Tiefschürfen in Datenbanken

Data Mining mit bloßem Auge

Von Daniel A. Kaim

Das menschliche Wahrnehmungssystem verfügt für Zwecke des Data Mining über bemerkenswerte Qualitäten. In einem Bild, das aus lauter kleinen Punkten (Pixeln) zusammengesetzt ist, erkennen wir mühelos großräumige Strukturen, auch wenn sie teilweise verdeckt, unscharf, unvollständig oder verzerrt sind. Nichts anderes tun wir, wenn wir ein Fernsehbild betrachten.

Die einzelnen Bildpunkte müssen nicht aus einer Fernsehkamera stammen. Die Farbe jedes einzelnen Pixels kann Ergebnis einer Berechnung sein, in die Millionen bis Milliarden von Einzeldaten eingegangen sind. Ein gutes Data-Mining-Verfahren macht aus einer ungeheuren Menge von Rohdaten die Farben der ungefähr eine Million Pixel, die ein größerer Computerbildschirm darstellen kann. Wenn man es geschickt anstellt, springt dann dem Betrachter des Bildschirms vielleicht eine Struktur ins Auge. Das kann durchaus eine Struktur sein, die der Data-Mining-Algorithmus gar nicht ausfindig gemacht hat. "Visuelles Data Mining" vereinigt also auf vorteilhafte Weise die enormen Speicherkapazitäten und Rechenleistungen moderner Computersysteme mit Fähigkeiten des Menschen, vor allem Flexibilität, Kreativität und Allgemeinverständnis.

Der Einsatz solcher Verfahren ist immer dann sinnvoll, wenn wenig über die Daten bekannt ist und man nicht genau weiß, wonach man sucht. Ein Mensch sieht - noch undeutlich - eine Struktur in den Daten, so wie das Programm sie ihm präsentiert, wählt eine andere Präsentationsform, die diese Struktur klarer zeigen soll, sieht daraufhin mehr Struktur, und so weiter. Dadurch stellt er Hypothesen über die Daten auf, die das Data-Mining-Programm im nächsten Datendurchlauf bestätigt - oder auch nicht. Da in der Regel viele Datendurchläufe erforderlich sind, muss die Interaktion des Benutzers mit der Maschine bequem und schnell gestaltet werden.

Visuelle Datenexploration ist also ein Prozess zur Erzeugung von Hypothesen. Im Verein mit automatischen Algorithmen aus den Bereichen Statistik und Künstliche Intelligenz ist sie zu einem unentbehrlichen Verfahren zur Exploration großer Datenbanken geworden. Durch die unmittelbare Rückkopplung über die bildliche Darstellung kommt der Benutzer ohne spezielle mathematische und statistische Kenntnisse aus.

Wenn die darzustellenden Daten selbst aus ungefähr einer Million einzelner Zahlenwerte bestehen, liegt die Grundidee nahe: Jeder Datenwert bestimmt die Farbe genau eines Pixels. Im Einzelnen ist die Sache nicht so einfach. Die Tageskurse der hundert Aktien des FAZ-Index über die reichlich zwanzig Jahre von Januar 1974 bis April 1995 sind zwar ungefähr eine Million Zahlen, und die Wahl einer geeigneten Einfärbevorschrift ist nicht schwer: Man nehme helle Farben für hohe Kurse und dunkle Farben für niedrige. Aber wie ordnet man sie an, damit Strukturen erkennbar werden?

Die Daten legen eine Tabellenstruktur nahe: Jede Zeile entspricht einer Aktie, jede Spalte einem Tag, und jeder Tabelleneintrag besteht aus einem einzigen Pixel. Das ergibt einen wenig instruktiven schmalen Zeitstreifen, der sich über mehr als sieben Bildschirmbreiten zieht. Besser ist es, die Werte jedes Monats übereinander anzuordnen, die Spalten für jeden Monat nebeneinander, die (dann deutlich kürzeren) Zeitverläufe jeder Aktie übereinander und, wenn dann der Platz nicht reicht, Teilgruppen von Aktien-Zeitverläufen wieder nebeneinander.

