

Data Mining-Tools – Eine neue Ära für das Data Mining ?

- Artikel in der Computerwoche -

Dr. Peter Gentsch – Business Intelligence Group

Einleitung

Für den Data-Mining-Markt wird von der Meta Group für das laufende Jahr ein Umsatz in Höhe von circa 9 Milliarden US-Dollar geschätzt. Die Gartner Group kommt in ihren Untersuchungen zu der Einschätzung, dass noch dieses Jahr mindestens 50 Prozent aller Fortune 1000 Companies Data-Mining-Technologien nutzen werden. Da dieser attraktive und zukunftssträchtige Markt von vielen Herstellern der Computerbranche adressiert wird, existiert eine kaum noch überschaubare Anzahl von Software-Werkzeugen für Data Mining.

Der inflationäre und meist wenig qualifizierte Umgang mit Thema Data Mining hat zu zahlreichen Spekulationen, übersteigerten Erwartungen und unseriösen Versprechungen seitens der Softwareanbieter geführt. Dies birgt die Gefahr in sich, dass eine Methode, die bei einem fundierten und zielgerichteten Einsatz eine große Hilfe bei der Bewältigung der Informationsflut bieten kann, zu Unrecht zu Grabe getragen wird. Denn Data Mining hat enorme Nutzenpotenziale in verschiedenen Anwendungsfeldern (zum Beispiel Controlling, Zielkunden-Marketing oder Innovationsmanagement) und für Business Szenarien (zum Beispiel Warenkorbanalysen, Identifikation abwanderungsgefährdeter Kunden, Cross-/Up-Selling-Potenziale oder Prozessanalysen).

Eine neue Vergleichsstudie, die von Peter Gentsch als Hauptautor in Zusammenarbeit mit dem Business Application Research Center (BARC) erstellt wurde, untersucht zwölf verschiedene am Markt angebotene Data-Mining-Lösungen anhand eines differenzierten Bewertungsschemas.

Motivation für den Einsatz von Data Mining

Es wird geschätzt, dass sich die weltweit vorhandene Informationsmenge alle 20 Monate verdoppelt. Gerade in den informationsintensiven Branchen stellt die schnelle und zielgerichtete Auswertung dieser Daten einen erfolgskritischen Faktor in einem dynamischen und kompetitiven Umfeld dar. Die betriebliche Praxis zeigt jedoch, dass die umfangreich gespeicherten Daten nur rudimentär bzw. oft nur für das operative, kurzfristige Tagesgeschäft genutzt werden.

Lediglich 5 bis 10 Prozent der in Unternehmen gesammelten und generierten Daten werden analysiert. Damit bleibt eine im Unternehmen bereits vorhandene Werte weitgehend ungenutzt. Fehlendes Know-how und knappe zeitliche Ressourcen in Unternehmen sind häufig genannte Gründe für die ausbleibende Analyse der Daten.

Neben diesen Pull-Kriterien sind die gestiegene Performance von Data-Mining-Infrastrukturen, die drastisch steigenden Datenmengen sowie die Vorstrukturierung von Daten im Rahmen von Data-Warehouse-Projekten die Push-Faktoren für das Data Mining. Data Mining soll als innovatives Konzept zur Bewältigung dieser Problematik verstanden werden, indem es die computergestützte Analyse von umfangreichen Datenbeständen ermöglicht.

