NR. 8

Heinz Lothar Grob
Frank Bensberg

Das Data-Mining-Konzept

INSTITUT FÜR WIRTSCHAFTSINFORMATIK DER WESTFÄLISCHEN WILHELMS-UNIVERSITÄT MÜNSTER
STEINFURTER STR. 107, 48149 MÜNSTER, FON (0251) 83-38000, FAX (0251) 83-38009
E-MAIL: GROB@UNI-MUENSTER.DE
http://www-wi.uni-muenster.de/aw/

Juni 1999

Inhalt

1 Das Data-Mining-Konzept

Der Einsatz betriebswirtschaftlicher Anwendungssysteme zur Unterstützung und Durchführung operativer Geschäftsprozesse konfrontiert Unternehmungen zunehmend mit dem Phänomen der *Datenflut*.[1] Dieses Phänomen ist auf mehrere Einflußfaktoren zurückzuführen. So unterliegen die Datenbanken der operativen Systeme einem „natürlichen Wachstum". Während die Gewinnung von Neukunden durch Marketingmaßnahmen eine Vergrößerung der Kundendatenbank verursacht, führt das Abwandern von Kunden i. d. R. *nicht* zum Löschen des entsprechenden Kundendatensatzes. Dieser Sachverhalt, der in vielen operativen Subsystemen[2] zu beobachten ist und die Asymmetrie zwischen Datengewinnung und Datenvernichtung verdeutlicht, fördert das *vertikale Wachstum* der operativen Datenbanken.[3] Neben dieser Wachstumsursache, die sich auf die Quantität der abgebildeten Sachverhalte bezieht, ist auch das *horizontale Wachstum* betrieblicher Datenbanken zu beobachten. Dieses Phänomen ist darauf zurückzuführen, daß aufgrund der zunehmenden Umweltkomplexität die Anzahl der zu erfassenden Fakten steigt.[4] Dadurch erhöht sich die Menge der Attribute bzw. der Dimensionen, die zur adäquaten Beschreibung von unternehmensrelevanten Sachverhalten erforderlich sind.[5] Die beiden dargestellten Phänomene beschreiben das Wachstum strukturierter unternehmens*interner* Datenbanken.

Neben den intern erzeugten Daten stehen Unternehmungen auch unternehmensexterne Datenquellen zur Verfügung. So weist das Internet mit über 5 Mio. Web-Servern[6] ein informatorisches Potential auf, das einem exponentiellen Wachstum unterliegt.[7] Zusammenfassend sind die Typen des Datenwachstums in Abb. 1 dargestellt worden.

---

1 Gebräuchlich ist in diesem Zusammenhang auch der Begriff der *Informations*flut. Vgl. Grob, H. L., Bieletzke, S. (1997), S. 29. In angloamerikanischen Publikationen wird bisweilen der Begriff der Datenüberschwemmung (data glut) genutzt, der die Dramatik der Situation unterstreicht. Vgl. Fayyad, U., Djorgovski, S. G., Weir, N. (1996), S. 471.

2 Anzuführen ist hier beispielsweise der Personalbereich oder das Rechnungswesen. Aufgrund rechtlicher Bestimmungsfaktoren erfolgt hier die Speicherung und Archivierung historischer Daten.

3 Der Begriff des vertikalen Wachstums verdeutlicht, daß es sich faktisch um eine *Verlängerung* der bestehenden Datenbanken handelt.

4 Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996b), S. 39.

5 So ist im Kontext der Kundendatenbank die Erfassung neuer Kontaktdaten (z. B. E-Mail-Adressen) oder die Abbildung neuer Zahlungsformen (z. B. Kreditkarten) erforderlich geworden.

6 Diese Angabe beruht auf einer Online-Analyse im April 1999. Vgl o. V. (1999).

7 Vgl. Behme, W., Muksch, H. (1998), S. 85.

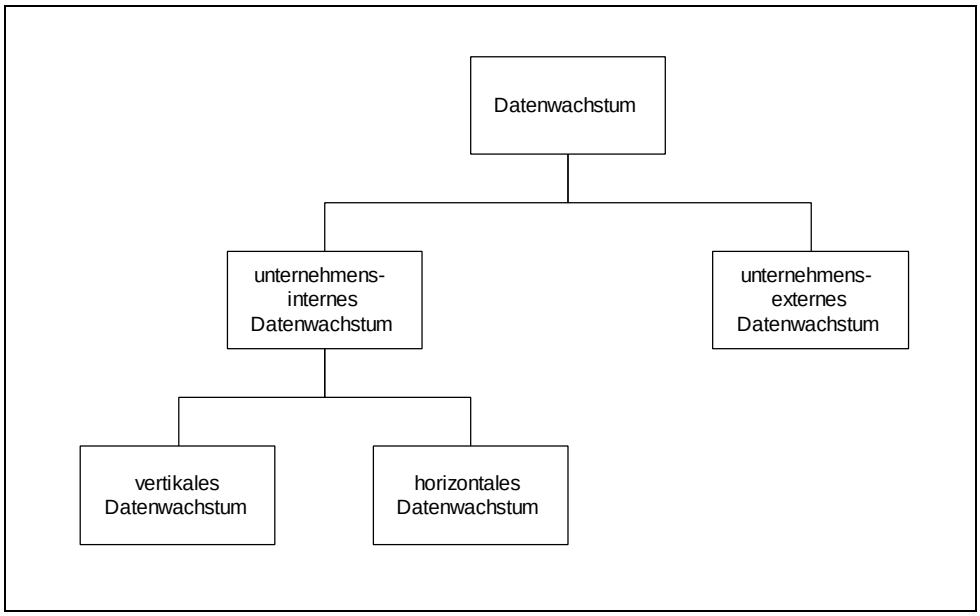

Abb. 1: Typen des Datenwachstums

Das Phänomen der Datenflut führt dazu, daß das Informationsangebot für das Management zunimmt und die Versorgung des Management mit relevanten Daten zunehmend schwieriger wird. Die Evolution der entscheidungsunterstützenden Systeme hat in den 80er Jahren zu der Klasse der Executive Information Systems (EIS)[1] geführt, die sich als datenorientierte Entscheidungsunterstützungssysteme einer gewissen Akzeptanz beim Management erfreuen durften.[2] Bei umfangreichen Datenvolumina und einer hohen Anzahl relevanter Dimensionen führen datenorientierte Systeme jedoch zur Informationsüberlastung[3] des Entscheidungsträgers, die potentiell eine geringere Entscheidungsqualität zur Folge hat. Hinzu kommt, daß durch das explosive Wachstum des Internet ein externer Datenpool zur Verfügung steht, dessen Daten i. d. R. in unstrukturierter Form vorliegen. Für die Aufbereitung und Nutzung dieser externen Daten sind EIS-Systeme nur bedingt geeignet.[4]

Vor diesem Hintergrund steigt die Attraktivität methodenorientierter Systeme, mit denen große Datenbestände analysiert und relevante Zusammenhänge identifiziert werden können. Mit ihrer Entwicklung und Anwendung beschäftigt sich der Forschungsbereich des Data Mining.

Der Begriff des Data Mining stammt aus dem Bereich der Statistik und besitzt in seiner ursprünglichen Verwendung abwertende Bedeutung, da unter Data Mining die selektive Metho-

---

1 Vgl. Henneböle, J. (1995), S. 2.

2 Vgl. Gluchowski, P., Gabriel, R., Chamoni, P. (1997), S. 201 ff.

3 Vgl. Picot, A., Reichwald, R., Wigand, R. T. (1998), S. 86-87.

4 Der Einsatz von EIS zur Integration von internetbasierten Daten ist dann sinnvoll, wenn es sich um wohlstrukturierte Daten handelt, die importiert und weiterverarbeitet werden können.

denanwendung zur *Bestätigung* vorformulierter Hypothesen verstanden wurde.[1] In diesem Sinne besaß Data Mining lediglich die Funktion, Aussagen pseudowissenschaftlich zu bestätigen und unternehmenspolitische Handlungen oder sogar Theorien zu rechtfertigen. Die Bedeutung des Data-Mining-Begriffs hat sich mittlerweile einem tiefgreifenden Wandel unterzogen. So wird Data Mining von MARKOWITZ und LIN XU im Rahmen der Portfoliotheorie angewendet, um Methoden zur Mischung von Portfolios auf der Basis von Vergangenheitsdaten zu bewerten.[2] „The practice of examining many methods and recommending the one that historically did best is commonly referred to as data mining."[3] Indes ist dieser Fokus für die Definition des Data-Mining-Begriffs unzweckmäßig, da in historischen Daten Informationen enthalten sein können, die ohne den Kontext der Methodenbewertung für den Anwender relevant sind.

Eine alternative Definition liefern MERTENS u. a., die den Begriff des Data Mining mit Datenmustererkennung gleichsetzen. Unter Data Mining wird demnach ein Prozeß verstanden, „... der aus einer Datenmenge implizit vorhandene, aber bisher unentdeckte, nützliche Informationen extrahiert."[4] Diese Definition impliziert einen breiteren Anwendungskontext, schränkt aber Data Mining auf die Entdeckung von Mustern ein, über die der Anwender bis zum Analysezeitpunkt kein Wissen besitzt. Da die Analyse vermuteter oder beobachteter Zusammenhänge zu den Aufgaben entscheidungsunterstützender Systeme zu rechnen ist, erscheint diese Definition als zu eng. Ein weiteres Problem stellt sich aus dem Anforderungskriterium der Nützlichkeit der gewonnenen Informationen. Da der subjektive Informationsnutzen nur ex post im Anwendungskontext bewertet werden kann, erscheint es nicht sinnvoll, das Kriterium der Nützlichkeit im Data-Mining-Begriff definitorisch zu verankern.[5]

BERRY und LINOFF sehen Data Mining als Exploration und Analyse großer Datenmengen mit automatischen oder semi-automatischen Werkzeugen, um bedeutungsvolle Muster und Regeln aufzufinden.[6] Hier findet die Einschränkung des Anwendungsbereiches von Data Mining auf große Datenmengen statt. Dieses Merkmal trifft zwar häufig auf empirische Anwendungskontexte zu, doch wird eine definitorische Verankerung nicht für adäquat gehalten, da auch kleine Datenmengen für den Anwender bedeutungsvolle Muster enthalten können. Die Definition erlaubt keine Abgrenzung zu traditionellen Auswertungs- und Berichtssystemen, die ebenfalls den allgemeinen Aufgabenstellungen der Exploration und der Analyse dienen.

---

1 Vgl. Berry, J. A., Linoff, G. (1997), S. 4; Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 3-4.

2 Vgl. Markowitz, H. M., Lin Xu, G. (1994), S. 60 ff.

3 Markowitz, H. M., Lin Xu, G. (1994), S. 61.

4 Mertens, P. u. a. (1994), S. 740.

5 Einen Überblick über Ansätze zur Bewertung der Interessantheit von Mustern liefern Müller, M., Hausdorf, C., Schneeberger, J. (1998), S. 248 ff.

6 Vgl. Berry, J. A., Linoff, G. (1997), S. 5.

FAYYAD, PIATETSKY-SHAPIRO und SMYTH definieren Data Mining als die Anwendung von Algorithmen auf Daten mit der Zielsetzung, Muster aus den Daten zu extrahieren.[1] Da keine einschränkenden Spezifika verwendet werden, weist diese Definition zwar das höchste Maß an Allgemeingültigkeit auf, erlaubt allerdings keine Abgrenzung zu verwandten Forschungsbereichen, wie beispielsweise der explorativen Statistik.

Auf der Basis der angeführten Definitionen soll Data Mining hier generisch und prozeßorientiert definiert werden. So wird Data Mining als *integrierter Prozeß* verstanden, der durch die Anwendung von Methoden auf einen Datenbestand Muster identifiziert. Der Integrationsaspekt bedeutet in diesem Zusammenhang, daß alle erforderlichen Schritte von der Datenbeschaffung über die Methodenanwendung bis hin zur Präsentation der Muster dem Data-Mining-Prozeß zuzurechnen sind. Diese Definition macht deutlich, daß Data-Mining-Systeme einen größeren Problembereich und damit einen breiteren Adressatenkreis abdecken als klassische Systeme der computergestützten Datenanalyse, deren Primärfunktion in der Methodenanwendung liegt und zu deren Anwenderkreis Personen mit fundierten statistischen Kenntnissen zu rechnen sind. So wurden im Rahmen des Data Mining effiziente Algorithmen entwikkelt, die die Analyse vollständiger Populationen ermöglichen und das Stichprobenparadigma der Statistik obsolet machen.[2] Darüber hinaus erlauben Data-Mining-Systeme nicht nur die Überprüfung vorformulierter Hypothesen, sondern können bei Anwendung adäquater Methoden diese auch selbständig generieren.[3] Zur Präsentation von Mustern hat sich der Bereich des grafischen Data Mining herausgebildet, der eigene Visualisierungstechniken einsetzt. Die Beschränkung der klassischen Visualisierungstechniken auf drei Dimensionen wird durch den Einsatz von Parallelkoordinatensystemen oder Glyphen[4] aufgehoben.

