständlich auch im Sinne des Datenschutzgesetzes sein. Jedoch wirkt diese Methode auf viele Benutzer abschreckend, da die Fragebogen sehr lang sein können.

- Nichtreaktive Verfahren (indirekte Datenerhebung) (Siehe auch Abschnitte 2.6.4, 2.6.5):

  Diese Verfahren kommen ohne Kooperationsbereitschaft des Surfers aus. Das System nimmt praktisch „im Vorbeigehen" Daten von Benutzern auf, ohne direkten Einfluß auf die Auswahl der vorgelegten Objekte zu nehmen oder nehmen zu können bzw. wollen. In der Praxis wird häufig das Interaktionsverhalten des Benutzers beobachtet. Über eine sogenannte „Click-Stream-Analyse" (die Liste der Seiten die ein Benutzer besucht hat) wird festgestellt, welche Inhalte von Webseiten ein Benutzer abgerufen hat oder für welche Produkte er sich interessiert. Die Zeit, die der Surfer auf den jeweiligen Seiten verbracht hat, zeigt sein Interesse noch genauer. Diese Daten lassen sich problemlos in Logfiles festhalten, die Zuordnung zu einem bestimmten Benutzer ist aber nicht ohne zusätzlichen Aufwand möglich. Z.B. kann man Session-ID oder Cookies benutzen. Anhand dieser Daten werden über den Benutzer Profile erstellt.

- Externe Daten (Abschnitt 2.6.3): Adressendatenbanken, Sammlungen von Click Stream.

  Die Daten können aber auch anderweitig erhoben werden und dann dem Personalisierungssystem zugeführt werden.

2.6 Haupttypen der Personalisierung

Die folgende Unterscheidung der Personalisierungstypen ist nur grob, oft werden Mischtypen verwendet. Damit versucht man, die Nachteile der einzelnen Typen oder zumindest deren Auswirkungen zu verhindern bzw. verringern.

Die Verschiedenen Typen unterscheiden sich in der Art und Weise wie die Daten erhoben werden, und auch in dem, was mit den gesammelten Daten möglich ist.

2.6.1 Namenserkennung

Bei der Namenserkennung wird versucht, den Namen des Benutzers zu ermitteln, um ihn persönlich zu begrüßen. Im einfachsten Fall wird der Name beim ersten Besuch abgefragt und gespeichert (z.B. in Cookies). Die Betreiber der Seite hoffen damit ein Gefühl der Vertraulichkeit zu erzeugen und den Kunden zu binden. Manchmal wird aber eine umgekehrte Wirkung erreicht. Die Kunden fühlen sich ausspioniert, gerade wenn die Namenserkennung in diesem Fall die einzige Anwendung der Personalisierung ist.

2.6.2 Check-Box-Personalisierung

Hierunter versteht man alle Arten von Personalisierung, die auf freiwilligen Angaben des Benutzers beruhen. Möchte ein Kunde das System benutzen, werden die Daten mit einem Formular abgefragt. Der Vorteil dieser Methode ist, daß der Kunde selbst entscheidet, welche Daten er preisgeben möchte. Der Aufwand an Rechenzeit und Speicherplatz ist sehr gering. Nachdem der Benutzer die Angaben gemacht hat, können ihm entsprechende Empfehlungen angeboten werden. Z.B. hat der Benutzer angegeben, daß er gerne surft, wird ihm ein Meeresurlaub angeboten. Für einen Bergsteiger werden andere Reiseziele vorgeschlagen [Sonntag, 1998].

Die Nachteile dieser Methode sind aber ziemlich schwerwiegend. Die Personalisierung ist relativ ungenau, da nur eine begrenzte Anzahl von Informationen abgefragt werden kann, weil Kunden nicht gerne längere Formulare bearbeiten. Oft machen sie unrichtige Angaben. Meistens kann es passieren, daß der Kunde einen Punkt einfach vergessen hat, oder er gibt an, woran er interessiert sein sollte (z.B. beruflich), nicht aber, woran er tatsächlich interessiert ist.

Aber auch wenn alle Angaben richtig und vollständig von den Benutzer gemacht wurden, gibt es Probleme. Der Geschmack eines Menschen kann und wird sich meistens auch ändern. Die Angaben bleiben aber die gleichen, es sei denn der Benutzer denkt daran, sie zu ändern.

2.6.3 Erstellung von Benutzerprofilen durch Segmentierung und Regeln

Unter Segmentation versteht man die Einteilung des Benutzers in homogene Gruppen anhand von Attributen. Zuerst werden dem Benutzer nur einige wenige Fragen gestellt, die sich kaum oder gar nicht auf den eigentlichen Inhalt der Personalisierung beziehen (z.B. Alter, Geschlecht, Einkommen). Danach kann der Benutzer in eine von mehreren vorher festgelegten Gruppen eingeordnet werden. Für jede Gruppe existiert eine Reihe von festgelegten Interessen. Diese können ein sehr breites Spektrum erfassen, und wurden oft durch die Mitwirkung von Psychologen und Statistiken entwickelt. Diese Interessen werden ebenfalls für den neuen Benutzer als zutreffend angenommen. Hat das System z. B. ermittelt, daß ein Benutzer ein älterer Hausbesitzer ist, könnte es aus vorgegebenen Regeln folgern, daß er sich für Alarmanlagen, Reisen für Senioren oder Gartenmöbel interessiert.

Regelbasierte Ansätze können sehr wissensintensiv, aber auch mächtig und präzise sein. Manuel erstellte Regeln und Profildaten müssen erhoben oder auf dem Markt gekauft werden.

Vorteile:

- Nur einige wenige allgemeine Fragen werden dem Benutzer gestellt. Er muß folglich auch keine langen Formulare bearbeiten.
- Die Strategie kann sehr vielfältige und detaillierte Interessen liefern, je nachdem wie fein und präzise die Gruppeneinteilung und dem Gruppen zugeordnete Interessen sind.
- Vorhersagen über komplett andere Sachrichtungen sind möglich.

Nachteile:

- Um eine sichere Gruppenzuordnung zu ermöglichen, kann die Erstellung des Fragenkatalogs sehr schwierig sein.
- Nur mit einer enormen Datenmenge und großen Erfahrungen ist es möglich, vernünftige Gruppen zu bilden.
- Im schlimmsten Fall kann es passieren, daß gerade in den, den Betreiber interessierenden Punkten das Profil falsch ist, da niemand den Durchschnitt genau entspricht, und das Profil in einigen Punkten mit Sicherheit falsch sein wird.
- Da die Einordnung in die Gruppen oft nur einmal geschieht, kann sie nicht mehr geändert werden. Dabei kann jeder Mensch sich im Laufe der Zeit mehrmals verändern.
- Kein Eingehen auf den Benutzer als Einzelperson, d.h. kein Hinzulernen.

2.6.4 Präsenzbasierte Personalisierung

Die Funktionsweise von den Systemen, die präsenzbasierte Personalisierung benutzen, kann in drei Schritten beschrieben werden [Rozycka et al., 2000]:

- Zuerst bewertet der neue Besucher einige Produkte, z. B. Filme, unter Beachtung des Datenschutzes. Diese Bewertung kann sowohl bewußt erfolgen (Reaktive Datenerhebung, dies wird auch als explizite Wertung bezeichnet, da der Benutzer zusätzliche Aktionen unternehmen muß, um eine Wertung abzugeben [Janetzko und Zugenmaier, 2000].), als auch unbewußt. In diesem Fall kann ein Zugriff auf ein Produkt als positive Bewertung interpretiert werden (nichtreaktive Dateerhebung).
- Dann werden die Bewertungen mit denen der übrigen Besucher, die bereits bei früheren Bewertungen gespeichert wurden, verglichen. Ziel dieses Vergleichs ist das Finden von einen oder mehreren Besuchern, deren Bewertungen mit den des neuen Besuchers möglichst gut übereinstimmen.

Oder aber der Unterschied zwischen Bewertungen eines Produkts von zwei Besuchern wird als Distanzmaß genommen, die Besuchen, für die die Summe der Distanzmaße kleiner als ein Schwellwert ist, bilden dann eine Gruppe.

- Danach werden nur noch Kaufvorschläge unterbreitet, die in das Profil des Nutzers passen, d.h. die Produkte, die von der im vorigen Schritt ausgewählten Besuchern positiv bewertet wurden.

Solche Systeme heißen Empfehlungssysteme (siehe Abschnitt 2.7) und arbeiten desto treffsicherer, je länger es sie gibt, und je mehr Anwender ihre Bewertungen abgegeben haben. Außerdem ist die Qualität der Informationen, die der Benutzer im Gegenzug erhält, desto besser, je ehrlicher seine Angaben sind [Rozycka et al., 2000].

Diese Form der Personalisierung basiert auf aktiven Kollaborativen Filtern (siehe Abschnitt 2.7.2.1). Die Wertungen bzw. Empfehlungen, die von einem Benutzer für ein Dokument, Webseite, Produkt o.ä. gegeben erden als Leitfaden für andere Benutzer bereitgestellt.

Das System, das diese Art des Kollaborativen Filterns nutzt lernt ständig hinzu, d.h. je mehr Benutzer teilnehmen und je mehr Elemente diese bewerten, um so besser wird die Personalisierung. Obwohl die Personalisierung nicht auf einzelne Personen bezogen ist, entspricht sie dennoch bei fast allen Teilnehmern sehr gut den tatsächlichen Interessen [Sonntag, 1998]. Auch ist es möglich Vorhersagen über Interessen über die Gebiete zu machen, von deren Existenz der Benutzer nicht einmal weiß, Voraussetzung dafür ist die Tatsache, daß die Präferenzen anderer Benutzer für dieses Gebiet bekannt sind.

Wenn es eine große Datenbasis vorliegt, können bei neuen Benutzern relativ gute Vorhersagen getroffen werden, auch wenn sie nur wenige Bewertungen durgeführt haben.

Als Nachteil der Systemen die diese Klasse von Methoden benutzen wirkt sich die Tatsache aus, daß eine größere Anzahl von regelmäßigen Benutzern notwendig ist, um sinnvolle und gute Ergebnisse zu erhalten.

Es können keine Vorhersagen über neue Produkte , die von keinem Benutzer noch bewertet wwurden, gemacht werden.

Außerden eignen sich solche Verfahren nicht für die Daten, die nur vom kurzfristigen Interesse sind, z.B. Nachrichtenartikel, da die Information veraltet ist, bevor genügend Bewertungen zur Personalisierung vorliegen.

2.6.5 Verhaltensbeobachtung

Im Gegensatz zu den vorher erwähnten Typen von Personalisierung wird hier von einer unpersonalisierten Seite ausgegangen. Die Aktionen der Benutzern werden beobachtet. Aus der Auswahl der Seiten und der Verweilzeit wird versucht, die Interessen des Benutzers zu identifizieren (ebenso über sonstige Aktionen wie Lesezeichen erstellen, die Seite lokal speichern, ausdrucken, ...). Diese werden dann über manuell erstellte Regeln dazu verwendet, den Inhalt persönlich zu gestalten.

Vorteile

- Die herausgefundenen Interessen sind meistens wirklich die tatsächlichen Interessen, sie treffen also fast immer zu.
- Der Benutzer muß keinen Fragenkatalog beantworten, sondern kann gleich beginnen.
- Es liegt eine „echte“ Personalisierung vor, da auf genau diesen einen Benutzer spezifisch eingegangen wird.
- Da die Strategie dauernd hinzulernt, kann sie auch auf sich ändernde Interessen reagieren.

Nachteile

- Hoher Aufwand für die Speicherung und Verarbeitung von Daten ist notwendig, da große Mengen an Daten anfallen (Klicks, Verweilzeit usw.).
- Nicht alle Aktionen der Benutzer bedeuten auch ein echtes Interesse. Eine lange Verweilzeit z.B. bedeutet noch kein besonderes Interesse: u. U. wird gerade eine Tasse Kaffe getrunken.
- Die Informationsgewinnung erfolgt nur sehr langsam, da aus einmaligen Aktionen nicht sehr viel abgeleitet werden kann. Sie ist daher nur für regelmäßige Benutzer mit starker Interaktion geeignet.
- Insbesondere bei bereits personalisiertem Inhalt kann es schwierig sein, aus den Aktionen bestimmte Vorlieben abzuleiten. Das Interesse der Benutzer kann sich sogar verringern.

  Z.B. bei einem Experiment mußten die Benutzer mehrere Witze bewerten. Entsprechend der Bewertung wurden dann die nächsten Witze angeboten, die dann auch bewertet wurden. Es hat sich herausgestellt, daß die „guten“ Witze nur dann als solche erkannt werden, wenn sie sich inmitten der schlechten befinden. Werden nur „gute“ Witze angeboten, legt sich die Begeisterung sehr schnell [Daum, 2000].
- Zu Beginn hat der Benutzer noch keinen Vorteil, da erst eine Lernphase notwendig ist.
- Eine Vorhersage über neue Gebiete ist nicht möglich.

2.7 Empfehlungssysteme

In den vorangegangenen Abschnitten wurden Methoden erklärt, mit deren Hilfe Systeme Vorschläge (Empfehlungen) an den Benutzer eines WWW-Dienstes liefern können. Falls die Abgabe von Empfehlungen für den Benutzer die hauptsächliche Aufgabe eines Systems ist, so werden solche Systeme unter dem Begriff Empfehlungssysteme zusammengefaßt.

Der Benutzer interagiert also wissentlich mit dem System oft mit der Motivation, eine Empfehlung für Objekte aus einer bestimmten Klasse zu erhalten [Runte, 2000].

Die Aufgabe der Verkäufer kann darin bestehen, solche Produkte zu finden, welche die Kunden wirklich brauchen oder sich wirklich wünschen. Mit diesen Produkten glauben die Verkäufer irh profit maximieren zu können. (Sicher gibt es auch unter den Verkäufern „schwarze Schafe", für die das erste Ziel Profitmaximierung um jeden preis ist, aber ich hoffe das sind die wenigsten.) Am besten für jeden Kunden passende Produkte. Deswegen standen schon immer die Vorlieben und die Abneigungen der Konsumenten bzw. Nachfrager im Vordergrund einer guten Unternehmensstrategie. Die Empfehlungssysteme helfen den Unternehmungen, diese Frage zu beantworten. Damit stellen sie eine hervorragende Ergänzung für den Handel im Web dar und spielen bei der Optimierung des Angebotes im Internet als Marketinginstrument eine große Rolle [Runte, 2000].

Mit Hilfe des Kollaborativen Filterns (Siehe Abschnitt 2.7.2 lassen sich Empfehlungssysteme auch zum Aufbau von Communities[2] nutzen und auch dazu, die Verbindung zu anderen Benutzern herzustellen, die die gleichen Vorlieben haben wie man selbst [Rozycka et al., 2000].

Empfehlungssysteme lassen sich nach ihrer Funktionsweise in nichtindividualisierte und individualisierte Systeme einteilen (Abb. 2.5).

Um sich die Funktionsweise der nichtindividualisierten Empfehlungssysteme vorstellen zu können, stellt man sich z.B. ein System vor, welches die Kinofilm–Präferenzen einer breiten Masse von Benutzern speichert und Präferenz-Mittelwerte für jeden Kinofilm berechnet. Die Kinofilme werden nach Ihrer Präferenz absteigend geordnet und die 100 Filme mit der höchsten mittleren Präferenz ausgegeben.

Wie man sieht, kommen bei den nichtindividualisierten Empfehlungssystemen vergleichsweise einfache Methoden der Empfehlungsgenerierung zum Einsatz. Die über solche Systeme abgegebenen Empfehlungen sind für jeden Benutzer identisch. Nichtindividualisierte Systeme nutzen die Möglichkeiten der Interaktiven Medien damit nur beschränkt.

Mit den in individualisierten Empfehlungssystemen verwendeten Methoden lassen sich typischerweise drei unterschiedliche Zielsetzungen verfolgen:

- Bei der ersten Zielsetzung handelt es sich um ein sogenanntes Filterproblem. Eine Menge von Objekten muß in „geeignete" und „ungeeignete" Objekte eingeteilt werden. Das Entscheidungskriterium für eine Filterung läßt sich unterschiedlich festlegen. Oft wird die Präferenz des Benutzers oder aber auch Interesse an den Objekten, oder Relevanz des Objektes für den Benutzer verwendet. Liegt die Präferenz für ein Objekt bzw. das Interesse an einem Objekt oder die Relevanz eines Objekten unter einem bestimmten Niveau, so wird das Objekt ausgefiltert.

[2]Mitglieder einer Communities (virtuellen Gemeinschaft) besitzen bestimmte Gemeinsamkeiten und sind nicht auf einen begrenzten geographischen Bereich beschränkt. Durch Nutzung modernster Kommunikationsmedien auf Basis Internets erfolgt Informationsaustausch und Kommunikation innerhalb einer virtuellen Gemeinschaft somit unabhängig von Ort und Zeit.

- Ein ähnliches Ziel wird verfolgt, wenn die Objekte in eine geordnete Liste, die mit der tatsächlichen Ranking durch den Benutzer möglichst gut übereinstimmen soll, eingefügt werden. Das Hauptinteresse gilt hier der relativen Präferenz der Objekte untereinander. Teilt man die geordnete Liste in einen oberen und einen unteren Teil, so erhält man wieder das mit der ersten Zielsetzung verfolgte Ergebnis, nämlich die Filterung der Objekte.

- Eine dritte Zielsetzung bildet die direkte Vorhersage, welche Präferenz ein bestimmter Benutzer für ein konkretes Objekt hat. Hier wird versucht aufgrund von vorhandenen Daten (über den aktuellen, wie auch über die anderen Benutzer) auf die fehlenden Daten zu schließen. Eine über den Einsatz in Empfehlungssystemen hinausgehende Anwendung liegt in der Schätzung von individuellen Kaufwahrscheinlichkeiten für Produkte. Dies ist vor allem für Betreiber von E-Commerce–Angeboten interessant, die eine Individualisierung ihres Angebots anstreben.

Um die oben dargestellten Ziele von Empfehlungssystemen erreichen zu können, ist es notwendig, daß Empfehlungssysteme einen Prognose-Mechanismus besitzen. In der Literatur beschreibt man hier zwei prinzipiell unterschiedliche methodische Ansätze (Balabanovic/Shoham 1997): Eigenschaftsbasiertes Filtern (Feature Based Filtering; auch Content Based Filtering genannt) und kollaboratives Filtern (Collaborative Filtering) (Abb. 2.5).

Der Unterschied zwischen eigenschaftsbasierten Systemen und Empfehler-Systemen liegt in dem Mechanismus, der zur Individualisierung der Empfehlung verwendet wird. Während in eigenschaftsbasierten Systemen die Eigenschaften der prinzipiell empfehlbaren Objekte untersucht werden, und damit eigenschaftsbasierendes Filtern zum Einsatz kommt, beruht die Empfehlung in Empfehler-Systemen auf einer Mehrzahl anderer Benutzer, die als „Empfehler“ (Recommender) tätig werden. Diese Systeme können persönliche Präferenzen des einzelnen Benutzers berücksichtigen, und dadurch individuell angepaßte Angebote generieren.

Die Verfahren der Empfehler-Systeme lassen sich immer dann einsetzen, wenn Konsumenten mit rechnergesteuerten Umgebungen in Kontakt kommen. So ist individuelle Empfehlung eines Buches beim Internet-Buchhändler BOL auf Basis eines gespeicherten Benutzerprofils nichts anderes als eine Anwendung einer Technik aus einem Empfehler-System.

In individualisierten elektronischen Zeitschriften kann die Schlagzeile eingeblendet werden, welche die höchste Aussicht auf das Weiterlesen und auf das Binden des Kunden an das Angebot erzeugt. In diesen Beispielen merkt der Kunde nicht direkt, daß er mit einem Empfehler-System interagiert. Die eingesetzten Algorithmen sind jedoch identisch [Runte, 2000].

Die Verfahren sowie deren spezifische Vor- und Nachteile werden in den folgenden Abschnitten näher betrachtet.

2.7.1 Eigenschaftsbasierte Systeme

Bei eigenschaftsbasierten Systemen wird für die betrachtete Klasse von Objekten (Objektdomäne) ein Eigenschaftsraum entwickelt. Objekte werden anhand einer Reihe von objektiven (und nicht etwa durch persönliche Meinung eines Benutzers gebildeten) Eigenschaften klassifiziert bzw. bewertet. Diese Eigenschaften können unterschiedlicher Art sein. Zur Darstellung objektiver Eigenschaftsräume wird beispielhaft oft die Domäne „Kraftfahrzeuge“ verwendet [Brockhoff, 1993]. Man könnte Kraftfahrzeuge z. B. nach der Eigenschaft „Aufbau“ bewerten. Mögliche Ausprägungen wären in diesem Falle u. a. „Limousine“, „Coupé“, „Cabriolet“ und „Kombi“. Weitere mögliche Eigenschaften wären die Leistung des Motors, Volumen des Kofferraums oder die Anzahl der Türen.

Man erkennt, daß sich Kraftfahrzeuge zumindest zu einem Teil anhand von Produkteigenschaften beschreiben lassen. Produkte können demnach als Eigenschaftsbündel beschrieben werden [Brockhoff, 1993]. Es gibt jedoch auch eine Reihe von Einschränkungen. Nur unter Schwierigkeiten ließe sich die Eigenschaft „Image der Automarke“ operationalisieren. Hier würde sich insbesondere die Frage der Objektivität stellen. Die Aussage, ob die Automarke Volkswagen als sportlich gilt, ist auch keine objektiv zu beantwortende Frage. Ziel des eigenschaftsbasierten Filterns ist es nun, anhand der objektiven Eigenschaften eines Objektes auf die Präferenz eines Benutzers für dieses Objekt zu schließen. Dies geschieht durch die Filterung von Objekten, die auf Basis ihrer Eigenschaften durch den Benutzer ausgeschlossen wurden. Ein solches Verfahren wird bereits bei Brockhoff beschrieben [Brockhoff, 1987]. Hier werden Gebrauchtwagen aus einem Gesamtangebot von Fahrzeugen gefiltert, indem Fahrzeuge nach bestimmten Kriterien ausgeschlossen werden können (z.B. Kilometerstand max. 50.000). Den herausgefilterten Objekten wird die Präferenz „null“ zugewiesen. Die verbleibenden Objekte werden in eine Rangordnung gebracht.

Allgemein lassen sich Einschränkungen bei der Anwendung eigenschaftsbasierten Filtern an folgenden Kriterien festmachen [Shardanand und Maes, 1995]:

- Entweder müssen die Objekte durch Maschinen „abtastbar “ sein (z. B. Text), oder es müssen ihnen Attribute oder Eigenschaftsausprägungen mit vertretbarem Aufwand von Hand zugewiesen werden können. Beim derzeitigen Stand der Technologie können Medien wie Klang, Bilder oder Video nicht hinreichend automatisch analysiert werden, um ihnen aussagekräftige Attribute zuweisen zu können. Weiterhin ist es in vielen Fällen nicht praktikabel, Eigenschaften von Objekten erstens zu erheben und zweitens diesen Objekten „von Hand“ zuzuweisen. Dies gilt insbesondere für Domänen, in denen es eine sehr große Anzahl für den Benutzer prinzipiell in Frage kommender Objekte gibt.

- Eigenschaftsbasierte Verfahren sind in der Regel nicht in der Lage, Objekte zu finden, die sich von bislang gefundenen Objekten in wesentlichen Attributen unterscheiden, aber trotzdem mit hoher Präferenz für den Benutzer verbunden sind. Es werden vielmehr Objekte gefunden, die den bislang gefundenen und vom Benutzer als hoch präferiert eingestuften

Objekten ähneln.

- Eigenschaftsbasierte Verfahren können Objekte nicht auf Basis von Kriterien wie „Qualität“, „Stil“ oder individueller „Blickwinkel“ empfehlen, da diese Eigenschaften keine objektive Merkmale sind. So kann ein eigenschaftsbasiertes System zwei Textdokumente nicht anhand ihrer Qualität oder Aussagekraft beurteilen, wenn die gleichen Begriffe verwendet werden.

- In vielen Objektkategorien ist die Beschreibung von Objekten anhand von objektiven Eigenschaften praktisch nicht durchführbar. Hierzu zählt u. a. der Bereich der Unterhaltung. Kinofilme, Bücher oder Musiktitel lassen sich nur äußerst eingeschränkt oder gar nicht anhand objektiver Eigenschaften beschreiben. Selbst wenn man bestimmte Eigenschaften wie „Länge des Films“ heranziehen würde, wäre der Erklärungsgehalt für die Globalpräferenz vermutlich nicht besonders hoch.

2.7.2 Empfehler-Systeme

Die Einschränkungen des eigenschaftsbasierten Filterns versucht man in Empfehler-Systemen mit Hilfe von Verfahren des „Kollaboratives Filtern“ zu umgehen.

Damit diese Systeme funktionieren, muß eine Vielezahl von Benutzern vorhanden sein. Diese Benutzer geben bewußt oder unbewußt ihre Präferenzen für eine Anzahl von Objekten. Aufgrund von diesen Präferenzen werden dann die Empfehlungen für andere Benutzer generiert. Dafür werden zuerst alle Benutzer in Gruppen eingeteilt, und zwar so, daß die Benutzer mit „ähnlichem “ Verhalten in einer und derselben Gruppe sind. Die Bewertungen dieser Benutzer werden dann dazu verwendet, für andere Benutzer dieser Gruppe Empfehlungen zu generieren. Solche Verfahren gehören zur Klasse „kollaboratives Filtern“ und werden in diesem Abschnitt umfassend behandelt.

Der Ausdruck „Collaborative Filtering“ wurde zum ersten Mal in einem Aufsatz über das System „Tapestry“ verwendet [Goldberg et al., 1992]. Dieses System ist ein Email–Filter–System, eingesetzt im Xerox Palo Alto Research Center. Dieses System filterte aus einer großen Menge von Emails die für einen Benutzer relevanten. Dabei helfen sich die Benutzer gegenseitig beim Informationsfiltern.

Kollaboratives Filtern bzw. Empfehler-Systeme gehören zu den adaptiven Techniken, die den Benutzern bei der Informationsbeschaffung helfen sollen. Basierend auf früheren Entscheidungen des Benutzers macht das System Vorschläge, die mit hoher Wahrscheinlichkeit den Präferenzen des Benutzers genügen sollen. In einem Online-Shop können solche Empfehlungssysteme eingesetzt werden, um den nicht vorhandenen menschlichen Berater (Verkäufer) zu ersetzen [Daum, 2000].

Wie bereits erwähnt benötigt Kollaboratives Filtern eine Startperiode, bis genügend statistische Daten vorhanden sind, um eine Empfehlung zu machen, und bis die Präferenzen eines individuellen Benutzers bestimmt werden können.

Sobald ein Benutzer als einer Gruppe zugehörig erkannt wird, können bei ihm als anfängliche Präferenzen die der Gruppe verwendet werden.

Kollaboratives Filtern weist gegenüber eigenschaftsbasiertem Filtern sowohl Vorteile, als auch Nachteile auf. Die Eignung der Verfahrens hängt vor allem von der betrachteten Klasse der zu filternden Objekte ab, also der Objektdomäne, aber auch von der Anzahl der regelmäßigen Benutzer, von den aktuellen Datenschutzbestimmungen usw.

Die Voraussetzungen der beiden Verfahren sind prinzipiell verschieden. Eigenschaftsbasiertes Filtern fordert, daß Objekte anhand ihrer Eigenschaften gut beschreibbar sein müssen. Nur diese Eigenschaften sind für die Präferenz der Benutzer relevant. Der spezielle „Geschmack“ des Benutzers spielt bei diesem Verfahren keine Rolle und ist sogar nicht darstellbar. Um den Aufwand des Verfahrens im Grenzen zu halten, müssen die Eigenschaften entweder maschinell oder mit wenig Mühe von Hand erhebbar sein. Typischen Einsatzbereich des eigenschaftsbasierten Filterns sind technische Produkte (Kraftfahrzeuge, Hi-Fi-Anlagen usw.).

Beim kollaborativen Filtern sind die Präferenzen dagegen subjektiv geprägt, die Präferenzbeziehungen existieren über die Objekt- und Benutzergrenzen hinaus. Die Objekte sind typischerweise überhaupt nicht anhand objektiver Eigenschaften beschreibbar, und wenn doch, dann ist deren Erhebung zu aufwendig. Um die Funktionalität des Verfahrens zu sichern, ist eine ausreichende Anzahl von Benutzergruppen mit änlichen Präferenzen für bestimmte Objekte oder Objektgruppen notwendig. Der Aufwand für die Startphase muß angemessen sein. Angewendet wird Kollaboratives Filtern z.B. beim Informationsmanagement oder in verschiedenen Empfehler-Systemen.

Zu den Vorteilen des eigenschaftsbasierten Filterns zählt, daß das Verfahren sofort funktionsbereit ist, nachdem der Benutzer seine Präferenzen abgegeben hat. Die Berechnung der Präferenzen geschieht völlig transparent, d.h. der Ergebnis der Prognose kann immer vollständig erklärt werden. Möchte man die Prognoseergebnisse beeinflüssen, kann man das leicht durch die Änderung der Eigenschaftsausprägungen erreichen. Außerdem gibt es die Möglichkeit, die Objektdomäne nach Objekten mit bestimmten Eigenschaften zu durchsuchen.

Gegen eigenschaftsbasiertes Filtern sprechen die Tatsachen, daß subjektive Eigenschaften nicht in die Bewertung einbeziehbar sind. Außerdem müssen für jedes Objekt deren Eigenschaften erhoben und ggf. gepflegt werden.

