Tobias Aubele

BIG DATA IN DER WEBANALYSE

Die Aussage „Gott würfelt nicht“, welche Albert Einstein zugeschrieben wird, ist sicher gewagt, dennoch zeigt sie, welche mächtigen Möglichkeiten der Statistik seit jeher zugeschrieben werden. Aufgrund einer kontinuierlichen Entwicklung der technischen und statistischen Möglichkeiten erschließen sich durch Big Data für die Webanalyse ungeahnte Einblicke in die Kunden bzw. potenziellen Kunden. Sei es die Optimierung von Marketingausgaben, die Prognose von Zahlungsstörungen oder das frühzeitige Erkennen und Bewerten von einschneidenden Veränderungen: Big Data kann hierfür eine ausgezeichnete Quelle sein.

Webanalyse ist seit Jahren die Basis aller E-Commerce-Aktivitäten und ihr Detaillierungsgrad bietet einen großen Vorteil gegenüber der traditionellen Werbung via Katalog oder Zeitungsanzeige. Mittels technischer Systeme lassen sich Online-Marketing-Maßnahmen sehr granular messen. Jede Mausbewegung, jeder Klick bzw. jedes ausgespielte Banner (View) kann aufgezeichnet werden und liefert dadurch eine umfassende Erfolgsbetrachtung sowie die Basis für zukünftige Entscheidungen. Dennoch wird in vielen Unternehmen nur ein Bruchteil der zur Verfügung stehenden Daten genutzt und schwache Signale, welche in den Daten versteckt sind, bleiben unerkannt.

Analysten erzeugen aus Daten Erklärungen

Die besondere Macht von Daten wird deutlich, wenn nicht nur über die Vergangenheit berichtet wird, sondern im Idealfall Erklärungen möglich sind und Prognosen erstellt werden können. Analysten haben in den Unternehmen die Möglichkeit, aus Tausenden von Interaktionen wie bspw. Einkäufen Profile abzuleiten und anschließend sehr passende, personalisierte Angebote zu unterbreiten. So wusste der Analyst eines Einzelhändlers noch vor dem Vater, dass dessen Tochter schwanger war. Basierend auf einer Veränderung ihres Einkaufverhaltens (v. a. unparfümierte Seife, viele Wattebällchen) konnte ihre Schwangerschaftswoche relativ genau prognostiziert werden. Sie bekam anschließend automatisch über den Algorithmus passende Gutscheine für Babyprodukte zugesendet. Der Vater beschwerte sich beim Händler über die sehr unpassende Werbung. Zu Unrecht, wie sich für ihn später herausstellte. Tipp: Wei-

DER AUTOR

Tobias Aubele ist wissenschaftlicher Mitarbeiter und Dozent im Studiengang E-Commerce an der Hochschule Würzburg-Schweinfurt sowie Doktorand im Bereich Konsumpsychologie an der University of Gloucestershire.

