OLAP-Technologie und mathematisch-statistische Verfahren des Data Mining

Rajnish Tiwari

(Rajnish.Tiwari@uni-hamburg.de)

http://www.rrz.uni-hamburg.de/RRZ/R.Tiwari/

April 2002

Universität Hamburg

Fachbereich Wirtschaftswissenschaften

INHALTSVERZEICHNIS

1. EINLEITUNG 3

2. BEGRIFFSBESTIMMUNG 4

2.1 BUSINESS INTELLIGENCE 4

2.2 KNOWLEDGE DISCOVERY IN DATABASES (KDD) 4

3. OLAP-TECHNOLOGIE 5

3.1 MULTIDIMENSIONALITÄT VON DATEN 6

3.2 AUSPRÄGUNGEN DES OLAP 6

3.3 EINSATZMÖGLICHKEITEN DES OLAP 7

3.4 GRENZEN DES OLAP 7

4. DATA MINING 7

4.1 METHODEN DES DATA MINING 8

4.1.1. Strukturen-prüfende Verfahren 8

4.1.2. Strukturen-entdeckende Verfahren 11

4.2 EINSATZMÖGLICHKEITEN DES DATA MINING 13

4.2.1. Customer Relations Management (CRM) 14

4.2.2. Text-Mining 14

4.2.3. Web-Mining 15

4.3 GRENZEN DES DATA MINING 16

5. VERGLEICH ZWISCHEN OLAP UND DATA MINING 17

6. SCHLUSSBETRACHTUNG 18

1. Einleitung

Die Unternehmenssituation ist heute gekennzeichnet durch Wettbewerbsverschärfung, zunehmenden Kostendruck, steigende Preis- und Service-Sensibilität der Kunden bei gleichzeitig abnehmender Kundenloyalität. Der globale Wettbewerb zwingt Firmen, die Informationen über ihre Kunden, Lieferanten und Geschäftsprozesse immer schneller zu analysieren und in Marktvorteile umzuwandeln. Der technische Fortschritt zwingt/ermuntert zur Qualitätsverbesserung und Produktinnovation.

Es herrscht aber auch ein komischer „Informationsmangel trotz Daten-Flut“, weil es immer schwieriger wird, von den exponential wachsenden Datenmengen relevante Informationen „herauszufischen“. Nach einer Untersuchung der Zeitschrift „SAP Info“ beklagen 47 Prozent aller befragten Manager, die Suche nach Informationen halte sie von ihren eigentlichen Aufgaben ab.[1] Die zunehmende Datenmenge und -komplexität erfordert intelligente Retrieval- und Analyseinstrumente. In der vorliegenden Arbeit werden die OLAP-Technologie und mathematisch-statistische Methoden des Data Mining untersucht, die immer häufiger für diese Zwecke in Unternehmen eingesetzt werden.

Zuerst werden die Begriffe „Business Intelligence“ und „Knowledge Discovery in Databases“ (im folgenden meistens kurz: KDD) und deren Bezug zu Data Warehouse sowie zur OLAP-Technologie und Data Mining definiert. Im 3. Kapitel wird die OLAP-Technologie behandelt. Es werden neben Begriffsbestimmung, die Ausprägungen sowie die Einsatzmöglichkeiten und Grenzen des OLAP gezeigt. Im 4. Kapitel wird das Thema Data Mining behandelt. Nach Begriffsbestimmung werden die mathematisch-statistischen Methoden des Data Mining vorgestellt. Anschließend werden die Einsatzmöglichkeiten insbesondere in den Bereichen „Customer Relations Management“ (im folgenden kurz: CRM), Web-Mining und Text-Mining, sowie die Grenzen des Data Mining gezeigt. Kapitel 5 vergleicht die beiden Technologien. Das 6. Kapitel. beinhaltet die Schlussbetrachtungen.

---

[1] Vgl. Gentsch, Peter: „Wie aus Daten Wissen wird“, Online im Internet http://www.sapinfo.net/public/de/article.php4/comvArticle-193353c63adb9bc782/de, Stand 16.05.00, Abruf 06.04.02, S. 1.

2. Begriffsbestimmung

2.1 Business Intelligence

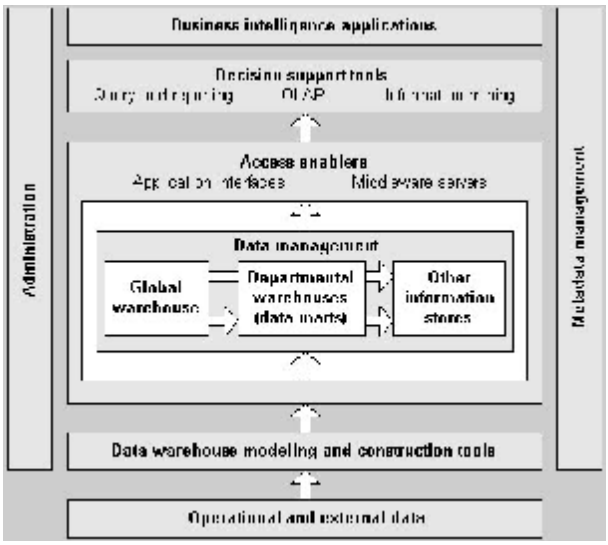

Abb. 1 (Quelle: IBM 2001)
BI is about making well informed decisions, using information that is based on data.

„Business Intelligence ist das Entdecken geschäftsrelevanten Wissens in Form von Strukturen und Mustern durch die intelligente Kombination menschlichen Kalküls mit moderner Informationstechnologie."[2] Das Entdecken ist dabei als einen Wertschöpfungsprozess zu verstehen, der Unternehmens-, Markt- und Wettbewerbsdaten in handlungsgerichtetes, strategisches Wissen transformiert und zwar über die Position, Performance, Fähigkeiten und Intentionen des eigenen wie auch der Konkurrenz.