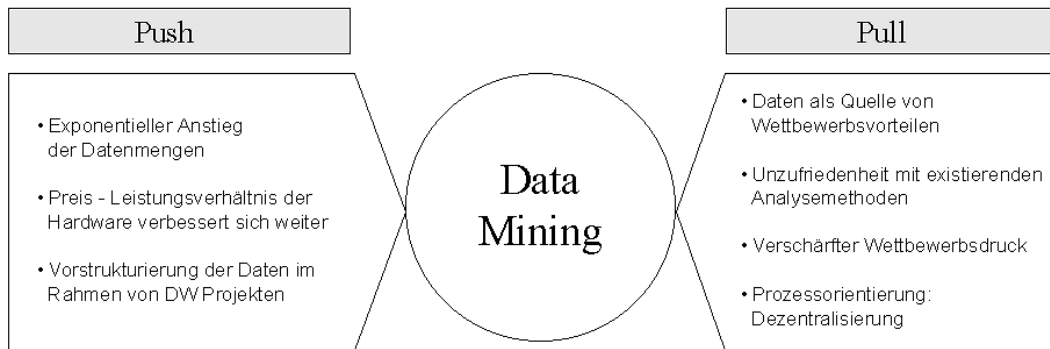










Abbildung 1: „Push“ und „Pull“ beim Data Mining

12 Data Mining-Tools im Vergleich

Zwölf verschiedene marktgängige Data-Mining-Tools in der Preisklasse von 600 bis 250.000 € werden in der Studie anhand eines differenzierten Bewertungsschemas in Bezug auf Firmenprofil, technische Spezifikationen, Datenmanagement, Preprocessing, Data Mining Funktionalität, Output Features und Besonderheiten untersucht. Die in Abbildung 2 aufgeführten 12 Hersteller von Data Mining-Tools erhielten einen Fragebogen zur Produktbeschreibung stellten ihr Tool temporär für einen Testzeitraum zur Verfügung. Zur Klärung von Detailfragen zu den Produkten wurden Expertengespräche mit den Herstellern geführt. Es standen verschiedene Real- und Simulationsdatensätze zur Verfügung, die sich hinsichtlich der Anzahl der Datensätze, Anzahl der Attribute, der Skalierungen (Nominal-/Ordinal- sowie Intervall-/Ratio-Skalierung), der Anzahl von Missing Values sowie dem Grad des Rauschens unterscheiden.

 	Client ca. 40.000 €; Client/server ab 75.000 €
 SAS/Enterprise Miner	Ca. 75.000 – 250.000 €, abhängig von der Serverkonfiguration
 ORACLE Darwin <small>SOFTWARE POWERS THE INTERNET™</small>	Ca. 40.000 €
 DataEngine	Terminalserver inkl. 3 Clients 30.000 € Einzelplatzlizenz 2.500 €
 	Ca. 60.000 €
 humanIT <small>Human Information Technologies GmbH</small> InfoZoom®	5 zeitlich unbegrenzte Lizenzen ca. 30 T€, 3 Jahre Support ca. 10 T€, Rabattstaffelung ab 5 Lizenzen




 DIALOGIS <small>SOFTWARE & SERVICES GMBH</small> D-MINER	Ca. 30.000 €
 COGNOS Scenario 	Ca. 1.500 €
 BUSINESS OBJECTS BusinessMiner	Ca. 700 €
 BISSANTZ DeltaMiner	Ca. 10.000 bis 40.000 €
 ANGOS Knowledge STUDIO <small>KNOWLEDGE ENGINEERING</small>	Ab 21.000 €
 prudsys <small>the power of data mining</small> DISCOVERER	Ca. 6.000 €

Abbildung 2: Data Mining-Tools

In der Studie wird der häufig nicht thematisierte Zusammenhang von betriebswirtschaftlicher Problemstellung und entsprechender Tool- bzw. Methodenwahl aufgezeigt. Die Tools werden daraufhin untersucht, in welcher Form und Intensität sie den gesamten Prozess des Data Mining unterstützen. Konzentriert sich die Unterstützung nur auf die eigentliche Mining-Phase, müssen zusätzliche Tools wie Datencleaning-Tools oder Report-Generatoren ergänzend zu den Data-Mining-Tools zum Einsatz kommen.

Um dem Anwender bei der Auswahl eines geeigneten Data-Mining-Tools einen möglichst großen Lösungsraum zu präsentieren, wurden bewusst Micro-Mining- und Macro-Mining-Tools bzw. Low-Budget- und High-End-Tools gemeinsam betrachtet. Je nach Hintergrund und Zielsetzung des Anwenders, je nach Dauer und Intensität des Data-Mining-Einsatzes sowie je nach Verfügbarkeit humaner, finanzieller und datenbezogener Ressourcen kann der Anwender „sein“ Tool auswählen.

