3 Hypothesenfreie Datenauswertung – Data Mining

„Who could be expected to digest billions of records, each with tens or hundreds of fields? Yet the true value of such data lies in the users' ability to extract useful reports, spot interesting events and trends, support decisions and policy based on statistical analysis and inference, and exploit the data to achieve business, operational, or scientific goals." [80]

Gegenstand von Kapitel 3 ist das Gebiet des Data Mining, das sich mit der computergestützten Suche nach zuvor unbekannten Mustern in Daten befasst. Kap. 3.1 dient zur Einführung in dieses Forschungsgebiet, seiner Einbettung in das übergeordnete Gebiet Knowledge Discovery in Databases (KDD) sowie zur Abgrenzung der Data Mining-Untergebiete Text Mining, Multimedia Data Mining und Web Mining. In Kap. 3.2 werden mit Segmentierung, Klassifikation und Assoziation typische Verfahren des Data Mining sowie einige der hierfür eingesetzten Methoden vorgestellt. Kap. 3.3 stellt eine Auswahl von Anwendungsbeispielen des Data Mining zusammen; das Untergebiet Web Mining mit seinen Teilbereichen Web Content Mining, Web Structure Mining und Web Usage Mining wird in Kap. 3.4 behandelt. Kap. 3.5 geht auf die Bedeutung der Datenbereitstellung und Datenaufbereitung ein und verweist auf die datenschutzrechtlich kritisch zu bewertenden Tendenzen der Anwendung von Konzepten wie Data Mining und Data Warehouse auf personenbezogene Daten. Ein abschließendes Fazit wird in Kap. 3.6 gezogen.

3.1 Einführung

Vor dem Hintergrund beständig ansteigender Datenvolumen erweist sich eine zielgerichtete, auf vorhandenen Hypothesen basierende Auswertung von Daten zunehmend als ungeeignet. „Klassische" hypothesengestützte Formen der Datenauswertung, wie sie bspw. Anfragen an Datenbanken sowie den meisten statistischen Ansätzen zugrunde liegen, stoßen hier an Grenzen: Eine manuelle Auswertung großer Datenmengen ist zeitaufwendig, kostspielig und birgt die Gefahr einer hochgradigen Subjektivität; sie wird bei steigender Komplexität[81] der auszuwertenden Datenmengen zunehmend impraktikabel [Fayyad et al. 1996c].

Unter dem Begriff *Data Mining* [Fayyad et al. 1996b] werden Techniken zur computergestützten hypothesenfreien Suche nach verwertbaren Mustern oder Strukturen in Daten zusammengefasst. Data Mining hat in den letzten Jahren zunehmende Verbreitung bspw. zur Analyse von Kundendaten gefunden (vgl. Kap. 3.3) und kann als Ergänzung zu Auswertungen über SQL oder OLAP betrachtet werden [Multhaupt 2000, 47f.] [Steinecke 1999].

3.1.1 Einflussfaktoren

Data Mining ist nicht als völlige Neuentwicklung, sondern eher als eine Kombination von schon zuvor verfügbaren Technologien und Beiträgen vielfältiger anderer Forschungsgebiete aufzufassen [Schinzer et al. 1999, 99]. So bilden Entwicklungen im Bereich von Datenbanktechnologien und schneller Hardware eine Basis für eine effiziente Speicherung, Verwaltung und Verarbeitung großer Datenmengen; für die eigentliche Auswertung finden bspw. mit maschinellem Lernen, neuronalen Netzen oder genetischen Algorithmen Methoden aus der Künstlichen Intelligenz ebenso Eingang wie klassische statistische Verfahren

[80] Fayyad, U.; Piatesky-Shapiro, G.; Smyth, P.: *The KDD Process for Extracting Useful Knowledge from Volumes of Data* [Fayyad et al. 1996c].

[81] Datensammlungen werden sowohl bezogen auf die Anzahl der Datensätze wie auf die Attribute, die die einzelnen Datensätze charakterisieren, immer komplexer. So können bspw. astronomische Datenbanken Milliarden von Datensätzen enthalten und Datensätze zur medizinischen Diagnose aus hunderten oder tausenden Attributen bestehen [Fayyad et al. 1996c].

[Alpar, Niedereichholz 2000, 5]. Statistische Methoden sind ferner zur Überprüfung von Ergebnissen des Data Mining bedeutsam [Schinzer et al. 1999, 99f.]. Visualisierungstechniken (vgl. Kap. 4) wiederum können sowohl für eine erste Orientierung über die Ausgangsdaten [Multhaupt 2000, 103f.] wie für die Interpretation der gefundenen Muster eingesetzt werden [Fayyad et al. 1996b]; Konzepte der Informationsvisualisierung können zur visuellen Auffindung unbekannter Muster in Daten herangezogen werden und so eine Alternative oder Ergänzung der algorithmischen Mustersuche darstellen [Schumann, Müller 2000, 342].

3.1.2 Knowledge Discovery in Databases (KDD)

Der Prozess des rechnergestützten Auffindens unbekannter Strukturen in Daten wird seit Ende der 80er Jahre des 20. Jahrhunderts von der interdisziplinären Forschungsrichtung *Knowledge Discovery in Databases* (KDD) untersucht [Wiedmann et al. 2001]. Eine vielzitierte Definition des Begriffes KDD stammt von Usama Fayyad, Gregory Piatetsky-Shapiro und Padhraic Smyth:

- „*KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*" [82]

Gefordert ist damit der Einsatz von Algorithmen, die über Datenbankanfragen oder einfache statistische Auswertungen hinausgehen (*nontrivial*) und zur Entdeckung von Mustern (*patterns*) in den Daten führen, die über die untersuchte Datenmenge hinaus Gültigkeit besitzen (*valid*), zudem zuvor unbekannt waren (*novel*) sowie potentiell nützlich (*potentially useful*) und schließlich auch verständlich (*ultimately understandable*) sind [Alpar, Niedereichholz 2000, 4]. KDD umfasst neben der eigentlichen Musterfindung eine Vielzahl vorbereitender und nachfolgender Schritte; unter dem eigentlichen Data Mining wird dabei zunächst derjenige Teilschritt des gesamten KDD-Prozesses verstanden, der der algorithmischen Auffindung von Datenmustern dient (vgl. Kap. 3.1.3) :

- „*Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data.*" [Fayyad et al. 1996b]

In der Praxis wird allerdings unter Data Mining mittlerweile zumeist nicht mehr nur der eigentliche Analyseprozess verstanden; vielmehr wird der Begriff Data Mining nun für *alle* Teilschritte des KDD verwendet (vgl. bspw. [Alpar, Niedereichholz 2000, 4] [Wiedmann et al. 2001, 19]).

3.1.3 Der KDD-Prozess

Der KDD-Prozess gliedert sich in eine Vielzahl von Einzelschritten [Fayyad et al. 1996b] [Fayyad et al. 1996c], für deren Organisation verschiedene Prozessmodelle vorgeschlagen werden[83]. Abb. 3.1 gibt einen Überblick über wichtige Teilschritte, die nachfolgend kurz beschrieben werden.

- **Verständnis**

Auch für eine ungerichtete Datenanalyse ist ein Klärungsprozess über die angestrebten Ziele unerlässlich, der von [Fayyad et al. 1996c] auch als „*Learning the application domain*" bezeichnet wird. Den ersten grundlegenden Schritt des KDD-Prozesses bildet daher die

[82] Zitiert nach [Fayyad et al. 1996b]; als Originalquelle dieser Definition wird dort [Fayyad et al. 1996a] angegeben.
[83] Eine vergleichende Übersicht verschiedener KDD-Prozessmodelle findet sich in [Säuberlich 2000, 22ff.].

Entwicklung eines Verständnisses der Anwendungsdomäne, von bereits vorhandenem, relevanten Wissen sowie über das Ziel des Prozesses [Fayyad et al. 1996b].

- **Selektion**

Der zweite zentrale Schritt (Selektion) besteht in der Auswahl derjenigen Daten, auf denen die Analyse durchgeführt werden soll [Fayyad et al. 1996b]. Dies kann sowohl durch Einschränkung auf bestimmte Datensätze wie bestimmte Attribute erfolgen; zu beachten ist dabei allerdings, dass hier nicht solche Variablen ausgeschlossen werden, die zu einem nicht vermutetem Erkenntnisgewinn beitragen können. Die oft erforderliche weitere Reduktion des Datenvolumens kann durch eine Auswahl geeigneter Stichproben erreicht werden [Alpar, Niedereichholz 2000, 6] [Schinzer et al. 1999, 102].

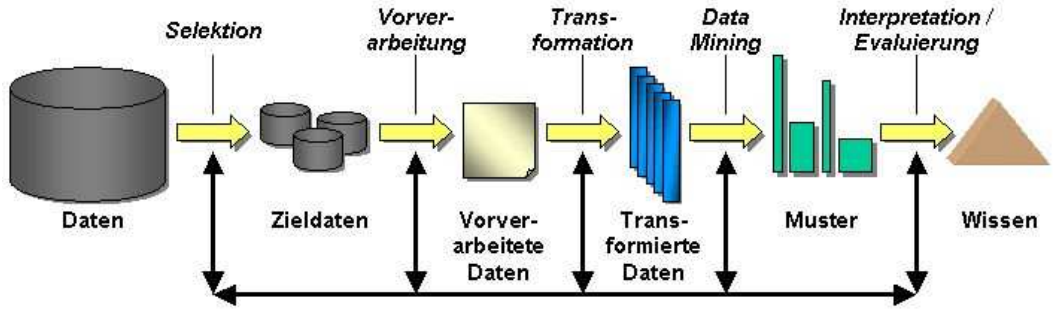

Abb. 3.1 - Überblick über die Schritte des KDD-Prozesses[84].

