Zeitschrift: Comtec : Informations- und Telekommunikationstechnologie = information and telecommunication technology

Herausgeber: Swisscom

Band: 78 (2000)

Heft: 4

Artikel: Data Mining : Schwerarbeit in Datenhalden

Autor: Achermann, Bernhard / Schmidt, Manfred / Loémbé, André

DOI: https://doi.org/10.5169/seals-876431

Exploration Programmes:
Corporate Technology explores
Future Telecommunications

Data Mining – Schwerarbeit in Datenhalden

Data Mining verspricht viel: bedeutende Einsparungen im Marketing durch Identifikation von Zielgruppen, bessere Kundenbindung durch präzise Vorhersage möglicher Kündigungen, Betrugserkennung durch Künstliche Intelligenz, und dies soll alles ganz einfach sein. Im Rahmen des Explorationsprojekts EP0025 wurde diese Technologie im Detail untersucht. Ein Team von Softwareingenieuren, Statistikern und Datenbankspezialisten hat in diesem Bereich Erfahrungen gesammelt, welche Swisscom heute in die Lage versetzen, diesen Versprechungen realistische Einschätzungen entgegenzusetzen. Dieses Wissen wird bereits in verschiedenen Bereichen innerhalb von Swisscom genutzt.

The Exploration Programme Customer Care (CC) and Service Management Platforms (SMP) deals with:
- methods and technologies supporting Customer Relationship Management (CRM), in particular technologies (Customer Contact Centre, Knowledge Management) enabling efficient customer touch points, with a strong focus on the e-channel (Web, e-mail);
- methods and technologies (Knowledge Discovery, Data Mining) that support the fast recognizing of market opportunities (e.g. Up/Cross-Selling) and of the customer behaviour (e.g. Churn Prediction, Customer Segmentation);
- the processes and the channels for provisioning and configuration of IP-VPN services with QoS are investigated and improved. An automation of service delivery is required to increase its accuracy and for reducing cost and delivery time;
- solutions for IP Billing will be delivered since innovative billing models for IP Services are crucial for successful introduction of new IP services.

With its Exploration Programmes, Corporate Technology is exploring telecommunication technologies and new service possibilities with a long-term view of 2–5 years. Further, the expertise built up in the course of this activity enables active support of business innovation projects.

Data Mining

BERNARD ACHERMANN, MANFRED SCHMIDT UND ANDRÉ LOÉMBÉ, BERN

Data Mining, das «Graben in Daten», ist ein noch wenig erforschter Bereich. Data Mining verbindet Methoden der Künstlichen Intelligenz und der Statistik und ermöglicht, aus den Informationen – abgelegt in sehr grossen Datenbanken – zu lernen. Vier Faktoren sind für die rasche Ausbreitung dieser Methoden verantwortlich:

- In den vergangenen Jahren wurde in vielen Unternehmen mit grossem Aufwand der Aufbau so genannter Data Warehouses vorangetrieben. Diese Datensammlungen sollten das «Gedächtnis» der Unternehmen sein. Sie enthalten riesige Datenmengen, die in TByte gemessen werden. Data Mining verspricht, diese Daten vermehrt für das Unternehmen nutzbar zu machen.
- Im Marketing hat sich in den 90er-Jahren das Konzept des 1-to-1-Marketing durchgesetzt, das nicht das Produkt, sondern die Kundenbeziehung ins Zentrum der Bemühungen setzt. Um aber den Massenmarkt mit personalisierten Angeboten versorgen zu können, ist eine sehr gute Kundensicht nötig. Data Mining bietet sich hier an, um Fragen zu beantworten wie: «Bei welchen Kunden lohnt sich Angebot xy?»
- Die gestiegene Leistungsfähigkeit der Computerhardware ermöglicht Data Mining buchstäblich für jedermann, denn GBytes von Daten können und werden heute in einem Laptop verwaltet und analysiert.
- Schliesslich stellt die vermeintliche Anonymität der Kundenbeziehung im Internet die Unternehmen vor die neue Herausforderung, aus dem Surfverhalten der Kunden auf das Kundenprofil und die Kundenprofitabilität zu schliessen. Das ist nur mit einem automatisierten Data-Mining-Prozess (click stream analysis) möglich.

Diesen Herausforderungen muss mit effizienten und skalierbaren Algorithmen begegnet werden. Skalierbarkeit heisst, dass die Rechenzeit linear mit dem Umfang des Inputs wächst. Die Verfahren der klassischen Statistik sind zwar prinzipiell anwendbar, brauchen aber sehr viel Rechenzeit, weil sie in der Regel nicht skalierbar und ihre Ergebnisse zum Teil nur schwer vermittelbar (p-Werte) sind. Statt optimaler Verfahren sind in diesem Bereich schnelle und praktikable Lösungen gefragt. In den letzten Jahren haben sich somit eine Handvoll Methoden herausgebildet, die heute in vielen Bereichen angewandt werden (Bild 1).

Der Begriff Data Mining suggeriert Bergbauaktivitäten in riesigen «Datensteinbrüchen». Informatikingenieure, angetan mit Schutzhelmen und «Datenpickeln», tauchen vor dem geistigen Auge auf und die Nase will Schweiss und Kohlenstaub gerochen haben; oder ist es wie im Dilbert-Comic (Bild. 2)? Erstaunliche und ungeahnte Zusammenhänge werden plötzlich ans Tageslicht gehoben. Sind es wahre Goldklumpen für das Business? Zuerst werden die gängigsten Verfahren und Produkte vorgestellt, die im Data Mining zum Einsatz kommen. Dann werden einige praktische Erfahrungen diskutiert