Ergebnis

Der Mythos des Data Mining als reine „Plug and Play“-Lösung, die vollständig autonom interessante Muster in Datenbanken findet, muss entkräftet werden. Data Mining ist zu komplex, als dass es sich auf Knopfdruck durch ein Tool durchführen ließe. In der praktischen Anwendung der Tools für verschiedene Fragestellungen zeigt sich, dass Data Mining eine Kombination aus Verifikationsmodell und Entdeckungsmodell ist. Ist der Lösungsraum zu einem bestimmten Grad händisch aufbereitet worden, kann das Tool mit einem hohen Autonomiegrad diesen Lösungsraum nach interessanten Auffälligkeiten untersuchen. Abbildung 3 klassifiziert die verschiedenen Data Mining-Tools gemäß den Dimensionen „Ability to Execute“ und „Completeness of Vision“.

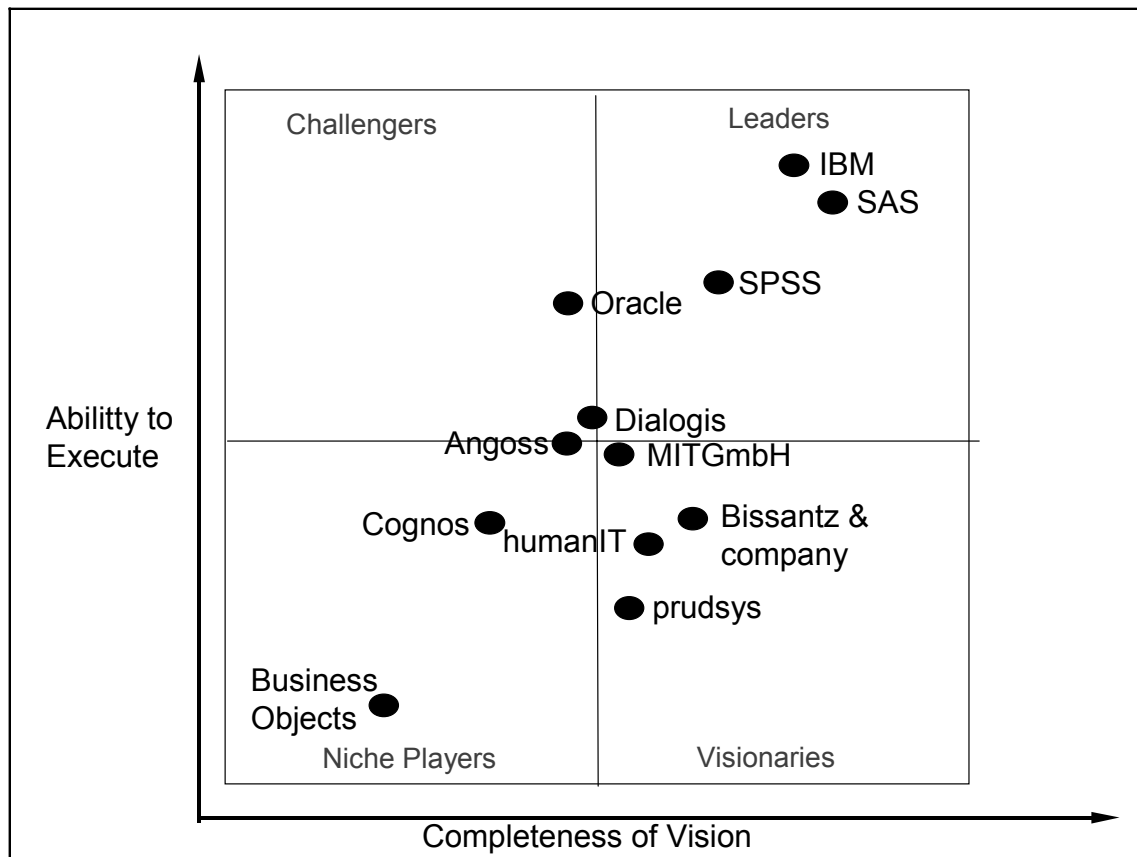


Abbildung 3: Klassifikation der verschiedenen Data-Mining-Unternehmen

Wie die Ergebnisse der BARC-Studie weiterhin zeigen, hat sich im Vergleich zu den ersten Tool-Generationen der Autonomiegrad, mit dem die Tools in der Mining-Phase die Datensätze nach interessanten Mustern durchsuchen, deutlich erhöht.