2.2 Knowledge Discovery in Databases (KDD)

KDD bezeichnet den Prozess, gültige, bisher unbekannte, verständliche und nützliche Informationen aus Datenbeständen zu gewinnen.[3] KDD beschreibt automatisierte Verfahren, mit denen Gesetzmäßigkeiten in Mengen von Datensätzen gefunden werden.

Es wird in der Literatur auch als „Knowledge Extraction", „Data Analysis" oder „Data Archaeology" bezeichnet.

Der KDD-Prozess beginnt mit einer Vorstufe, in welcher relevantes, schon vorhandenes Wissen über den gewünschten Anwendungsbereich gesammelt wird. Daher ist eine konsistente, qualitativ hochwertige Datenbasis in der Form eines Data Warehouse[4] (im folgenden kurz: DW) für Erfolg eines KDD-Systems (sei es

---

[2] Gentsch, Peter, a.a.O., S. 1

[3] Fayyad, Usama et al.: "From Data Mining to Knowledge Discovery: An Overview", in: Fayyad, Usama et al.: *Advances in Knowledge Discovery and Data Mining*, Menlo Park u.a. 1996, S. 1-34

[4] Ein Data Warehouse ist ein subjektorientierter, integrierter, zeitvarianter und dauerhafter Datenpool, der online jederzeit für entscheidungsunterstützende Maßnahmen herangezogen werden.

OLAP oder Data Mining) von großer Bedeutung. Ein Data Warehouse besteht aus den drei Grundelementen[5]:

i) Datenmanagement,

ii) Datenorganisation, und

iii) Datenauswertung sowie –Aufbereitung

Das wichtigste Instrument der Auswertung/Aufbereitung von Daten stellt Data Mining dar, das durch Analyse des Datenbestandes noch unbekannte Zusammenhänge von Daten aufzudecken versucht. Ein anderes, ebenfalls wichtiges Instrument für das Data Warehouse und Management-Support-Systeme ist OLAP, das sich als einen guten Baustein innerhalb dieser Auswertung-/Aufbereitungskomponente einsetzen lässt[6].

KDD wird manchmal auch als Synonym für Data Mining verwendet, was aber nicht völlig richtig ist. KDD ist der gesamte Informationsgewinnungsprozess von Fragenformulierung bis zur Ergebnisinterpretation, Data Mining hingegen nur der Prozess der Muster- und Trenderkennung. Der Unterschied liegt darin, dass:

„KDD refers to the overall process of discovering useful Knowledge from data while data while Data Mining refers to a particular step in this process."[7]

3. OLAP-Technologie

OLAP ist ein Oberbegriff für Technologien und Methoden, die der Analyse multidimensionaler Informationen dienen und den Entscheidungsträgern relevante Informationen zur Verfügung stellen.

Es umfasst die Bereitstellung und Aufbereitung von Informationen zu Zwecken der betrieblichen Analyse und Berichtserstellung hinsichtlich Ressourceneinsatz, Prozessleistung und Unternehmenserfolg. OLAP erlaubt eine mehrdimensionale Betrachtung der Daten (z.B. Produkt, Region, Zeit etc.). Dadurch wird eine Analyse nach verschiedenen Kriterien ermöglicht. Somit bilden OLAP-Systeme den Kern von den sog. „Executive Information Systems".[8]

---

5 Vgl. Mertens, Peter et al.: „Grundzüge der Wirtschaftsinformatik", 7. Auflage, Berlin u.a., 2001, S. 75 ff.

6 Vgl. Mertens, Peter et al., a.a.O. S. 76 ff.

7 Fayyad, Usama: "Knowldege Discovery in Database: An Overview", in Saso Dzeroski; Nada Lavrac (ed.): *Relational Data Mining*, Berlin u.a. 2001, S. 32.

8 Vgl. Hummeltenberg, Wilhem: „Data Warehousing: Management des Produktionsfaktors Information – eine Idee und ihr Weg zum Kunden", in Martin, Wolfgang (Hrsg.): *Data Warehousing / Data Mining – OLAP*, 1. Auflage, Bonn, 1998, S. 55.

„While OLAP systems have the ability to answer 'who?' And 'what?' questions, it is their ability to answer 'what if?' And 'why?' that sets them apart from data warehouses. OLAP enables decision making about future actions."[9]

OLAP-Werkzeuge unterstützen den schrittweisen Prozess der Analyse mit einer Folge von Fragen und dem Beurteilen von Auswerteergebnissen. Der Anwender kann seine Fragen modifizieren, bis er brauchbare Ergebnisse erhält.[10]

3.1 Multidimensionalität von Daten

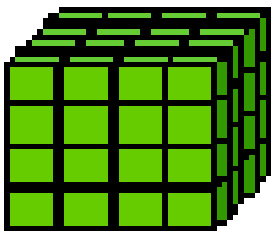

Abb. 2 OLAP-Dimensionen [Quelle: IBM]

Eine konzeptionelle Sicht der betriebswirtschaftlichen Analyse muss multidimensional sein, denn Kosten- und Umsatzgrößen sind nur in ihrem Bezug auf Kunden, Produkt, Region etc. aussagekräftig, daher arbeitet OLAP extensiv mit Dimensionen[11]. Um komplexe betriebwirtschaftliche Analysen durchzuführen, werden Daten entweder in relationalen Tabellen oder in sog. Cubes (deutsch: Würfeln) gespeichert. Die Cubes können Daten über verschiedene Kriterien/Dimensionen enthalten, z.B. Quartal, Umsatz, Region, Verkäufer etc. Diese Cubes lassen sich dann beliebig manipulieren. Unterschieden wird zwischen dem Relational Online Analytical Processing (im folgenden kurz: ROLAP) und dem Multidimensional Online Analytical Processing (im folgenden kurz: MOLAP), die hier aber aus Platzgründen nicht näher behandelt werden. Die Multidimensionalität der Daten bildet den Kernpunkt vom OLAP und ermöglicht die Berücksichtigung von hierarchischen Strukturen in Daten.

