Ersch. in: HMD : Praxis der Wirtschaftsinformatik ; 46 (2009), 268. - S. 4-5

Michael R. Berthold

Data Mining – Daten suchen und finden?

Data Mining ist im Prinzip ein ungeschickt gewählter Begriff, denn es geht ja nicht darum, Daten zutage zu fördern, das Interesse gilt vielmehr dem Auffinden von Zusammenhängen, die zu neuen Erkenntnissen führen. Viel eher trifft also der leider sehr unhandliche englische Begriff *Knowledge Discovery in Databases* (KDD) zu, allerdings natürlich auch hier mit der unnötigen Einschränkung auf Datenbanken. Da es bei den meisten Data-Mining-Anwendungen um die Entdeckung neuen Wissens oder neuer Information geht, träfe *Knowledge* oder *Information Mining* eigentlich viel besser zu. Data Mining ist aber mittlerweile in den USA ohnehin ein verbrannter Begriff geworden, nachdem es in den Ruf gekommen ist, dazu benutzt zu werden, in der Privatsphäre der Bevölkerung zu schnüffeln. Das hindert aber kaum jemand daran, mit Begeisterung Kundenkarten zu nutzen, die für wenige Gummipunkte präzise Auswertungen des eigenen Kaufverhaltens erlauben. Noch verbreiteter ist die Freigabe persönlichster Informationen auf allerlei sozialen Netzwerken im Internet oder bei irgendwelchen Angeboten, um »freie« Dienste in Anspruch nehmen zu können. Welche Datenschutzrichtlinien dabei gelten und was aus diesen Informationen für Schlüsse gezogen werden können, interessiert dort dann allerdings eher selten.

Alter Wein in neuen Schläuchen?

Wie so oft in neuen, vielversprechenden Gebieten stagnierte die Euphorie nach einiger Zeit aber ein wenig, da viele der initialen Versprechungen doch nicht ganz eingehalten werden konnten. Weder haben die Kundenforscher perfekt herausfinden können, wie man Kunden wirklich zufriedenstellen kann, noch ist es gelungen, die Gesetzmäßigkeiten der Finanzmärkte oder biologischer Systeme komplett zu verstehen – und das, obwohl gerade in all diesen Beispielen gigantische Datenvorkommen vorliegen. Oft werden dann lieber einfachste Kennzahlen oder grafische Zusammenfassungen als Data Mining verkauft, und die Visualisierung über bunte Plots ersetzt die gewissenhafte Auswertung der vorliegenden Datenbestände. Mittlerweile bleibt einem forschenden Data Miner eigentlich nur, sich mit fundierter Forschung zu beschäftigen (und milde belächelt zu werden) oder – und das ist leider die oftmals vorgezogene Lösung – einfach einen neuen Begriff zu erfinden, unter dem man dann – völlig unbeeinflusst von tatsächlich interessanten Problemstellungen – seine alten Arbeiten neu verkaufen kann. Das führt zurzeit zu einem Aufblühen »neuer« Forschungsrichtungen, die eigentlich alle nach wie vor die gleichen Probleme bearbeiten. Teilweise werden diese Arbeiten dann auch noch an andere Gebiete angegliedert und ermöglichen so das Wiedererfinden größtenteils schon bekannter Methoden unter neuer Flagge.

Information und Wissen entdecken helfen!

Aber lassen wir die Wortspielereien und politischen Polarisierungen beiseite: Das Gebiet des Data Mining hat sich in den letzten Jahren kräftig gemausert – vom einfachen, klar vordefinierten Musterfinden sind wir mittlerweile zu komplexen Verfahren gekommen, die in großen, heterogenen Datenquellen potenziell interessante Muster finden. Zunehmend wird allerdings problematisch, dass die klassischen Verfahren nach recht strikt vorgegebenen Mustern suchen. Denn der Benutzer hat immer öfter keine genaue Vorstellung von den zu fin-

denden Mustern, daher muss das Data-Mining-System also beim Auffinden beliebiger, interessanter Zusammenhänge unterstützend wirken (können). Bis zum interaktiven Herausschälen neuer, überraschender und in der Tat interessanter Informationsstückchen ist es allerdings immer noch ein weiter Weg. Es gibt in diesem Bereich jedoch bereits einige sehr interessante Ansätze. Wichtig sind dabei zunehmend zwei Aspekte: die Einbeziehung des Benutzers, um die Konzentration auf das Wesentliche, zurzeit Interessante zu ermöglichen, und die Lernfähigkeit des Systems selbst. Ersteres ist nötig, da es praktisch unmöglich ist, »Interessantheit« zu modellieren. Das ist insbesondere in all denjenigen Fällen schwierig, in denen der Benutzer zu Beginn des Prozesses noch gar nicht exakt erklären kann, welche Arten von Mustern interessant sein könnten, und bei Problemstellungen, bei denen sich die interessanten Aspekte täglich ändern. Letzteres ist wichtig, um dem Benutzer zu ermöglichen, auch deutlich komplexere Anfragen zu stellen. So will man beispielsweise nach bestimmten Zusammenhängen in Bildstrukturen suchen. Dazu muss der Benutzer die interessanten Strukturen einfach skizzieren können und nicht gezwungen sein, sie computerkonform zu beschreiben. Das erfordert aber natürlich ein interaktives (maschinelles) Lernen der für diese Art von Bildstrukturen wichtigen Beschreibungsebenen neben der Modellierung des passenden Klassifikationsmodells.

Zunehmend entwickeln sich echte Data-Mining-Systeme also in Richtung interaktiver, explorativer Toolboxen, die es erlauben, unterschiedlichste, oft hochkomplexe Data-Mining-Algorithmen ohne Spezialistenwissen einfach und schnell einsetzen zu können. Aber letztlich gilt auch hier die alte Regel: Ohne eine einfache und intuitive Integration der vielen Datenquellen gehen auch solchen Systemen schnell die Daten aus. Diese Hausaufgabe müssen in der Tat noch viele Hersteller von Data-Mining-Systemen in Angriff nehmen, bevor wir existierende und neue Data-Mining-Methoden *wirklich* ausreizen können.

Prof. Dr. Michael R. Berthold
Nycomed-Lehrstuhl für Bioinformatik und Information Mining
Universität Konstanz
FB Informatik und Informationswissenschaften
78484 Konstanz
Michael.Berthold@uni-konstanz.de
www.informatik.uni-konstanz.de