Als wesentliche Entwicklung der Data-Mining-Software kann die Integration von immer mehr Methoden, insbesondere auch grafische Methoden sowie die Einführung fensterbasierter Oberflächen auch bei den früher „kryptischen“ High-End-Tools gesehen werden. Hierbei muss kritisch angemerkt werden, dass die fensterbasierte Oberflächentechnik die Anwendung der Data-Mining-Systeme zwar bequemer gestaltet, sie jedoch nicht in methodischer Hinsicht erleichtert. Fehlt beim Anwender das erforderliche methodische Wissen, birgt dies die Gefahr in sich, dass Methoden falsch auf korrekte Daten angewendet werden mit der Konsequenz, dass man bestenfalls Antworten auf nicht gestellte Fragen erhält.

Die Studie macht darüber hinaus deutlich, dass – obwohl alle getesteten Tools das Label „Data Mining“ tragen – erhebliche Unterschiede hinsichtlich Funktionsumfang, Benutzerführung und Vision bestehen. Zum Teil verfolgen die Unternehmen mehr die Strategie eines Produktgeschäftes, zum Teil eher die eines Beratungsgeschäftes. Insgesamt ist die Tendenz zu höherer Lösungsorientierung aus betriebswirtschaftlicher Sicht zu erkennen. Die Werkzeuge fokussieren immer mehr die zu bearbeitenden Business Cases in Form vordefinierter Templates, Makros und Analyse-Workflows und weniger konkrete Methoden und Algorithmen.

Als weiteres wichtiges Ergebnis muss beachtet werden, dass vor der Auswahl eines Data-Mining-Tools zunächst Klarheit über die betriebswirtschaftlichen Ziele und Aufgabenstellungen herrschen muss, die mit Data Mining angegangen werden sollen. Data Mining darf kein Selbstzweck sein und sollte nicht allein durch die DV- oder IT-Abteilung angestoßen werden. Der Antrieb zu Data Mining darf nicht allein die Faszination einer innovativen Technologie sein, sondern vielmehr der Wunsch, betriebswirtschaftlichen Problemstellungen auf den Grund zu gehen.

Ausblick: E-Business - Eine neue Ära für das Data Mining ?

„The Web analytics market grows from \$425 million in 2000 sales to a projected \$4 billion in 2004“ (David Rickard, 2001)

E-Business revolutioniert die Art und Weise, wie Produkte und Dienstleistungen vermarktet und vertrieben werden. Das Internet verändert die Einkaufsgewohnheiten der Verbraucher, und das dauerhaft. Fehlende Loyalität ist ein prägnantes Kennzeichen der Verbraucher im E-Commerce.

Dies bedeutet, dass Unternehmen ihre „ausgetrampelten“ Marketingpfade verlassen müssen, und neue Wege zum Kunden suchen müssen. Die Grundlage zum Erfolg und zur langfristigen Profitabilität des Online-Marketings ist neben einem ausgewogenem Kosten-/Nutzen-Verhältnis, die personalisierte Kundenorientierung. Der Grad der Kundenbindung im Online-Markt steht in direkter Relation zur Reflektion der individuellen Bedürfnisse der Kunden.

Das Internet bietet zum einen eine einzigartige Vielzahl von Informationen, die es ermöglichen, den Puls des Kunden in Realtime zu messen. Zum anderen stellen Privacy-Restriktionen und Datenvolumina besonders hohe Anforderungen an die Analyse von Kundendaten im Internet. Es stellt sich nun die Frage, inwieweit E-Business in dem Spannungsfeld zwischen Möglichkeiten und Restriktionen eine neue Ära für das Data Mining bedeutet.

