Visualisierung großer Datenbanken

Prof. Dr. Hans-Peter Kriegel, Institut für Informatik, Universität München und Prof. Dr. Daniel A. Keim, Institut für Informatik, Universität Halle-Wittenberg

Durch den rasanten technologischen Fortschritt steigt die Menge an Daten, die in heutigen Computersystemen gespeichert ist, sehr schnell an. Damit wird die Suche nach interessanter Information innerhalb der Datenbestände immer schwieriger. Für sehr große Datenbanken werden deshalb auf neuartigen Visualisierungstechniken basierende Ansätze zur Datenexploration und Datenanalyse benötigt.

Bei Entscheidungen ist es wichtig, im richtigen Augenblick die richtigen Informationen zur Hand zu haben. Die Menge an Information, die in gespeicherter Form verfügbar und für die Entscheidungsfindung potentiell von Bedeutung ist, nimmt rasend schnell zu. Nach neuesten Schätzungen verdoppelt sich die Menge der weltweit vorhandenen Information alle 20 Monate.

Eine Ursache für die ständig ansteigenden Datenmengen ist die Automatisierung fast aller Vorgänge in Wirtschaft, Wissenschaft und Verwaltung. Selbst einfache Vorgänge, wie das Bezahlen mit Kreditkarte oder das Telefonieren, werden heute durch Computer erfaßt. Versuchsreihen in Physik, Chemie und Medizin erzeugen große Mengen an Daten, die zumeist automatisch mit Hilfe von Sensoren gesammelt werden. Und Beobachtungssatelliten werden schon bald täglich Datenmengen im Terabytebereich erheben und zur Erde übermitteln.

Heute verfügbare Datenbankmanagementsysteme unterstützen den Benutzer bei der Speicherung und Verwaltung der Daten sowie bei der Suche nach exakt spezifizierten Daten. Sie sind im allgemeinen aber ungeeignet, um die unexakt spezifizierte Suche nach interessanten Zusammenhängen (Data Mining-Prozeß) herauszufinden. Für das Data Mining (Datenexploration und -analyse) verwendet man Techniken aus den Bereichen multivariate Statistik (z. B. Hauptkomponenten-, Faktor- und Clusteranalyse), Knowledge Discovery sowie Information Retrieval. Die in diesen Bereichen entwickelten Techniken eignen sich im allgemeinen jedoch nicht für die Datenexploration und -analyse von großen Datenbanken mit Hunderttausenden oder sogar Millionen von Datensätzen.

Eine effektive Unterstützung der Datenexploration und -analyse bei großen Datenmengen ist derzeit nur unter Einbeziehung des Menschen und seiner Fähigkeiten möglich. Insbesondere seine unübertroffenen Fähigkeiten der Wahrnehmung erlauben

es dem Menschen, in kürzester Zeit komplexe Sachverhalte zu analysieren, wichtige Informationen zu erkennen und Entscheidungen zu treffen. Das menschliche Wahrnehmungssystem kann flexibel die verschiedensten Arten von Daten verarbeiten, wobei es intuitiv ungewöhnliche Eigenschaften erkennt, bekannte Eigenschaften dagegen ignoriert. Menschen können leichter und besser mit vagen Beschreibungen und unscharfem Wissen umgehen als heutige IT-Systeme, und ihr Allgemeinwissen erlaubt es ihnen, ohne geistige Anstrengung komplexe Schlußfolgerungen zu ziehen.

Das Ziel eines modernen Ansatzes der Datenexploration und -analyse muß es deshalb sein, den Menschen in den Data Mining-Prozeß einzubeziehen. Es gilt, die immense Speicherkapazität und Rechenleistung heutiger Computer mit Intuition, Flexibilität, Kreativität und Allgemeinwissen des Menschen zu vereinen. Dabei ist die Entwicklung von Techniken wichtig, die den Menschen nicht einfach mit Daten überhäufen, sondern einen guten Überblick über die wesentlichen Daten ermöglichen.

