Seminar

Business Intelligence II
Data Mining & Knowledge Discovery

Was ist Data Mining?

23. Januar 2004

Sabine Queckbörner
s_queckb@informatik.uni-kl.de

Betreuer: J.Göres

Inhalt

1 Einleitung

Durch den schnellen Fortschritt in der Informationstechnologie ist es möglich geworden, verschiedene Arten von Informationen in Datenbanken und Data-Warehouses zu speichern. Die zunehmende Automatisierung von Geschäftsprozessen, sowie das automatische Erfassen und Verarbeiten einer Vielzahl von alltäglichen Vorgängen wie Telefongespräche, Kreditkartentransaktionen und Einkäufe in Supermärkten durch Scannerkassen führt zu immer größeren Datenbeständen. Durch Fortschritte in der Speichertechnologie, billigere Speichermedien und vor allem leistungsfähigere Datenbanksysteme können auch immer größere Datenmenge über längere Zeiträume gespeichert werden.
War es früher noch möglich, Daten „manuell“ zu analysieren, da die Datenmenge überschaubar war, so ist dies heute durch die ständig anwachsende Menge und Vielfalt der Daten zu einer für einen Menschen sehr zeitraubenden und kaum bezahlbaren Angelegenheit geworden. Es sind immer effizientere und schnellere Analyseverfahren notwendig, um aus dieser riesigen und ständig wachsenden Informationsmenge einen Nutzen zu ziehen, zum Beispiel in Form eines Wettbewerbsvorteils für ein Unternehmen. Solche Verfahren werden zum Beispiel im Data-Mining und im *Knowledge Discovery in Databases* (KDD) umgesetzt. Ohne Data-Mining bleibt Wissen ungenutzt, das aus großen Datenmengen extrahiert werden könnte. Datenbanken und Data-Warehouses werden dann nur unvollständig eingesetzt.
Im Folgenden der Begriff Data-Mining präzisiert, indem es mit seinen Funktionen und Anwendungen beschrieben wird. Außerdem wird Data-Mining als Teilschritt im KDD-Prozess betrachtet und dem Online Analytical Processing (OLAP) gegenübergestellt.

Diese Ausarbeitung wurde im Rahmen des Seminars Business Intelligenz – Data-Mining und Knowledge Discovery verfasst, einer weiterführenden Veranstaltung des Seminars Business Intelligenz – OLAP und Data-Warehousing.

2 Data-Mining

2.1 Motivation

Data-Mining kann als Ergebnis des Fortschritts in der Informationsverarbeitung angesehen werden. Die Entwicklung von Datenbanktechnologien ermöglichte ein komfortables und strukturiertes Ablegen von Daten und vereinfachte und beschleunigte die Suche nach ihnen. Mit steigender Leistungsfähigkeit der Datenbankmanagementsysteme wuchsen die Datenmengen jedoch explosionsartig an, so dass die darin enthaltenen Datensätze sowie ihre Zusammenhänge für den Menschen nicht mehr überschaubar und auswertbar waren [HaKa01]. Dies führte zu einem verstärkten Nachdenken über Verfahren zur automatisierten Wissensfindung.

Um eine Wissensgewinnung aus einer Menge von Daten möglichst flexibel zu halten, sollte eine Analyse unabhängig von der Art der Datenorganisation (Datenbank, Data-Warehouse, WWW, etc.) durchführbar sein [HaKa01].

Dazu müssen die ausgewählten Daten zunächst aufbereitet werden. Zur Aufbereitung werden Verfahren eingesetzt, die zum Beispiel fehlende Attributwerte in Datensätzen ergänzen oder stark abweichende Werte entfernen. Anschließend werden die Daten so transformiert, dass die anstehende Analyseaufgabe möglichst gut unterstützt wird. Darunter fällt zum Beispiel die Reduktion des Datenbestandes auf eine für diese Aufgabe relevante Teilmenge. Nach dieser Datenvorbereitung erfolgt die Auswahl der Analysefunktion, mit deren Hilfe die Daten auf Muster untersucht werden. Die tatsächliche Durchführung der Analyse bezeichnet man dann als Data-Mining. Anschließend werden die gefundenen Muster interpretiert, was zu einer erneuten Analyse oder zu einer Darstellung der um irrelevante sowie redundante Elemente reduzierten Ergebnismuster führen kann [FaPiSm96a].

Diesen Prozess der Wissensgewinnung aus gesammelten Daten bezeichnet man auch als *Knowledge Discovery in Databases* oder kurz als KDD. Data-Mining ist also eigentlich nur ein Schritt des ganzen Prozesses, Wissen aus den Daten zu extrahieren, obwohl es begrifflich in der Industrie, in den Medien und auch im Datenbankforschungsbereich mit dem ganzen Prozess der Wissensgewinnung aus Daten (KDD) gleichgesetzt wird (vgl. [HaKa01]).

2.2 Begriffsdefinition

Die Bezeichnung *Data-Mining* stammt ursprünglich aus dem Bereich der Statistik und kennzeichnet dort die selektive Methodenanwendung zur Bestätigung vorformulierter Hypothesen [GrBe99]. Noch heute beruhen daher zahlreiche Data-Mining-Methoden auf statistischen Verfahren [FaPiSm96].

Im informationstechnischen Kontext existieren mehrere unterschiedliche Definitionsvarianten. So definieren Berry und Linoff Data-Mining als Erforschung und Analyse großer Datenmengen mit automatischen oder halbautomatischen Werkzeugen, um bedeutungsvolle Muster und Regeln aufzufinden [BeLi97].

Decker und Focardy beschreiben Data-Mining als eine Methodik zur Problemlösung um logische oder mathematische, zum Teil komplexe Beschreibungen von Mustern und Regelmäßigkeiten in Datensätzen zu entdecken [DeFo95].
Eine weitere allgemeine Definition liefern Fayyad, Piatetsky-Shapiro und Smyth, in der Data-Mining als Teilschritt des KDD-Prozesses angesehen wird, der aus der Anwendung von Datenanalysealgorithmen besteht und zu einer Auflistung von Mustern, die aus den Daten gewonnen wurden, führt [FaPiSm96].

Aus den vorgestellten Definitionen kann Data-Mining zusammenfassend also als die Anwendung von Algorithmen auf Daten mit der Zielsetzung, Muster aus den Daten zu extrahieren, verstanden werden.

