Seminararbeit zum Thema

Anwendungen des Data Mining in der Praxis

von Holger Dürr

Seminar Data Mining – Wintersemester 2003/2004
Professor Dr. Franz Schweiggert – Universität Ulm

Inhaltsverzeichnis

1 Einleitung

Bei der Informationssuche zum Thema Data Mining trifft man immer wieder auf die bekannte These des Trendforschers John Naisbett „We are drowning in information, but starving for knowledge“ – „Wir ertrinken in Informationen, aber hungern nach Wissen“ [Hudec].
Die historische Entdeckung unseres Planetensystems verdeutlicht, wie eng das Sammeln von Daten und die Entdeckung von neuem Wissen zusammenhängen kann. Allerdings zeigt dieses Beispiel auch, wie schwer es trotz der genauen Kenntnis von Daten sein kann, daraus neue Erkenntnisse zu ziehen. Der dänische Astronom Tycho Brahe erstellte in jahrelangen Beobachtungen mit Hilfe einer aus heutiger Sicht primitiven Technik die bis dahin genaueste Aufzeichnung über die Stellung der Himmelskörper. Allerdings gelang es dem dänischen Astronom nicht, ein schlüssiges Konzept für die Bewegung der Himmelskörper zu entdecken. Erst nach Brahes Tod gelang es dem deutschen Astronom Johannes Kepler, aus Brahes Aufzeichnungen unser heutiges Planetensystem herzuleiten.
Heute sind dank der elektronischen Mittel riesige Mengen von Daten gespeichert. Und so wie Johannes Kepler in den Aufzeichnungen von Brahe nach einem Muster für die Planetenbewegung suchte, so ist Data Mining der Ansatz in den riesigen Datenmengen, die uns heute vorliegen, mit Hilfe von modernen Rechnern neues Wissen zu entdecken [Bristol].
In dieser Arbeit wird ein Überblick über die Funktionsweise, die Verfahren, den Ablauf, die Anwendungen und Probleme des Data Minings in der Praxis gegeben.

2 Data Mining

2.1 Definition des Data Mining

Unter Data Mining versteht man eine Menge von Datenanalysemethoden. Umstritten bleibt jedoch welche konkreten Verfahren dem Data Mining zuzuordnen sind. Eine allgemein anerkannte Definition beschreibt Data Mining als nicht triviale Entdeckung gültiger, neuer, potentiell nützlicher und verständlicher Muster in großen Datenbeständen [KnobWeid].

2.2 Einordnung des Data Mining

Datenanalyseprobleme lassen sich in zwei verschiedene Klassen einteilen. Das Kriterium für die Einteilung ist das Ausmaß, in dem die Hypothesen des Anwenders eine Rolle spielen. Man unterscheidet zwischen hypothesengetriebenen und hypothesenfreien Problemen. Das Ziel bei hypothesengetriebenen Problemen ist die Verifizierung oder Falsifikation von Annahmen oder Theorien anhand von Datenbeständen. Man bezeichnet die zugehörigen Verfahren auch als Top-Down-Ansätze, da sie die Datenbestände von einer Hypothese ausgehend untersuchen.
Im Gegensatz dazu wird bei hypothesenfreien Problemen keine Hypothese überprüft, sondern ausgehend von den Daten werden neue Erkenntnisse erzeugt. Man bezeichnet die zugehörigen Verfahren auch als Bottom-Up-Ansätze oder datengetriebene Analysen. Die Data Mining-Verfahren gehören somit in die Klasse der datengetriebenen oder hypothesenfreien Analysen [KnobWeid].

2.3 Arten des Data Mining

Die klassische Form des Data Mining ist die Suche nach Mustern in Datenbeständen, die in tabellarischer Form vorliegen. Man möchte z.B. unbekannte Zusammenhänge zwischen einer Zielvariablen und den weiteren Merkmalen von Objekten erkennen oder die verschiedenen Objekte auf Ähnlichkeiten untersuchen. Die Analyse von Kundendatenbanken eines Unternehmens ist ein Beispiel für diese Form des Data Mining.

Allerdings liegen heutzutage viele Unternehmensdaten nicht in tabellarischer Form, sondern als Texte, im Web oder in Form von Bildern und Filmen vor. Deshalb entstanden mittlerweile neben dem klassischen Data Mining in Tabellen auch Text und Web Mining. Beim Text Mining werden nun Texte auf Ähnlichkeiten analysiert und klassifiziert. Eine Anwendung des Text Mining ist die automatische Einordnung und Weiterleitung von E-Mails. Beim Web Mining steht die Analyse von Internetseiten und das Navigationsverhalten der Benutzer im Mittelpunkt.
Außerdem wird zur Zeit zum Beispiel am Fraunhofer Institut im Bereich des Multimedia Mining geforscht. Das Multimedia Mining analysiert und klassifiziert Bilder und Filme und kann so zum Beispiel in der Zukunft die Archivierung erleichtern.

2.4 Die Bedeutung von Data Mining

Welche Bedeutung Data Mining heute schon hat und welche Rolle es in der Zukunft noch spielen könnte, verdeutlichen die folgenden Studien. Bei IBM wird davon ausgegangen, dass sich die weltweit vorhandene Informationsmenge alle 20 Monate verdoppelt [IBMNews]. Somit erscheint es fast unumgänglich auch für die Informationsgewinnung die modernen Rechner zu verwenden.
Auch ein Blick in die Unternehmenswelt zeigt, dass das Interesse an Data Mining weiterhin sehr groß ist. Laut einer Studie der Gartner Group waren im Jahr 2000 mindestens bei der Hälfte der sogenannten Fortune-1000-Unternehmen Data Mining-Technologien im Einsatz [IBMNews]. Eine Studie der Katholischen Universität Eichstätt-Ingolstadt ergab, dass in Deutschland im Jahr 2002 ebenfalls knapp 50% der 500 größten Unternehmen Data Mining oder multivariate Statistik zur Analyse ihrer Kunden benutzten.
Die Studie zeigt auch das Wachstumspotenzial das Data Mining noch besitzt. Fast alle der Unternehmen, bei denen Data Mining-Techniken angewandt werden, wollen in Zukunft diesen Einsatz noch erhöhen, und 87% dieser Unternehmen berichten über eine hohe Rentabilität ihrer Data Mining-Projekte [Golem]. Das Beratungsunternehmen NHConsult schätzt zudem, dass zur Zeit nur ca. 10% der in den Unternehmen gespeicherten Datenbestände überhaupt analysiert werden [NHConsult].