Der hier vorgestellte Ansatz zur Datenexploration und -analyse großer Datenbanken basiert auf neuartigen Visualisierungstechniken für multidimensionale Daten. Die prinzipielle Idee ist die gleichzeitige Darstellung möglichst vieler Datenobjekte am Bildschirm, wobei jeder Datenwert durch ein Pixel des Bildschirms repräsentiert wird. Die Farbe des Pixels entspricht jeweils einem Datenwert (dem Abstand zwischen Datenwert und Anfragewert). Die Anordnung der Pixel hängt von der gewählten Visualisierungstechnik ab und ist entweder durch eine vorgegebene Sortierung der Daten oder durch die Gesamtdistanz der Datensätze in bezug auf die Anfrage gegeben. Durch ein graphisches Benutzerinterface kann der Benutzer seine Anfragen schrittweise ändern, wobei er durch das visuelle Feedback, das er bei Änderungen bekommt, in der Verfeinerung seiner Anfragen unterstützt wird.

Bei den angestellten Betrachtungen wird zunächst von einer einfachen Strukturierung der Daten, dem relationalen Datenmodell, ausgegangen. Eine simple relationale Datenbank (z. B. eine Personendatenbank) kann man sich als eine große Tabelle vorstellen, in der die Zeilen den Datensätzen (Personen) entsprechen und die Spalten der über die Personen gespeicherten Information (z. B. Name, Geburtsdatum, Alter etc.). In unserem Kontext werden die Spalten auch als Attribute der Datensätze bezeichnet.

Visualisierungstechniken multidimensionaler Daten

In vielen Bereichen von Forschung und Industrie werden Visualisierungen von Daten, die eine inhärente zwei- oder dreidimensionale Semantik haben, verwendet. Eine

Übersicht über solche Techniken ist bei Tufte[1] zu finden. Bis vor kurzem gab es jedoch nur wenige Techniken, die eine Visualisierung multidimensionaler Daten ohne inhärente zwei- oder dreidimensionale Semantik erlauben. Erste Ansätze sind beispielsweise Matrizen von X-Y-Diagrammen oder die Chernoff'sche Gesichterdarstellung.[2]

Durch die zunehmende Verfügbarkeit von Grafik-Workstations mit hoher Rechenleistung wurden in den letzten Jahren zahlreiche neue Visualisierungstechniken entwickelt. Zwei bekannte Techniken, die auch im Rahmen des VisDB-Systems implementiert wurden, sind die Parallele Koordinaten-Technik und die Strichmännchen-Technik.

Die Technik der parallelen Koordinaten ist eine geometrische Projektionstechnik, bei der der k-dimensionale Raum mit Hilfe von k äquidistanten Achsen, die parallel zu einer der Bildschirmachsen liegen, dargestellt wird.[3] Die Achsen entsprechen den Dimensionen und sind vom Minimum- bis zum Maximumwert der Dimensionen linear skaliert. Jeder Datensatz wird als polygonale Linie dargestellt, die jede Achse an dem Punkt schneidet, dessen Wert der jeweiligen Dimension entspricht (vgl. Abb. 1).

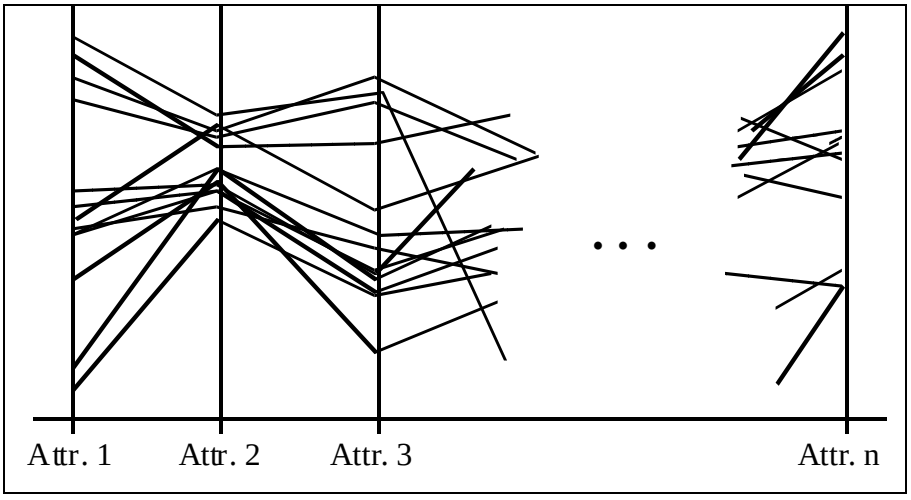

Abb. 1: Parallele Koordinaten-Technik

Die Parallele Koordinaten-Technik ermöglicht das Erkennen eines weiten Spektrums

[1] Vgl. Tufte (1990).
[2] Vgl. Tufte (1990).
[3] Vgl. Inselberg/Dimsdale (1990).

von Datencharakteristika, wie z. B. verschiedene Datenverteilungen und funktionale Abhängigkeiten. Wegen der Linienüberlappungen ist jedoch die Anzahl der gleichzeitig visuell darstellbaren Datensätze auf ca. 1.000 begrenzt.