- **Vorverarbeitung**

Die Vorverarbeitung[85] dient zur Aufbereitung und Bereinigung der selektierten Daten. Je nach Qualität der Ausgangsdaten sind hierbei bspw. die Beseitigung doppelter Datensätze, die Behandlung von Ausreißern und Fehlwerten sowie inhaltliche Angleichungen erforderlich [Fayyad et al. 1996b] [Alpar, Niedereichholz 2000, 6] [Schinzer et al. 1999, 102].

- **Transformation**

Die Transformation umfasst Schritte zur Umwandlung der ausgewählten und bereinigten Daten in eine für die Analyse geeignete Form. Hierzu zählen bspw. die Normierung von Variablen, die Zusammenfassung inhaltlich abhängiger Attribute, die Erzeugung neuer Attribute wie Summen, Abweichungs- und Durchschnittswerte sowie die Konvertierung von Datenwerten in die von den Data Mining-Techniken benötigten Strukturen [Schinzer et al. 1999, 102]. Durch die hierzu erforderlichen Schritte geht in der Regel Information verloren, so dass sie sorgfältig durchgeführt werden müssen [Alpar, Niedereichholz 2000, 6f.]; die gewählte Transformation besitzt überdies Einfluss auf die nachfolgende Musterentdeckung (vgl. [Chamoni 1998b, 305]).

- **Data Mining**

An die Auswahl und Aufbereitung der Datenbasis schließt sich als nächster Schritt das eigentliche Data Mining an, das sich aus der Auswahl eines geeigneten Verfahrens und der Anwendung einer zu diesem Verfahren passenden Methode (vgl. Kap. 3.2) auf die Datenbasis zusammensetzt [Schinzer et al. 1999, 103] [Fayyad et al. 1996b].

- **Interpretation / Evaluierung**

Die gefundenen Muster müssen nun interpretiert und evaluiert werden; die so erhaltenen Ergebnisse können zu einer erneuten Iteration über den Prozess durch Rückkehr zu jedem

[84] Abbildung erstellt in Anlehnung an [Fayyad et al. 1996b, Fig.1] sowie [Fayyad et al.1996c, Fig.1].
[85] Der Begriff *Vorverarbeitung* wird von [Alpar, Niedereichholz 2000, 6] verwendet; [Fayyad et al. 1996b] sprechen von *data cleaning* sowie *preprocessing* (Datenreinigung und Vorverarbeitung), [Schinzer et al. 1999, 102] beschreiben diesen Schritt des KDD-Prozesses als *Datenfilterung* und -*reinigung*.

der bisher beschriebenen Schritte führen [Fayyad et al. 1996b]. So ist es nicht selten erforderlich, verschiedene Data Mining-Techniken mehrfach oder hintereinander einzusetzen; erhaltene Ergebnisse lassen bspw. Rückschlüsse auf die Qualität der Datenbasis zu oder liefern Erkenntnisse, die als Basis für den Einsatz anderer Techniken genutzt werden können [Schinzer et al. 1999, 103].

- **Verwendung**

Das über den KDD-Prozess gewonnenen Erkenntnisse müssen zunächst einer konsequenten Validierung unterzogen werden; hierzu zählt auch der Abgleich mit bisher vorhandenen Erkenntnissen und die Auflösung eventueller Konflikte. Die Nutzung der Ergebnisse kann schließlich bspw. in einer direkten Verwendung, ihrer Integration in die Datenbasis anderer Systeme, in einer reinen Dokumentation sowie ihrer Kommunikation an interessierte Dritte bestehen [Fayyad et al. 1996b] [Schinzer et al. 1999, 103].

3.1.4 Varianten des Data Mining

Während in der Literatur der Begriff Data Mining zumeist für die Analyse strukturierter Daten verwendet wird, haben sich daneben mit Text Mining, Multimedia Data Mining und Web Mining weitere Bereiche herausgebildet, die jeweils spezielle Arten von Daten adressieren.

- **Text Mining**

Text Mining bezeichnet die Anwendung von Data Mining auf Textsammlungen mit dem Ziel, in diesen implizit enthaltene, zuvor unbekannte Muster zu entdecken. Text Mining kann als Erweiterung des Information Retrieval angesehen werden [Alpar, Niedereichholz 2000, 5]; ebenfalls bestehen Bezüge zwischen Text Mining und dem Web Content Mining, einem der Unterbereiche des Web Mining [Pal et al. 2002] [Kosala, Blockeel 2000]. Multhaupt untersucht die Einsatzmöglichkeiten von Data Mining und Text Mining im strategischen Controlling. Er betont die enge Verwandtschaft zwischen Data Mining und Text Mining und führt aus, dass sich typische Data Mining-Verfahren wie Assoziation, Klassifikation oder Segmentierung (vgl. Kap. 3.2) unter Berücksichtigung der Besonderheiten von Texten auch auf diesen einsetzen lassen [Multhaupt 2000, 60ff.]. Gentsch und Diercks verweisen darauf, dass nach Schätzungen 80 Prozent der Unternehmensdaten in unstrukturierter Form vorliegen und dass durch die gemeinsame Analyse von strukturierten und unstrukturierten Daten (Kundendaten, Korrespondenzen, Aktennotizen, Börsennotierungen der Kunden aus Datenbanken, Dokumentenmanagementsystemen, HTML-Seiten etc.) ein erhöhter Erkenntnisgewinn zu erwarten sei. Sie prognostizieren, dass daher integrierte Systeme, die Data Mining und Text Mining verbinden, an Bedeutung gewinnen werden [Gentsch, Diercks 1999].

- **Multimedia Data Mining**

Als *Multimedia Data Mining* wird die Anwendung von Data Mining auf Multimedia-Daten bezeichnet, also bspw. die Suche nach Mustern bspw. in Bildern, Filmen oder Audiodaten. Kosala und Blockeel zufolge befindet sich Multimedia Data Mining trotz verschiedener Ansätze „*still in it's infancy*" [Kosala, Blockeel 2000]. Für einen - allerdings kurzen - Überblick über einige Ansätze sei auf [Zaiane et al. 1998] verwiesen, die auch einen eigenen, als MultiMediaMiner bezeichneten Prototyp zur Auswertung von Bildern und Videodaten vorstellen; auf zwei Anwendungen von Data Mining für die Auswertung wissenschaftlicher Bilddaten wird in Kap. 3.3 (Anwendungsbeispiele / Wissenschaftliche Daten) hingewiesen.

- **Web Mining**

Die Anwendung von Data Mining auf Daten des World Wide Web wird als *Web Mining* bezeichnet. Web Mining adressiert dabei ebenso ein verbessertes Auffinden von Informationen im World Wide Web (Web Content Mining), die Analyse der Hyperlinkstruktur zwischen Webseiten (Web Structure Mining) wie die Suche nach Nutzungsmustern in Daten von

Web-Servern (Web Usage Mining). In diesem Bereich kommen neben speziellen Ansätzen ebenfalls typische Data Mining Verfahren zum Einsatz, allerdings sind aufgrund der Spezifika der betrachteten Daten eine Vielzahl von Besonderheiten zu beachten. Auf das Gebiet des Web Mining wird daher in Kap. 3.4 ausführlicher eingegangen.

3.2 Verfahren und Methoden

Das Instrumentarium des Data Mining wird üblicherweise auf zwei Ebenen betrachtet, die hier als *Verfahren* sowie *Methoden* zu deren Umsetzung bezeichnet werden sollen[86]. Zunächst lassen sich verschiedene Verfahren des Data Mining unterscheiden, für deren Umsetzung wiederum unterschiedliche Methoden zur Verfügung stehen. Da einzelne dieser Methoden jeweils für unterschiedliche Fragestellungen verwendet werden können, finden sich in der Literatur z.T. uneinheitliche Abgrenzungen (vgl. bspw. [Alpar, Niedereichholz 2000, 9] sowie [Säuberlich 2000, 40ff.]). Auch hinsichtlich der Verfahren, die jeweils dem Data Mining zugerechnet werden, finden sich unterschiedliche Auffassungen[87]. Nach [Schinzer et al. 1999, 104ff.] lassen sich mit

- Segmentierung,
- Klassifikation und
- Assoziation

drei wesentliche Verfahren identifizieren, die nachfolgend gemeinsam mit einigen der hierfür eingesetzten Methoden skizziert werden.

3.2.1 Segmentierung

Ziel der *Segmentierung* ist die Aufteilung einer Menge von Objekten anhand ihrer Eigenschaften in zuvor nicht bekannte Gruppen. Auf diese Weise können in der Folge die Eigenschaften der identifizierten Gruppen analysiert werden (vgl. Kap. 3.3, Anwendungsbeispiele); derartigen Analysen verdanken bspw. die Begriffe Dinks und Yuppies[88] ihren Ursprung [Alpar, Niedereichholz 2000, 10]. Segmentierung eignet sich zudem auch als Vorstufe für weitere Data Mining-Verfahren, die dann auf kleineren, homogenen Ausschnitten arbeiten können [Schinzer et al. 1999, 104]. Nachfolgend sollen mit Clusteranalyse und Kohonennetzen zwei Methoden zur Segmentierung beispielhaft vorgestellt werden.