2.3 Nach welchen Mustern wird gesucht?

Beim Data-Mining wird in einer ausgewählten Datenmenge nach Mustern gesucht. Diese Muster sind Ausdrücke, die eine Teilmenge dieser Daten beschreiben [FaPiSm96] und das zu extrahierende oder bereits gewonnene Wissen repräsentieren [Pet97]. Man unterscheidet zwischen Regeln und Abhängigkeiten, Gruppen (Cluster), Verbindungsmuster (Link), zeitliche Muster (Sequence), Abweichungen, Formeln und Gesetzmäßigkeiten.
Nach welcher Art von Mustern gesucht wird, hängt von der vorliegenden Analyseaufgabe ab. Dazu unterscheidet man hauptsächlich zwischen beschreibender und vorhersagender Analyse. Bei der beschreibenden Analyse versucht man generelle Auffälligkeiten der vorhandenen Daten zu erfassen, während bei der vorhersagenden Analyse Trends aus der gegebenen Datenmenge abgeleitet werden sollen [HaKa01]. Im Fall der beschreibenden Analyse würde man also zum Beispiel nach Gruppen suchen, während man bei der vorhersagenden Analyse eher an zeitlichen Mustern, Regeln und Abhängigkeiten sowie Formeln und Gesetzmäßigkeiten interessiert wäre.
Eine konkrete Analyseaufgabe ist die Klassifikation. Hierbei werden Daten in eine oder mehrere vordefinierte Kategorien oder Gruppen eingeteilt. Eine Aufgabenstellung, die diese Funktion verdeutlicht, wäre zum Beispiel die Fragestellung, ob ein Kunde einen Kredit zurückzahlen wird. Im Gegensatz dazu wird eine Einteilung der Daten in Gruppen, die nicht vorher bekannt beziehungsweise vordefiniert sind, sondern aus den Daten abgeleitet werden, durch ein sogenanntes *Clustering* erreicht. Die Muster, die bei der Klassifikation gesucht werden, können unter anderem Gruppen sein, in denen ähnliche Objekte in eben diesen Klassen zusammengefasst werden. Die durch Clustering gefundenen Muster sind neben den gefundenen Gruppen auch die Regeln und Abhängigkeiten, welche die Gruppen beschreiben.
Bei der Abhängigkeitsanalyse wird nach Abhängigkeiten zwischen Attributen gesucht. Die Muster, nach denen gesucht wird, sind folglich Regeln und Abhängigkeiten, also Zusammenhänge zwischen verschiedenen Attributen eines Objektes. Man unterscheidet zwischen einer strukturellen und einer quantitativen Ebene der Abhängigkeiten. Während man auf der strukturellen Ebene untersucht, *welche*

Attribute zusammenhängen, interessiert man sich auf der quantitativen Ebene eher dafür, *wie stark* diese Zusammenhänge sind. Eine typische Fragestellung, die mit Hilfe der Abhängigkeitsanalyse untersucht werden kann, ist: „Welche Produkte werden zusammen gekauft?“

Die Verbindungsanalyse ist eine Aufgabenstellung, bei der man nach Verbindungsmustern in den vorbereiteten Daten suchen möchte. Diese Muster beschreiben Verknüpfungen und Regelmäßigkeiten zwischen verschiedenen Objekten. Bei dieser Analyse werden Beziehungen zwischen Attributen ermittelt, wobei der Schwerpunkt hier im Gegensatz zur Abhängigkeitsanalyse auf Korrelationen zwischen *mehreren* Attributen liegt.

Eine weitere Aufgabenstellung ist die Sequenzanalyse, die der eigentlichen Aufgabe, der Erstellung von Prognosen zugrunde liegt. Hierbei werden zeitliche Abfolgen erfasst und auf Abhängigkeiten untersucht. Die gesuchten Muster sind also zeitliche Muster, sogenannte Sequenzen, die häufig wiederkehrende Abfolgen in den Daten beschreiben. Eine denkbare Fragestellung einer solchen Analyse wäre zum Beispiel: „Wie entwickelt sich der Dollarkurs?“ [FaPiSm96a].

Dies ist nur eine Auswahl der möglichen Aufgabenstellungen. Sie können mit Hilfe von neuronalen Netzen, Kohonen-Netzen, klassischen statistischen Verfahren, Verfahren des maschinellen Lernens oder genetischen Algorithmen umgesetzt werden [HaKa01, Das03]. Zum Beispiel werden neuronale Netze, lineare Regression und CHAID häufig bei Fragestellungen mit Prognosecharakter verwendet. Kohonen-Netze und regelbasierte Systeme werden hingegen oftmals beim Clustering eingesetzt, eine eindeutige Zuordnung dieser Verfahren zu den Aufgabenstellungen gibt es jedoch nicht, da mehrere Data-Mining-Technologien zur Lösung einer Aufgabe angewandt werden können.

Abschließend sollte gesagt werden, dass weder alle relevanten Muster durch Data-Mining-Verfahren gefunden werden können, noch alle gefundenen Muster wichtig sind. Ob ein Muster für einen Benutzer interessant ist, hängt davon ab, ob das Muster von ihm verstanden wird, ob es für neue Daten auch in einem gewissen Grade zutrifft, ob es potentiell nutzbar und vor allem neu ist. Ein Muster ist auch dann wichtig, wenn es eine vorher aufgestellte Hypothese bestätigt, die der Benutzer überprüfen wollte [HaKa01].

Statistische Maße für die Relevanz von Mustern sind *Support* und *Confidence*. Während Support ein Maß für den Anteil der Datensätze ist, welche die Regel erfüllen, beschreibt Confidence die Wahrscheinlichkeit, dass eine Regel X=>Y auf einen Datensatz, der X erfüllt zutrifft. Untersucht man zum Beispiel eine Assoziationsregel, die beschreibt, wie der Kauf von Elektrogeräten vom Wohnort München abhängig ist (also Elektrogeräte => München), so wäre Support der Anteil aus der Gesamtmenge aller Transaktionen, bei denen Käufer aus München beteiligt sind und Elektrogeräte gekauft werden. Confidence wäre hierbei die Wahrscheinlichkeit, dass der Kunde aus München kommt, wenn ein Elektrogerät verkauft wird. Man kann für Support und Confidence bestimmte Grenzen festlegen, und solche Regeln oder Muster verwerfen, welche diese Werte nicht erreichen.