Die sogenannte Strichmännchen-Technik ist eine Pixeldiagramm-Technik.[4] Wie der Name bereits sagt, sind die verwendeten Icons eine Art Strichmännchen, wobei die Winkel und Strichlängen die Datendimensionen repräsentieren (vgl. Abb. 2).

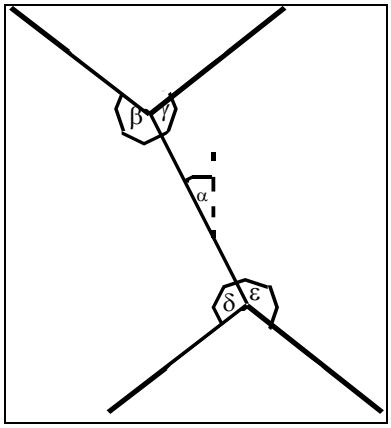

Abb. 2: Strichmännchen-Technik

Wenn die Datensätze in bezug auf die Bildschirmdimensionen verhältnismäßig dicht beieinander liegen, zeigt die resultierende Visualisierung Strukturmuster, die gemäß der Datencharakteristika variieren. Als Strichmännchen können verschiedene Icons mit unterschiedlicher Dimensionalität verwendet werden (vgl. Abb. 3).

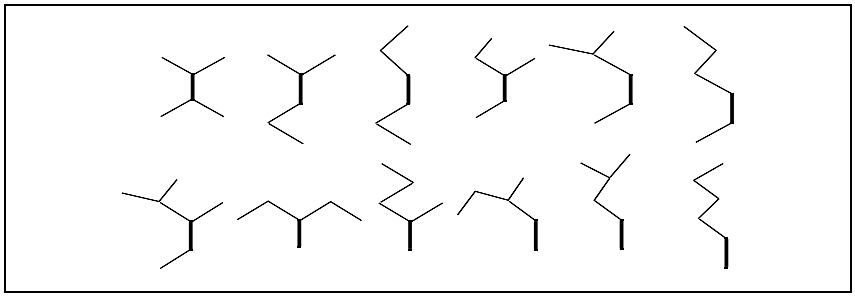

Abb. 3: Beispiele für „Strichmännchen"

[4] Vgl. Pickett/Grinstein (1988).

Pixel-orientierte Visualisierung

In den bisher vorgestellten Visualisierungstechniken für multidimensionale Daten ist die Anzahl der gleichzeitig am Bildschirm darstellbaren Datensätze auf maximal 100 bis 1.000 begrenzt. Im folgenden sollen einige neu entwickelte Visualisierungstechniken, die sich auch für Datenmengen bis 1.000.000 Datensätze eignen, vorgestellt werden.

Bei den pixel-orientierten Visualisierungstechniken wird jeder Datenwert durch ein farbiges Pixel dargestellt. Die Pixel für die einzelnen Attribute werden in separaten Fenstern dargestellt (vgl. Abb. 4).