- **Clusteranalyse**

Clusteranalyse (vgl. bspw. [Multhaupt 2000, 81ff.] [Chamoni 1998b, 307ff.] [Schinzer et al. 1999, 108f.]) ist ein statistisches Verfahren zur Segmentierung. Es soll dabei eine Menge von Objekten so in Gruppen (Cluster) aufgeteilt werden, dass die Objekte in einem Cluster einander möglichst ähnlich und zugleich zu Objekten anderer Cluster möglichst unähnlich sind. Die Zuordnung von Objekten zu Clustern erfolgt durch Auswertung ihrer Ähnlichkeit zueinander, die durch Vergleiche der zugehörigen Werteausprägungen ihrer einzelnen Attribute ermittelt wird. Die Aufteilung der Objekte in Cluster kann dabei auf mehreren Wegen erreicht werden, wobei zwischen partitionierenden und hierarchischen Methoden unterschieden wird. Sog. *partitionierende Methoden* beginnen mit einer zufälligen Anfangsauf-

---

[86] Die Bezeichnung der beiden Ebenen wird dabei in der Literatur keineswegs einheitlich vorgenommen. So sprechen [Fayyad et al. 1996b] von *data-mining methods* und ihrer Umsetzung durch *data-mining algorithms* und *selection methods*, [Schinzer et al. 1999, 104ff.] unterscheiden zwischen *Verfahren* und *Techniken*, [Alpar, Niedereichholz 2000, 9ff.] zwischen *Aufgaben* und *Methoden*, [Multhaupt 2000, 60ff.] hingegen zwischen *Kernfunktionen* und *Methoden*.

[87] So unterscheiden etwa [Alpar, Niedereichholz 2000, 9] die *Aufgaben* Klassifikation, Segmentierung, Prognose, Abhängigkeitsanalyse und Abweichungsanalyse, [Multhaupt 2000, 60ff.] hingegen zwischen den *Kernfunktionen* Assoziation, Klassifizierung, Segmentierung, Modellbildung, Prognose und Zeitreihenanalyse.

[88] Dinks steht für Double Income No Kids, also kinderlose Doppelverdiener; Yuppies sind Young Urban Professionals.

teilung der Objekte in Cluster, die dann schrittweise durch Austausch von Objekten verbessert wird. *Hierarchi*sche Methoden arbeiten entweder agglomerativ, d.h. verschmelzend, oder divisiv, d.h. aufteilend. *Agglomerative* Methoden ordnen jedes Objekt zunächst einem eigenen Cluster zu und fassen danach ähnliche Cluster schrittweise zusammen; *divisive* Methoden ordnen hingegen zu Beginn alle Objekte einem einzigen Cluster zu, der dann schrittweise aufgeteilt wird (vgl. Abb. 3.2).

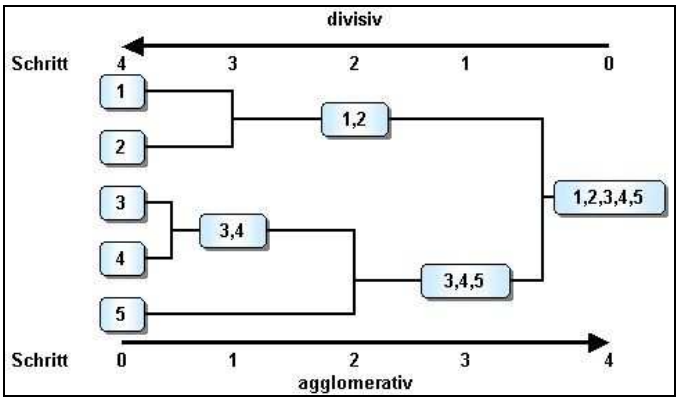

Abb. 3.2 - Varianten des hierarchischen Clustering: Divisives und agglomeratives Clustering[89].

Zusätzlich finden sich verschiedene weitere Methoden der Clusteranalyse. So ordnen Bayes-Verfahren (vgl. [Multhaupt 2000, 86f.]) Objekte den Clustern nicht aufgrund von Ähnlichkeiten, sondern mit Hilfe von Wahrscheinlichkeiten zu. Im Konzeptionellen Clustern (vgl. [Multhaupt 2000, 87ff.]), das zum Bereich des maschinellen Lernens gezählt wird, erfolgt die Zuordnung hingegen auf Basis von Konzepten, die es ermöglichen, zusätzliches Hintergrundwissen in die Klassifizierung einzubeziehen; Objekte werden hier nur in einem Cluster zusammengefasst, wenn sie durch dasselbe Konzept beschrieben werden. Fuzzy-Clustering-Methoden (vgl. [Multhaupt 2000, 109f.]) wiederum können eingesetzt werden, wenn eine trennscharfe Aufteilung der Objekte nicht geeignet ist.

- **Kohonennetze**

Eine weitere Methode, die im Rahmen des Data Mining zur Segmentierung eingesetzt wird, ist die Verwendung einer bestimmten Art neuronaler Netze, der sog. selbstorganisierenden Karten oder *Kohonennetze* (vgl. bspw. [Schinzer et al. 1999, 117] [Chamoni 1998b, 316ff.]). Kohonennetze basieren auf unüberwachtem Lernen; sie bestehen in der Regel aus einer sog. Eingabeschicht und eine zweidimensionalen Neuronenkarte, wobei der Lage der Neuronen eine entscheidende Bedeutung zukommt. Jedes Neuron der Eingabeschicht ist mit jedem Neuron der Neuronenkarte gewichtet verbunden; die Gewichte werden als den Abstand bestimmend interpretiert und iterativ berechnet, so dass sich die Neuronen auf der Karte anordnen und aufgrund ihrer Nähe zueinander als Cluster angesehen werden können.

3.2.2 Klassifikation

Durch das Verfahren der *Klassifikation*, auch als *Klassifizierung* bezeichnet, werden Objekte zu vorhandenen Klassen zugeordnet. Dies erfolgt basierend auf den Eigenschaften der Objekte wie der Merkmale der Klassen [Alpar, Niedereichholz 2000, 9]. Da den Klassen zuvor Eigenschaften zugewiesen wurden - bspw. potentiell zahlungsfreudige oder eher zahlungsunwillige Kunden - kann aus der Zuordnung eines Objektes zu einer Klasse auf die Eigenschaften dieses Objektes gefolgert werden (vgl. Kap. 3.3, Anwendungsbeispiele).

[89] Abbildung erstellt in Anlehnung an [Chamoni 1998b, 307, Abb. 2].

- **Entscheidungsbäume**

Entscheidungsbäume (vgl. bspw. [Multhaupt 2000, 71ff.] [Schinzer et al. 1999, 109ff.]) basieren auf induktivem Lernen und gehören damit in den Bereich des maschinellen Lernens. Entscheidungsbäume teilen dabei die einzelnen Objekte anhand von Regeln einer Klasse zu. Sie können dazu eingesetzt werden, um bisher unbeobachtete Objekte zu vorhandenen Klassen zuzuordnen; zudem können die einzelnen Klassen anhand der gefundenen Regeln beschrieben werden, so dass die Eigenschaften einer Klasse zugeordneten Objekte besser erkannt werden können. Ein Entscheidungsbaum (vgl. Abb. 3.3) besteht aus einem Wurzelknoten, diversen weiteren Entscheidungsknoten sowie Endknoten, die die Blätter des Baumes darstellen. Jeder Knoten repräsentiert ein Attribut, nach dessen Werten die Objekte aufgeteilt werden; jedes Blatt steht hingegen für eine Klasse. Objekte werden dadurch einer Klasse zugeteilt, in dem sie an jedem Knoten entsprechend ihres Wertes für dieses Attribut an den entsprechenden Unterbaum weitergeleitet werden, an dem sich das Verfahren mit einem anderen Attributwert wiederholt. Jedes Objekt wird so schließlich einem eindeutigen Blatt - und damit einer Klasse - zugeordnet; aus den Pfaden, die von der Wurzel des Baumes zu einem bestimmten Blatt führen, lassen sich dann die Regeln zur Klassifizierung ableiten. So wird ein Objekt dann zur Klasse 2 in Abb. 3.3 zugeordnet, wenn es für das Attribut Geschlecht den Wert *männlich* und zugleich für das Attribut Einkommen einen Wert größer als 50.000 besitzt.

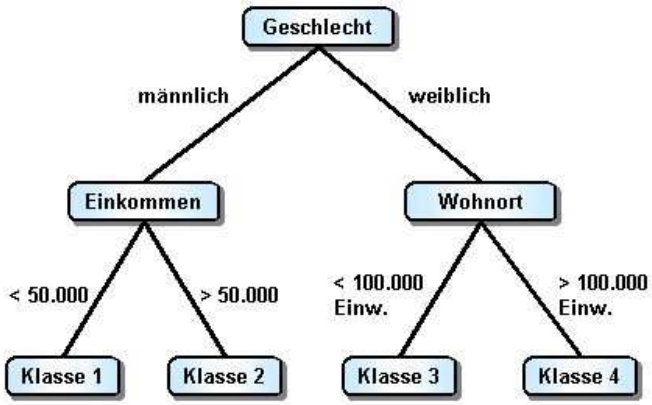

Abb. 3.3 - Schematische Darstellung eines binären Entscheidungsbaumes[90].

Entscheidungsbäume werden anhand entsprechender Algorithmen[91] konstruiert. Es wird dabei der kleinste Baum gesucht, der einen Satz von Trainingsdaten richtig klassifiziert. Dazu wird zunächst ein Attribut gesucht, das alle Objekte der Trainingsdaten richtig klassifiziert oder möglichst gleichmäßig aufteilt. Die Werte des so identifizierten Attributes werden in Intervalle aufgeteilt, um so die Verzweigung zu den nächsten Knoten zu bestimmen. Auf diese Weise werden die Trainingsdaten in Teilmengen aufgeteilt, für die der Prozess wiederholt wird. Der Prozess wird iterativ ausgeführt, bis alle in einer Teilmenge enthaltenen Fälle einer Klasse zugeordnet wurden oder die Teilmenge nicht weiter aufgeteilt werden kann.

- **Weitere Methoden**

Weitere Methoden, die im Rahmen des Data Mining zur Klassifikation eingesetzt werden, sind überwacht lernende neuronale Netze (sog. Multilayer-Perzeptron mit Backpropaga-

[90] Abbildung erstellt in Anlehnung an [Schinzer et al. 1999, 110, Abb. 6/6].