Datenelemente werden also zu Teilgruppen zusammengefasst, jede Teilgruppe ist ein Element der nächsthöheren Stufe, und zusammengefasst wird abwechselnd zeilen- und spaltenweise. Dieses Verfahren, das aus kleinen Strukturen durch Wiederholung einer Aktion ("rekursiv") immer größere aufbaut, hat den Namen recursive pattern erhalten.

Sein Vorteil ist, dass es den langen, unübersichtlichen Faden, der einer eindimensionalen Datenreihe (den Kursen einer einzelnen Aktie) entspricht, zu einem breiteren, kürzeren Band verwebt, in dem der ursprüngliche Faden quer verläuft (die Weber würden es den Schussfaden nennen). Die optimale Lösung der Aufgabe, eine - beispielsweise - quadratische Fläche gleichmäßig mit einem Faden zu bedecken, wäre die berüchtigte, fraktale Peano-Kurve (Spektrum der Wissenschaft 03/1992, S. 72). Aber diese Anordnung verschafft dem Betrachter keinen klaren Überblick, obgleich benachbarte Datenpunkte benachbart bleiben (die Peano-Kurve ist stetig). Die Technik recursive pattern bietet eine gute Annäherung an das mathematische Ideal, ermöglicht aber ein besseres Verständnis der Daten.

Ihr Nachteil ist, dass Werte, die - in unserem Beispiel - zum gleichen Zeitpunkt gehören, über das Bildfeld verstreut und für das Auge kaum als eine Einheit wahrnehmbar sind. Besser ist es, die monatsweise zusammengefassten Zeitverläufe im Kreis statt unter- und nebeneinander anzuordnen. In den Kreissektoren (diesmal für fünfzig statt hundert Aktien) verläuft die Zeit von innen nach außen. Hier sind Hochpreisphasen als helle kreisförmige Ringe auf den ersten Blick zu erkennen.

Noch instruktiver wäre es, wenn Aktien mit ähnlichen Verläufen auch nahe beieinander angeordnet würden. Dann würde auf den ersten Blick offensichtlich, wenn Teile des Aktienmarktes einem gemeinsamen Trend folgen und wie groß dieser Trend wäre. Dazu ist ein Maß für den "Abstand" zweier Aktienverläufe zu definieren - je unähnlicher, desto größer der Abstand - und dann die Reihenfolge zu finden, welche die Summe der Abstände aufeinander folgender Verläufe minimiert. Das ist im Prinzip dasselbe wie das Problem des Handlungsreisenden (travelling salesman problem, SdW 04/1999, S. 76), der die Reihenfolge der Städte so wählen möchte, dass die Summe der Abstände aufeinander folgender Städte minimal wird. Da die exakte Lösung dieses Problems für große Anzahlen extrem aufwendig wird, begnügen sich die Visualisierungs-Algorithmen mit einer näherungsweisen Lösung.

Eine einfache Form der visuellen Datenexploration besteht darin, die Knoten eines Graphen so anzuordnen, dass seine Struktur auf den ersten Blick sichtbar wird. Schon für einen kleinen Graphen ist dabei die Unterstützung durch den menschlichen Betrachter hilfreich; denn eine

vollautomatische Analyse der Daten kann zwar die Knoten mit der höchsten Zahl an Kanten oder der ausgeprägtesten Vermittlerfunktion ermitteln, aber nur schwer ein Gesamtbild der Abhängigkeiten verdeutlichen.

Nicht immer sind die Datenmengen so klein wie im Beispiel der Flugzeugentführer (34 Personen mit weniger als hundert Beziehungen). Betrachtet man beispielsweise Verbindungen in Telefon- oder Computernetzwerken, das Kaufverhalten von Kunden im Internethandel oder die Hyperlinks im World Wide Web, so erhält man Graphen mit Millionen von Knoten und Billionen von Beziehungen. Den gesamten Graphen des Internets auf dem Bildschirm darzustellen ist ein hoffnungsloses Unterfangen; aber mit geeigneten Techniken bekommt man gleichwohl einen gewissen Überblick.