3 Aufgabentypen des Data Mining

3.1 Assoziationsanalysen

Bei der Assoziationsanalyse suchen die Data Mining Verfahren nach interessanten Abhängigkeiten zwischen einzelnen Untersuchungsobjekten. Die identifizierten Muster können dann zum Beispiel in Form von Wenn-Dann-Regeln sprachlich formuliert werden.
Ein verbreiteter Anwendungsbereich der Assoziationsanalyse ist die Warenkorbanalyse bei Supermärkten. Man untersucht die von den Kunden gekauften Waren in einem bestimmten Zeitraum um das Kaufverhalten kennen zu lernen. Ein klassisches Ergebnis einer Warenkorbanalyse ist das folgende Muster: Wenn Kunden Brot und Butter kaufen, dann kaufen 70% der Kunden auch Marmelade. Die einfache Warenkorbanalyse erfasst allerdings nur die Abhängigkeiten zu einem bestimmten Zeitpunkt. Um Abhängigkeiten zu unterschiedlichen Zeitpunkten zu erkennen, kann man die Analyse noch um die Dimension Zeit erweitern. Das Ziel dieser sogenannten Sequenzanalyse ist es zeitliche Distanzen bei Einkäufen zu entdecken. Man benötigt dafür jedoch die Daten der einzelnen Kunden über einen längeren Zeitraum. Allerdings sind diese Informationen heutzutage mit Kunden- und Pay-Back-Karten einfach zu bekommen. Ein Ergebnis einer Sequenzanalyse wäre beispielsweise folgende Aussage: Wenn Kunden im Winter einen Fernseher kaufen, kaufen 50% von ihnen nach spätestens 10 Wochen auch einen Videorecorder [GrobBens].

3.2 Klassifikationsanalysen

Die Verfahren zur Klassifikation ermitteln Muster, um Aussagen über Objekte anhand von vorhandenen Informationen zu treffen. Zuerst werden die bereits vorhandenen Objekte nach ihrem bekannten Merkmal oder Verhalten bezüglich des zu analysierenden Problems in verschiedenen

Klassen zusammengefasst. Aus dieser Menge von Objekten wird nun ein Klassifikationsmodell entwickelt, mit dem man dann die Klassenzugehörigkeit eines neuen Objekts vorhersagen kann. Bei der Klassifikationsanalyse können Methoden der künstlichen Intelligenz oder entscheidungsbaumorientierte Methoden eingesetzt werden.
Eine Anwendung der Klassifikationsanalyse ist zum Beispiel die Beurteilung der Kreditwürdigkeit von Bankkunden. Hier werden zuerst die bisherigen Kreditnehmer, die ihren Kredit zurückbezahlten, in der Gruppe „kreditwürdig“ zusammengefasst. Die Kunden, die ihren Kredit nicht zurückbezahlten, bilden dann die Gruppe „nicht kreditwürdig“. Anhand dieser bekannten Testdaten entwickelt das Verfahren nun einen Entscheidungsbaum mit zum Beispiel folgenden Komponenten: Verschuldungsgrad, Alter, Einkommen, Sicherheiten..... Anhand des Entscheidungsbaumes kann man nun bei der Kreditvergabe in Zukunft den Kunden anhand der Informationen Alter, Verschuldungsgrad, Einkommen in die Gruppe „kreditwürdig“ oder „nicht kreditwürdig“ einordnen.
Abb. 1 zeigt einen möglichen Entscheidungsbaum. Hier entspricht die Klasse 0 der Gruppe „nicht kreditwürdig“ und Klasse 1 der Gruppe „kreditwürdig“.

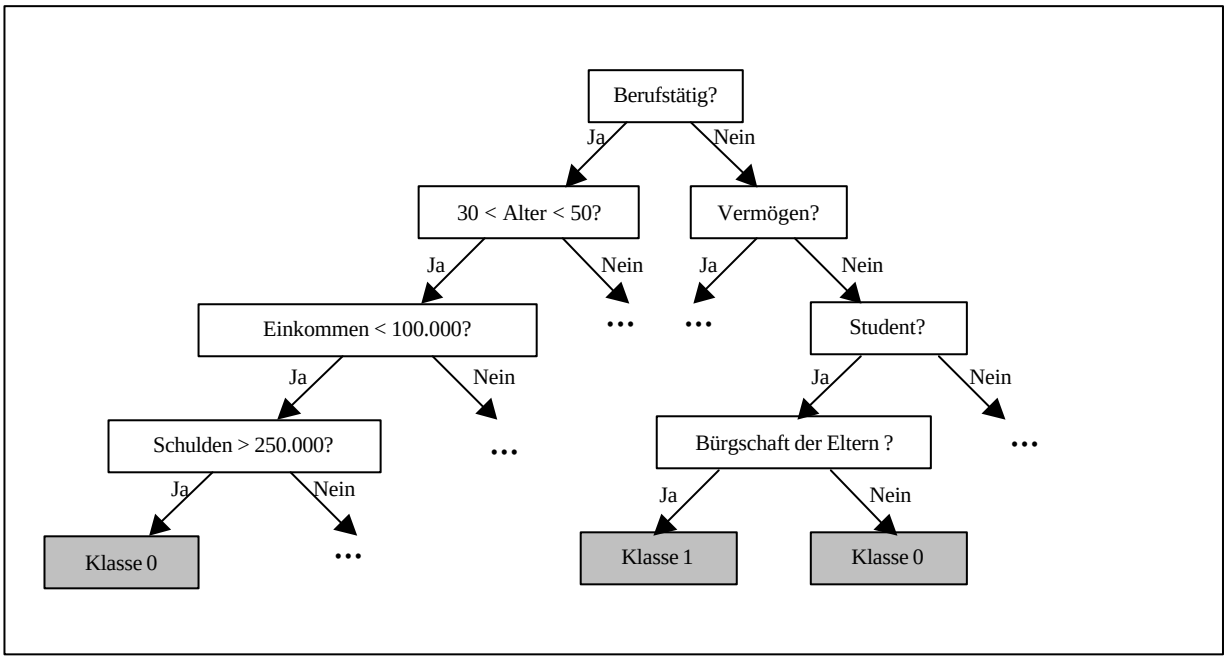

Abb. 1: Entscheidungsbaum für Kreditwürdigkeit

Dieser in Abb.1 dargestellte Entscheidungsbaum sagt voraus, dass jemand der berufstätig ist, zwischen 30 und 50 Jahre alt ist, sein Einkommen maximal 100.000 Euro beträgt und seine Schulden größer als 250.000 Euro sind, sehr wahrscheinlich einen weiteren Kredit nicht zurückzahlen kann.
Dagegen legt das Modell nahe, einem Studenten, der kein Vermögen besitzt, aber eine Bürgschaft seiner Eltern vorweisen kann, einen Kredit zu gewähren [Exper].