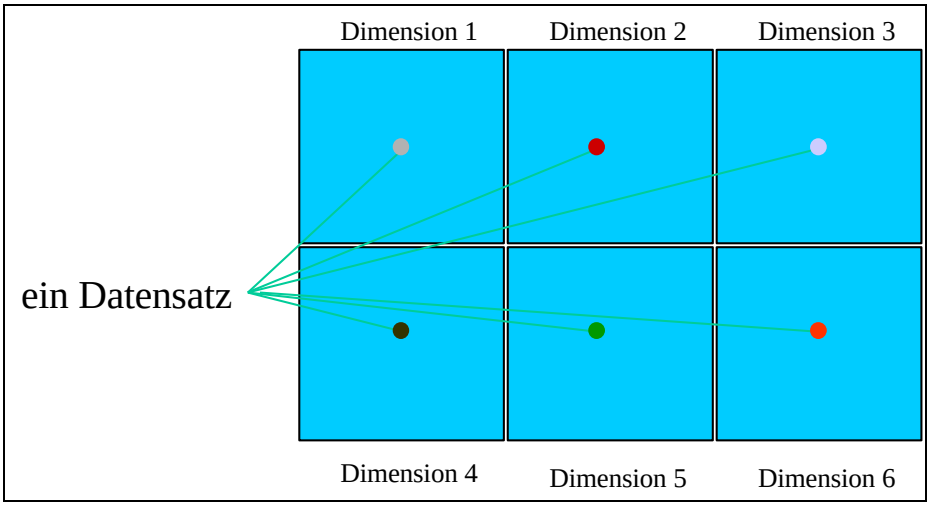

Abb. 4: Pixel-orientierte Visualisierung

Eine wichtige Frage ist, wie die einem Attribut entsprechenden Pixel innerhalb des Fensters angeordnet werden. Eine einfache Anordnung wäre beispielsweise zeilen- oder spaltenweise. Diese führt in der Regel aber zu wenig aussagekräftigen Visualisierungen. Bei der Recursive Pattern-Visualisierungstechnik wird deshalb eine strukturierte Anordnung verwendet, die vom Benutzer gemäß der Semantik der Daten wählbar ist.[5]

Die Anordnung ist eine rekursive Verallgemeinerung zeilen- und spaltenorientierter Anordnungen. Auf der ersten Rekursionsebene werden h_1 mal w_1 Pixel abwechselnd von links nach rechts und von rechts nach links angeordnet. Das entstehende

[5] Vgl. zur Recursive Pattern-Technik Keim/Kriegel/Ankerst (1995).

Pixelmuster sei ein (Ebene-1)-Pixelmuster. Auf den nächsten Rekursionsebenen werden die (Ebene-1)-Pixelmuster wieder gemäß dem gleichen Schema angeordnet. Sei w_i die Anzahl der Pixelmuster, die auf Rekursionsebene i in Links-rechts-Richtung angeordnet werden sollen und h_i die Anzahl der Zeilen auf Rekursionsebene i. Auf jeder Rekursionebene werden w_i(Ebene-1)-Pixelmuster h_i-mal abwechselnd von links nach rechts und von rechts nach links angeordnet. Die w_i und h_i müssen vom Benutzer vorgegeben werden.

Durchgeführt wurde eine Recursive Pattern-Visualisierung der täglichen Aktienkurse des Frankfurter Aktienindexes von Januar 1994 bis April 1995, was insgesamt 532.900 Datenwerten entspricht. Die gewählten Parameterwerte für w_i und h_i sind (w_1, h_1) = (1, 22) und (w_2, h_2) = (243, 1). Eine Spalte mit 22 Pixel entspricht dabei ungefähr einem Monat. Die Farbkodierung der Datenwerte bildet hohe Datenwerte auf helle Farben und niedrige Werte auf dunkle Farben ab. Das Farbspektrum durchläuft die Farben gelb, hellgrün, blau, rot bis dunkelbraun.

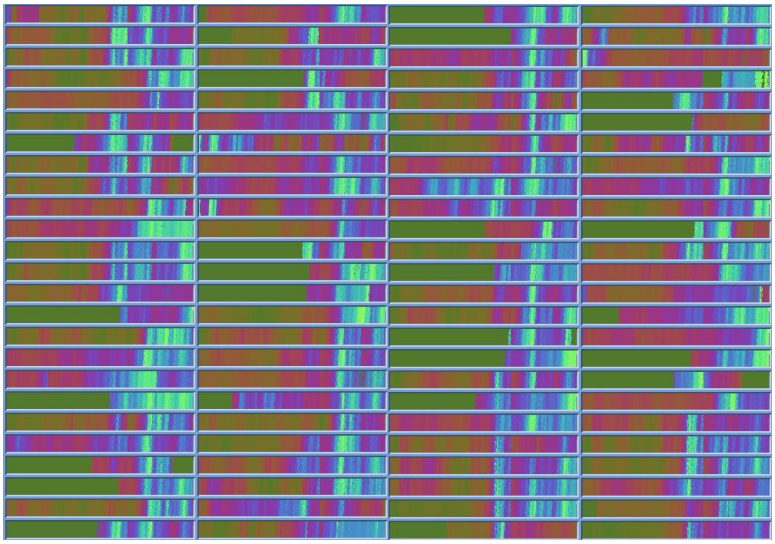

Abb. 5: "Recursive Pattern"-Visualisierung der täglichen Aktienkurse des FAZ-Index für den Zeitraum Jaunar ´74 bis April `95 (insgesamt 532.900 Datenwerte)