2.4 Wie wird gesucht?

2.4.1 Möglichkeiten

Es gibt verschiedene Möglichkeiten, nach Mustern in den vorhandenen Daten zu suchen. Man kann beispielsweise während einer Analyse nach mehreren Mustern parallel suchen. Dies ist insbesondere dann sinnvoll, wenn noch keine Vorstellungen über die Arten der anzutreffenden Muster vorhanden sind [HaKa01].
Außerdem ist es möglich, in verschiedenen Abstraktionsebenen nach Auffälligkeiten zu forschen. So kann ein bundesweit tätiger Händler, die Absatzdaten seiner Filialen so analysieren, dass er den Umsatztrend des Bundesgebiets mit den einzelnen Trends der Bundesländer vergleicht. Dadurch kann ermittelt werden, in welchen Gebieten der Umsatz verhältnismäßig zurückgeht. In diesen Regionen kann er dann gezielt Werbung für sein Unternehmen betreiben. Dabei sind Bundesgebiet und Bundesländer einzelne Abstraktionsebenen [HaKa01]. Diese Analyse kann dann selbstverständlich auch noch weiter bis zur Abstraktionsebene der einzelnen Städte fortgesetzt werden.
Es ist weiterhin möglich, dass nach verschiedenen, eventuell vom Benutzer vorgegebenen Schwerpunkten gesucht wird. Hier kann es sein, dass der Benutzer eine Hypothese bezüglich der Daten aufgestellt hat und diese verifizieren möchte [HaKa01].

2.4.2 Verfahren

Die klassischen Data-Mining-Verfahren sind statistische Verfahren, da ehemals speziell ausgebildete Statistiker damit beauftragt wurden, mittels Formeln und einfacher Software die gesammelten Datenmengen zu analysieren. Erst mit der Entwicklung leistungsfähigerer Computer entstanden Data-Mining-Technologien, die auf künstlicher Intelligenz basieren. Im Folgenden sind einige gängige Data-Mining-Verfahren angeführt [Das03].

- Künstliche neuronale Netze

  Künstliche neuronale Netze sind lineare Prognoseverfahren, die der biologischen Informationsverarbeitung nachempfunden wurden und in der Lage sind, selbständig zu lernen.
- Kohonen-Netze

  Kohonen-Netze bilden die Grundlagen für ein Segmentierungsverfahren, das auf den Prinzipien neuronaler Netze basiert und selbständig Gruppen innerhalb eines Datensatzes bildet.
- Lineare Regression

  Die lineare Regression ist ein klassisches Prognoseverfahren mit unabhängigen Variablen zur Erklärung von Verhaltensweisen.
- Genetische Algorithmen

  Genetische Algorithmen basieren auf den Grundlagen der biologischen Evolution. Sie suchen innerhalb eines Lösungsraumes nach einer optimalen Lösung.

- CHAID

  Chi-squared Automatic Interaction Detection ist eine Methode, die eine Menge von Datensätzen nach einer abhängigen Variable in Gruppen einteilt.
- Regelbasierte Systeme

  Regelbasierte Systeme sind Methoden, die zum Herausfiltern und Ausfindigmachen von „Wenn-Dann“-Regeln dienen.

Welche dieser Methoden letztendlich zur Analyse ausgewählt wird, hängt von der Aufgabenstellung und dem gewünschten Ergebnis (welche Arten von Mustern gefunden werden sollen, warum analysiert wird) ab. Auch werden mehrere Lösungen für die selbe Aufgabenstellung entwickelt und getestet, um bessere Ergebnisse zu erzielen. Um ein bestimmtes Ergebnis oder Ziel zu erreichen, können mehrere Verfahren (auch innerhalb einer Data-Mining-Lösung) kombiniert werden [HaKa01, Das03].

3 Data-Mining und KDD

3.1 Begriffsdefinition KDD

Der Begriff des *Knowledge Discovery in Databases* (KDD) wird in der Literatur relativ einheitlich beschrieben. Gemeint ist damit ein mehrere Stufen umfassender Prozess, in dem Wissen aus gesammelten Daten gelernt beziehungsweise extrahiert wird. Fayyad, Piatetsky-Shapiro und Smyth liefern folgende, oft zitierte Definition: Der KDD-Prozess ist ein nichttrivialer Prozess zur Identifikation gültiger, neuartiger, potentiell nützlicher und verständlicher Muster in Daten (übersetzt aus [FaPiSm96a]).
KDD wird hier als Prozess bezeichnet, da die Wissensgewinnung aus den Daten in vielen Schritten von der Datenauswahl über deren Analyse bis hin zu ihrer Auswertung abläuft, die in mehrfachen Wiederholungen durchlaufen werden können.
Dieser Prozess wird deshalb als nichttrivial bezeichnet, da er mehr können soll, als lediglich die gegebenen Daten zusammenzufassen, nämlich Beziehungsmuster, Regeln und Abhängigkeiten aufzeigen [FaPiSm96, FaPiSm96a].
Mit Gültigkeit der Muster ist gemeint, dass gefundene Beziehungen und Abhängigkeiten in den gegebenen Daten auch mit einer gewissen Sicherheit in neuen Daten zu finden sein werden. Es sollte sich bei den gefundenen Mustern also nicht um „zufällige“ Auffälligkeiten handeln. Die gefundenen Muster beziehungsweise die daraus gewonnenen Erkenntnisse sollten außerdem für das System und den Benutzer unbekannt, zusätzlich aber auch für die Aufgabenstellung (zum Beispiel Prognose) verwertbar, also nützlich sein. Letztendlich sollten die Auffälligkeiten, Regeln und Beziehungen, die als Muster aus den Daten gewonnen wurden, verständlich sein. Dies muss nicht nach dem ersten Durchgang der Fall sein, sondern kann auch erst nach einer oder mehreren Wiederholungen zutreffen [FaPiSm96, FaPiSm96a].
Der detailliertere Ablauf der Wissensgewinnung soll im nächsten Kapitel in Form einer Übersicht der einzelnen Stufen des KDD-Prozesses erläutert werden.

3.2 Der KDD-Prozess

Der Prozess des *Knowledge Discovery in Databases* (KDD-Prozess) umfasst wie bereits erläutert, das ganze Verfahren der Wissensgewinnung oder Wissensextraktion aus Daten, von der Auswahl der Datenmenge über die Vorbereitung und Analyse bis hin zur Auswertung und Interpretation der Daten.