3.3 Clustering

Beim Aufgabentyp des Clustering sind im Gegensatz zu den Klassifikationsanalysen die Klassen der Objekte nicht im voraus bekannt. Statt dessen versucht man auf der Basis von Distanzmaßen und unter Berücksichtigung vieler Merkmale die Objekte in möglichst homogene Gruppen aufzuteilen. Anschließend können dann Modelle entwickelt werden, die neue Objekte den gefundenen Clustern zuordnen [GrobBens].
In Abb. 2 werden in einem vereinfachten Beispiel die Kunden einer Telefongesellschaft anhand von der Anzahl von Orts- oder Ferngesprächen in vier Gruppen eingeteilt. Man hat in diesem Fall eine Gruppe mit wenigen Orts- und Ferngesprächen, eine weitere Gruppe mit vielen Orts- und wenig

Ferngesprächen, eine dritte mit vielen Fern- und wenig Ortsgesprächen und noch die Gruppe, die sowohl viele Orts- als auch Ferngespräche führt. Die Anzahl in wieviel Gruppen die Objekte aufgeteilt werden ist im voraus nicht bekannt, sondern ergibt sich als die bestmögliche Gruppeneinteilung durch das Verfahren [Exper].

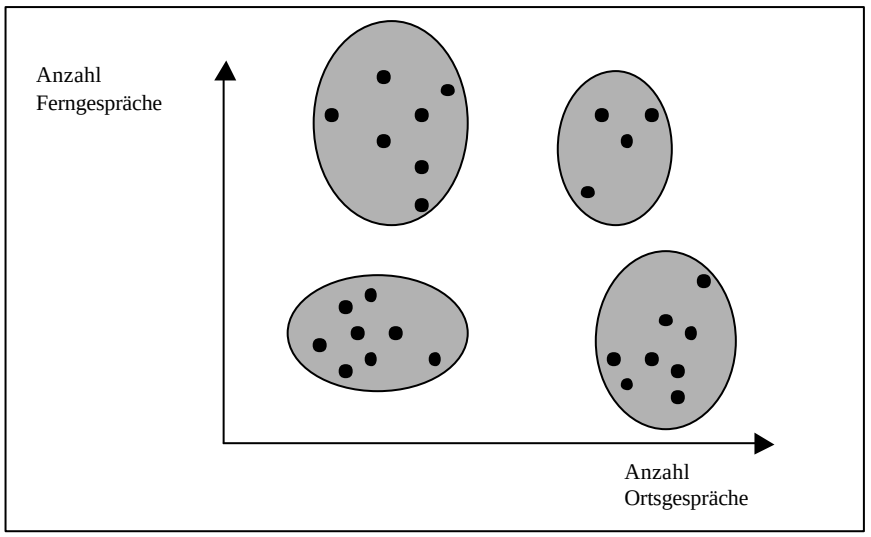

Abb.2: Cluster der Kunden einer Telefongesellschaft

4 Der Analyseprozess in der Praxis

Für die Durchführung von Data Mining-Untersuchungen in der Praxis gibt es verschiedene Prozessmodelle. Im Folgenden werden die wichtigsten Prozessphasen anhand des CRISP-DM-Modells (Cross Industry Standard Process for Data Mining) etwas näher betrachtet.

4.1 Business Understanding

In dieser Phase werden die Ziele, Erwartungen, Ressourcen und Restriktionen in eine Data Mining-Problemstellung transformiert. Hierzu gehört die genaue Formulierung des Ziels sowie der damit verbundenen Fragestellungen. Beispielsweise müsste die Zielformulierung „Umsatzsteigerung im nächsten Jahr“ sicherlich konkretisiert werden. Denn man könnte den Umsatz durch höhere Ausschöpfung des bisherigen Kundenpotentials oder durch die Gewinnung neuer Kunden erreichen.
Außerdem werden hier schon Erfolgskriterien festgelegt, um später beispielsweise anhand von Kontrollgruppen zu erkennen, ob das Projekt den gewünschten Erfolg bringt.

4.2 Data Understanding

Im zweiten Schritt wird nun analysiert, welche Daten für die Untersuchung nützlich sind und welche Daten zur Verfügung stehen. Als Datenquellen dienen die internen Unternehmensdatenbanken, aber man kann gegebenenfalls noch auf weitere externe Daten wie beispielsweise Marktforschungsdaten zurückgreifen. Zudem wird hier die Datenqualität betrachtet, die später sehr entscheidend für den Erfolg des Projektes ist. Man untersucht, ob die Daten aktuell und valide sind.

4.3 Data Preperation

In der Data Preperation-Phase werden die Datenbestände aus den verschiedenen Quellen zusammengefügt und die Daten bereinigt, so dass keine doppelten, fehlerhaften oder unvollständigen Einträge

vorliegen. In der Praxis wird damit gerechnet, dass 1% - 5% der Einträge des Datenbestandes Fehler aufweisen [GrobBens]. Zudem werden die Daten in die benötigten Formate übertragen.

4.4 Modeling

Hier beginnt nun der eigentliche analytische Teil des Data Mining-Prozesses. Trotzdem sollten die vorhergehenden Phasen nicht unterschätzt werden, denn sie haben entscheidenden Einfluss für die Qualität der Aussagen, die man später erhält. Außerdem wird geschätzt, dass die vorhergehenden Phasen der Datenvorbereitung bis zu 80% der Zeit eines Data Mining-Projektes benötigen und somit hohe Kosten verursachen [GrobBens].
In diesem Teil des Prozesses werden nun die Methoden zur Mustererkennung ausgewählt und eingesetzt. Anschließend wird das Modell auf Genauigkeit und Allgemeingültigkeit untersucht, beispielsweise indem man es auf nicht zur Modellbildung verwendete Testdatensätze anwendet. Durch verschiedene Änderungen am Modell und seinen Parametern nähert man sich iterativ an das endgültige Modell an.