3.2 Ausprägungen des OLAP

1993 wurden von Ted Codd 12 OLAP-Regeln veröffentlicht, die er später um 12 weiter erweiterte und die noch immer als Basis zur Beurteilung von OLAP-Systemen dienen[12]. Aber da diese relativ viel sind und damit zur Unübersichtlichkeit führen, wird ein FASMI-Ansatz (**F**ast **A**nalysis of **S**hared **M**ultidimensional

---

[9] IBM Redbook: „Mining Your Own Business in Banking / Using DB2 Intelligent Miner for Data", 1. Auflage, San Jose, 2001, S. 18.

[10] Vgl. Lezgus, Andreas : „Neue Ansätze zur Gewinnung von Führungsinformationen", Seminararbeit, im Internet http://www.lezgus.com/lezgusftp/datawarehouse.zip, Stand April 1999, Abruf 20.03.02 S. 11.

[11] Vgl. Whitehorn, Mark; Whitehorn, Mary: „Business Intelligence: The IBM Solution: Datawarehousing and OLAP", London, 1999, S. 182.

Information) angewendet. FASMI-Ansatz besagt, dass die Antwortzeiten bei einfachen Abfragen 1 bis 2 Sekunden und bei komplexen Abfragen maximal 20 Sekunden betragen sollen, der Anwender muss seine Untersuchungen ad-hoc formulieren und intuitiv durchführen können. Ein System ist geeignet wenn es mehr Datenelemente bei stabilen Antwortzeiten analysiert.[13]

OLAP ermöglicht interaktiven Zugriff auf Daten zur analytischen multidimensionalen Datenauswertung. Der Zugriff in vertikaler Richtung ist entweder *Drill Down* oder *Drill Up*. In horizontaler Richtung gelten *Drill Across*, *Data Slicing*.

3.3 Einsatzmöglichkeiten des OLAP

Da OLAP in der Lage ist, komplexe Geschäftsanalysen vorzunehmen und damit ein unbeschränktes Datenvolumen „transparent" zu machen, ist es geeignet für sehr komplexe Anwendungen des Vertriebs- oder Finanzcontrollings sowie für Kunden- bzw. Interessenten-Analyse.

3.4 Grenzen des OLAP

Da OLAP ein Analyse- und kein Prognose-Tool ist, kann es keine zukunftbezogene Aussagen machen. Einpaar Beispiele typischer Fragen, die OLAP nicht beantworten kann, sind:

- Welche Kunden werden auf die nächste Marketing-Aktion reagieren?
- Welche Kunden werden wahrscheinlich demnächst den Vertrag kündigen?
- Wie kann ich meine profitabelsten Kunden charakterisieren?

4. Data Mining

"*Data Mining, the central activity in the process of KDD is concerned with finding patterns in Data*".[14]

Fayyad definiert Data Mining als einen mechanisierten Prozess, der nützliche Strukturen in Daten identifiziert oder entdeckt.[15]

---

[12] Eine Liste dieser Regeln befindet sich im Anhang – A.

[13] Vg. Chamoni, Peter: „Entwicklungslinien und Architekturkonzepte des On-Line Analytical Processing", in: Chamoni, Peter; Gluchowski (Hrsg.): *Analytische Informationssysteme : Data Warehouse, On-Line Analytical Processing, Data Mining*, Berlin, Heidelberg (u.a), 1998, S. 237.

[14] Dzeroski, Saso: „Data Mining in a Nutshell", in: Dzeroski, Saso; Lavrac, Nada, a.a.O. S. 3

[15] Vgl. Fayyad, Usama, Grinstein, George: „Introduction to Information Visualization in Data Mining und Knowledge Discovery in Databases", in: Fayyad, Usama, et al. (eds.): *Information Visualization in Data Mining and Knowledge Discovery in Databases*, 1. Auflage, San Jose, 2002.

Data Mining sucht nach bisher unbekannten Zusammenhängen und Mustern, die in riesigen Datenbanken versteckt sind. Dadurch ermöglicht es die Transformation von rohen Daten in nützliche Informationen.

4.1 Methoden des Data Mining

Es gibt verschiedene Methoden und Techniken, die in Data Mining verwendet werden. Die wichtigsten können aber 2 folgenden Bereichen zugeordnet werden:

- Techniken der künstlichen Intelligenz („*Artificial Intelligence*“), und
- Statistische Methoden

Die Techniken der künstlichen Intelligenz sind aus Platzgründen nicht Gegenstand dieser Arbeit. Im Folgenden werden wir uns mit statistischen Methoden beschäftigen. Betrachtet man die Verfahren hinsichtlich der anwendungsbezogenen Fragestellungen, so lassen sich die statistischen Methoden in primär ***strukturen-entdeckende*** und primär ***strukturen-prüfende*** Verfahren unterteilen. Wobei zu beachten ist, dass eine überschneidungsfreie, eindeutige Einordnung aller Verfahren nicht immer möglich ist, da die Zielsetzung der praktischen Fragestellungen dabei eine entscheidende Rolle spielt[16].