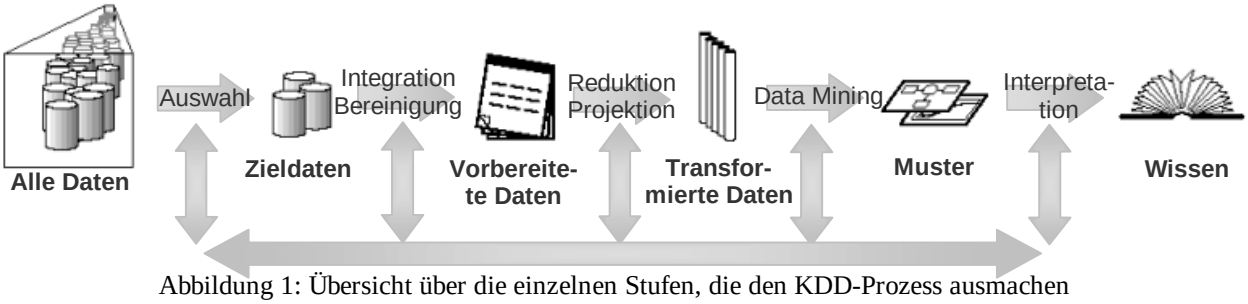


Abbildung 1: Übersicht über die einzelnen Stufen, die den KDD-Prozess ausmachen (angelehnt an [FaPiSm96a])

Vor diesem Prozess steht das Festlegen und Verstehen der Aufgabenstellung. Bevor mit der Vorbereitung der Daten und der Analyse begonnen wird, sollte bestimmt werden, welche Ziele erreicht werden sollen und welche Ergebnisse erwartet werden können [FaPiSm96].

Zu Beginn des KDD-Prozesses wird die Analyse im Hinblick auf die Aufgabenstellung vorbereitet. Dies beinhaltet eine *Selektion* der Daten, sowie im Rahmen des von Piatetsky-Shapiro und Smyth als *Preprocessing* bezeichneten Vorgangs die Integration und Bereinigung der Daten, eine anschließende *Transformation* beinhaltet eine Reduktion und Projektion der Daten.

Zunächst werden die gesammelten Daten und Informationen so vorbereitet, dass eine Art Zieldatensatz ausgewählt wird, der aus einer Teilmenge aller zur Verfügung stehenden Daten besteht (Selektion, [FaPiSm96]). Kommen die Daten aus verschiedenen Quellen, zum Beispiel aus verschiedenen Datenbanksystemen oder Anwendungen, ist nun eine Datenintegration sinnvoll, während der die Datensätze aneinander angepasst werden um Kompatibilitätsprobleme (zum Beispiel durch unterschiedliche Datenformate) zu beheben [Pet97]. Danach findet eine Bereinigung der ausgewählten Daten (Data Cleaning, [FaPiSm96]) statt, bei der inkonsistente, redundante und fehlerhafte Datensätze korrigiert oder aus der Analysemenge entfernt werden. Als nächstes findet eine Transformation der Daten durch eine Datenreduktion und Datenprojektion statt. Hierbei werden Funktionen angewandt, welche die Daten im Hinblick auf das Analyseziel geeignet darstellen. Eine Mengenreduktion kann zum Beispiel unter Berücksichtigung der Analyseaufgabe und nach Abwägung die effektive Anzahl der Variablen reduzieren [FaPiSm96], sodass die Analyse anschließend wesentlich schneller durchzuführen ist. In diesem Transformationsschritt kann auch die Kodierung der einzelnen Attribute verändert werden, wenn sich aus den Ergebnissen früherer Analyseschritte oder aufgrund der Anforderungen des verwendeten Analyseprogramms die Notwendigkeit dafür ergibt [Pet97].

Nun werden im nächsten Schritt die Ziele des gesamten KDD-Prozesses auf die Analyseziele übertragen und angepasst. Analysefunktionen, wie zum Beispiel Klassifikation oder Clustering und Analysealgorithmen werden ausgewählt, die in der sich nun anschließenden Analyse- oder Data-Mining-Phase verwendet werden [FaPiSm96, FaPiSm96a].
Im Anschluss daran findet die eigentliche, bereits erwähnte Analyse beziehungsweise das Data-Mining statt. Nun wird in der vorbereiteten Datenbasis anhand der Analysefunktionen und Algorithmen nach interessanten Mustern, den gesuchten Regeln und Beziehungen gesucht. Dabei kann der Benutzer maßgeblich die Data-Mining-Funktionen unterstützen, in dem er die vorangehenden Schritte richtig und gewissenhaft durchführt [FaPiSm96].
Nach der Analyse können die gefundenen Muster in den Daten dargestellt und vom Benutzer interpretiert werden. Alternativ dazu kann der Prozess erneut mit den gefundenen Mustern als Datenbasis durchlaufen werden. Da die gefundenen Muster noch kein explizites Wissen darstellen, kann der Benutzer abschließend aus den dargestellten und interpretierten Beziehungen und Regeln der untersuchten Daten Wissen gewinnen und zum Beispiel in Prognosen anwenden und weiterverarbeiten, oder in Form von Informationen erneut in der Datenbank festhalten [FaPiSm96].

Der KDD-Prozess wird oft als ein interaktives und iteratives Verfahren bezeichnet. Interaktiv ist er dadurch, dass in den einzelnen Schritten Entscheidungen vom Benutzer getroffen werden müssen, wie zum Beispiel bei der Auswahl der Daten oder der Analysefunktionen. Die Iterativität zeigt sich darin, dass sich die Stufen des KDD-Prozesses immer wieder durchlaufen lassen, in dem man zum Beispiel die im ersten Durchgang gefundenen Muster als Datenbasis für einen zweiten Durchlauf auswählt [FaPiSm96].

3.3 Gegenüberstellung Data-Mining und KDD

Wie eingangs bereits erwähnt, werden die Begriffe Data-Mining und KDD (Knowledge Discovery in Databases) in der Praxis oft synonym verwendet.

Eine begriffliche Abgrenzung des Data-Mining von KDD in der Praxis begründet sich in den verschiedenen Verwendungs- und Entwicklungsbereichen. So wird der Begriff *Data-Mining* häufiger von Statistikern, Datenanalytikern und im Rahmen von Management-Informationssystemen (Management Information Systems, MIS) verwendet. Der Begriff des *Knowledge Discovery in Databases* wurde während des ersten KDD-Workshops 1989 geprägt, um zu betonen, dass Wissen das Endprodukt eines Auswertungsprozesses von Daten ist. Der Begriff Data-Mining wird also mehr im Zusammenhang mit Datenbanken genannt, während KDD mehr der künstlichen Intelligenz sowie dem maschinellen Lernen zugeordnet wird [FaPiSm96].

In den meisten Dokumenten wird zwischen den beiden Begriffen allerdings so unterschieden, dass Data-Mining als ein Teilschritt des KDD-Prozesses angesehen wird. Dieser Prozess beginnt wie oben beschrieben, nach der Planung mit der Auswahl der Daten (*Selektion*) und wird mit der Aufbereitung (*Preprocessing*) und Transformation der Daten fortgesetzt. Anschließend findet die Analyse der nun vorbereiteten Daten statt, deren Ergebnisse am Ende interpretiert und ausgewertet werden. Die erwähnte Analyse der Daten ist der Teilschritt, der das Data-Mining verkörpert (vergleiche [FaPiSm96, FaPiSm96a, Liu02, HaKa01]).