Die angeführten Eigenschaften zeigen, daß Data Mining nicht mit der traditionellen computergestützten Datenanalyse konkurriert, sondern die Synthese derselben mit Teildisziplinen der Wirtschaftsinformatik als Zielsetzung hat. Dieser integrative Charakter des Data Mining hat dazu geführt, daß sich mittlerweile ein Markt für Data-Mining-Systeme als auch für entsprechende Beratungsleistungen herausgebildet hat.[5] Im wissenschaftlichen Bereich läßt sich eine sprunghafte Zunahme der Publikationen beobachten, die sich mit dieser Thematik beschäftigen.[6] Diese Entwicklung hat allerdings dazu geführt, daß sich mehrere Definitionen herausgebildet haben. Neben den Bezeichnungen Knowledge Extraction, Information Harvesting, und Data Archaeology wird häufig der Begriff des Knowledge Discovery in Databases

---

1 Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 4.

2 Vgl. Fayyad, U. (1997), S. 6.

3 Vgl. Berry, J. A., Linoff, G. (1997), S. 64.

4 Glyphen stellen Eigenschaften von Objekten durch geeignete visuelle Codierungen (Farb- oder Formgebung) dar. Vgl. Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 608.

5 Eine Auswahl der verfügbaren kommerziellen und nicht-kommerziellen Data-Mining-Systeme ist zu finden unter http://www.kdnuggets.com.

6 Vgl. Tuzhilin (1997), S. 1; Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 3.

(KDD) verwendet, um den dargestellten Objektbereich zu kennzeichnen.[1] Die Vertreter der KDD-Anhänger verstehen unter Data Mining lediglich den Vorgang der Methodenanwendung auf den Datenbestand und integrieren Data Mining in einen übergreifenden KDD-Prozeß.[2]

Um den Ausdrücken Data Mining und Knowledge Discovery in Databases eine sinnvolle, konfliktfreie Bedeutung zuzuweisen, sollten die Begriffe Daten („data") und Wissen („knowledge") voneinander abgegrenzt werden. Zu diesem Zweck erfolgt die Einordnung in den Bezugsrahmen der Semiotik, unter der die Lehre von den Zeichensystemen, ihren Strukturen und den Beziehungen zu den dargestellten Gegenständen zu verstehen ist.[3] Die Begriffsabgrenzungen der Semiotik und die aufeinander aufbauenden semiotischen Dimensionen werden in Abb. 2 dargestellt.

Bezugsebene	Semiotische Dimension	Wissenschaftstheoretischer Aspekt	Begriffe
Physikalische Ebene	Syntaktik	formalwissenschaftlich	Zeichen
Bedeutungsebene	Sigmatik	formalwissenschaftlich	Daten
Bedeutungsebene	Semantik	realwissenschaftlich	Nachricht
Wirkungsebene	Pragmatik	realwissenschaftlich	Information

Abb. 2: Semiotische Begriffsabgrenzung[4]

Die Formulierung von Mustern wie beispielsweise „Wenn Produkt Nr. x gekauft wird, dann wird auch Produkt Nr. y mit einer Wahrscheinlichkeit von z % gekauft" geschieht letztlich durch die Nutzung von Zeichensystemen, die eine definierte syntaktische Struktur besitzen. Auf der nächsthöheren Ebene, der Sigmatik, tritt die formale Zuordnung der bezeichneten Objekte hinzu. So kann „Produkt Nr. x" den Artikel „Hemd" bezeichnen, während „Produkt Nr. y" für den Artikel „Krawatte" steht. Auf dieser Ebene wird von der inhaltlichen Bedeutung der verwendeten Begriffe abstrahiert, d. h. es erfolgt eine rein formale Zusammenfassung der Zeichenfolgen zu Daten. Auf dieser Ebene erfolgt also die Einordnung des Datenbegriffs Daten.

Die inhaltliche Bedeutungsbildung erfolgt erst auf der semantischen Ebene. Auf dieser Ebene manifestiert sich die inhaltliche Aussage des Musters, daß beim Kauf eines Hemdes mit einer bedingten Wahrscheinlichkeit von 66 % auch eine Krawatte erworben wird. Die Integration des Anwenders in den Prozeß der Zeichenbildung führt zur pragmatischen Ebene, bei der die Wirkung von Daten im Mittelpunkt des Interesses steht. So kann die kognitive Verarbeitung

---

1 Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 3.

2 Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 39.

3 Vgl. Grob, H. L., Bieletzke, S. (1997), S. 6; Wessling, E. (1991), S. 13-18.

4 Wessling, E. (1991), S. 18.

von Daten zu einer effektiven Wissensveränderung führen, wodurch die Daten die Qualität von Information[1] erlangen. Die Wissensveränderung impliziert dabei auch eine potentielle Reaktion des Anwenders. Nimmt z. B. der Anwender eines Data-Mining-Systems das oben angeführte Beispielmuster wahr und handelt es sich um einen Sachverhalt, der vorher unbekannt war, kann es aufgrund der erfolgten Wissensvermehrung zu Handlungen im betrieblichen Planungs- und Entscheidungskontext kommen. Beispielsweise können Maßnahmen der Verkaufsraumgestaltung ergriffen werden, die zu einem räumlichen Zusammenhang zwischen dem Angebot von Hemden und Krawatten führen.

Da der Data-Mining-Prozeß Muster generiert, die Daten im Sinne von objektivem Wissen darstellen, liegt die Einordnung des Data-Mining-Konzepts auf der sigmatischen Ebene der Semiotik nahe. Das KDD-Konzept intendiert dagegen die Vermittlung von Information im Sinne eines subjektbezogenen Wissenszuwachses. Aus diesem Grund ist dieses Konzept auf der pragmatischen Ebene anzusiedeln. Dem Data-Mining-Konzept fällt damit die Aufgabe eines formalen Mustergenerators zu, während das KDD-Konzept darüber hinaus Sorge zu tragen hat, daß die Muster durch den Anwender interpretiert werden und ihren Nutzen im betrieblichen Planungs- und Kontrollsystem entfalten können. Dieser Sachverhalt wird durch die folgende Abbildung zusammenfassend dargestellt.

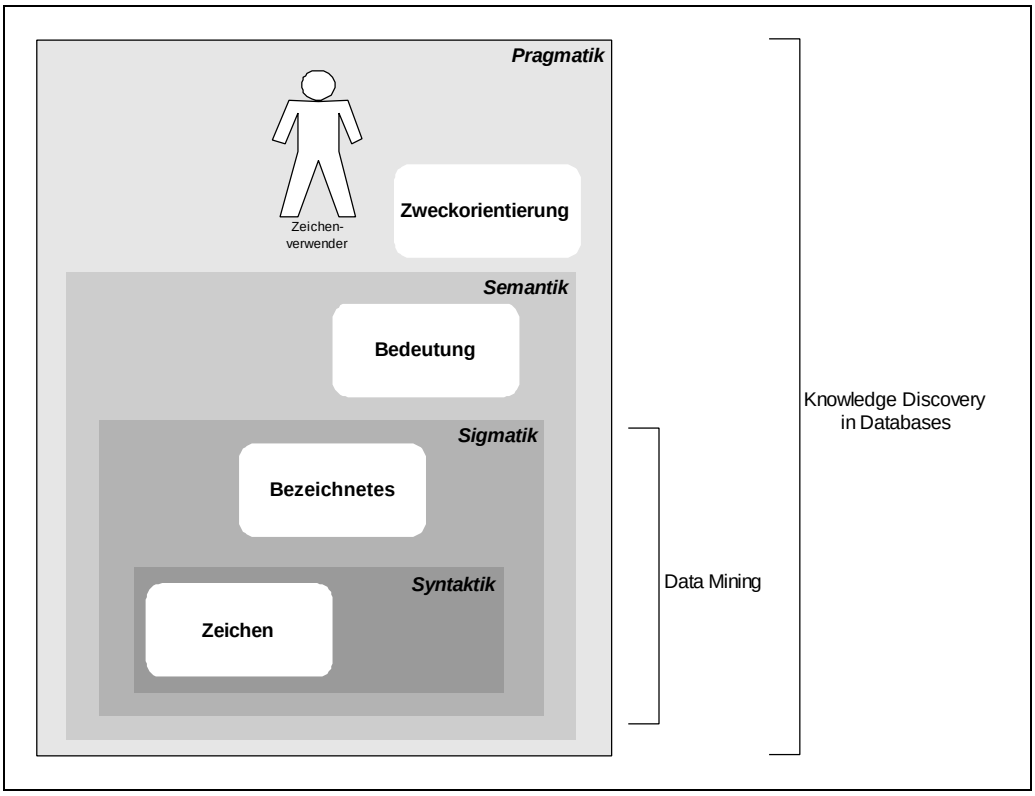

Abb. 3: Die Abgrenzung von Data Mining und Knowledge Discovery in Databases

[1] So wird Information von Wessling als prozeßhaft, wissensverändernd, subjektbezogen und nicht objektivierbar definiert. Vgl. Wessling, E. (1991), S. 19.

Nach der Darstellung des Data-Mining-Konzepts und seiner Abgrenzung vom KDD-Konzept wird ein Prozeßmodell vorgestellt, das die Teilaktivitäten des Data-Mining-Prozesses zeitlich strukturiert.

2 Das Prozeßmodell des Data Mining

Für die Strukturierung des Data-Mining-Prozesses gibt es in der Literatur unterschiedliche Vorschläge.[1] Der Data-Mining-Prozeß soll hier als iterativer Prozeß erfaßt werden, der sich aus fünf Phasen zusammensetzt.

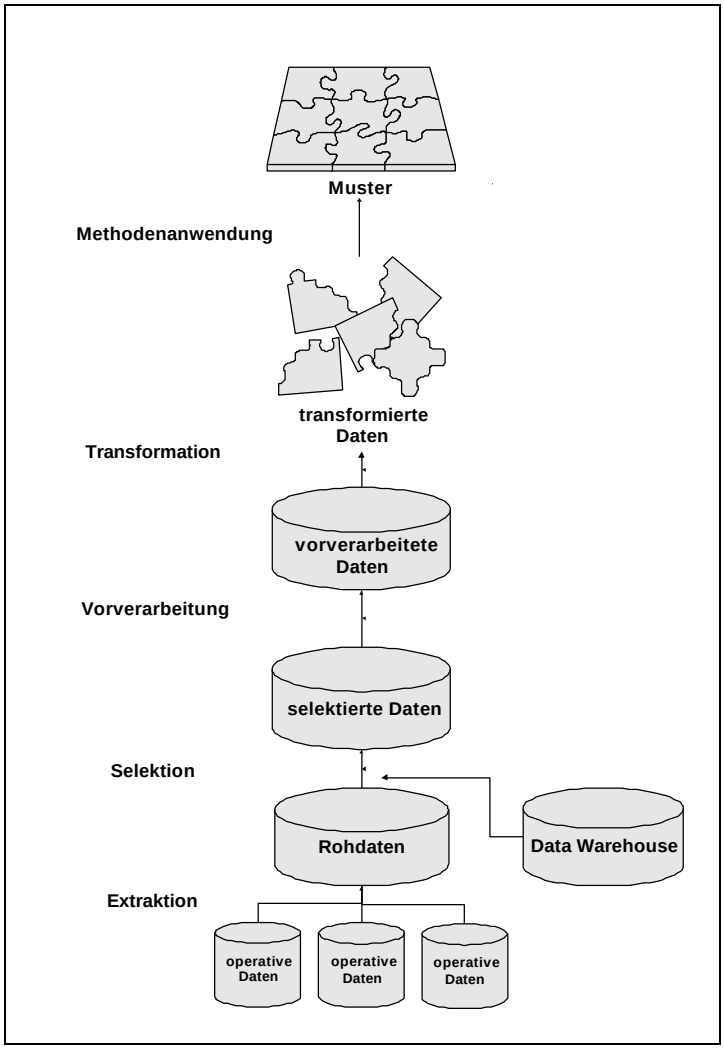

Abb. 4: Der Data-Mining-Prozeß[2]

1 Vgl. Flanagan, T., Safdie, E. (1997), S. 15; Heiting, M. (1996), S. 12.

2 Vgl. Heiting, M. (1996), S. 12.

In der ersten Phase des Data-Mining-Prozesses findet die *Extraktion* der relevanten Daten aus einer geeigneten Datenquelle statt. Data-Mining-Systeme können sinnvollerweise mit vorhandenen Data Warehouse-Systemen gekoppelt werden, sind aber nicht auf den ausschließlichen Zugriff auf ein Data Warehouse beschränkt. So können die Daten auch direkt aus den operativen Datenbanken oder aus externen Datenquellen beschafft werden. Diese Phase ist für die Mustererkennung von ausschlaggebender Bedeutung, da die Datengrundlage für den gesamten Data-Mining-Prozeß festgelegt wird.