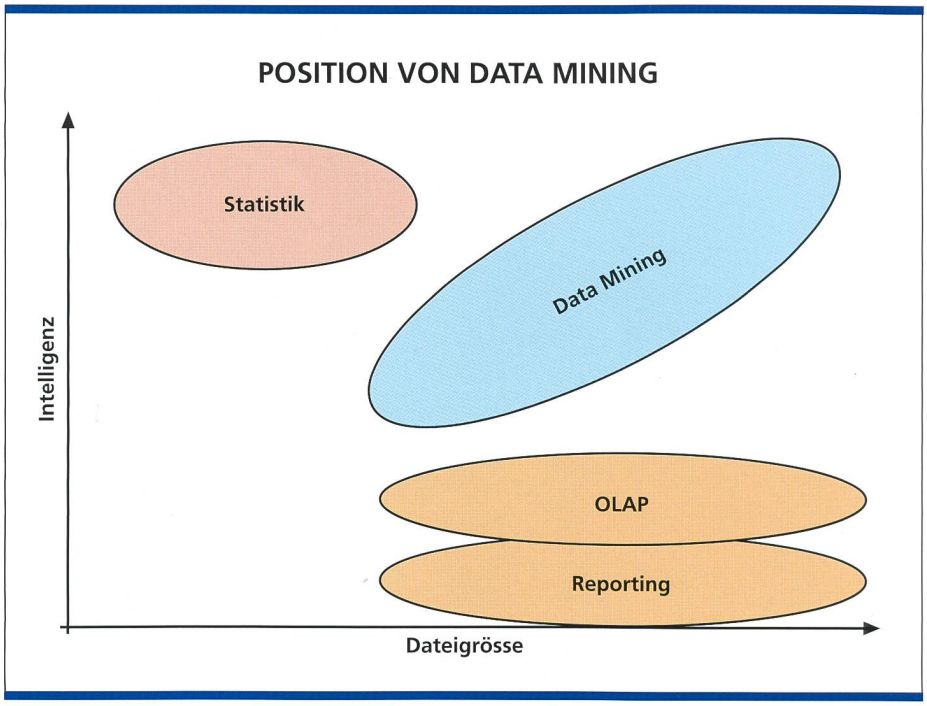

Bild 1. Während die Statistik mit relativ kleinen Datenmengen «intelligent» umgeht, vermögen so genannte Reporting- und OLAP-Tools auch einfache Analysefunktionen auf grosse Datenmengen anzuwenden. Data Mining schliesst die Lücke zwischen diesen Konzepten.

Bild 2. Data Mining at work?

und Erfolgsfaktoren oder Hemmnisse identifiziert. Zur Sprache kommt dann die Beziehung von Data Mining zum Datenschutz und abschliessend werden einige zukünftige Entwicklungen skizziert.

Data-Mining-Methoden

Wenn Data Mining die Suche nach geeigneten Informationen in grossen Datenbeständen ist, dann müssen auch entsprechend leistungsfähige Werkzeuge eingesetzt werden, Methoden, die es erlauben, grosse Datenmengen automatisiert und – zumindest ansatzweise – intelligent zu bearbeiten. Es sind dies in der Regel Algorithmen, die aus dem Bereich der Künstlichen Intelligenz und aus der Statistik stammen.

Assoziationsregeln

Eine der erfolgreichsten industriellen Anwendungen von Knowledge Discovery ist die «Warenkorbanalyse».
Mit der Einführung der Scannerkassen in Supermärkten Ende der 80er-Jahre entstand erstmals die Möglichkeit jede einzelne Transaktion maschinell zu erfassen. Neben der Optimierung der Warenflüsse stellte sich dann die Frage, wie man mit diesen Daten lernen kann, erfolgreicher zu verkaufen.
Zu diesem Zweck wurden Verfahren entwickelt, die so genannten Assoziationsregeln, aus denen sich Transaktionsdaten ableiten lassen. Eine solche Regel hat etwa die Form:
[Spargel, Schinken] => [Sauce hollandaise] (3,2%, 76%).
Diese Regel besagt, wer Spargel und Schinken kauft, nimmt auch Sauce hollandaise mit; das heisst genauer, 3,2% sämtlicher Kunden kaufen alle drei Produkte und 76% aller Kunden, die Spargeln und Schinken kaufen, greifen auch zur Sauce. In diesem Fall hat die Regel einen «Support» von 3,2% und eine «Confidence» von 76%. Eine Konsequenz dieser Erkenntnis kann in der Platzierung der Sauce neben den Spargeln oder dem Schinken liegen.
Algorithmen, die solche Regeln erzeugen, liefern dann alle Regeln, deren Support und/oder Confidence einen vorgegebenen Schwellenwert überschreiten. Das ist bereits eine sehr anspruchsvolle Aufgabe. Dies wird sofort einsichtig, wenn man von etwa 2000 Artikeln und einer grossen Anzahl Transaktionen ausgeht. In diesem Fall sind bereits 22 000 Warenkörbe möglich – eine sechshundertstellige Zahl. Daraus ergibt sich, dass die Algorithmen vor allem die unwichtigen, aber möglichen Regeln ausser Acht lassen müssen [10].
Die Verwendung von Kundenkarten (z.B. M-Cumulus) eröffnet zusätzlich die Möglichkeit, den Warenkorb einer bestimmten Person zuzuordnen, um so Massnahmen zu treffen, die auf bestimmte Einkaufsgewohnheiten abgestimmt sind.

Entscheidungsbäume (Decision Trees)

Klassifikationen im Alltag greifen oft auf eine gewisse Regelhaftigkeit zurück. Oft schliesst man aus Aussagen wie «Wenn jemand Gliederschmerzen, eine tropfende Nase, einen schmerzenden Hals und Fieber hat», dass er dann mit grosser Wahrscheinlichkeit eine Grippe erwischt hat. Solcher Art strukturiertes Wissen lässt sich in so genannten Entscheidungsbäumen darstellen. In den Knoten stehen jeweils Regeln, welche die Ausgangsmenge auf Grund gewisser Kriterien in geeignete Teilmengen zerlegen.
Der Aufbau von Entscheidungsbäumen kann eine langwierige und komplizierte Aufgabe sein, insbesondere wenn noch Randbedingungen dazu kommen, wie etwa die Minimierung der Stufen im Baum. Interessant wurden die Entscheidungsbäume eigentlich erst dadurch, dass es eine Reihe von Algorithmen gibt, die basierend auf einem Datenbestand und Zielmerkmalen automatisch Entscheidungsbäume aufbauen können. In Bild 3 ist ein Entscheidungsbaum dargestellt.
Wenn zum Beispiel ein Datensatz von neuen Kunden zur Verfügung steht, auf Grund dessen man Regeln bestimmen möchte, die beschreiben, welche Kunden lukrativ sein mögen, ist dies eine Aufgabe, die mit Entscheidungsbäumen angegangen werden kann. Zunächst muss bekannt sein, welche Kunden in der Vergangenheit lukrativ waren. Hier ist ausschliesslich Geschäftswissen gefragt.
Diese Entscheidung kann an keinen Algorithmus delegiert werden. Der Datensatz wird also für jeden bisherigen Kunden um die Information erweitert, ob er die Lukrativitätskriterien erfüllt oder nicht. Anschliessend wird der Datensatz dem Entscheidungsbaumalgorithmus übergeben mit dem Hinweis, dass die Lukrativität des Kunden das Zielmerkmal ist. Als Resultat liefert der Algorithmus einen Satz von Regeln, der eine Einteilung in lukrative und weniger lukrative Kunden nach bester Möglichkeit vornimmt. Der Entscheidungsbaum wird also mit den Daten «trainiert», bei denen bereits bekannt ist, ob der Kunde lukrativ war oder nicht. Das Ergebnis des Entscheidungsbaums sollte dann möglichst gut die vorgegebene Bewertung rekonstruieren.
Der so gewonnene Regelsatz kann nun zur Vorhersage eingesetzt werden, das heisst, er kann auf Fälle von Kunden angewandt werden, die nicht «trainiert» wurden. Es reicht zudem aus, nur einen relativ kleinen Ausschnitt aus der Gesamtmenge für Trainingszwecke auszu-