4 Data-Mining und OLAP

4.1 Begriffsbestimmung OLAP

Der Begriff des *Online Analytical Processing* (OLAP) wurde 1993 von E.F. Codd eingeführt. Dieser beschreibt OLAP als eine dynamische Analyse, die erforderlich ist, um Informationen aus erklärenden, anschaulichen und formelhaften Analysemodellen zu erzeugen, zu manipulieren, darzustellen und aufzubauen (Übersetzt aus [CoCoSa93]). Codd kam zu dem Schluss, dass die bisher verwendeten Analysemethoden und relationalen Datenbanken für die immer schneller anwachsenden Mengen von Daten nicht mehr geeignet waren, da schon relativ einfache SQL-Anfragen die Systeme an die Grenzen ihrer Leistungsfähigkeit bringen. Außerdem stellte er fest, dass operationale Daten nicht ausreichen, um die immer anspruchsvoller werdenden betriebswirtschaftlichen Fragen zu beantworten. Es sollte also eine Analysemethode entwickelt werden, die einem Manager eine schnelle Auswertung gesammelter betriebswirtschaftlicher Daten (über mehrere Abstraktionsebenen) ermöglicht, um zum Beispiel auftretende Veränderungen der Marktlage schneller erfassen zu können [Liu02]. Das die Informationen in verschiedenen Abstraktionsebenen darstellbar und manipulierbar sein sollten, bedeutet, dass die betreffende Führungskraft zum Beispiel in der Lage sein sollte, Betriebsergebnisse eines Jahres einzusehen, genauso wie sie in die eines Quartals, eines Monats oder einer Woche hineinschauen können soll. Um ein solches Problem lösen zu können, müssen die Unternehmensdaten aus der "flachen" relationalen Form (Tabellen) in eine Darstellung gebracht werden, die nach den Bedürfnissen des Benutzer ausgerichtet ist, nämlich in multidimensionale Sichten.

Data-Warehouses und OLAP-Funktionen basieren auf einem solchen multidimensionalen Datenmodell. Dieses Modell stellt die Informationen zum Beispiel in sogenannten Datenwürfeln dar, die auch als Hypercubes oder Decision Cubes bezeichnet werden (vergleiche [HaKa01, Han97]). Auf Datenwürfel soll im nächsten Kapitel im Rahmen der durchführbaren OLAP-Funktionen etwas genauer eingegangen werden.

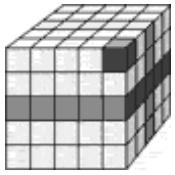


Abb.2: Datenwürfel, aus [PuSo03]

OLAP wird heute als ein dynamisches Analyseverfahren verstanden, das es einem Benutzer (zum Beispiel einem Manager) erlaubt, mittels interaktiver Datenbankzugriffe („online“) eine Vielzahl von Sichten und Darstellungsweisen über bestimmte Daten zu erhalten und damit einen schnellen und umfassenden Überblick zu verschaffen [PuSo03].

In der Praxis findet OLAP vor allem in sogenannten *Decision-Support-Systemen* Verwendung, die Entscheidungsträger wie Manager oder Controlling-Mitarbeiter durch Analyse der Unternehmensdaten besser mit relevanten Daten versorgen und bei der Entscheidungsfindung unterstützen sollen.

4.2 OLAP-Funktionen

Da in großen Datenmengen herkömmliche Methoden für den Datenzugriff, wie zum Beispiel SQL-Anfragen, sehr zeitintensiv sind und Rechnersysteme bis an ihre Leistungsgrenze bringen, sind neue Zugriffsmethoden und -funktionen gefragt. Die multidimensionale Datenhaltung in Data-Warehouses und OLAP-Datenbanken ermöglicht verschiedene Wege, zum Beispiel mit Hilfe eines Datenwürfels eine multidimensionale Datenanalyse ad hoc und schnell durchzuführen. Dabei wird die Multidimensionalität der Datenmodelle und die Möglichkeit, Hierarchien bilden zu können, ausgenutzt. Auf die Datenmodelle wird im Folgenden noch vor den tatsächlichen Operationen eingegangen werden.

4.2.1 Mehrdimensionales Datenmodell

Eine Möglichkeit eines mehrdimensionales Datenmodells ist der bereits erwähnte Datenwürfel, wie er in [HaKa01] eingeführt wird. Dieses Modell ermöglicht die Darstellung von Datensätzen in der Art, dass jedes Attribut als eigene Dimension des Würfels dargestellt werden kann. Die Wertebereiche der Dimensionen können kontinuierlich sein, oder diskrete Werte enthalten. Ein dreidimensionaler Datenwürfel kann zum Beispiel die Dimensionen Zeit, Ort und Produkt enthalten.
Ein Würfel darf auch aus mehreren kleineren Teilwürfeln zusammengesetzt sein. Um auf diese detaillierteren Informationen strukturiert zugreifen zu können, kann man innerhalb einer jeden Dimension des Würfels eine oder mehrere Hierarchien definieren. Stellt man eine solche Hierarchie als Baum dar, umfasst die Wurzel den gesamten Wertebereich, während die inneren Knoten den Wertebereich rekursiv in immer kleinere Intervalle oder diskrete Abschnitte unterteilen. Die Blätter enthalten (bei diskretem Wertebereich) die einzelnen Werte. Durch Bildung der Summe oder des Mittelwertes können dann die detaillierteren Informationen zu einem Repräsentanten zusammengefasst auf einer höheren Abstraktionsebene dargestellt werden. In eben genanntem Beispiel könnte eine Hierarchie für das Attribut *Ort* so aussehen, dass die detailliertesten Angaben aus Städten bestehen, die zu Bundesländern zusammengefasst, die wiederum zu Ländern gruppiert werden können, und so weiter (Abb. 3).