Die anschließende Phase der *Selektion* hat die Aufgabe, aus dem extrahierten Datenbestand eine Menge von Datensätzen (vertikale Selektion) und Attributen (horizontale Selektion) auszuwählen. In Abhängigkeit vom Anwendungskontext kann es im Rahmen der vertikalen Selektion hinreichend sein, lediglich eine begrenzte, repräsentative Menge von Datensätzen auszuwählen, auf welche die Methoden angewendet werden. Wird die Grundgesamtheit der Daten zu einer Stichprobe reduziert, stellt sich potentiell die Gefahr, daß eine subjektive Auswahl durch den Anwender zu einer Manipulation der Datengrundlage führt. Die analyserelevanten Attribute werden in Abhängigkeit von der betriebswirtschaftlichen Problemstellung im Rahmen der horizontalen Selektion bestimmt.

In der Phase der *Vorverarbeitung (Preprocessing)* wird die Datenqualität des selektierten Datenpools untersucht. Aufgrund technischer oder menschlicher Fehler können die Daten operativer Systeme fehlerhafte Elemente enthalten. In der Praxis wird damit gerechnet, daß 1 %-5 % der Felder des Datenbestands falsche Angaben aufweisen.[1] Eine häufige, leicht zu identifizierende Fehlerart besteht in *fehlenden* Werten (missing values). Zur Behandlung von fehlenden Werten stehen unterschiedliche Techniken zur Verfügung.[2] So kann der Datensatz, der fehlende Werte aufweist, aus dem Datenbestand gelöscht werden. Dies kann jedoch zu einer Verfälschung der Analyseergebnisse führen. Eine weitere Möglichkeit besteht in der Schätzung oder nachträglichen Erhebung des fehlenden Datums. Während die Schätzung bei quantitativen Daten über die Bildung des Mittelwerts technisch einfach realisierbar ist, kann die nachträgliche manuelle Erhebung zu einem hohen Aufwand führen. Eine weitere potentielle Fehlerart wird durch Ausreißer hervorgerufen.[3] Dabei handelt es sich um Wertausprägungen, die deutlich vom Niveau der übrigen Werte abweichen. Bei diesen Ausprägungen kann es sich um korrekt erfaßte Daten handeln, die damit Eingang in die Analyse finden, oder aber um falsche Angaben, die nicht berücksichtigt werden dürfen und daher aus dem Datenbestand zu löschen sind. Die Erkenntnisse, die der Nutzer eines Data-Mining-Systems in dieser Phase der Vorverarbeitung über den Datenbestand gewinnt, kann Hinweise auf die Verbesserung der Datenqualität der operativen Systemen geben.

---

1 Vgl. Redman, T. C. (1998), S. 80.

2 Vgl. Berry, J. A., Linoff, G. (1997), S. 70 f.

3 Vgl. Berry, J. A., Linoff, G. (1997), S. 77 f.

Die vierte Phase besteht in der *Transformation* der Daten. Diese Phase wird benötigt, um die analyserelevanten Daten in ein Datenbankschema zu transformieren, das von dem verwendeten Data-Mining-System verarbeitet werden kann. Verfügt das Data-Mining-System über eine standardisierte Datenbankschnittstelle wie beispielsweise Open Database Connectivity (ODBC[1]), so stellt die Transformationsphase eine optionale Teilaktivität dar. Da die ODBC-Schnittstelle häufig eine geringere Zugriffsgeschwindigkeit als der direkte Zugriff aufweist, bietet die Schematransformation eine Möglichkeit zur Verkürzung der Analysezeit.

In der fünften Phase der *Methodenanwendung* erfolgt die Methodenauswahl und deren Einsatz zur Identifikation von Mustern auf der Basis des vorbereiteten Datenbestands. Zwar dienen die der Mustererkennung vorgelagerten Phasen „nur" der Datenvorbereitung, doch verbrauchen diese Aktivitäten einen erheblichen Teil der Gesamtressourcen des Data-Mining-Prozesses. So werden nach Expertenschätzungen ca. 80 % der Zeit und Kosten des Data Mining für die Vorarbeiten aufgewendet.[2] Die Methodenanwendung wird von der Art der gesuchten Muster bestimmt und weist daher eine starke situativ bedingte Problemorientierung auf. Im folgenden findet deshalb eine Beschränkung auf die zentralen Methodenklassen des Data Mining statt.[3]

Die Methoden zur *Klassifikation* ermitteln Muster, die es erlauben, noch nicht bekannte Aussagen über Objekte zu treffen. In einem ersten Schritt werden Objekte in Gruppen zusammengefaßt, die sich durch charakteristische Attribute und gleiches Verhalten bezüglich des zu untersuchenden Problems auszeichnen. Diese Menge von Objekten stellt die Trainingsdatenmenge dar, auf deren Basis in einem zweiten Schritt ein Klassifikationsmodell (Klassifikator, Klassifikationsfunktion) entwickelt wird. Dieses Modell erlaubt dem Anwender die Prädiktion der Klassenzugehörigkeit von neuen Objekten. In der folgenden Abbildung wird eine lineare Klassifikationsfunktion dargestellt, die zur Trennung kreditwürdiger und nichtkreditwürdiger Kunden dient.[4]

---

1 ODBC stellt eine Datenbank-Middleware dar, die auf vorhandenen Datenbanksystemen aufsetzt und so den Datenaustausch zwischen heterogenen Systemen ermöglicht. Zum Zugriff stellt ODBC eine standardisierte API-Schnittstelle bereit, die Anfragen von Client-Systemen in die Sprache des Serversystems übersetzt.

2 Vgl. Alpar, P. u. a. (1998), S. 38.

3 Einen Überblick über die Methoden des Data Mining liefern Chen, M.-S., Han, J., Yu, P. S. (1997), S. 4-33 und Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996a), S. 13-16.

4 Zur Generierung einer linearen Klassifikationsfunktion kann beispielsweise die Diskriminanzanalyse herangezogen werden. Vgl. Backhaus, K. u.a. (1996), S. 96 ff. Aufgrund der zu erfüllenden Prämissen wird dieses Verfahren in der Praxis allerdings wenig angewandt. Vgl. Krahl, D., Windheuser, U., Zick, F.-K. (1998), S. 76.

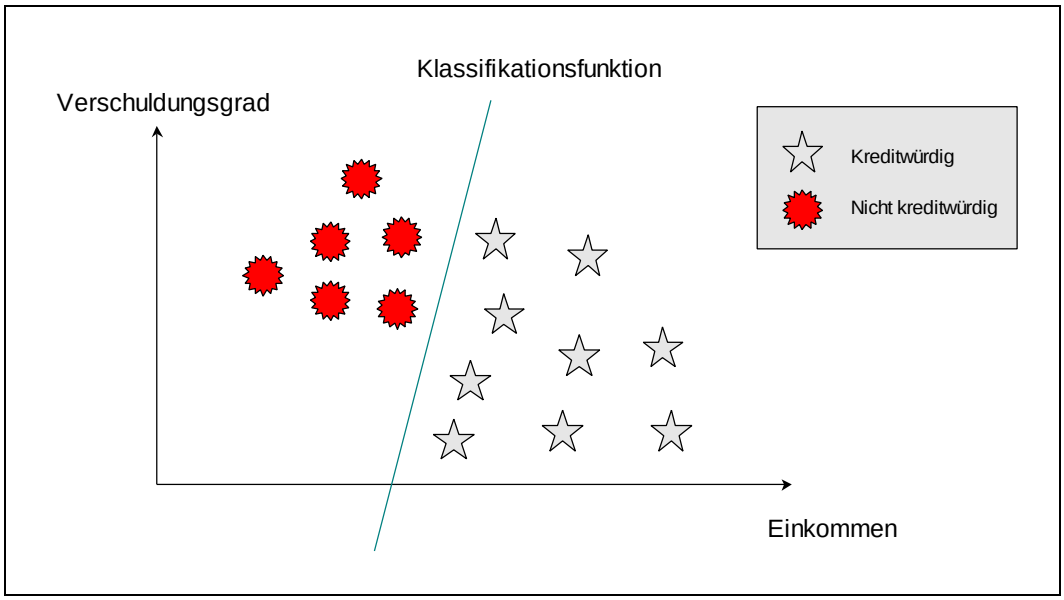

Abb. 5: Klassifikation von Objekten (Beispiel)

Zur Klassifikation von Objekten werden im Rahmen von Data-Mining-Systemen häufig entscheidungsbaumorientierte Methoden eingesetzt. Zwar können auch Methoden der Künstlichen Intelligenz zur Klassifikation eingesetzt werden, doch verfügen Entscheidungsbäume über ein höheres Maß an Transparenz, da die Klassifikationsregeln direkt aus dem Entscheidungsbaum ablesbar sind. [1] Die folgende Abbildung stellt einen binären Entscheidungsbaum dar.

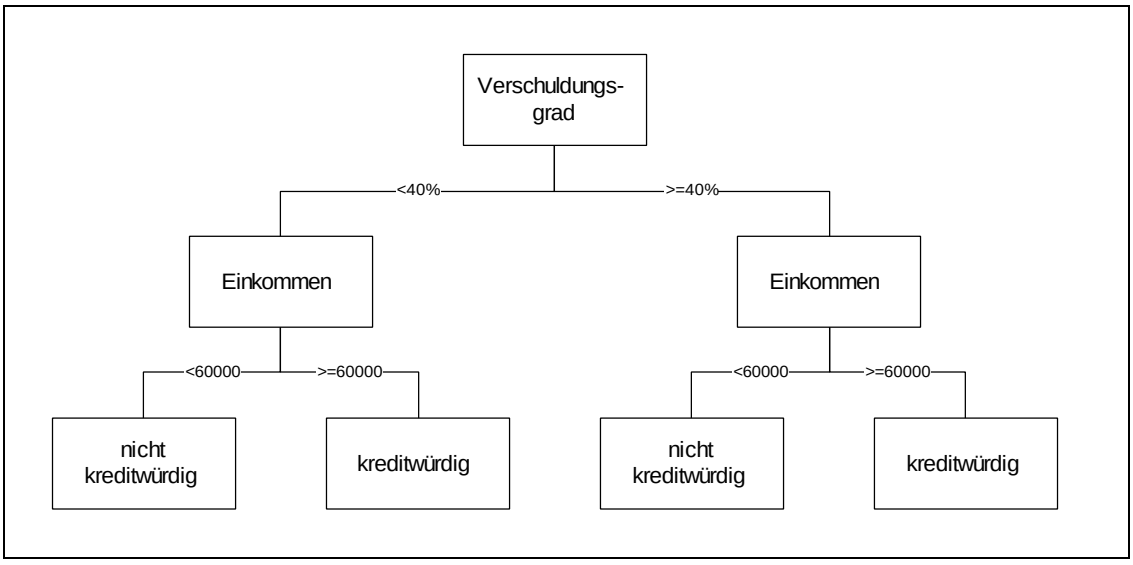

Abb. 6: Entscheidungsbaum (Beispiel)

[1] Vgl. Berry, J. A., Linoff, G. (1997), S. 243.

Die Generierung von Entscheidungsbäumen erfolgt häufig nach dem Top-Down-Prinzip.[1] Dabei werden sukkzessiv diejenigen Attribute zur Baumkonstruktion verwendet, welche die Trainingsdatensätze am besten klassifizieren. Auf der untersten Ebene des Entscheidungsbaums sind schließlich die Klassenzugehörigkeiten abzulesen. Aus einem Entscheidungsbaum lassen sich direkt die generierten Regeln ablesen. Für das dargestellte Beispiel gilt:

Wenn(Verschuldungsgrad<40 %) und (Einkommen<60000) dann (nicht kreditwürdig)

Problematisch wird die Anwendung von Entscheidungsbäumen, wenn der generierte Baum aus vielen Ebenen besteht und durch die Vielzahl von Verzweigungen unübersichtlich wird. In diesem Fall ist die Tiefe des Baums durch entsprechende Pruning-Verfahren zu reduzieren.[2] Da zur Generierung von Entscheidungsbäumen effiziente Algorithmen zur Verfügung stehen, wird dieses Verfahren von vielen Data-Mining-Systemen unterstützt.[3]

Sind die Klassenzugehörigkeiten von Objekten a priori nicht gegeben, sondern steht die Partitionierung der Objektmenge im Mittelpunkt des Interesses, so findet das Verfahren der Clusteranalyse Anwendung.[4] Die Clusteranalyse berücksichtigt alle Merkmale der Objektmenge und generiert auf der Basis von Distanzmaßen (möglichst) homogene Gruppen (Cluster). Die folgende Abbildung zeigt das mögliche Ergebnis einer Clusteranalyse, die auf das angeführte Beispiel aus dem Bankenbereich angewendet wurde. Die identifizierten Muster repräsentieren im dargestellten Beispiel Kundensegmente und können eine informatorische Grundlage für die clusterspezifische Marktbearbeitung darstellen.