4.1.1. Strukturen-prüfende Verfahren

Das sind Verfahren, deren primäres Ziel ist es, Zusammenhänge zwischen Merkmalen (Variablen) zu überprüfen. Sie werden primär zur Durchführung von Kausalanalysen eingesetzt. Der Benutzer muss konkrete Vorstellungen über mögliche Zusammenhänge auf sachlogischer und theoretischer Ebene haben, d.h. es wird eine Hypothese bzgl. ihrer Richtigkeit überprüft. Die Aufgabe besteht darin, Hypothese anhand des zu Verfügung stehenden Datenmaterials zu verifizieren oder zu verwerfen. Es gibt verschiedene strukturen-prüfende Techniken, einige wichtige werden hier vorgestellt.

4.1.1.1. Regressionsanalyse

Die Regressionsanalyse wird verwendet, wenn Zusammenhänge von Variablen erkannt oder erklärt bzw. Werte einer abhängigen Variablen geschätzt bzw. prognostiziert werden sollen. Wenn es um mehrere abhängigen Variablen handelt, wird eine multiple Regressionsanalyse durchgeführt. Zwischen *der* bzw. *den* er-

[16] Vgl. Backhaus, Klaus et al.: „Multivariate Analysemethoden: eine anwendungsorientierte Einführung“, Berlin, Heidelberg u.a., 2000, S. XXI – XXVII.

klärenden und *der* zu erklärenden Variablen wird eine lineare oder nicht lineare Beziehung unterstellt.

Als Optimalitätskriterium dient i.d.R. die „Summe der quadrierten Residuen". Die Regressionsanalyse liefert Maßzahlen, die Aussagen über die Güte der Anpassung und die Bedeutung der abhängigen Variablen u.a. machen. Wenn die verwendeten Daten als „Stichprobe" betrachtet werden, dann kann man die vermutete Hypothese bewerten.

Die Schwierigkeit dieses Verfahrens liegt in Erkennung abhängiger und unabhängiger Variablen und der Auswahl geeigneter Variablen. Problematisch ist auch, dass zwischen den Variablen eine Beziehung nicht nur bestehen sondern auch von dem Benutzer vermutet werden muss. Dazu kommt, dass vor allem die lineare Regressionsanalyse Prämissen unterstellt, die oft nicht erfüllt sind.

Trotz dieser Probleme wird dieses Verfahren gegenüber den Verfahren der künstlichen Intelligenz oft vorgezogen – aufgrund seiner einfachen Algorithmen, Schnelligkeit und Transparenz. Die Regressionsanalyse ist „ein außerordentlich flexibles Verfahren, das sowohl für die Erklärung von Zusammenhängen wie auch für die Durchführung von Prognosen große Bedeutung besitzt"[17].

4.1.1.2. Diskriminanzanalyse

In der Diskriminanzanalyse wird untersucht, ob sich die Gruppen (z.B. Benutzer- oder Kundengruppen) hinsichtlich der Variablen signifikant unterscheiden und welche Variablen zur Unterscheidung zwischen den Gruppen geeignet sind. Ein typisches Anwendungsbeispiel ist die Kreditwürdigkeitsprüfung von Privatkunden, die anhand

a) soziodemographischer Merkmalen (Alter, Einkommen, Familienstand usw.),
b) Anzahl weiterer Kredite,
c) Beschäftigung und Beschäftigungsdauer etc.

in Klassen von hohem oder niedrigem Risiko einteilt.

Die Koeffizienten einer Linearkombination der Merkmalsvariablen werden durch die Diskriminanzanalyse so bestimmt, dass der Quotient aus der Streuung zwischen den Gruppen und der Streuung innerhalb der Gruppen maximiert wird.

[17] Backhaus, et al. a.a.O. S. XXII

Die Diskriminanzanalyse setzt eine Gruppenbildung voraus. Zudem muss die Gruppenzugehörigkeit durch eine nominal skalierte Variable ausgedrückt werden. Außerdem müssen die Merkmalsvariablen metrisch skaliert und die Daten für die Merkmalsvariablen von Elementen und deren Gruppenzugehörigkeit vorhanden sein. Die Anzahl der Merkmalsvariablen soll höchstens halb so groß wie der Umfang der Stichprobe und mindestens so groß wie die Anzahl der Gruppen sein[18].

4.1.1.3. Kontingenzanalyse

Kontingenzanalyse analysiert Beziehungen zwischen ausschließlich nominalen Variablen, d.h. es werden Zusammenhänge zwischen nominal skalierten Variablen aufgedeckt und untersucht. Ein Anwendungsbeispiel wäre die Untersuchung, ob der Besitz einer Kreditkarte im Zusammenhang mit dem Beschäftigungsverhältnis (Arbeiter, Angestellter, Beamter, Selbstständiger) steht.[19]

Es wird lediglich ein statistischer Zusammenhang zwischen Variablen gemessen. Aussagen über Kausalität können mit den durch diese Analyse gewonnen Informationen nicht getroffen werden.[20]

4.1.1.4. Entscheidungsbäume

Entscheidungsbäume finden die Attribute, die am besten zwischen Klassen unterscheiden und können als Menge von *„wenn ...; dann ...“* Regeln beschrieben werden. Sie konstruieren durch wiederholte Spezialisierung hierarchische Regelbäume. Zum Bilden solcher Bäume werden unabhängige Variablen im Hinblick auf ihre Trennschärfe sortiert und in einem Baum angeordnet. Die Knoten dieses Entscheidungsbaumes werden durch Attribute (z.B. Geschlecht), die Kanten durch deren Ausprägungen (z.B. Alter) und die Blätter durch ihre Klassenzugehörigkeit (z.B. Einkommen) gebildet.[21] Das Entscheidungsbaum-Verfahren ist nicht einer einzigen Klasse eindeutig einzuordnen, da es in verschiedenen Weisen zur Anwendung kommt.