Im Gegenteil zum eigenschaftsbasierten Filtern deckt das kollaborative die Beziehungen zwischen Benutzern und Objekten auf, die nicht mit objektiven Eigenschaften beschreibbar sind. Das Verfahren ermöglicht Erfahrungsaustausch zwischen einer hohen Anzahl von Benutzern, die sich nicht zwingend persönlich kennen müssen. Dadurch nutzt es menschliches Wissen und Urteilsvermögen. Deshalb findet Kollaboratives Filtern auch die interessanten Objekte, die nicht anhand von Eigenschaften gefunden worden wären, und zwar auch dann, wenn der Benutzer nicht speziell nach diesen Objekten sucht.

Als Nachteil des kollaborativen Filterns wirkt sich die Tatsache aus, daß keine Prognose für neue Benutzer, die noch keiner Gruppe zugehören, möglich ist. Auch für neue Objekte ist keine Prognose möglich, da noch kein einziger Benutzer die Möglichkeit hatte, seine Präferenzen für dieses Objekt abzugeben.

Das Prognoseergebniss kann nur schwer erklärt werden, da das Kollaborative Filtern ein Black-Box-System darstellt. Schwerwiegend ist auch, daß die Eigenschaften von Objekten nicht mit einbezogen werden können, selbst wenn diese verfügbar und relevant sind.

Es ist also sinnvoll, in einem System die Vorteile von beiden Verfahren zu nutzen. Solche Systeme heißen „hybride Systeme", und werden in Abschitt 2.7.3 behandelt.

Je nachdem, ob die Benutzer ihre Präferenzen aktiv abgeben, oder ihre Aktionen beobachtet werden, kann man aktives und automatisches kollaboratives Filtern unterscheiden(siehe Abb. 2.6).

2.7.2.1 Aktives kollaboratives Filtern

Der Begriff des aktiven kollaborativen Filterns wurde in einem Aufsatz aus dem Jahre 1995 eingeführt [Maltz und Ehrlich, 1995]. Aktives kollaboratives Filtern bringt zum Ausdruck, daß Benutzer sich gegenseitig aktiv bestimmte Objekte empfehlen (Push-Kommunikation). Ein Beispiel bildet das oben erwähnte Tapestry–System.

Der Nachteil dieses Systems sehe ich vor allem darin, daß die Benutzer sich aktiv darum kümmern müssen, daß das System funktioniert. Je nachdem wie aufwendig die Bewertung ist, kann es leicht geschehen, daß die Benutzer die Bewertung teilweise oder vollständig ignorieren. Dann wird das ganze Verfahren in Frage gestellt.

Der Vorteil besteht darin, daß es keine Zweifel gibt, ob die erhobenen Daten auch den tatsächlichen Interessen entsprechen, angenommen alle Angaben werden nach besten Gewissen gemacht. Gibt der Benutzer sein Interesse bekannt, so ist die Wahrscheinlichkeit, daß diese Daten falsch sind, ziemlich klein.

2.7.2.2 Automatisches kollaboratives Filtern

Unter dem Begriff „Automated Collaborative Filtering" versteht man Verfahren, die die Eigenschaft der automatisierten Empfehlungsabgabe aufweisen [Baldi, 1998]. Dabei ist es nicht (mehr) notwendig, daß die Benutzer ihre Präferenzen bewußt abgeben.

Mit Hilfe dieser Verfahrensgruppe sind Systeme darstellbar, welche eine prinzipiell unbegrenzt große Menge von Benutzern aufnehmen können. Damit werden neue Marketingstrategien wie Mass Customization[3] und One-to-One Marketing[4] möglich [Snäbele, 1997]; [Pine, 1993].

[3] Mass Customization – Kundenindividuelle Massenproduktion - ist eine umfassende Wettbewerbsstrategie, die eine Fertigung abnehmerspezifischer Produkte zu einem Preis ermöglicht, der dem einer Produktion entspricht. In Märkten mit hohem Wettbewerbsdruck wird die erfolgreiche Umsetzung Strategie zu einem entscheidenden Wettbewerbsvorteil. Dabei bildet die effiziente Abwicklung informationstechnischen Ansprüche in der ganzen Wertkette eine der wesentlichen Erfolgsbedingungen [Piller, 1999].

[4] Dem One-to-One Marketing liegt in erster Linie das Ziel zu Grunde, den Umsatzanteil bei einem Kunden (share of customer) zu steigern. Die Steigerung des Marktanteils (share of market) steht dagegen erst an zweiter Stelle. Die Organisation des Unternehmens wird auf die Kunden ausgerichtet, mit dem Ziel, die individuellen Bedürfnisse der Kunden zu erkennen und gewinnmaximierend zu befriedigen [KPMG, 2000].

Die Verfahren der Klasse des Automatischen Kollaborativen Filterns lassen sich nach ihrem Prinzip in zwei Gruppen einteilen [Breese, 1998]. Die speicherbasierten Algorithmen (Memory Based Algorithms) nehmen bei jeder Anfrage Berechnungen über die gesamte Datenmenge vor. Im Gegensatz dazu benutzen modellbasierte Algorithmen (Model Based Algorithms) diese Datenmenge, um die Parameter eines Modells zu schätzen. Dieses wird nachfolgend dazu verwendet, um Empfehlungen und Prognosen für individuelle Benutzer abzugeben, ohne bei jeder Empfehlungsabgabe auf die gesamten oder große Teile der Datenbank zugreifen zu müssen [Runte, 2000].

Das ist der entscheidende Vorteil dieser Methode. Die aufwendige Benutzermodellierung kann also offline erfolgen, d.h. bei einer Anfrage an das Empfehler-System entsteht kein Zeitdruck. Das bereits berechnete Modell kann dann mit deutlich geringerem Rechenaufwand angewendet werden. Andererseits kann bei modellbasierten kollaborativen Filtern auch ein Informationsverlust entstehen. Während bei den speicherbasierten Verfahren die Berechnung immer auf aktuellen Daten stattfindet, geschieht es bei modellbasierten Verfahren in bestimmten Zeitabständen. Sind diese Abstände zu kurz entsteht ein unnötiger Zeit- und Rechenaufwand. Sind dagegen die gewählten Zeitabstände zu lang, so kann es passieren, daß das tatsächliche Benutzerverhalten sich bereits stark geändert hat, das Modell und somit die Empfehlungen aber noch nicht.

2.7.3 Hybride Systeme

Durch die Kombination von Eigenschaftsbasierten und Kollaborativen Filtern versucht man die spezifischen Nachteile der beiden Methoden zu beheben. Man spricht bei diesem Ansatz von Feature Guided Collaborative Filtering oder auch Content Based Collaborative Filtering [Pazzani, 1998].

In diesen Systemen wird zuerst die Menge der Objekte durch einen Eigenschaftsraum partitioniert. Diese vergleichsweise trennscharfe Vorauswahl geschieht Basis der Eigenschaften der Objekte. Dann werden für die restlichen Objekte Präferenzen über Kollaboratives Filtern ermittelt. Zur Ähnlichkeitsbestimmung zwischen den Benutzern können hier neben den vorliegenden Ratingdaten für Objekte auch die Partialpräferenzen der Benutzer verwendet werden [Runte, 2000].

Der Vorteil dieser Methode liegt daran, daß zum einen durch das Eigenschaftsbasierte Filtern eine effiziente Vorselektion von Objekten vorgenommen wird, soweit dies anhand von Eigenschaften möglich ist. Zum anderen übernimmt der Ansatz des Kollaborativen Filterns anschließend die Einbringung menschlicher Erfahrung und Urteilsvermögen in den Bereichen, in denen Objekteigenschaften versagen.

Um die hybride Systeme anwenden zu können, müssen die allgemeinen Voraussetzungen für beide Ansätze zumindest teilweise erfüllt sein. So muß die verfügbare Eigenschaftsinformation für die Objekte für die Partitionierung relevant sein. Analog gilt dies für die oben beschriebenen Voraussetzungen für das Kollaborative Filtern. Somit sind die typischen Anwendungsgebiete von Feature Guided Collaborative Filtering subjektiv geprägte Domänen mit einer breiten Auswahl von Objekten, in denen eine Reihe von Eigenschaften leicht erhebbar oder bereits verfügbar sind (z. B. Websites, Bücher und Restaurants) [Runte, 2000].

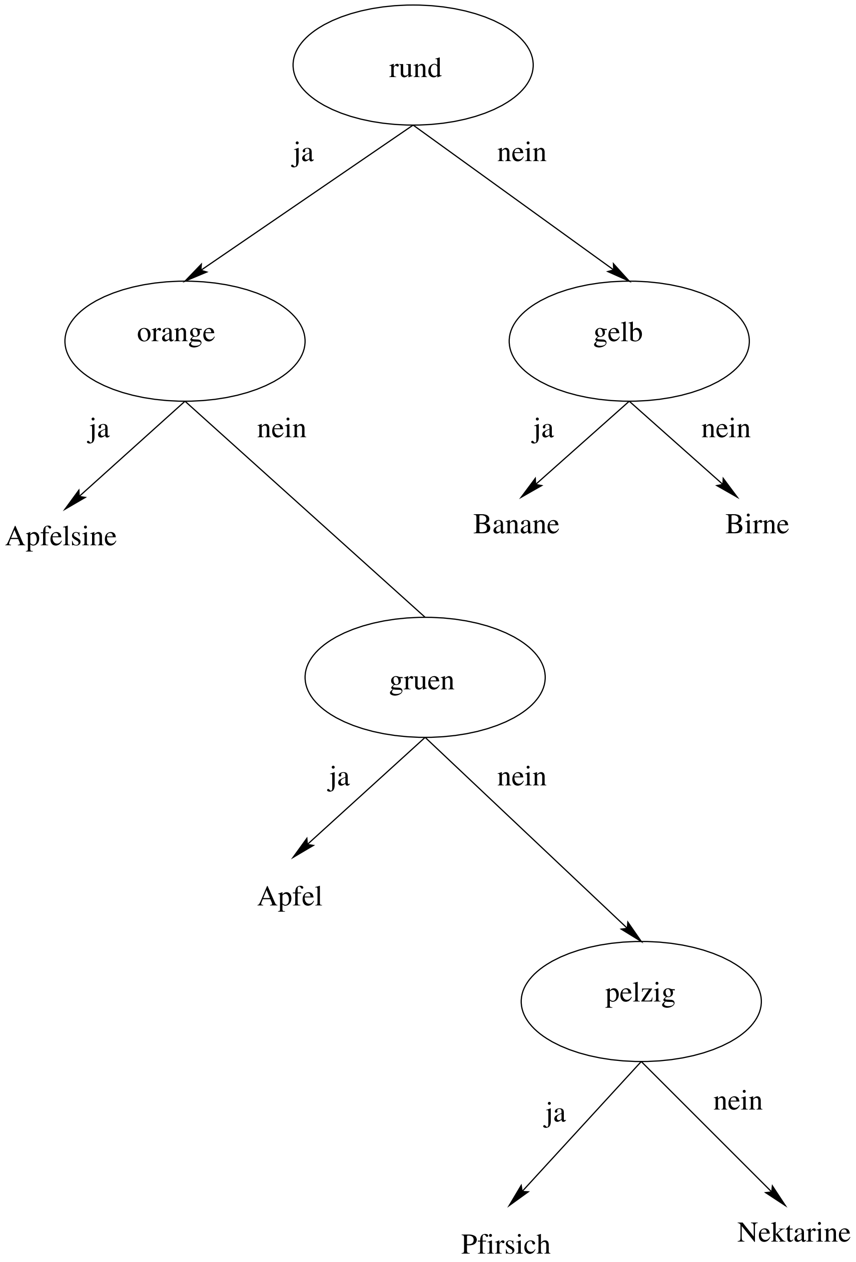

Abb. 2.3: Beispiel für einen Entscheidungsbaum [Schneider, 2000]

Sei E die Beispielmenge und T die Menge aller aus den Attributen und Attributwerden der Beispielmenge konstruierbaren Knotentests.

$TDIDT(E, T)$

- Falls E nur Beispiele einer Klasse enthält, liefere einen Blattknoten mit dieser Klasse zurück. Andernfalls:
- Für jeden $t \in T$, berechne $Qualit\ddot{a}t(t, E)$.
- Wähle den Test t' mit der höchsten Qualität für den aktuellen Knoten aus.
- Teile E anhand dieses Tests in 2 oder mehr Teilmengen $E_1, ..., E_k$ auf.
- Für $i = 1, ..., k$ rufe rekursiv: $T_i := TDIDT(E_i, T \backslash \{t'\})$.
- Liefere als Resultat den aktuellen Knoten mit den darunter hängenden Teilbäumen $T_1, ..., T_k$ zurück.

Abb. 2.4: Entscheidungsbaumverfahren [Morik et al., 2000]

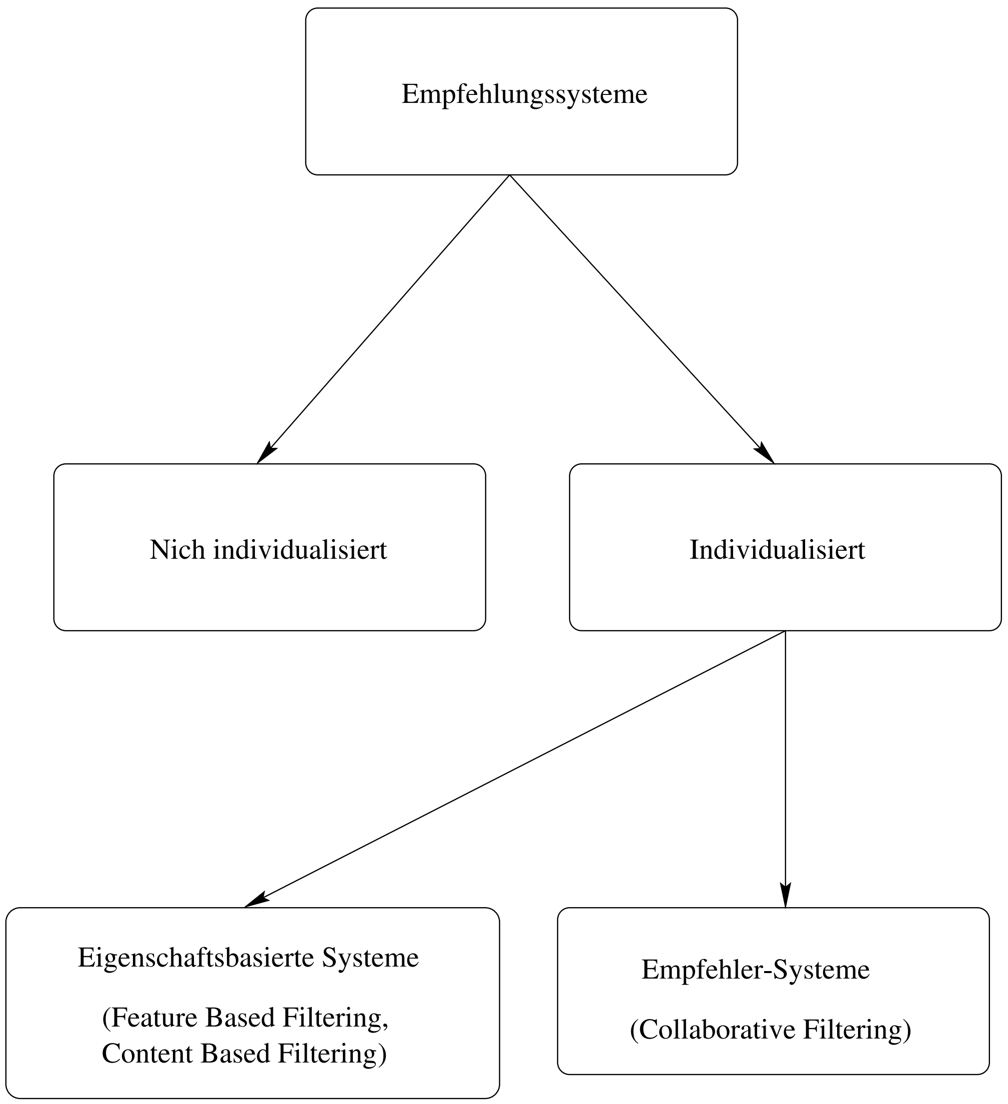

Abb. 2.5: Arten von Empfehlungssystemen [Runte, 2000].

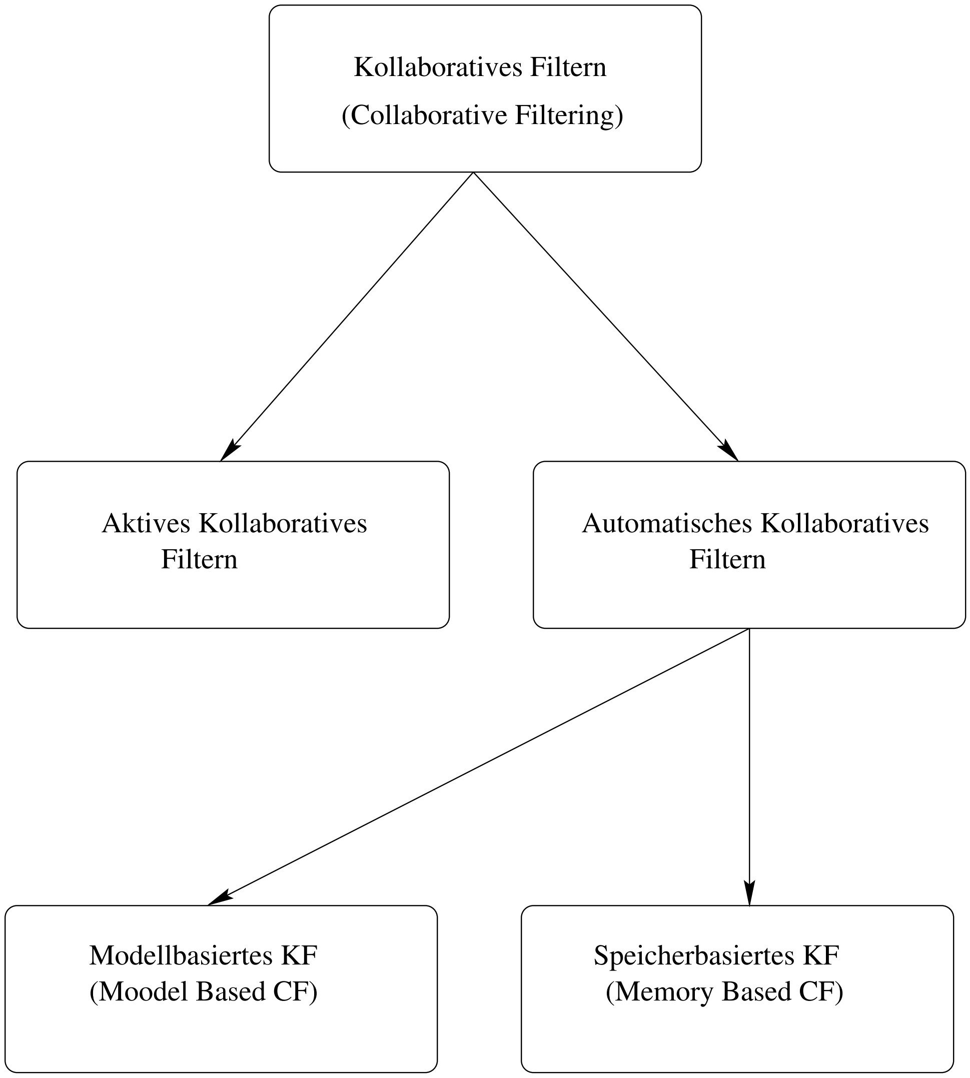

Abb. 2.6: Aktives und Automatisches Kollaboratives Filtern [Runte, 2000].

Kapitel 3

Kollaborative Datenanalyse

3.1 Ansatz

In dieser Diplomarbeit möchte ich die modellbasierten Algorithmen des Kollaborativen Filterns einsetzen. Dabei müssen sowohl die Vorteile wie geringer Rechenaufwand als auch die Nachteile (mögliche „Veralterung“ des Modells) dieses Verfahrens beachtet werden.

Die Datenerhebung erfolgt indirekt, die Kunden machen also ihre Bewertungen unbewußt. Die höchste Bewertung, die ein Kunde für ein Produkt abgeben kann, ist im Falle eines Kaufes abgegeben.

Nach der Datensammlung sollen dieKunden eines Internet-Shops in Gruppen eingeteilt werden. In einer Gruppe sollen die Kunden mit möglichst ähnlichem Kaufverhalten sein. Die Regeln mit denen die Benutzer in Gruppen eingeteilt werden, werden von den Lernverfahren apriori, Midos und c4.5 geliefert. Für jede Gruppe kann dann ein Vorschlag, der aus mehreren Produkten besteht, berechnet werden.

Um die Gruppeneinteilung zu testen, werden den Kunden die Vorschläge zum Kaufen angeboten.

3.2 Datenerhebung

Idealerweise sollte eine Kundenverhaltensanalyse auf realen Einkaufsdaten basieren. Deshalb habe ich als Erstes versucht, die Daten in Form einer Datensammlung zu finden. Ich konnte aber weder freiverfügliche noch freiverkäufliche Verkaufsdaten im Internet finden.

Danach habe ich verschiedene Internet-Händler angeschrieben mit der Bitte, mir ihre Daten in irgendeiner Form zu Verfügung zu stellen. Selbstverständlich habe ich versprochen die Anonymität des Händlers, der Produkte und der Kunden zu gewährleisten. Im Gegenzug habe ich den Händlern angeboten, die Ergebnisse der Diplomarbeit zu Verfügung zu stellen, was meiner Meinung nach einen nicht unerheblichen Nutzen für die Händler bedeutet hätte. Doch die verantwortlichen Personen waren nicht zu überreden. Sie waren nicht bereit, ihre Daten weder in aggregierter noch in anonymisierter, noch in irgendeiner anderen Form preiszugeben.

Ich kann mir vorstellen, daß aus der Sichtweise der Internet-Händler, der Nutzen sehr fraglich wäre, der Schaden aber, allein wenn bekannt worden wäre, daß diese vertrauliche Daten an dritte Personen weitergeleitet worden wäre, sehr groß sein könnte.

Also blieb mir keine andere Möglichkeit, als die Daten selbst zu erfassen. Dabei war der Grundgedanke, daß die Daten vermutlich realistischer ausfallen, wenn mehrere Personen in einem Internet-Shop „virtuell einkaufen", als wenn es nur eine Person täte (d.h. wenn ich alleine die Daten selbst generieren würde). Ich habe mich entschlossen einen Lebensmittelshop zu implementieren. Dabei habe ich mich für Lebensmittel entschieden, weil der Sachbereich wohl strukturiert sein kann, und allen Testpersonen bekannt ist. Wegen der einfachen und schönen Strukturierung habe ich als Vorlage den Online-Supermarkt „Freude am Kaufen" (http://www.freude-am-kaufen.de/) genomen.

3.2.1 Lebensmittelshop

Der Shop wurde auf einer Plattform von Vodafone TeleCommerce entwickelt, die eine bequeme Möglichkeit bietet, neue Online-Shops mit geringen Zeitaufwand zu entwickeln. Auch eine Produkt- und Benutzerverwaltung war bereits vorhanden.

Der Shop besteht aus 8 großen Abteilungen, die in insgesamt 38 Kategorien unterteilt wurden (siehe Abbildung 3.1). Insgesamt besteht der Shop aus ca. 500 einzelnen Produkten. Von der Startseite aus, wie auch mit Hilfe der Navigationsleiste im oberen Teil des Shops (Abbildung 3.1) kann sich der Kunde in dem Shop „bewegen". Befindet er sich in einer der Kategorien (siehe Abbildung 3.2), so kann er wie in einer Abteilung eines echten Geschäftes die einzelnen Produkte auswählen (genau so wie wir in einem realen Lebensmittelgeschäft ein Produkt in die Hand nehmen) und, falls diese Artikel seinen Vorstellungen entsprechen, diese kaufen (in den Einkaufswagen legen) (siehe Abbildung 3.3) oder aber auch nicht (wie wir die einzelnen Produkte wieder in die Regale stellen, falls wir uns doch anders entscheiden).

52 Leute haben sich bereit erklärt, in meinem Shop einzukaufen. Von diesen 52 Personen haben 22 mehr als einmal und 17 haben sogar mehr als zweimal eingekauft.

Insgesamt wurden 107 „Einkaufsgänge" gemacht, und 233 verschiedene Produkte in 36 Kategorien wurden virtuell gekauft. Im Durchschnitt wurden bei jedem Einkauf ca. 7 Produkte gekauft.

Die Daten wurden im Form vom Clickstreams in die Datenbank geschrieben, d.h. für jede Aktion wurden Kunden ID, der eindeutige Produktname, die Nummer des aktuellen Einkaufs und Einkaufsganges gespeichert. In diesem Fall bezieht sich ein Einkauf auf ein einzelnes Produkt. Ein Einkaufsgang umfaßt alle Einkäufe, die ein Kunde während eines Shopbesuchs macht. Außerdem wurde die Information, ob ein Produkt wirklich „gekauft", oder ob der Kunde nur sein Interesse für dieses Produkt gezeigt hat, es sich dann aber doch anders überlegt hat, festgehalten.

Die entsprechende Tabelle in der Datenbank sieht wie die Tabelle 3.1 aus.

Abb. 3.1: Die Startseite des Lebensmittelshops

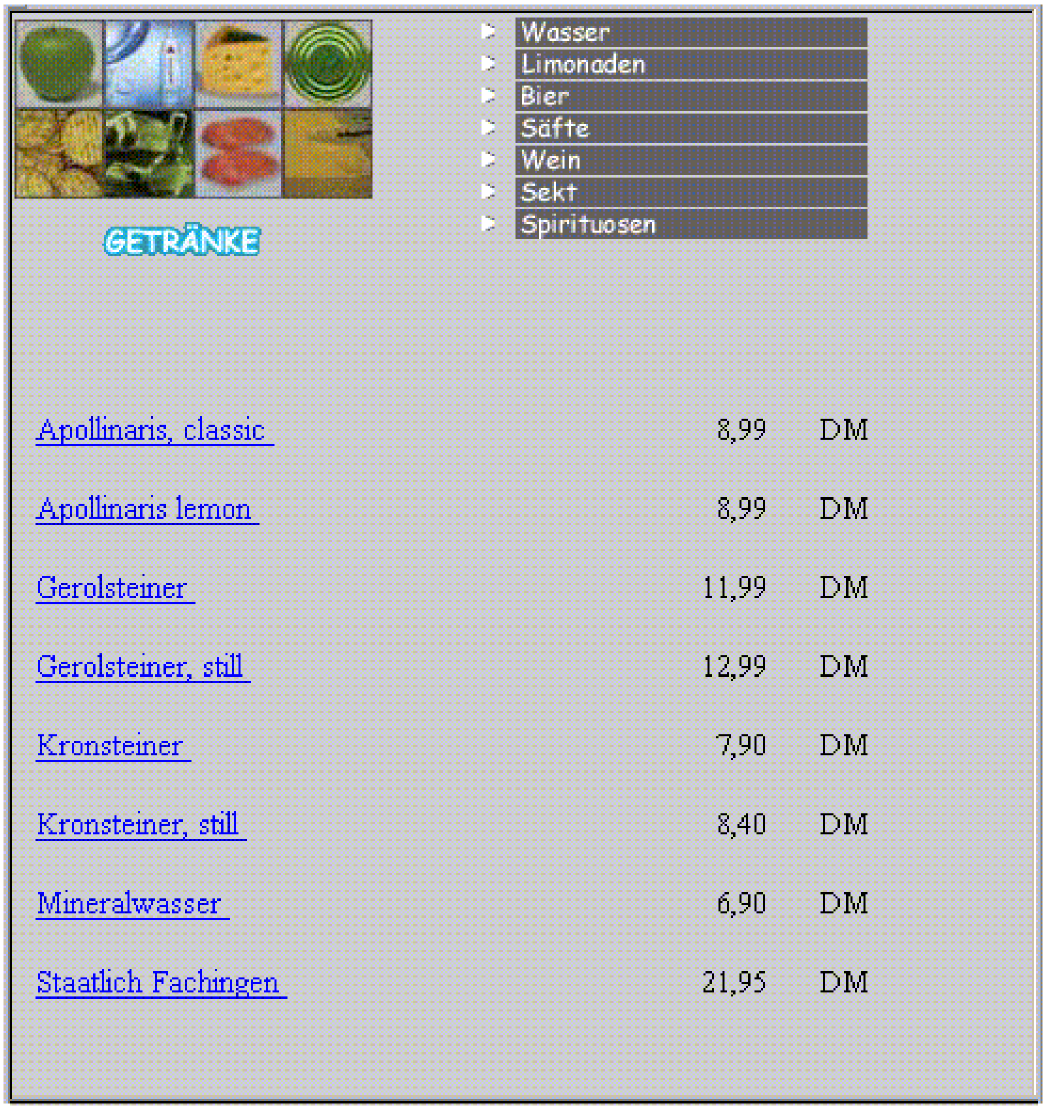

Abb. 3.2: Kategorie „Wasser“ in der Abteilung Getränke

Abb. 3.3: Beispiel für ein Produkt. Unten sieht man die Knöpfe für das endgültige Einkaufen oder für das Zurücklegen

Tabelle 3.1: Clickstream-Tabelle

Kunde	Produkt	EinkaufsNr	EinkaufsgangNr	gekauft?
Achim	Apfel	1	1	ja
Achim	Orange	2	1	nein
Anna	Orange	1	1	ja
Achim	Birne	3	1	ja
Alla	Apfel	1	1	nein
Anna	Butter	2	1	ja
Alla	Milch	2	1	ja
Alla	Wurst	1	2	ja
Alla	Joghurt	2	2	nein
Achim	Brot	1	2	ja
Achim	Apfel	2	2	ja
Achim	Butter	3	2	nein
Achim	Margarine	1	3	ja
...				