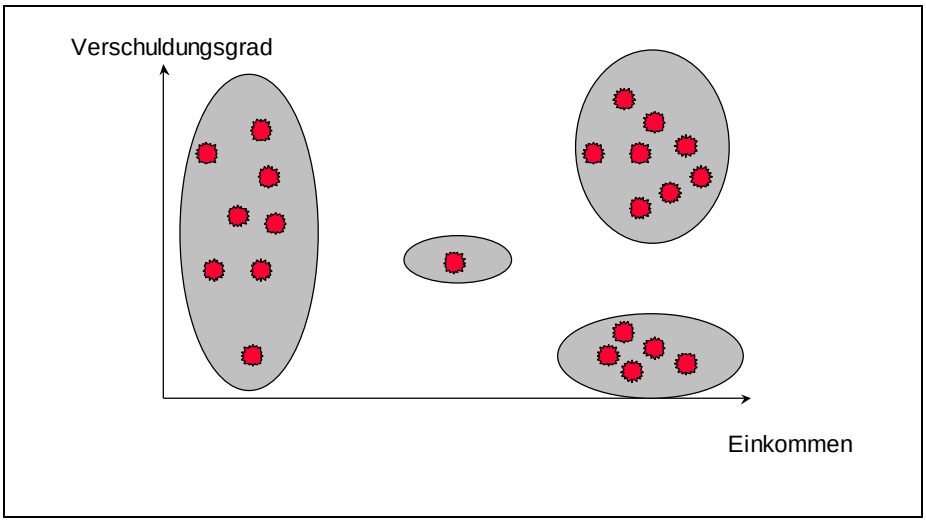

Abb. 7: Clustern von Objekten (Beispiel)

1 Vgl. Bissantz, N. (1996), S. 62.

2 Vgl. Berry, J. A., Linoff, G. (1997), S. 256.

3 Anzuführen sind die die Algorithmen ID-3, C4.5, CHAID und CART. Vgl. Chen, M.-S., Han, J., Yu, P. S. (1997), S. 20 f.; Berry, J. A., Linoff, G. (1997), S. 244.

4 Vgl. Chen, M.-S., Han, J., Yu, P. S. (1997), S. 22-30.

Die *Abweichungsanalyse* (*deviation detection)* beschäftigt sich mit Objekten, die sich keinem Muster eindeutig zuordnen lassen. Bei diesen „Ausreißern“ (outlier) kann es sich um fehlerfreie, interessante Merkmalsausprägungen handeln oder aber um fehlerhafte Daten, die keine realen Sachverhalte beschreiben. Die Zielsetzung der Abweichungsanalyse besteht darin, die Ursachen für die untypischen Merkmalsausprägungen des Ausreißers aufzudecken.[1] Wird ein Ausreißer im Datenbestand identifiziert, so durchsucht das Data-Mining-Tool alle assoziierten Datenbestände, um die Einflußfaktoren zu erklären, die zu einer abweichenden Merkmalsausprägung geführt haben.[2] Handelt es sich bei einem Ausreißer um einen fehlerhaften Wert, wird dieser aus dem Datenbestand eliminiert. Da auf diese Weise die Datenqualität gesteigert wird, werden Methoden zur Abweichungsanalyse bereits in der Phase der Vorverarbeitung eingesetzt.

Bei der *Assoziationsanalyse (association rule mining)* suchen Data-Mining-Werkzeuge nach signifikanten Abhängigkeiten zwischen einzelnen Feldern der Analyseobjekte.[3] Die identifizierten Muster werden sprachlich in Form von Wenn-dann-Regeln dargestellt oder grafisch präsentiert. Ein verbreiteter Anwendungsbereich für die Assoziationsanalyse ist die *Warenkorbanalyse (market basket analysis).*[4] Dabei wird eine Menge von Kaufakten analysiert, um Aussagen über das Käuferverhalten zu entwickeln und sachliche Verbundeffekte aufzudekken. Eine klassisches Ergebnis einer Warenkorbanalyse ist beispielsweise das Muster: *Wenn Kunden Brötchen und Margarine kaufen, kaufen 70 % der Kunden auch Marmelade.* Die einfache Warenkorbanalyse erfaßt jedoch nur gleichzeitig vorhandene Abhängigkeiten innerhalb des Warenkorbes.[5] Um *zeitliche Verbundeffekte* aufzudecken, wird die Warenkorbanalyse um die Dimension Zeit erweitert. Bei dieser Untersuchung handelt es sich um eine *Reihenfolgeanalyse* (*Sequenzanalyse*). Das Ziel der Reihenfolgeanalyse besteht darin, einzelne Phasen und die zeitlichen Distanzen zwischen wiederkehrenden Prozessen zu entdecken.[6] Voraussetzung hierfür ist, daß die Daten einzelner Kunden über einen größeren Zeitraum gesammelt werden. Ein Ergebnis einer Sequenzanalyse ist beispielsweise: *Wenn Kunden im Winter einen Fernseher kaufen, kaufen 50 % von ihnen nach spätestens 10 Wochen auch einen Videorekorder.*

---

1 Vgl. Alpar, P. u. a. (1998), S. 38.

2 Vgl. Matheus, C. J.; Piatetsky-Shapiro, G., McNeill, D. (1996), S. 495 f.; Guyon, I., Matic, N., Vapnik, V. (1996), S. 181.

3 Vgl. Agrawal, R., Imielinski, T., Swami, A. (1993), S. 207.

4 Vgl. Berry, J. A., Linoff, G. (1997), S. 124 ff.

5 Im Rahmen der Warenkorbanalyse werden intratransaktionale Muster identifiziert. Die Assoziationsanalyse kann für die Entdeckung intertransaktionaler Muster erweitert werden und dadurch auch für andere Anwendungsfelder, wie beispielsweise der Analyse von Börsenkursbewegungen nutzbar gemacht werden. Vgl. Lu, H., Han J., Feng, L. (1998), S. 1.

6 Vgl. Alpar, P. u. a. (1998), S. 38 f.; Agrawal, R., Srikant, R. (1994), S. 1; Berry, J. A., Linoff, G. (1997), S. 149 ff.

Um eine Informationsüberlastung des Anwenders zu vermeiden, sind nutzlose Muster auszufiltern. Zu der Kategorie der nutzlosen Muster sind auch Sachverhalte zu zählen, die dem Anwender bereits bekannt sind.[1] Dabei kann es sich um Muster handeln, die einem vorhergehenden Analysedurchlauf identifiziert und präsentiert wurden oder aber um Trivialaussagen wie „Alle Liefermengen sind positiv".[2] Werden Muster auf mehreren Hierarchieebenen ermittelt, so können redundante Muster auftreten. Im Rahmen der Warenkorbanalyse kann die Regel „Wenn ein Kunde Brot kauft, dann kauft er auch Butter" bis auf die Artikelebene heruntergebrochen werden, so daß die detaillierte Regel lautet: „Wenn ein Kunde das Brot x kauft, dann kauft er auch die Butter y". Besitzen beide Muster die gleiche statistische Güte, kann *ein* Muster aus der Ergebnismenge gestrichen werden. Kritisch sind auch Muster zu betrachten, die eine geringe statistische Signifikanz aufweisen und sich nur auf wenige Objekte des Datenbestands beziehen.[3] Diese Muster können für den Entscheider irrelevant sein, doch kann das Eliminieren dieser Muster auch dazu führen, daß nützliche Informationen vernichtet werden, beispielsweise wenn diese Muster selten auftretende Betrugsfälle beschreiben.

Die Darstellung des Prozeßmodells zeigt, daß Data-Mining-Systeme in Abhängigkeit vom Methodenvorrat unterschiedliche Arten von Mustern generieren und präsentieren können. An dieser Stelle endet nach der hier vorgetragenen Auffassung die Aufgabe des Data Mining. Weiterführenden Funktionen, wie die Vermittlung des identifizierten objektiven Wissens an den Anwender des Systems, sind von einem übergreifenden KDD-System zu übernehmen. So hat der Anwender das Informationsangebot des KDD-Systems ökonomisch zu bewerten, in eigene ökonomische Modelle einzubauen und den Nutzen für betriebliche Entscheidungsprozesse zu ermitteln.[4]

3 Anwendungsgebiete des Data Mining

Die Forschungsdisziplin Data Mining ist nicht etwa das Ergebnis der Spezialisierung bestehender Wissenschaftsbereiche, sondern zeichnet sich durch die Synthese bereits etablierter Disziplinen mit der Informatik und insbesondere - sofern es um Anwendungsorientierung geht - der Wirtschaftsinformatik aus. So werden im Rahmen des Data Mining die Erkenntnisse der Datenbankforschung, der Statistik und der Künstlichen Intelligenz herangezogen und kombi-

---

1 Diese Muster stellen keine Information dar, da sie nicht dem Anforderungskriterium der Neuartigkeit genügen. Vgl. Grob, H. L., Bieletzke, S. (1997), S. 11.

2 Vgl. Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 603.

3 Vgl. Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 603.

4 Da die Bestimmung des Informationswertes aufgrund der Eigenschaften des Produktionsfaktors Information notwendigerweise ex post erfolgt, wird hier von den Kosten der Informationsbeschaffung, die durch die Realisierung und den Betrieb eines Data-Mining-Systems anfallen, abstrahiert. Zur Bewertung von Information vgl. Picot, A., Reichwald, R., Wigand, R. T. (1998), S. 109 f.

niert, um effiziente Systeme zur Informationsgewinnung zu gestalten.[1] Der Erfolg derartiger Systeme läßt sich in einer Vielzahl wissenschaftlicher Disziplinen nachweisen. So werden Data-Mining-Systeme in der Astronomie zur Klassifikation von Himmelskörpern[2] eingesetzt oder sie dienen im Rahmen molekularbiologischer Forschungen der Identifikation von Genen in DNA-Sequenzen.[3] Im betriebswirtschaftlichen Bereich haben Data-Mining-Systeme ein breites Anwendungsfeld im Einzelhandel gefunden. So führt die US-amerikanische Supermarktkette Wal-Mart täglich alle POS-Transaktionen (ca. 20 Mio.) in einer zentralen Datenbank zusammen und verfügt somit über einen tagesaktuellen Datenpool.[4] Durch die Anwendung von Data Mining können in den Daten Zusammenhänge entdeckt werden. Die so gewonnenen Hypothesen können in die weitere Planung (z. B. im Marketingcontrolling) integriert werden. So können mit Warenkorbanalysen Aussagen der Art „Wenn ein Kunde Produkt A kauft, dann kauft er mit einer Wahrscheinlichkeit von x % auch Produkt B“ generiert werden. Diese Aussage liefert einen Beitrag für die Verkaufsraumgestaltung oder die Bestellmengenplanung. Der Nutzen von Data-Mining-Systemen im Handel wird dabei hoch eingeschätzt.[5] Aber auch in anderen betriebswirtschaftlichen Anwendungsbereichen ist der erfolgreiche Einsatz von Data-Mining-Systemen zu beobachten, wie beispielsweise bei der Marktsegmentierung, der Antwortvorhersage, der Kündigungsanalyse und der Kreditwürdigkeitsprüfung.[6]

Da in jedem Funktionsbereich der Unternehmung Daten anfallen, die möglicherweise interessante Muster enthalten, stellt Data Mining *grundsätzlich* eine Querschnittstechnologie dar. Somit ergibt sich die Frage, welche Unternehmensbereiche als betriebswirtschaftliche „hot spots“ anzusehen sind, in denen der Einsatz von Data Mining den höchsten Erfolgsbeitrag liefert. In praxi dominieren derzeit Anwendungen im *Controlling* und insbesondere im *Marketingcontrolling*.[7] Aus sektoraler Perspektive ist der Einsatz in den dienstleistungsorientier-

---

1 Vgl. Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 604.

2 In diesem Anwendungsbereich werden Datenbestände, die beispielsweise von terrestrischen Radioteleskopen oder Satelliten generiert werden, automatisch nach Himmelskörpern durchsucht. Diese Datenbestände umfassen mehrere Terabyte und sind mit klassischen Werkzeugen der wissenschaftlichen Analyse nicht mehr zugänglich. Vgl. Fayyad, U., Djorgovski, S. G., Weir, N. (1996), S. 471 ff.

3 Vgl. Fayyad, U., Haussler, D., Stolorz, P. (1996), S. 53 f.

4 Vgl. Reese Hedberg, S. (1995), S. 83.

5 So werden für Referenzinstallationen im Einzelhandel ROI-Werte von 1000 %-7000 % angeführt. Vgl. Krivda, C. D. (1995), S. 97-99. Diese Angaben sind jedoch kritisch zu betrachten, da derzeit keine empirisch gesicherten Wirtschaftlichkeitsstudien für Data-Mining-Systeme verfügbar sind.

6 Vgl. Berry, J. A., Linoff, G. (1997), S. 10-16.

7 Eine Übersicht über Data-Mining-Anwendungen liefern Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 604 f.

ten und informationsintensiven Branchen *Banken, Versicherungen, Telekommunikation* und *Handel* zu beobachten.[1]

Im Controlling ist der Einsatz von Data-Mining-Verfahren dann sinnvoll, wenn betriebswirtschaftlich relevante Muster entdeckt werden können, die nicht von traditionellen Controllinginstrumenten identifiziert werden können. Ein derartiges Anwendungsfeld ist üblicherweise bei Controllingsystemen zu sehen, die durch eine starke Berichts- und Kennzahlenorientierung gekennzeichnet sind. Die Erstellung von Berichten aus hochaggregierten Daten enthält die Gefahr, daß relevante Zusammenhänge in den Elementardaten in verdichteten Kennzahlen untergehen. Der Einsatz von Data-Mining-Verfahren bietet das Potential, die „blinden Flekken" eines kennzahlenorientierten Berichtswesens durch die Entdeckung betriebswirtschaftlich relevanter Muster zu reduzieren und dadurch die Entscheidungsqualität zu steigern.