---

[18] Vgl. Wohlschiess, Jan: „Einsatz von Data Warehouse und Data Mining zur Kundenanalyse und Kundenbewertung im Direktbanking“, in Swoboda, Uwe (Hrsg.): *Direct Banking: Wie virtuelle Institute das Bankgeschäft revolutionieren*, 1.Auflage, Wiesbaden, 2000 S. 126 ff.

[19] Vgl. Backhaus, et al. a.a.O. S. XXII.

[20] Vgl. Wohlschiess, Jan, a.a.O. S. 128.

[21] Vgl. Wittmann, Thomas: „Wissensentdeckung in Datenbanken mit adaptiven Regelsystemen“, Frankfurt am Main, 2000, S. 81 ff

4.1.2. Strukturen-entdeckende Verfahren

Die strukturen-entdeckenden Verfahren haben vor allem zum Ziel, ohne präzise Fragestellung verborgene und unbekannte Informationsstrukturen, strategische Erkenntnisse, Regeln und Zusammenhänge aufzudecken. Der Anwender hat zu Beginn keine Vorstellung von Zusammenhängen zwischen den Variablen oder Objekten.

4.1.2.1. Faktoranalyse

In diesem Verfahren wird eine größere Anzahl metrischer Variablen auf eine kleiner Anzahl von Einflussgrößen, sog. Faktoren, reduziert. Dieses Verfahren dient oft als eine Vorstufe zu den anderen Verfahren, um „Abhängigkeiten zwischen den unabhängigen Variablen auszuschalten“.[22]

Es wird als erstes aus den Korrelationen einzelner Variablen ein Maß berechnet, das erkennen lässt, ob Zusammenhänge zwischen Paaren von Variablen bestehen, und somit ob eine Bündlung von Variablen überhaupt sinnvoll ist.[23]

Ergibt sich eine Bündlung als sinnvoll, so wird in einem Faktorextraktionsprozess die Anzahl der Faktoren (Bündeln) bestimmt und in einem Faktorrotationsprozess die Variablen den einzelnen Faktoren zugewiesen. Danach werden Faktorenwerte z.B. durch regressionsanalytische Verfahren bestimmt.

Ein wichtiges Anwendungsbeispiel von der Faktoranalyse sind Positionierungsanalysen, die zeigen, wie Produktmarken, Politiker etc. durch Konsumenten, Wählern etc. (subjektiv) wahrgenommen und beurteilt werden.

Ein Vorteil der Faktoranalyse ist, dass der Anwender keine Vorkenntnis über die möglichen Beziehungen der Variablen unter einander haben muss. Die Voraussetzungen für die Anwendung der Faktoranalyse sind ein metrisches Skalenniveau und eine ausreichende Fallzahl, die mindestens so groß wie die Variablenzahl sein muss.

Zwar kann die u.U. schwierige Interpretation der gefundenen Faktoren die Anschaulichkeit der Ergebnisse verzerren, ein Einsatz der Faktoranalyse bleibt aber notwendig bei großer Anzahl stark korrelierender Variablen, insbesondere wenn für ein anderes Verfahren eine kleinere Anzahl von Einflussgrößen gebraucht wird, z.B. bei Anwendung der Regressionsanalyse.

---

[22] Wohlschiess, Jan, a.a.O. S. 128
[23] Vgl. Backhaus, et al., a.a.O. S. 262

4.1.2.2. Clusteranalyse

Im Gegensatz zu Faktoranalyse, die die Variablen bündelt, wird in der Clusteranalyse eine Bündlung von Objekten vorgenommen. D.h. es werden sog. Cluster (Gruppen) gebildet. Die Objekte in einer Gruppe sollen möglichst ähnlich (homogen), die Gruppen untereinander aber möglichst unähnlich (inhomogen) sein. Die Merkmalsvariablen können entweder von nominalem oder von metrischem Skalenniveau sein.

Typische Anwendungsbeispiele sind Bildung von Käufersegmenten und Bildung von Persönlichkeitstypen. Die Güte der Ergebnisse der Clusteranalyse kann/muss i.d.R. durch die Diskriminanzanalyse überprüft werden[24]. Somit ergänzen sich die beiden Verfahren.

Eine sinnvolle Interpretierbarkeit der Ergebnisse ist eine wichtige Voraussetzung für Anwendung der Clusteranalyse, d.h. ein Einsatz der Clusteranalyse ist bei einer sehr heterogenen Grundgesamtheit nicht empfehlenswert, da die so gebildeten Gruppen über kaum Elemente verfügen würden. Die Clusteranalyse ist ein wichtiges Data Mining Verfahren und wird häufig verwendet.[25]

4.1.2.3. CHAID (Chi-squared Automatic Detection)

Dieses Verfahren basiert auf die sog. χ^2-Tests. Es werden aus einer Stichprobe unter Beachtung des Zielkriteriums homogene Segmente gebildet. Es wird nach einer Variablen gesucht, die zur Trennung von *vorab* definierten Gruppen am Besten passt. Es werden diejenigen Ausprägungen dieser Variablen zusammengefasst, zwischen denen sich die Gruppen nicht stark unterscheiden. Danach werden die so gebildeten Segmente analog überprüft. Dieser Prozess wird iterativ solange fortgesetzt, bis vorab festgelegte Abbruchkriterien erreicht sind oder keine Variablen mehr vorhanden sind, die die Trennung verbessern können.