[91] Die verbreitetsten Algorithmen zum Aufbau von Entscheidungsbäumen sind nach Schinzer et al. CART (Classification and Regression Trees), CHAID (Chi-squared Automatic Interaction Detection) sowie ID3 (Iterative Dichotomiser 3) und dessen Nachfolger C4.5. Den Autoren zufolge ist zumindest einer dieser Algorithmen in den meisten Data Mining-Produkten enthalten [Schinzer et al. 1999, 112].

tion-Regel[92]) (vgl. [Wiedmann, Buckler 2001a] [Uerkvitz 2001] [Schinzer et al. 1999, 107ff.] [Chamoni 1998b]), fallbasiertes Schließen (Case Based Reasoning) (vgl. [Pfuhl 2000]) oder genetische Algorithmen (vgl. [Deventer, van Hoof 1998]).

3.2.3 Assoziation

Ziel der *Assoziation,* auch als *Assoziierung* oder *Abhängigkeitsanalyse* bezeichnet (vgl. [Multhaupt 2000, 63ff.] [Schinzer et al. 1999, 106] [Alpar, Niedereichholz 2000, 10]) ist es, Beziehungen entweder zwischen den Merkmalen eines Objektes oder zwischen mehreren Objekten zu identifizieren. Das typische Anwendungsfeld der Assoziation ist die Warenkorb-Analyse, bei der nach Mustern in den Einkäufen von Kunden gesucht wird (vgl. Kap. 3.3, Anwendungsbeispiele).

- **Assoziationsregeln**

Assoziation wird mit Hilfe von *Assoziationsregeln* (vgl. [Schinzer et al. 1999, 117ff.] [Multhaupt 2000, 63ff.]) ermittelt, deren Arbeitsweise hier anhand eines einfachen Beispieles verdeutlicht werden soll. Einer Getränkekette liegen Daten über die Einkäufe ihrer Kunden vor, es ist also bekannt, ob einzelne Kunden bspw. Saft, Whisky, Wein, Bier, Wasser etc. bei einem Einkauf erworben haben. Gesucht ist nun nach Mustern in diesen Daten, aus denen gefolgert werden kann, welche Produkte (etwa Saft und Wein) von Kunden häufig zusammen eingekauft werden. Es wird also in den Daten nach Regeln der Form *wenn X dann Y* gesucht (X → Y)[93], wobei der linke Teil der Regel als *Prämisse* und der rechte Teil als *Konklusio* bezeichnet wird. Zum automatischen Auffindung geeigneter Regeln kommen entsprechende Algorithmen[94] zum Einsatz. Um dabei solche Regeln zu identifizieren, die nicht nur für einige wenige Datensätze gelten, werden zwei Parameter zur Steuerung verwendet: Der *Support* einer Regel drückt aus, wie groß der Anteil von Datensätzen ist, für die die gefundene Regel gilt; je höher also der Support einer Regel ist, desto mehr Datensätze erfüllen den durch diese ausgedrückten Sachverhalt. Der Support einer Assoziationsregel X → Y errechnet sich dabei aus der Zahl der Datensätze, in denen sowohl X wie Y vorkommen, geteilt durch die Anzahl aller Datensätze. Die *Konfidenz* einer Regel drückt hingegen aus, wie viele Datensätze, die die Prämisse erfüllen, auch die Konklusio erfüllen. Die Konfidenz einer Assoziationsregel X → Y errechnet sich entsprechend durch den Support der Regel (X → Y) geteilt durch den Support von X. Wurden also bspw. bei 1000 betrachteten Einkäufen 600mal Saft und 200mal sowohl Saft wie Wein erworben, dann liegt der Support der Regel (*Saft → Wein*) bei 200/1000 und damit bei 20 Prozent, ihre Konfidenz bei 200/600 und damit bei 33 Prozent. Durch Festlegung von Werten für Mindestsupport und Mindestkonfidenz können so als hinreichend relevant erachtete Regeln automatisch aus den Daten herausgefiltert werden.

3.2.4 Anmerkungen

Die hier kurz dargestellten Methoden stellen nur einen Ausschnitt dar. Für einen ausführlichen Überblick über die verschiedenen Methoden des Data Mining sei bspw. auf [Säuberlich 2000] und [Multhaupt 2000] verwiesen. Ferner ist anzumerken, dass einige Methoden jeweils für bestimmte Data Mining Verfahren angewandt werden; so die Clusteranalyse für das Verfahren Segmentierung, Entscheidungsbäume für das Verfahren Klassifikation und Assoziationsregeln für das Verfahren Assoziation. Andere Methoden werden hingegen für mehrere Verfahren eingesetzt; so finden bspw. die unterschiedlichen Arten neuronaler

[92] Der Backpropagation-Algorithmus ist eine der am weitesten verbreiteten Lernmethoden für neuronale Netze (vgl. dazu detailliert [Rojas 1994, 72ff.]).
[93] Die Regel „wenn Saft gekauft wird, wird auch Wein gekauft", wird entsprechend durch *Saft → Wein* ausgedrückt. Die Regeln können komplexer sein, bspw. *Bier und Wasser → Grappa.*
[94] Ein typischer Vertreter ist der Apriori-Algorithmus (vgl. bspw. [Multhaupt 2000, 64ff.]).

Netze sowohl bei der Segmentierung wie der Klassifizierung Verwendung.

Bei der Auswahl von Methoden für ein Verfahren sind ferner weitere Faktoren zu berücksichtigen. Dazu zählen der jeweils erforderliche Vorbereitungs- und Transformationsaufwand für die einzelnen Methoden; so können bspw. Entscheidungsbäume auch nichtnumerische Werte verarbeiten, neuronale Netze bieten wiederum Fehlertoleranz bei verrauschten und verunreinigten Daten. Der Einsatz neuronaler Netze geht wiederum mit erhöhtem Aufwand einher und produziert Ergebnisse, die weniger leicht nachvollziehbar sind als die Resultate von Entscheidungsbäumen oder Assoziationsregeln [Schinzer et al. 1999, 107f.]. So stoßen nach [Multhaupt 2000, 80] Ergebnisse, die mit neuronalen Netzen ermittelt wurden, aufgrund ihrer geringeren Nachvollziehbarkeit oft auch auf geringe Akzeptanz.

3.3 Anwendungsbeispiele

Usama Fayyad stellt 1998 fest: „*The best applications of KDD are in fields that involve information-rich, rapidly changing environments, require knowledge-based decisions that have high payoff, and have sufficient and accessible data*" [Fayyad 1998]. Er sieht die wesentliche kommerzielle Anwendung in der Analyse von Kundendaten, um darauf basierend Angebote zuzuschneiden. In der Tat finden sich in der Literatur vielfältige Beispiele für den Einsatz von Data Mining zur Optimierung von Marketing und Vertrieb. Bevor typische kommerzielle Einsatzfelder vorgestellt werden, sollen zuvor einige andere Anwendungsgebiete erwähnt werden.

- **Wissenschaftliche Daten**

[Fayyad et al. 1996d, 1996e] beschreiben kurz einige Anwendungen des Data Mining zur Analyse wissenschaftlicher Daten. Zwei Anwendungen dienen zur Auswertung jeweils großer Mengen von Bilddaten der Weltraumforschung, der Identifikation von Vulkanen auf der Venus und der Klassifikation von Sternen und Galaxien. Ferner werden von den Autoren kurz Data Mining-Anwendungen zur Identifikation tektonischer Aktivitäten aus Satellitendaten, zum Auffinden bekannter und unbekannter Muster (Zyklone oder Tornados) in Atmosphärenmodell-Daten sowie zur Suche von Mustern in DNA-Sequenzen vorgestellt. Eine weitere Anwendung aus der Molekularbiologie - zur Untersuchung von HIV-Viren - wird in [Hofacker et al. 1996] beschrieben; zu Data Mining im Kontext von Medizin und Bioinformatik vgl. bspw. auch [Kastenmüller et al. 1998] [Houston et al. 1999] [Cook et al. 2001] [Geschwind 2002] [Deitcher 2002].

- **Pflegekinder und Basketball-Statistiken**

Als Beispiele für eher untypische Anwendungen des Data Mining verweist [Fayyad 1998] auf ein nicht näher bezeichnetes System zur Analyse der Akten von Pflegekindern im Staat Washington / USA mit dem Ziel, bessere Sozialeinrichtungen zu entwickeln, sowie auf den von mehreren NBA-Teams eingesetzten IBM Advanced Scout zur Analyse von Basketball-Statistiken.

- **Produktionsplanung und Produktionssteuerung**

Mögliche Einsatzbereiche des Data Mining im Bereich von Produktionsplanung und Produktionssteuerung liegen nach [Schinzer et al. 1999, 124f.] in der Prozess- und Qualitätskontrolle (Interpretation der Messergebnisse von Prüfverfahren sowie Identifikation von Einflussfaktoren auf die Produktqualität), in Anlagenüberwachung und -instandhaltung (genauere Abschätzung von Wartungszyklen sowie präventive Fehlererkennung) sowie in Bedarfsermittlung und Absatzprognose zur genaueren Abschätzung von Materialbedarf und Produktionsauslastung.

Den betriebswirtschaftlichen Anwendungsschwerpunkt von Data Mining sehen auch Schinzer et al. im Bereich von Vertrieb und Marketing, wo der Einsatz neuer Techniken zur Klas-

sifizierung, Segmentierung, Assoziiierung einen komperativen Wettbewerbsvorteil bedeuten könne [Schinzer et al. 1999, 123]. Die Mehrzahl der gesichteten Literatur behandelt Anwendungen in diesem Bereich; nachfolgend werden einige ausgewählte Einsatzbeispiele aufgeführt.