3.2.2 Testverhalten und Erfahrungen der Kunden(Testkäufer)

Alle Shop-Benutzer haben mir versichert, daß sie ihr Kaufverhalten in einem realen Lebensmittelgeschäft nachzuahmen versuchten. Einige haben sogar ihre Kassenbons aufgehoben, und dann ihre wirklichen Einkäufe dupliziert. Andere haben sich ein oder mehrere Gerichte vorgestellt, die sie gerne essen würden, und danach die nötigen Zutaten virtuell gekauft. Und die anderen Kunden haben mehr oder weniger das gekauft, was ihnen gerade eingefallen war. Diese Leute haben sich aber meistens auch in einem echten Geschäft so benommen, d.h. es wurde nach Lust und Laune gekauft, worauf man gerade Appetit hat.

In einem realen Online-Shop könnte der Verhalten der Kunden aber teilweise oder sogar ganz anders sein: Auch wenn mehrere Lebensmittelhändler gleiche Produkte anbieten, bieten sie diese zu unterschiedlichen Preisen, und auch die Qualität kann schwanken.

Der Preis ist aber ein wichtiger Bestandteil der Analyse. Es würden vermutlich die Produkte verstärkt eingekauft, die bei der Konkurrenz teuerer wären, da aber keiner der Kunden wirklich bezahlen mußte, konnte man sich in diesem Fall nicht wirklich auf den Zusammenhang zwischen dem Preis und dem Kauf eines Produktes verlassen. Auch die Gespräche mit den „Kunden“ haben gezeigt, daß manche Produkte nur deshalb gekauft wurden, weil sie nicht wirklich bezahlt werden mußten. Außerdem wäre bei einem realen Shop wichtig, ob die Kunden mit den Lieferbedingungen und letztendlich mit der Qualität der Lebensmittel zufrieden waren. In diesem Fall wäre z.B. die Anzahl der Kunden, die mehr als einmal einkaufen, ein Kriterium für Kundenzufriedenheit, aber da wir es hier nicht mit realen Produkten zu tun haben, können wir es nicht in Betracht ziehen.

Also für die Experimente wurde die Annahme gemacht, daß nur die Tat-

Tabelle 3.2: Kunde-Produkt Repräsentation

Kunde	Produkt	gekauft?
Achim	Apfel	ja
Achim	Orange	nein
Anna	Orange	ja
Achim	Birne	ja
Alla	Apfel	ja
Anna	Butter	ja
Alla	Milch	ja
Alla	Wurst	ja
...		

sache, das sich ein Kunde für ein Produkt interessiert, von Wichtigkeit ist. Selbstverständlich ist die Interesse eines Kunden größer, falls ein Produkt gekauft wurde, als wenn der Einkauf dann doch nicht zustande gekommen ist.

Alle anderen für einen realen Online-Shop möglichen Randbedingungen wie der Preis, die Anzahl der Produkte, die Qualität, der Lieferfrist und Lieferbedingungen, die kundenfreundliche Benutzeroberfläche bleiben bei den Experimenten unberücksichtigt.

3.3 Datenrepräsentation

Nachdem die Daten gesammelt wurden, stellte sich die Frage wie sie repräsentiert werden sollen. Da die Daten mit dem System Kepler (siehe Abschnitt 3.4.3), das sie in Form einer oder mehreren Datenbanktabellen verlangt, weiter verarbeitet wurden, mußten die Daten entsprechend vorbereitet werden.

3.3.1 Naiver Ansatz: Kunde-Produkt Repräsentation

Der nächstliegende Gedanke war, die Daten so zu nehmen, wie sie in der Datenbank vorhanden waren (siehe Tabelle 3.2). In jeder Zeile der Tabelle steht ein Kundenname, das Produkt auf das der Kunde zugegriffen hat und das Attribut „gekauft?“ daß zeigt ob bei diesem Zugriff das Produkt gekauft worden ist oder nicht.

Diese Repräsentation führte bei keinem der drei Verfahren (apriori, Midos, c4.5) zu einem Ergebnis.

3.3.1.1 Apriori

Das apriori-Algorithmus wurde in Abschnitt 2.2.1 bereits beschrieben. Wenn wir das apriori-Algorithmus betrachten wird uns klar, warum es in diesem Fall keine Ergebnisse liefern konnte: Es erwartet eine Transaktion als Eingabe. Typischerweise ist eine Transaktion eine Menge von Produkten, die von einem Kunden

Tabelle 3.3: Kundenvektoren

Kunde	Wasser	Bier	Apfelsaft	Apfel	Tomate	Weißwein	Brot	Tee	Frischmilch
Achim	**1**	**1**	0	0	0	0	**1**	0	**1**
Alla	1	0	0	1	0	0	0	1	1
Anna	**1**	**1**	0	0	0	0	**1**	1	**1**
Anton	0	1	1	0	0	1	1	1	0
Bruno	0	0	1	1	1	0	0	1	1
Igor	**1**	**1**	1	0	1	0	**1**	0	**1**

Tabelle 3.4: Produktvektoren

Produkt	Achim	Alla	Anna	Anton	Bruno	Igor
Wasser	**1**	1	**1**	0	0	**1**
Bier	**1**	0	**1**	1	0	**1**
Apfelsaft	0	0	0	1	1	1
Apfel	0	1	0	0	1	0
Tomate	0	0	0	0	1	1
Weißwein	0	0	0	1	0	0
Brot	**1**	0	**1**	1	0	**1**
Tee	0	1	1	1	1	0
Frischmilch	**1**	1	**1**	0	1	**1**

bei einem Einkauf gekauft wurden (Kundenvektor). Mit einer Repräsentation wie „Kunde-Produkt" kann apriori-Algorithmus nichts anfangen und liefert demnach auch keine Regeln.

Kann aber eine Transaktion eine Menge von Kunden oder besser gesagt eine Menge von Ausprägungen des Verhaltens eines Kunden sein (also Produktvektor)? Diese Frage versuche ich hierzu beantworten.

In den beiden Fällen sind für diese Arbeit vor allem die „Large Item Sets" (LIS) von Interesse. Zur Erinnerung: Large Item Sets heißen die Itemmengen, deren Support $s \geq s_{min}$ ist. Betrachten wir die beiden Tabellen (Tab. 3.3, Tab. 3.4). Sie bestehen aus untereinander geschriebenen Vektoren stellen also Matrizen dar. Die Tabelle 3.3 entsteht, wenn die Tabelle (Matrix) 3.4 transponiert wird und umgekehrt.

In der ersten Tabelle können wir z.B. ein LIS „Wasser, Bier, Brot, Frischmilch" sehen (fett). Supportwert für diesen LIS ist 3/6.

Wenn wir die entsprechenden Zeilen in der zweiten Tabelle betrachten, finden wir auch ein LIS: „Achim, Anna, Igor". Wie man sieht, sind es genau die Kunden, die Wasser, Bier, Brot und Frischmilch gekauft haben. Der Supportwert ist hier selbstverständlich ein anderer: 4/9. Man sieht also, daß jedem LIS in der ersten Tabelle ein LIS in der zweiten entspricht und umgekehrt. Da die

Supportwerte verschieden sind, werden auch die gefundenen LIS zwar bei einem bestimmten minimalen Support gefunden, die Zusammensetzung der LIS wird aber unterschiedlich sein.

Wie man sieht, stellen die LIS bei Kundenvektoren Mengen von Produkten dar, bei Produktvektoren sind die LIS — Mengen von Kunden.

Diese Kundenmengen sind die gesuchten Kundengruppen. Im Falle der Mengen von Produkten, müssen daraus noch Kundengruppen extrahiert werden. Dieses Problem wird gelöst, indem für jeden gefundenen LIS die Kunden als Bestandteile einer Gruppe genommen werden, die diese Produkte gekauft haben.

3.3.1.2 Midos

Bei dieser Art der Repräsentation kann man „Kunde“ oder „Produkt“ als Zielattribut bestimmen. Als Ergebnis bekommt man jeweils eine Subgruppe, die der Eingangspopulation gleicht.

Das bedeutet, daß diese Repräsentation nicht für das Algorithmus geeignet ist. Bei Midos stellte sich die grundlegende Frage, welche Ziele gestellt werden müssen, und welche der Algorithmus verfolgen kann. Offensichtlich kann Midos die Aufgabe

> *Teile alle Benutzer in Gruppen nach ihrem Kaufverhalten.*

gar nicht oder nicht effizient lösen. Dagegen ist die Frage

> *Was kaufen Kunden, die häufig Obst kaufen?*

zu beantworten. Es ist also sehr wichtig, daß nicht der Merkmal sondern der Merkmalwert (in unserem Fall „Obst“) als Spaltenname in der Tabelle vorkommt. Selbstverständlich soll diese Frage für jedes Produkt, das mindestens einmal gekauft wurden, gestellt werden.

Als Ergebnis werden Subgruppen ausgegeben, die wie in der Abbildung 3.4 aussehen können. Dabei stellt der erste Kreis die Eingagngspopulation, d.h. alle Einkäufe dar. Der hellere Bereich repräsentiert den Anteil der Einkäufe, in denen kein Obst vorkam, der dunklere Bereich — die Obst enthaltende Einkäufe.

Wenn wir z.B. die mittlere obere Subgruppe betrachten können wir sehen, daß 95 % der Kunden, die Brot und Käse gekauft haben, haben auch Obst gekauft. Die dazu gehörende Regel lautet:

> *Ein Kunde kauft Obst, falls er Brot und Käse gekauft hat.*

Diese Regel ist in also in 95% der Fälle erfüllt.

Betrachten wir die linke obere Subbgruppe, so sehen wir, daß die wenigsten Kunden (etwa 8 %), die keine Limonade, kein Bier, kein Sekt und keine Käse gekauften, Obst gekauft haben. Hieraus dürfen wir aber nicht schließen, daß bei den Kunden, die alle diese Dinge gerne kaufen, auch häufig Obst dabei ist.

Wie bei apriori kann man dann die Kunden als eine Kundengruppe betrachten, die sich der Regel konform benehmen. Allerdings kann hier die Anzahl der gefundenen Regeln variieren, je nachdem wie groß die man die minimale Prozentzahl der erfüllten Fälle setzt.

3.3.1.3 C4.5

Bei dieser Repräsentation kann man wie bei Midos als Ziel der Klassifikation entweder das Attribut „Kunde“ oder das Attribut „Produkt“ wählen. In beiden Fällen ist das Ergebnis ein breites Entscheidungsbaum der Höhe 1, in den Knoten stehen dann die Kunden bzw. die Produkte, es kann also keine neue Information gewonnen werden.

Auch c4.5 kann nur z.B. für ein bestimmtes Produkt ein Entscheidungsbaum liefern (siehe Abbbildung 3.5).

Wie man sieht, muß auch dieser Algorithmus für jedes Produkt einmal durchlaufen. Aus jedem Zweig eines Baumes werden Regeln generiert z.B. für Abbildung 3.5:

Obst; Sekt;
Obst; nich Sekt; Brot; Käse;
nicht Obst; nicht Sekt; nicht Brot;
nicht Obst; nicht Sekt; Brot; nicht Käse;

Wie bei Midos sind die Regeln – Kriterien, aufgrund deren Kunden in eine Gruppe zusammengefaßt werden.

3.3.2 Kunden- und Produktvektoren

Die nächsten Repräsentationen, die ausprobiert wurden, waren Kundenvektoren und Produktvektoren. Dabei wurde für jeden Kunden und jeden seiner Einkäufe ein N-dimensionaler Vektor erzeugt (Kundenvektor), wobei N die Anzahl der Produkte ist. Für jedes Produkt existiert ein M-dimensionaler Vektor (Produktvektor), M ist die Anzahl der Einkäufe. Wenn man alle Kundenvektoren untereinander schreiben würde, hätte man eine NxM – dimensionale Matrix bekommen. Würde man genau dasselbe mit Produktvektoren tun, so hätte die Matrix die Dimensionen MxN (Es ist genau die transponierte Matrix aus Kundenvektoren).

Anders ausgedrückt, falls

p : Index Produkt; P: Anzahl Produkte $(p = 1, ..., P)$

k : Index Kunde; K Anzahl Kunde $(k = 1, ..., K)$

e_k : Index Einkauf eines Kunden k; E_k : Anzahl Einkäufe eines Kunden$(e_k = 1, ..., E_k)$

dann kann ein Einkauf eines Kunden e eines Kunden k, also ein Kundenvektor folgendermaßen dargestellt werden:

$$a_{e_k} = \{a_{e_k,1}, ..., a_{e_k,p}, ..., a_{e_k,P}\},$$

wobei

$$a_{e_k,p} = \begin{cases} 2 & \text{:Kunde } k \text{ hat beim Einkauf } e \text{ das Produkt } p \text{ gekauft} \\ 1 & \text{:Produkt } p \text{ wurde von Kunden } k \text{ beim } e\text{-tem Einkauf angeschaut} \\ 0 & \text{:sonst} \end{cases}$$

Auch ein Produktvektor ist genau so aufgebaut. Er enthällt alle Informationen über ein Produkt:

$$a_p = \{a_{p,1}, ..., a_{p,E_1}, ..., a_{p,e_k}, ..., a_{p,E_K}\}$$

hier ist

$$a_{p,e_k} = \begin{cases} 2 & \text{:Kunde } k \text{ hat beim Einkauf } e \text{ das Produkt } p \text{ gekauft} \\ 1 & \text{:Produkt } p \text{ wurde von Kunden } k \text{ beim } e\text{-tem Einkauf angeschaut} \\ 0 & \text{:sonst} \end{cases}$$

Die Ergebnisse:

- apriori: Es konnten auch bei kleinen minimalen Support und kurzer Länge der Large Item Sets keine Regeln gefunden werden.
- c.45: Nur für einige wenige am häufigsten gekaufte Produkte konnten Entscheidungsbäume und Regeln gebildet werden. Folglich konnten nur wenige Kunden in Gruppen eingeteilt werden.
- Midos: Nur für wenige Produkte konnten Subgruppen gefunden werden, und auch dann wurden die dazugehörigen Regeln nie in mehr als 50% der Fälle erfüllt.

Nachdem ich also auch hier nichts Gutes bekommen habe, bin ich zu binären Vektoren übergegangen.

3.3.3 Binäre Kunden- und Produktvektoren

Alle von „0“ verschiedenen Stellen der Kunden- und Produktvektoren wurden auf „1“ gesetzt. Es gab also keinen Unterschied mehr, ob das Produkt wirklich gekauft wurde oder ob der Kunde sich doch dagegen entschied. Alle anderen Stellen wurden auf „0“ gesetzt (Tabellen 3.5 und 3.6).

In diesem Fall sind Einkäufe, Kunden- und Produktvektoren also Folgen von Nullen und Einsen.

Ein Einkauf eines Kunden (Kundenvektor) wird so definiert:

$$a_{e_k} = \{a_{e_k,1}, ..., a_{e_k,p}, ..., a_{e_k,P}\},$$

Produktvektor:

$$a_p = \{a_{p,1}, ..., a_{p,e_k}, ..., a_{p,E_K}\},$$

wobei

$$\begin{matrix} a_{e_k,p} \\ \text{bzw.} \\ a_{p,e_k} \end{matrix} = \begin{cases} 1 & : \begin{matrix}\text{Produkt } p \text{ wurde von Kunden } k \\ \text{beim } e\text{-tem Einkauf gekauft oder angeschaut}\end{matrix} \\ 0 & : \text{sonst} \end{cases}$$

Auch diese Repräsentation brachte keine vernünftigen Ergebnisse:

- apriori: lieferte keine Regeln bei $s_{min} \geq 5\%$ und $k \geq 4$.
- c.45: Wie in dem vorigen Abschnitt wurden hier nur für häufig gekaufte Produkte Regeln gefunden. Für die mesten Produkte, auch wenn es für sie Regeln gab, konnten nur ein oder höchstens zwei Kunden in eine Gruppe eingeteilt werden.
- Midos: Die Ergebnisse sind wie die im vorigen Abschnitt beschriebenen.

Tabelle 3.5: Datenrepräsentation als binäre Kundenvektoren

Kunde	Apfel	Orange	Birne	Milch	...
Achim	1	0	1	0	...
Alla	0	0	0	1	...
Anna	0	1	0	0	...
...					

Tabelle 3.6: Datenrepräsentation als binäre Produktvektoren

Produkt	Achim	Alla	Anna	...
Apfel	1	0	0	...
Orange	0	0	1	...
Birne	1	0	0	...
Milch	0	1	0	...
...				

3.3.4 Aggregierte Kunden- und Produktvektoren

Um die Anzahl der von Nullen verschiedenen Einträgen zu erhöhen, habe ich im weiteren Verlauf pro Kunden bei Kundenvektoren bzw. pro Produkt bei Produktvektoren nur jeweils einen Vektor betrachtet, in dem die Information über einen Kunden bzw. über ein Produkt aggregiert wurde. Ein Produktvektor hat dann genau so viele Stellen, wie es Kunden gab, also K.

Ein Kunden- bzw. ein Produktvektor sind somit:

$$a_k = \{a_{k,1}, ..., a_{k,p}, ..., a_{k,P}\}$$

$$a_{k,p} = \sum_{e=1}^{E} a_{e_k,p} \text{ bzw.}$$

$$a_p = \{a_{p,1}, ..., a_{p,k}, ..., a_{p,K}\}$$

$$a_{p,k} = \sum_{e=1}^{E} a_{p,e_k},$$

wobei

$a_{e_k,p}$ bzw. $a_{p,e_k} = 0$, wenn Produkt p von Kunde k während des Einkaufs e nicht gekauft,

$a_{e_k,p}$ bzw. $a_{p,e_k} = 1$, wenn Produkt p von Kunde k während des Einkaufs e angeschaut und

$a_{e_k,p}$ bzw. $a_{p,e_k} = 2$, wenn Produkt p von Kunde k während des Einkaufs e gekauft wurde.

D.h. wurde das Produkt von einem Kunden mehrmals gekauft bzw. hat der Kunde mehrmals überlegt, das Produkt zu kaufen, wurden die entsprechenden Zahlen addiert.

- apriori: Immer noch zu kleine Supportwerte.
- c.45: Die Ergebnisse sind wie die, im vorigen Abschnitt beschriebenen. Wenn es auch Regeln gab, dann konnten nur wenige Kunden in Gruppen eingeteilt werden.
- Midos: Die Ergebnisse sind wie die im vorigen Abschnitt beschriebenen.

3.3.5 Aggregieren nach Maximum-Prinzip

Nachdem alle Versuche mit vorherigen Repräsentationen gescheitert waren, habe ich nur Einsen und Zweien unterschieden: Wurde ein Produkt mindestens einmal gekauft, wurde die entsprechende Stelle der Kunden- und Produktvektoren mit „2" gekennzeichnet, bei Interesse – nur mit „1" und sonst – mit „0".

Die Definizion von Kunden- und Produktvektoren bleibt somit unverändert:

$$a_k = \{a_{k,1}, ..., a_{k,p}, ..., a_{k,P}\}$$

$$a_p = \{a_{p,1}, ..., a_{p,k}, ..., a_{p,K}\},$$

ihre Bestandteile werden aber anders definiert:

$$a_{k,p} = \max_{e=1}^{E} a_{e_k,p} \text{ bzw. } a_{p,k} = \max_{e=1}^{E} a_{p,e_k}$$

Die Ergebnisse von allen drei Lernverfahren blieben im Großen ung Ganzen unverändert. C4.5 und Midos lieferten zu wenige Regeln (es konnten nur einzelne wenige Kunden in Gruppen eingeteilt werden), bei apriori mußten die Supportwerte immer noch unter 8% und der Länge der Large Item Sets unter 3 gesetzt werden, damit das Algorithmus Regeln liefern konnte.

3.3.6 Aggregierte binäre Kunden- und Produktvektoren

Auch bei den aggregierten Vektoren wurde die binäre Variante ausprobiert. Der Einkauf ist wie oben definiert, $a_{k,p}$ bzw. $a_{p,k}$ können nur zwei Werte annehmen:

$$\begin{matrix} a_{k,p} \\ \text{bzw.} \\ a_{p,k} \end{matrix} = \begin{cases} 1 & \text{:Produkt } p \text{ wurde von Kunden } k \text{ gekauft oder angeschaut} \\ 0 & \text{: sonst} \end{cases}$$

- apriori: Bei Support-Werten von mehr als 10% und Lange der LIS gröser 3 immer noch keine Regeln.
- c.45: Die Qualität der Ergebnisse hat sich nicht verändert.
- Midos: Die Ergebnisse sind wie die im vorigen Abschnitt beschriebenen.

Tabelle 3.7: Übergang von Produkten zu Kategorien

Kunde	Apfel	Birne	Wasser	Apfelsaft
Achim	1	1	0	1
Alla	1	2	1	1
Anna	0	0	2	0
...				

→

Kunde	Obst	Getränke
Achim	2	1
Alla	3	2
Anna	0	2
...		

3.3.7 Übergang von Produkten zur Kategorien

Der Grund für das Scheitern an dieser Stelle sehe ich vor allem in der Datenmenge. Die Menge der Daten, die ich sammeln konnte, war offensichtlich nicht ausreichend. Die Zahl der angebotenen Produkte war mit der Menge verschiedener Produkten in einem echten Online-Shop vergleichbar — die Anzahl der Kunden aber bei weitem nicht. Die Anzahl der Kunden ist im Vergleich zu der Anzahl der Produkte viel zu klein, deshalb konnten auch keine oder nicht genug „ähnlichen“ Kunden gefunden werden.

Ich konnte aber auch nicht, um mehr Daten zu bekommen, die vorhandenen Beispiele einfach duplizieren. Beim Duplizieren würden auch die „Ausreißer“ vervielfacht, und damit zu Regeln mutiert.

Auch das manuelle Erzeugen von neuen Beispielen war nicht möglich, wegen der damit verbundenen Annahmen über die Zusammensetzung der Einkäufe. Das könnte die Lernergebnisse stark verfälschen, weil sich die gemachten Annahmen vermutlich im Lernergebnis wiederfinden, egal ob diese Annahmen zulässig waren oder nicht.

Genauso wenig konnte ich annehmen, daß die einzelnen gekauften Produkte unabhängig von einander ausgewählt worden waren, und die Einkäufe einfach durch Zufallsauswahl selbst zusammensetzen. Meistens kauft man ja für ein bestimmtes Gericht oder Gerichte ein, und da hängen die Zutaten in einer ganz bestimmten Art und Weise von einander ab. Es mußte also eine andere Art von Repräsentation gefunden werden, damit die eingesetzten Lernverfahren mit den vorhandenen Datenmenge die Ergebnisse, mit denen dann letztendlich die Empfehlungabgabe möglich wird, liefern können.

Wenn also eine Erhöhung der Kundenzahl nicht möglich war, mußte die Zahl der Produkte verringert werden. Das konnte durch den Übergang von Produkten zu Kategorien erreicht werden. Dafür wurden für jeden Kunden und jeden Einkauf die Stellen, die Produkten aus einer und denselben Kategorie entsprechen addiert. Dadurch wurde zwar die genaue Information über die einzelnen Produkte verloren, die Anzahl der Kunden, die „Gleiches“ gekauft haben aber erhöht (siehe Tabelle 3.7).

3.3.8 Aggregierte Kunden- und Kategorievektoren

Bei dieser Repräsentation ist k wie oben ein Kundenindex ($k = (1, ..., K)$); t ist Index der Kategorien ($t = (1, ..., T)$).

Pro Kunde bzw. pro Kategorie gibt es jeweils nur ein Vektor. Ein Kundenvektor ist somit:

$$a_k = \{a_{k,1}, ..., a_{k,t}, ..., a_{k,T}\},$$

Kategorievektoren sind analog dan Produktvektoren aufgebaut:

$$a_t = \{a_{t,1}, ..., a_{t,k}, ..., a_{t,K}\},$$

wobei

$a_{k,t}$ bzw. $a_{t,k} = 1$, wenn mindestens ein Produkt aus Kategorie t von Kunde k während dangeschaut wurde

$a_{k,t}$ bzw. $a_{t,k} = 2$, wenn mind. ein Produkt aus Kategorie t von Kunde k gekauft wurde.

$a_{k,t}$ bzw. $a_{t,k} = 0$, sonst.

Die Ergebnisse waren durchaus weiter verwendbar. Apriori lieferte bei verschiedenen Support und k-Werten Regeln.

Auch c4.5 hat für die meisten Kategorien Regeln bilden können, die Anzahl der Regeln für eine Kategorie konnte mit der nächsten Repräsentation noch verringert werden. Zu viele Regeln bedeuten auch zu viele Kundengruppen, was wiederum große Speicheraufwand bedeutet.

Bei Midos wäre es aber wünschenswert, wenn die Anteil der Subgruppen, bei denen die Regeln in mehr als 50% der Fälle erfüllt waren, erhöht werden könnte. Das wurde mit der nächsten Repäsentation erreicht.

3.3.9 Binäre Kunden- und Kategorievektoren

Repräsentation analog wie bei binären Kunden- und Produktvektoren.

$$a_k = \{a_{k,1}, ..., a_{k,t}, ..., a_{k,T}\}\text{bzw.}$$

$$a_t = \{a_{t,1}, ..., a_{t,k}, ..., a_{t,K}\},$$

$$\begin{array}{c} a_{k,t} \\ \text{bzw.} \\ a_{t,k} \end{array} = \begin{cases} 1 & : \begin{array}{l}\text{mind. ein Produkt aus Kategorie}t \text{ wurde von} \\ \text{Kunden } k \text{ gekauft oder angeschaut}\end{array} \\ 0 & : \text{sonst} \end{cases}$$

Diese Repräsentation lieferte die besten Ergebnisse. Das Einsetzen und die Ergebnisse der einzelnen Lernverfahren sind in Abschnitten 3.4.4, 3.4.5 und 3.4.6 dargestellt.

3.4 Einsetzen der Lernverfahren

3.4.1 Mittelwertverfahren

Um die Güte der berechneten Vorschläge zu bewerten, braucht man ein Vergleichsmaß, anhand dessen man die Leistungsfähigkeit des untersuchten Verfahrens an bestimmten Kriterien mißt. Es ist zwar erfreulich, wenn die Kunden die berechneten Vorschläge besser finden, als einen Vorschlag, der aus zufällig ausgewählten Produkten besteht, aber es ist keine große Herausforderung. Für diese Diplomarbeit wurde deshalb ein zweites Maß verwendet. Das ist der arithmetische Rating-Mittelwert für jedes Objekt über alle Benutzer.

Im Fokus der Betrachtungen steht jeweils der „aktive Kunde“, k^*. Es handelt sich um den Benutzer, für den im konkreten Fall Empfehlungen zu erstellen sind. Formal geschieht die Prognose der Interesse f_{k^*p} des aktiven Benutzers k^* für ein Produkt $p \in P$ dabei auf Basis des arithmetischen Mittelwertes $\bar{u_p}$ für dieses Produkt über alle Benutzer $k \in K$, die das Produkt gekauft oder ihr Interesse für dieses Produkt gezeigt haben.

Es sei wie oben $a_{k,p} = 1$, wenn Kunde k Produkt p gekauft bzw. sein Interesse gezeigt hat und damit ein Wert für $a_{k,p}$ existiert; sonst ist $a_{k,p} = 0$:

$$f_{kp} = \frac{\sum_{k \in K} a_{k,p}}{|K|} = \bar{u_p}$$

Offensichtlich liefert f_{kp} für alle aktiven Benutzer identische Werte die zwischen 0 und 1 liegen. Bei dem Verfahren handelt es sich damit um eine nichtindividualisierte Prognose. Für jedes Objekt läßt sich ein Wert errechnen [Runte, 2000].

Als eine konkrete Empfehlung kann man die Produkte auswählen, deren Rating höher als 0,5 liegt. Dabei werden die Produkte empfohlen, für die angenommen wird, daß sie mit einer Wahrscheinlichkeit von höher als 50% von diesem Kunden gekauft werden.

Mittelwertverfahren wird in der Literatur zu Collaborative Filtering an verschiedenen Stellen als Vergleichsmaßstab angewendet (z. B. [Shardanand, 1994]). M. Runte ist der Meinung, daß ein Individualisierungsverfahren nur dann Sinn macht, wenn es bessere Ergebnisse liefern kann als Mittelwertverfahren [Runte, 2000]. Und tatsächlich, falls die Ergebnisse der Lernverfahren schlechter sind als die vom Mittelwertverfahren gelieferten, so könnte man sich die Mühe sparen und das Mittelwertverfahren benutzen.

Es gibt jedoch auch Literaturquellen, die geringere Maßstäbe anlegen. Beispielsweise vergleicht Breese die von ihm getesteten Verfahren mit zufälligen Prognosen [Breese, 1998]. Weitere Quellen verzichten vollständig auf diese Überprüfung [Gupta et al., 1999].

3.4.2 Bewertungskriterien

Um die Ergebnisse von den drei eingesetzten Lernverfahren bereits auf diesem Stand der Arbeit zu bewerten, wurden für sie und für das Mittelwertverfahren die Fehlerquote (F) und die Konfidenzmatrix (siehe Abbildung 3.6) mit den Accuracy- (A), Recall- (R) und Precision-Werten (P), sowie die Standardabweichungen (S_f, S_a, S_r, S_p) berechnet.

In der Abbildung 3.6 bedeutet

- P_+: die Anzahl der korrekt vorhergesagten positiven Einträge (echter Eintrag – „1“, vorhergesagter Eintrag „1“)

Tabelle 3.8: Fehlerberechnung

	Wasser	Bier	Saft	Obst	Gemüse	Wein	Brot	Tee	Milch
Anton	0	1	1	0	0	1	1	1	0
Vorschlag	1	1	0	1	0	0	1	1	1
Fehler	f	r	f	f	r	f	r	r	f
Anzahl der Fehler:									5
Fehlerquote:									56 %

- P_-: die Anzahl der falsch vorhergesagten negativen Einträge (echter Eintrag – „1“, vorhergesagter Eintrag „0“)
- N_+: die Anzahl der falsch vorhergesagten negativen Einträge (echter Eintrag – „0“, vorhergesagter Eintrag „1“)
- N_-: die Anzahl der korrekt vorhergesagten negativen Einträge (echter Eintrag – „0“, vorhergesagter Eintrag „0“)

Die Fehlerquote wird berechnet, indem ein generierter Vorschlag mit jedem Einkauf eines jeden Kunden verglichen wird, die Anzahl aller falschen Vorschläge addiert wird, und die Prozentzahl daraus berechnet wird. Sein Beispiel dazu siehe in der Tabelle 3.8

Die Accuracy-Werte wurden berechnet, indem die Anzahl korrekter Vorschläge (P_+ und N_-) durch die Anzahl aller Vorschläge geteilt wurde. (Als „Vorschläge“ sind hier die einzelnen als Empfehlung vorgeschlagenen Produkte gemeint.) Die Recall-Werte geben Auskunft darüber, wie groß der Anteil der richtig vorhergesagten positiven Einträgen im Vergleich zu allen wirklich positiven ist, und Precision sagt aus, wie groß die Anzahl korrekterweise als positiv klassifizierten Einträge im Vergleich zu der Anzahl aller als positiv klassifizierten ist. Wenn man also die Konfidenzmatrix ansieht, werden die Werte A, R und P folgendermaßen berechnet:

$$A = \frac{P_+ + N_-}{P_+ + P_- + N_+ + N_-}$$

$$R = \frac{P_+}{P_+ + P_-}$$

$$P = \frac{P_+}{P_+ + N_+}$$

Für jedes Verfahren werden diese Werte für jeden Kunden berechnet. Dann wird für jedes Wert der Durchschnitt und die Standardabweichung berechnet. Für das Mittelwertverfahren sind diese Werte in der Tabelle 3.9 zu sehen.

Tabelle 3.9: Werte nach Mittelwertverfahren

$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
10,0 ± 5,4	75,0 ± 13,1	52,3 ± 37,2	19,5 ± 18,1

3.4.3 Kepler

Da die für die Datenanalyse verwendeten Lernverfahren in das System Kepler eingebunden sind möchte ich hier eine kurze Übersicht über dieses System geben.

Laut seiner Hersteller ist Kepler ein erweiterbares Softwaresystem, daß die Benutzer bei der optimalen Analyse der Datenbestände unterstützt. Erweiterbarkeit bedeutet, daß Kepler mit den Anforderungen mitwachsen kann: Erfordern die neuen Fragestellungen andere Analyseverfahren, so kann man diese leicht in Kepler einbinden. Kepler gibt Unterstützung bei der Erzeugung, Interpretation und Organisation von Datenanalyseprojekten. Er enthält eine einheitliche Schnittstelle und grafische Oberfläche zu verschiedenen Datenquellen und einer Großzahl von Analysealgorithmen. Damit können sich die Benutzer auf das zu bearbeitenden Analyse- bzw. Data Mining Projekte konzentrieren, anstatt sich mit komplexen Benutzerschnittstellen auseinandersetzen zu müssen.

Kepler wurde am GMD – Forschungszentrum für Informationstechnik in Sankt Augustin bei Bonn von der Forschungsgruppe Maschinelles Lernen entwickelt. Die Erfahrungen, die diese Gruppe gesammelt hatte zeigten, daß man zur optimalen Analyse eines Datensatzes eine Vielzahl von Analysealgorithmen heranziehen mußte.

Leider war es damals noch notwendig, die Daten für einen jeden Analysealgorithmus erneut aufzubereiten, da die Ein- und Ausgabeformate oft sehr unterschiedlich waren. Zusätzlich gab es häufig das Problem, daß die Benutzer nach einiger Zeit nicht mehr wußten, mit welchen Parametereinstellungen sie bestimmte Ergebnisse erzielt hatten bzw. wie man ein bestimmtes Analyseverfahren überhaupt benutzt. Daher haben die Entwickler von Kepler großen Wert darauf gelegt, daß das System erweiterbar ist, und daß man jederzeit seine Analyseschritte nachvollziehen kann.

Das Kernsystem von Kepler stellt grundlegende Funktionalitäten bereit, die unabhängig von bestimmten Analyseverfahren bei jeder Analyse benötigt werden, wie z.B. der Import und Export von Daten.

Nachdem ein Datensatz geladen und unter Umständen genauer betrachtet und aufbereitet wurde, kann er analysiert werden. Dies ist das eigentliche Data Mining.

Eine Besonderheit von Kepler liegt in der Fähigkeit zu multirelationalen Analysen. Diese Verfahren sind nicht darauf angewiesen, daß alle Informationen in einer einzigen Datentabelle vorliegen, sondern analysieren auch Informationen, die auf verschiedene Relationen verteilt sind.

Beispielsweise könnte die Kundennummer aus einer Bestelldatei auf die auch in der Kundendatei auftauchende Kundennummer verweisen. Bei einer multi-

Tabelle 3.10: Eine Übersicht der im Kepler integrierten Verfahren

Verfahren	Module
Generierung von Entscheidungsbäumen	C4.5, C5.0, Tilde, DTI
Generierung von Clustern	AutoClass, k-means, hierarchisches Clustering
Generierung von multidimensionalen Rechtecken	NGE, BNGE
Regression	MARS ; RT, M5'
Nächste Nachbarn-Verfahren	KNN
Regellerner zur Erstellung kompakter Datenmodelle	Foil, C4.5-Rules, C5.0-Rules, Progol, Claudien
Automatische Suche nach auffälligen Untergruppen	Midos (Exklusiv im Kepler eingebundener und patentierter Algorithmus)
Assoziations-Regellerner	Apriori, Apriori TID, Hybrid TID

relationalen Analyse der Bestelldatei ließe sich auf dieses Hintergrundwissen verweisen, was dazu führt, daß auch die Informationen aus der Kundendatei berücksichtigt werden.

Leider hat sich die Integration von apriori-Algorithmus im Kepler als fehlerhaft erwiesen, so daß ich dieses Algorithmus wie im Abschnitt 2.2.1 beschrieben selbst implementiert habe.

3.4.4 Apriori

Für die beiden Arten der Darstellung (Kunden- und Kategorievektoren) wurden mehrere Versuche mit verschiedenen Support- und k-Werten durchgefürt.(k ist die minimale Länge des LIS. Für das weitere Vorgehen reichte aber nur eine Ergebnis, die die besten Aussichten auf Erfolg hatte, d.h. die Vorschläge, die die für die Gruppen empfohlenen Artikel, würden bei den Kunden die größtmögliche Kaufbereitschaft finden.

Der Vergleich der Ergebnisse erfolgte, indem für jede Kundengruppe ein Vorschlag generiert wurde. Dieser bestand aus den Produkten, die von über 50% der Kunden aus dieser Gruppe gekauft wurden. Danach wurde dieser Vorschlag mit den einzelnen Einkäufen der Kunden verglichen und die Fehlerquote (F), die Accuracy- (A), Recall- (R) und Precision-Werte (P) sowie die Standardabweichungen (S_f, S_a, S_r, S_p) berechnet.

Diese Werte werden in Tabellen 3.12 und 3.13 dargestellt. Außer den o.g. Werten mußte man auch die Tatsache beachten, daß nicht zu wenige Kunden mit

Tabelle 3.11: Anzahl der Regeln und Kunden in Abhängigkeit von Support- und k-Werten

Kundenvektoren:

Regeln	Kunden	Support	k_{min}
35	31	0,21-0,20	3
8	21	0,19-0,18	4
35	25	0,17-0,16	4
11	17	0,15-0,14	5
2	8	0,13-0,12	6
31	20	0,11-0,10	6
33	15	0,09-0,08	7

Kategorievektoren:

Regeln	Kunden	Support	k_{min}
5	7	0,26-0,27	3
35	14	0,25	3
4	6	0,23	4
27	12	0,22	4
4	7	0,2	5
21	**15**	**0,18-0,17**	**5**
38	20	0,14-0,16	6
12	18	0,13-0,12	8
54	20	0,11	10

Tabelle 3.12: Werte aus Kundenvektoren

# Regeln	# Kunden	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
35	31	10,3 ± 0,5	76,3 ± 1,1	75,5 ± 4,1	65,1 ± 4,3
8	21	10,3 ± 0,4	77,5 ± 1,3	75,3 ± 1,9	69,3 ± 2,1
35	25	9,9 ± 0,8	77,1 ± 3,2	77,4 ± 3,2	68,2 ± 4,2
11	17	9,5 ± 0,4	78,8 ± 1,8	75,7 ± 1,4	75,5 ± 2,3
2	8	7,8 ± 0,08	82,1 ± 0,5	85,8 ± 0,3	72,4 ± 0,7
31	20	7,8 ± 0,8	82,5 ± 1,8	79,2 ± 1,7	80,6 ± 2,8
33	15	6,9 ± 0,8	84,0 ± 1,8	88,7 ± 1,3	75,7 ± 2,6

Tabelle 3.13: Werte aus Kategorievektoren

# Regeln	# Kunden	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
4	7	5,6 ±0,2	87,0 ± 0,5	92,1 ±0,4	81,3 ±0,8
35	14	6,9 ± 1,3	84,6 ± 1,3	86,2 ±1,5	83,1 ±1,7
4	6	6,9 ±0,3	84,4 ± 0,3	80,9 ±0,7	88,5 ±0,6
27	12	7,3 ± 1,3	84,6 ± 1,0	80,1 ±1,6	86,2 ±1,9
4	7	6,9 ±0,2	84,0 ± 0,7	89,1 ±0,5	75,5 ±0,8
21	**15**	**6,9 ± 0,4**	**84,7±1,2**	**86,2±1,1**	**77,5±2,1**
38	20	8,2 ± 1,3	81,0 ± 1,7	78,0 ± 2,1	80,1 ±2,8
12	18	9,4 ± 0,4	78,3 ±15	75,6 ±1,5	76,4 ±2,2
54	20	9,9 ± 0,9	77,6 ± 30,4	74,2 ±4,7	70,3 ± 1,9

den Regeln abgedeckt werden sollen, weil sonst die Testdurchführung schwierig wurde, und auch der praktische Nutzen mit der Verringerung der Kundenzahl, für die die Regeln angewendet werden können, um so mehr verschwindet.

Nach dem alle o.g. Kriterien verglichen und bewertet wurden, wurde die fettgedruckte Zeile ausgewählt und weiter für den Test benutzt.

3.4.5 Midos

Wie man in der Tabelle 3.14 sehen kann, sind die Fehlerwerte nicht so gut, wie bei apriori. Die Werte für Accuracy, Recall und Precision sind insgesamt etwas schlechter als bei apriory, aber durchaus noch akzeptabel.

Weniger gut fallen die Standardabweichungswerte für Accuracy, Recall und Precision aus, obwohl sie mit steigender Abdeckung insgesamt besser werden.

Für weitere Arbeit wählte ich die letzte Zeile der Tabelle 3.14, da dieses Ergebnis die besten Werte in allen Spalten hat.

Tabelle 3.14: Midos–Werte

Abdeckung, %	# Kunden	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
50%	43	9,9 ±3	77,1 ± 25,3	73,2 ±42,0	70 ± 42,5
75%	42	9,9 ± 2,6	78,0 ± 23,6	73,1 ±39,9	73 ± 41,7
85%	42	9,4 ± 0,9	78,1 ± 19,9	73,5 ±32,3	68 ±36,5
95%	**35**	**9,4 ±0,5**	**78,1±9,5**	**74,4±16,3**	**73,0±16,6**

Tabelle 3.15: C4.5–Werte

c	# Kunden	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
25	52	21 ±7	78,4 ±10,2	70,5 ±21,8	66,4 ±21,5
20	**47**	**21±5**	**77,6±10,0**	**71,3± 18,1**	**69,6±19,2**
5	47	21 ± 6	78,5 ± 11,1	58,9 ± 38,7	33,4 ± 32,2

3.4.6 C4.5

C4.5 gibt die Möglichkeit den Entscheidungsbaum (und die Regeln) kompakter zu machen, d.h. abzuschneiden. Je kleiner der Parameter c gesetzt wird, desto stärker werden die Bäume abgeschnitten. Das füht dazu, daß die Regeln und die Bäume schwächer klassifiziert werden. Obwohl das auf den ersten Blick zu mehr Fehlern führt, kann diese Tatsache beim späteren Testen Vorteile bringen.

Die Ergebnisse der Analyse sind in der Tabelle 3.15 zusammengefaßt. Auch hier sieht man, daß die Standardabweichungen sehr groß ist, obwohl die Durchschnittswerte besser als beim Mittelwertverfahren sind. Also habe ich mich für die Ergebnisse mit kleinsten Standardabweichungen entschieden, damit die beim nachfolgenden Testen gewonnenen Erkenntnisse mit größerer Wahrscheinlichkeit signifikant werden.

Allen drei angewandten Methoden (apriori, Midos, c4.5) ist gemein, daß die Fehler-, Accuracy-, Recall- und Precisionwerte in Vergleich zum Mittelwertverfahren bessere Ergebnisse liefern. Diese Erkenntnis lässt hoffen, daß die Ergebnisse des nachfolgenden Testens der Sinn und Zweckmäßigkeit einer solchen Generierung der Empfehlungen bestätigen.

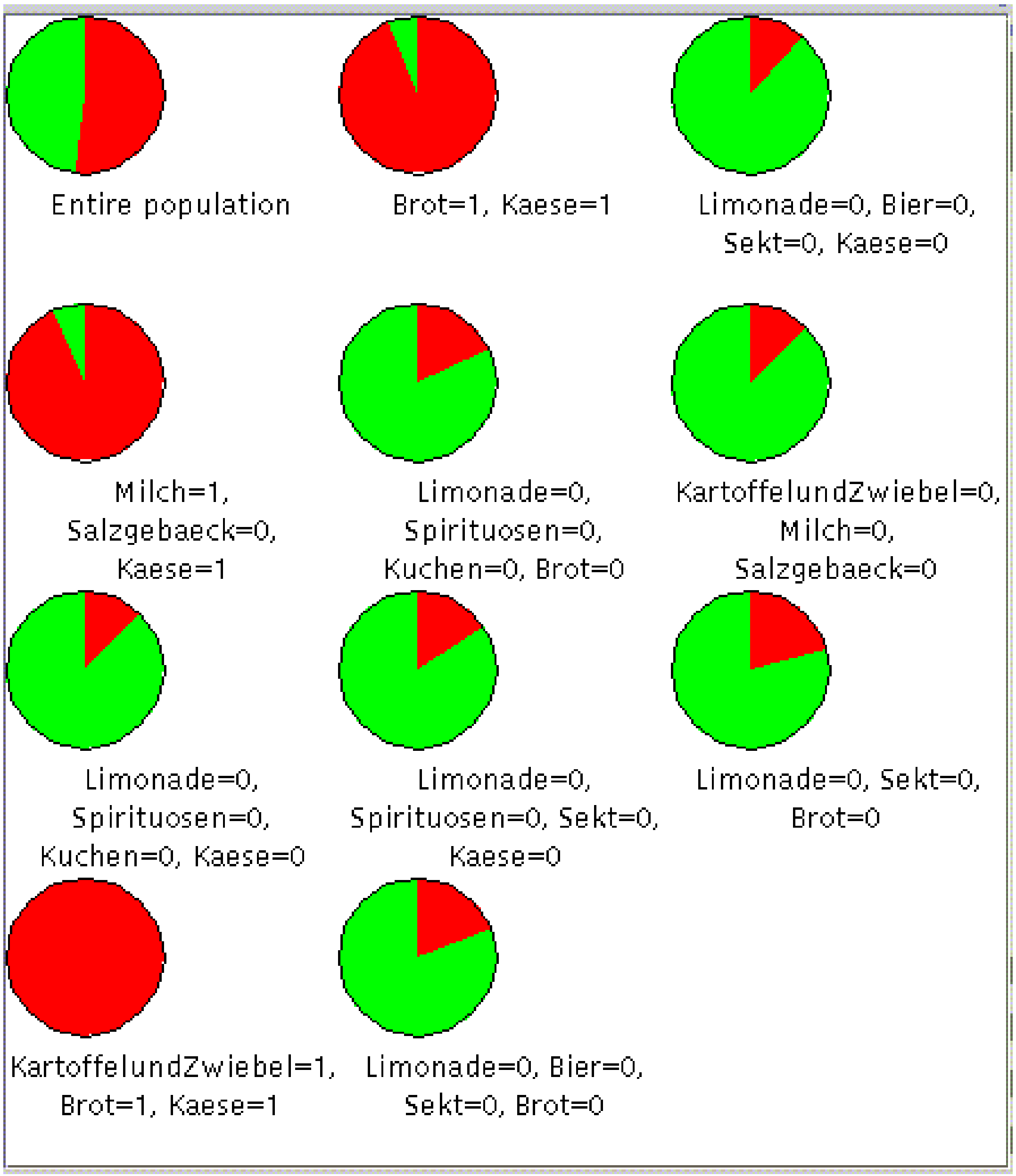

Abb. 3.4: Midos-Subgruppen

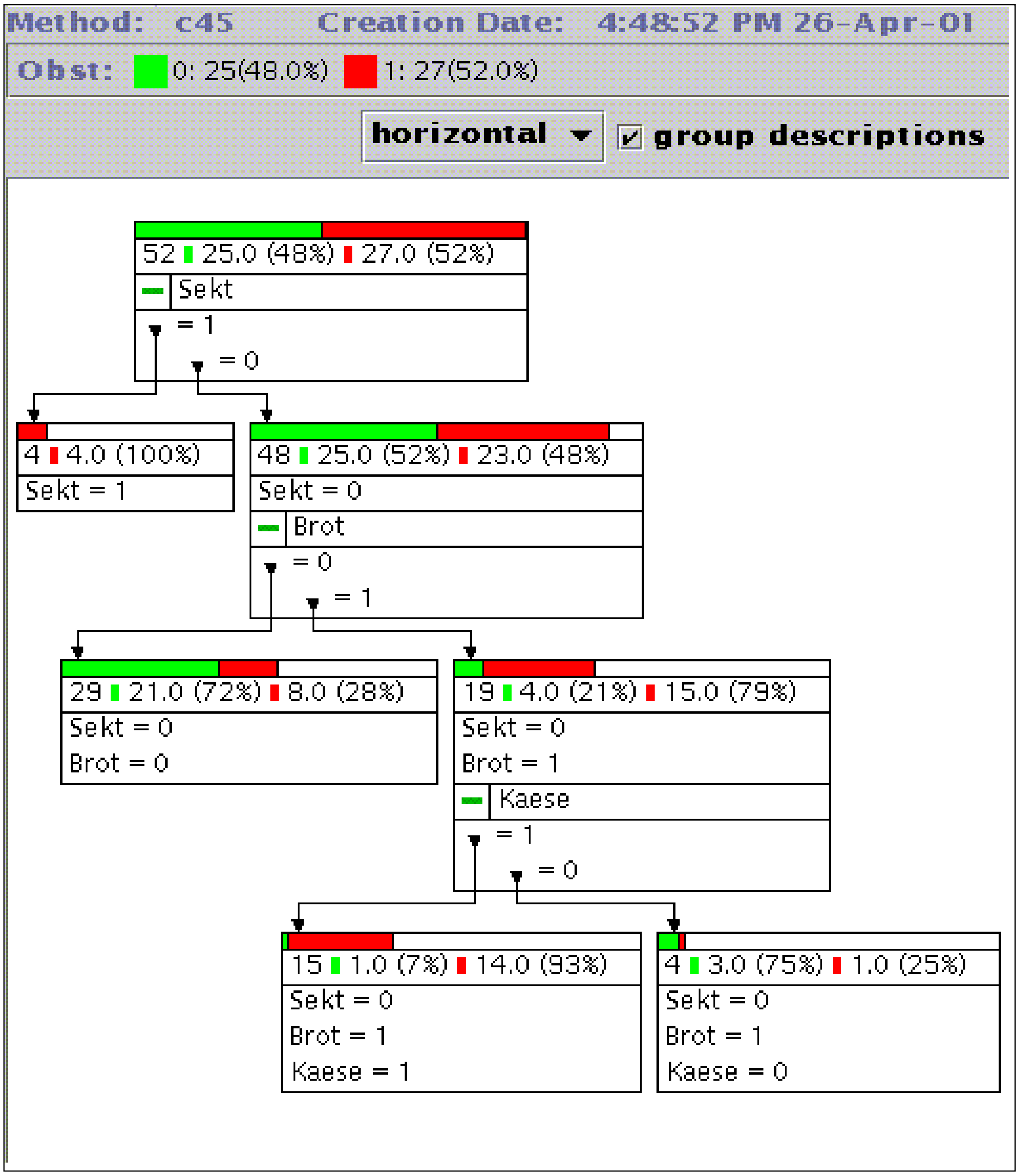

Abb. 3.5: Entscheidungsbaum für „Obst = 1“

		echte Klasse	
		1	0
vorhergesagte	1	P_+	N_+
Klasse	0	P_-	N_-

Abb. 3.6: Konfidenzmatrix

Kapitel 4

Testdurchführung

4.1 Ziele

Nachdem die Ergebnisse von allen drei Lernverfahren (apriori, Midos und c4.5) gesammelt wurden, mußten sie getestet werden.

Das erste Ziel des Testens war die Feststellung, ob die mit den drei Lernverfahren gefundenen Ergebnisse besser sind als die zwei Vergleichsverfahren. Um dies zu beweisen reicht zu zeigen, daß die meisten Produkte (z.B. mehr als die Hälfte) mit Hilfe eines der Vorschläge gemacht wurden, und das der Anteil der über Vergleichsverfahren gekauften Produkte verhältnismäßig klein ist.

Ferner war festzustellen, ob eines oder mehrere von den Lernverfahren generierte Vorschläge mehr benutzt wurden als der Shop selbst, d.h. hier stellt sich die Frage, ob diese Art der Personalisierung sich überhaupt lohnt. Wenn der Kunde z.B. einen von einem der benutzten Lernverfahren generierten Vorschlag zwar mehr benutzt als die Vergleichsvorschläge, das Meiste aber doch im Shop kauft, ist es vielleicht nicht wirklich gewinnbringend, diese Verfahren einzusetzen.

Als drittes wurde das subjektive Empfinden der Kunden abgefragt. Sie hatten die Möglichkeit, das ihnen am besten zusagenden Vorschlag auszuwählen. Die dritte Ziel wird erreicht, wenn die Kunden das Gefühl haben, das eins der Vorschläge besonders gut war, und zwar genau der, mit deren Hilfe die Kunden die meisten Produkte gekauft haben. Anders gesagt sollte der ausgezeichnete Vorschlag die, für den Kunden interessanten Produkte enthalten.

4.2 Testvorbereitung

Bei einem realen Betrieb sollte eine derart personalisierte Seite so aussehen, daß an der Shop-Einstiegsseite ein Vorschlag und der Shop selbst erscheinen. Der Kunde sollte sich von den Vorschlägen nicht gestört fühlen, sie sollten zwar sichtbar sein, aber das Vorgehen des Kunden nicht stören oder sogar hindern. Noch besser wäre, wenn der Kunde den Vorschlag immer vor Augen hat, während er sich innerhalb des Shops befindet. Z.B. erscheinen die vorgeschlagenen Produkte immer in oberen Teil der Seite, es bleibt für den Shop selbs aber ausreichend Platz.

Tabelle 4.1: Beispiel der Testdarstellung in der Datenbank

Kunde	Produkt	Einkaufsgang Nr.	Methode
Achim	Apfel	3	apriori
Achim	Apfel	3	midos
Anna	Orange	2	midos
Anna	Orange	2	mittelwert
Alla	Apfel	5	random
Alla	Apfel	5	mittelwert
Alla	Milch	5	random
Alla	Wurst	5	c4.5
Alla	Joghurt	6	c4.5
Achim	Brot	4	apriori
Achim	Brot	4	midos
Achim	Brot	4	c4.5
Achim	Margarine	4	random
...			

Also mußte bei diesem Art des Testens der Test 5 mal durchgefürt werden.

Da ich aber die Testbereitschaft der Benutzer nicht überstrapazieren wollte, kam für mich diese Art des Testens nicht im Frage. Alle 5 Vorschläge mußten auf einmal getestet werden. Aus Platzgrunden ist jeder Vorschlag eine Liste von Produkten. Die Aufbau der Seite ist in der Abbildung 4.1 gezeigt.

Dabei kann der Benutzer die Produkte, die er einzukaufen beabsichtigt, in einer der Listen auswählen. Die Möglichkeit des „nicht Einkaufens“ wie sie in dem Shop selbst gegeben war, gibt es hier nicht. Ein einmal gewählter (angeklickter) Produkt gilt sofort als gekauft.

Falls ein Produkt nicht nur in einer Liste vorkommt, spielt es für den Kunden keine Rolle, in welcher der Listen er die Produkte anklickt. Ein Zugriff auf ein solches Produkt führt dazu, daß ein Einkauf dieses Produktes in allen Produktlisten, in denen es vorkommt registriert wird. Die Information wird direkt zur Datenbank übertragen. In einer Tabelle wird der Kunden ID, Produktname, der Nummer des Einkaufsganges und die Methode, mit der die Liste erstellt worden ist (apriori, Midos, c4.5, Zufallsauswahl oder Mittelwertverfahren), gespeichert z.B. wie in der Tabelle 4.1.

Falls die Produkte, die ein Kunde kaufen möchte, in keiner der Listen vorkommen, so kann dieser Kunde sie in dem Shop wie gewönlich kaufen. Für die Kunden funktioniert das Einkaufen wie beim ersten Mal, die Verkaufsdaten werden aber nicht wie beim ersten Mal in die Click-Stream Tabelle eingetragen, sondern für sie ist eine separate Datenbank-Tabelle angelegt worden.

Tabelle 4.2: Statistik über drei Benutzergruppen

Nr.	Anzahl Vorschläge	Anzahl Benutzer	Anzahl Einkäufe
Gruppe 1	5	13	20
Gruppe 2	4	16	16
Gruppe 3	3	11	12

Diese Änderung wurde aus folgenden Gründen gemacht: Die Click-Streams werden benutzt, um die für jede Kundengruppe meistgekauften Produkte auszuwählen. Falls aber auch die neuen Einkäufe in dieselbe Tabelle eingetragen werden, werden sie in den nächsten Einkaufsgängen auch in die Berechnung der Vorschläge eingezogen. Dann ist die Chancengleichheit für alle Benutzer aber nicht gewährt. Aus derselben Grund werden auch die Einkäufe, die über die Listen mit Vorschlägen gamacht wurden, separat gespeichert und nicht in den Click-Streams.

4.3 Testdurchführung

Von 52 Testkunden, die zum Zeitpunkt des Tests bereits einige Produkte im Lebensmittelshop eingekauft haben, haben 40 am Test teilgenommen. Von den 12 Kunden, die nicht am Test teilgenommen haben, konnten für 5 keine „echten“ Vorschläge generiert werden, da sie nur wenige Produkte bei dem ersten Durchlauf gekauft haben. Die anderen 7 waren entweder nicht zu finden (umgezogen, nicht zu erreichen) oder sie konnten die für das Testen nötige Zeit nicht aufbringen.

48 Eikaufsgänge wurden gemacht. Bei jedem Einkaufsgang wurden durchschnittlich 15 verschiedene Produkte eingekauft.

Da die Kunden verschiedene Anzahl von Vorschlägen testen konnten, wurden sie zunächst in drei Gruppen eingeteilt. Die erste Gruppe besteht aus den Benutzern, für die alle fünf Vorschläge generiert werden konnten (apriori, Midos, c4.5, Mittelwertverfahren und Zufallsauswahl). Für die zweite Gruppe gab es vier Vorschläge (apriori konnte für diese Benutzer keine Regeln finden). Und die dritte Gruppe bestand aus den Benutzern, die nur 3 Vorschläge bekamen. Für diese Benutzer konnte apriori keine Regeln finden bzw. diese Benutzer paßten in keine von Midos gefundenen Gruppen. Die Einteilung der Benutzer und die Zahl der Einkäufe sind in der Tabelle 4.2 dargestellt.