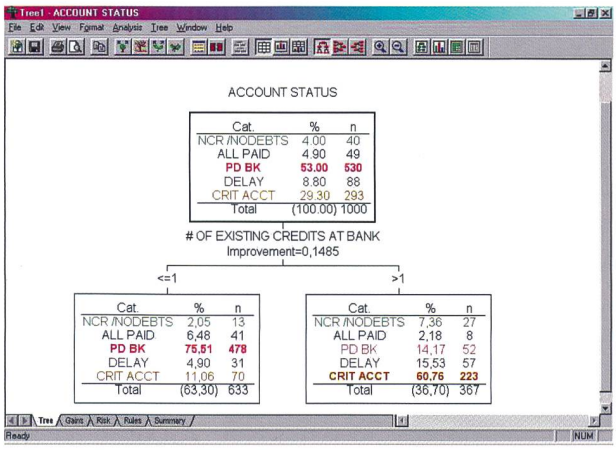

Bild 3. Ein einfacher Entscheidungsbaum: Aus einer Vielzahl von Faktoren wurde die Anzahl bestehender Kredite als entscheidender Einflussfaktor erkannt (AnswerTree, SPSS).

wählen. Bedingung ist lediglich, dass ein repräsentativer Ausschnitt aus der Kundensammlung verwendet wird. Der Algorithmus liefert in diesem Falle ein Modell für die Bestimmung der vorgegebenen Zielmerkmale.

Künstliche neuronale Netzwerke (Artificial Neural Networks)

Die Lern- und Abstraktionsfähigkeit des Menschen war schon seit jeher eines seiner auffälligsten und faszinierendsten Merkmale. Ganz offensichtlich kann der Mensch auf Grund von Beispielen generalisierte Konzepte ableiten (lernen), die er später in adäquaten Situationen wieder abrufen und verwenden kann. Noch erstaunlicher ist, dass diese komplexen Vorgänge offenbar auf sehr einfachen Bausteinen beruhen, den Nervenzellen, welche je nach Erregungspotenzial senden oder nicht senden. Die Vernetzung solcher einfachen Bausteine erlaubt es, ein hochkomplexes System zu bilden.
Ein (bescheidener) Versuch, diese Funktionsweise mit Computern nachzubilden, sind die sogenannten künstlichen neuronalen Netzwerke (Bild 4). Der Computer simuliert hierbei ein Netz von einfachen Neuronen. Diese Neuronen «feuern» bei bestimmtem Input, während sie bei anderem Input nicht «feuern». Die Bestimmung allerdings, wann ein Neuron «feuert», geschieht im Training, indem dem Netz eine Serie von Fällen vorgelegt wird, aus denen automatisch abgeleitet wird, wie das Netz zu reagieren hat (der gewünschte Output ist bei diesen Trainingsfällen bekannt). Die Art und Weise, wie diese automatische Ableitung vonstatten geht, wird etwa auch Lernregel genannt. Üblicherweise wird ein Trainingssatz einem neuronalen Netz mehrmals vorgelegt, bis eine weit gehende Adaption sichergestellt ist.
Zurück zum Beispiel mit den lukrativen Kunden: Wenn einem neuronalen Netz beigebracht werden soll, was ein lukrativer Kunde ist, dann ist das Vorgehen ganz ähnlich wie bei einem Entscheidungsbaum. Wiederum wird ein Satz von Kunden um die Information erweitert, ob es sich bei jedem Einzelnen davon um einen lukrativen Kunden handelt oder nicht. Dem neuronalen Netzwerk wird diese Sammlung von Daten zum «Lernen» präsentiert, mit dem Hinweis, dass es lernen soll, was ein lukrativer Kunde ist, welche Konstellation auf einen lukrativen Kunden hinweist.
Als Resultat des «Lernvorgangs» erhält man wiederum eine Art Modell eines lukrativen Kunden. Wenn dem neuronalen Netzwerk ein (bis anhin unbekannter) Kunde präsentiert wird, ist es in der Lage, mit einer gewissen Wahrscheinlichkeit zu ermitteln, ob es sich hierbei eher um einen lukrativen Kunden handelt oder nicht. Das neuronale Netz hat gewissermassen ein Modell davon entwickelt, wie ein lukrativer Kunde typischerweise aussieht.
Nachteilig bei neuronalen Netzwerken ist, dass die gelernte Information der menschlichen Anschauung kaum mehr zugänglich sind. Sie ist in Schwell- und Gewichtswerten im neuronalen Netz versteckt, ganz im Gegensatz zu den Regeln eines Entscheidungsbaumes, die dem menschlichen Denken nach wie vor zugänglich sind. Unter Umständen liefern aber neuronale Netze bessere prädiktive Ergebnisse als Entscheidungsbäume. Es lassen sich allerdings keine Regeln dazu aufstellen, wann welche Methode besser geeignet ist. Im Einzelfall sind beide einzusetzen und anschliessend zu vergleichen.