4.5 Evaluation

Nach der technischen Überprüfung des Modells wird nun die fachliche Angemessenheit und Relevanz der erarbeiteten Lösung überprüft. Es wir untersucht, ob die Ergebnisse zur Lösung des gestellten Geschäftsproblems beitragen und ob das Ergebnis fachlich plausibel erscheint oder ob man z.B. widersprüchliche Aussagen erhalten hat.

4.6 Deployment

In dieser Phase wird das Verfahren eingesetzt, um in der betrieblichen Praxis auftretende Fragestellungen zu beantworten. Häufig werden zur Beurteilung der Ergebnisse Kontrollgruppen gebildet, um den Unterschied zwischen dem herkömmlichen und dem neuen Verfahren zu erkennen [Feldkirch].

5 Anwendungen des Data Mining – Praxisbeispiele

5.1 Überblick über die Anwendungen in der Praxis

Data Mining-Techniken werden in den verschiedensten Unternehmungen angewandt. Ein große Anzahl von Anwendungen findet man branchenübergreifend im Marketing und Custumer Relationship Management (CRM). Zudem wird Data Mining in der Risikoanalyse von Banken und Versicherungen eingesetzt. Im Handel ergeben sich Einsatzmöglichkeiten bei der Analyse des Kaufverhaltens oder der Erkennung von zahlungsunfähigen Kunden. Für die Webportale der Handelsunternehmen können Web Mining-Methoden zur Kundenanalyse benutzt werden. Im nichtbetriebswirtschaftlichen Bereich findet man Data Mining beispielsweise in der Verbrechensbekämpfung wieder. Aber auch in der Wissenschaft kann Data Mining zum Beispiel in der Astronomie oder der molekularbiologischen Forschung eingesetzt werden. Ein für die Zukunft vielleicht bedeutendes Gebiet in der Wissenschaft ist die Analyse der großen Mengen von Fachliteratur mit Hilfe von Text Mining Methoden.
Bezüglich der Sektoren ist der Einsatz vor allem in den dienstleistungsorientierten und informationsintensiven Branchen von Banken, Versicherungen, Telekommunikation und Handel zu beobachten[GrobBens].
Die folgenden Praxisbeispiele wurden bei einer gründlichen Internetrecherche ausgewählt, um die typischen Fragestellungen und Anwendungen von Data Mining vorzustellen. Da allerdings wenig wissenschaftliche Arbeiten zu konkreten Anwendungen vorliegen, stammen die Informationen und Berichte über die Projekte fast ausschließlich von Softwareanbietern oder Beratungsfirmen. Diese Unternehmen stellen der Öffentlichkeit leider nur die erfolgreichen Ergebnisse, die sie bei

verschiedenen Projekten erzielt haben, dar. Über die Probleme und gescheiterten Projekte ist dagegen nur wenig zu erfahren. Auf die Probleme, die es bei der Umsetzung von Data Mining Projekten geben kann, wird im Beispiel aus dem Pharmagroßhandel näher eingegangen.

5.2 Anwendungen im CRM und Marketing

Unter **C**ustumer **R**elationship **M**anagement versteht man die Gestaltung der Beziehungen zwischen einem Unternehmen und seinen Kunden und es hat das Ziel diese Beziehungen effizienter und effektiver zu organisieren [Netlex], [Publi]. Die Analyseprobleme in diesem Bereich beziehen sich oft auf Fragen folgender Art:

?? Welchem Kunden biete ich welches Produkt an?
?? Wer ist ein potentieller Kandidat für eine Direktmailing-Kampagne?
?? Bei welchem Kundenprofil lohnt sich ein Außendienstbesuch?

Diese Analysen werden vor allem bei Banken, Versicherungen oder im Handel durchgeführt. Im Folgenden wird anhand eines Kundenclusterings eines Kreditinstituts der Prozess näher dargestellt. Diese Analyse wurde mit der Unterstützung der Datenanalysefirma Dataengine durchgeführt.
Das Ziel der Analyse war die Zusammenfassung von Kunden des Kreditinstitutes zu verschiedenen Kundengruppen, um darauf aufbauend ähnlichen Kunden die gleichen und für sie am besten geeigneten Produkte anzubieten und die passenden Marketingstrategien anzuwenden. Man erhofft sich dann durch die bessere Kundenansprache eine Umsatzsteigerung.
Für das Kundenclustering wurde aus der Datenbank des Instituts eine repräsentative Stichprobe der Kundendaten ausgewählt. Aus den so vorliegenden Merkmalen wie z.B. Alter, Geschlecht, Familienstand, Anzahl der Kinder, Einkommen, Vermögen, Deckungsbeitrag des Kunden....wurden durch Gespräche mit Kundenbetreuern die relevanten Merkmale selektiert. In diesem Fall waren es Alter, Nettoeinkommen pro Monat, angelegtes Geldvermögen und der Deckungsbeitrag den dieser der Kunde der Bank brachte.
Danach wurde aus mehreren Clusterungen mit verschiedenen Klassenanzahlen die Segmentierung ausgewählt, die geeigneten Gütekriterien am besten entsprach. In diesem Fall wurden die Kunden in sieben sinnvolle Klassen eingeteilt. Die für das Institut gewinnbringendste Gruppe, d.h. die mit dem höchsten Deckungsbeitrag, hatte folgende Eigenschaften: zwischen 40 und 50 Jahren, kein übermäßiges Vermögen, hohe Einkommen aber auch hohe Verbindlichkeiten, die vor allem in Immobilien oder Investitionen in der eigenen Firma verwendet wurden [Dataeng].
Neben dem allgemeinen Kundenclustering interessieren bei gezielten Werbeaktionen wie zum Beispiel einer Mailingkampagne die Antwortwahrscheinlichkeiten (Response-Wahrscheinlichkeiten) der angeschriebenen Kunden. Für Versandhandel und Spendenorganisationen sind diese Analysen von großer Bedeutung. Das Marketing-Beratungsunternehmen OgilvyOne führte für UNICEF Deutschland eine solche Untersuchung durch. Das Ziel der Untersuchung war es den Mailing-Rücklauf zu optimieren, indem man nur Adressaten mit hoher Antwortwahrscheinlichkeit anschreibt.
Bisher benutzte UNICEF zur Auswahl der Adressaten ein sogenanntes RFM-Modell. Dabei verwendet man die drei Merkmale „letzte Spende“ (recency), „Anzahl der Spenden“ (frequency) und „Höhe der Spende“ (monetary value) der Adressaten, um zu entscheiden, ob ein Spender auf die Anfrage reagieren würde oder nicht. Allerdings war das Modell nicht geeignet, noch weitere Variablen aufzunehmen.
Im neuen Modell wurden der Datenbestand, der aus den Mailingergebnissen der letzten fünf Jahre gewonnen wurde, zuerst anhand von ca. 30 Variablen eindimensional auf deren Einfluss auf die Zielvariable „spenden bzw. nicht spenden“ untersucht. Neben den drei oben genannten Variablen wurden nun auch Alter, Geschlecht, Wohnort und weitere unicef-spezifische Variablen aufgenommen. Durch diese Analyse wurde schon eine deutliche Themenaffinität aufgedeckt, d.h. es gab einen Zusammenhang zwischen den Merkmalen der Spender und den verschiedenen Themen der Anschreiben. Einige reagierten auf „Hunger in Afrika“ eher als auf „Katastrophenhilfe in Asien“.
Danach wurden mit Hilfe von einem statistischen Segmentierungsmodell die besten Vorhersagevariablen bezüglich der Zielvariable „spenden bzw. nicht spenden“ herausgefunden. Anschließend wurden zwei Mailingaktionen durchgeführt, eine nach dem neuen Segmentierungsverfahren und eine nach dem bisherigen RFM-Verfahren. Das neue Modell hatte bis zu 80% höhere Antwortquoten. Der Return on Investment war 65% höher als beim herkömmlichen Verfahren [vonLühe].