Die Visualisierung erlaubte eine Reihe interessanter Beobachtungen, die an dieser Stelle nur angedeutet werden können. Aufschlußreich waren beispielsweise die ähnlichen Entwicklungen der Aktienkurse von Südzucker, Thyssen, Veba, Volkswagen und Bayrische Hypothekenbank. Bemerkenswert war dabei insbesondere,

daß es sich um Firmen handelt, die in völlig unterschiedlichen Branchen arbeiten. Eine weitere interessante Beobachtung war, daß in mehr als 50 Prozent der Fälle ein heller grüner Streifen an ungefähr der gleichen Position zu sehen war. Der grüne Streifen signalisiert eine allgemeine Phase besonders hoher Aktienkurse, wie sie etwa im Frühjahr 1990 zu beobachten war. Der Visualisierung war leicht zu entnehmen, daß sich viele Aktien nicht mehr vollständig von einem Einbruch der Kurse erholen konnten, der auf die Phase der hohen Aktienkurse folgte. Ebenso leicht waren die Aktien zu identifizieren, die sich gegen den allgemeinen Trend entwickelt hatten.

Anfrageabhängige Visualisierungstechniken

Neben einer statischen Visualisierung der Daten ist es beim Data Mining aber auch wichtig, die Daten in Abhängigkeit von einer gestellten Anfrage zu visualisieren. Im folgenden wird eine anfrageabhängige Visualisierungstechnik, die sogenannte Spiraltechnik, vorgestellt. Um diese Technik beschreiben zu können, betrachten wir die Relationen einer relationalen Datenbank als Mengen von Tupeln (Vektoren) der Form (a1, a2, ..., ak), wobei a1, a2,..., ak die Attributswerte eines Datensatzes darstellen.

Anfragen an relationale Datenbanken können als Anfrageregion(en) im k-dimensionalen Raum, der durch die k Attribute einer Relation aufgespannt wird, verstanden werden. Alle Datensätze, die innerhalb der Anfrageregion(en) liegen, stellen die Antworten auf die Anfrage dar und werden als Ergebnis der Anfrage ermittelt. Die Menge der Antworten kann sehr groß aber auch leer sein. In beiden Fällen ist es für den Benutzer schwierig, die Antwort zu verstehen und die Anfrage entsprechend zu modifizieren. Um dem Benutzer mehr Feedback auf seine Anfrage zu geben, werden durch unsere Visualisierungstechniken nicht nur die Datensätze visualisiert, die innerhalb der Anfrageregion(en) liegen und damit die Anfrage erfüllen, sondern auch solche, die in der Nähe der Anfrageregion(en) liegen und damit die Anfrage nur approximativ erfüllen.

Unabhängig davon, ob ein Datensatz die Anfrage erfüllt oder nicht, kann für jedes Attribut der Abstand von dem vorgegebenen Anfragewert oder -intervall berechnet werden. Macht man dies für jedes Attribut, so erhält man Tupel (d1, d2,..., dk), die die Distanzen der Datenwerte bezüglich der Anfrage beinhalten. Verändert man die Anfrageregion, so ändern sich die Distanztupel entsprechend. Das Distanztupel kann um einen (k+1)-ten Wert erweitert werden, der die Gesamtdistanz des Datensatzes bezüglich der Anfrage darstellt. Der Wert von dk+1 ist 0, falls der Datensatz die Anfrage erfüllt; ansonsten gibt dk+1 den Abstand des Datensatzes bezüglich der Anfrage wieder. Die Menge der Distanztupel (d1, d2,..., dk, dk+1) wird nach dem Wert dk+1 (Resultat) aufsteigend sortiert, d. h. am Anfang stehen die Tupel mit dk+1

= 0 (falls vorhanden) und am Schluß die Tupel mit den größten Distanzen.