Eine gewöhnliche Landkarte der Erde schöpft die zwei Dimensionen, die ein Bildschirm oder ein Stück Papier bieten, bereits vollständig aus (von den Schwierigkeiten der Landkartenprojektion ganz abgesehen); für weitere Informationen zu einem Internet-Knoten - über dessen geografischen Standort hinaus - ist kaum noch Platz. Die im Projekt Skitter verwendete Visualisierungstechnik löst das Problem, indem sie eine der beiden Dimensionen, nämlich die geografische Breite, schlicht nicht wiedergibt und den gewonnenen Freiraum für eine instruktive Anordnung der Knoten nutzt: Je wichtiger der Knoten, desto weiter innen im Darstellungskreis wird er angesiedelt (Kasten unten). Obgleich die Nord-Süd-Richtung dabei völlig vernachlässigt wird, erlaubt die Visualisierung wichtige Rückschlüsse auf die technische Realisierung der Netzwerke, wie zum Beispiel Verkabelung und Router-Platzierung, sowie politische Faktoren wie enge wirtschaftliche und politische Verflechtungen zwischen Ländern.

Selbst wenn die Verbindungen zwischen den Knoten nicht das gesamte Bild bis zur Unkenntlichkeit zudecken würden, kann eine geografisch getreue Darstellung problematisch bis unverständlich sein - dann nämlich, wenn sich an einzelnen Stellen die Knoten so häufen, dass ihre Darstellungen sich zu sehr überlappen. Die Visualisierungstechnik Gridfit schafft Abhilfe, indem sie in dicht besetzten Gebieten die Punkte systematisch, aber nur wenig auseinanderrückt. Nachdem eine überlagerungsfreie Darstellung der Pixel gefunden ist, kann die Entwicklung der zugehörigen Daten über die Zeit betrachtet werden.

Viele Daten, die in industriellen oder wissenschaftlichen Datenbanken abgespeichert sind, haben Hunderte oder sogar Tausende von Attributen. In den meisten Fällen gibt es nicht die zwei oder drei wichtigsten Attribute (oder Kombinationen von Attributen), zu deren Gunsten man alle anderen vernachlässigen könnte; deswegen sind die Daten in einem zwei- oder dreidimensionalen Koordinatensystem nicht angemessen darstellbar. Für solche Fälle ist die Parallele-Koordinaten-Technik geeignet: Jede Koordinate (jedes Attribut) entspricht einer von beliebig vielen parallelen Achsen, die jeweils so skaliert sind, dass alle für diese Koordinate vorkommenden Datenwerte gerade hineinpassen. Ein Datensatz enthält für jede Koordinate einen Zahlenwert; der entspricht einem Punkt auf der zugehörigen Achse, und der Datensatz wird veranschaulicht durch die gebrochene Linie, die diese Punkte verbindet.

Wenn die Daten eine hierarchische Ordnung haben, also in Abteilungen, Unterabteilungen

und möglicherweise deren Unterabteilungen gegliedert sind, wünscht man sich ihre Darstellungen entsprechend dieser Hierarchie gegliedert. Ein nahe liegendes Mittel ist die klassische Baumstruktur: Die Äste geben die Hierarchie wieder, und die Blätter sind die einzelnen Daten. Ein Visualisierungsverfahren namens Treemap ordnet diesen Baum so an, dass die - stets rechteckigen - Blätter so groß sind, wie es ihrer Bedeutung entspricht, in ihrer Anordnung die Hierarchie wiedergeben (Blätter aus derselben Abteilung sind benachbart) und durch ihre Farbe weitere Information tragen.