Das intelligente und effiziente Gestalten und Managen von Internet-basierten Kunden- und Geschäftsbeziehungen erfordert zunehmend die systematische Analyse der zugrundeliegenden Daten.

Aufgrund der immensen Datenfülle kommen konventionelle, ‚manuelle‘ Analysen schnell an ihre Grenzen. Genau hier bietet sich der intelligente Einsatz des Computers an: Durch den Einsatz des Data Mining lassen sich trotz Datenflut Strukturen und Muster finden, die wichtigen Input sowohl für die strategische Ausrichtung des E-Business als auch für die konkrete Gestaltung der Personalisierung liefern können. Der hohe Automatisierungsgrad des Mining ermöglicht zudem die relativ einfache Erfassung von Veränderungen. Die Berücksichtigung der Dynamik in Kundenbeziehungen ist von großer Bedeutung. So verändern sich z.B. Kundenpräferenzen, das Nachfrageverhalten oder die adressierte Zielgruppe. Ein leistungsfähiges und flexibles Customer Relationship Management muß diese Veränderungen schnell aufnehmen, um dann mit einem angepaßten Angebot kundengerecht reagieren zu können.

Betrachtet man den Markt für Data Mining-Tools und den Markt für Personalisierungs-Tools (siehe nächste CW-Ausgabe ...) wird deutlich, dass die Grenze zwischen Herstellern von Personalisierungs-Tools und Data Mining-Tools zunehmend verschwimmen. So bewegen sich klassische Data Mining-Tool-Hersteller wie SAS oder SPSS zunehmend im Personalisierungsmarkt und umgekehrt nehmen Hersteller von Personalisierungslösungen wie ATG, Broadvision oder Epiphany, zunehmend Mining-Funktionalität in ihre Produkte auf.

Neben dem traditionellen Mining auf Basis strukturierter Daten werden für das E-Business im allgemeinen und für die Personalisierung im besonderen zunehmend Mining-Technologien interessant, die auch auf Basis unstrukturierter, textueller Daten arbeiten: Personalisierungs-Interaktionen im Internet erzeugen eine Vielzahl von sowohl strukturierten und unstrukturierten als auch von Content-bezogenen und Transaktions-orientierten Daten. Aus Sicht einer umfassenden Personalisierung werden neben den strukturierten Daten zunehmend die unstrukturierten, qualitativen Daten im Internet wichtig werden. So enthalten Web-Seiten, E-Mails sowie Äußerungen in Chats und Newsforen wertvollen Input zur Analyse von Kunden und deren Verhaltensweisen.

Insbesondere die Integration der verschiedenen Mining-Ansätze wird es zukünftig ermöglichen, systematisch automatisierte Daten- und Marktforschungsanalysen im E-Business durchzuführen zu können. Ziel dieser Analysen ist die möglichst umfassende individuelle Ansprache des Kunden, die Optimierung des Leistungsangebotes sowie die Erhöhung der Kundenzufriedenheit und -bindung. Der Kern des integrierten Data Mining-Einsatzes liegt in der Ausrichtung aller Aktivitäten entlang der digitalen Wertschöpfungskette an dem individuellen Profil und den individuellen Bedürfnissen des Kunden.

Mit zunehmenden internen und externen Daten- und Dokumentenvolumina wird sich das Potential des integrierten Mining für die Personalisierung weiter erhöhen. Die zunehmende Verbreitung des XML-Standards wird die Semantik der unstrukturierten Daten erhöhen und damit insbesondere die Anwendungsmöglichkeiten und ‚Entdeckungspotentiale‘ des integrierten Mining-Ansatzes im Internet deutlich vergrößern. Dies wird die Transparenz der Internetgemeinde ‚tierisch‘ erhöhen, was eine kritische Reflexion des gläsernen Kunden respektive Hundes zunehmend erforderlich macht: