PREDICTIVE ANALYTICS – DATA MINING

Data Mining und „Big Data“

Mit den technischen Möglichkeiten, sehr grosse Datenmengen in einem attraktiven Zeitfenster zu speichern und zu verarbeiten, ergeben sich neue Anwendungen für Datenanalyse und Data Mining und damit interessante Perspektiven vor allem in den Bereichen Qualitätsmanagement und Optimierung von Effizienz.

Seit einiger Zeit hört und liest man immer mehr zu den Begriffen „Big Data“, „Hadoop“, „Map Reduce“ und Ähnliches. Konferenzen beschäftigen sich zunehmend mit diesem Thema, und Unternehmen machen sich zunehmend Gedanken, ob „Big Data“ auch für sie wichtig wäre.
Was hat es nun aber mit „Big Data“ auf sich? Einerseits lässt sich „Big Data“ – wie der Name schon sagt – durch eine grosse Menge von Daten charakterisieren, oft ein Vielfaches, was traditionellerweise in Unternehmen vorhanden war. Dies ist aber noch nicht alles – „Big Data“ entsteht häufig ausserhalb der bisher üblichen Unternehmensanwendungen, z.B. durch Daten, die während industrialisierter Prozesse anfallen, z.B. mittels Sensoren. Vor allem in der industriellen Fertigung, durch einen gesteigerten Fokus auf die Verminderung von Ausfallzeiten und die Minimierung von Wartungs- und Garantiefällen, drängt sich die Analyse und Nutzbarmachung dieser maschinengenerierten Daten nachgerade auf.
Des Weiteren besteht Big Data häufig auch aus Daten, die bisher noch selten ausgewertet werden, wie unstrukturierten oder halb strukturierten Informationen, z.B. aus Texten, Audio- oder Videodateien.

„Big Data“ ist mit dem Wachsen und der zunehmenden Bedeutung des Internets entstanden – Firmen wie Google, Facebook oder eBay sahen sich mit enormen Datenmengen konfrontiert, die sie speichern und schnell verarbeiten mussten.
Dies erforderte neue technische Ansätze – und diese haben sich eben in Hadoop (Framework für skalierbare, verteilt arbeitende Software) und dem Map Reduce Algorithmus von Google materialisiert. HBase ist eine skalierbare Datenbank für sehr grosse Datenmengen innerhalb eines Hadoop-Clusters, und Hive erweitert Hadoop um Data-Warehouse Funktionalitäten.

> „‚Traditionelle‘ Data Mining-Ansätze haben aber auch schon vieles vorweggenommen, was heute unter dem Etikett ‚Big Data‘ verkauft wird.“

Die innerhalb dieses Frameworks gespeicherten Daten sollten natürlich auch für Analysen zur Verfügung stehen, um sie sinnvoll nutzen zu können – dafür gibt es auch schon verschiedene Ansätze vom Zugriff auf HDFS und Hive-Daten über ODBC bis hin zu massiv paralleler Verarbeitung.

Bei Millionen von Datensätzen spricht man noch nicht von „Big Data“

Allerdings sollte man davon absehen, bereits an Hadoop etc. zu denken, wenn die Abfragen aus dem Warehouse zu lange dauern – auch bei einigen Millionen Datensätzen redet man in der Regel noch nicht von „Big Data“. Performance Probleme können dann mit wesentlich geringerem Aufwand als einer komplexen Hadoop-Implementation gelöst werden, z.B. mit In-Memory Technologien oder schlicht einem besseren Warehouse-Design.
„Big Data“ wird in Zukunft auch für Datenanalysen mit Sicherheit eine wichtige Rolle spielen. Text Mining, das ja schon seit längerer Zeit etabliert ist, gehört eigentlich auch schon zum Big Data-Thema. Verteilte Verarbeitung ist mit Sicherheit eine Technologie, von der wir noch viel hören werden.
„Traditionelle“ Data Mining-Ansätze haben aber auch schon vieles vorweggenommen, was heute unter dem Etikett „Big Data“ verkauft wird: Schon seit geraumer Zeit ist es in Data Mining-Projekten sozusagen selbstverständlich, dass die verschiedensten Datenquellen miteinander verknüpft werden (z.B. Sensordaten von Produktionsgeräten), dass Freitext aufbereitet und in die Analysen miteinbezogen wird oder dass auch Bild- und Tondaten integriert werden. Insofern stellt „Big Data“ eine fast logische Weiterführung der seit einigen Jahren stattfindenden Entwicklung im Analysebereich dar und erweitert diese technologisch erheblich. ●