- **Markt- / Kundensegmentierung**

Eine Identifikation von Kundensegmenten kann bspw. dazu beitragen, das Verständnis eines Unternehmens für den Markt zu verbessern, den für ein Unternehmen relevanten Markt abzugrenzen, Kunden zielgruppengerecht anzusprechen, Instrumente des Marketing gezielter einzusetzen oder Neuprodukte effektiver zu positionieren [Saathoff 2000]. Die Segmentierung wird mit unterschiedlichen Methoden und auf unterschiedlichen Arten von Daten durchgeführt. Beispiele sind der kombinierte Einsatz von neuronalen Netzen und Clusteranalyse auf Kassenbon-Daten [Saathoff 2000] oder von Clusteranalyse auf Daten über das Nutzungsverhalten bei Produkten aus dem Bereich der Telekommunikation [Gossens 2000]. Auf die durch Segmentierung identifizierten Zielgruppen kann dann bspw. die Marketing-Kommunikation zugeschnitten werden, etwa durch den Einsatz von Spezialkatalogen für bestimmte Kundengruppen oder individualisierten Katalogen, die nur Teilsegmente des Gesamtangebotes enthalten [Schinzer et al. 1999, 127]; zum Einsatz neuronaler Netze zur Unterstützung des Zielkundenmarketing in der Automobilindustrie siehe [Wiedmann, Jung 2001].

- **Bonitätsprüfung**

Bonitätsprüfungen dienen dazu, Voraussagen über Zahlungswilligkeit bzw. Zahlungsfähigkeit von Kunden zu treffen. Hierzu wird in der Regel Klassifikation eingesetzt, um Kunden oder Firmen vorgegebenen Bonitätsklassen zuzuordnen und auf diese Weise zwischen „*guten und problembehafteten Kreditarrangements*" zu trennen [Schierreich 2000]. So wurde beim Versandhaus Neckermann zur täglichen Bonitätsprüfung von rund 8.000 Neukunden die bisher eingesetzten multivariaten statistischen Verfahren durch ein neuronales Netz ersetzt; auf diese Weise konnte die Vorhersagegenauigkeit von Zahlungsausfällen von 78 Prozent auf 79 bis 80 Prozent gesteigert werden. Dieser Wert mutet nur auf den ersten Blick gering an: Im Vergleich zu den zuvor verwendeten Verfahren werden nun pro Tag zusätzliche 80 Kunden richtig klassifiziert. Davon erhalten vier Kunden, die zuvor positiv bewertet worden wären, nunmehr keinen Kredit; hingegen werden 76 Kunden, die zuvor keinen Kredit erhalten hätten, nun als kreditwürdig eingestuft. Bei einem durchschnittlichem Bestellwert von 250 bis 2500 DM und 300 Bestelltagen im Jahr wird so durch Rückgang von Forderungsausfällen wie durch erhöhten Umsatz ein Mehrgewinn von mehreren hunderttausend bis mehreren Millionen DM im Jahr erzielt [Meinitz, Müller 1996] [Strüby 2001].

- **Betrugsentdeckung (fraud detection)**

Ein weiteres Einsatzgebiet liegt in der Suche nach Mustern betrügerischer Transaktionen in anfallenden Geschäftsdaten. So können bspw. Kreditkartenfirmen zunächst durch Segmentierung Kartennutzungsprofile erstellen und diese nachfolgend mit den aktuellen Geschäftsvorgängen abgleichen, um bei erkannten Abweichungen vom Profil, etwa im Ort der Nutzung oder im Verfügungsvolumen, entsprechende Maßnahmen einzuleiten. Nach Schinzer et al. werden an der New Yorker Börse täglich rund 900.000 Transaktionen durch ein Data Mining-System überwacht; innerhalb eines Jahres konnten so 400 Prozent mehr Betrugsfälle aufgedeckt werden als zuvor [Schinzer et al. 1999, 130].

- **Adressabgleich**

Im Versandhandel wird Adressmaterial aus unterschiedlichen Quellen verwendet. Ein Abgleich von Adressen durch Klassifikation mit neuronalen Netzen kann hier dazu eingesetzt werden, um die Adressen eigener Kunden nicht erneut anzumieten - ebenso, um Werbekosten zu sparen, wie um Kunden nicht durch Mehrfachansprache zu irritieren -, aber auch,

um vorhandene Adresslisten nach Kunden zu durchsuchen, die zuvor als zahlungsunwillig oder zahlungsunfähig eingestuft wurden [Schikowsky 2000].

- **Kaufbereitschaft und Zuordnung zu Einkommenskategorien**

Data Mining kann zu einer auf Klassifizierung basierenden Untersuchung des Einflusses verschiedener Produktattribute auf die Kaufbereitschaft eingesetzt werden; als Faktoren werden dabei die aktuelle Marktsituation, Eigenschaften des Produktes und Eigenschaften der Kunden einbezogen [Schinzer et al. 1999, 126]. Ebenfalls anhand von Klassifizierung kann eine Zuordnung von Kunden zu bestimmten Einkommenskategorien erfolgen. [Uerkvitz 2001] beschreibt den Einsatz neuronaler Netze, um potentielle Kunden anhand von Daten über Familienstand, Geschlecht, Altersgruppe und Position im Haushalt (der Kunde ist Haushaltsvorstand oder nicht) einzuordnen.

- **Kundenbindung**

Ein weiteres Einsatzgebiet des Data Mining liegt in der Vorhersage der Abwanderungswahrscheinlichkeit von Kunden durch Klassifikation. Basierend auf typischen Mustern in vorhandenen Daten bereits abgewanderter Kunden wird zunächst ein Vorhersagemodell entwickelt. Werden nachfolgend anhand einer regelmäßigen Datenanalyse für einen Kunden Abweichung vom gewohnten Verhalten festgestellt, wird daraus auf eine künftige Abwanderungswilligkeit geschlossen; damit besteht eine Option, eine Abwanderung durch gezielte Anreize zu verhindern [Schinzer et al. 1999, 129]. [Schneider et al. 2001] vergleichen in diesem Kontext den Einsatz unterschiedlicher Data Mining-Methoden für die Abwanderungsvorhersage im Kontext eines nicht näher bezeichneten Finanzdienstleisters.

- **Verbundkäufe / Warenkorbanalyse**

Eines der bekanntesten Einsatzgebiete des Data Mining ist die Warenkorbanalyse durch das Verfahren der Assoziation. Die Warenkorbanalyse strebt die Identifikation von Abhängigkeiten zwischen Produkten, Sortimenten oder Abteilungen an. Dazu wird bspw. durch Einsatz von Assoziationsregeln, zumeist basierend auf Daten, die von elektronischen Kassensystemen generiert werden, nach Abhängigkeiten zwischen Artikeln, Sortimenten oder Abteilungen gesucht [Schinzer et al. 1999, 128]. Die Ergebnisse können bspw. zur Ladengestaltung, für Personaleinsatz, Sortimentspolitik oder zielgerichtete Marketingaktionen eingesetzt werden [Schwarz 2000]. Nach [Schinzer et al. 1999, 128] ist der Einsatz von Data Mining zur Verkaufsförderung durch effektive Auswahl von Aktionsartikeln sowie die Platzierung von Artikeln besonders rentabel[95]: So sei es beispielsweise der Handelskette Walmart gelungen, nachdem eine Assoziation von Kosmetikartikeln und Grußkarten entdeckt wurde, durch eine gemeinsame Platzierung eine Umsatzsteigerung von rund 30 Prozent für diese Artikel zu erzielen.

- **Weitere Beispiele**

Weitere Beispiele für den Einsatz von Data Mining im betriebswirtschaftlichen Bereich sind Kundenunterstützung im Internet (Klassifikation anhand von Case Based Reasoning [Pfuhl 2000]), die Abstimmung der kommunikativen Maßnahmen zur Zielgruppenoptimierung in der Pharmaindustrie [Wiedmann, Buckler 2001b] oder die Ableitung von Prognosen über das künftige Kaufverhalten eines Kunden anhand seines Kaufverhaltens in der Vorsaison [Dastani 2000]. Zum Abschluss sei ein Ergebnis aus der Versicherungsindustrie genannt, dass auf einer Klassifikation über Entscheidungsbäume basiert[96] – hier lieferte Data Mining einen Hinweis darauf, dass Kunden, die mehr Versicherungsverträge abschließen, pro

[95] Nach [Schinzer et al. 1999] werden 50 bis 70 Prozent aller Kaufentscheidungen erst vor Ort während des Einkaufs getroffen.

[96] Der Aufbau des Entscheidungsbaumes basierte auf 407.392 Datensätzen mit 60 Attributen und benötigte 3,5 Stunden auf einem schnellem Parallelrechner [Aksu, Wittemann 2000].

Vertrag auch mehr Schäden verursachen [Aksu, Wittemann 2000].

3.4 Web Mining

3.4.1 Ziele und Untergebiete

Die Anwendung von Data Mining Techniken auf das World Wide Web zur Auffindung und Analyse nützlicher Informationen wird als *Web Mining* bezeichnet [Cooley et al. 1997a]. Zu den von Web Mining adressierten Zielen zählen beispielsweise [Kosala, Blockeel 2000]:

- **Verbessertes Auffinden relevanter Informationen im World Wide Web**

Die Suche nach spezifischen Informationen im World Wide Web erfolgt in der Regel entweder über ein interaktives Navigieren von Seite zu Seite (Browsing) oder durch die Verwendung von Suchmaschinen (Search Engines). Beide Ansätze erschließen die enorme Fülle der verfügbaren Quellen nur bedingt. Während ein interaktives Navigieren nur einen Bruchteil der Web-Ressourcen einbeziehen kann, sind Suchmaschinen sowohl durch ihre geringe Suchpräzision, die ihre Nutzer mit einer Vielzahl irrelevanter Ergebnisse konfrontiert, wie die nicht hinreichende Ausschöpfung der verfügbaren Quellen[97] limitiert.