Zwei Aufgabenstellungen lassen sich beim Einsatz von Data-Mining-Verfahren im Controlling identifizieren.[2] Zum einen können im Rahmen eines Top-Down-Ansatzes für wichtige Zielobjekte (z. B. eine Deckungsbeitragsabweichung) diejenigen Controlling-Objekte ermittelt werden, die das Zielobjekt am besten erklären. In diesem Fall leistet das Data-Mining-System eine Navigationshilfe beim Drill-Down in bestehenden Bezugsgrößenhierarchien. Zum anderen können Data-Mining-Verfahren im Rahmen eines Bottom-Up-Ansatzes versuchen, betriebswirtschaftlich relevante Muster auf der Basis der elementaren unverdichteten Datenbasis zu identifizieren. Relevante Controlling-Muster sind beispielsweise:[3]

- Abweichungen (z. B. zwischen Plan- und Istgrößen),
- Konzentration eines großen Anteils einer Mengen- oder Wertgröße auf relativ wenige Bezugsobjekte (z. B. bei ABC-Analysen),
- Abhängigkeiten (z. B. zwischen Kundengruppen und Umsätzen),
- Ähnlichkeiten (z. B. bei der Klassifikation von Geschäftsbereichen im Rahmen der Portfolioanalyse),
- Rangfolgen (z. B. Produkte nach Rentabilitätskennzahlen) und
- zeitliche Entwicklungen (z. B. Zeitreihen zum Aufdecken von Diskontinuitäten).

Voraussetzung für den Bottom-Up-Ansatz sind atomare Controllingdaten, die aus den operativen Systemen des Rechnungswesens zu extrahieren sind. In diesen meist hochdimensionalen Ergebnisdaten können beispielsweise Konzentrationen mit Hilfe clusternder Verfahren aufge-

---

1 Vgl. Küppers, B. (1998), S. 142 ff.; Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J. (1991), S. 17 f.

2 Vgl. Küppers, B. (1999), S. 135 f.

3 Vgl. hierzu Bissantz, N. (1996), S. 45 ff.

deckt werden.[1] Die folgende Abbildung stellt für den Objektbereich des Ergebniscontrolling zwei typische Konzentrationsmuster dar.

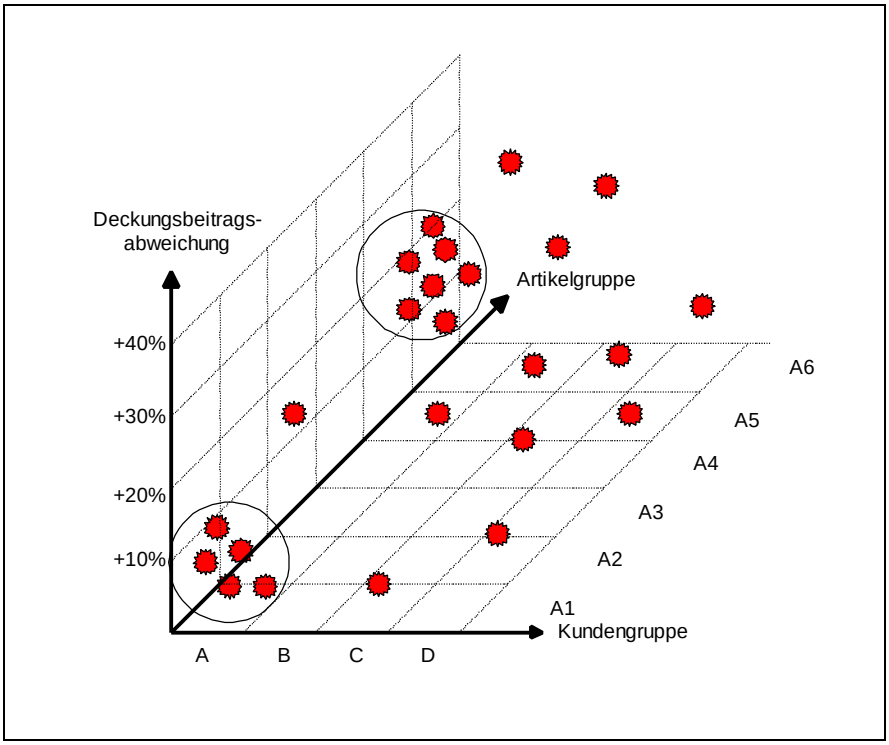

Abb. 8: Clusteranalyse im Ergebniscontrolling

Die Controlling-Objekte stellen in diesem Beispiel Geschäftsvorfälle dar, die allesamt zu einer positiven Deckungsbeitragsabweichung geführt haben. Die Interpretation der beiden identifzierten Cluster führt zu den Aussagen „An die Kundengruppe A wurde die Artikelgruppe A1 mit einem 0-10 % höherem Deckungsbeitrag verkauft als geplant“ bzw. „An die Kundengruppe B wurde die Artikelgruppe A4 mit einem 20-30 % höheren Deckungsbeitrag verkauft als geplant“. Obwohl die Clusteranalyse dem Controller interessante Konstellationen in den relevanten Dimensionen liefert, sind die Muster durch eine Ursachenanalyse zu erklären. Zu diesem Zweck sollte die Deckungsbeitragsabweichung für die identifizierten Marktsegmente einer Abweichungsanalyse unterzogen werden. Das Beispiel zeigt, daß der Einsatz mustergenerierender Bottom-Up-Verfahren zu weiteren Fragestellungen führen kann, die durch den Einsatz von Top-Down-Ansätzen beantwortet werden können. Als Instrumente finden beispielsweise Berichtssysteme Anwendung, die komplexe Kennzahlensysteme darstellen können und dem Anwender den Durchgriff auf die Elementardaten (Drill-Down) erlauben.

Im Marketingcontrolling finden Data-Mining-Ansätze zunehmend hohe Akzeptanz. Dies ist nicht zuletzt darauf zurückzuführen, daß viele Unternehmen zur Abwicklung von Geschäfts-

[1] Eine umfassende Darstellung dieses Ansatzes ist zu finden bei Bissantz, N. (1996), S. 50 f., S. 91 ff.

prozessen mittlerweile über Standardsoftwaresysteme (z. B. SAP R/3, Baan IV) verfügen, in denen detaillierte *kundenbezogene* Stamm- und Bewegungsdaten gespeichert sind. Zur Analyse dieser historischen Daten wurden bislang Führungsinformationssysteme (EIS) angewandt, die allerdings nur über ein beschränktes Methodenarsenal verfügten und nicht in der Lage waren, komplexe Zusammenhänge in den gegebenen Datenbeständen zu entdecken. Auch neuere Entwicklungen wie das Online Analytical Processing (OLAP), das die interaktive und multidimensionale Datenanalyse ermöglicht, konnten diesen Nachteil letztlich nicht ausräumen.[1] Systeme der statistischen Datenanalyse (z. B. SPSS, SAS) decken zwar hohe methodische Anforderungen ab,[2] doch waren diese Systeme nicht auf den Einsatz mit großvolumigen, operativen Daten vorbereitet. Hinzu kommt, daß derartige Systeme relativ hohe Anforderungen an den Anwender stellen. Gleichzeitig stehen vor allem Unternehmen, die informationsintensive Produkte vermarkten, einem erhöhten Wettbewerbsdruck gegenüber. Vor diesem Hintergrund wird Data Mining von marktorientierten Unternehmungen sogar als *strategisches* Instrument zur Realisierung von Konkurrenzvorteilen gesehen.

Das Data-Mining-Konzept kann im Marketingcontrolling zwei grundlegende Aufgaben erfüllen. Zum einen stellt es ein Informationsinstrument dar, das im Rahmen der Marketingforschung zur Analyse und zur Prognose von Marktreaktionen eingesetzt werden kann. In diesem Sinne stellt Data Mining eine Erweiterung des „klassischen" Analyseinstrumentariums des Marketingforschers dar. Zum anderen kann Data Mining auch zur *operativen Steuerung* von Marketinginstrumenten eingesetzt werden. Als konzeptioneller Rahmen dient das *Database Marketing*, das durch die konsequente Nutzung marktbezogener Daten die Effektivitätssteigerung der Marketinginstrumente intendiert. Im Mittelpunkt steht dabei nicht die Bearbeitung eines „anonymen" Kundensegments, sondern die Realisierung eines möglichst kundenindividuellen Marketing. Die Konzeption des Database Marketing wird anhand des RADAR-Modells (Reaction, Analysis, Detection, Action, Reaction) verdeutlicht.[3]

---

1 Zum OLAP-Konzept vgl. Alpar, P. u. a. (1998), S. 171 ff.

2 So verfügen Systeme zur statistischen Datenanalyse über explorative und konfirmatorische Methoden, die auch Bestandteil von Data-Mining-Werkzeugen sind.

3 Vgl. Link, J.; Hildebrand, V. (1993), S. 30 ff.

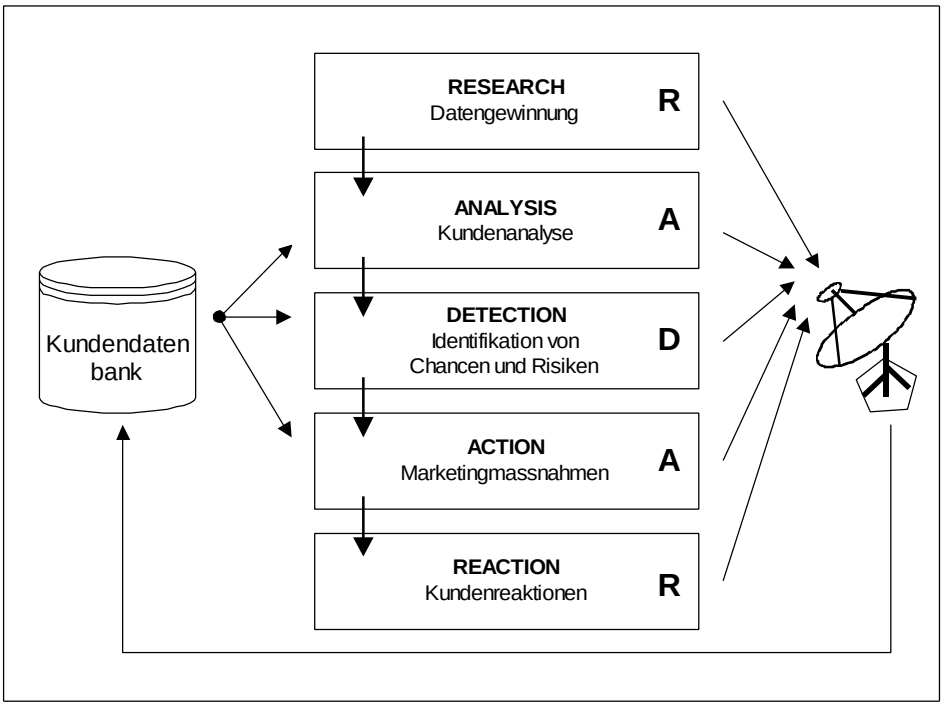

Abb. 9: Das RADAR-Modell des Database Marketing[1]

Die zentrale Infrastruktur des Database Marketing ist eine Kundendatenbank, die in der Phase der Datengewinnung (research) aufgebaut wird. Die Daten der Kundendatenbank werden in Grund-, Potential-, Aktions- und Reaktionsdaten differenziert. Während unter **Grunddaten** zeitstabile, kundenspezifische Daten (z. B. Adreßdaten) zu verstehen sind, geben **Potentialdaten** Aufschluß über den kundenindividuellen Bedarf und den Bedarfszeitpunkt (z. B. Art und Menge bereits erworbener Produkte). Die kundenbezogenen Marketingmaßnahmen werden durch **Aktionsdaten** (z. B. Kundenkontakte) dokumentiert. Die Effektivität der eingesetzten Marketingmaßnahmen (z. B. erteilte Kundenaufträge) wird schließlich durch **Reaktionsdaten** quantifiziert.

Auf dieser Datengrundlage können in der Analysephase (analysis) Data-Mining-Werkzeuge eingesetzt werden. Die Analyseergebnisse decken potentiell Chancen und Risiken für die Ausgestaltung der Marketinginstrumente in bezug auf einzelne Kunden, Marktsegmente oder den gesamten Markt auf (detection). Durch die Transformation dieser Erkenntnisse in marketingpolitische Maßnahmen (action) werden Kundenreaktionen (reaction) induziert. Wesentlich für den Prozeß des Database Marketing ist die systematische und kontinuierliche Erfassung kundenbezogener Daten: die Ergebnisse jeder Phase werden in der Kundendatenbank dokumentiert und stehen für spätere Analysen zur Verfügung.