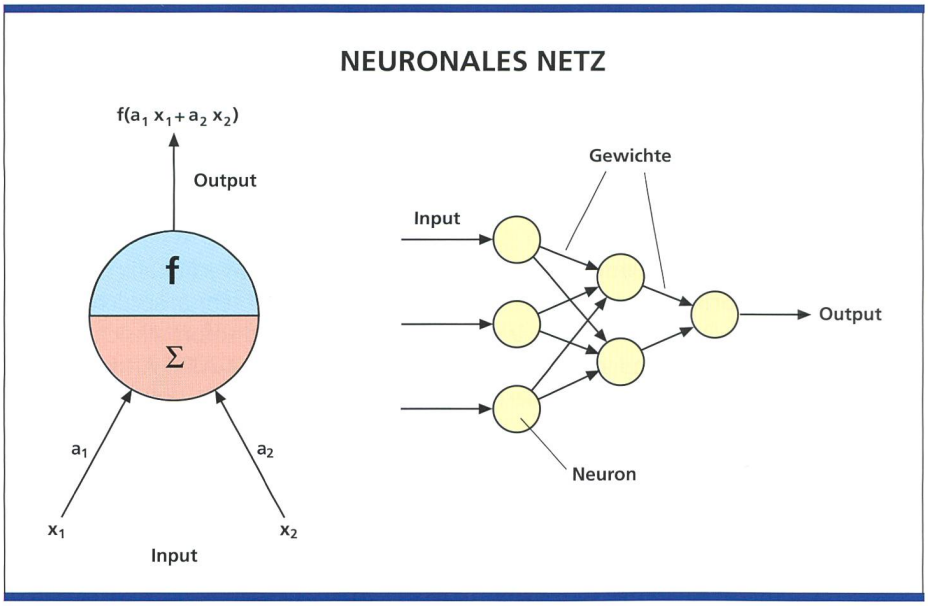

Bild 4. Links ein einzelnes Neuron: Die Eingangswerte (Input) werden gewichtet, aufaddiert und in eine Funktion eingesetzt. Der Output kann wieder als Input für andere Neuronen dienen. Durch Kombination von Neuronen entsteht ein neuronales Netz (rechte Seite).

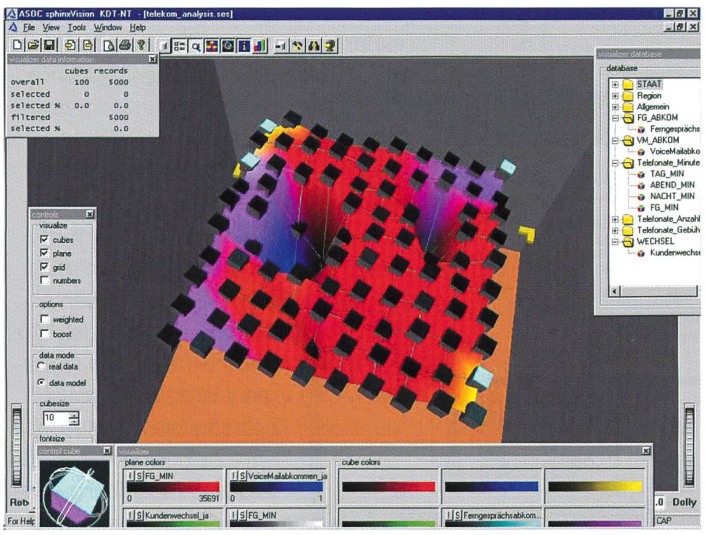

Bild 5. Die ganze Datenbank auf einen Blick: SphinxVision von ASOC.

Weitere Data-Mining-Methoden

Selbstverständlich ist die Menge an Werkzeugen, die für Data Mining zur Verfügung stehen, damit nicht abschliessend dargestellt. Namentlich gehören

noch Clusteringalgorithmen, genetische Algorithmen oder Memory Based Reasoning dazu. Allerdings kann aus Platzgründen nicht weiter auf diese Techniken eingegangen werden. Für weitergehende Informationen sei hierzu auf [1] verwiesen.
Neben diesen quantitativen Methoden ist die Visualisierung der Informationen entscheidend. Interaktive Grafiken können sowohl bei der Analyse als auch bei der Präsentation der Resultate beim Kunden sinnvoll eingesetzt werden. So erlaubt das Produkt «Mine Set» von Sillicon Graphics die Darstellung von Entscheidungsbäumen als dreidimensionale Landschaft, die überflogen werden kann. Dass das alles andere als eine Spielerei ist, wird schnell klar, wenn man versucht, einen komplexen Entscheidungsbaum zu verstehen oder anderen zu erklären. Ganz auf Visualisierung setzt das Produkt «SphinxVision» der Firma Asoc AG (Bild 5). Mit Kohonennetzen, einer Variante neuronaler Netze, wird die gesamte Datenbasis erfasst und abgebildet. Durch Manipulation der Grafik werden dann einzelne Kundensegmente identifiziert und können dann etwa in MS-Office-Produkten direkt verwendet werden.

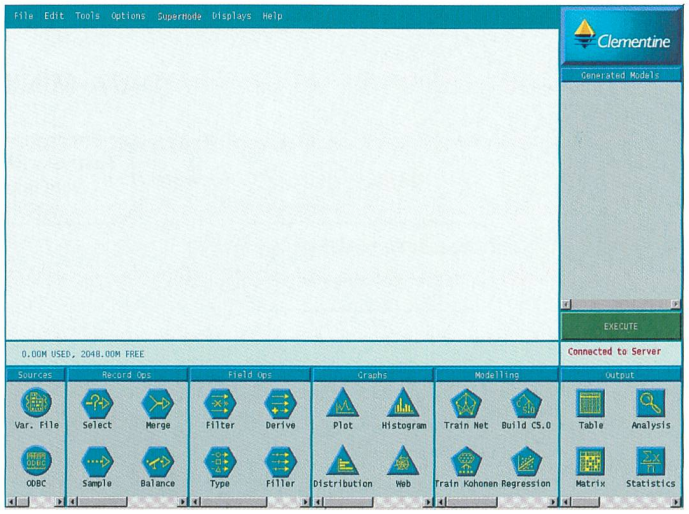

Bild 6. Ein «Stream» in Clementine 5.1. Erzeugte Modelle (gelbe Diamanten können als Knoten verwendet werden.

Tools für Data Mining

Ein kurzer Überblicksartikel kann keine vollständige Marktübersicht über Data Mining Tools anbieten. In der Folge seien aber die drei am häufigsten genutzten Produkte kurz dargestellt:

Clementine 5.1 (SPSS)

Das Produkt «Clementine» (Bild 6) wurde von ISL entwickelt und hat Massstäbe gesetzt; insbesondere bei der grafischen Darstellung der einzelnen Analyseschritte (Stream) und der Integration vieler Data-Mining-Algorithmen. Seit ISL im Juni 1999 von SPSS erworben wurde, wird das Produkt durch SPSS vertrieben und weiterentwickelt. Durch die Kombination mit dem Statistikprogramm «SPSS für Windows», das jetzt wie Clementine als Client/Server-Version vorliegt, ergibt sich eine vollständige Data-Mining-Plattform. Leider ist Clementine für Windows NT zurzeit nur als X-Windows-Emulationen erhältlich, was sich aber mit der Version 6 ändern soll.