CHAID-Ergebnisse besitzen eine prognostische Relevanz und müssen nicht durch eine weitere Technik überprüft werden.[26] Bei der CHAID-Analyse kann es u.U. zu einem Informationsverlust kommen, da metrisch skalierte Variablen auf das (niedrigere) nominale oder ordinale Datenniveau transformiert werden. Was aller-

---

24 Der Unterschied zwischen den beiden Verfahren liegt darin, dass die Diskriminanzanalyse mit bereits definierten Gruppen arbeitet, während die Clusteranalyse diese erst bildet. . Somit ergänzen sich die beiden Verfahren.

25 Vgl. Wohlschiess, Jan, a.a.O. S. 130

26 Was z.B. bei der Clusteranalyse der Fall ist.

dings kein großes Problem darstellt, da in der Praxis ohnehin überwiegend nominale Variablen über die Merkmalsträger vorliegen.[27] Ein weiteres und komplizierteres Problem ist, dass die Zusammenhänge zwischen der abhängigen Variablen und den einzelnen unabhängigen Variablen isoliert untersucht werden. Das hat zur Folge, dass Zusammenhänge, die erst durch die Kombination zweier oder mehrer erklärender Variablen entstehen, unentdeckt bleiben.

In der Praxis ist CHAID wegen übersichtlicher Darstellung und leicht Interpretierbarer Ergebnisse vor allem im Bereich Database-Marketing weitverbreitet, was allerdings auch damit zusammenhängt, dass große Softwarehersteller z.B. SPSS und SAS CHAID in ihre Software in einer auch für Nicht-Statistiker verständlichen Form anbieten.

4.1.2.4. Assoziationsanalyse

Die Assoziationsanalyse beschäftigt sich mit Finden von Beziehungen zwischen Attributen. Zunächst werden alle Attributkombinationen gesucht, die mit einer zuvor festgelegten Häufigkeit auftreten.

Die Assoziationsanalysen werden häufig für Warenkorbanalyse verwendet. Warenkorbanalyse entdeckt Kombinationen von durch unterschiedliche Kunden gekauften Produkten. Durch Assoziierung dieser Daten kann festgestellt werden, welche Produkte i.d.R. zusammen gekauft werden.[28] Der Vorteil dieses Verfahrens ist, dass es mit vergleichsweise geringem Aufwand umfangreiche Datenbestände analysieren kann, und verständliche Ergebnisse liefert.

4.2 Einsatzmöglichkeiten des Data Mining

Typische Data Mining Anwender sind Banken, Kreditkarteninstitute, Versicherungen usw. Data Mining wird für effizientere Ausrichtung auf potentielle Kunden durch Entdeckung von Kundenwünschen und neuen Märkten, Verhindern von Kundenabwanderungen durch z.B. Identifizierung von „Cross-Selling"-Möglichkeiten, aber auch Verkleinern von Betrugsrisiken verwendet. Seinen Anwendungen gehören aber auch medizinische Diagnose z.B. durch automatisierte Diagnostik in der klinischen Tumordiagnostik, Genexpressionsanalyse und Analyse dynamischer Prozesse in lebenden Zellen.[29] Data Mining ermöglichte sogar

---

[27] Vgl. Wohlschiess, Jan, a.a.O. S. 131
[28] Vgl. IBM Redbook, a.a.O. S. 27
[29] Vgl. Information Angebot des Deutschen Krebsforschungszentrum, Heidelberg, Online im Internet, http://www.dkfz-heidelberg.de/i.pdf, Stand Feb 1999, Abruf 01.03.02

Entdeckung von „Agbami Oil Fields“ auf nigerischer Küste, die 1,45 Mrd. Barrel Erdöl enthalten und durch Satelliten verpasst wurden.[30]

4.2.1. Customer Relations Management (CRM)

Kunden sind besser informiert und ihre Erwartungen sind vor allem im Dienstleistungsbereich erheblich gestiegen. Ein Unternehmen muss wissen, welche Produkte welchen Kunden verkauft werden können und wie. CRM hat zum Ziel, Kunden langfristig an ein Unternehmen zu binden. Das bedeutet, die profitablen Kunden, ihr Verhalten und ihre Wünsche sollen systematisch identifiziert und entsprechende Geschäftsstrategien entwickelt werden. Durch Erstellung und Analyse von Kundenprofilen lassen sich diese Informationen entdecken.

4.2.2. Text-Mining

Text-Mining ist ein Instrument, um die wachsende Menge von Dokumenten effizient zu organisieren und nutzen. Text-Mining versucht aus den in der Form von unstrukturierten Dokumenten vorliegenden Daten Informationen auf maschinelle Art und Weise zu extrahieren. Motiv und Zielsetzung ist es, in Dokumentenbeständen automatisiert nach versteckten, interessanten Strukturen und Mustern zu suchen.

Text-Mining verwendet verschiedene Verfahren, z.B. Klassifikation, automatisches Zusammenfassen von Dokumenten, Abstraktion, Clustering und Methoden des Information Retrieval, z.B. Volltextsuche.

In einer Wissensgesellschaft ist das Wissensmanagement von hoher Bedeutung. Wissen als Ware hat einen wesentlichen Vorteil, da seine Ausbreitung mehrwertbehaftet ist und das Gesetz der abnehmenden Erträge in seiner klassischen Form nicht gilt.