Der Einsatz von Data-Mining-Verfahren im Rahmen des **Database Marketing** erfolgen vor allem bei solchen Märkten, bei denen die marktlichen Transaktionsprozesse „automatisch“

[1] Entnommen aus Link, J.; Hildebrand, V. (1993), S. 31.

detaillierte kundenbezogene Daten generieren. So werden im Einzelhandel durch den Einsatz integrierter Warenwirtschaftssysteme am Point of Sale (POS) die Warenkörbe der Konsumenten erfaßt und weiterverarbeitet (vgl. Abb. 10). Diese Reaktionsdaten werden herangezogen, um Hypothesen über das Kaufverhalten der Konsumenten zu generieren. So können Muster, die Komplementaritätsbeziehungen zwischen Produkten dokumentieren, als informatorische Grundlage für marketingpolitische Maßnahmen verwendet werden.[1] Beschreibt ein Muster eine starke Komplementarität zwischen zwei Produkten, so kann durch Maßnahmen der Verkaufsraumgestaltung und Produktpositionierung eine erhöhte Präsenz der komplementären Güter erreicht werden. Die Gefahr, daß ein Konsument ein Komplementärprodukt „vergißt", kann somit verringert werden. Auch für die Artikeldisposition lassen sich Handlungsmaßnahmen ableiten. Da sich bei der Nachdisposition von Kernartikeln der Bedarf der komplementären Produkte ableiten läßt, können potentielle Präsenzlücken vermieden werden.[2]

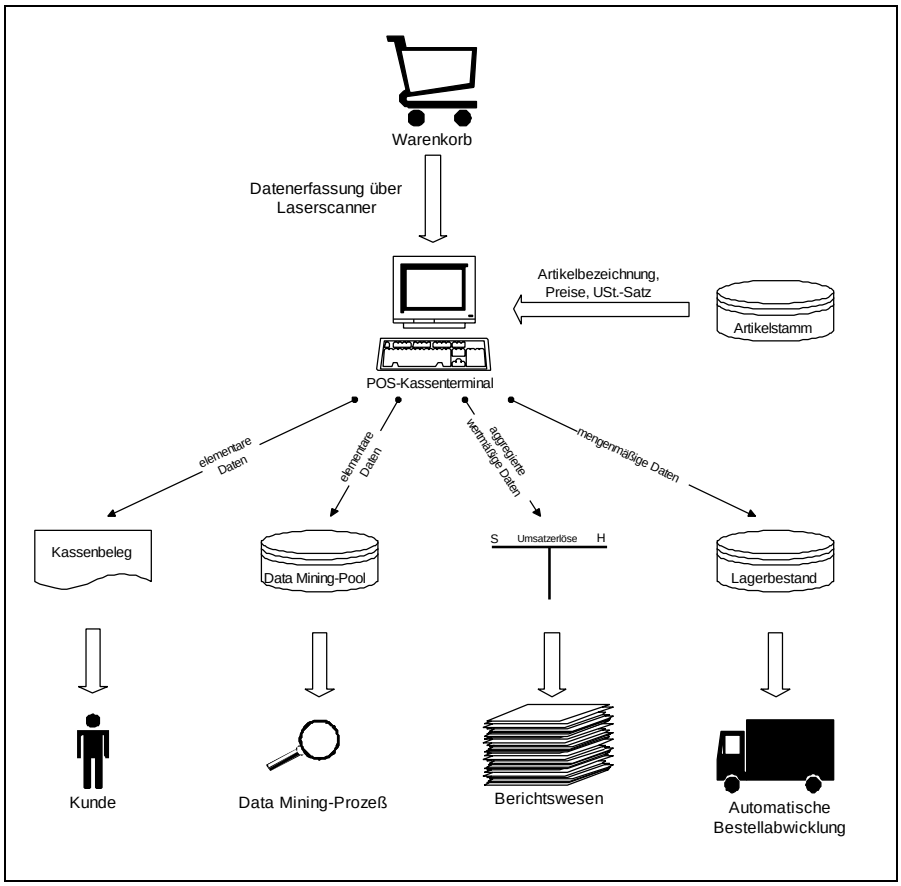

Abb. 10: Datenfluß in einem integrierten Warenwirtschaftssystem[3]

---

1 Die Zielsetzung der Assoziationsanalyse besteht in der Ableitung einer Menge von Regeln in der Form $A_1 \Lambda ... \Lambda A_m \Rightarrow B_1 \Lambda ... \Lambda B_n$ wobei A_i und B_j im Kontext der Warenkorbanalyse unterschiedliche Produkte (Items) darstellen. Vgl. Agrawal, R., Imielinski, T., Swami, A. (1993), S. 207 ff.

2 Vgl. Michels, E. (1995), S. 38.

3 Entnommen aus Alpar, P. u. a. (1998), S. 157.

Auch internetbasierte Märkte bieten ein hohes Potential für den Einsatz von Data-Mining-Verfahren. Da die marktlichen Transaktionsprozesse im Medium „Internet“ stattfinden, können kundenbezogene Daten kostengünstig und zeitnah erfaßt werden. So zeichnen Web-Server Nutzungsdaten über die Besucher von Web-Präsenzen in Form von Protokolldateien auf. Diese Protokolldateien beschreiben das Interaktionsverhalten von Konsumenten mit der Web-Präsenz in elementarer Form. Die folgende Abbildung liefert einen Auszug einer typischen Protokolldatei.

```
pcwi184.uni-muenster.de - - [17/Jun/1998:16:10:00 +0200] „GET /Sporthotel.html HTTP/1.0“ 200 3462
pcwi184.uni-muenster.de - - [17/Jun/1998:16:10:11 +0200] „GET /Reiten.html HTTP/1.0“ 200 18588
pcwi184.uni-muenster.de - - [17/Jun/1998:16:10:11 +0200] „GET /Tennis.gif HTTP/1.0“ 200 11138
pcwi184.uni-muenster.de - - [17/Jun/1998:16:11:11 +0200] „GET /Golf.html HTTP/1.0“ 200 12424
pcwi184.uni-muenster.de - - [17/Jun/1998:16:12:30 +0200] „GET /Bogen.html HTTP/1.0“ 200 15766
```

Abb. 11: Auszug aus einer typischen Protokolldatei

Aus dieser Protokolldatei wird deutlich, daß der Besucher mit der Internetadresse pcwi184.uni-muenster.de am 17. Juni 1998 ab 16:10 Uhr Informationsangebote von der Web-Präsenz eines bestimmten Sporthotels abgerufen hat.[1] Aus diesen Elementardaten können durch Anwendung geeigneter Verfahren Muster gewonnen werden, die als Entscheidungsgrundlage für die Ausgestaltung des Marketing in elektronischen Märkten dienen können. Einige Anwendungsfälle seien hier skizziert:

- Durch die Analyse der Aufrufreihenfolge der Informationsangebote können die Navigationspfade von Anwendern ermittelt werden (Pfadanalyse). Diese lassen Aussagen darüber zu, welche Informationsangebote einer Web-Präsenz von den Konsumenten als interessant erachtet werden und welche zum Verlassen der Web-Präsenz stimulieren.

- Die Assoziationsanalyse liefert Auskunft darüber, welche Informationsangebote häufig gemeinsam abgerufen werden. Derartige Muster indizieren Verbundbeziehungen zwischen Produkten und liefern Hinweise für die Gestaltung von Web-Präsenzen. So ist dem Konsumenten die Navigation zwischen komplementären Produkten durch das Einfügen von Verweisen zu erleichtern.

- Sequenzanalytische Verfahren identifizieren, nach wie vielen Besuchen der Web-Präsenz im Durchschnitt eine On-line-Bestellung erfolgt. Für das Marketing besteht hier die Zielsetzung, durch den Einsatz verkaufsfördernder Maßnahmen die erforderliche Kontakthäufigkeit zu reduzieren. Maßnahmen sind beispielsweise das dynamische Generieren von Sonderangeboten oder das Angebot von Testversionen zu Software oder anderen Medien, die on line distribuiert werden.

Die dargestellten Beispiele zeigen, daß Data Mining im Kontext internetbasierter Marktsysteme seinen Nutzen erst dann voll entfalten kann, wenn es zur *direkten* Steuerung des elektronischen Marketinginstrumentariums eingesetzt wird. Zur Realisierung eines derartigen

---

[1] Eine Beschreibung der protokollierten Attribute findet sich bei Bensberg, F., Weiß, T. (1998), S. 197 ff.

kundenindividuellen *Echtzeit-Marketings* werden dynamische Content Management-Systeme eingesetzt, die durch Analyse von Nutzungsdaten individuelle Informations- und Produktangebote konfigurieren können.[1] Dabei ist jedoch zu beachten, daß die Aufzeichnung potentiell personenbezogener Daten Schutzrechte verletzen kann. Insbesondere die Erstellung personenbezogener Kundenprofile ist nach den Rechtsvorschriften des Bundesdatenschutzgesetzes und des Gesetzes zur Regelung der Rahmenbedingungen für Informations- und Kommunikationsdienste (IuKDG) unzulässig.[2]

4 Data-Mining-Software - Darstellung eines Beispiels

Zur Darstellung eines Beispiels wird hier das Data-Mining-System Intelligent Miner for Data (Version 2.1.3) von der International Business Machines Corporation (IBM) verwendet. Dieses Softwareprodukt ist eine integrierte Data-Mining-Plattform, die alle Phasen des in Abschnitt 2 dargestellten Prozeßmodells realisiert und durch die Unterstützung paralleler Hardwarearchitekturen eine hohe Skalierbarkeit aufweist. Die Architektur des Systems orientiert sich am Client/Server-Prinzip und besteht aus den Komponenten des Intelligent Miner Server und des Intelligent Miner Client. Serverseitig werden alle Data-Mining-Schritte von der Datenextraktion bis hin zur Methodenanwendung ausgeführt.[3] Der Intelligent Miner Client ist eine Java-Anwendung, die Funktionen zur Administration, Visualisierung (Visualizer) und zum Export von Ergebnissen bereitstellt.

Die Methodenbank des Intelligent Miner erlaubt die Bearbeitung der folgenden Aufgabenstellungen:

- Assoziationsanalyse (Associations)
- Klassifikation (Classification)
- Clusteranalyse (Clustering)
- Sequenzanalyse (Sequential Patterns)
- Prognoseverfahren (Prediction)

Für die Bearbeitung dieser Aufgabenstellungen verfügt Intelligent Miner über unterschiedliche Verfahren, die in Abhängigkeit vom Anwendungskontext auszuwählen sind. Außerdem besitzt das System eine Reihe statistischer Funktionen (z. B. Faktorenanalyse, Regressions-

---

1 Vgl. Luedi, A. F. (1997), S. 22 ff.

2 Vgl. Kargl, H., Guba, A. (1999), S. 347.

3 Das Datenmanagement des Intelligent Miner Server basiert auf dem Datenbankmanagementsystem DB2 Universal Database (IBM).

verfahren, Hauptkomponentenanalyse), die sich auch in den Analysesystemen SPSS und SAS befinden.

Im folgenden wird die Anwendung des Intelligent Miner anhand eines Beispiels dargestellt. Das betriebswirtschaftliche Ziel des Beispiels sei die Kundensegmentierung. Die Kunden werden in der Beispieldatenbank durch die Attribute Geschlecht, Familienstand und Einkommen beschrieben. Als Data-Mining-Methode wird das Verfahren des Demographischen Clusterns angewendet. Bei dieser Methode handelt es sich um ein von IBM entwickeltes Clusteranalyseverfahren, das im Gegensatz zu den klassischen Verfahren der multivariaten Statistik auch bei großen Datenmengen effizient eingesetzt werden kann. An dieser Stelle wird angenommen, daß die notwendigen Daten aus den operativen Systemen extrahiert wurden und in bereinigter Form vorliegen. Die Oberfläche des Intelligent Miner Client stellt diesen Sachverhalt wie folgt dar:

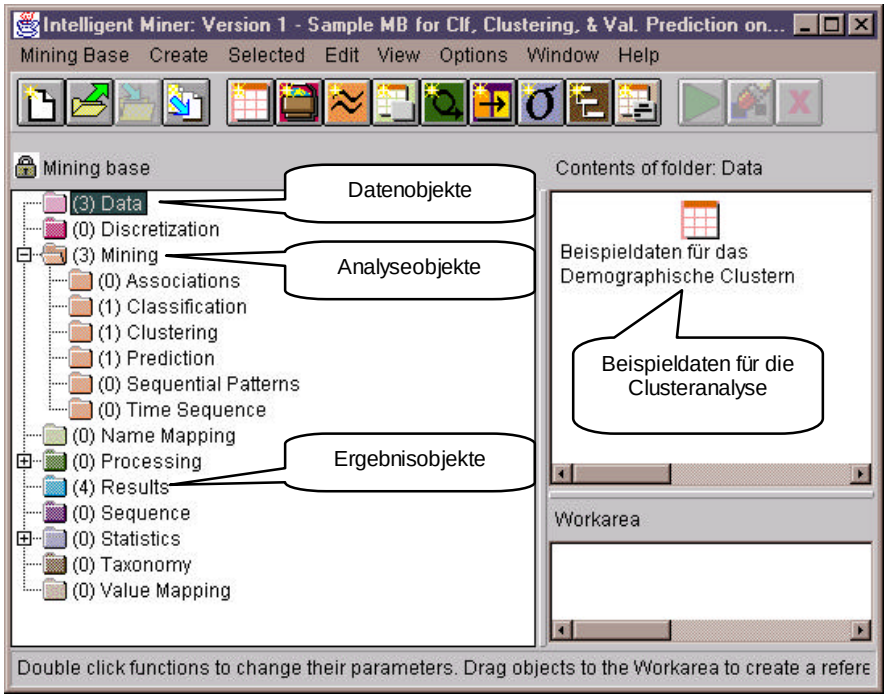

Abb. 12: Die Oberfläche des Intelligent Miner Client

Die Verwaltung der Arbeitsumgebung erfolgt unter Intelligent Miner in Form einer sog. Mining Base. Eine Mining Base ist ein Objektcontainer, in dem der Anwender Objekte anlegt. Zu den zentralen Objekttypen sind Datenobjekte (Data), Analyseobjekte (Mining) und Ergebnisobjekte (Results) zu zählen. Bei der Definition von Datenobjekten legt der Anwender den Datenbestand für die Analyse fest. Die anzuwendende Methode wird in Form von Analyseobjekten (Associations, Classification, Clustering usw.) bestimmt und assistentengesteuert parametrisiert. Außerdem wird festgelegt, welche Datenbestände zur Analyse verwendet und welche Ergebnisobjekte generiert werden. Die Beziehungen zwischen Datenbestand, Analyseobjekt und Ergebnisobjekt werden von Intelligent Miner grafisch dargestellt (vgl. Abb. 13). Zur Automatisierung von komplexen Arbeitsabläufen können Sequenzobjekte angelegt wer-

den, denen z. B. mehrere Analyseobjekte zugewiesen werden können. Wird das Sequenzobjekt vom Anwender aktiviert, werden alle Analyseobjekte gestartet und die Mustererkennung durchgeführt.