- **Ableitung neuen Wissens aus im World Wide Web verfügbaren Informationen**

Eine weitere Herausforderung neben dem Auffinden relevanter Informationen im World Wide Web stellt ihre Auswertung zur Extrahierung potentiell nützlichen Wissens dar; mögliche Anwendungen liegen hier bspw. in der Erschließung von im World Wide Web vorhandenen Informationen zur Entscheidungsunterstützung.

- **Personalisierung der im World Wide Web verfügbaren Informationen**

Einen anderen Schwerpunkt des Web Mining bildet die Entwicklung von Werkzeugen, die eine den Vorlieben individueller Nutzer gemäße Erschließung verfügbarer Web-Ressourcen unterstützen, etwa durch automatisches Auffinden und Bereitstellen von Webseiten, die für die Interessen des Anwenders relevant sein können.

- **Verbessertes Verständnis von Website-Benutzern**

Während die bisher beschriebenen Aufgaben die Erschließung der potentiell über das World Wide Web verfügbaren Ressourcen adressieren, beschäftigt sich ein weiterer Schwerpunkt des Web Mining mit der Analyse des Nutzerverhaltens auf Websites. Dieser Bereich dient insbesondere kommerziellen Interessen, etwa zur Optimierung des Aufbaus der Struktur von Online-Shops oder zur automatischen Bereitstellung eines auf Kundengruppen oder individuelle Nutzer zugeschnittenen Angebotes.

Entsprechend der Bandbreite seiner Ziele setzt sich Web Mining aus unterschiedlichen Teilbereichen zusammen. Der Ansatz von [Cooley et al. 1997a] zur Systematisierung der verschiedenen Ansätze des Web Mining unterscheidet zunächst zwischen

- *Web Content Mining* (der automatischen Informationssuche im Web) sowie
- *Web Usage Mining* (dem Auffinden von Nutzungsmustern in Daten von Web-Servern).

In neueren Arbeiten [Kosala, Blockeel 2000] [Pal et al. 2002] wird neben diesen zusätzlich noch der Bereich des

- *Web Structure Mining* (Analyse der Hyperlinkstruktur zwischen Webseiten)

unterschieden. Nachfolgend werden diese drei Untergebiete des Web Mining charakterisiert.

[97] Das World Wide Web wird aufgrund seiner Größe von Suchmaschinen nicht erst dann durchsucht, wenn eine Suchanfrage gestellt wird. Die Datenbasis von Suchmaschinen bildet eine vorab erfolgte Indizierung von Webseiten; die Grenzen dieser Indizierung sind damit zugleich die Grenzen des von ihnen in Anfragen einbezogenen Ausschnitts des gesamten World Wide Web. Dabei wird offensichtlich jeweils nur ein Bruchteil des World Wide Web abgedeckt; so kommen [Lawrence, Giles 1999] zu dem Ergebnis, dass keine Suchmaschine mehr als 16 Prozent des World Wide Web indiziert.

3.4.2 Web Content Mining

In den Bereich des *Web Content Mining* fallen Ansätze zum automatischen Auffinden von Informationen im World Wide Web [Cooley et al. 1997a], wobei sowohl eine direkte Analyse des Inhaltes von Webseiten (Web Page Content Mining*)* wie eine Weiterverarbeitung von Suchmaschinen-Ergebnissen *(*Search Result Mining*)* erfolgen kann (vgl. [Säuberlich 2001]). Damit bestehen beispielsweise sowohl enge Bezüge des Web Content Mining zu den Aufgaben von Suchmaschinen [Pal et al. 2002] wie zum Bereich des Information Retrieval und des Text Mining [Kosala, Blockeel 2000] [Pal et al. 2002]. Ein Beispiel für die Vielfalt der sich bei der Analyse von Webseiten stellenden Aufgaben ist die Arbeit von [Wang, Hu 2002], die eine automatische Unterscheidung von HTML-Tabellen adressiert, um Tabellen, für deren Inhalt eine tabellarische Struktur semantisch bedeutsam ist, gegen andere, die nur einem übersichtlichen Layout dienen, abzugrenzen. Die Techniken des Web Content Mining können nach [Cooley et al. 1997a] in agentenbasierte Ansätze und Datenbank-Ansätze unterschieden werden:

- **Agentenbasierte Ansätze**

Agentenbasierte Ansätze verfolgen die Entwicklung autonomer oder halbautonomer Softwaresysteme, die zur Unterstützung konkreter Nutzeranforderungen Informationen im World Wide Web entdecken und aufbereiten sollen [Pal et al. 2002]. Diese Ansätze lassen sich dabei in drei Kategorien aufteilen: „Intelligente" Suchagenten durchforsten das World Wide Web nach relevanten Informationen unter Einbeziehung von Nutzerprofilen sowie Hintergrundwissen über die abzufragenden Quellen, eine zweite Gruppe von Agenten adressiert die automatische Filterung und Kategorisierung von Informationen aus dem World Wide Web. Personalisierte Webagenten hingegen sollen die Vorlieben von Nutzern erlernen und darauf basierend relevante Informationen im World Wide Web entdecken [Cooley et al. 1997a].

- **Datenbank-Ansätze**

Datenbank-Ansätze adressieren die Transformation von semistrukturierten Daten im World Wide Web in strukturiertere Datensammlungen, gegen die nachfolgend Datenbankabfrage- oder Data Mining-Techniken zur Analyse angewendet werden können [Cooley et al. 1997a]. Ansätze aus diesem Bereich umfassen bspw. die Extraktion von Schemata oder die Ableitung strukturierter Zusammenfassungen aus semistrukturierten Daten (sog. DataGuides), die Erstellung von mehrschichtigen Datenbanken mit einem von Schicht zu Schicht zunehmenden Generalisierungsgrad der gespeicherten Informationen oder die Entwicklung von speziellen Anfragesprachen für das World Wide Web (für einen Überblick über entsprechende Arbeiten siehe [Kosala, Blockeel 2000]).

3.4.3 Web Structure Mining

Während Web Content Mining die Analyse der Webdokumente in den Vordergrund stellt, ist das Ziel des *Web Structure Mining* [Kosala, Blockeel 2000] [Pal et al. 2002] die Auswertung der Hyperlinkstruktur, die die Verbindungen zwischen den Seiten im World Wide Web darstellt (Inter Document Structure). So kann - in Analogie zu bibliographischen Zitaten - bspw. eine große Zahl von Links, die auf eine bestimmte Seite zeigen, einen Indikator für ihre Popularität darstellen; zum anderen können viele Links, die von einer Seite ausgehen, einen Hinweis auf die eventuelle Reichhaltigkeit oder Vielschichtigkeit ihres Inhaltes geben. Derartige Informationen können bspw. als Basis für eine Bewertung der Relevanz von Webseiten benutzt werden; ein entsprechender Ansatz liegt dem PageRank-Algorithmus zugrunde, der in der Suchmaschine Google zur Festlegung einer Rangliste von Suchergebnissen verwendet wird [Brin, Page 1998].

Die Auswertung spezifischer Links auf einer Seite kann überdies Rückschlüsse auf die Organisation von Information geben; so kann bspw. eine hohe Zahl von lokalen Links[98] indizieren, dass sich integrierte Informationen auf einer Website, aber verteilt auf verschiedene Dateien befinden, während eine hohe Zahl von internen Links[99] auf das Vorhandensein relevanter Informationen in der gleichen Datei hindeuten [Madria et al. 1999]. Weitere Anwendungsgebiete des Web Structure Mining bilden bspw. die Analyse von Hyperlinkstrukturen zur Identifikation von Nutzergruppen (Communities) im World Wide Web und der für diese bedeutsamen Webseiten [Kumar et al. 1999] oder die Entwicklung spezieller Web Crawler[100], die nur solche Hyperlinks verfolgen, die anhand von Beispieldokumenten als relevant für vorgegebene Themen identifiziert werden (sog. Focused Crawling, vgl. [Pal et al. 2002]).

3.4.4 Web Usage Mining

Web Usage Mining beschäftigt sich mit der Analyse des Nutzungsverhaltens von Besuchern einer Website. Typische Anforderungen sind bspw. die Identifikation der am häufigsten aufgesuchten Seiten, von bevorzugten Navigationspfaden oder typischen Merkmalen zur Beschreibung der Nutzer [Säuberlich 2001]. Solche Informationen werden insbesondere von E-Commerce-Anbietern eingesetzt [Pal et al. 2002] und können bspw. für die Optimierung der Struktur eines Webauftritts, die Auswertung der Effizienz von bestimmten Marketingstrategien oder Angebotskampagnen sowie für die Personalisierung von Websites durch automatische Bereitstellung von auf spezifische Kundengruppen zugeschnittenen Angeboten verwendet werden [Cooley et al. 1997a]. Bedeutsam im Kontext einer solchen Personalisierung sind auch sog. Collaborative Recommender, die zur Bereitstellung maßgeschneiderter Angebote durch eine Analyse von Ähnlichkeiten und Unterschieden in den Vorlieben von Nutzern einer Website dienen sollen [Pal et al. 2002].

- **Anwendungen und Methoden**

Anwendungen des Web Usage Mining lassen sich in das Erkennen von Nutzungsmustern sowie ihre Analyse unterteilen [Cooley et al. 1997a] [Srivastava et al. 2000] [Kosala, Blockeel 2000]. Nach Säuberlich lassen sich mit der Analyse von Zugriffspfaden, der Identifikation von Nutzerprofilen sowie der Beschreibung und Prognose von Nutzertypen drei wichtige Fragestellungen unterscheiden und jeweils durch bestimmte Data-Mining-Methoden adressieren. Um Muster bei der Nutzung von Websites zu extrahieren, werden die jeweils gewählten Zugriffspfade, also die Navigation zwischen einzelnen Seiten analysiert. Hierzu können Assoziationsverfahren herangezogen werden, um bspw. Beziehungen zwischen verschiedenen Angeboten eines Webauftritts zu analysieren. Eine Identifikation von Nutzerprofilen kann hingegen mittels Segmentierung durch Clusteranalyse oder selbstorganisierende Kohonen-Karten erfolgen. Um eine genauere Beschreibung der erkannten Nutzergruppen zu erhalten, können Entscheidungsbäume verwendet werden; für die Zuordnung neuer Nutzer zu bestehenden Klassen schließlich können Regressionsverfahren oder neuronale Netze eingesetzt werden [Säuberlich 2001].