Eine ähnliche Untersuchung wurde bei der Raiffeisenlandesbank Niederösterreich-Wien AG durchgeführt. Hier sollte eine Mailingaktion durchgeführt werden, um neue Kunden für Wertpapiere und Aktienfonds zu gewinnen. Durch Data Mining-Methoden wurden hier die Verkaufszahlen für Wertpapiere um 65% und für Fonds um 42% gesteigert [Sailer].

5.3 Anwendungen in Banken und Versicherungen

Neben den Anwendungen im CRM und Marketing gibt es bei Banken und Versicherungen noch die Einsatzmöglichkeit von Data Mining in der Risikoanalyse. Beispielsweise bei der Entscheidung, ob einem Kunden ein Kredit beziehungsweise eine Kfz- oder Lebensversicherung angeboten werden soll. Bei Versicherungen können die Datenbestände auf Kunden untersucht werden, die ein hohes oder niedriges Schadensaufkommen haben. Hier kann man nun wieder mit Klassifizierungsverfahren nach Merkmalen suchen, die diese beiden Kundengruppen haben.
Nach dem gleichen Schema können auch Probleme aus dem Bereich Betrugsaufdeckung analysiert werden und Entscheidungsunterstützung liefern. Beispielsweise, ob bei Abbuchungen Kreditkartenmissbrauch vorliegt, oder ob es sich um eine gerechtfertigte Versicherungsleistung handelt [Exper].

5.4 Anwendungen im Handel

Auch im Einzelhandel wird Data Mining nicht nur im CRM, sondern auch bei der oben schon kurz vorgestellten Warenkorbanalyse eingesetzt. So sammelt die US-amerikanische Supermarktkette Wal-Mart täglich alle Transaktionen (ca. 20 Mio.) in einer zentralen Datenbank. Die dann erstellten Warenkorbanalysen können dann Entscheidungsunterstützung für die Verkaufsraumgestaltung oder die Bestellmengenplanung liefern [GrobBens].

5.5 Anwendung bei der Verbrechensbekämpfung

Bei der Aufklärung von ungelösten Kriminalfällen kommen mittlerweile auch Data Mining-Verfahren zum Einsatz. So versuchte die Polizei in Großbritannien mit Hilfe von Data Mining, die Muster von Täterbeschreibungen und Täterverhalten mit Clusterverfahren in Gruppen einzuteilen. Anschließend wurde untersucht, ob sich Tätergruppen mit ähnlichen physischen Eigenschaften mit den Tätergruppen mit ähnlichem Verhalten decken. Ist eine hohe Übereinstimmung vorhanden und ist ein Täter für diese Art von Straftat bekannt, ist er auch für diese Taten verdächtig. Ist kein Täter bekannt, deuten die Cluster aber auf einen Serienstraftäter hin, kann man die verschiedenen Indizien kombinieren und möglicherweise neue Erkenntnisse sammeln [SPSSPoli].
Nach einem Bericht der Südwestpresse verwenden auch CIA und FBI Text Mining Verfahren, um Geheimdienstinformationen auszuwerten und Verbindungen zwischen verschiedenen Gruppen zu erkennen.

5.6 Anwendungen des Text Minings

Auch in der Unternehmenswelt finden Text Mining-Methoden in den verschiedensten Branchen und Bereichen Einsatzmöglichkeiten. Im Gegensatz zu Suchmaschinen werden bei den Text Mining Verfahren nicht nur die Texte nach Schlagwörtern abgesucht, sondern ebenso der Satzbau und die Wortarten analysiert. Mit den Text Mining Verfahren können auf diese Weise über 250.000 Seiten Text pro Stunde analysiert werden.
In der Produktentwicklung dient Text Mining der Analyse von Fachartikeln, Projektberichten und Patentdatenbanken. Auf diese Weise kann man frühzeitig Informationen über laufende Projekte und Patentanmeldungen bekommen.
Im Marketing ermöglicht Text Mining, durch die Analyse von Webseiten und Pressetexten der Konkurrenten, eine Sondierung der Konkurrenz. Die gesammelten Informationen können dann in die eigene Produkt- und Preispolitik einfließen.
Eine weitere Einsatzmöglichkeit bietet auch das Kundenmanagement. Beispielsweise nutzen Telekommunikationsunternehmen Text Mining, um Kundenanfragen oder Beschwerden schneller zu identifizieren und zu bearbeiten.

Da über 80% der Unternehmensinformationen in Textform vorliegen, gehen im Rahmen des Wissensmanagements Unternehmen wie der Automobilhersteller Peugeot oder der Pharmakonzern AstraZenca dazu über, möglichst viele Wissensquellen in nutzbare Informationen umzuwandeln. Dieser Prozess befindet sich allerdings noch in einer Entwicklungsphase[Horny].