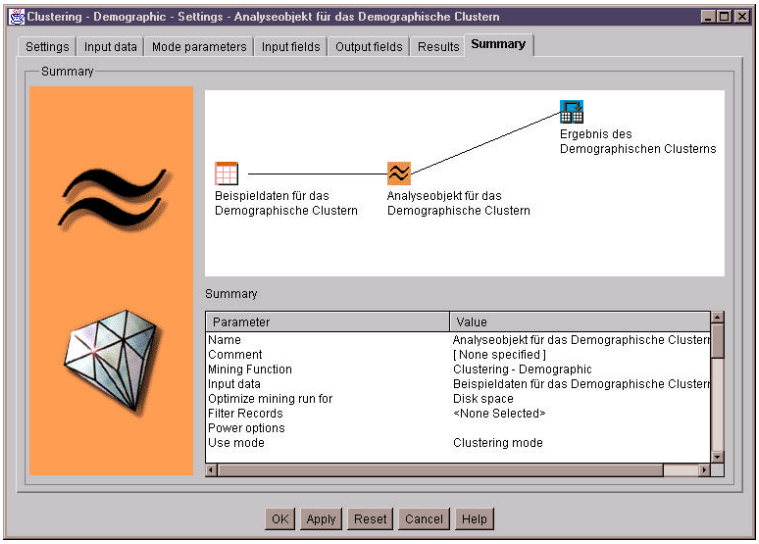

Abb. 13: Grafische Darstellung der parametrisierten Objekte (Intelligent Miner)

Nach der Parametrisierung des Analyseobjekts kann die Methodenanwendung erfolgen. Das hier gewählte Verfahren des Demographischen Clusterns berechnet die Ähnlichkeit der Objekte des Datenbestands anhand der Attributwerte und ermittelt die Anzahl der Cluster.[1] Nach der Methodenanwendung erfolgt die Visualisierung des Ergebnisobjektes, das die Gesamtheit der identifizierten Cluster beschreibt.

[1] Vgl. o. V. (1998), S. 127.

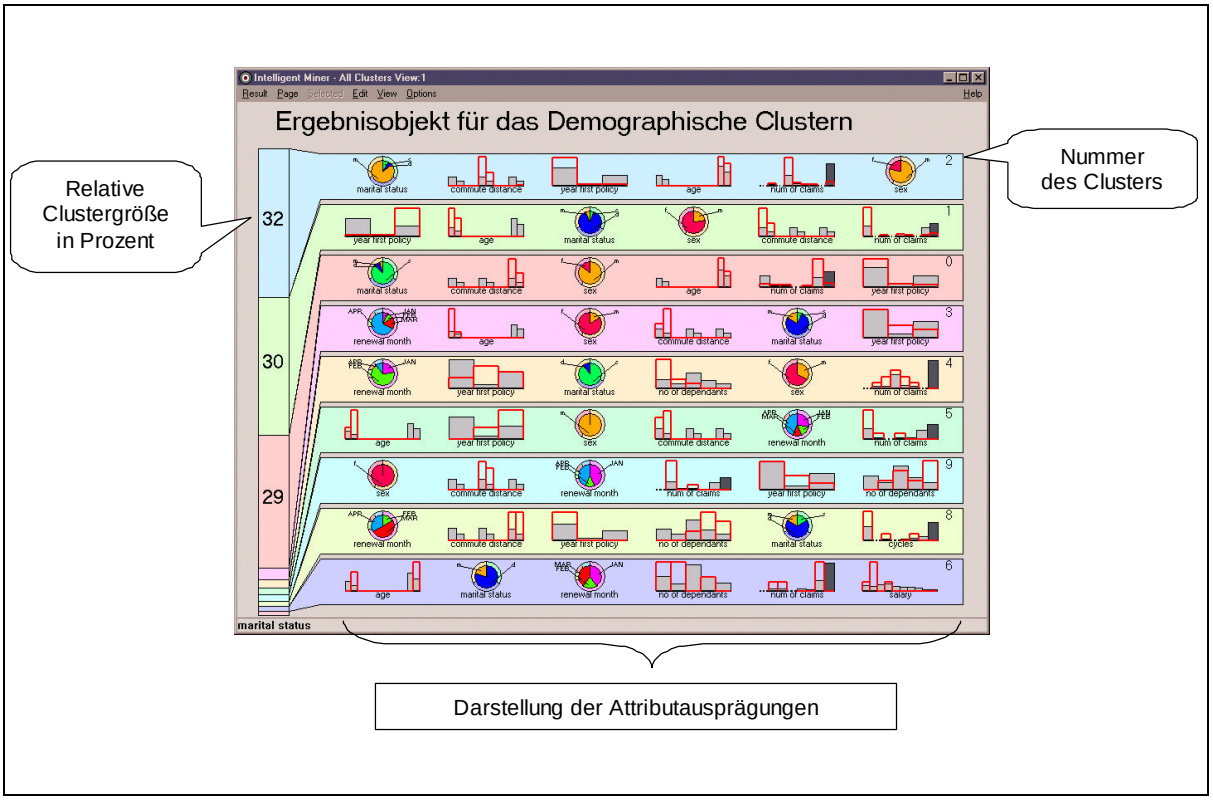

Abb. 14: Ergebnisobjekt des Demographischen Clusterns (Intelligent Miner)

Die Methodenanwendung hat insgesamt neun Cluster identifiziert, die im Ergebnisobjekt horizontal dargestellt werden. Die Sortierung der Cluster erfolgt dabei nach der relativen Größe (z. B. 32% für den Cluster Nr. 2). Für jeden Cluster werden außerdem die Attributausprägungen grafisch dargestellt. Die Darstellung kategorieller Attribute (z. B. Familienstand) erfolgt als Kreisdiagramm, während numerische Attribute (z. B. Alter) in Form zweidimensionaler Balkendiagramme dargestellt werden. Bei der Attributdarstellung werden in jedem Diagramm die Werte für den jeweiligen Cluster sowie für alle Kunden dargestellt. Die Sortierung der Attribute erfolgt in Abhängigkeit von dem Einfluß auf die Clusterbildung (mit abnehmendem Einfluß von links nach rechts).

Die zusammenfassende Darstellung der Cluster liefert dem Anwender eine erste Übersicht über die Ergebnisse der Analyse. Zur Untersuchung einzelner Cluster ist die detaillierte Darstellung von Clustern und Attributen erforderlich. Zu diesem Zweck verfügt Intelligent Miner über einen Drill-Down-Mechanismus, der den Durchgriff bis auf die elementaren Analyseergebnisse ermöglicht. In Abb. 15 wird diese Detaillierung bis hin zu den numerischen Analyseergebnissen dargestellt.

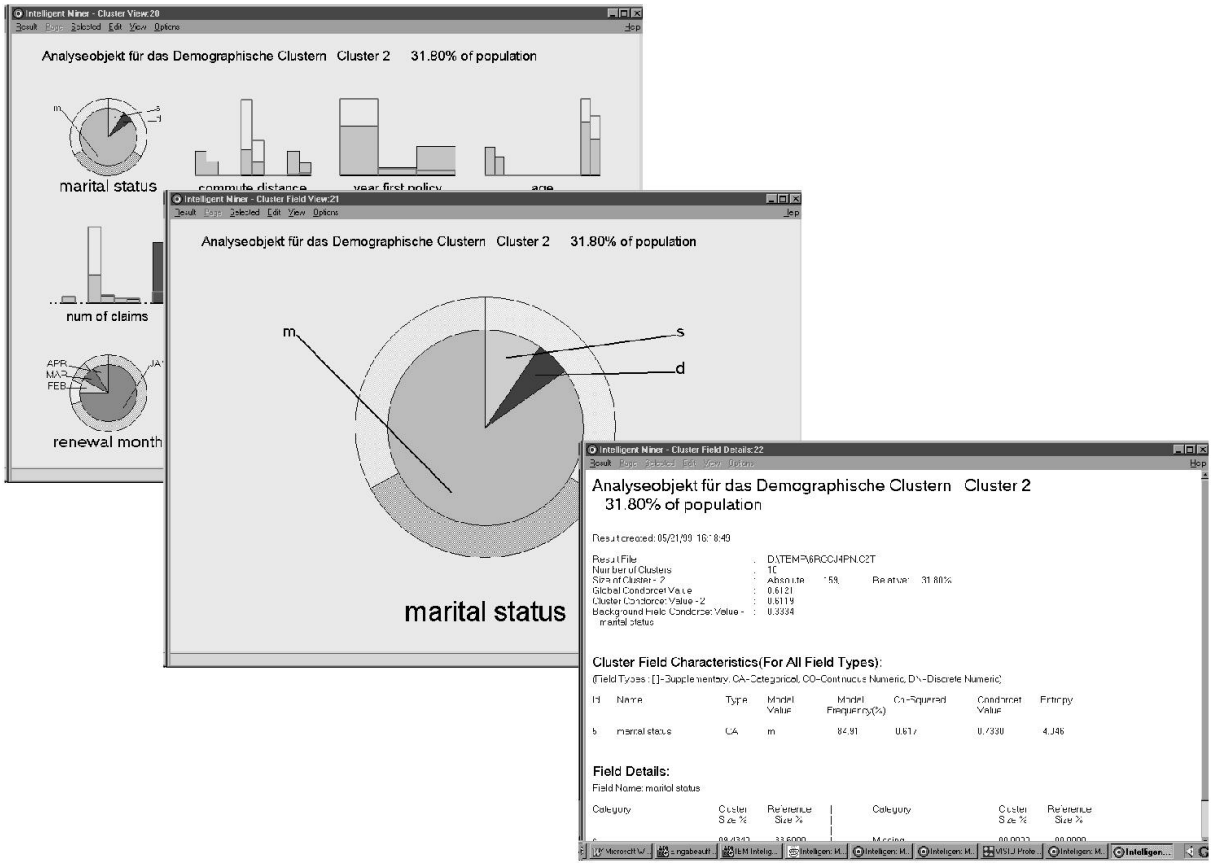

Abb. 15: Detaillierung der Analyseergebnisse (Intelligent Miner)

Zur Weiterverarbeitung der Analyseergebnisse bietet Intelligent Miner den Export im Post-Script-Format an. Auf diese Weise können die Ergebnisse manuell in das betriebliche Berichtswesen übernommen werden. Um ein „automatisches“ Data Mining zu ermöglichen, das nicht mehr auf die Benutzerinteraktion angewiesen ist, steht eine API-Schnittstelle zur Verfügung. Über diese Schnittstelle kann der Zugriff auf alle Funktionen des Data-Mining-Prozesses erfolgen und das System in die Geschäftsprozesse integriert werden. Die Nutzung dieser Schnittstelle ist vor allem dann erforderlich, wenn der relevante Datenbestand des Anwendungskontextes einer stetigen Änderung unterliegt und die zeitnahe, automatische Mustererkennung notwendig ist.[1]

5 Ausblick

Das Forschungsgebiet des Data Mining weist ein hohes Maß an interdisziplinär fundierter Anwendungsorientierung auf. Die aktuellen Entwicklungen im Bereich kommerzieller und nicht-kommerzieller Data-Mining-Systeme machen deutlich, daß diese Forschungsstrategie in

[1] Beispielhafte Anwendungsfälle sind die Überprüfung von Kreditkartentransaktionen oder die Überwachung von Netzwerken.

vergleichsweise kurzer Zeit eine Klasse von Informationssystemen hervorgebracht hat, die eine hohe praktische Akzeptanz genießen.[1] Trotzdem sind im Forschungsprogramm des Data Mining wesentliche Fragestellungen bisher noch unbeantwortet geblieben.