Anders als bei den bisher beschriebenen Web Mining-Bereichen werden beim Web Usage Mining sog. *sekundäre* Webdaten analysiert, die aus der Interaktion von Nutzern mit dem World Wide Web entstehen [Kosala, Blockeel 2000]. Zu diesen Daten zählen bspw. die Log-Dateien von Web-Servern oder Proxy-Servern, ebenso Aufzeichnungen von Nutzer-

[98] Lokale Hyperlinks verweisen von einem Webdokument auf andere Dokumente, die über den selben Web-Server zur Verfügung gestellt werden.
[99] Interne Hyperlinks verweisen auf andere Stellen innerhalb des selben Webdokuments.
[100] Web Crawler sind Programme, die in einem rekursiven Prozess das Web durchwandern, indem sie die Hyperlinks eines Webdokumentes entdecken, auf die so referenzierten Dokumente zugreifen, deren Hyperlinks entdecken und verfolgen etc.

profilen, Bookmarks, Mausklicks, Scrollingbewegungen etc. [Pal et al. 2002]. Dabei nimmt - wie generell beim Data Mining - auch beim Web Usage Mining die Aufbereitung der Daten vor der eigentlichen Analyse einen erheblichen Anteil des Gesamtaufwandes ein [Säuberlich 2001]. Dass die Spezifika der hier einzubeziehenden Daten zudem spezielle Formen der Aufbereitung erfordern, verdeutlichen die drei Aufgabenbereiche Datenbereinigung, Identifikation der Zugriffe sowie Identifikation der Nutzer.

- **Bereinigung**

Relevant für die Analyse im Web Usage Mining sind nur solche Ressourcen einer Website, die vom Benutzer explizit angefordert wurden. Dabei ist zu beachten, dass der Abruf einer HTML-Seite von einem Web-Server zu mehr als einer Anforderung an diesen - und entsprechend zu mehr dokumentierten Zugriffen - führen kann. Wird bspw. eine HTML-Seite angefordert, in die drei Bilder eingebunden sind, werden diese Bilder ebenfalls vom Web-Server angefordert, so dass dieser für einen Abruf dieser Seite insgesamt vier Zugriffe verzeichnet [Pitkow 1997]. Entsprechend müssen vor einer Analyse zunächst diese sog. auxilary requests, die automatisch beim Aufruf einer Seite mit zum Client geladen werden, herausgefiltert werden [Säuberlich 2001]. Eine solche Bereinigung kann durch Überprüfung der Suffixe der angeforderten URLs auf Endungen wie .gif oder .jpg etc. erreicht werden [Cooley et al. 1997a].

- **Identifikation der Zugriffe**

Ein wesentliches Problem im Kontext des Web Usage Mining bildet die Identifikation solcher Zugriffe auf Webdokumente, die nicht vom Web-Server aufgezeichnet werden. Dies kann zunächst durch lokales Caching von Webdokumenten durch den Browser des Nutzers bedingt sein; in diesem Fall kann ein Nutzer eine gegebene Seite mehrfach aufrufen, ohne dass ein Zugriff über den Web-Server der entsprechenden Site erforderlich ist. Eine weitere Ursache für nicht vom Web-Server dokumentierte Zugriffe bilden die Caches von Proxy-Servern, über die mehrere Nutzern auf ein Webdokument zugreifen können, das nur einmal vom Web-Server angefordert wurde [Pitkow 1997]. Für die Vervollständigung von Zugriffspfaden, die bspw. auch aufgrund von Sprüngen durch den Zugriff über Bookmarks notwendig sein kann, können sog. Referer-Logs verwendet werden, in denen dokumentiert ist, welches Webdokument zuvor aufgerufen wurde [Säuberlich 2001].

- **Identifikation der Nutzer**

Auch die Zuordnung von Zugriffen zu jeweiligen Nutzern anhand der in den Logfile-Daten des Web-Servers dokumentierten IP-Adressen, von denen aus jeweilige Seiten angefordert wurden, ist aus mehreren Gründen erschwert. So ist es möglich, dass mehrere Nutzer vom selben Rechner aus zugreifen oder Zugriffe von dynamisch verteilten IP-Adressen aus erfolgen [Säuberlich 2001]. Zudem erschweren Proxy-Server die Identifikation individueller Nutzer, so dass die Zuordnung von Requests zu einzelnen Nutzern allein über die anfordernde IP-Adresse zu einer fälschlichen Zusammenfassung mehrerer Nutzer zu einem einzigen führen kann [Cooley et al. 1997a]. Zur Unterscheidung von Nutzern mit der gleichen IP-Adresse kann einbezogen werden, ob jeweils von unterschiedlichen Betriebssystemen oder verschiedenen Browsern zugegriffen wurde [Säuberlich 2001]. Der Ansatz von [Pirolli et al. 1996] ordnet eine angeforderte Seite dann dem selben Nutzer zu, wenn sie über einen Hyperlink von einer Seite erreichbar ist, die zuvor innerhalb einer bestimmten Zeitspanne von der selben IP-Adresse angefordert wurde. [Cooley et al. 1997b] stellen in diesem Zusammenhang einen Ansatz vor, um Web-Serverzugriffe, die der Navigation dienen, von solchen, die inhaltsbezogen erfolgen, zu unterscheiden. Eine verbesserte Nutzeridentifikation wird auch durch die Verwendung von Cookies[101] oder Nutzerregistrierun-

[101] Cookies bieten die Möglichkeit, auf dem Rechner eines im World Wide Web navigierenden

gen angestrebt; allerdings erweisen sich beide Ansätze als deutlich limitiert – Cookies können nutzerseitig deaktiviert oder gelöscht werden, die Verlässlichkeit von freiwilligen Nutzerangaben im World Wide Web ist begrenzt[102] [Pitkow 1997].

3.5 Data Mining, Data Warehouse und Datenschutz

„Kritiker des Data Mining merken zurecht an, daß in historisch gewachsenem Datenmüll keine Nuggets zu finden sind." [103]

„Die Europäische Datenschutzrichtlinie spricht grundsätzlich jeder Person das Recht zu, keiner belastenden automatisierten Einzelentscheidung unterworfen zu werden (Art. 15). ‚Data Mining' ist ein Instrument, das für solche Entscheidungen herangezogen werden kann." [104]

3.5.1 Data Warehouse als Ausgangsbasis für Data Mining

Die Bereitstellung einer geeigneten Datenbasis ist für erfolgreiches Data Mining von wesentlicher Bedeutung. Vor dem Hintergrund vorhandener Datenheterogenitäten verwundert es daher nicht, dass das eigentliche Data Mining zwar den methodisch anspruchsvollsten Teil des KDD-Prozesses darstellt, jedoch die Datenvorbereitung und -bereitstellung als dessen aufwendigster Bestandteil eingestuft wird [Multhaupt 2000, 57]. Für diesbezügliche Aufgaben sind nach [Steinecke 1999] 80 Prozent des Zeitbudgets, nach [Aksu, Wittemann 2000] etwa 80 Prozent des Gesamtaufwandes, nach [Alpar, Niedereichholz 2000] etwa 75 bis 85 Prozent der Gesamtanstrengungen und nach [Schinzer et al. 1999, 101] nicht selten bis zu 90 Prozent des Zeitaufwandes und der Kosten eines Data Mining-Projektes zu veranschlagen.

Um den Aufwand für Data Mining-Projekte zu reduzieren, verweisen viele Autoren daher übereinstimmend auf die zentrale Bedeutung eines Data Warehouse[105] zur Bereitstellung einer geeigneten Datenbasis (vgl. bspw. [Schinzer et al. 1999, 102f.] [Multhaupt 2000, 57ff.] [Alpar, Niedereichholz 2000, 14ff.] [Wiedmann et al. 2001, 29ff.]).

3.5.2 Interessenkollisionen

Eine Behandlung von neuen und zunehmend eingesetzten Technologien wie Data Warehouse und Data Mining wäre nicht vollständig ohne einige Anmerkungen zu den datenschutzrechtlichen Herausforderungen, die durch diese aufgeworfen werden. Bereits 1991 weist der Berliner Datenschutzbeauftragte darauf hin, dass die sich abzeichnenden technologischen Entwicklungen mit ihren sich beständig steigernden Möglichkeiten der Datenverarbeitung den Menschen „*zunehmend transparent*" machen [Gesetze zum Datenschutz 1991, 7]. So heißt es etwa im Bundesdatenschutzgesetz (BDSG), das den Datenschutz in den öffentlichen Stellen des Bundes sowie die Verarbeitung personenbezogener Daten im gesamten Bereich der Privatwirtschaft regelt[106], in der Fassung vom 20.12.1990, Paragraph 1:

Anwenders Informationen abzulegen und wieder auszulesen; auf diese Weise können individuelle Nutzerdaten, bspw. das Datum des letzen Besuches einer spezifischen Webseite, ausgewertet werden.

[102] Eine Befragung von über 14.500 Web-Nutzern ergab, dass 33 Prozent der Befragten mindestens einmal und über 10 Prozent der Befragten sogar bei einem Viertel der von ihnen ausgefüllten Online-Registrierungen falsche Angaben gemacht haben [Pitkow 1997].

[103] Chamoni,P; Gluchowski,P.: *Analytische Informationssysteme – Einordnung und Überblick* [Chamoni, Gluchowski 1998, 20] (Schreibweise im Original).