5.7 Anwendung im Pharmagroßhandel

Die bisherigen Beispiele zeigen verschiedenste Einsatzmöglichkeiten des Data Mining in der Praxis, bei denen gute Ergebnisse erzielt wurden. Doch eine Diplomarbeit, die am „Institut für Parallele und Verteilte Höchstleistungsrechner“ der Universität Stuttgart in Zusammenarbeit mit der „Gehe Pharmahandel GmbH“ geschrieben wurde, zeigt auch die Probleme die beim Einsatz von Data Mining auftreten können.
Die „Gehe Pharmahandel GmbH“ ist der zweitgrößte Pharmagroßhandel in Deutschland. Das Unternehmen übernimmt die Verteilungs- und Lagerfunktion zwischen der herstellenden Industrie und den Apotheken. Im Rahmen der Diplomarbeit sollten mit Hilfe von kommerzie ller Software der Datenbestand des Data Warehouses auf neue Erkenntnisse untersucht und herausgefunden werden, ob sich Data Mining in dieser Branche lohnt und neue Ergebnisse liefern kann.
Zuerst wurden die vorhandenen Probleme und offenen Fragestellungen gemeinsam mit dem Marketing und der Vertriebsunterstützung analysiert. Man erkannte folgende Themen als potentielle Einsatzgebiete von Data Mining Verfahren.
Zum ersten die Klassifikation von erfolgreichen bzw. umsatzstarken Kunden (Apotheken), um durch gezieltes Marketing eine Erhöhung der Bestellungen bei Gehe zu erreichen. Diese Identifikation könnte man anhand von demografischen Merkmalen, wie beispielsweise Anzahl der Einwohner, Ärzte und Krankenhäuser im Umkreis durchführen. Eine andere Möglichkeit ergibt sich durch die Analyse der kundenspezifischen Merkmale wie z.B. Alter des Apothekers, Anzahl der Angestellten, Qualifikation der Angestellten...
Als weiteres Einsatzgebiet sah man das Clustering von Apotheken in sortimentspezifische Kundengruppen, um die Apotheken zukünftig zielgerichteter ansprechen zu können. Eine Erkennung von Sortimentschwerpunkten anhand von geographischen und demographischen Merkmalen erschien interessant, genauso wie die Analyse von saisonalen Schwerpunkten.
Durch eine Warenkorbanalyse könnte man nützliche Ergebnisse für die Bildung von Aktionspaketen im Marketing oder im Einkauf gewinnen. Ebenso könnten die Ergebnisse der Warenkorbanalyse zur Optimierung der Lagerhaltung beitragen.
Außerdem betrachtete man eine Analyse des Kundenverhaltens beim Onlinedienst der Gehe als interessant. Zum einen könnte man die Segmentierung der Kunden bezüglich der Verwendung des Onlinedienstes, wie Informations- oder Bestellmedium, untersuchen. Genauso würde auch die Erkennung von Themenschwerpunkten zu einer Verbesserung des Onlinedienstes beitragen.
Als letzter Punkt wurde dann noch die Klassifizierung von zahlungsunfähigen Kunden in die Liste der möglichen Einsatzbereiche aufgenommen. Falls es möglich wäre Indikatoren zu finden, die eine Apotheke mit Zahlungsschwierigkeiten kennzeichnen, könnte man die Lieferungen an diese Apotheke schon vor der wirtschaftlichen Notsituation einstellen und somit die zum Teil sehr beträchtlichen Zahlungsausfälle reduzieren.
Als Datenbestand waren zum einen die Informationen aus dem eigenen Data Warehouse und Informationen des Institutes für Medizinische Statistik verfügbar.
Bei der Auswahl der durchzuführenden Projekte wurden die ersten praktischen Probleme deutlich. Die Identifikation von erfolgreichen Kunden anhand von kundenspezifischen Merkmalen konnte wegen dem Mangel an Informationen nicht durchgeführt werden. Bei einzelnen Niederlassungen waren die benötigten Daten nur zu 20% vorhanden oder veraltet. Bei der Analyse von zahlungsunfähigen Kunden bestand das Problem darin, dass die Daten des Buchhaltungssystems damals noch nicht im Data Warehouse integriert waren und so der Zeitaufwand für die Datenaufbereitung zu hoch gewesen wäre. Auch auf die Untersuchung des Onlinedienstes wurde verzichtet. Um das Navigationsverhalten der Kunden aufzeichnen zu können, hätte man zuerst die technischen Voraussetzungen schaffen müssen. Somit entschied man sich für die Durchführung der Warenkorbanalyse, der Klassifikation von umsatzstarken Kunden aus demografischen Merkmalen und des Clusterings der Kunden anhand der Sortimentzusammensetzung.