Enterprise Miner 3.0 (SAS)

Aufbauend auf der bekannten SAS-Statistik-Suite bietet SAS ebenfalls ein Data Mining Tool an. Wie bei Clementine/SPSS entsteht durch die Kombination der Fülle der statistischen Funktionen mit den effizienten Algorithmen des Data Mining eine vollständige Data-Mining-Plattform. Auch der Enterprise Miner bietet einen grafischen Workflow an. Neben der grossen Zahl an Algorithmen, die in der Version 3 implementiert sind bietet SAS ein eigenes Data-Mining-Konzept an (SEMMA: Sample, Explore, Modify, Model, Assess), das mit dem EM realisierbar ist. Insbesondere für die ersten Schritte ist es aber notwendig, sich detailliert mit der SAS-Programmiersprache auseinander zu setzen. Eine ausführliche Schulung der Mitarbeiter für dieses Produkt ist ein Muss.

DB2 Intelligent Miner for Dat 6.1 (IBM)

Der IM 6.1 von IBM beinhaltet nicht so viele Algorithmen wie die ersten beiden Produkte. Dafür sind die implementierten Methoden einfach zu bedienen und unterstützen auch Mehrprozessorsysteme. Besonders gelungen ist die Visualisierung von Assoziationsregeln. Die Integration in die Datenbank vereinfacht das Datenmanagement. Allerdings ist die Installation nicht ganz so einfach. Dieses Produkt eignet sich für eine schnelle Automatisierung gängiger Data-Mining-Routinen.
Generell ist zu sagen, dass der Erfolg eines Data-Mining-Projekts nur in geringem Masse vom eingesetzten Tool abhängt. Wichtiger sind die Erfahrung der Mitarbeiter im Umgang mit einem bestimmten Produkt und die Möglichkeiten der Integration in die bestehende IT-Umgebung.

Data Mining in der Praxis

Grundlegend für die Anwendung von Data Mining ist eine Businessfrage und eine Kosten-Nutzen-Analyse. Da diese Methoden Geld zu sparen bzw. zu gewinnen versprechen, gilt es das unter Beweis zu stellen. Bei einer Erfolgskontrolle, bei welcher der Nutzen in Franken und Rappen beziffert wird, sollte immer versucht werden, auch wenn es nicht immer möglich ist, etwa den Kundenwert genau zu bestimmen. Was ist ein Neukunde wert? Ab welchem Umsatz wird ein Kunde rentabel?

Ablauf

In Bild 7 ist ein Data-Mining-Prozess dargestellt, der in einen Geschäftsprozess integriert ist. Der Prozess gliedert sich im Wesentlichen in drei Phasen (Set-up, Data Mining, Implementation), welche je wiederum eine Reihe von Unterprozessen umfassen. Im Folgenden werden diese Teilprozesse kurz skizziert.

Phase Set-up

In dieser Phase wird viel diskutiert. Da der gesamte Prozess in der Regel «quer» zum operativen Geschäft liegt und viele unterschiedliche Stellen involviert sind, sind Widerstände und Missverständnisse mehr die Regel als die Ausnahme. Simple Fragen wie «Was ist ein Modell?» oder «Was ist ein Kunde?» sind zu klären, werden aber oft nicht gestellt.
Wichtigstes Ergebnis der erste Phase ist die Entscheidung, ob mit der Analyse begonnen wird. Dazu muss vom Auftraggeber ein klares Ziel vorgegeben werden und im Rahmen einer Vorstudie muss geklärt werden, ob die Ressourcen (Personal, Daten, Zeit, Hardware) dazu ausreichen. Die Beurteilung der Datenqualität

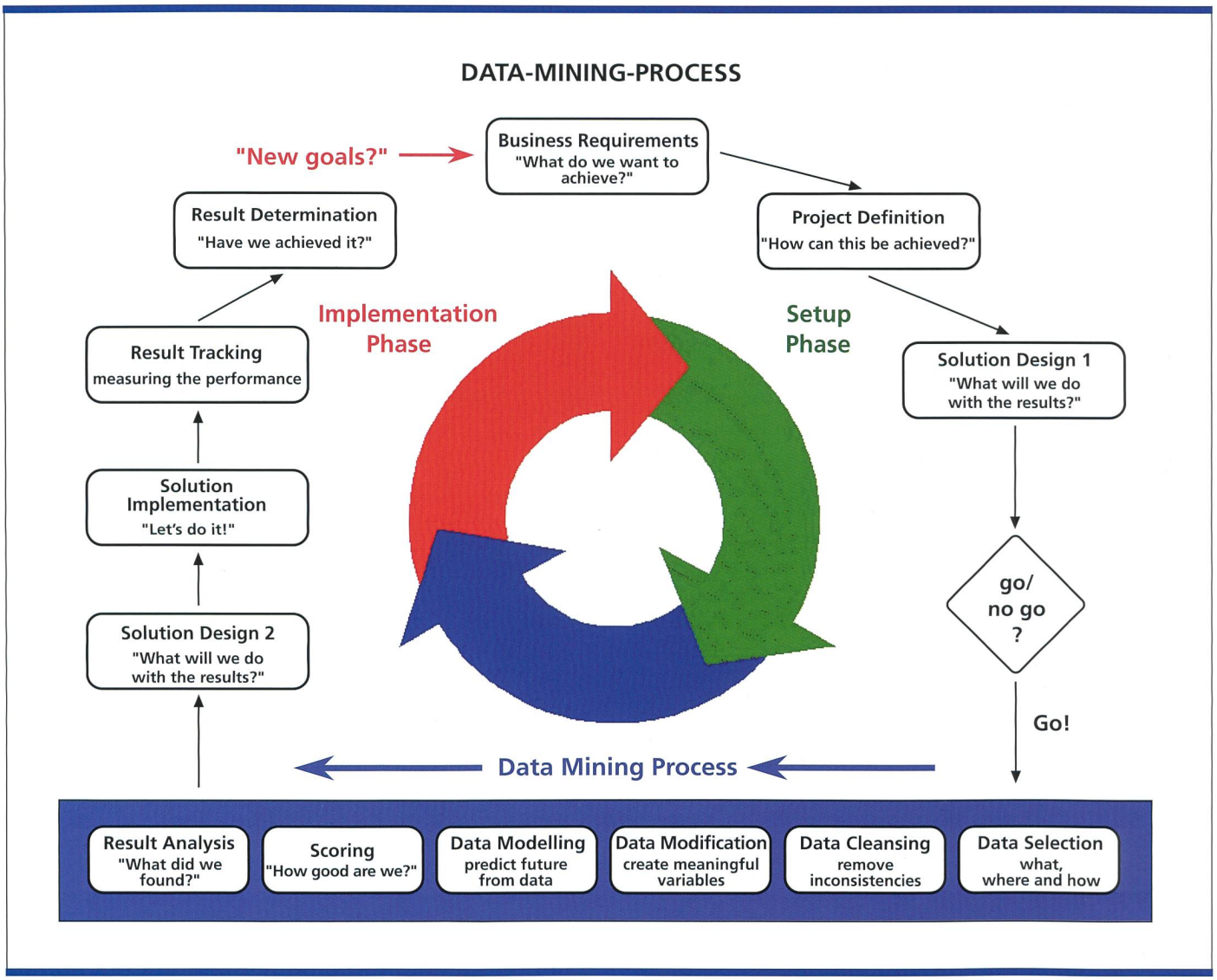

Bild 7. Der «erweiterte» Data-Mining-Prozess.