4.4 Testauswertung

4.4.1 Ergebnisse mit apriori

18 Kunden haben einige Produkte in dem von apriori generierten Vorschlag gekauft. Zwei anderen Kunden haben alle Produkte über das Lebensmittelshop

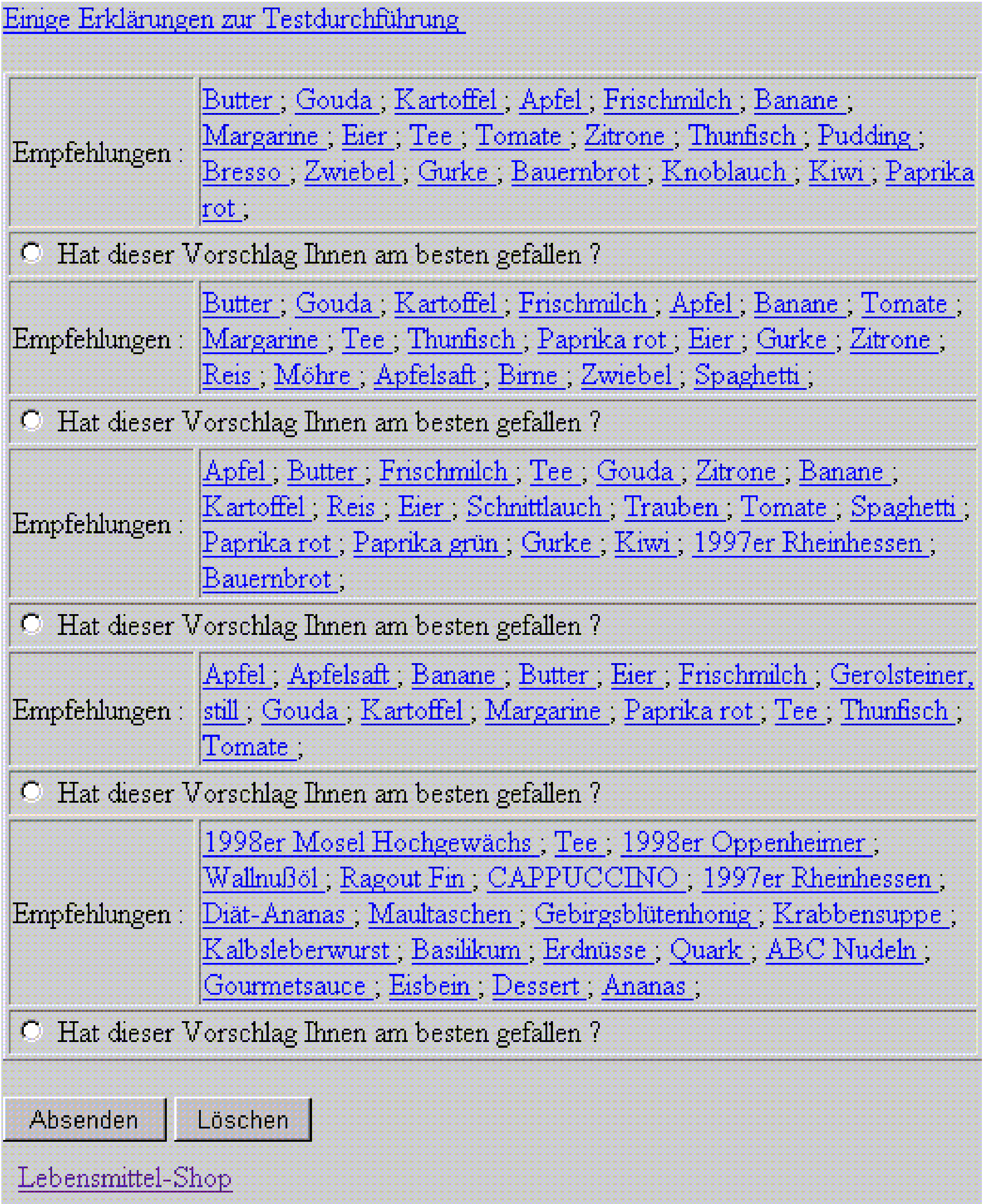

Abb. 4.1: Testseite

gekauft, keiner von dem Vorschlägen wurde von diesen Kunden benutzt. Wenn man aber die von diesen beiden Kunden gekauften Produkte genauer betrachtet, kann man feststellen, das einige davon auch in den apriori-Vorschlägen vorkamen. Das diese Kunden den Vorschlag nicht benutzt haben, kann bedeuten, daß sie die Produkte in den Vorschlägen einfach nicht bemerkt haben, oder sie wollten tatsächlich keine Vorschläge nutzen. Da aber das Nutzen der Vorschläge unter anderen einen Zeitersparnis bedeutet (Kauft man im Shop, so mus man zuerst die richtige Produktkategorie finden, dann Produkt auswählen und dann bestätigen. Nutzt man die Vorschläge, so ist ein Produktkauf ein einfacher Klick), kann ich vorschtellen, daß der Grung eher der Erstegenannte ist. Deswegen wurden diese beiden Einkäufe hier auch berücksichtigt.

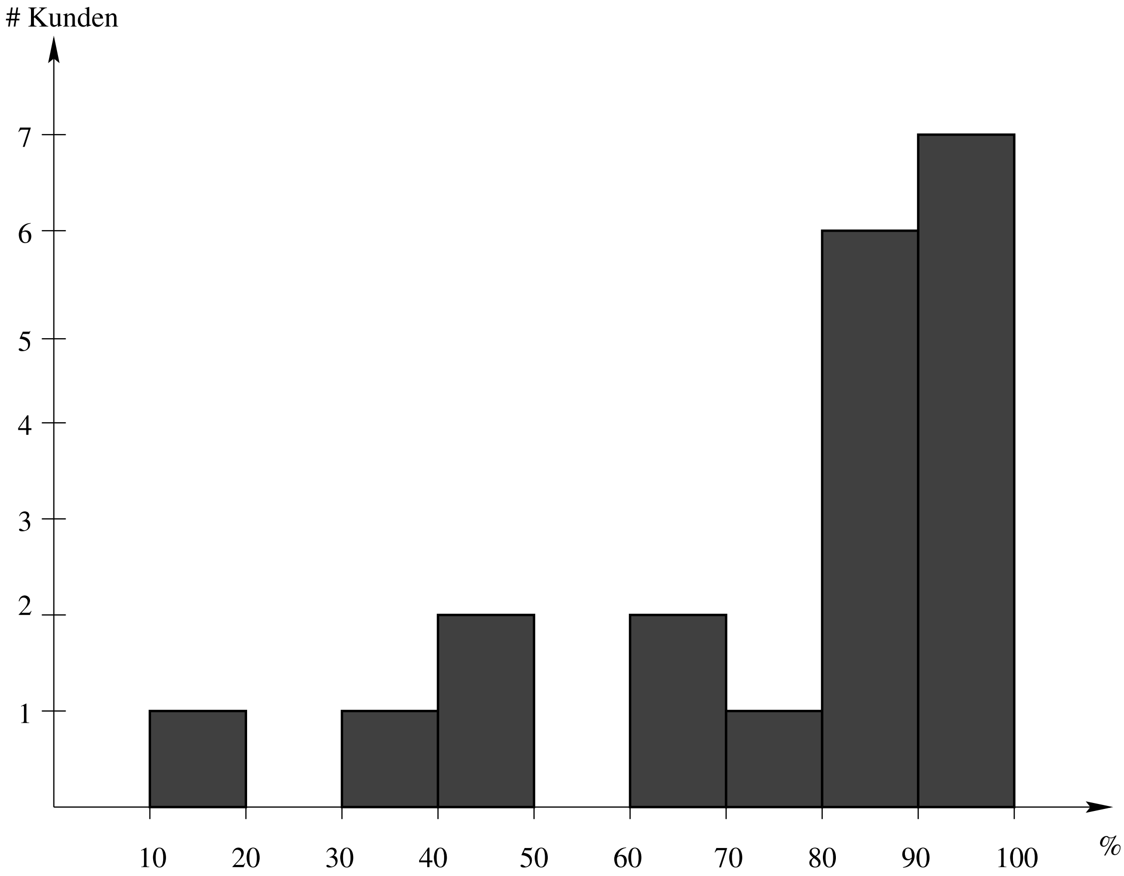

Abb. 4.2: Verteilung der „über apriori“ gekauften Produkte

Wenn man die Anzahl der Produkte, die mit Hilfe von über apriori generierten Vorschlag gekauft wurden, analysiert, sieht man, daß in 15 Einkäufen (75%) meht als die Hälfte aller Produkte über diesen Vorschlag gekauft wurden. In 45% der Einkäufe wurden sogar über 80% der Produkte „bei apriori“ gekauft (siehe Abbildung 4.2).

Tabelle 4.3: Werte der apriori-Gruppe

# Eunkäufe	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
20	$4,2 \pm 1,6$	$95,9 \pm 1,7$	$93,0 \pm 13,2$	$54,0 \pm 16,3$

Betrachtet man die Fehler-, Accuracy-, Recall- und Prezisionwerte und deren Standartabweichungen (Tabelle 4.3), so kann man feststellen, daß die Fehlerwerte relativ klein sind (durchschnittlich knapp 10 oder 4,2% bei 233 Produkte). Die Accuracy- und Recallwerte liegen bei über 90%. Nur die Prezisionwerte sind nicht so gut asgefallen. Wenn wir diese Werte mit den Werten aus Abschnitt 3.4.4, Tabelle 3.13 vergleichen, so kann man feststellen, daß nur der Prezisionwert sich verschlechtert hat. Diese Verschlechterung stammt daher, daß die Werte im Abschnitt 3.4.4 mit derselben Datenmenge berechnet wurden, mit denen das Algorithmus gearbeitet hat, also mit der Trenningsmenge. Deshalb waren die Prezisionwerte besser.

Werden die, von diesen Kunden als „beste“ markierten Vorschläge betrachtet, so kann festgestellt werden, daß die Kunden für 14 aus 18 tatsächlich überwiegend mit diesem Vorschlag gemachten Einkäufen apriori als bester Vorschlag gewählt haben. In allen diesen Einkäufen wurden auch die meisten Produkte mit Hilfe von apriori gekauft. Bei 4 Einkäufen wurde Midos und einmal wurde die Zufallsauswahl als bester Vorschlag markiert. Bei einem Einkauf wurden keine Eingaben gemacht.

In der Abbildung 4.3 sind diese Angaben veranschaulicht. Dabei steht in dieser Abbildung (und auch in den Abbildungen 4.5 für Midos und 4.7 für c4.5)die dunkle Farbe für die für das tatsächliche Verhalten von Kunden. Die weiße gemusterte Rechtecke darstellen die Angaben von Kunden. Gibt es eine übereinstimmung von Tatsache und Angeben so werden die entsprechenden Rechtecke dunkel und gemustert.

4.4.2 Ergebnisse mit Midos

In 34 Einkäufen wurden die Vorschläge, die für die von Midos generierten Subgruppen generiert wurden, benutzt. Die Statistik ist in der Abbildung 4.4 dargestellt.

Hier wurden in 29 Fällen über 50% der Produkte über den entsprechenden Vorschlag gekauft. In 17 Einkäufen wurden mehr als 70% Produkte mit den Vorschlägen gekauft und in 9 Fällen sogar mehr als 80%.

Die Analyse der Fehler-, Accuracy-, Recall- und Precision-Werten (siehe Tabelle 4.4) zeigt, daß die Fehler- und Precision-Werte etwas schlechter als bei apriori ausgefallen sind, Accuracy-Werte sind genauso gut und Recall-Werte

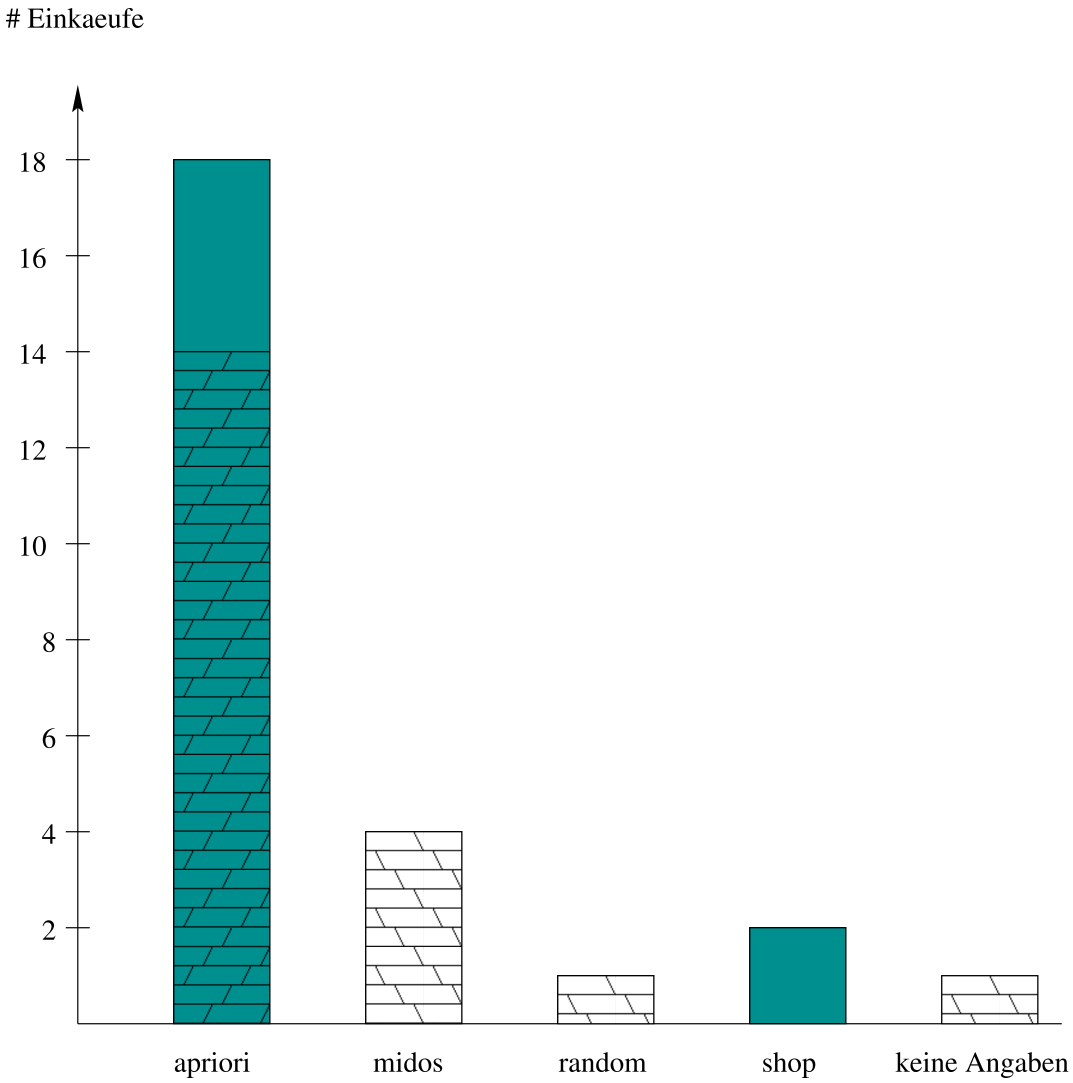

Abb. 4.3: Die besten bei apriori

sogar noch besser. Werden die Werte mit den in der Abschnitt 3.4.5, Tabelle 3.14 vergliechen, so sehen wir, wie bei apriori, daß alle Werte mit Ausnahme von Precision besser geworden sind.

7 Kunden haben Midos als den besten Vorschlag markiert, dabei stimmt es tatsächlich nur in zwei Fällen. In Wirklichkeit haben die Kunden c4.5 und apriori häufiger benutzt als den von Midos generierten Vorschlag, so daß in 14 Fällen apriori tatsächlich und nach Angaben der Kunden der beste Vorschlag war. C4.5 wurde in 12 Fällen als bestes Vorschlag erkannt, in Wirklichkeit war es das sogar 18 Mal.

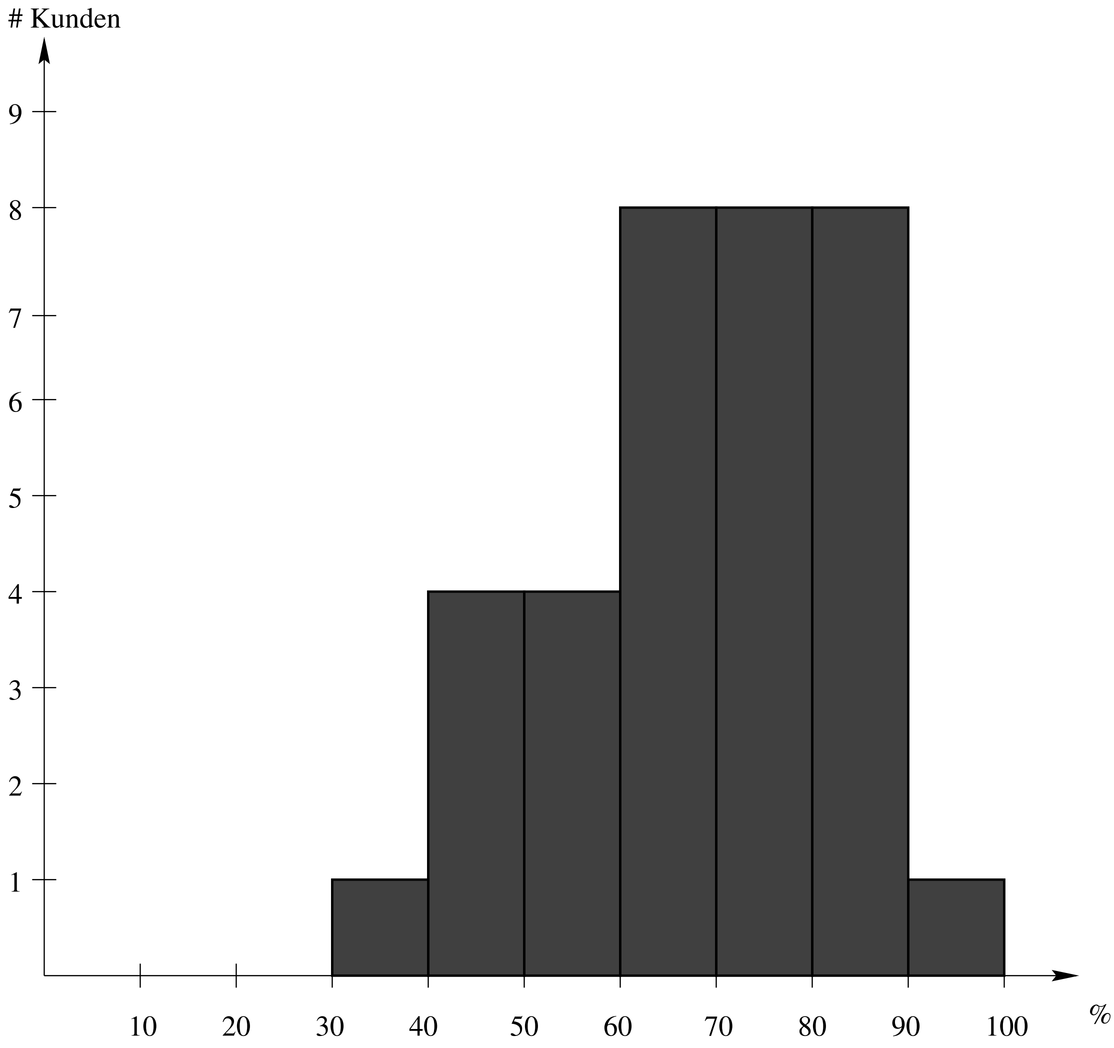

Abb. 4.4: Verteilung der „über Midos“ gekauften Produkte

4.4.3 Ergebnisse mit c4.5

Da mit Hilfe der c4.5 die meisten Kunde in Gruppen eingeteilt werden konnten, ist die Anzahl der Einkäufe, die mit Hilfe von diesem Lernverfahren gemacht wurden, am größten: 46. Die Verteilung der Produkte kann man in der Abbildung 4.6 sehen.

In 41 Einkäufen wurden über 50% der Produkte und in 20 Fällen über 80% mit diesem Vorschlag gekauft.

Wenn wir die Fehler-, Accuracy-, Recall- und Prezisionwerte mit den Werten aus dem Abschnitt 3.4.6, Tabelle 3.15 vergleichen, sehen wir, daß alle Werte sich verbessert haben. Alle Durchschtswerte sind sogar besser als es bei apriori und bei Midos der Fall ist.

Tabelle 4.4: Werte der Midos-Gruppe

# Einkäufe	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
34	$4,2 \pm 1,2$	$95,9 \pm 1,1$	$100,0$	$51,6 \pm 13,3$

Tabelle 4.5: Werte der c4.5-Gruppe

# Einkäufe	$F \pm S_f$, %	$A \pm S_a$, %	$R \pm S_r$, %	$P \pm S_p$, %
46	$3,8 \pm 1,6$	$96,2 \pm 1,6$	$100,0$	$56,2 \pm 19,1$

Betrachten wir die als beste ausgezeichnete Vorschläge, so sehen wir, daß nur bei 24 Einkäufen c4.5 tatsächlich und nach Angaben der beste Vorschlag war. 14 Kunden haben sich für apriori entschieden (haben auch alle das Meiste „bei apriori“ gekauft) und 7 wählten Midos. Einmal wurde das Mittelwertverfahren gewählt, der Kunde hat sich allerdings anders verhalten, und einmal wurden keine Angeben gemacht.

4.4.4 Vergleich der Ergebnisse

Wie bereits früher erwähnt, konnten einige Produkte in mehreren Produktlisten vorkommen. Es kann deshalb sein, das ein Kunde bestimmte Produkte mehrfach kauft auch wenn er dies nur einmal beabsichtigte. Die Mengen der Produkte, die ein Kunde mit Hilfe verschiedener Verfahren gekauft hatte, sind also nicht disjunkt. Diese Situation hat sich sogar als die Regel herausgestellt.

Es kann also sein, daß ein Kunde mit Hilfe z.B. von apriori 10 Produkte gekauft hat und mit Hilfe von Midos 8. Davon sind aber 6 Produkte bereits „über apriori“ gekauft worden. Also wurden „über Midos“ 2 zusätzliche Produkte gekauft und „über a priori“ 4.

Um die Ergebnisse von allen drei Lernverfahren zu vergleichen, habe ich also folgendermaßen vorgegangen: Für jeden Einkaufsgang wurden die Produkte, die in mehreren Vorschlägen vorkamen nicht betrachtet. Für die Bewertung sind nur die in jedem Vorschlag verschiedene Produkte wichtig.

Von den Kunden, die die Möglichkeit hatten „bei apriori“ einzukaufen, wurden in 18 aus 20 Einkaufsgängen die meisten Produkte mit Hilfe von apriori gemacht, nur bei 3 Einkäufen betrug die Anzahl der über apriori gekauften Produkte weniger als 50%. Hier kann man also feststellen, daß falls apriori für einen Kunden Regeln liefert, dann nutzen die Kunden diese Vorschläge auch gerne.

Von den Kunden, die außer Vergleichsvorschlägen Vorschläge von Midos und c4.5 bekamen, waren in 14 aus 16 aller Fälle die meisten Produkte mit Hilfe von c4.5 gekauft. Nur zwei Benutzer haben das Meiste über Midos gekauft.

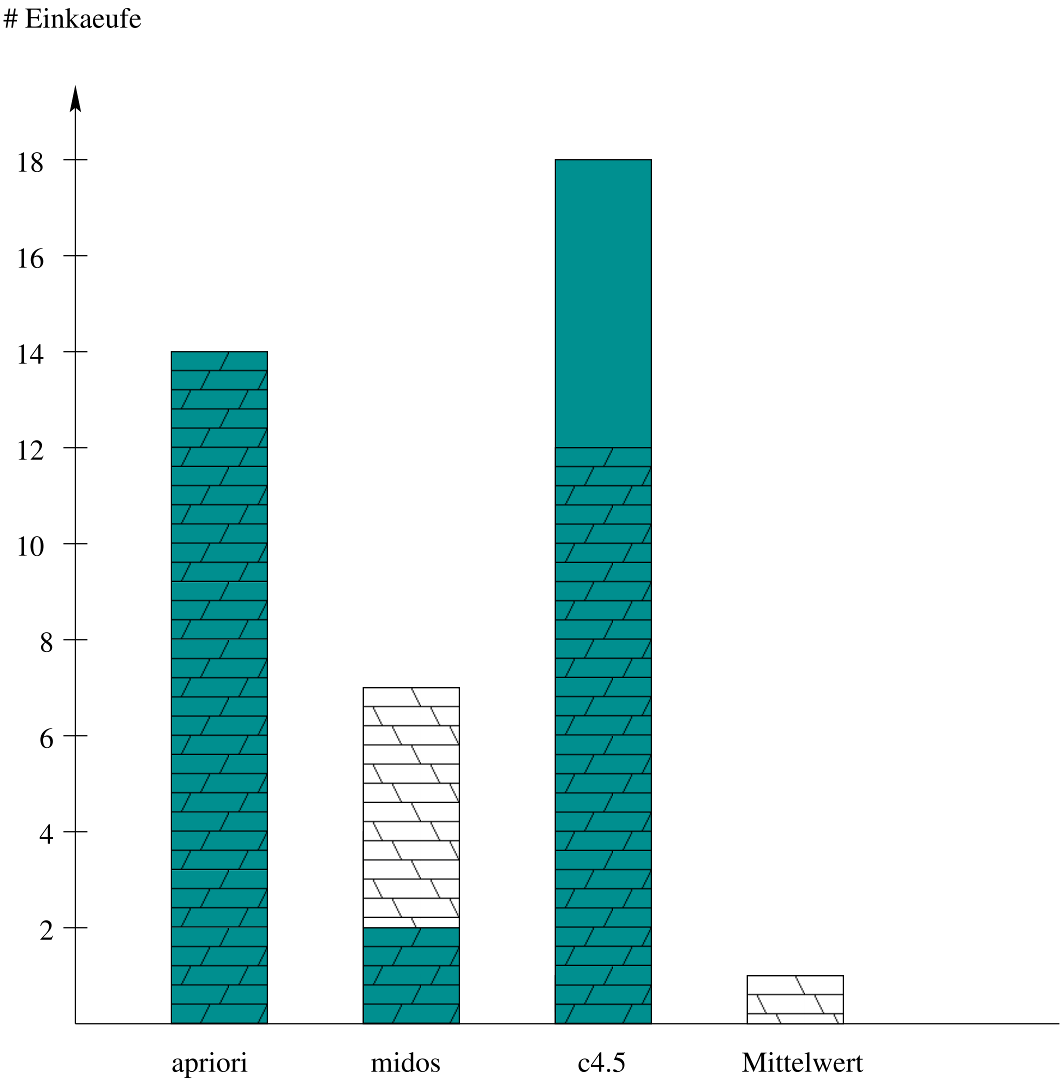

Abb. 4.5: Die besten bei Midos

Daraus kann man schließen, daß in diesem Fall c4.5 dem, was die Testpersonen auswählten am nächsten kommt.

Von den restlichen Kunden, die nur drei Vorschläge bekamen, kauften alle Kunden mit einer Ausnahme die meisten Produkte über c.45. Ein Kunde hat die meisten Produkte aus dem Zufallsauswahl gekauft.

Zusammenfassend kann man sagen, daß während des gesamten Tests die meisten Produkte über die mit Hilfe der drei Lernverfahren generierten Vorschläge gemacht worden sind.

Der große Mehrheit der Kunden haben die Vorschläge zum Einkaufen benutzt. Nur in drei Einkäufen (also in 6,25%) wurde das Meiste ohne Nutzung der Vorschläge gekauft.

Der Test hat deutlich gezeigt, daß die Kunden die Vorteile der Vorschläge bereit sind zu nutzen. Nur zwei Kunden haben es vorgezogen, im Lebensmittel-Shop zu kaufen. Alle anderen haben nur wenige Produkte, die nicht in den Vorschlägen angeboten wurden gekauft. Damit ist das erste und auch das zweite Ziele erreicht.

Die dritte Zielsetzung ist auch bewiesen, da die meisten Kunden sich im Klaren darüber waren, wie sie sich beim Einkaufen verhalten. Insgesamt wählten 38 Kunden den Vorschlag als bestes, der auch am häufigsten von ihnen benutzt wurde.

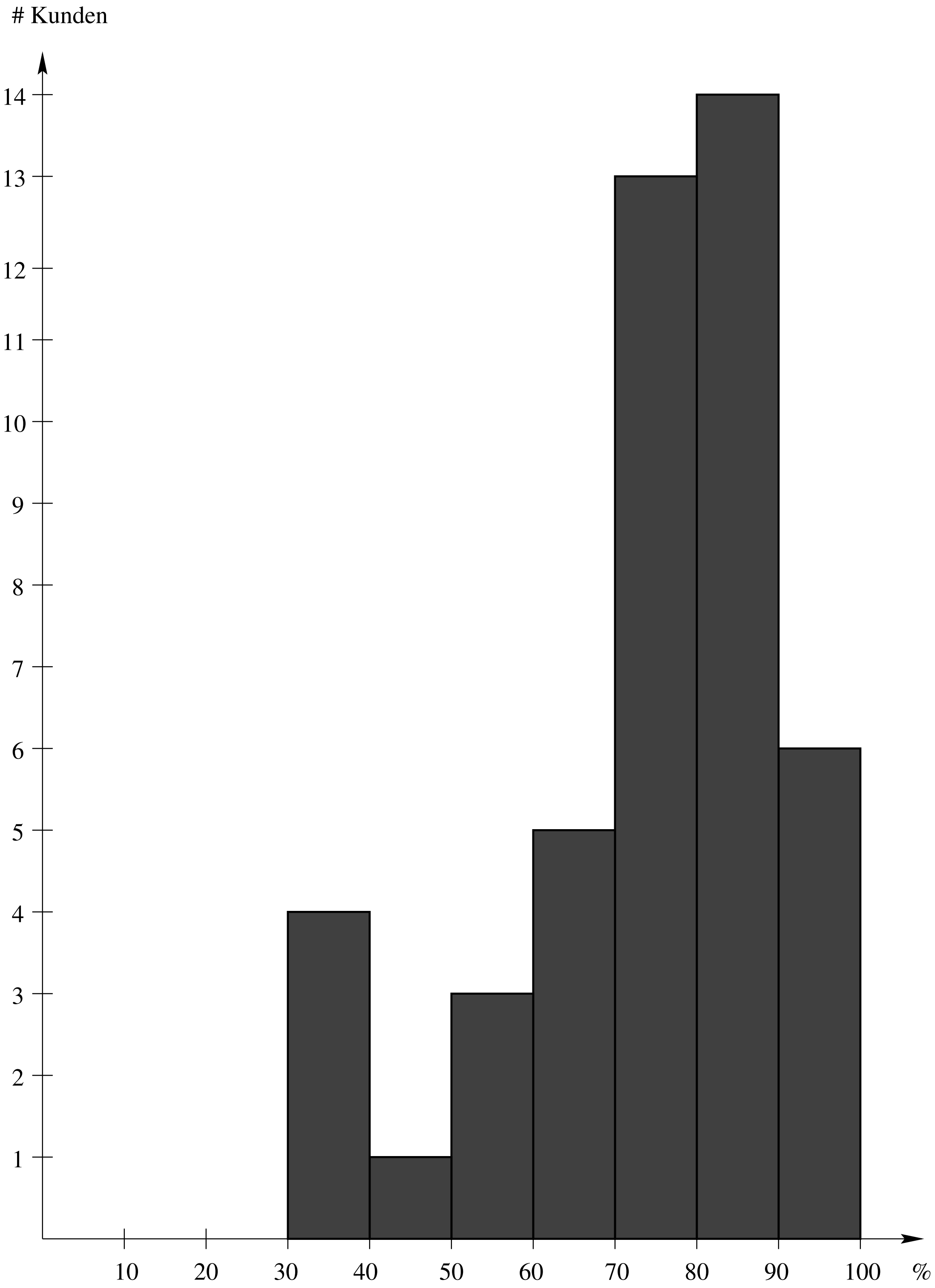

Abb. 4.6: Verteilung der „über c4.5“ gekauften Produkte

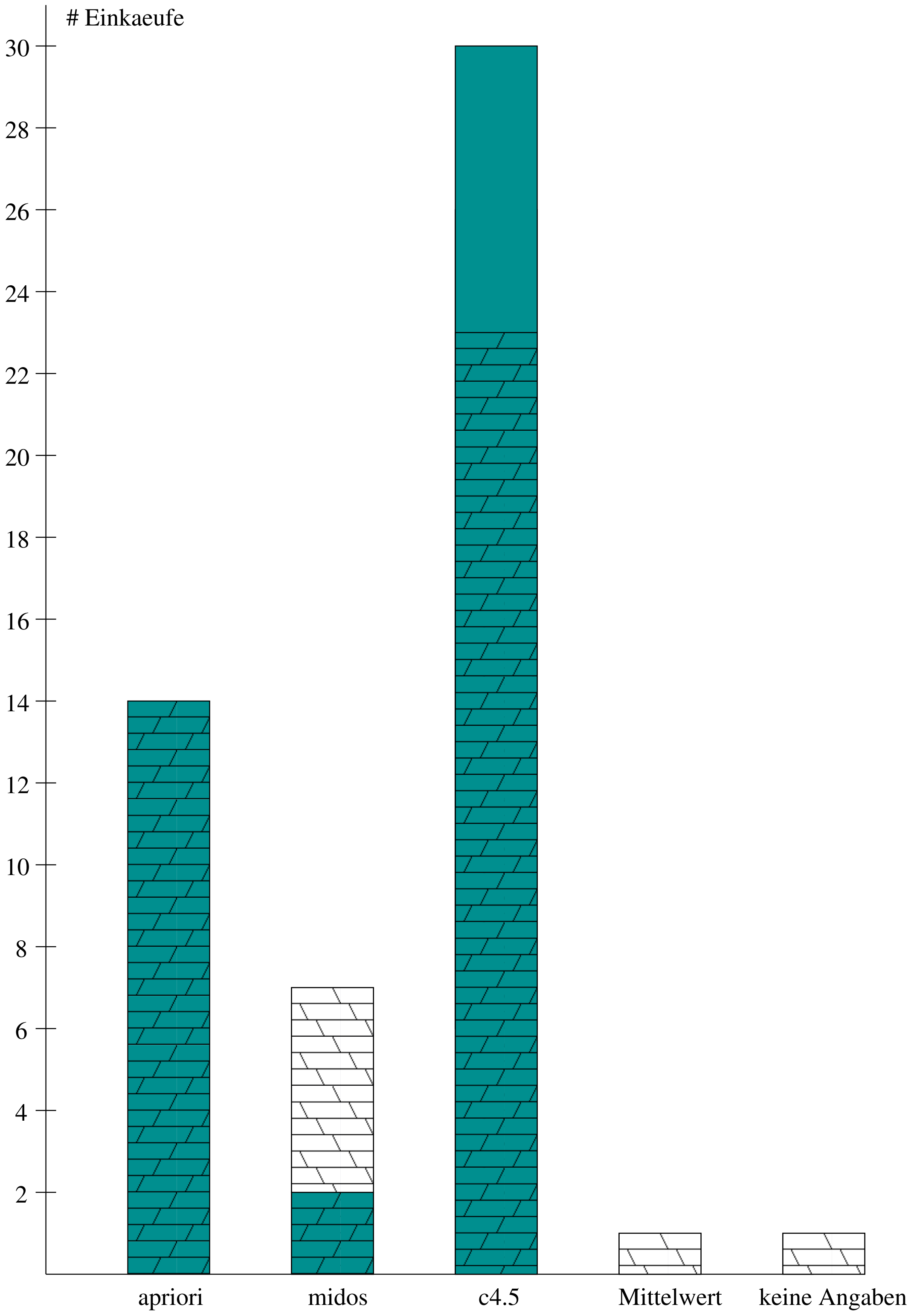

Abb. 4.7: Die besten bei c4.5

Kapitel 5

Zusammenfassung

5.1 Fazit

In dieser Diplomarbeit werden im wesentlichen zwei Fragen gestellt. Hier versuche ich, die einzelnen Fragen mit den Antworten zusammenfassend darzustellen.

- Können die Methode des Maschinellen Lernens sinnvoll zur Wissensentdeckung auf E-Commerce Daten benutzt werden?

  Die Ergebnisse der Arbeit zeigen, daß alle drei Lernverfahren dafür sich einsetzen lassen. Die Aktivität der Kunden in dem Online-Shop mit Vorschlägen im Vergleich zu dem Shop ohne Vorschlägen ist gewachsen. Die Vorschläge selbst waren bei den Testkunden sehr beliebt, die Kunden haben viele Produkte über die Vorschläge gekauft.

- Kann man eine Lernmethode als die beste wählen?

  Die besten Ergebnisse waren mit apriori und c4.5 zu erzielen. Die Antwort auf diese Frage ist aber nicht trivial. Der Test zeigt, daß die Kunden lieber „bei apriori“ einkaufen, falls sie diesen Vorschlag zur Verfügung haben. Die Ergebnisse, die apriori liefern, lassen aber weniger Kunden in Gruppen einteilen, als es bei c4.5 der Fall ist. An dieser Stelle muß das Unternehmen, daß die Lernverfahren einsetzen will, selbst entscheiden, welche Aspekte für ihn wichtiger sind.

Die Schwäche der Diplomarbeit sehe ich vor allem darin, daß die Verkaufsdaten, die die Grundlage für den Test bilden, keine echten sind. Man kann nicht ausschliesen, daß die Kunden sich in einem realen Geschäft anders verhalten, und damit die ganzen Ergebnisse der Diplomarbeit in Frage stellen. In diesem Punkt mußte ich mich auf Gewissen und Ehrlichkeit der Testkunden verlassen.

5.2 Ausblick

Der Punkt, der in dieser Diplomarbeit nicht betrachtet wurde, sind die Abstände in denen die Benutzermodellierung – also in diesem Fall die Gruppenbildung durchgeführt werden muß. Das Unternehmen, daß eines der hier beschriebenen Verfahren einsetzen möchte, muß diese Abstände evtl. empirisch bestimmen. Wie bereits erwähnt, dürfen diese nicht so groß sein, daß die Ähnlichkeit der Benutzer in einer Gruppe stark nachlässt, sie kann aber wegen der Zeit- und Rechenaufwandes nicht zu klein gewählt werden.

Anhang A

Gesetz zum Schutz personenbezogener Daten (Datenschutzgesetz Nordrhein-Westfalen - DSG NRW)

Erster Teil. Allgemeiner Datenschutz

Erster Abschnitt. Allgemeine Bestimmungen

§1 Aufgabe

Aufgabe dieses Gesetzes ist es, den Einzelnen davor zu schützen, dass er durch die Verarbeitung personenbezogener Daten durch öffentliche Stellen in unzulässiger Weise in seinem Recht beeinträchtigt wird, selbst über die Preisgabe und Verwendung seiner Daten zu bestimmen (informationelles elbstbestimmungsrecht).

§2 Anwendungsbereich

(1) Dieses Gesetz gilt für die Behörden, Einrichtungen und sonstigen öffentlichen Stellen des Landes, die Gemeinden und Gemeindeverbände sowie für die sonstigen der Aufsicht des Landes unterstehenden juristischen Personen des öffentlichen Rechts und deren Vereinigungen (öffentliche Stellen), soweit diese personenbezogene Daten verarbeiten. Für den Landtag und für die Gerichte sowie für die Behörden der Staatsanwaltschaft gilt dieses Gesetz, soweit sie Verwaltungsaufgaben wahrnehmen; darüber hinaus gelten für die Behörden der Staatsanwaltschaft, soweit sie keine Verwaltungsaufgaben wahrnehmen, nur die Vorschriften des Zweiten Teils dieses Gesetzes. Für den Landesrechnungshof und die Staatlichen Rechnungsprüfungsämter gelten der Dritte Abschnitt des Ersten Teils und der Zweite Teil sowie die §§8 und 32 a nur, soweit sie Verwaltungsaufgaben wahrnehmen. Für die Ausübung des Gnadenrechts findet das

Gesetz keine Anwendung.

(2) Von den Vorschriften dieses Gesetzes gelten nur die Vorschriften des Zweiten Teils sowie die §§8 und 28 bis 31 dieses Gesetzes, soweit

1. wirtschaftliche Unternehmen der Gemeinden oder Gemeindeverbände ohne eigene Rechtspersönlichkeit (Eigenbetriebe),
2. öffentliche Einrichtungen, die entsprechend den Vorschriften über die Eigenbetriebe geführt werden,
3. der Aufsicht des Landes unterstehende juristische Personen des öffentlichen Rechts, die am Wettbewerb teilnehmen,

personenbezogene Daten zu wirtschaftlichen Zwecken oder Zielen verarbeiten. Im Übrigen sind mit Ausnahme der §§32 sowie 36 bis 38 die für nicht-öffentliche Stellen geltenden Vorschriften des Bundesdatenschutzgesetzes einschließlich der Straf- und Bußgeldvorschriften anzuwenden. Unbeschadet der Regelung des Absatzes 1 Satz 1 gelten Schulen der Gemeinden und Gemeindeverbände, soweit sie in inneren Schulangelegenheiten personenbezogene Daten verarbeiten, als öffentliche Stellen im Sinne dieses Gesetzes.

(3) Soweit besondere Rechtsvorschriften auf die Verarbeitung personenbezogener Daten anzuwenden sind, gehen sie den Vorschriften dieses Gesetzes vor.

§3 Begriffsbestimmungen

(1) Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (betroffene Person).

(2) Datenverarbeitung ist das Erheben, Speichern, Verändern, Übermitteln, Sperren, Löschen sowie Nutzen personenbezogener Daten. Im Einzelnen ist

1. Erheben (Erhebung) das Beschaffen von Daten über die betroffene Person,
2. Speichern (Speicherung) das Erfassen, Aufnehmen oder Aufbewahren von Daten auf einem Datenträger zum Zwecke ihrer weiteren Verarbeitung,
3. Verändern (Veränderung) das inhaltliche Umgestalten gespeicherter Daten,
4. Übermitteln (Übermittlung) das Bekanntgeben gespeicherter oder durch Datenverarbeitung gewonnener Daten an einen Dritten in der Weise, dass die Daten durch die verantwortliche Stelle weitergegeben oder zur Einsichtnahme bereitgehalten werden oder dass der Dritte zum Abruf in einem automatisierten Verfahren bereitgehaltene Daten abruft,
5. Sperren (Sperrung) das Verhindern weiterer Verarbeitung gespeicherter Daten,
6. Löschen (Löschung) das Unkenntlichmachen gespeicherter Daten,
7. Nutzen (Nutzung) jede sonstige Verwendung personenbezogener Daten,

ungeachtet der dabei angewendeten Verfahren.

(3) Verantwortliche Stelle ist die Stelle im Sinne des §2 Abs. 1, die personenbezogene Daten in eigener Verantwortung selbst verarbeitet oder in ihrem Auftrag von einer anderen Stelle verarbeiten lässt.

(4) Empfänger ist jede Person oder Stelle, die Daten erhält. Dritter ist jede Person oder Stelle außerhalb der verantwortlichen Stelle. Dritte sind nicht die betroffene Person sowie diejenigen Personen oder Stellen, die im Inland oder im übrigen Geltungsbereich der Rechtsvorschriften zum Schutz personenbezogener Daten der Mitgliedstaaten der Europäischen Union personenbezogene Daten im Auftrag verarbeiten.

(5) Automatisiert ist eine Datenverarbeitung, wenn sie durch Einsatz eines gesteuerten technischen Verfahrens selbsttätig abläuft.

(6) Eine Akte ist jede der Aufgabenerfüllung dienende Unterlage, die nicht Teil der automatisierten Datenverarbeitung ist.

(7) Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßigen Aufwand einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.

(8) Pseudonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse ohne Nutzung der Zuordnungsfunktion nicht oder nur mit einem unverhältnismäßigen Aufwand einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können. Die Daten verarbeitende Stelle darf keinen Zugriff auf die Zuordnungsfunktion haben; diese ist an dritter Stelle zu verwahren.

§4 Zulässigkeit der Datenverarbeitung

(1) Die Verarbeitung personenbezogener Daten ist nur zulässig, wenn

a) dieses Gesetz oder eine andere Rechtsvorschrift sie erlaubt oder

b) die betroffene Person eingewilligt hat.

Die Einwilligung ist die widerrufliche, freiwillige und eindeutige Willenserklärung der betroffenen Person, einer bestimmten Datenverarbeitung zuzustimmen. Sie bedarf der Schriftform, soweit nicht wegen besonderer Umstände eine andere Form angemessen ist. Soll die Einwilligung zusammen mit anderen Erklärungen schriftlich erteilt werden, ist die betroffene Person auf die Einwilligung schriftlich besonders hinzuweisen. Sie ist in geeigneter Weise über die Bedeutung der Einwilligung, insbesondere über den Verwendungszweck der Daten, bei einer beabsichtigten Übermittlung über die Empfänger der Daten aufzuklären; sie ist unter Darlegung der Rechtsfolgen darauf hinzuweisen, dass sie die Einwilligung verweigern und mit Wirkung für die Zukunft widerrufen kann. Die Einwilligung kann auch elektronisch erklärt werden, wenn sichergestellt ist, dass

1. sie nur durch eine eindeutige und bewusste Handlung der handelnden Person erfolgen kann,

2. sie nicht unerkennbar verändert werden kann,
3. ihr Urheber erkannt werden kann,
4. die Einwilligung bei der verarbeitenden Stelle protokolliert wird und
5. der betroffenen Person jederzeit Auskunft über den Inhalt ihrer Einwilligung gegeben werden kann.

(2) Die Planung, Gestaltung und Auswahl informationstechnischer Produkte und Verfahren haben sich an dem Ziel auszurichten, so wenig personenbezogene Daten wie möglich zu erheben und weiterzuverarbeiten (Datenvermeidung). Produkte und Verfahren, deren Vereinbarkeit mit den Vorschriften über den Datenschutz und die Datensicherheit in einem förmlichen Verfahren (Datenschutzaudit) festgestellt wurde, sollen vorrangig berücksichtigt werden.

(3) Die Verarbeitung personenbezogener Daten über die rassische oder ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen, die Gewerkschaftszugehörigkeit, die Gesundheit oder das Sexualleben ist nur zulässig, wenn sie in einer Rechtsvorschrift geregelt ist, die den Zweck der Verarbeitung bestimmt sowie angemessene Garantien zum Schutz des Rechtes auf informationelle Selbstbestimmung vorsieht. Darüber hinaus ist die Verarbeitung dieser Daten zulässig, wenn

1. die betroffene Person eingewilligt hat,
2. sie ausschließlich im Interesse der betroffenen Person liegt,
3. sie sich auf Daten bezieht, die die betroffene Person selbst öffentlich gemacht hat,
4. sie
   (a) auf der Grundlage der §§15, 28 und 29,
   (b) zur Geltendmachung rechtlicher Ansprüche vor Gericht oder
   (c) für die Abwehr von Gefahren für die öffentliche Sicherheit, für Zwecke der Strafrechtspflege oder zum Schutz vergleichbarer Rechtsgüter

   erforderlich ist.

(4) Soweit gesetzlich unter Wahrung der berechtigten Interessen der betroffenen Person nichts anderes bestimmt ist, dürfen Entscheidungen, die für die betroffene Person eine rechtliche Folge nach sich ziehen oder sie erheblich beeinträchtigen, nicht ausschließlich auf eine automatisierte Verarbeitung personenbezogener Daten zum Zweck der Bewertung einzelner Persönlichkeitsmerkmale gestützt werden, ohne dass der betroffenen Person die Geltendmachung der eigenen Interessen möglich gemacht worden ist.

(5) Wenn die betroffene Person schriftlich begründet, dass der im Übrigen rechtmäßigen Verarbeitung ihrer Daten oder einer bestimmten Datenverarbeitungsform ein schutzwürdiges besonderes persönliches Interesse entgegensteht, erfolgt die Verarbeitung ihrer personenbezogenen Daten nur, wenn eine

Abwägung im Einzelfall ergibt, dass das Interesse der datenverarbeitenden Stelle gegenüber dem Interesse der betroffenen Person überwiegt. Die betroffene Person ist über das Ergebnis zu unterrichten.

(6) Die Datenverarbeitung soll so organisiert sein, dass bei der Verarbeitung, insbesondere der Übermittlung, der Kenntnisnahme im Rahmen der Aufgabenerfüllung und der Einsichtnahme, die Trennung der Daten nach den jeweils verfolgten Zwecken und nach unterschiedlichen Betroffenen möglich ist. Sind personenbezogene Daten in Akten derart verbunden, dass ihre Trennung nach erforderlichen und nicht erforderlichen Daten auch durch Vervielfältigung und Unkenntlichmachung nicht oder nur mit unverhältnismäßigem Aufwand möglich ist, sind auch die Kenntnisnahme, die Weitergabe innerhalb der datenverarbeitenden Stelle und die Übermittlung der Daten, die nicht zur Erfüllung der jeweiligen Aufgaben erforderlich sind, zulässig, soweit nicht schutzwürdige Belange der betroffenen Person oder Dritter überwiegen. Die nicht erforderlichen Daten unterliegen insoweit einem Verwertungsverbot.

§4a Verbunddateien

(1) Die Einrichtung gemeinsamer oder verbundener automatisierter Verfahren, in und aus denen mehrere öffentliche Stellen personenbezogene Daten verarbeiten sollen, ist nur zulässig, wenn dies unter Berücksichtigung der schutzwürdigen Belange der betroffenen Personen und der Aufgaben der beteiligten Stellen angemessen ist. Die Vorschriften über die Zulässigkeit des einzelnen Abrufs bleiben unberührt. Die beteiligten Stellen haben die Datenart, die Aufgaben jeder beteiligten Stelle, den Zweck und den Umfang ihrer Verarbeitungsbefugnis sowie diejenige Stelle festzulegen, welche die datenschutzrechtliche Verantwortung gegenüber den betroffenen Personen trägt. Der Landesbeauftragte für den Datenschutz ist vorab zu unterrichten.

(2) Innerhalb einer öffentlichen Stelle bedarf die Einrichtung gemeinsamer oder verbundener automatisierter Verfahren, mit denen personenbezogene Daten aus unterschiedlichen Aufgabengebieten verarbeitet werden sollen, der Zulassung durch die Leitung der Stelle. Für die Zulässigkeit gilt Absatz 1 Satz 1 und 2 entsprechend.

§5 Rechte der betroffenen Person

Jeder hat nach Maßgabe dieses Gesetzes ein Recht auf

1. Auskunft, Einsichtnahme (§18),
2. Widerspruch aus besonderem Grund (§4 Abs. 5),
3. Unterrichtung (§§12 Abs. 2, 13 Abs. 2 Satz 2, 16 Abs. 1 Satz 2 und 3),
4. Berichtigung, Sperrung oder Löschung (§19),
5. Schadensersatz (§20),
6. Anrufung des Landesbeauftragten für den Datenschutz (§25 Abs. 1),

7. Auskunft aus dem beim zuständigen behördlichen Datenschutzbeauftragten geführten Verfahrensverzeichnis (§8).

Diese Rechte können auch durch die Einwilligung der betroffenen Person nicht ausgeschlossen oder beschränkt werden.

§6 Datengeheimnis

Denjenigen Personen, die bei öffentlichen Stellen oder ihren Auftragnehmern dienstlichen Zugang zu personenbezogenen Daten haben, ist es untersagt, solche Daten unbefugt zu einem anderen als dem zur jeweiligen rechtmäßigen Aufgabenerfüllung gehörenden Zweck zu verarbeiten oder zu offenbaren; dies gilt auch nach Beendigung ihrer Tätigkeit.

§7 Sicherstellung des Datenschutzes

Die obersten Landesbehörden, die Gemeinden und Gemeindeverbände sowie die sonstigen der Aufsicht des Landes unterstehenden juristischen Personen des öffentlichen Rechts und deren Vereinigungen ungeachtet ihrer Rechtsform haben jeweils für ihren Bereich die Ausführung dieses Gesetzes sowie anderer Rechtsvorschriften über den Datenschutz sicherzustellen.

§8 Verfahrensverzeichnis

(1) Jede datenverarbeitende Stelle, die für den Einsatz eines Verfahrens zur automatisierten Verarbeitung personenbezogener Daten verantwortlich ist, hat in einem für den behördlichen Datenschutzbeauftragten bestimmten Verzeichnis festzulegen:

1. Name und Anschrift der datenverarbeitenden Stelle,
2. die Zweckbestimmung und die Rechtsgrundlage der Datenverarbeitung,
3. die Art der gespeicherten Daten,
4. den Kreis der Betroffenen,
5. die Art regelmäßig zu übermittelnder Daten, deren Empfänger sowie die Art und Herkunft regelmäßig empfangener Daten,
6. die zugriffsberechtigten Personen oder Personengruppen,
7. die technischen und organisatorischen Maßnahmen nach §10,
8. die Technik des Verfahrens, einschließlich der eingesetzten Hard- und Software,
9. Fristen für die Sperrung und Löschung nach §19 Abs. 2 und Abs. 3,
10. eine beabsichtigte Datenübermittlung an Drittstaaten nach §17 Abs. 2 und Abs. 3,

11. die begründeten Ergebnisse der Vorabkontrollen nach §10 Abs. 3 Satz 1.

(2) Die Angaben des Verfahrensverzeichnisses können bei der datenverarbeitenden Stelle von jeder Person eingesehen werden; dies gilt für die Angaben zu den Nummern 7, 8 und 11 nur, soweit dadurch die Sicherheit des technischen Verfahrens nicht beeinträchtigt wird. Satz 1 gilt nicht für

1. Verfahren nach dem Verfassungsschutzgesetz Nordrhein-Westfalen,
2. Verfahren, die der Gefahrenabwehr oder der Strafrechtspflege dienen,
3. Verfahren der Steuerfahndung,

soweit die datenverarbeitende Stelle eine Einsichtnahme im Einzelfall mit der Erfüllung ihrer Aufgaben für unvereinbar erklärt. Die Gründe dafür sind aktenkundig zu machen und die antragstellende Person ist darauf hinzuweisen, dass sie sich an den Landesbeauftragten für den Datenschutz wenden kann. Dem Landesbeauftragten für den Datenschutz ist auf sein Verlangen Einsicht zu gewähren.

§9 Automatisiertes Abrufverfahren und regelmäßige Datenübermittlung

(1) Die Einrichtung eines automatisierten Verfahrens, das die Übermittlung personenbezogener Daten durch Abruf ermöglicht, ist nur zulässig, soweit dies durch Bundes- oder Landesrecht bestimmt ist.

(2) Die Ministerien werden ermächtigt, für die Behörden und Einrichtungen ihres Geschäftsbereichs sowie für die der Rechtsaufsicht des Landes unterliegenden sonstigen öffentlichen Stellen die Einrichtung automatisierter Abrufverfahren durch Rechtsverordnung zuzulassen. Ein solches Verfahren darf nur eingerichtet werden, soweit dies unter Berücksichtigung des informationellen Selbstbestimmungsrechts des betroffenen Personenkreises und der Aufgaben der beteiligten Stellen angemessen ist. Die Vorschriften über die Zulässigkeit des einzelnen Abrufs bleiben unberührt. Die Datenempfänger, die Datenart und der Zweck des Abrufs sind festzulegen. Der Landesbeauftragte für den Datenschutz ist zu unterrichten.

(3) Die am Abrufverfahren beteiligten Stellen haben die nach §10 erforderlichen Maßnahmen zu treffen.

(4) Für die Einrichtung automatisierter Abrufverfahren innerhalb einer öffentlichen Stelle gelten nur Absatz 2 Satz 2 bis 4 sowie Absatz 3 entsprechend.

(5) Personenbezogene Daten dürfen für Stellen außerhalb des öffentlichen Bereichs zum automatisierten Abruf nicht bereitgehalten werden; dies gilt nicht für die betroffene Person.

(6) Die Absätze 1 bis 5 gelten nicht für Datenbestände, die jedermann ohne oder nach besonderer Zulassung zur Benutzung offenstehen oder deren Veröffentlichung zulässig wäre.

(7) Absatz 1 und Absatz 2 Satz 1 und 5 sowie Absatz 5 finden keine Anwendung, soweit die zur Übermittlung vorgesehenen Daten mit schriftlicher

Einwilligung der betroffenen Personen zum Zwecke der Übermittlung im automatisierten Abrufverfahren gespeichert sind. §4 Abs. 1 Satz 4 und 5 gilt entsprechend.

(8) Die Absätze 1 bis 7 sind auf die Zulassung regelmäßiger Datenübermittlungen entsprechend anzuwenden.

§10 Technische und organisatorische Maßnahmen

(1) Die Ausführung der Vorschriften dieses Gesetzes sowie anderer Vorschriften über den Datenschutz ist durch technische und organisatorische Maßnahmen sicherzustellen.

(2) Dabei sind Maßnahmen zu treffen, die geeignet sind zu gewährleisten, dass

1. nur Befugte personenbezogene Daten zur Kenntnis nehmen können (Vertraulichkeit),
2. personenbezogene Daten während der Verarbeitung unversehrt, vollständig und aktuell bleiben (Integrität),
3. personenbezogene Daten zeitgerecht zur Verfügung stehen und ordnungsgemäß verarbeitet werden können (Verfügbarkeit),
4. jederzeit personenbezogene Daten ihrem Ursprung zugeordnet werden können (Authentizität),
5. festgestellt werden kann, wer wann welche personenbezogenen Daten in welcher Weise verarbeitet hat (Revisionsfähigkeit),
6. die Verfahrensweisen bei der Verarbeitung personenbezogener Daten vollständig, aktuell und in einer Weise dokumentiert sind, dass sie in zumutbarer Zeit nachvollzogen werden können (Transparenz).

(3) Die zu treffenden technischen und organisatorischen Maßnahmen sind auf der Grundlage eines zu dokumentierenden Sicherheitskonzepts zu ermitteln, zu dessen Bestandteilen die Vorabkontrolle hinsichtlich möglicher Gefahren für das in §1 geschützte Recht auf informationelle Selbstbestimmung gehört, die vor der Entscheidung über den Einsatz oder einer wesentlichen Änderung eines automatisierten Verfahrens durchzuführen ist. Das Verfahren darf nur eingesetzt werden, wenn diese Gefahren nicht bestehen oder durch Maßnahmen nach den Absätzen 1 und 2 verhindert werden können. Das Ergebnis der Vorabkontrolle ist aufzuzeichnen. Die Wirksamkeit der Maßnahmen ist unter Berücksichtigung sich verändernder Rahmenbedingungen und Entwicklungen der Technik zu überprüfen. Die sich daraus ergebenden notwendigen Anpassungen sind zeitnah umzusetzen.

(4) Der Landesrechnungshof kann von der zu prüfenden Stelle verlangen, dass für ein konkretes Prüfungsverfahren die notwendigen Maßnahmen nach den Absätzen 1 bis 3 zeitnah geschaffen werden.

§10a Datenschutzaudit

Die öffentlichen Stellen können zur Verbesserung von Datenschutz und Datensicherheit sowie zum Erreichen größtmöglicher Datensparsamkeit ihr Datenschutzkonzept sowie ihre technischen Einrichtungen durch unabhängige und zugelassene Gutachter prüfen und bewerten sowie das Ergebnis der Prüfung veröffentlichen lassen. Sie können auch bereits geprüfte und bewertete Datenschutzkonzepte und Programme zum Einsatz bringen. Die näheren Anforderungen an die Prüfung und Bewertung, das Verfahren sowie die Auswahl und Zulassung der Gutachter werden durch besonderes Gesetz geregelt.

§11 Verarbeitung personenbezogener Daten im Auftrag

(1) Werden personenbezogene Daten im Auftrag einer öffentlichen Stelle verarbeitet, bleibt der Auftraggeber für die Einhaltung der Vorschriften dieses Gesetzes und anderer Vorschriften über den Datenschutz verantwortlich. Der Auftraggeber ist verantwortliche Stelle im Sinne dieses Gesetzes; die in §5 genannten Rechte sind ihm gegenüber geltend zu machen. Der Auftragnehmer darf personenbezogene Daten nur im Rahmen der Weisungen des Auftraggebers verarbeiten. Der Auftraggeber hat den Auftragnehmer unter besonderer Berücksichtigung seiner Eignung für die Gewährleistung der nach §10 notwendigen technischen und organisatorischen Maßnahmen sorgfältig auszuwählen. Der Auftrag ist schriftlich zu erteilen, wobei erforderlichenfalls ergänzende Weisungen zu technischen und organisatorischen Maßnahmen und etwaige Unterauftragsverhältnisse festzulegen sind.