Aus betriebswirtschaftlicher Perspektive stellt sich die Frage nach der Bewertung von Data-Mining-Systemen und der organisatorischen Gestaltung ihrer Anwendung. Zur Bewertung von Data-Mining-Systemen sind Methoden notwendig, die die Ergebnisse eines Data-Mining-Systems einer theoretisch fundierten Beurteilung zuführen können. Die Entwicklung und Evaluation multikriterieller Bewertungsansätze für die objektive oder subjektive „Interessantheit" von Mustern befindet sich derzeit noch in den Anfängen.[2] Mit dem zunehmenden Einsatz von Data-Mining-Systemen ist jedoch damit zu rechnen, daß die Frage der Wirtschaftlichkeit durch empirische Methoden der Evaluationsforschung beantwortet wird. In Bezug auf den Organisationsaspekt sind derzeit Bestrebungen zu beobachten, die die Entwicklung eines standardisierten Data-Mining-Referenzmodells verfolgen.[3] Dieses Modell verfeinert den hier dargestellten Data-Mining-Prozeß und liefert einen Bezugsrahmen für die organisatorische Ausgestaltung des Data Mining.

Aus Sicht der Informatik hat sich Data Mining zu den „Grand Challenges" an die derzeit verfügbaren Hochleistungsrechner entwickelt.[4] Begründet wird dies durch die Tatsache, daß Datenbanken zur Mustererkennung mehrfach traversiert werden müssen. Bei Datenvolumen im Giga- und Terabyte-Bereich, die von betrieblichen Data Warehouse-Systemen mittlerweile erreicht werden, weisen Data-Mining-Systeme immer noch Effizienzprobleme auf, die erst durch die Weiterentwicklung der zugrundeliegenden Algorithmen gelöst werden können.[5]

Für den Wissenschaftler stellt das Data-Mining-Konzept ein methodenorientiertes Werkzeug zur Entdeckung und Erklärung empirischer Phänomene dar. Ein wissenschaftstheoretischer Bezugsrahmen für den methodologisch fundierten Einsatz von Data Mining ist derzeit allerdings noch nicht vorhanden. Erste Arbeiten im Bereich der Entdeckungswissenschaft (discovery science) deuten jedoch darauf hin, daß Data Mining einen theoretisch begründeten Platz im Instrumenatrium der empirischen Forschung einnehmen wird.

---

1 Der Aspekt der Vergleichbarkeit bezieht sich auf die Forschungsaktiväten im Bereich entscheidungsunterstützender Systeme in den 70er und 80er Jahren.

2 Vgl. Müller, M., Hausdorf, C., Schneeberger, J. (1998), S. 248 ff.

3 Vgl. Chapman, P. et al. (1999), S. 1 ff.

4 Vgl. Cap, C. H. (1998), S. 50 ff.

5 Vgl. Nakhaeizadeh, G., Reinartz, T., Wirth, R. (1998), S. 27.

Literatur

Agrawal, R., Imielinski, T., Swami, A. (1993), Mining Association Rules between Sets of Items in Large Databases, in: SIGMOD Record, 5/1993, S. 207-216.

Agrawal, R., Srikant, R., Mining Sequential Patterns, IBM Research Report RJ9910, San Jose 1994, im WWW unter http://www.almaden.ibm.com/cs/people/ragrawal/papers/-icde95_rj.ps [10.05.1999].

Alpar, P., Grob, H. L., Weimann, P., Winter, R. (1998), Unternehmensorientierte Wirtschaftsinformatik, Braunschweig, Wiesbaden 1998.

Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (1996), Multivariate Analysemethoden – Eine anwendungsorientierte Einführung, 8. Aufl., Berlin u. a. 1996.

Behme, W., Muksch, H. (1998), Auswahl und Klassifizierung externer Informationen zur Integration in ein Data Warehouse, in: Integration externer Informationen in Management Support Systems, Hrsg.: Uhr, W., Breuer, S. E., Dresden 1998, S. 85-104.

Bensberg, F., Weiss, T. (1998), Web Log Mining als Analyseinstrument des Electronic Commerce, in: Integration externer Informationen in Management Support Systems, Hrsg.: Uhr, W., Breuer, S. E., Dresden 1998, S. 197-214.

Berry, M. J. A., Linoff, G. (1997), Data Mining Techniques : For Marketing, Sales and Customer Support, New York 1997.

Bissantz, N. (1996), CLUSMIN - Ein Beitrag zur Analyse von Daten des Ergebniscontrollings mit Datenmustererkennung (Data Mining), Diss. Univ. Erlangen-Nürnberg 1996.

Cap, C. H. (1998), Wirtschaftliche Anwendungen: Die neuen Grand Challenges der Informatik?, in: HMD 203/1998, S. 50-57.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., Wirth, R. (1999), The CRISP-DM Process Model, o. O. 1999. Im WWW unter: http://www.ncr.dk/CRISP [10.05.1999].

Chen, M.-S., Han, J., Yu, P. S. (1997), Data Mining: An Overview from Database Perspective, 1997, im WWW unter ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/survey97.ps [10.05.1999].

Fayyad, U. M. (1997), Editorial, in: Data Mining and Knowledge Discovery, Vol. 1 Issue 1 1997, S. 5-10.

Fayyad, U. M., Djorgovski, S. G., Weir, N. (1996), Automating the Analysis and Cataloging of Sky Surveys, in: Advances in Knowledge Discovery and Data Mining, Hrsg.: Fayyad, U. M. u. a., Cambridge 1996, S. 471-493.

Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P. (1996a), From Data Mining to Knowledge Discovery: An Overview, in: Advances in Knowledge Discovery and Data Mining, Hrsg.: Fayyad, U. M. u. a., Cambridge 1996, S. 1-33.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. (1996b), From Data Mining to Knowledge Discovery, in: AI Magazine, Vol. 17 No. 3 1996, S. 37-51.

Fayyad, U., Haussler, D., Stolorz, P. (1996), KDD for Science Data Analysis: Issues and Examples, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Hrsg.: Simoudis, E., Han, J., Fayyad, U., Menlo Park 1996, S. 50-56.

Flanagan, T.; Safdie, E. (1997), Leveraging Visual and Analytical Data Mining, 1997, im WWW unter http://www.techguide.com/forum/dw/visual/index.html [10.05.1999].

Frawley, W. J., Piatetsky-Shapiro, G., Matheus, C. J. (1991), Knowledge Discovery in Databases: An Overview, in: Knowledge Discovery in Databases, Hrsg.: Piatetsky-Shapiro, G., Frawley, W. J., Menlo Park 1991, S. 1-27.

Guyon, I., Matic, N., Vapnik, V. (1996), Discovering Informative Patterns and Data Cleaning, in: Advances in Knowledge Discovery and Data Mining, Hrsg.: Fayyad, U. M. u. a., Cambridge 1996, S. 181-203.

Grob, H. L., Bieletzke, S. (1997), Aufbruch in die Informationsgesellschaft, Münster 1997.

Gluchowski, P., Gabriel, R., Chamoni, P. (1997), Management Support Systeme : computergestützte Informationssysteme für Führungskräfte und Entscheidungsträger, Berlin 1997.

Hagedorn, J.; Bissantz, N.; Mertens, P. (1997), Data Mining (Datenmustererkennung): Stand der Forschung und Entwicklung, in: Wirtschaftsinformatik, 39/1997, S. 601-612.

Heiting, M. (1996), Die Suche nach versteckten Informationen - Data Mining, in: it Management, Nr. 09-10/1996, S. 10-12.

Henneböle, J. (1995), Executive Information Systems für Unternehmensführung und Controlling: Strategie – Konzeption – Realisierung, Wiesbaden 1995.

Kargl, H., Guba, A. (1999), Online-Monitoring – Gewinnung und Verwendung von Online-Daten, in: WISU 3/1999, S. 345-351.

Krahl, D., Windheuser, U., Zick, F.-K. (1998), Data Mining : Einsatz in der Praxis, Bonn 1998.

Krivda, C. D. (1995), Data Mining Dynamite. Blasting loose those buried nuggets of information requires clean data, warehousing strategies, powerful parallel processors, and heaps of hard disk space, in: Byte 10/1995, S. 97-102, im WWW unter http://www.byte.com/art/9510/sec8/art9.htm [10.05.1999].

Küppers, B. (1998), Data Mining in der Praxis – Ein Ansatz zur Nutzung der Potentiale von Data Mining im betrieblichen Umfeld. Frankfurt a. M. u. a. 1998.

Lu, H., Han, J., Feng, L. (1998), Stock Movement Prediction and N-Dimensional Inter-Transaction Association Rules, 1998, im WWW unter ftp://ftp.fas.sfu.ca/pub/cs/han/kdd/intertran98.ps [10.05.1999].

Luedi, A. F. (1997), Personalize or Perish, in: Electronic Markets, Vol. 7 No. 3 1997, S. 22-25.

Link, J., Hildebrand, V. (1993), Database Marketing und Computer aided selling: strategische Wettbewerbsvorteile durch neue informationstechnologische Systemkonzeptionen, München 1993.

Markowitz, H. M.; Lin Xu, G. (1994), Data Mining Corrections - Simple and plausible, in: The Journal of Portfolio Management, Vol. 21 1994, S. 60-69.

Matheus, C. J.; Piatetsky-Shapiro, G., McNeill, D. (1996), Selecting and Reporting what is interesting, in: Advances in Knowledge Discovery and Data Mining. Hrsg.: U. M. Fayyad et al. Cambridge 1996, S. 495-515.

Mertens, P. u. a.(1994), Datenmustererkennung in der Ergebnisrechnung mit Hilfe der Clusteranalyse, in: Die Betriebswirtschaft, 54/1994, S. 739-753.

Michels, E. (1995), Datenanalyse mit Data Mining, in: Dynamik im Handel, 11/1995, S. 37-43.

Müller, M., Hausdorf, C., Schneeberger, J. (1998), Zur Interessantheit bei der Entdeckung von Wissen in Datenbanken, in: Data Mining – Theoretische Aspekte und Anwendungen. Hrsg.: Nakhaeizadeh, G., Heidelberg 1998, S. 248-264.

Nakhaeizadeh, G., Reinartz, T., Wirth, R. (1998), Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick, in: Data Mining – Theoretische Aspekte und Anwendungen, Hrsg.: Nakhaeizadeh, G., Heidelberg 1998, S. 1-33.

o. V. (1999), The Netcraft Web Server Survey, 1999, im WWW unter http://www.netcraft.com/Survey [17.4.1999].

o. V. (1998), Using the Intelligent Miner for Data, Benutzerhandbuch zum Intelligent Miner for Data Version 2 Release 1, Hrsg.: International Business Machines Corporation, 3. Aufl., o. O. 1998.

Picot, A., Reichwald, R., Wigand, R. T. (1998), Die grenzenlose Unternehmung : Information, Organisation und Management, 3. Aufl., Wiesbaden 1998.

Redman, T. C. (1998), The Impact of Poor Data Quality on the Typical Enterprise, in: Communications of the ACM, Vol. 41 No. 2 1998, S. 79-82.

Reese Hedberg, S. (1995), The Data Gold Rush - Smart data miners are cashing in on valuable information buried in private and public data sources, in: Byte 10/1995, S. 83-88, im WWW unter http://www.byte.com/art/9510/sec8/art2.htm [10.05.1999].

Tuzhilin, A. (1997), Editor’s introduction to the special issue on knowledge discovery and its applications to business decision-making, in: Decision Support Systems, 21/1997, Special issue on knowledge discovery and its applications to business decision making, S. 1-2.

Wessling, E. (1991), Individuum und Information: die Erfassung von Information und Wissen in ökonomischen Handlungstheorien, Tübingen 1991.

Arbeitsberichte der Reihe Computergestütztes Controlling

Nr. 1 Grob, H. L., Positionsbestimmung des Controlling, Arbeitsbericht Nr. 1, Münster 1996.

Nr. 2 Grob, H. L., Weigel, L., Flexible Investitionsplanung mit VOFI - Integration von VOFI und DPL, Arbeitsbericht Nr. 2, Münster 1996.

Nr. 3 Meininger, P., Differenzanalyse bei LP-Modellen, Arbeitsbericht Nr. 3, Münster 1996.

Nr. 4 Borkenfeld, A., Fuzzy VOFI, Arbeitsbericht Nr. 4, Münster 1996.

Nr. 5 Ziegenbein, R., CriterEUS - Ein multikriterielles Entscheidungsunterstützungssystem unter Excel, Arbeitsbericht Nr. 5, Münster 1996.

Nr. 6 Schulenburg, K., Liquiditätsplanung mit VOFI, Arbeitsbericht Nr. 6, Münster 1997.

Nr. 7 Grob, H. L., Mrzyk, A., Risiko-Chancen-Analyse in der Investitionsrechnung - Integration von VOFI und Crystal Ball - Arbeitsbericht Nr. 7, Münster 1997.

Nr. 8 Grob, H. L., Bensberg, F., Das Data-Mining-Konzept, Arbeitsbericht Nr. 8, Münster 1999.