(garbage in – gargabe out) ist in hohem Masse von der Erfahrung der Mitarbeiter abhängig. Bereits in dieser Phase muss auch die Implementation der Resultate geplant werden.

Phase Data-Mining-Kernprozess
Dies ist die Prozessphase, die hauptsächlich der Informationstechnologie obliegt. Hier findet der eigentliche Kernprozess des Data Mining statt. Es ist auch der Prozessabschnitt, der teilweise von den als Data-Mining-Tools verkauften Paketen unterstützt wird.
Oft werden allerdings die vorgängig notwendigen Schritte Data Cleansing und Datenaufbereitung massiv unterschätzt. Konkrete Erfahrungen haben gezeigt, dass die ersten drei Subprozesse dieser Phase mit ungefähr 80% Zeitaufwand (!) am Gesamtprozess zu Buche schlagen. Insbesondere widersetzt sich das Data Cleansing bis heute einer geeigneten Automatisierung. Es ist in dieser Phase also viel «Handarbeit» notwendig. Die anschliessenden Phasen hingegen können dann dank der verfügbaren Tools in kürzerer Zeit erledigt werden.

Phase Implementierung
Der Prozesskreis schliesst sich nun. Die gewonnenen Erkenntnisse sind im Businessprozess zu implementieren. In der Regel ist mit dem gefundenen Modell eine Datenbank zu bearbeiten und die so erhaltenen Werte werden dann in weiteren Applikationen verwendet.

Lessons Learned
Folgende Faktoren tragen zum Erfolg eines Data-Mining-Projektes bei:

Erfahrung und Teamplay
Da nicht nur 80% der Arbeitszeit, sondern auch 80% der Probleme nicht mit der eigentlichen Modellierung, sondern mit der Datenbeschaffung und -aufbereitung entstehen, ist Erfahrung sehr wertvoll und ein flexibles Team ein grosser Vorteil.

Samples nutzen
Auf Grund der sehr grossen Datenmengen kann sehr viel Zeit gespart werden, indem so weit wie möglich mit kleinen Stichproben gearbeitet wird.

Schneller Datenzugriff
Die Extraktion grosser Datenmengen aus relationalen Datenbanken ist bereits sehr zeitaufwendig. Kommen noch administrative Hürden hinzu, wird es zu einem Albtraum. Daher ist der gesicherte, aber selbstständige Zugriff auf die Daten anzustreben. Unterstützt wird diese Aufgabe auch durch Software wie «Work-

Hersteller	Produkt	Website
SAS Software	Enterprise Miner	www.sas.com/software/data_mining/
SPSS	Clementine	www.spss.com/software/clementine/
SGI	Mine Set	www.sgi.com/software/mineset/
IBM	Intelligent Miner	www.software.ibm.com/data/iminer/
Oracle	Darwin	www.oracle.com/datawarehouse/products/datamining/
Angoss	Knowledgeseeker	www.angoss.com/
Data Destilleries	Data Surveyor	www.ddi.nl/
ASOC	SphinxVision	www.asoc.de/

Bild 8. Das Data-Mining-Team von CT: Souheil Ben Yacoub, Manfred Schmidt, André Loémbé, Marcel Reitmann, Bernard Achermann (v.l.).

bench» der Firma Systemfabrik oder den SAS-«Administrator».

Start small

Aus Sicht von Swisscom empfiehlt es sich, zunächst in einem kleinen Rahmen zu beginnen. Viele Hindernisse werden erst in der praktischen Arbeit aufgedeckt. Es ist ökonomischer, diese Probleme einzeln zu identifizieren, als sie in der Vorphase ausschliessen zu wollen. Damit entsteht ein mehrstufiger Aufbau:

1. Pilotprojekt: Einmal den Weg von den Daten bis zum Modell gehen. Die Mitarbeiter lernen, Bottlenecks werden identifiziert, Prozeduren für die Datenbereinigung werden entwickelt und erste Modelle stehen zur Verfügung.
2. Produktion: Der Data-Mining-Prozess wird beherrscht und es werden regelmässig Modelle erzeugt und angewendet.
3. Automatisierung: Nach und nach werden Teile der Datenanalyse automatisiert, wie Extraktion, Bereinigung und Modellierung.

Datenschutz und Data Mining

Mit einigem Recht können Data Warehousing und Data Mining als «Anti-datenschutz-Technologien» bezeichnet werden. Das Zusammenführen personenbezogener Daten aus unterschiedlichen Datenquellen in ein Data Warehouse erfolgt in der Regel ohne das Wissen der Betroffenen, ebenso die Erstellung von individuellen Verhaltensmustern und die Vorhersage über das Konsumverhalten.

Nicht nur zur Akquirierung von Neukunden reichen die intern vorhandenen Kundeninformationen für ein 1-to-1-Marketing nicht aus; die vorhandenen Kundendaten werden daher mit weiteren personenbezogenen Daten «angereichert», um ein möglichst vollständiges Bild vom Kunden zu erhalten. Diese Daten werden von Informationsagenturen gesammelt und vermarktet – in der Regel ohne Wissen der Betroffenen.

Data Mining ermöglicht auch solche Informationen zu erhalten, die der Kunde nicht preisgeben will: die Affinität zu bestimmten Produkten und Unternehmen, die finanzielle Leistungsfähigkeit, Vorlieben und Gewohnheiten. Der Anspruch des Data Mining, unbekannte Muster und Informationen zu entdecken, kann nur schwer mit der Anforderung des Bundesgesetzes über den Datenschutz (DSG) vereinbart werden. Gemäss diesem Bundesgesetz dürfen Daten nur zu dem Zweck verarbeitet werden, der bei der Beschaffung angegeben wurde (DSG, Art. 4(3)) .