Bei der Durchführung des Kundenclusterings anhand des Sortiments wurden zuerst die Umsatzzahlen der verschiedenen Apotheken in bestimmten Warengruppen auf ihren Zusammenhang mit demografischen Faktoren untersucht. Allerdings konnte auf dieser Ebene keine sinnvolle Gruppeneinteilung gefunden werden. Daraufhin wurde die Untersuchung nur bei einer Niederlassung und auf weniger und somit größeren Warengruppen durchgeführt. Nun konnten drei Kundengruppen identifiziert werden. Es stellte sich heraus, dass 20% der Gehe-Kunden über alle Warengruppen hohe oder mittlere Umsätze verbuchen. Die restlichen 80% der Kunden verbuchen über alle Warengruppen nur niedere Umsätze. Es ergab sich also keine Einteilung nach Warengruppenschwerpunkten, sondern nur nach umsatzstarken und umsatzschwachen Kunden. Die fehlenden Zusammenhänge in den Daten zwischen Warengruppenumsatz und demografischen Faktoren wurden auf die mangelhafte Genauigkeit der Daten zurückgeführt. Denn es lagen nur die Informationen über die Fachärzte in Gebieten vor, die mehrere Postleitzahlenbezirke umfassen. Für den Verkauf von Medikamenten einer Apotheke ist aber wahrscheinlich nur die Anzahl von Fachärzten der verschiedenen Bereiche in einer sehr kleinen Umgebung entscheidend.
Danach wurden die Daten auf saisonale Schwerpunkte analysiert. Das Resultat lieferte auch einige Schwerpunkte, allerdings waren es keine grundlegenden neuen Erkenntnisse, sondern nur Bestätigungen von bekannten Zusammenhängen. Die Analyse lieferte beispielsweise folgende leicht erklärbaren Resultate: Die hohen Umsätze von Grippemitteln in Herbst und Winter ließen sich durch das vermehrte Auftreten dieser Erkrankungen in diesen Monaten erklären. Der höhere Umsatz von Asthmamitteln in den Sommermonaten durch die erhöhte Pollen- und Ozonbelastung war zuvor auch schon bekannt. Zumindest konnten Apotheken gefunden werden, die diesen saisonalen Trends besonders deutlich folgen.
Bei der Warenkorbanalyse lagen mehrere Millionen Datensätze vor, so dass man zuerst die Testmenge einschränken musste, da die verschiedenen Softwaretools diese Datenmengen nicht bearbeiten konnten. Man untersuchte nun nur die umsatzstarken Kunden, die nicht apothekenpflichtigen Produkte und den Zeitraum von zwei Monaten. Auch die Ergebnisse der Analyse waren nicht von allzu großer Bedeutung. Zwar gab es Artikel, die mit großer Häufigkeit gemeinsam bestellt wurden, aber es handelte sich hier meist um Artikel die nur sehr selten bestellt wurden. Die Zusammenhänge waren also nicht geeignet, um Maßnahmen im Lager- oder Logistikbereich zu ergreifen. Auch für den Marketingbereich waren die Ergebnisse nicht allzu aufschlussreich.
Bei der Klassifikation von Kunden mit hohem Umsatzpotential anhand von demografischen Daten konnte auch nicht direkt auf die Daten des Data Warehouses zugegriffen werden. Da eine Apotheke bei mehreren Niederlassungen Bestellungen tätigen konnte und in jeder Niederlassung eine andere Kundennummer vorlag, wurde im Data Warehouse jeder Apotheke eine Hauptnummer vergeben. Diese eindeutige Zuordnung konnte leider doch nicht benutzt werden, da bei einer großen Anzahl von Apotheken diese Hauptnummer nicht eingetragen war. Man konnte allerdings für diese Analyse die eindeutige Nummer des Bundesgesundheitsministeriums verwenden.
Das Umsatzpotential einer Apotheke setzt sich aus dem Umsatz bei Gehe und dem Umsatz bei anderen Pharmagroßhändlern zusammen. Der Umsatz bei der Gehe lag als numerischer Wert vor. Als prozentualen Umsatzanteil der anderen Anbieter diente eine Schätzung der Vertriebsmitarbeiter. Aus den Daten dieser Apotheken wurde nun der Entscheidungsbaum in Abhängigkeit von den demografischen Merkmalen generiert und anhand von Testdaten geprüft. Doch auch hier konnte trotz mehrmaliger Verbesserungen und Modellanpassungen nur eine maximale Genauigkeit der Vorhersage von 33% generiert werden. Auch hier wurde für das schlechte Resultat die mangelhafte räumliche Genauigkeit der demografischen Bezirke verantwortlich gemacht.
Insgesamt wurde in der Arbeit festgestellt, dass es Möglichkeiten für Data Mining auch im Pharmagroßhandel gibt und auch Resultate erzielt wurden. Aber es wurde ebenso betont, dass die Datenqualität und die kundenspezifischen Daten nicht in der benötigten Form vorliegen, um weitere Zusammenhänge analysieren zu können [Kimmerle].

6 Probleme des Data Mining in der Praxis

6.1 Probleme

6.1.1 Datenqualität

Wie schon oben erwähnt, ist die Qualität der Daten wesentlich für das Ergebnis des Data Mining. Allerdings werden jedoch in der Praxis Data Mining Verfahren teilweise direkt auf die Datenbanken der Unternehmen angewandt, ohne davor die Daten in entsprechender Weise zu bearbeiten. In diesem Fall erscheint es unwahrscheinlich, neues nützliches und relevantes Wissen zu entdecken. Ein typischer Mangel der in Datenbanken vorliegenden Daten, der die Anwendbarkeit von Data Mining Verfahren einschränkt, ist das Fehlen von relevanten Daten und somit die mangelnde Repräsentativität der Daten. Noch problematischer ist das systematische Fehlen von Daten einer bestimmten Klasse, da so das Ergebnis sicherlich verfälscht wird. Auch wenn in den vorliegenden Datenbeständen wichtige Variablen nicht erfasst wurden, ist ein aussagefähiges Ergebnis des Data Mining Verfahrens kaum zu erwarten. Zudem können laufende Veränderungen von Strukturen in den datengenerierenden Prozessen zu falschen Aussagen führen.

6.1.2 Softwarequalität

Durch die höhere Verfügbarkeit von Datenmengen in Datenbanken, stieg in den letzten Jahren auch die Nachfrage nach Software zur Analyse dieser Daten. Inzwischen gibt es zahlreiche Anbieter für Data Mining Software. Allerdings ist die Anwendung von vorgefertigten Algorithmen auf konkrete Probleme ohne genauere Betrachtung wohl kaum sinnvoll. Auch sollten Data Mining Algorithmen auf extrem große Datenmengen angewandt werden können. Hier ergeben sich in der Praxis von kommerziellen Softwarepaketen jedoch häufig Speicherprobleme oder lange Laufzeiten.

6.1.3 Aussagekraft der Ergebnisse

Da die von Data Mining Verfahren generierten Aussagen keine wahren Gesetzmäßigkeiten, sondern nützliche Ergebnisse für das analysierte Problem sind, kann man die Ergebnisse nicht beliebig verallgemeinern oder auf andere Situationen übertragen. Das bedeutet, dass man trotz des hohen Aufwands nur Aussagen für ein spezielles Problem bekommt, und bei ähnlichen Problemen oder anderen Umständen seine Ergebnisse nicht oder nur nach näherer Betrachtung einsetzten kann. Zudem können Fehlaussagen durch mangelnde Kenntnis der Daten und fehlender Modellüberprüfung entstehen und danach als erwiesene Fakten angesehen werden.

6.1.4 Datenschutz

Mit der wachsenden Verfügbarkeit von elektronisch gespeicherten Daten steigen neben den Anwendungsmöglichkeiten für Data Mining Verfahren auch die Gefahren für missbräuchliche Verwendungen der Daten. Zum Beispiel würde die Verknüpfung von Individualdaten einen Verstoß gegen den Datenschutz darstellen [hudec].

6.2 Verbesserungsmöglichkeiten

Um den Prozess des Data Mining zu verbessern und die Daten in homogener Form vorliegen zu haben, verwenden die meisten Großunternehmen Data Warehouse Lösungen. Nach einer Studie der Meta Group überstiegen die Ausgaben für Data Warehouse Lösungen im Jahr 2000 der „Global-3000"- Unternehmen im Mittel drei Millionen Euro [Meta].
Unter einem Data Warehouse versteht man eine von den operativen Datenverarbeitungssystemen getrennte Datenbank, die einen effizienten Zugriff auf Informationen von verschiedenen Informationsquellen erlaubt und als unternehmensweite Datenbasis dient [Mantel].