Daten, die nicht "von Natur aus" zahlenmäßig sind, vor allem Texte, lassen sich durch eine Reihe von Transformationen auf zahlenmäßige Daten abbilden und dadurch in ein Koordinatensystem bringen. Die einfachste Abbildung dieser Art ist das Auszählen aller nicht-trivialen Wörter im Text. Aus der Häufigkeit, in der zwei Wörter in enger Nachbarschaft in einem Text vorkommen, schließt man auf ihre thematische Verwandtschaft; die wiederum wird in ein abstraktes Abstandsmaß umgesetzt - je enger die Verwandtschaft, desto geringer der Abstand. Dann platziert man die Wörter so in einen abstrakten Raum, dass die Abstände stimmen. Dabei ergeben sich Cluster ("Familien") von eng verwandten Begriffen. In diesem - vieldimensionalen - Raum wählt man eine Perspektive, unter der die Cluster einigermaßen gleichmäßig verteilt erscheinen. Am Ende stellt die in den Pacific Northwest National Laboratories entwickelte Visualisierungstechnik ThemeView eine große Menge von Textdokumenten als eine Landschaft dar, deren Berge den am meisten angesprochenen Themengebieten entsprechen.

Wie auch immer die Daten zusammengefasst auf dem Bildschirm des Benutzers landen - er wird sich häufig eine Lupe wünschen: ein Hilfsmittel, das ihn an bestimmten Stellen genauer hinsehen lässt, ohne dass er die Umgebung dieser Stelle aus dem Blick verliert. Verschiedene Techniken des visuellen Data Mining stellen solche Lupen bereit. Wenn der Benutzer nun die gedachte Lupe über das Gesamtobjekt schweifen lässt, muss das Programm ihm in Echtzeit eine etwas andere Detailansicht liefern, was die Fähigkeiten des Systems aufs äußerste fordern kann.

Eine Lupe für Tabellen ist das Programm TableLens, das auf Wunsch des Benutzers bestimmte Zeilen oder Spalten einer Tabelle sichtbar macht, während alle anderen Tabelleneinträge nur durch kleine Balken angedeutet werden. Für geografische Daten oder allgemein große Graphen bietet sich die hyperbolische Verzerrung an: Die interessierende Stelle wird vergrößert, während die Umgebung nicht etwa, wie bei einer echten Lupe, ausgeblendet, sondern an den Rand des Bildes gedrückt wird. So bleibt der Überblick über den Gesamtkontext erhalten.

Hilfreich ist häufig auch eine Kombination verschiedener Visualisierungstechniken. Eine weit verbreitete Technik ist das "Linking und Brushing" (Verknüpfung und Einfärbung). In mehreren Darstellungen derselben Datenmenge werden einander entsprechende Elemente auf gleiche Weise hervorgehoben - typischerweise durch gleiche Einfärbung. Dadurch werden Abhängigkeiten und Korrelationen in den Daten erkennbar. Interaktive Veränderungen in einer Darstellung werden in den anderen sofort sichtbar. Zwei durch "Linking and Brushing" verknüpfte Bilder lassen im Allgemeinen mehr erkennen als die beiden Bilder für sich.

Visuelle Datenexploration wird heute schon erfolgreich eingesetzt in der Betrugserkennung, im Marketing und beim Data Mining in biomolekularen Datenbanken. Fast alle kommerziellen Data-Mining-Systeme sind derzeit dabei, zur Steigerung von Qualität und Effizienz neuartige Visualisierungstechniken in ihre Software zu integrieren. Die Anbindung an traditionelle Techniken aus den Bereichen Statistik, maschinelles Lernen und Operations Research ist eine Aufgabe für die Zukunft. Fernziel ist ein integriertes, leicht bedienbares und verständliches System, das eine schnelle Exploration großer Datenmengen ermöglicht.

Literaturhinweise

Designing pixel-oriented visualization techniques: Theory and applications. Von Daniel Keim in: Transactions on Visualization and Computer Graphics, Bd. 6, Nr. 1, S. 59, Jan.–März 2000.

The gridfit approach: An efficient and effective approach to visualizing large amounts of spatial data. Von D. Keim und A. Herrmann in: Proceedings Visualiza-tion 98, Research Triangle Park, NC. IEEE, 1998, S. 181.

Aus: Spektrum der Wissenschaft 11 / 2002, Seite 88