Eine Lösung für diese Probleme könnten Produkte wie NCR «Privacy-Builder» sein, das – angepasst an die Teradata-Datenbanksysteme von NCR – Funktionalitäten zur Verfügung stellt, die den Schutz personenbezogener Daten ermöglicht. Alex Schweizer schreibt im Vorwort zu seinem Buch [9]: «Noch nicht morgen, aber in absehbarer Zeit werden Unternehmungen, welche Data-Mining- und Data-Warehouse-Technologien verwenden, ernsthaft mit einer wahren Prozesslawine wegen Persönlichkeitsverletzung rechnen müssen.»

Website	Kommentar
www.acm.org/sigkdd/	Special Interest Group Knowledge Discovery and Data Mining
www.kdnuggets.com/	eine Fundgrube, Newsletter «KDNuggets»
www.cs.bham.ac.uk/~anp/TheDataMine.html	Andy Pryke's Data Mining Site, viele Informationen
www.dw-insttute.com/buyersguide99/solutions/datamining/maindm.html	Informationen zu DM-Produkten
www.almaden.ibm.com/cs/quest/publications.html	Publikationen Forschungszentrum IBM, Almaden
datawarehouse.dci.com/links.htm	weitere Links zu DM und Data Warehousing
ftp.sas.com/pub/neural/FAQ.html	viele Informationen über neuronale Netze
www.patents.ibm.com/details?&pn=US05937422__	Patent der Echelon-Technologie
www.co.umist.ac.uk/~hamid/bookmark.html#Dat	viele weitere Data Mining Links

Referenzen

[1] Berry, M.J.A., Linoff, G. Data Mining Techniques. For Marketing, Sales, and Customer Support. Wiley 1997.

[2] Bigus, J.P. Data Mining with Neural Networks. McGraw-Hill Companies, Inc. 1996.

[3] Cios, K.J., Pedrycz, W. and Swiniarski, R.W. Data Mining Methods for Knowledge Discovery. Kluwer Academic 1998.

[4] Groth, R. Data Mining. A Hands-on Approach for Business Professionals. Prentice Hall 1998.

[5] Mattison, R. Data Warehousing and Data Mining for Telecommunications. Artech House, Inc. 1997.

[6] Thuraisingham, B. Data Mining. Technologies, Techniques, Tools, and Trends. CRC Press 1999.

[7] Zhong, N. and Zhou, L. (eds.). Methodologies for Knowledge Discovery and Data Mining. Springer 1999 (Lecture notes in computer science; 1574. Lecture notes in artificial intelligence).

[8] Zytkow, J.M. and Quafafou, M. (eds.). Principles of Data Mining and Knowledge Discovery. Springer 1998 (Lecture notes in computer science; Vol. 1510. Lecture notes in artificial intelligence).

[9] Schweizer, Alex. Data Mining, Data Warehousing: Datenschutzrechtliche Orientierungshilfen. Orell-Füssli 1999.

[10] Agrawal, R., Imielinski, T. and Swami A. Mining Association Rules between Sets of Items in Large Databases. Proc. of the ACM SIGMOD Conference on Management of Data. 1993.

Trends

Vertikale Integration

Data-Mining-Lösungen als Stand-alone-Product stellen den Anwender vor verschiedene Probleme: Diese Tools sind zum Teil nicht einfach zu bedienen und sie sind in die bestehende IT-Infrastruktur zu integrieren. Daher ist es naheliegend, diese Funktionalität in die Produkte zu integrieren, die heute mit den Data Mining Tools zusammenarbeiten. So werden bereits heute einige E-Commerce-Produkte mit integrierter Data-Mining-Funktionalität angeboten (z.B. von Blue Martini) und Datenbankhersteller sind dabei, Data-Mining-Funktionen in ihre Datenbanksoftware zu integrieren (z.B. «NonStop SQL/MX» von Compaq). Diese Entwicklung wird dazu führen, dass Data-Mining-Funktionen allgegenwärtig, aber dennoch unsichtbar werden.

Text Mining

Da sehr viel «Knowledge» in Textdokumenten enthalten ist, gibt es eine starke Entwicklung zu Tools, die eine automatische Klassifikation und Erkennung von Texten ermöglichen. Eines der ersten kommerziellen Produkte ist der «Intelligent Miner for Text» von IBM. Solche Produkte können verwendet werden für die automatische Analyse von Kundenreaktionen oder E-Mails. Eben diese Technologie kommt auch bei dem globalen Überwachungssystem Echelon zu Anwendung.

Standards

Um die Schwierigkeiten mit unterschiedlichen proprietären Modellformaten zu überwinden und die erstellten Modelle in verschiedenen Applikationen zu verwenden, sind offene Datenformate, welche die Meta-Informationen (welche Daten, welche Algorithmen, Formate usw.) enthalten, gesucht. SPSS versucht mit

Glossar

Data Mining:	«Graben in Daten», automatisierte Analyse grosser Datenbanken.
Data Warehouse:	Unternehmensweite Datenbank, entlastet operative Systeme von Reportingaufgaben.
Data Warehouse:	Unternehmensweite Datenbank, welche die Daten der operativen Systeme zusammenführt, aufbereitet und zur Verfügung stellt.
OLAP:	Im Unterschied zu Data Mining Tools steht bei OLAP-Tools nicht der automatisierte Aspekt im Vordergrund, vielmehr unterstützen sie einen Anwender bei der manuellen Datenanalyse, indem sie ihm erlauben, die Daten im Warehouse nach verschiedenen Dimensionen und in verschiedene Granularitätsstufen (Aggregationen) aufgeteilt zu betrachten.
CRM:	Customer Relationship Management, zusammenfassender Begriff für das Identifizieren, Gewinnen und Binden von Kunden.

Summary

Data Mining

As Data Mining is an important enabler of Customer Relationship Management, this technology is explored in EP0025. The most common methods and tools are presented in this article. Critical success factors for Data Mining are:
– a clear goal from business
– experienced employees and team play
– start small
– do not underestimate data retrieval and preprocessing
– use samples whenever possible
Furthermore, it is shown how Data Mining might endanger customers privacy in several ways. Finally, new trends in Data Mining such as «vertical integration», «textmining» and upcoming standards are briefly indicated.

PMML (Predictive Modelling Markup Language) einen solchen Standard auf Grundlage von XML zu etablieren. Eine Standardisierung ist auch für das Data-Mining-Projekt- bzw. -Prozessmanagement zu erwarten. Das Projekt CRISP-DM (www.crisp-dm.org/) ist ein erster Versuch dazu.