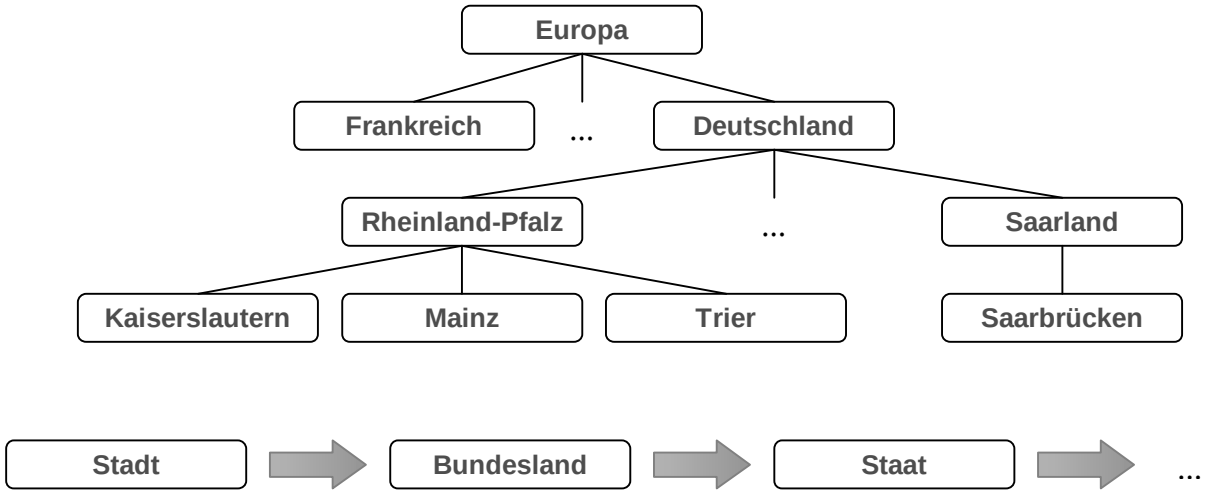


Abbildung 3: Beispiel für eine Hierarchie im Attribut *Ort*

Das Zusammensetzten mehrerer Datenwürfel zu einem großen Würfel unterstützt das Auswerten der gegebenen Datensätze in verschiedenen Abstraktionsebenen. Dabei navigiert man innerhalb der verschiedenen Ebenen und Hierarchien mit Operationen wie *Roll-up*, *Drill-down*, *Slice*, *Dice* und *Rotate*.

4.2.2 Operationen auf einen Datenwürfel

Auf n-dimensionalen Datenwürfeln stehen verschiedene Operationen zur Verfügung. Diese können die Daten auf verschiedene Art und Weise darstellen, aggregieren oder manipulieren. Die klassischen Funktionen, die auch in herkömmlichen Datenbanken nachgebildet werden können, dienen der Navigation [Han97, PuSo03]:

- Roll-up

  Roll-up ist eine Funktion, die ausgehend von einem Abstraktionsgrad, also einer Stufe in einer auf den Wertebereich einer Dimension definierten Hierarchie, eine Abstraktion vornimmt, so dass man sich in der Hierarchie eine Stufe nach oben bewegt. Eine solche Abstraktion kann beispielsweise so aussehen, dass die Daten durch aufsummieren der einzelnen Werte verdichtet werden. Ein Beispiel für ein Roll-up in der Dimension Ort wäre das Zusammenfassen der Städte Ludwigshafen, Kaiserslautern und Mainz zu dem Bundesland Rheinlandpfalz.

- Drill-down

  Drill-down ermöglicht es, innerhalb einer Dimension auf detailliertere Daten zuzugreifen. Entlang einer Attribut-Hierarchie werden Daten, die vorher verdichtet wurden, wieder detailliert dargestellt.

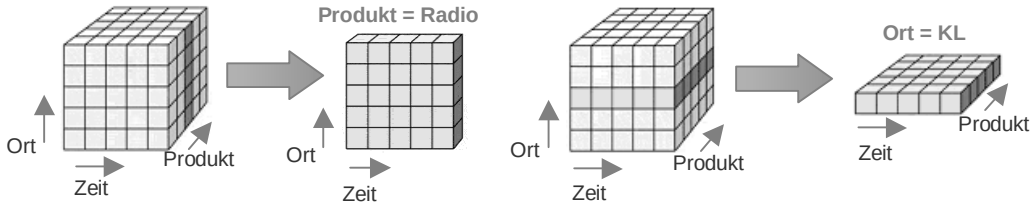
Roll-up und Drill-down können entlang einer, mehrerer oder aller Attribute beziehungsweise Dimensionen durchgeführt werden, in einer bestimmten Reihenfolge oder gleichzeitig.

Ergänzend zu diesen Funktionen sind die analytischen Funktionen *Slice*, *Dice* und *Rotate* zu nennen. Durch das eingeführte multidimensionale Datenmodell besteht die Möglichkeit, die gesammelten Un-

ternehmensdaten von mehreren Positionen, mit Hauptaugenmerk auf verschiedene Dimensionen beziehungsweise Attribute, zu betrachten und die Datenmenge einzuschränken. Das geschieht nun durch die nachfolgend aufgeführten Funktionen so, dass der Datenwürfel bildlich gesehen auseinander geschnitten oder um die verschiedenen Achsen gedreht wird [Han97, PuSo03].

- Slice

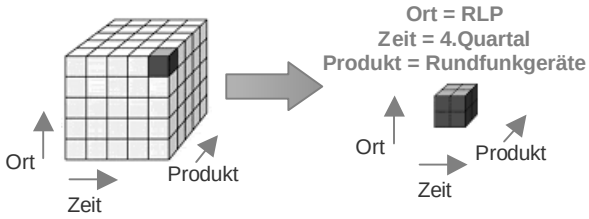
  Slice ist eine Operation, bei der die Menge der zu analysierenden Daten reduziert wird. Bei dieser Funktion wird der Wertebereich einer Dimension eingeschränkt. Bildlich kann man sich das so vorstellen, dass man eine Scheibe (Slice = Scheibe) aus einem großen Datenwürfel herausschneidet und genauer untersucht.



Beispiele zur Slice-Operation, aus [PuSo03]

- Dice

  Dice ist ebenfalls eine manipulierende Operation, welche die Menge der Daten einschränkt und kann als eine Verallgemeinerung von Slice betrachtet werden. Diese Funktion schränkt den Wertebereich mehrerer Dimensionen ein. Bildlich gesehen werden kleine Würfel (Dice = Würfel) aus der gesamten Datenmenge geschnitten. Zu einer Teilmenge der Dimensionen werden Bedingungen formuliert, welchen die Daten in der resultierenden Darstellung genügen müssen.



Beispiel zur Dice-Operation, aus [PuSo03]

- Rotate

  Rotate (auch *Pivoting* genannt) ist eine Operation, bei der Daten nicht manipuliert, sondern lediglich ihre Darstellung verändert wird. Diese Funktion ermöglicht es, die Attribute im Würfel anders darzustellen, die Dimensionen zu vertauschen und die Daten dadurch neu anzuordnen. Hier wird der Datenwürfel bildlich gesehen um die verschiedenen Achsen gedreht, beziehungsweise werden Achsen des Würfels vertauscht.