Text-Mining has a very high commercial potential. A recent study indicated that 80% of a company's information was contained in text documents, such as emails, memos, customer correspondence, and reports[31]*. The ability to distil this untapped source of information provides substantial competitive advantages for a*

---

[30] Vgl. Port, Otis: „Virtual Prospecting“, Online im Internet, http://www.businessweek.com:/print/bw50/content/mar2001/bf20010323.htm, Stand. 23.03.01 Abruf 10.02.02 S. 2

[31] Vgl. Tan, Ah-Hwee: Text Mining: Promises And Challenges, Online im Internet, http://textmining.krdl.org.sg/docs/ TM_searcc99.pdf, Stand: 1999, Abruf: 05.04.02 S. 1 ff

company to succeed in the era of a knowledge-based economy. There are many possible applications of text mining technology. We briefly highlight a few below.

- ***Customer profile analysis**, e.g., mining incoming emails for customers' complaint and feedback.*
- ***Patent analysis**, e.g., analyzing patent databases for major technology players, trends and opportunities.*
- ***Information dissemination**, e.g., organizing and summarizing trade news and reports for personalized information services.*
- ***Company resource planning**, e.g., mining a company's reports and correspondences for activities, status, and problems reported.*

4.2.3. Web-Mining

Unter Web Mining werden die Analyse von WWW-Dokumenten (Web Content Mining) sowie WWW-Zugriffsdaten (Web Usage Mining) zusammengefasst. Letztere sind für Marketingfragestellungen im Bereich E-Commerce von besonderer Bedeutung. Der große Umfang der Server-Nutzungsdaten erfordert die Anwendung von Data Mining Verfahren.

Daten über Surfer bzw. Konsumenten werden hauptsächlich über Log Files, Cookies und sog. Forms (Anmeldeformulare) gesammelt. Dabei werden Daten über Domain, *time of access*, Keywords, verwendete Suchmaschinen, gerufene Seiten sowie Besuchshäufigkeiten etc. gesammelt und gespeichert.

The Web provides companies with an unprecedented opportunity to analyze customer behavior and preferences. Every visit to a Web site generates important consumer behavioral data, regardless of whether or not a sale is made. Every visitor action is a digital gesture exhibiting habits, preferences, and tendencies. These interactions reveal important trends and patterns that can help a company design a Web site that effectively communicates and markets its products and services. Companies can aggregate, enhance, and mine Web data to learn what sells, what works and what doesn't, and who is or isn't buying. The Web data generated with a single sale is of more value then the sale itself, because it can lead to a long and profitable relationship with that customer. The goal of marketers today is not to capture market share but to capture a share of a customer over a long period of time. Data mining in this context enables you to address such business questions as, "Who is buying what items and at what rates?"

You should also know what is selling so you can adjust your inventory and plan your orders and shipping. You need to know how to sell, what incentives, offers, and ads work, and how you should design your site to optimize your profits. Data mining algorithms can search for relationships in Web data to determine if patterns exist that can yield actionable business and marketing intelligence.

„Web data mining focuses on identifying customer attributes and consumer behaviour. The goals are generally to find out who is likely to buy your products and services and identify the features of your most loyal and profitable customers so that you can find more like them". [32]

Web-Mining ermöglicht den Unternehmen

- mit seinen Angebote die „richtigen" Nutzer gezielt zu erreichen oder
- • den Nutzern die „richtigen" Angebote gezielt zu unterbreiten.

In beiden Fällen ist damit zu klären, wer mit dem Webauftritt erreicht wird und wie dieser genutzt wird. Websites bieten die Möglichkeit, die Nutzung des Angebots genauer zu analysieren und unter Umständen nutzerspezifisch auszugestalten. Web-Mining-Tools sollen Antworten auf folgende Fragestellungen liefern (Auswahl):

- Wer besucht die Website (Herkunft: Land, Provider, Unternehmen etc.)?
- Wann wird die Website am häufigsten besucht?
- Wie oft wird auf welche Seite zugegriffen?
- Welche Seiten werden von welchen Besuchern (Kunden, Konkurrenten) abgerufen?

4.3 Grenzen des Data Mining

Die Qualität der Ergebnisse hängt von der Qualität der gelieferten Daten ab. Damit erst verstärkt - und nicht etwa reduziert - der Einsatz von Data Mining die Anforderung an Datenqualität. Der Anwender muss exakte Kenntnis der zugrundeliegenden Daten haben, und ein tiefes Verständnis des Business ist auch unentbehrlich. Data Mining kann den Entscheidungsträger unterstützen aber nicht er-

[32] Vgl. Mena, Jesus: „Mining E-Customer Behavior", Online im Internet, http://www.db2mag.com/db_area/archives/1999/q4/mena.shtml, Stand: Winter 1999, Abruf: 04.04.02.

setzen: „*There can be many blunders that one might committ, working with Data Mining.*“[33]

5. Vergleich zwischen OLAP und Data Mining

OLAP-Tools sind grundsätzlich Analyse-Tools. Ein Beispiel für eine OLAP-Anfrage könnte sein: *Hängt die Kündigung eines Vertrages vom Alter, Einkommen und Familienstand des Kunden ab?* Data Mining Tools suchen nach bisher unbekannten Zusammenhängen und können somit Informationen entdecken, nach denen man nicht einmal bewusst gesucht hat. Ein Beispiel für eine Data Mining Anfrage könnte sein: ***welche** Eigenschaften eines Kunden lassen auf eine bevorstehende Vertragskündigung schließen?*

OLAP Vs. Data Mining	
Interaktiv	Möglichst automatisierte Suche
Benutzer muss selber wissen, wonach er sucht	Benutzer muss nicht von Anfang an wissen, wonach er sucht
Nur Aggregatikon	Sucht versteckte Muster
Hilfswerkzeug zur Analyseunterstützung	Kompliziertere Analyse-Tools
Verifikationsmodell	Verifikationsmodell

Tabelle 1, OLAP vs. Data Mining

OLAP-Tools werden von manchen Experten als „workhorses of the analysis tool family“ und DM-Tools als „race horses“ [34] beschrieben.