(2) Soweit das Landesamt für Datenverarbeitung und Statistik (Landesdatenverarbeitungszentrale), die Gemeinsamen Gebietsrechenzentren, die Fachrechenzentren, die Hochschulrechenzentren und die kommunalen Datenverarbeitungseinrichtungen personenbezogene Daten im Auftrag öffentlicher Stellen verarbeiten, gelten für sie außer §§6 und 10 auch §22 sowie §§24 und 25 dieses Gesetzes unmittelbar.

(3) Sofern die Vorschriften dieses Gesetzes auf den Auftragnehmer keine Anwendung finden, ist der Auftraggeber verpflichtet sicherzustellen, dass der Auftragnehmer die Bestimmungen dieses Gesetzes befolgt und sich, sofern die Datenverarbeitung im Geltungsbereich dieses Gesetzes durchgeführt wird, der Kontrolle des Landesbeauftragten für den Datenschutz unterwirft. Bei einer Auftragsdurchführung außerhalb des Geltungsbereichs dieses Gesetzes ist die zuständige Datenschutzkontrollbehörde zu unterrichten.

(4) Externe Personen und Stellen, die mit der Wartung und Systembetreuung von Einrichtungen zur automatisierten Datenverarbeitung beauftragt sind, unterliegen den Regelungen der Datenverarbeitung im Auftrag. Sie müssen die notwendige fachliche Qualifikation und Zuverlässigkeit aufweisen. Der Auftraggeber hat vor Beginn der Arbeiten sicherzustellen, dass der Auftragnehmer personenbezogene Daten nur zur Kenntnis nehmen kann, soweit dies unvermeidlich ist. Dies gilt auch für die Kenntnisnahme von Daten, die Berufs- oder besonderen Amtsgeheimnissen unterliegen. Der Auftragnehmer hat dem Auftraggeber zuzuordnende personenbezogene Daten unverzüglich nach Erledigung

des Auftrages zu löschen. Die Dokumentation der Maßnahmen ist zum Zweck der Datenschutzkontrolle drei Jahre aufzubewahren.

Rechtsgrundlagen der Datenverarbeitung

§12 Erhebung

(1) Das Erheben personenbezogener Daten ist nur insoweit zulässig, als ihre Kenntnis zur rechtmäßigen Erfüllung der Aufgaben der erhebenden Stelle erforderlich ist. Durch die Art und Weise der Erhebung darf das allgemeine Persönlichkeitsrecht der betroffenen Person nicht beeinträchtigt werden. Personenbezogene Daten sind bei der betroffenen Person mit ihrer Kenntnis zu erheben; bei anderen Stellen oder Personen dürfen sie ohne ihre Kenntnis nur unter den Voraussetzungen des §13 Abs. 2 Satz 1 Buchstabe a und c bis g oder i erhoben werden.

(2) Werden Daten bei der betroffenen Person erhoben, so ist sie über den Verwendungszweck und eine etwaige beabsichtigte Übermittlung aufzuklären. Werden Daten aufgrund einer Rechtsvorschrift erhoben, so ist die betroffene Person in geeigneter Weise über diese aufzuklären. Soweit eine Auskunftspflicht besteht oder die Angaben Voraussetzung für die Gewährung von Rechtsvorteilen sind, ist die betroffene Person hierauf, sonst auf die Freiwilligkeit ihrer Angaben hinzuweisen. Werden Daten ohne Kenntnis der betroffenen Person erstmals erhoben, so ist sie bei Beginn der Speicherung oder im Fall einer vorgesehenen Übermittlung bei der ersten Übermittlung davon zu benachrichtigen, wenn die Erfüllung der Aufgaben dadurch nicht wesentlich beeinträchtigt wird. Satz 4 gilt nicht, wenn die betroffene Person auf andere Weise Kenntnis erhält, die Übermittlung durch Gesetz oder eine andere Rechtsvorschrift ausdrücklich vorgesehen ist oder die Daten für Zwecke von Statistiken, die durch Gesetz oder eine andere Rechtsvorschrift vorgeschrieben sind, verarbeitet werden. Mitzuteilen ist, welche Daten von welcher Stelle zu welchem Zweck auf welcher Rechtsgrundlage erhoben oder an wen sie übermittelt worden sind.

(3) Werden Daten bei einer dritten Person oder einer nicht-öffentlichen Stelle erhoben, so ist diese auf Verlangen über den Verwendungszweck aufzuklären. Soweit eine Auskunftspflicht besteht, ist sie hierauf, sonst auf die Freiwilligkeit ihrer Angaben hinzuweisen.

§13 Zweckbindung bei Speicherung, Veränderung und Nutzung

(1) Das Speichern, Verändern und Nutzen personenbezogener Daten ist zulässig, wenn es zur rechtmäßigen Erfüllung der Aufgaben der öffentlichen Stelle erforderlich ist. Die Daten dürfen nur für Zwecke weiterverarbeitet werden, für die sie erhoben worden sind. Daten, von denen die Stelle ohne Erhebung Kenntnis erlangt hat, dürfen nur für Zwecke genutzt werden, für die sie erstmals gespeichert worden sind.

(2) Sollen personenbezogene Daten zu Zwecken weiterverarbeitet werden, für die sie nicht erhoben oder erstmals gespeichert worden sind, ist dies nur zulässig, wenn

a. eine Rechtsvorschrift dies erlaubt oder die Wahrnehmung einer durch Gesetz oder Rechtsverordnung zugewiesenen einzelnen Aufgabe die Verarbeitung dieser Daten zwingend voraussetzt,

b. die betroffene Person eingewilligt hat,

c. Angaben der betroffenen Person überprüft werden müssen, weil tatsächliche Anhaltspunkte für deren Unrichtigkeit bestehen,

d. es zur Abwehr erheblicher Nachteile für das Gemeinwohl oder einer sonst unmittelbar drohenden Gefahr für die öffentliche Sicherheit oder zur Abwehr einer schwerwiegenden Beeinträchtigung der Rechte einer anderen Person erforderlich ist,

e. die Einholung der Einwilligung der betroffenen Person nicht möglich ist oder mit unverhältnismäßig hohem Aufwand verbunden wäre, aber offensichtlich ist, dass es in ihrem Interesse liegt und sie in Kenntnis des anderen Zwecks ihre Einwilligung erteilen würde,

f. sie aus allgemein zugänglichen Quellen entnommen werden können oder die speichernde Stelle sie veröffentlichen dürfte, es sei denn, dass das Interesse der betroffenen Person an dem Ausschluss der Speicherung oder einer Veröffentlichung der gespeicherten Daten offensichtlich überwiegt,

g. es zu Zwecken einer öffentlichen Auszeichnung oder Ehrung der betroffenen Person erforderlich ist,

h. sich bei Gelegenheit der rechtmäßigen Aufgabenerfüllung Anhaltspunkte für Straftaten oder Ordnungswidrigkeiten ergeben und die Unterrichtung der für die Verfolgung oder Vollstreckung zuständigen Behörden geboten erscheint oder

i. zur Durchsetzung öffentlich-rechtlicher Geldforderungen ein rechtliches Interesse an der Kenntnis der zu verarbeitenden Daten vorliegt und kein Grund zu der Annahme besteht, dass das schutzwürdige Interesse der betroffenen Person an der Geheimhaltung überwiegt.

Die betroffene Person ist außer im Fall des Buchstaben b davon in geeigneter Weise zu unterrichten, sofern nicht die Aufgabenerfüllung wesentlich beeinträchtigt wird. Unterliegen die personenbezogenen Daten einem Berufs- oder besonderen Amtsgeheimnis und sind sie der verantwortlichen Stelle von der zur Verschwiegenheit verpflichteten Person in Ausübung ihrer Berufs- oder Amtspflicht übermittelt worden, findet Satz 1 Buchstabe c bis i keine Anwendung.

(3) Eine Verarbeitung zu anderen Zwecken liegt nicht vor, wenn sie der Wahrnehmung von Aufsichts- und Kontrollbefugnissen, der Rechnungsprüfung oder der Durchführung von Organisationsuntersuchungen dient. Zulässig ist auch die Verarbeitung zu Ausbildungs- und Prüfungszwecken, soweit nicht berechtigte Interessen der betroffenen Person an der Geheimhaltung der Daten offensichtlich überwiegen.

§14 Übermittlung innerhalb des öffentlichen Bereichs

(1) Die Übermittlung personenbezogener Daten an öffentliche Stellen ist zulässig, wenn sie zur rechtmäßigen Erfüllung der Aufgaben der übermittelnden Stelle oder des Empfängers erforderlich ist und die Voraussetzungen des §13 Abs. 1 Satz 2 oder Satz 3 oder des Absatzes 2 Satz 1 vorliegen, sowie zur Wahrnehmung von Aufgaben nach §13 Abs. 3. Die Übermittlung ist ferner zulässig, soweit es zur Entscheidung in einem Verwaltungsverfahren der Beteiligung mehrerer öffentlicher Stellen bedarf.

(2) Die Verantwortung für die Übermittlung trägt die übermittelnde Stelle. Erfolgt die Übermittlung auf Grund eines Ersuchens des Empfängers, hat die übermittelnde Stelle lediglich zu prüfen, ob das Übermittlungsersuchen im Rahmen der Aufgaben des Empfängers liegt. Die Rechtmäßigkeit des Ersuchens prüft sie nur, wenn im Einzelfall hierzu Anlass besteht; der Empfänger hat der übermittelnden Stelle die für diese Prüfung erforderlichen Angaben zu machen. Erfolgt die Übermittlung durch automatisierten Abruf (§9), so trägt die Verantwortung für die Rechtmäßigkeit des Abrufs der Empfänger.

(3) Der Empfänger darf die übermittelten Daten nur für die Zwecke verarbeiten, zu deren Erfüllung sie ihm übermittelt worden sind; §13 Abs. 2 findet entsprechende Anwendung.

(4) Die Absätze 1 bis 3 gelten entsprechend, wenn personenbezogene Daten innerhalb einer öffentlichen Stelle weitergegeben werden.

§15 Übermittlung an öffentlich-rechtliche Religionsgesellschaften

Die Übermittlung personenbezogener Daten an Stellen der öffentlich-rechtlichen Religionsgesellschaften ist in entsprechender Anwendung der Vorschriften über die Datenübermittlung an öffentliche Stellen zulässig, sofern sichergestellt ist, dass bei dem Empfänger ausreichende Datenschutzmaßnahmen getroffen sind.

§16 Übermittlung an Personen oder Stellen außerhalb des öffentlichen Bereichs

(1) Die Übermittlung personenbezogener Daten an Personen oder Stellen außerhalb des öffentlichen Bereichs ist zulässig, wenn

a. sie zur rechtmäßigen Erfüllung der in der Zuständigkeit der übermittelnden Stelle liegenden Aufgaben erforderlich ist und die Voraussetzungen des §13 Abs. 1 vorliegen,

b. die Voraussetzungen des §13 Abs. 2 Satz 1 Buchstabe a, b, d, f oder i vorliegen,

c. der Auskunftsbegehrende ein rechtliches Interesse an der Kenntnis der zu übermittelnden Daten glaubhaft macht und kein Grund zu der Annahme besteht, dass das Geheimhaltungsinteresse der betroffenen Person überwiegt, oder

d. sie im öffentlichen Interesse liegt oder hierfür ein berechtigtes Interesse geltend gemacht wird und die betroffene Person in diesen Fällen der Datenübermittlung nicht widersprochen hat.

Bei Übermittlungen nach Satz 1 Buchstabe b, soweit sie unter den Voraussetzungen des §13 Abs. 2 Satz 1 Buchstabe i erfolgen, sowie in den Fällen des Satzes 1 Buchstabe c wird die betroffene Person vor der Mitteilung gehört, es sei denn, es ist zu besorgen, dass dadurch die Verfolgung des Interesses vereitelt oder wesentlich erschwert würde, und eine Abwägung ergibt, dass dieses Interesse das Interesse der betroffenen Person an ihrer vorherigen Anhörung überwiegt; ist die Anhörung unterblieben, wird die betroffene Person nachträglich unterrichtet. In den übrigen Fällen des Satzes 1 ist die betroffene Person über die beabsichtigte Übermittlung, die Art der zu übermittelnden Daten und den Verwendungszweck in geeigneter Weise zu unterrichten, sofern nicht die Aufgabenerfüllung wesentlich beeinträchtigt wird.

(2) Der Empfänger darf die übermittelten Daten nur für die Zwecke verarbeiten, zu denen sie ihm übermittelt wurden. Hierauf ist er bei der Übermittlung hinzuweisen.

§17 Übermittlung an ausländische Stellen

(1) Die Zulässigkeit der Übermittlung an öffentliche und nicht-öffentliche Stellen außerhalb des Geltungsbereichs des Grundgesetzes richtet sich nach den §§14 und 16. Die Übermittlung an Stellen außerhalb der Mitgliedstaaten der Europäischen Union ist nur zulässig, wenn dort ein angemessenes Datenschutzniveau gewährleistet ist. Vor der Entscheidung über die Angemessenheit des Datenschutzniveaus ist der Landesbeauftragte für den Datenschutz zu hören.

(2) Fehlt es an einem angemessenen Datenschutzniveau, so ist die Übermittlung nur zulässig, wenn

1. die betroffene Person in die Übermittlung eingewilligt hat,
2. die Übermittlung zur Wahrung eines überwiegenden öffentlichen Interesses oder zur Geltendmachung, Ausübung oder Verteidigung eines rechtlichen Interesses erforderlich ist,
3. die Übermittlung zur Wahrung lebenswichtiger Interessen der betroffenen Person erforderlich ist,
4. die Übermittlung aus einem für die Öffentlichkeit bestimmten Register erfolgt oder
5. die Übermittlung genehmigt wird, wenn die empfangende Stelle ausreichende Garantien hinsichtlich des Schutzes der informationellen Selbstbestimmung bietet. Die für die Genehmigungserteilung zuständige Stelle oder zuständigen Stellen bestimmt die Landesregierung durch Rechtsverordnung.

(3) Die empfangende Stelle ist darauf hinzuweisen, dass die Daten nur zu den Zwecken verarbeitet werden dürfen, für die sie übermittelt wurden.

Rechte der betroffenen Person

§18 Auskunft, Einsichtnahme

(1) Der betroffenen Person ist von der verantwortlichen Stelle auf Antrag Auskunft zu erteilen über

1. die zu ihrer Person verarbeiteten Daten,
2. den Zweck und die Rechtsgrundlage der Verarbeitung,
3. die Herkunft der Daten und die Empfänger von Übermittlungen sowie
4. die allgemeinen technischen Bedingungen der automatisierten Verarbeitung der zur eigenen Person verarbeiteten Daten.

Dies gilt nicht für personenbezogene Daten, die ausschließlich zu Zwecken der Datensicherung oder der Datenschutzkontrolle gespeichert sind.

(2) Auskunft oder Einsichtnahme sind zu gewähren, soweit die betroffene Person Angaben macht, die das Auffinden der Daten mit angemessenem Aufwand ermöglichen. Auskunftserteilungen und Einsichtnahme sind gebührenfrei, die Erstattung von Auslagen kann verlangt werden.

(3) Die Verpflichtung zur Auskunftserteilung oder zur Gewährung der Einsichtnahme entfällt, soweit

a. dies die ordnungsgemäße Erfüllung der Aufgaben der verantwortlichen Stelle erheblich gefährden würde,

b. dies die öffentliche Sicherheit gefährden oder sonst dem Wohle des Bundes oder eines Landes Nachteile bereiten würde,

c. die personenbezogenen Daten oder die Tatsache ihrer Speicherung nach einer Rechtsvorschrift oder wegen der berechtigten Interessen einer dritten Person geheimgehalten werden müssen.

(4) Einer Begründung für die Auskunftsverweigerung bedarf es nur dann nicht, wenn durch die Mitteilung der Gründe der mit der Auskunftsverweigerung verfolgte Zweck gefährdet würde. In diesem Fall sind die wesentlichen Gründe für die Entscheidung aufzuzeichnen.

(5) Bezieht sich die Auskunftserteilung oder die Einsichtnahme auf die Herkunft personenbezogener Daten von Behörden des Verfassungsschutzes, der Staatsanwaltschaft und der Polizei, von Landesfinanzbehörden, soweit diese personenbezogene Daten in Erfüllung ihrer gesetzlichen Aufgaben im Anwendungsbereich der Abgabenordnung zur Überwachung und Prüfung speichern, sowie von den in §19 Abs. 3 Bundesdatenschutzgesetz genannten Behörden, ist sie nur mit Zustimmung dieser Stellen zulässig. Gleiches gilt für die Übermittlung personenbezogener Daten an diese Behörden. Für die Versagung der Zustimmung gelten, soweit dieses Gesetz auf die genannten Behörden Anwendung findet, die Absätze 3 und 4 entsprechend.

(6) Werden Auskunft oder Einsichtnahme nicht gewährt, ist die betroffene Person darauf hinzuweisen, dass sie sich an den Landesbeauftragten für den Datenschutz wenden kann.

§19 Berichtigung, Sperrung und Löschung

(1) Personenbezogene Daten sind zu berichtigen, wenn sie unrichtig sind. Sind personenbezogene Daten zu berichtigen, so ist in geeigneter Weise kenntlich zu machen, zu welchem Zeitpunkt und aus welchem Grund diese Daten unrichtig waren oder geworden sind.

(2) Personenbezogene Daten sind zu sperren, wenn

a. ihre Richtigkeit von der betroffenen Person bestritten wird und sich weder die Richtigkeit noch die Unrichtigkeit feststellen lässt,

b. die betroffene Person an Stelle der Löschung nach Absatz 3 Satz 1 Buchstabe a die Sperrung verlangt,

c. die weitere Speicherung im Interesse der betroffenen Person geboten ist,

d. sie nur zu Zwecken der Datensicherung oder der Datenschutzkontrolle gespeichert sind.

In den Fällen nach Satz 1 Buchstabe c und d sind die Gründe aufzuzeichnen. Bei automatisierten Dateien ist die Sperrung grundsätzlich durch technische Maßnahmen sicherzustellen; im Übrigen ist ein entsprechender Vermerk anzubringen. Gesperrte Daten dürfen über die Speicherung hinaus nicht mehr weiterverarbeitet werden, es sei denn, dass dies zur Behebung einer bestehenden Beweisnot oder aus sonstigen im überwiegenden Interesse der verantwortlichen Stelle oder eines Dritten liegenden Gründen unerlässlich ist oder die betroffene Person eingewilligt hat.

(3) Personenbezogene Daten sind zu löschen, wenn

a. ihre Speicherung unzulässig ist oder

b. ihre Kenntnis für die speichernde Stelle zur Aufgabenerfüllung nicht mehr erforderlich ist.

Sind personenbezogene Daten in Akten gespeichert und ist die nach §4 Abs. 6 vorgesehene Abtrennung nicht möglich, ist die Löschung nach Satz 1 Buchstabe b nur durchzuführen, wenn die gesamte Akte zur Aufgabenerfüllung nicht mehr erforderlich ist, es sei denn, dass die betroffene Person die Löschung verlangt und die weitere Speicherung sie in unangemessener Weise beeinträchtigen würde. Soweit hiernach eine Löschung nicht in Betracht kommt, sind die personenbezogenen Daten auf Antrag der betroffenen Person zu sperren.

(4) Abgesehen von den Fällen des Absatzes 3 Satz 1 Buchstabe a ist von einer Löschung abzusehen, soweit die gespeicherten Daten auf Grund von Rechtsvorschriften einem Archiv zur Übernahme anzubieten oder von einem Archiv zu übernehmen sind.

(5) Über die Berichtigung unrichtiger Daten, die Sperrung bestrittener Daten und die Löschung oder Sperrung unzulässig gespeicherter Daten sind unverzüglich die betroffene Person und die Stellen zu unterrichten, denen die Daten übermittelt worden sind. Die Unterrichtung kann unterbleiben, wenn sie einen erheblichen Aufwand erfordern würde und nachteilige Folgen für die betroffene Person nicht zu befürchten sind.

§20 Schadensersatz

(1) Wird der betroffenen Person durch eine nach den Vorschriften dieses Gesetzes oder nach anderen Vorschriften über den Datenschutz unzulässige oder unrichtige Verarbeitung ihrer personenbezogenen Daten ein Schaden zugefügt, so ist ihr der Träger der verantwortlichen Stelle zum Schadensersatz verpflichtet. In schweren Fällen kann die betroffene Person auch wegen des Schadens, der nicht Vermögensschaden ist, eine angemessene Entschädigung in Geld verlangen.

(2) Ist der Schaden durch Verarbeitung der Daten in einer automatisierten Datei entstanden, besteht die Entschädigungspflicht unabhängig von einem Verschulden der verantwortlichen Stelle. In diesem Fall haftet der Ersatzpflichtige gegenüber der betroffenen Person für jedes schädigende Ereignis bis zu einem Betrag von 500.000 Deutsche Mark oder 250.000 Euro. Im Übrigen setzt die Verpflichtung zum Schadensersatz Verschulden voraus. Der verantwortlichen Stelle obliegt in Fällen des Satzes 3 die Beweislast, dass sie die unzulässige oder unrichtige Verarbeitung der Daten nicht zu vertreten hat. Mehrere Ersatzpflichtige haften als Gesamtschuldner.

(3) Auf eine schuldhafte Mitverursachung des Schadens durch die betroffene Person und die Verjährung des Entschädigungsanspruchs sind die §§254, 839 Abs. 3 und §852 des Bürgerlichen Gesetzbuches entsprechend anzuwenden.

(4) Weitergehende sonstige Schadensersatzansprüche bleiben unberührt.

Landesbeauftragter für den Datenschutz

§21 Berufung und Rechtsstellung

(1) Der Landtag wählt auf Vorschlag der Landesregierung einen Landesbeauftragten für den Datenschutz mit mehr als der Hälfte der gesetzlichen Zahl seiner Mitglieder. Dieser muss die Befähigung zum Richteramt oder zum höheren Dienst haben und die zur Erfüllung seiner Aufgaben erforderliche Fachkunde besitzen. Die Amts- und Funktionsbezeichnung "Der Landesbeauftragte für den Datenschutz wird in männlicher oder weiblicher Form geführt.

(2) Der Landesbeauftragte für den Datenschutz wird jeweils für die Dauer von acht Jahren in ein Beamtenverhältnis auf Zeit berufen. Nach Ende der Amtszeit bleibt er bis zur Ernennung eines Nachfolgers im Amt. Die Wiederwahl ist zulässig. Der Landesbeauftragte für den Datenschutz ist in Ausübung seines Amtes unabhängig und nur dem Gesetz unterworfen. Der Landesbeauftragte für den Datenschutz bestellt eine Mitarbeiterin oder einen Mitarbeiter zur Stellvertreterin oder zum Stellvertreter. Diese oder dieser führt die Geschäfte im Verhinderungsfall.

(3) Der Landesbeauftragte für den Datenschutz ist dem Innenministerium angegliedert. Er ist oberste Dienstbehörde im Sinne des §96 der Strafprozessordnung und trifft Entscheidungen nach §§64 und 65 des Landesbeamtengesetzes für das Land Nordrhein-Westfalen für sich und seine Bediensteten in eigener Verantwortung. Im Übrigen untersteht er der Dienstaufsicht des Innenministeriums.

(4) Dem Landesbeauftragten für den Datenschutz ist die für die Erfüllung seiner Aufgaben notwendige Personal- und Sachausstattung zur Verfügung zu stellen; sie ist im Einzelplan des Innenministeriums in einem eigenen Kapitel auszuweisen.

(5) In Personalangelegenheiten hat der Landesbeauftragte für den Datenschutz ein Vorschlagsrecht. Die Stellen sind im Einvernehmen mit ihm zu besetzen. Die Bediensteten können nur im Einvernehmen mit ihm versetzt oder abgeordnet werden; sie unterstehen seinen Weisungen.

(6) Der Landesbeauftragte für den Datenschutz kann sich jederzeit an den Landtag wenden.

§22 Aufgaben und Befugnisse

(1) Der Landesbeauftragte für den Datenschutz überwacht die Einhaltung der Vorschriften dieses Gesetzes sowie anderer Vorschriften über den Datenschutz bei den öffentlichen Stellen. Den Stellen kann der Landesbeauftragte für den Datenschutz auch Empfehlungen zur Verbesserung des Datenschutzes geben, insbesondere die Landesregierung und einzelne Ministerien, Gemeinden und Gemeindeverbände sowie die übrigen öffentlichen Stellen in Fragen des Datenschutzes beraten.

(2) Die öffentlichen Stellen sind verpflichtet, den Landesbeauftragten für den Datenschutz bei der Aufgabenerfüllung zu unterstützen und Amtshilfe zu leisten. Gesetzliche Geheimhaltungsvorschriften können einem Auskunfts- oder Einsichtsverlangen nicht entgegengehalten werden. Dem Landesbeauftragten für den Datenschutz sind insbesondere

1. Auskunft über die Fragen zu erteilen sowie Einsicht in alle Datenverarbeitungsvorgänge, Dokumentationen und Aufzeichnungen zu gewähren, die im Zusammenhang mit der Verarbeitung personenbezogener Daten stehen, namentlich auch in die gespeicherten Daten,

2. jederzeit Zutritt zu allen Diensträumen und Zugriff auf elektronische Dienste zu gewähren und

3. Kopien von Unterlagen, von automatisierten Dateien, von deren Verfahren und von organisatorischen Regelungen zur Mitnahme zur Verfügung zu stellen, soweit nicht die Aufgabenerfüllung der verantwortlichen Stelle wesentlich gefährdet wird. Die Gefährdung ist schriftlich zu begründen.

Die Rechte nach Satz 3 dürfen nur vom Landesbeauftragten für den Datenschutz persönlich ausgeübt werden, wenn die oberste Landesbehörde im Einzelfall feststellt, dass die Sicherheit des Bundes oder eines Landes dies gebietet. In diesem Fall müssen personenbezogene Daten einer betroffenen Person, der von der datenverarbeitenden Stelle Vertraulichkeit besonders zugesichert worden ist, auch ihm gegenüber nicht offenbart werden.

(3) Der Landesbeauftragte für den Datenschutz ist frühzeitig über Planungen zur Entwicklung, zum Aufbau oder zur wesentlichen Veränderung automatisierter Datenverarbeitungs- und Informationssysteme zu unterrichten, sofern

in dem jeweiligen System personenbezogene Daten verarbeitet werden sollen. Dasselbe gilt bei Entwürfen für Rechts- oder Verwaltungsvorschriften des Landes, wenn sie eine Verarbeitung personenbezogener Daten vorsehen.

(4) Der Landtag und die Landesregierung können den Landesbeauftragten für den Datenschutz mit der Erstattung von Gutachten und Stellungnahmen oder der Durchführung von Untersuchungen in Datenschutzfragen betrauen.

(5) Der Landesbeauftragte für den Datenschutz ist befugt, personenbezogene Daten, die ihm durch Beschwerden, Anfragen, Hinweise und Beratungswünsche bekannt werden, zu verarbeiten, soweit dies zur Erfüllung seiner Aufgaben erforderlich ist. Er darf im Rahmen von Kontrollmaßnahmen personenbezogene Daten auch ohne Kenntnis der betroffenen Person erheben. Von einer Benachrichtigung der betroffenen Person kann nach pflichtgemäßem Ermessen abgesehen werden. Die nach den Sätzen 1 und 2 erhobenen und verarbeiteten Daten dürfen nicht zu anderen Zwecken weiterverarbeitet werden.

(6) Der Landesbeauftragte für den Datenschutz arbeitet mit den Behörden und sonstigen Stellen zusammen, die für die Kontrolle der Einhaltung der Vorschriften über den Datenschutz in der Europäischen Union, im Bund und in den Ländern zuständig sind. Aufsichtsbehörde im Sinne des §38 Bundesdatenschutzgesetz ist der Landesbeauftragte für den Datenschutz. Insofern untersteht er der Aufsicht des Innenministeriums. Führt er die Weisungen nicht aus, kann ihn das Innenministerium erneut anweisen. Kommt er der neuerlichen Weisung nicht binnen einer Woche nach, steht zur Prüfung der Rechtmäßigkeit der Weisung der Rechtsweg vor dem Verwaltungsgericht offen. Kommt der Landesbeauftragte für den Datenschutz der Weisung auch nach Bestätigung ihrer Rechtmäßigkeit durch das Verwaltungsgericht nicht nach, kann das Innenministerium den Vertreter anweisen; entgegenstehende Weisungen des Landesbeauftragten für den Datenschutz sind unbeachtlich. Das Innenministerium und der Landesbeauftragte für den Datenschutz werden ermächtigt, Regelungen zum weiteren Verfahren der Aufsicht im nicht-öffentlichen Bereich zu vereinbaren.

§23

(aufgehoben)

§24 Beanstandungen durch den Landesbeauftragten

(1) Stellt der Landesbeauftragte für den Datenschutz Verstöße gegen die Vorschriften dieses Gesetzes, gegen andere Vorschriften über den Datenschutz oder sonstige Mängel bei der Verarbeitung personenbezogener Daten fest, so beanstandet er diese

1. bei der Landesverwaltung gegenüber der zuständigen obersten Landesbehörde, beim Landesrechnungshof gegenüber der Präsidentin oder dem Präsidenten,

2. bei der Kommunalverwaltung gegenüber der jeweils verantwortlichen Gemeinde oder dem verantwortlichen Gemeindeverband,

3. bei den wissenschaftlichen Hochschulen, Gesamthochschulen und Fachhochschulen gegenüber dem Hochschulpräsidenten oder dem Rektor, bei öffentlichen Schulen gegenüber dem Leiter der Schule,

4. bei den sonstigen Körperschaften, Anstalten und Stiftungen des öffentlichen Rechts gegenüber dem Vorstand oder dem sonst vertretungsberechtigten Organ

und fordert zur Stellungnahme innerhalb einer von ihm zu bestimmenden Frist auf. In den Fällen von Satz 1 Nr. 2 bis 4 unterrichtet der Landesbeauftragte für den Datenschutz gleichzeitig auch die zuständige Aufsichtsbehörde.