4.3 Gegenüberstellung OLAP und Data-Mining

Es gibt mehrere Ansätze, wie OLAP und Data-Mining zueinander in Beziehung stehen können. Griffin vergleicht in [Gri00] OLAP-Anwendungen mit Arbeitspferden und Data-Mining-Anwendungen mit Rennpferden. OLAP-Tools bieten neben den Standardanfragen auch Funktionen wie zum Beispiel Slice und Dice, die nötig sind, um komplexere Anfragen zu stellen und Vergleiche zu erstellen. Data-Mining-Tools gehen nach Meinung von Griffin weiter: Sie bieten Informationen an, von denen der Benutzer, noch nicht wusste, dass er danach suchte. Der Nachteil von Data-Mining-Verfahren im Gegensatz zu OLAP-Techniken wird hier so beschrieben, dass zwar eine Menge Informationen gefunden werden, die nicht zwingend nützlich sein müssen. Außerdem kann man durch die Masse der gefundenen Muster die relevanten Auffälligkeiten übersieht.

Einen weiteren Ansatz, wie Data-Mining zu OLAP steht, stellt Han in [HaKa01] vor: Er sieht Data-Mining als eine Erweiterung des OLAP an. Es geht seiner Meinung nach über den Bereich der zusammenfassenden Art des OLAP hinaus und erweitert es um fortgeschrittene Techniken zum Verstehen der Daten. Weitere Punkte sprechen dafür, Data-Mining als Erweiterung des OLAP anzusehen:

Ist bei einer Analyse mittels OLAP-Techniken ein Analyseziel notwendig, zum Beispiel in Form der Bestätigung einer Hypothese, so wird bei einer Auswertung der Daten mit Hilfe von Data-Mining-Verfahren nicht unbedingt ein konkretes Ziel vorausgesetzt, das Verfahren kann eine Hypothese liefern. Dadurch, dass das Ziel oder eine Hypothese dem Benutzer bei OLAP-Verfahren vorher bekannt sein muss, steht auch das Ergebnis – zum Beispiel in Form einer bestätigten Hypothese – fest. Dies ist beim Data-Mining nicht der Fall. Es können Muster und Beziehungen in den Daten gefunden werden, von denen der Benutzer gar nicht wusste, dass er nach solchen suchen könnte. Diese unbekannten Muster fördern das Aufstellen neuer Hypothesen, die mit Hilfe der Data-Mining-Verfahren weiter untersucht werden können. Auch wenn eine Hypothese vorlag und diese in dem Ergebnis der Analyse bestätigt wurde, kann es doch sein, dass bei der Auswertung der Daten mehr als diese Bestätigung herauskommt.

Auch die Analyse mittels Data-Mining-Techniken kann als eine Erweiterung der Datenauswertung mit Hilfe von OLAP-Verfahren angesehen werden: Dadurch, dass bei OLAP-Werkzeugen ein Benutzer konkret Daten auswählen muss, erfordert eine Analyse mit Hilfe dieser Technik Kenntnisse über die Daten und ihre Zusammenhänge. Bei Data-Mining-Verfahren hingegen kann es durchaus möglich sein, dass die Daten nicht bekannt sind und blind gesucht wird. Ebenso ist bei OLAP eine konkrete Interaktion des Benutzers gefragt, der während der Analyse die Daten aktiv mit Hilfe der vorgestellten Funktionen manipuliert und anders darstellt. Bei Data-Mining-Verfahren ist zwar eine Interaktion des Benutzers möglich, allerdings läuft der Prozess der Mustersuche aus den Daten mit Hilfe von computergestützten Algorithmen weitestgehend automatisch ab.

Als eine weitere Erweiterung des OLAP durch Data-Mining kann das Analyseergebnis angesehen werden. Besteht das Ergebnis der Auswertung bei OLAP-Verfahren aus Daten, ist das Resultat einer

Analyse mittels Data-Mining-Techniken ein Muster, das zum Beispiel Zusammenhänge und Abhängigkeiten in den Daten aufzeigt. Während bei der Analyse mittels OLAP-Tools Anfragen gestellt werden, die Daten anders dargestellt liefern, können die Ergebnisse als Basis für Data-Mining-Verfahren verwendet werden, die aus diesen Darstellungen Zusammenhänge liefern.

Han zeigt in [HaKa01] eine weitere Möglichkeit auf, wie die Beziehung zwischen OLAP und Data-Mining aussehen könnte: Aufgrund der Tatsache, dass Data-Mining als Erweiterung des OLAP angesehen werden kann, ist es auch möglich, OLAP in Data-Mining-Verfahren zu integrieren. Als Gründe für das von ihm benannte *OLAP-Mining* führt er unter anderem an, dass Data-Mining integrierte, konsistente und gesäuberte Daten voraussetzt, dass Benutzer interaktiv aus mehreren Aggregationsebenen Daten untersuchen möchten und dass OLAP durch die Datenhaltung in Data-Warehouses und das mehrdimensionale Datenmodell genau dies bietet. Er schlägt Methoden vor, welche die Besonderheiten beider Analyseverfahren berücksichtigen, in denen Data-Mining-Algorithmen zum Beispiel auf Datenwürfel angewandt werden („Cubing then mining“) oder Data-Mining-Methoden verwendet werden und deren Ergebnisse mittels Datenwürfeln visualisiert werden („Mining then cubing“).

5 Problemfelder des Data-Mining

Trotz all der Vorzüge, die Data-Mining in seiner heutigen Form bietet, besteht weiterer Forschungsbedarf, der unter anderem weiterhin aus dem rasanten Anwachsen der Datenmengen und der Weiterentwicklung der informationsverarbeitenden Techniken resultiert. Grossman et al stellen in [GrKaMo98] einige Herausforderungen für Data-Mining-Lösungen vor.

Da die Datenmengen – und damit auch für eine Analyse interessante Daten – exponentiell zunehmen, die Kapazität der Speichermedien jedoch nicht ganz in dieser Geschwindigkeit anwächst, wird es immer wichtiger, dass Data-Mining-Algorithmen unabhängig von der Datenmenge die Analyseaufgabe bewältigen und somit skalierbar sind. So ist es beispielsweise nötig, neue Algorithmen zu entwerfen, die eine Analyse in mehreren Durchgängen ermöglichen, so dass die Analyse unabhängig von der Datenmenge im Speicher durchgeführt werden kann.

Da die Anforderungen an die Antwortzeiten der Systeme steigen, aber durch die zunehmende Datenmenge immer schwieriger erfüllbar sind, besteht zum Beispiel ein Bedarf für parallele Analysetechniken. Solche verteilte Methoden sollten es auch ermöglichen, Daten aus verschiedenen Quellen zu analysieren, ohne diese zuvor auf einem Rechner beziehungsweise an einem Ort zu sammeln und aneinander anzupassen.