Anstatt der Datenwerte werden bei den anfrageabhängigen Techniken die Distanzen bezüglich der Anfrage visualisiert. Die Distanzen für jedes Attribut inklusive des Gesamtresultats werden dabei auf eine spezielle Farbskala abgebildet. Die Farbskala ist so entworfen, daß dem Distanzwert 0 die Farbe gelb zugeordnet ist; Distanzwerte größer 0 werden in aufsteigender Reihenfolge immer dunkler. Die gelbe Farbe ist besonders hervorgehoben und zeigt an, daß der zugehörige Datenwert innerhalb des vorgegebenen Anfrageintervalls liegt; die übrigen Farben zeigen die relative Entfernung des Attributwertes von dem Intervall an. Für eine einfache Zuordnung von Datenwerten zu den Farbpixeln sorgt eine Option des interaktiven Interfaces (vgl. Keim/Kriegel/Seidl 1994 und Keim 1994). Durch Anklicken von Pixeln können die zugehörigen Datenwerte abgefragt werden.

Ausblick

Visualisierungstechniken können bei der Exploration und Analyse sehr großer multidimensionaler Daten hilfreich sein, um interessante Daten und ihre Eigenschaften zu finden. Unser Ansatz der Datenexploration zielt auf eine adäquate Unterstützung des Menschen durch den Computer ab und kombiniert Datenbankanfrage- und Information-Retrieval-Techniken mit neuartigen Visualisierungstechniken. Die Anzahl der Datenwerte, die zu einem Zeitpunkt am Bildschirm dargestellt werden können, ist dabei nur durch die Auflösung des Bildschirms beschränkt und liegt für die heute zur Verfügung stehenden 19 Zoll Bildschirme mit einer Auflösung von 1.024 x 1.280 bei etwa 1,3 Millionen Pixel. Ziel zukünftiger Forschungsarbeiten ist es, die Menge der gleichzeitig darstellbaren Datensätze noch weiter zu erhöhen sowie die Qualität und Aussagekraft der Visualisierungen zu verbessern. Eine Möglichkeit ist beispielsweise, durch Verschieben der Anfrageregion im k-dimensionalen Raum Sequenzen von Visualisierungen zu erzeugen. Dadurch können zum einen größere Datenmengen visualisiert werden, zum anderen werden durch die Veränderung der Bilder aber auch Abhängigkeiten innerhalb der Daten besser wahrnehmbar.

Literatur

Inselberg, A./Dimsdale, B. (1990), Parallel Coordinates - A Tool for Visualizing Multi-Dimensional Geometry, in: Visualization ‘90, San Francisco (CA) 1990, S. 361 - 370

Keim, D. A. (1994), Visual Support for Query Specification and Data Mining, Dissertation, Ludwig-Maximilians-Universität, München 1994

Keim, D. A./Kriegel, H.-P./Seidl, T. (1993), Visual Feedback in Querying Large Databases, Procedure Visualization ‘93, San Jose (CA) 1993, S. 158 - 165

Keim, D. A./Kriegel, H.-P. (1994), VisDB - Database Exploration using Multidimensional Visualization, in: Computer Graphics & Applications, Sept. 1994, S. 40 - 49

Keim, D. A./Kriegel, H.-P./Seidl, T. (1994), Supporting Data Mining of Large Databases by Visual Feedback Queries, in: Procedure of the 10th International Conference on Data Engineering, Houston (TX) 1994, S. 302 - 313

Keim, D. A./Kriegel, H.-P. (1995), Visualisierungstechniken zur Exploration und Analyse sehr großer Datenbanken, Procedure zu Datenbanksysteme in Büro, Technik und Wissenschaft (BTW), Dresden 1995, in: Informatik Aktuell, Springer, Heidelberg 1995, S. 262 - 281

Keim, D. A./Kriegel, H.-P./Ankerst, M. (1995), Recursive Pattern - A Technique for Visualizing Very Large Amounts of Data, in: Visualization ‘95, Atlanta (GA) 1995, S. 279 - 286

Pickett, R. M./Grinstein, G. G. (1988), Iconographic Displays for Visualizing Multidimensional Data, in: Procedure of IEEE Conference on Systems, Man and Cybernetics, IEEE Press, Piscataway (NJ) 1988, S. 514 - 519

Tufte, E. R. (1990), Envisioning Information, Cheshire (CT), 1990