Automatisierung

Während heute Data Mining eine teure und zeitraubende Tätigkeit ist, kann 1-to-1-Marketing im Massenmarkt nur mit weit gehend automatisierten Prozessen realisiert werden. Auch wenn viele Spezialisten skeptisch sind, was eine völlig automatisierte Modellierung betrifft, bietet SLP mit dem Produkt «Churn/CPS» bereits eine weitgehend automatisierte Data-Mining-Lösung an. [7]

Bernard Achermann *arbeitet seit 1998 als Softwareingenieur bei CIT-CT-ITA. Zuvor war er an der Universität Bern im Bereich Bildverarbeitung und Künstliche Intelligenz tätig, wo er mit einer Arbeit im Gebiet der Tiefenbilder und Gesichtserkennung promoviert hat. Seine Schwerpunktgebiete bei Swisscom sind Data Mining, E-Commerce und Internet Computing, wo er in einer Reihe von Projekten involviert ist*

Manfred Schmidt *studierte in Dortmund Mathematik und Elektrotechnik, promovierte an der philosophisch-naturwissenschaftlichen Fakultät der Universität Bern und ist seit 1999 für Swisscom, CIT-CT-TPM, tätig. Neben der Anwendung von Data Mining in der Telekommunikation gilt sein Interesse vor allem Smart Cards und Sicherheit.*

André Loémbé *trat 1987 nach der Promotion als Elektroingenieur an der EPF Lausanne in den Dienst der PTT, wo er in der Gruppe für drahtlose Kommunikation arbeitete und an verschiedenen internationalen Projekten mitarbeitete. Seit er 1997 zu CIT-CT-ITA wechselte, hat er bei der Entwicklung ortsbasierter Dienste mit WAP und Java-Servlet-Technologie mitgearbeitet. Seine gegenwärtige Tätigkeit konzentriert sich auf die Bewertung und Anwendung verschiedener Data-Mining-Prozesse.*

19 Millionen Leuchtdioden erhellen den Times Square

...natürlich nicht, um die Strassenbeleuchtung zu ersetzen. Die NASDAQ, Technologiebörse der Amerikaner, hat an diesem strategischen Punkt in New York ein Flachdisplay für die aktuellen Börsennotierungen in Betrieb genommen. Mit seinen 30 × 40 m dürfte es das derzeit grösste LED-Display der Welt sein – es reicht acht Stockwerke hoch.

Chip-Produktion: Leasen statt kaufen?

Bei den japanischen Halbleiterfirmen scheint ein Umdenkprozess im Gange zu sein: NEC wird mehr als die Hälfte des geplanten neuen Halbleiterfertigungsequipments im Wirtschaftsjahr 1999 (endend 31. 3. 2000) leasen – insgesamt für 775 Mio. US-$. Toshiba will die gesamten Produktionsgeräte für ihre Fabrik in Virginia (Investition von rund 400 Mio. US-$) ebenfalls leasen. In beiden Fällen kaufen die Unternehmen das Equipment, verkaufen es dann an eine Leasingfirma (NEC z.B. an Comdisco in den USA) und leasen dann alles wieder zurück.

Qualcomm verkauft seinen CDMA-Mobilfunkbereich an Kyocera

Mehr als 1 Mia. US-$ zahlt Kyocera für den CDMA-Bereich der kalifornischen Qualcomm. Kyocera hat bereits CDMA-Ableger in Japan und Korea. Das Unternehmen baut gegenwärtig rund 4 Mio. Geräte im Jahr und will diese Menge mit Hilfe der US-Neuerwerbung auf 15 Mio. hochfahren.

Kyocera America Inc.
8611 Balboa Avenue
San Diego
CA 92153-1580
USA.

Patentstreit mit Zähnen und Klauen

Was früher einmal die legendären Kilby-Patente von Texas Instruments für die ersten integrierten Schaltkreise waren, das sind jetzt Patente für die RAMBUS-Speicher. Die Schnelligkeit der RAMBUS-Chips prädestiniert sie für die ultraschnellen neuen PCs und vor allem für leistungsfähige Workstations. Die von der Rambus Inc. gehaltenen Patente erschweren vielen Chipherstellern den Zugang zu der Technologie, es sei denn, sie zahlen Royalties. Die Rambus Inc. hat jetzt Hitachi verklagt, vier ihrer fundamentalen Patente verletzt zu haben. Die angeblichen Patentverletzungen erstrecken sich sowohl auf die Basistechnologie dieser Speicher als auch auf Verfahren zur Ansteuerung. Die Patente stammen alle von 1990, aus dem Gründungsjahr von Rambus, als deren Systemvorschlag noch sehr exotisch war: Alle Welt beschäftigte sich damals noch mit EDO-DRAMs und fing erst an, über synchrone DRAMs nachzudenken.

Auf dem Weg zu selbstkonfigurierenden Robotern

Im Palo Alto Research Center (PARC) von Xerox wird gegenwärtig an Grundbausteinen für dreidimensionale Roboter gearbeitet, die sich eines Tages entsprechend den jeweiligen Anforderungen zu komplexen Funktionen selbst zusammensetzen können. Das Kernproblem dabei ist, den Lernprozess für eine dreidimensionale Konfiguration zu beschreiben. Was zweidimensional heute mit Computerhilfe lösbar ist, wird dreidimensional zur Sisyphusarbeit. Wenn man den Gedanken weiterspinnt, dann könnten eines Tages aus scheinbar «nutzlosen» Basiselementen je nach Bedarf auch «tote» Gegenstände neu entstehen: Ein Hammer oder ein Stuhl würden aus ähnlichen Basiselementen bestehen. Xerox glaubt, dass die laufenden Arbeiten den Entwurf von Robotern in zehn Jahren entscheidend verändern werden.

Fuji Xerox PARC
3400 Hillview Ave, Bldg 4
Palo Alto CA 94304
U.S.A.
Tel. +1-650-813 7765
Fax +1-650-813 7081

Japanische Telekom- und Rundfunkbranche wuchs 1999 um fast 10%

Auf 192 Mia. US-$ ist nach Schätzungen des japanischen Postministeriums der Umsatz in der Telekommunikation und in der Rundfunk-/Kabelindustrie im Jahr 1999 gewachsen. Gegenüber dem Vorjahr waren die Investitionen rückläufig (-6%). Unter den verschiedenen Bereichen zeigt die Mobilfunkindustrie mit 14% Zuwachs deutlich die stärkste Zunahme. Die Zahlen sind erste Hochrechnungen aus 60% der befragten Unternehmen.