(2) Der Landesbeauftragte für den Datenschutz kann von einer Beanstandung absehen oder auf eine Stellungnahme der betroffenen Stelle verzichten, wenn es sich um unerhebliche Mängel handelt oder wenn ihre Behebung sichergestellt ist.

(3) Mit der Beanstandung kann der Landesbeauftragte für den Datenschutz Vorschläge zur Beseitigung der Mängel und zur sonstigen Verbesserung des Datenschutzes verbinden.

(4) Die gemäß Absatz 1 abzugebende Stellungnahme soll auch eine Darstellung der Maßnahmen enthalten, die auf Grund der Beanstandung des Landesbeauftragten für den Datenschutz getroffen worden sind. Die in Absatz 1 Nr. 2 bis 4 genannten Stellen leiten der zuständigen Aufsichtsbehörde eine Abschrift ihrer Stellungnahme an den Landesbeauftragten für den Datenschutz zu.

§25 Anrufungsrecht der betroffenen Person

(1) Wer der Ansicht ist, dass gegen Vorschriften dieses Gesetzes oder gegen andere Datenschutzvorschriften verstoßen worden ist oder ein solcher Verstoß bevorsteht, hat das Recht, sich unmittelbar an den Landesbeauftragten für den Datenschutz zu wenden; dies gilt auch für Bedienstete öffentlicher Stellen, ohne dass der Dienstweg eingehalten werden muss.

(2) Niemand darf deswegen benachteiligt oder gemaßregelt werden, weil er sich an den Landesbeauftragten für den Datenschutz wendet.

§26

(aufgehoben)

§27 Datenschutzbericht

Der Landesbeauftragte für den Datenschutz legt dem Landtag und der Landesregierung jeweils für zwei Kalenderjahre einen Bericht über seine Tätigkeit vor (Datenschutzbericht). Die Landesregierung nimmt hierzu gegenüber dem Landtag schriftlich Stellung. Der Landesbeauftragte für den Datenschutz berät und informiert mit dem Bericht und auf andere Weise die Bürger sowie die Öffentlichkeit zu Fragen des Datenschutzes.

Besonderer Datenschutz

§28 Datenverarbeitung für wissenschaftliche Zwecke

(1) Die Verarbeitung personenbezogener Daten zu wissenschaftlichen Zwecken soll in anonymisierter Form erfolgen. Stehen einer Anonymisierung wissenschaftliche Gründe entgegen, dürfen die Daten auch verarbeitet werden, wenn sie pseudonymisiert werden und der mit der Forschung befasste Personenkreis oder die empfangende Stelle oder Person keinen Zugriff auf die Zuordnungsfunktion hat. Datenerfassung, Anonymisierung oder Pseudonymisierung kann auch durch die mit der Forschung befassten Personen erfolgen, wenn sie zuvor nach dem Verpflichtungsgesetz zur Verschwiegenheit verpflichtet worden sind und unter der Aufsicht der übermittelnden Stelle stehen.

(2) Ist eine Anonymisierung oder Pseudonymisierung nicht möglich, so dürfen personenbezogene Daten für ein bestimmtes Forschungsvorhaben verarbeitet werden, wenn

1. die betroffene Person eingewilligt hat,
2. schutzwürdige Belange der betroffenen Person wegen der Art der Daten oder der Art der Verwendung nicht beeinträchtigt werden oder
3. der Zweck der Forschung auf andere Weise nicht oder nur mit unverhältnismäßig großem Aufwand erreicht werden kann und das öffentliche Interesse an der Durchführung des Forschungsvorhabens die schutzwürdigen Belange der betroffenen Person überwiegt.

(3) Sobald es der Forschungszweck gestattet, sind die Daten zu anonymisieren, hilfsweise zu pseudonymisieren. Die Merkmale, mit deren Hilfe ein Personenbezug wiederhergestellt werden kann, sind gesondert zu speichern; sie müssen gelöscht werden, sobald der Forschungszweck dies zulässt. Sollen personenbezogene Daten für einen anderen als den ursprünglichen Forschungszweck verarbeitet werden, ist dies nur nach Maßgabe der Absätze 1 und 2 zulässig.

(4) Die zu wissenschaftlichen Zwecken verarbeiteten Daten dürfen nur veröffentlicht werden, wenn

1. die betroffene Person eingewilligt hat oder
2. das öffentliche Interesse an der Darstellung des Forschungsergebnisses die schutzwürdigen Belange der betroffenen Person erheblich überwiegt.

(5) Soweit öffentliche Stellen personenbezogene Daten übermitteln, haben sie diejenigen empfangenden Stellen, auf die dieses Gesetz keine Anwendung findet, darauf zu verpflichten, die Vorschriften der Absätze 1 bis 4 einzuhalten und jederzeit Kontrollen durch den Landesbeauftragten für den Datenschutz zu ermöglichen. Bei einer Datenübermittlung an Stellen außerhalb des Geltungsbereichs dieses Gesetzes hat die übermittelnde Stelle die für den Empfänger zuständige Datenschutzkontrollbehörde zu unterrichten.

§29 Datenverarbeitung bei Dienst- und Arbeitsverhältnissen

(1) Daten von Bewerbern und Beschäftigten dürfen nur verarbeitet werden, wenn dies zur Eingehung, Durchführung, Beendigung oder Abwicklung des Dienst- oder Arbeitsverhältnisses oder zur Durchführung organisatorischer, personeller und sozialer Maßnahmen, insbesondere auch zu Zwecken der Personalplanung und des Personaleinsatzes, erforderlich ist oder eine Rechtsvorschrift, ein Tarifvertrag oder eine Dienstvereinbarung dies vorsieht. Abweichend von §16 Abs. 1 ist eine Übermittlung der Daten von Beschäftigten an Personen und Stellen außerhalb des öffentlichen Bereichs nur zulässig, wenn der Empfänger ein rechtliches Interesse darlegt, der Dienstverkehr es erfordert oder die betroffene Person eingewilligt hat. Die Datenübermittlung an einen künftigen Dienstherrn oder Arbeitgeber ist nur mit Einwilligung der betroffenen Person zulässig.

(2) Die beamtenrechtlichen Vorschriften über die Führung von Personalakten (§§102 ff. Landesbeamtengesetz) sind für alle nicht beamteten Beschäftigten einer öffentlichen Stelle entsprechend anzuwenden, soweit nicht die Besonderheiten des Tarif- und Arbeitsrechts hinsichtlich der Aufnahme und Entfernung von bestimmten Vorgängen und Vermerken eine abweichende Behandlung erfordern.

(3) Die Weiterverarbeitung der bei ärztlichen oder psychologischen Untersuchungen und Tests zum Zwecke der Eingehung eines Dienst- oder Arbeitsverhältnisses erhobenen Daten ist nur mit schriftlicher Einwilligung der betroffenen Person zulässig. Die Einstellungsbehörde darf vom untersuchenden Arzt in der Regel nur die Übermittlung des Ergebnisses der Eignungsuntersuchung und dabei festgestellter Risikofaktoren verlangen.

(4) Personenbezogene Daten, die vor der Eingehung eines Dienst- oder Arbeitsverhältnisses erhoben wurden, sind unverzüglich zu löschen, sobald feststeht, dass ein Dienst- oder Arbeitsverhältnis nicht zustande kommt, es sei denn, dass die betroffene Person in die weitere Speicherung eingewilligt hat. Nach Beendigung eines Dienst- oder Arbeitsverhältnisses sind personenbezogene Daten zu löschen, wenn diese Daten nicht mehr benötigt werden, es sei denn, dass Rechtsvorschriften entgegenstehen; §19 Abs. 3 Satz 2 und 3 sowie Abs. 4 finden Anwendung.

(5) Die Ergebnisse medizinischer oder psychologischer Untersuchungen und Tests der Beschäftigten dürfen automatisiert nur verarbeitet werden, wenn dies dem Schutz der Beschäftigten dient.

(6) Soweit Daten der Beschäftigten im Rahmen der Durchführung der technischen und organisatorischen Maßnahmen nach §10 gespeichert werden, dürfen sie nicht zu Zwecken der Verhaltens- oder Leistungskontrolle genutzt werden.

(7) Beurteilungen dürfen nicht allein auf Informationen gestützt werden, die unmittelbar durch automatisierte Datenverarbeitung gewonnen werden.

§29a Mobile personenbezogene Datenverarbeitungssysteme

(1) Informationstechnische Systeme zum Einsatz in automatisierten Verfahren, die an die Betroffenen ausgegeben werden und die über eine von der ausgebenden Stelle oder Dritten bereitgestellte Schnittstelle Daten automati-

siert austauschen können (mobile Datenverarbeitungssysteme, z. B. Chipkarten), dürfen nur mit Einwilligung der betroffenen Person nach ihrer vorherigen umfassenden Aufklärung eingesetzt werden.

(2) Für die Betroffenen muss jederzeit erkennbar sein,

1. ob und durch wen Datenverarbeitungsvorgänge auf dem mobilen Datenverarbeitungssystem oder durch dieses veranlasst stattfinden,

2. welche personenbezogenen Daten der betroffenen Person verarbeitet werden und

3. welcher Verarbeitungsvorgang im Einzelnen abläuft oder angestoßen wird.

Den Betroffenen müssen die Informationen nach Nummer 2 und 3 auf ihren Wunsch auch schriftlich in Papierform mitgeteilt werden.

(3) Die Betroffenen sind bei der Ausgabe des mobilen Datenverarbeitungssystems über die ihnen nach §5 zustehenden Rechte aufzuklären. Sofern zur Wahrnehmung der Informationsrechte besondere Geräte oder Einrichtungen erforderlich sind, hat die ausgebende Stelle dafür Sorge zu tragen, dass diese in angemessenem Umfang zur Verfügung stehen.

§29b Optisch-elektronische Überwachung

(1) Die nicht mit einer Speicherung verbundene Beobachtung öffentlich zugänglicher Bereiche mit optisch-elektronischen Einrichtungen ist zulässig, soweit dies der Wahrnehmung des Hausrechts dient und keine Anhaltspunkte dafür bestehen, dass schutzwürdige Interessen betroffener Personen überwiegen. Die Tatsache der Beobachtung ist, soweit nicht offenkundig, den Betroffenen durch geeignete Maßnahmen erkennbar zu machen.

(2) Die Speicherung von nach Absatz 1 Satz 1 erhobenen Daten ist nur bei einer konkreten Gefahr zu Beweiszwecken zulässig, wenn dies zum Erreichen der verfolgten Zwecke unverzichtbar ist. Die Daten sind unverzüglich zu löschen, wenn sie hierzu nicht mehr erforderlich sind; dies ist in angemessenen Zeitabständen zu prüfen.

(3) Werden die gespeicherten Daten einer bestimmten Person zugeordnet und verarbeitet, so ist diese jeweils davon zu benachrichtigen. Von einer Benachrichtigung kann abgesehen werden, solange das öffentliche Interesse an einer Strafverfolgung das Benachrichtigungsrecht der betroffenen Person erheblich überwiegt.

§30 Fernmessen und Fernwirken

(1) Öffentliche Stellen dürfen ferngesteuerte Messungen oder Beobachtungen (Fernmessdienste) in Wohnungen oder Geschäftsräumen nur vornehmen, wenn die betroffene Person zuvor über den Verwendungszweck sowie über Art, Umfang und Zeitraum des Einsatzes unterrichtet worden ist und nach der Unterrichtung schriftlich eingewilligt hat. Entsprechendes gilt, soweit eine Übertragungseinrichtung dazu dienen soll, in Wohnungen oder Geschäftsräumen andere

Wirkungen auszulösen (Fernwirkdienste). Die Einrichtung von Fernmess- und Fernwirkdiensten ist nur zulässig, wenn die betroffene Person erkennen kann, wann ein Dienst in Anspruch genommen wird und welcher Art dieser Dienst ist; dies gilt nicht für Fernmess- und Fernwirkdienste der Versorgungsunternehmen. Die betroffene Person kann ihre Einwilligung jederzeit widerrufen, soweit dies mit der Zweckbestimmung des Dienstes vereinbar ist. Das Abschalten eines Dienstes gilt im Zweifel als Widerruf der Einwilligung.

(2) Eine Leistung, der Abschluss oder die Abwicklung eines Vertragsverhältnisses dürfen nicht davon abhängig gemacht werden, dass die betroffene Person nach Absatz 1 Satz 1 oder 2 einwilligt. Verweigert oder widerruft sie ihre Einwilligung, so dürfen ihr keine Nachteile entstehen, die über die unmittelbaren Folgekosten hinausgehen.

(3) Soweit im Rahmen von Fernmess- oder Fernwirkdiensten personenbezogene Daten erhoben werden, dürfen diese nur zu den vereinbarten Zwecken verarbeitet werden. Sie sind zu löschen, sobald sie zur Erfüllung dieser Zwecke nicht mehr erforderlich sind.

§31 Nutzung von Verwaltungsdaten für die Erstellung von Statistiken

Für die Erstellung von Statistiken dürfen öffentliche Stellen personenbezogene Daten weiterverarbeiten, soweit diese bei der rechtmäßigen Erfüllung der in ihrer Zuständigkeit liegenden Aufgaben angefallen sind. Die Veröffentlichungen dürfen keine Angaben enthalten, die den Bezug auf eine bestimmte Person zulassen.

§32 Nutzung von Einzelangaben aus der amtlichen Statistik durch Gemeinden und Gemeindeverbände

(1) Dürfen den Gemeinden und Gemeindeverbänden auf Grund gesetzlicher Ermächtigungen zur Durchführung eigener statistischer Aufgaben Einzelangaben aus der amtlichen Statistik (Datensätze) für ihren Zuständigkeitsbereich übermittelt werden, so ist dies nur zulässig auf Datenträgern, die zur maschinellen Weiterverarbeitung bestimmt sind.

(2) Datenträger dürfen nur den für die Durchführung statistischer Aufgaben zuständigen Stellen der Gemeinden und Gemeindeverbände übermittelt werden, die organisatorisch und räumlich von den anderen Verwaltungsstellen der Körperschaft getrennt, gegen den Zutritt unbefugter Personen hinreichend geschützt und mit eigenem Personal ausgestattet sind, das die Gewähr für Zuverlässigkeit und Verschwiegenheit bietet, schriftlich auf das Statistikgeheimnis verpflichtet worden und während der Tätigkeit in der Statistikdienststelle nicht mit anderen Aufgaben des Verwaltungsvollzuges betraut ist.

(3) Die in den Statistikdienststellen der Gemeinden und Gemeindeverbände tätigen Personen dürfen die aus den nach Absatz 1 übermittelten Einzelangaben gewonnenen personenbezogenen Erkenntnisse während und nach ihrer Tätigkeit in der Statistikdienststelle nicht in anderen Verfahren oder für andere Zwecke verarbeiten oder offenbaren.

(4) Eine Durchführung eigener statistischer Aufgaben im Sinne des Absatzes

1 liegt nur vor, wenn aus den übermittelten Einzelangaben auf Grund vorgegebener sachlicher Kriterien Zahlensummen (Tabellen) erstellt werden, aus denen kein Bezug auf eine bestimmte Person hergestellt werden kann. Die Speicherung der übermittelten Einzelangaben in Dateien für andere als statistische Nutzungen und ihre Zusammenführung mit anderen Einzelangaben, aus denen ein Bezug zu personenbezogenen Daten hergestellt werden kann, sind unzulässig.

(5) Die Übermittlung nach Absatz 1 ist nach Zeitpunkt, Art der übermittelten Daten, Zweck der Übermittlung und Empfänger von der übermittelnden Dienststelle, nach Art und Zeitpunkt der Nutzung von der Dienststelle, die die Daten erhalten hat, aufzuzeichnen. Die Aufzeichnungen sind fünf Jahre aufzubewahren.

§32a Behördliche Datenschutzbeauftragte

(1) Öffentliche Stellen, die personenbezogene Daten verarbeiten, haben einen internen Beauftragten für den Datenschutz sowie einen Vertreter zu bestellen. Der Beauftragte muss die erforderliche Sachkenntnis und Zuverlässigkeit besitzen. Mehrere Stellen können gemeinsam einen Beauftragten für den Datenschutz bestellen, wenn dadurch die Erfüllung seiner Aufgabe nicht beeinträchtigt wird. Bei Bedarf kann eine Stelle auch mehrere Beauftragte sowie mehrere Vertreter bestellen. Der Beauftragte unterstützt die Stelle bei der Sicherstellung des Datenschutzes. Er berät die datenverarbeitende Stelle bei der Gestaltung und Auswahl von Verfahren zur Verarbeitung personenbezogener Daten und überwacht bei der Einführung neuer Verfahren oder der Änderung bestehender Verfahren die Einhaltung der einschlägigen Vorschriften. Er ist bei der Erarbeitung behördeninterner Regelungen und Maßnahmen zur Verarbeitung personenbezogener Daten frühzeitig zu beteiligen und hat die Einhaltung der datenschutzrechtlichen Vorschriften zu überwachen, die mit der Verarbeitung personenbezogener Daten befassten Personen mit den Bestimmungen dieses Gesetzes sowie den sonstigen Vorschriften über den Datenschutz vertraut zu machen und die Vorabkontrolle durchzuführen. Satz 5 findet auch Anwendung auf die Tätigkeit von Personalvertretungen, soweit bei diesen personenbezogene Daten verarbeitet werden.

(2) Der Beauftragte ist in seiner Eigenschaft als behördlicher Datenschutzbeauftragter der Leitung der öffentlichen Stelle unmittelbar zu unterstellen und in dieser Funktion weisungsfrei. Er darf wegen der Erfüllung seiner Aufgaben nicht benachteiligt werden. Während seiner Tätigkeit darf er mit keiner Aufgabe betraut sein, deren Wahrnehmung zu Interessenkollision führen könnte.

(3) Die verantwortliche Stelle ist verpflichtet, dem Beauftragten die Beschreibung aller automatisiert geführten Verfahren, in denen personenbezogene Daten verarbeitet werden, mit den nach §8 Abs. 1 vorgesehenen Angaben vorzulegen. Der Beauftragte führt das Verfahrensverzeichnis. Er gewährt jeder Person unentgeltlich nach Maßgabe des §8 Abs. 2 Einsicht in das Verfahrensverzeichnis. Das Einsichtsrecht in die Verfahrensverzeichnisse, die bei den in §2 Abs. 2 Satz 1 genannten Stellen geführt werden, kann verwehrt werden, soweit damit Betriebs- oder Geschäftsgeheimnisse offenbart würden. Wird keine Einsicht gewährt, ist in geeigneter Weise Auskunft zu erteilen; die Gründe

für die Verweigerung der Einsichtnahme sind aktenkundig zu machen und die einsichtverlangende Person ist darauf hinzuweisen, dass sie sich an den Landesbeauftragten für den Datenschutz wenden kann. Dem Landesbeauftragten für den Datenschutz ist auf sein Verlangen Einsicht in das Verfahrensverzeichnis zu gewähren.

(4) Bedienstete der öffentlichen Stellen können sich jederzeit in Angelegenheiten des Datenschutzes unmittelbar an den Beauftragten wenden. Der Beauftragte ist zur Verschwiegenheit über die Identität der betroffenen Person sowie über Umstände, die Rückschlüsse auf diese zulassen, verpflichtet, soweit er von der betroffenen Person davon nicht befreit wurde.

Straf- und Bußgeldvorschriften; Übergangsvorschriften

§33 Straftaten

(1) Wer gegen Entgelt oder in der Absicht, sich oder einen anderen zu bereichern oder einen anderen zu schädigen, entgegen den Vorschriften über den Datenschutz in diesem Gesetz oder in anderen Rechtsvorschriften des Landes Nordrhein-Westfalen personenbezogene Daten, die nicht offenkundig sind,

1. erhebt, speichert, zweckwidrig verwendet, verändert, weitergibt, zum Abruf bereithält oder löscht,

2. abruft, einsieht, sich verschafft oder durch Vortäuschung falscher Tatsachen ihre Weitergabe an sich oder andere veranlasst,

wird mit Freiheitsstrafe bis zu zwei Jahren oder mit Geldstrafe bestraft. Ebenso wird bestraft, wer unter den in Satz 1 genannten Voraussetzungen Einzelangaben über persönliche oder sachliche Verhältnisse einer nicht mehr bestimmbaren Person mit anderen Informationen zusammenführt und dadurch die betroffene Person wieder bestimmbar macht. Der Versuch ist strafbar.

(2) Absatz 1 findet nur Anwendung, soweit die Tat nicht nach anderen Vorschriften mit Strafe bedroht ist.

§34 Ordnungswidrigkeiten

(1) Ordnungswidrig handelt, wer entgegen den Vorschriften über den Datenschutz in diesem Gesetz oder in anderen Rechtsvorschriften des Landes Nordrhein-Westfalen personenbezogene Daten, die nicht offenkundig sind,

1. erhebt, speichert, zweckwidrig verwendet, verändert, weitergibt, zum Abruf bereithält oder löscht,

2. abruft, einsieht, sich verschafft oder durch Vortäuschung falscher Tatsachen ihre Weitergabe an sich oder andere veranlasst.

Ordnungswidrig handelt auch, wer unter den in Satz 1 genannten Voraussetzungen Einzelangaben über persönliche oder sachliche Verhältnisse einer nicht mehr bestimmbaren Person mit anderen Informationen zusammenführt und dadurch die betroffene Person wieder bestimmbar macht.

(2) Die Ordnungswidrigkeit kann mit einer Geldbuße bis zu 100.000 Deutschen Mark oder 50.000 Euro geahndet werden.

(3) Verwaltungsbehörde im Sinne des §36 Abs. 1 Nr. 1 des Gesetzes über Ordnungswidrigkeiten ist für die Verfolgung und Ahndung von Ordnungswidrigkeiten

a.nach den Absätzen 1 und 2 die Bezirksregierung,

b.nach §44 des Bundesdatenschutzgesetzes der Landesbeauftragte für den Datenschutz.

§35 Übergangsvorschriften

(1) Verarbeitungen personenbezogener Daten, die zum Zeitpunkt des In-Kraft-Tretens dieses Gesetzes1 bereits begonnen wurden, sind innerhalb von drei Jahren nach diesem Zeitpunkt mit den Vorschriften dieses Gesetzes in Übereinstimmung zu bringen.

(2) Für Behörden des Justizvollzuges gilt §18 mit der Maßgabe, dass die betroffene Person Auskunft oder Akteneinsicht erhält, soweit sie zur Wahrnehmung ihrer Rechte oder berechtigten Interessen auf die Kenntnis gespeicherter Daten angewiesen ist. §185 des Strafvollzugsgesetzes bleibt unberührt.

(3) Für Dateien, die bereits zum Register des Landesbeauftragten für den Datenschutz gemeldet sind, finden die Vorschriften des §8 Abs. 1 und des §32 a Abs. 3 erstmals in Fällen eintretender Veränderungen Anwendung. Im Übrigen wird die Dateienregisterverordnung vom 11. April 1989 (GV. NRW. S. 226) aufgehoben.

Literaturverzeichnis

[Agrawal et al., 1996] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., und Verkamo, A. I. (1996). Fast Discovery of Association Rules. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., und Uthurusamy, R., Hrsg., *Advances in Knowledge Discovery and Data Mining*, Kapitel 12, Seiten 307–328. AAAI Press/The MIT Press, Cambridge Massachusetts, London England.

[Baldi, 1998] Baldi, S. (1998). Electronic Commerce mit Software-Agenten. http://www.ebs.de/Lehrstuehle/Wirtschaftsinformatik/Lehre/E-Commerce/index98.htm, Abruf am 30.06.1999.

[Breese, 1998] Breese (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, Madison, WI, USA.

[Brockhausen und Morik, 1998] Brockhausen, P. und Morik, K. (1998). Wissensentdeckung in relationalen Datenbanken: Eine Herausforderung für das maschinelle Lernen. In Nakhaeizadeh, G., Hrsg., *Data Mining, theoretische Aspekte und Anwendungen*, Wirtschaftsinformatik, Seiten 193–211. Physica Verlag. http://www-ai.cs.uni-dortmund.de/DOKUMENTE/brockhausen_morik_98a.ps.gz.

[Brockhoff, 1987] Brockhoff, K. (1987). *Marketing durch Kunden-Informationssysteme*. Poeschel, Stuttgart.

[Brockhoff, 1993] Brockhoff, K. (1993). *Produktpolitik*. Fischer, Stuttgart, Jena, 3. Auflage.

[Daum, 2000] Daum, B. (2000). Anforderung an die Gestaltung von E-Business-Plattformen. http://www.utk.ch/archiv/2000/2/seit1922.htm, Abruf am 26.03.2000.

[Fayyad et al., 1996] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., und Uthurusamy, R., Hrsg. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press Series in Computer Science. A Bradford Book, The MIT Press, Cambridge Massachusetts, London England.

[Goldberg et al., 1992] Goldberg, D., Nichols, D., Oki, B., und Terry, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. In *Communications of the ACM, 35, 12, S. 61-70.*

[Gupta et al., 1999] Gupta, D., Digiovanni, M., Narita, H., und Goldberg, K. (1999). Jester 2.0: Evaluation of a New Linear Time Collaborative Filtering Algorithm. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval.*

[Janetzko und Zugenmaier, 2000] Janetzko, D. und Zugenmaier, D. (2000). Viele Gesichter. *Computer Technik*, (18).

[KPMG, 2000] KPMG (2000). One-To-One Marketing im Electronic Commerce – Status quo und Perspektiven. http://www.kpmg.de/library/surveys/, Abruf am 25.08.2001.

[Maltz und Ehrlich, 1995] Maltz, D. und Ehrlich, K. (1995). Pointing The Way: Active Collaborative Filtering, Human Factors in Computing Systems. In *CHI '95 Conference Proceedings*, Denver,Colorado, USA.

[Michalski, 1986] Michalski, R. (1986). Understanding the Nature of Learning. In Michalski, Carbonell, und Mitchell, Hrsg., *Machine Learning - An Artificial Intelligence Approach.* Morgan Kaufmann, Los Altos, California, USA.

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine Learning.* McGraw Hill, New York, USA.

[Morik et al., 2000] Morik, K., Wrobel, S., und Joachims, T. (2000). Maschinelles Lernen. In Görz, G., Hrsg., *Handbuch der Künstlichen Intelligenz.* Addison Wesley.

[Pazzani, 1998] Pazzani, M. (1998). A Framework for Collaborative, Content-Based and Demographic. *Artificial Intelligence Review.*

[Piller, 1999] Piller, F. T. (1999). Mass Customization als Wettbewerbstrategie. *it Industrielle Informationstechnik*, (1).

[Pine, 1993] Pine, B. (1993). *Mass Customization.* Harvard Business School Press, Boston, USA.

[PNP-Omline, 2000] PNP-Omline (2000). Click-Throuth: Bannerclick. www.pnp.de/magazin/comp/print/2000/2107/00013025.htm, Abruf am 25.08.2001.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Machine Learning. Morgan Kaufmann, San Mateo, CA.

[Rozycka et al., 2000] Rozycka, P., Zwirik, M., und Luczak, A. (2000). Collaborative Filtering — Kaufempfehlungen automatisiert. http://viadrina.euv-frankfurt–o.de/ sk/SS98/EC/coll_filter_the.html, Abruf am 26.03.2000.

[Runte, 2000] Runte, M. (2000). Personalisierung im Internet. Individualisierte Angebote mit Collaborative Filtering. Technical report, Lehrstuhl für Innovation, Neue Medien und Marketing an der

Christian-Albrechts-Universität zu Kiel, Kiel. http://linxx.bwl.uni-kiel.de/publications/runte/personalisierung_im_internet.pdf, Abruf am 16.10.2000.

[Schneider, 2000] Schneider, M. (2000). Seminararbeit Maschinelles Lernen in Data Mining und Information Retrieval. http://www.kbs.uni-hannover.de/ schneide/c45.html, Uni Hannover, Institut für Technische Informatik, Abruf am 03.05.2001.

[Shardanand, 1994] Shardanand, U. (1994). Social Information Filtering for Music Recommendation, Master's Thesis. Massachusetts Institute of Technology, Cambridge, MA, USA.

[Shardanand und Maes, 1995] Shardanand, U. und Maes, P. (1995). Social Information Filtering: Algorithms for Automating „Word of Mouth", Human Factors in Computing Systems. In *CHI '95 Conference Proceedings*, Denver, Colorado, USA.

[Simon, 1983] Simon, H. A. (1983). Why Should Machines Learn? In Michalski, R. S., Carbonell, J. G., und Mitchell, T. M., Hrsg., *Machine Learning — An Artificial Intelligence Approach*, Jgg. 1, Kapitel 2, Seiten 25–39. Morgan Kaufmann, Palo Alto, CA, USA.

[Snäbele, 1997] Snäbele, P. (1997). *Mass Customized Marketing*. Deutsche Universitatsverlag, Bamberg.

[Sonntag, 1998] Sonntag, M. (1998). Untersuchungen zur Personalisierung. Technical report, Johannes Kepler Universität Linz, Fachbereich Informatik, Linz. http://www.fim.uni-linz.ac.at/Aussendung10.98/Personalisierung.htm, Dissertation, Abruf am 16.10.2000.