Im Endeffekt ist entscheidend, ob man seine Kunden „nur“ behalten will, neue Kunden gewinnen will oder durch mehr Service die Kundenzufriedenheit steigern will. Wobei eine saubere Trennung nicht möglich ist, weil das eine das andere nicht ausschließt sondern ergänzt.

---

[33] Vgl. Skalak, David: „Data Mining Blunders Exposed“, Online im Internet: http://www.db2mag.com/db_area/archives/2001/q2/miner.shtml Stand Summer 2001, Abruf 04.04.02.

[34] Vgl. Griffin, Jane: „OLAP vs. Data Mining: Which One Is Right For Your Data Warehouse?“, Online im Internet: http://www.datawarehouse.com/iknowledge/articles/print.cfm?ContentId=281, Stand 22.08.00 Abruf 06.04.02.

„OLAP and Data Mining can use the same data, the same concepts, the same metadata and also the same tools, perform in synergy, and benefit from each other by integrating their results in the data warehouse.“[35]

6. Schlussbetrachtung

OLAP, Data Warehousing und Data Mining sind aus der Notwendigkeiten entstanden, dezentral gehaltene große Datenmengen effektiv zu verwalten und wirtschaftlich profitabel auszuschöpfen, indem aus den Informationenmengen wissenschaftlich wichtige, wirtschaftlich profitable und bisher unbekannte Informationen gewonnen werden.

Business Intelligence, insbesondere durch OLAP-Technologie und Data Mining, bietet Unternehmen eine Möglichkeit, Wissensmanagement und Reaktionsgeschwindigkeit zur Kernfähigkeiten in einer globalisierten Wirtschaft und einem E-Business-Environment zu machen und für Wettbewerbsvorteile einzusetzen. Die mathematisch-statistischen Methoden des Data Mining sind für diese Aufgaben besonders geeignet, da die Ergebnisse statistisch-empirisch nachprüfbar sind.

Es wird gar gesagt: „In wenigen Jahren wird es nur noch zwei Arten von Firmen geben: solche, die heute auf die Web-Technologie und eine integrierte Geschäftsdatenbank setzen und solche, die einfach bankrott sind.“[36] Diese Aussage mag übertrieben sein, aber sie deutet durchaus auf einen Trend hin, wie wichtig die Gewinnung von Informationen aus Info-Bergen geworden ist.

---

[35] IBM Redbook: „Mining Your Own Business in Banking / Using DB2 Intelligent Miner for Data“, 1. Auflage, San Jose, 2001, S. 22.

[36] McFarlan, Professor an der Harvard Business School, Wirtschaftswoche Nr. 24/1998, S. 105; zitiert in: Lezgus, Andreas, a.a.O. S. 24.

LITERATURVERZEICHNIS

Backhaus, Klaus / Erichson, Bernd (et al.): „Multivariate Analysemethoden: Eine anwendungsorientierte Einführung“, 9. Auflage, Berlin, Heidelberg (u.a.) 2000.

Chamoni, Peter / Gluchowski, Peter (Hrsg.): „Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining“, Berlin, Heidelberg (u.a.), 1998.

Ester, Martin / Sander, Jörg: „Knowledge Discovery in Databases: Techniken und Anwendungen“, Berlin, Heidelberg u.a. 2000.

Fayyad, Usama / Grinstein, Georges / Wierse, Andreas(eds.): „Information Visualization in Data Mining and Knowledge Discovery“, San Francisco, London (u.a.) 2002.

Martin, Wolfgang (Hrsg.): „Data Warehousing: Data Mining – OLAP“, 1. Auflage, Bonn, Albany (u.a.), 1998.

Mertens, Peter / Bodendorf, Freimut et al.: „Grundzüge der Wirtschaftsinformatik“, 7. Auflage, Berlin, Heidelberg etc. 2001.

Whitehorn, Mike / Whitehorn, Mary: „Business Intelligence: The IBM Solution: Data Warehousing & OLAP“, London, 1999.

Wittmann, Thomas: „Wissensentdeckung in Datenbanken mit adaptiven Regelsystemen : Entwicklung eines Data Mining Methodenbaukastens auf Basis von Neuro-Fuzzy-Systemen“, Frankfurt am Main, 2000.

Anmerkung: Internetquellen sind direkt auf der Stelle angegeben, an der sie verwendet werden.

ANHANG – A

OLAP-REGELN VON TED CODD

Basiseigenschaften

1. Multidimensionale Sichtweise
2. Intuitive Datenmanipulation
3. Variable Zugriffsmöglichkeit
4. Zwischenspeicherung von Daten sowie Zugriff auf Basisdaten
5. Vier OLAP-Analysemodelle
6. Client/Server-Architektur
7. Benutzer-Transparenz
8. Mehrbenutzerunterstützung auch bei „concurrent write“

Spezielle Eigenschaften

9. Integration unnormalisierter Daten
10. Getrennte Speicherung von Ergebnissen und Basisdaten
11. Vorhandensein von Nullwerte
12. Nullwerte werden bei der Analyse übergangen

Reports

13. Flexible Report-Generierung
14. Stabile Antwortzeiten
15. Automatische Anpassung der physischen Speicherung

Dimensionskontrolle

16. Keine Einschränkung der Multidimensionalität
17. Unbeschränkte Anzahl von Dimensionen (performance-abhängig)
18. Uneingeschränkte Operationen über Dimensionen hinweg

(Erschienen im „Unternehmensberater“, Ausgabe 1/2001, S. 56)