Ein weiteres Problemfeld stellt die zunehmende Vielfalt der Datentypen dar, die analysiert werden können. So gibt es mittlerweile die Möglichkeit, Multimediadaten wie Fotos und Filme in Datenbanken zu speichern, weshalb es auch entsprechende Verfahren geben sollte, die in der Lage sind, diese zu analysieren.

Des Weiteren lässt bei bisherigen Data-Mining-Systemen die Benutzerfreundlichkeit oft zu wünschen übrig, da die meisten dieser Systeme für die Benutzung durch Experten ausgelegt sind. Um sie allgemeiner einsetzbar zu machen, sollten diese Tools intuitiver anwendbar sein. Es könnten vollständige Data-Mining-Umgebungen entwickelt werden, vergleichbar mit Arbeitsumgebungen für die Softwareentwicklung.

Datenschutz und Sicherheit stellen eine weitere Herausforderung der Entwicklung im Bereich des Data-Mining dar. Da die Data-Mining-Methoden zwangsweise immer mächtiger werden wächst auch die Gefahr, dass Ergebnisse und gewonnenen Erkenntnisse missbraucht werden, stärker an, weshalb Schutzmechanismen benötigt werden.

6 Zusammenfassung

Durch den schnellen Fortschritt in der Informationstechnologie und das immer schnellere Anwachsen der Datenmengen werden die Anforderungen an Systeme, die Wissen in irgendeiner Form aus Daten extrahieren und darstellen, ständig steigen. Der stetige Wachstum der Datenbestände macht den Zugriff auf die gewünschten Informationen immer schwieriger, eine manuelle Analyse „von Hand“ wird immer zeitaufwendiger, immer personalintensiver und dadurch kostspieliger und für einen Menschen quasi unmöglich. Es werden daher immer zeitsparendere und effektivere Systeme und Methoden zur Wissensgewinnung gesucht.

Das im vierten Kapitel angesprochene *Online Analytical Processing* wird im Hinblick darauf weiterhin ein wichtiges Analysewerkzeug für Entscheidungsträger bleiben, da durch das multidimensionale Datenmodell wie zum Beispiel der angesprochene Datenwürfel und die multidimensionale Datenhaltung in OLAP-Datenbanken eine intuitive Datenauswahl und dadurch eine effektive Analyse stattfinden kann. Die dafür verwendeten Operationen sind unter anderem die vorgestellten Funktionen *Roll-up, Drill-down, Slice, Dice* und *Rotate,* mit denen in Datenbeständen navigiert und die Darstellung der Daten verändert werden kann. OLAP hat zwar den Vorteil, dass es interaktiv ist, allerdings besteht die Gefahr, dass wichtige Auffälligkeiten oder Zusammenhänge aufgrund der Größe des Suchraumes von der analysierenden Person übersehen werden.

Mit Hilfe der Analysemethode des Data-Mining, das im zweiten Kapitel genauer erläutert wurde, können Zusammenhänge und Regelmäßigkeiten gefunden werden, die der Benutzer nicht kannte oder erwartet und danach gesucht hat. Eingebettet in den im dritten Kapitel vorgestellten KDD-Prozess, in dem gesammelte Daten für diese Analyse vorbereitet werden, kann das Suchen nach Auffälligkeiten insofern beschleunigt werden, dass das Vorbereiten der Daten für eine Aufgabenstellung einmal geschieht, das Auswerten der Daten allerdings beliebig oft und durch die Vorbereitung um einiges beschleunigt ablaufen kann.

Im letzten Kapitel wurden Probleme des Data-Mining angesprochen, die in der bisherigen Praxis aufgetaucht sind und solche, die durch die Entwicklung im Bereich der Informationstechnik und die zunehmende Datenmenge absehbar sind.

7 Quellenangaben

[BeLi97] Berry, Linoff:
Data Mining Techniques: For Marketing, Sales, and Customer Support; John Wiley & Sons; 1997

[CoCoSa93] Codd, E. F., Codd, S. B., and Salley, C. T.:
Providing olap (on-line analytical processing) to user-analysts: An it mandate. Technical report, E.F.Codd & Associates 1993

[Das03] Dastani, P.:
Data Mining – Eine Einführung,
Forum Database Marketing & Mining, 2000;
http://www.data-mining.de, abgerufen Dezember 2003

[DeFo95] Decker, Focardi:
Technology overview: a report on data mining. Technical Report CSCS TR-95-02, CSCS-ETH, Swiss Scientific Computing Center, 1995

[FaPiSm96] Fayyad, Piatesky-Shapiro, Smyth:
From Data Mining to Knowledge Discovery: An Overview; In Advances in Knowledge Discovery and Data Mining, 1996

[FaPiSm96a] Fayyad, Piatesky-Shapiro, Smyth:
The KDD Process for Extracting Useful Knowledge from Volumes of Data; In Communications of the ACM, Vol. 39, No 11, November 1996

[For97] Forsman, S.:
OLAP Council White Paper, OLAP Council 1997
http://www.olapcouncil.org/research/whtpaply.htm, abgerufen Dezember 2003

[GrBe99] Grob, H. L., Bensberg, F.:
Das Data-Mining-Konzept, Arbeitsbericht Nr. 8,
Münster 1999

[Gri00] Griffin, J.:
OLAP Vs. Data Mining: Which One is Right for Your Data Warehouse?
dataWarehouse.com - The data warehousing community,
Arthur Andersen Business Consulting 2000
http://www.datawarehouse.com

[GrKaMo98] Grossman, R., Kasif, S., Moore, R., Rocke, D., Ullman, J.:
Data Mining Research: Opportunities and Challenges – A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data, 1998
http://www.rgrossman.com/reprints/dmr-v8-4-5.htm, abgerufen Januar 2004

[HaKa01] Han, J., Kamber, M.:
Data Mining - Concepts and Techniques, Morgan Kaufmann Publishers, 2001

[Han97] Han:
OLAP Mining, an Integration of OLAP with Data Mining, 1997

[Liu02] Liu, G.:
A Proposal of High Performance Data Mining System; Lecture Notes in Computer Science, Springer Verlag, Berlin Heidelberg 2002

[Pet97] Petrak, J.:
Data Mining - Methoden und Anwendungen. Technischer Report OEFAI-TR-97-15, Österreichisches Forschungsinstitut für Artificial Intelligence, 1997.

[PuSo03] Purgold-Software:
Was ist OLAP? , Hamburg 2003
http://www.purgold-software.de/info/olap.html, abgerufen Dezember 2003