Da ein Großteil der Zeit und Kosten bei einem Data Mining Projekt die Phase der Datenvorbereitung einnimmt, erweist sich der Zugriff auf die schon konsolidierten und bereinigten Daten eines Data Warehouses günstiger als der Zugriff auf Rohdaten [KnobWeid].

7 Zusammenfassung

Insgesamt zeigen die Praxisanwendungen, dass Data Mining prinzipiell zur Informationsgewinnung einen wertvollen Beitrag leisten kann. Vor allem die Ergebnisse im CRM und Marketing, sowie die Studien zum Einsatz von Data Mining in der Unternehmenswelt verdeutlichen die Chancen, die Data Mining Projekte bieten können.
Bei der Recherche wurde zudem festgestellt, dass mehr Anwendungen in der Praxis branchenübergreifend dem Aufgabentyp Klassifikation als dem Clustering zugeordnet werden konnten. Das liegt sicherlich an den vielseitigeren Einsatzmöglichkeiten der Klassifikationsanalysen. Während beim Clustering hauptsächlich die Fragestellung der allgemeinen Gruppeneinteilung von Kunden untersucht wird, gibt es bei den Klassifikationsanalysen durch die genaue Vorgabe der Zielvariablen mehrere interessante Analysemöglichkeiten. Die Anwendungen von Assoziationsanalysen werden hauptsächlich im Handel durchgeführt. Über die tatsächliche prozentuale Verbreitung der einzelnen Aufgabentypen in der Praxis kann jedoch nur schwer etwas gesagt werden, da keine Studien zu dieser Fragestellung vorliegen.
Bei allen drei Aufgabentypen wurden in der Praxis erfolgreiche Projekte durchgeführt. Allerdings bleibt festzuhalten, dass die Aussagekraft der Klassifikationsanalysen insgesamt höher ist als beim Clustering oder der Assoziationsanalyse. Durch die genaue Zielsetzung und die leichtere Überprüfbarkeit der Ergebnisse konnten hier konkrete Verbesserungsmaßnahmen erzielt werden. Beim Clustering dagegen bedeutet beispielsweise eine sinnvolle Kundengruppeneinteilung noch nicht, dass auch eine Umsatzsteigerung durch verbesserte Kundenansprache erreicht wird.
Außerdem zeigt das Beispiel aus dem Pharmagroßhandel deutlich, wie wichtig geeignete Daten, die in passender Form vorliegen, sowie geeignete Software und die nötigen technischen Voraussetzungen für den Erfolg eines Data Mining Projektes sind. Ansonsten ist es möglich, dass man die gewünschten Untersuchungsergebnisse kaum verwirklichen kann.

Literaturverzeichnis

[Hudec] Hudec, Marcus: Data Mining – Ein neues Paradigma der angewandten Statistik
http://www.statistik.tuwien.ac.at/oezstat/ausg021/papers/hudec.doc
Abruf am 15.01.04

[Bristol] Universität Bristol, Abteilung Maschinelles Lernen
http://www.cs.bris.ac.uk/Research/MachineLearning/Kepler/en/introduction.html
Abruf am 15.01.04

[IBMNews] IBM eNews, Ausgabe 06 2001
http://www-5.ibm.com/de/software/enews/essay/2001-06-15-ess-1.html
Abruf am 18.01.04

[Golem] Golem Newsletter, Ausgabe 15.08.2002
www.golem.de/0208/21218.html
Abruf am 04.01.04

[NHConsult] NHConsult, Data Mining
http://www.nhconsult.de/images/nhc_dm.pdf
Abruf am 18.01.04

[KnobWeid] Knobloch, Weidner: Eine kritische Betrachtung von Data Mining-Prozessen
http://pda15.seda.sowi.uni-bamberg.de/ceus/papers/%5BKnWe00%5D.pdf
Abruf am 04.01.04

[GrobBens] Grob, Bensberg: Das Data Mining-Konzept
http://www.wi.uni-muenster.de/aw/publikationen/CGC8.pdf
Abruf am 05.01.04

[Exper] Dr. Rieger: Der Aufwand zahlt sich aus, Fachartikel aus Experpraxis 99/2000
http://www.experteam.de/startd/publikationen/Artikel/Ber578.html?Themen+Datawarehouse
Abruf am 22.12.03

[Feldkirch] Feldkirch: Zwischen Goldesel und Sternschnuppe, SPSS in der Praxis
http://www.spss.com/de/praxis/CRISP.PDF
Abruf am 22.12.03

[Netlex] net-lexikon, Onlinelexikon
http://www.net-lexikon.de/Customer-Relationship-Management.html
Abruf am 20.12.03

[Publi] Publimax
http://www.publimax.de/Glossar/customer_relationship_management.html/
Abruf am 20.12.03

[Dataeng] Data Engine: 3 Anwendungen von Data Mining mit Intelligenten Technologien
http://www.dataengine.de/german/it/datenan23933/seite3.htm
Abruf am 03.01.04

[vonLühe] von Lühe: Neue Methoden der Zielgruppensegmentierung
http://www.spss.com/de/praxis/unicef.pdf
Abruf am 05.12.03

[Sailer] Sailer: Zielgerichtet Kundenansprache als Verkaufsmotor
http://www.spss.com/de/praxis/rlb.pdf
Abruf am 16.01.04

[SPSSPoli] SPSS in der Praxis: Westmidlands Police Department
http://www.spss.ch/_pdf/WestmidlandsPoliceData.pdf
Abruf am 06.01.04

[Kimmerle] Kimmerle: Data Mining im Pharmagroßhandel, Diplomarbeit
http://elib.uni-stuttgart.de/opus/volltexte/2000/719/pdf/DIP-1821.pdf
Abruf am 15.01.04

[Meta] Metagroup: Business Intelligence und Data Warehouse 2002
http://www.metagroup.de/studien/2002/businessintelligence
Abruf am 18.01.04

[Mantel] Stephan Mantel: Einführung in das Data Warehouse-Konzept
http://www.mik.com/WebSite/MIKWebArchiv.NSF/PDF/DataWarehouse/$File/DataWarehouse.pdf
Abruf am 18.01.04

[Horny] Dr. Alexandra Horny, SPSS in der Praxis
http://www.spss.com/de/praxis/Textmining.pdf
Abruf am 20.01.04