[104] Aus der Entschließung der 59. Konferenz der Datenschutzbeauftragten des Bundes und der Länder vom 14./15. März 2000 [KDBA 2000].

[105] Auch OLAP-Datenbanken können hierfür prinzipiell herangezogen werden; allerdings ist zu berücksichtigen, dass die Ausrichtung solcher Daten auf bestimmte Bearbeitungszwecke sowie ihr Verdichtungsgrad die Möglichkeiten eines Data Mining einschränken können [Multhaupt 2000, 59].

[106] Vgl. [Gesetze zum Datenschutz 1991, 46].

- *§1 (1) Zweck dieses Gesetzes ist es, den einzelnen davor zu schützen, daß er durch den Umgang mit seinen personenbezogenen Daten in seinem Persönlichkeitsrecht beeinträchtigt wird.* [107]

Die aufgeführten Anwendungsbeispiele verdeutlichen, dass Data Mining von Unternehmen häufig auf personenbezogenen Daten durchgeführt wird und damit das Recht auf informationelle Selbstbestimmung des Einzelnen tangiert. Hier können schnell die Interessen von Unternehmen mit den Interessen einzelner Personen kollidieren. So weist Möller darauf hin, dass beispielsweise eine Klassifizierung als mehr oder weniger zahlungswilliger Kunde schnell zu einer Stigmatisierung des Betroffenen führen kann, wenn dieser an immer mehr Stellen als unzuverlässig gespeichert ist. Als Beispiel für einen fragwürdigen Umgang mit Daten führt Möller den Online-Buchhandel Amazon.com an, der ohne Einwilligung der Betroffenen auf seiner Website Statistiken über die Lesegewohnheiten seiner Kunden veröffentlichte, so dass offengelegt wurde, welche Literatur bspw. von Mitarbeitern von Unternehmen oder Behörden bevorzugt erworben wurde. Möller zufolge wurden auf diese Weise recht private Details bekannt, bspw. dass Angestellte bei National Semiconductors angeblich häufig „*101 Nights of Grrreat Sex*" bestellten, während Angestellte von Larry Ellisons Datenbankkonzern Oracle bevorzugt auf „*The Difference between God and Larry Ellison*" zurückgriffen [Möller 1999].

3.5.3 Stellungnahme der Datenschutzbeauftragten

Die durch Data Warehousing und Data Mining gegebenen Möglichkeiten werden auch von den Datenschutzbeauftragten kritisch beobachtet. In der Entschließung der 59. Konferenz der Datenschutzbeauftragten des Bundes und der Länder vom 14./15. März 2000 heißt es:

> „[...] *Diese Entwicklung schafft neben Vorteilen neue Gefahren und Risiken für das Grundrecht auf informationelle Selbstbestimmung und für den Schutz der Privatheit: Persönlichkeitsprofile, automatisierte Vorhersagen von Verhaltens- und Handlungsweisen, Manipulationsmöglichkeiten und zu lange Speicherung sind befürchtete Gefahren.*
>
> *Die Konferenz der Datenschutzbeauftragten weist auf Folgendes hin:*
>
> - *Nach dem grundrechtlichen Gebot der Zweckbindung dürfen personenbezogene Daten nur im Rahmen der gesetzlich zugelassenen Zwecke oder der gegenseitigen Vereinbarungen verwendet werden. Eine personenbezogene Speicherung in einem allgemein verwendbaren Data Warehouse entfernt sich vom ursprünglichen Verwendungszweck und stellt eine Speicherung auf Vorrat ohne Zweckbindung dar. Personenbezogene Daten, die bei der öffentlichen Verwaltung vorhanden sind, sind in ihrer Zweckbestimmung grundrechtlich geschützt und dürfen nicht für unbestimmte Zwecke in einem ‚Daten-Lagerhaus' gesammelt werden.*
> - *Eine Zweckänderung ist nur mit Einwilligung der Betroffenen zulässig, nachdem diese über die Tragweite der Einwilligung aufgeklärt worden ist. Eine Einwilligung in unbestimmte und zeitlich unbegrenzte Zweckänderungen ist deswegen unwirksam.*
> - *Gestaltung und Auswahl von Datenverarbeitungs-Systemen haben sich an dem Ziel auszurichten, keine oder so wenig personenbezogene Daten wie möglich zu verarbeiten. Anonyme und pseudonyme Verfahren sind datenschutzrechtlich unbedenklich.*

[107] [Gesetze zum Datenschutz 1991, 52] (Schreibweise im Original).

- *Verfahren sind so zu gestalten, dass die Betroffenen hinreichend unterrichtet werden, damit sie jederzeit die Risiken abschätzen und ihre Rechte wahrnehmen können. Sie haben insbesondere das Recht, eine erteilte Einwilligung jederzeit zurückzuziehen.*
- *Die gesetzlichen Speicherfristen, nach deren Ablauf die Daten zwingend archiviert oder gelöscht werden müssen, sind strikt zu beachten. Deswegen ist die Einrichtung von permanenten ‚Daten-Lagerhäusern' rechtswidrig.*
- *Die Europäische Datenschutzrichtlinie spricht grundsätzlich jeder Person das Recht zu, keiner belastenden automatisierten Einzelentscheidung unterworfen zu werden (Art. 15). ‚Data Mining' ist ein Instrument, das für solche Entscheidungen herangezogen werden kann.*

Die Konferenz der Datenschutzbeauftragten des Bundes und der Länder ruft die Hersteller und Anwender von ‚Data Warehouse'- und ‚Data Mining'-Verfahren dazu auf, solchen Programmen den Vorzug zu geben, die unter Einsatz von datenschutzfreundlichen Technologien die Speicherung von personenbezogenen Daten durch Anonymisierung oder Pseudonymisierung vermeiden." [KDBA 2000].

3.5.4 Klassifizierung potentieller Terroristen – Total Information Awareness

Welche Ausmaße künftig der Einsatz von Data Mining-Technologien zur Auswertung personenbezogener Daten annehmen kann, verdeutlicht das von den USA in der Folge der Anschläge auf das World Trade Center initiierte Total Information Awareness (TIA) - Projekt, später umbenannt in Terrorism Information Awareness. Hierzu wurde die Defence Advanced Research Projects Agency (DARPA) vom Pentagon beauftragt, zur Terrorismusbekämpfung ein gigantisches Data Mining System zu etablieren, dass neben Daten der Geheimdienste und des FBI auch private in- und ausländische Datenbanken einbeziehen und private Kommunikation ebenso wie kommerzielle Transaktionen nach verdächtigen Mustern durchsuchen sollte:

„*Wir entwickeln Technologien und ein Prototyp-System, um die Möglichkeit der USA zu revolutionieren, ausländische Terroristen zu entdecken, zu klassifizieren und zu identifizieren, ihre Pläne zu entziffern und so die USA instand zu setzen, rechtzeitig zu handeln, um erfolgreich terroristischen Akten zuvorzukommen und sie nieder zu schlagen.*" [108]

„*Wenn Terrororganisationen Angriffe auf die USA planen und durchführen, müssen deren Mitglieder Transaktionen vornehmen. Dabei hinterlassen sie Spuren in diesem Informationsraum.*" [109]

Nachdem bereits im Jahr 2002 das geplante Terrorist Information and Prevention System (TIPS) eingestellt wurde, wurde nach anhaltenden Protesten auch dem TIA-Programm 2003 mit der Verabschiedung des Pentagon-Haushaltes für 2004 sämtliche finanziellen Mittel entzogen; Beobachter vermuten jedoch, dass an diesem und ähnlichen Konzepten in verschiedenen anderen Projekten weitergearbeitet werden wird [Pichet 2003ab] [Rötzer 2003, 2002abc] [Palm 2002].

[108] J. Walker, Sprecher der DARPA, zitiert nach [Rötzer 2002b] (Schreibweise im Original).
[109] J. Poindexter, Leiter der DARPA-Abteilung Information Awareness Office, zitiert nach [Rötzer 2002b].

3.6 Fazit

Data Mining kombiniert verschiedene Techniken, die jeweils eine ungerichtete, hypothesenfreie Datenauswertung unterstützen. Typische Auswertungsziele liegen in einer Aufteilung von Datenräumen durch Identifikation entsprechender Segmente, der automatischen Zuordnung neuer Datensätze zu vordefinierten Gruppen durch Klassifizierung oder dem Auffinden von Zusammenhängen durch Assoziation. Data Mining kann im Rahmen eines interaktiven Prozesses zur Entdeckung bisher unbekannter Mustern in Daten führen, die dann von jeweiligen Experten des Anwendungsgebietes zu bewerten sind. Hier wird am ehesten deutlich, welche Rolle Data Mining bei der Auffindung neuer Zusammenhänge spielen kann. Durch die entsprechenden Algorithmen können Datenmengen, die aufgrund von Größe oder Komplexität basierend auf Hypothesen nicht mehr geeignet ausgewertet werden können, schnell nach Zusammenhängen und Mustern durchsucht werden. Die Bewertung der Nützlichkeit der so in den Daten gefundenen Strukturen allerdings erfordert umfassende Fachkenntnisse über das jeweilige Anwendungsgebiet, um interessante und potentiell wertvolle Muster von unbrauchbaren zu unterscheiden und verwertbare Erkenntnisse ableiten zu können; Data Mining kann hier also allenfalls unterstützend wirken und potentiell interessante Hinweise geben, nicht jedoch Expertenwissen und menschliche Erfahrung ersetzen. Nicht zuletzt sind Konzepte wie Data Mining und Data Warehousing aufgrund von Tendenzen zu ihrer Anwendung auf personenbezogene Daten mit Zielen, die von der Absatzoptimierung bis zur automatischen Klassifizierung potentieller Straftäter reichen, aus datenschutzrechtlichen Gründen auch kritisch zu bewerten.