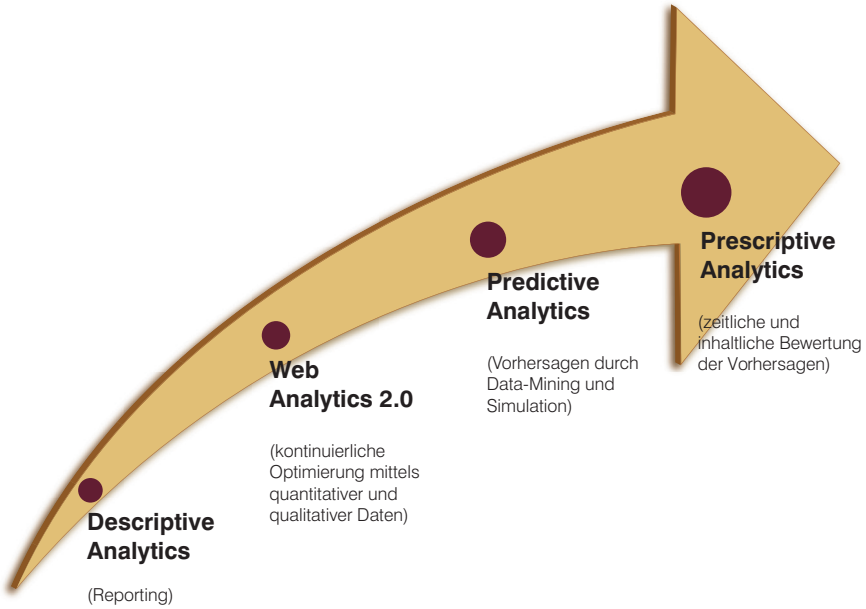

Abb. 1: Reifegrad der Webanalyse

tere interessante Einblicke zur Vorhersage von Kundenverhalten siehe *http://einfach.st/nyt3*.

Wie jede Technologie unterliegt die Webanalyse ebenfalls einem Wandel bzw. einer Reifung (siehe Abb. 1). Infolge der Vielfältigkeit verfügbarer Datenquellen sowie des technischen Fortschritts hat sich in den letzten Jahren die Webanalyse massiv verändert. Diese Entwicklung soll im weiteren Verlauf kurz beleuchtet werden. Anschließend wird auf die erweiterten Möglichkeiten durch Big Data eingegangen.

Reifegrad der Webanalyse

Die E-Commerce-Welt wird immer komplexer. Neue Online-Marketing-Möglichkeiten erblicken in immer kürzeren Zyklen das Licht der Welt. Damit wird die Chance, das Geld ohne umfassende Datenanalyse „richtig" zu investieren, statistisch gesehen immer geringer. Kommen durch Multi-Channel-Organisation noch Kataloge und Filialen ins Spiel, wird es für den Händler interessant: In welchen Kanal investiere ich welchen Anteil meines Budgets? Wo erreiche ich meine Zielgruppe am besten? Welchen prozentualen Anteil meiner Zielgruppe erreiche ich derzeit über welchen Kanal? Und anschließend die finale Frage im E-Commerce: Welcher Anteil des Budgets entfällt auf die Teilfunktionen CRO, SEA, SEO, RTB etc.?

Es stellt sich die Frage, ob in all dieser Komplexität das eingesetzte Business-Intelligence- oder Webanalyse-Tool die einzige Quelle der Wahrheit ist bzw. sein kann. Eher nicht, da das Bild des Kunden nur klarer wird, wenn er aus verschiedenen Blickwinkeln betrachtet wird. Google Analytics zeigt nur, welche Seiten aufgerufen wurden, jedoch nicht, was der Kunde gern aufgerufen hätte und was er in Zukunft voraussichtlich erwartet.

Multiplicity – Datenvielfalt als Schlüssel der Erkenntnis

Die (Daten-)Vielfalt – Multiplicity – ist nach Avinash Kaushik, Autor der Pflichtlektüre Web Analytics 2.0, das Zauberwort. Dieser Ansatz steht inhaltlich als krasser Gegensatz zum klassischen Data-Warehouse, bei dem versucht wird, alles zentral in einer Datenbank zu speichern. Für E-Commerce-Verantwortliche sind insbesondere unstrukturierte Daten anderer Quellen (Facebook, Twitter, Blogs, Bewertungen) von besonderer Bedeutung, da sie in großem Maße qualitative Aussagen zum Unternehmen bzw. zur eigenen Person treffen bzw. offenbaren. In den klassichen ERP-Systemen sind hierfür meist keine Eingabefelder vorgesehen ;-)

Nach Kaushik ist das Ziel klar: Im Idealfall wollen erfolgreiche Webseitenbetreiber zwei Fragen an den Besucher positiv beantwortet haben und ihn damit zufriedenstellen bzw. zum Kunden konvertieren:

1. Warum bist du hier?
2. Warst du in der Lage, dein Anliegen/deinen Wunsch zu erfüllen? Wenn nein, warum nicht?

That´s it! Wenn das gelingt, hat die Webseite die Aufgabe zu 100 % erfüllt Um dies tun zu können, müssen die Händler die Kunden verstehen: Sie müssen mit ihnen statt zu ihnen sprechen. Dieser Unterschied beeinflusst den Erfolg maßgeblich. Hier bedarf es der Analyse, der qualitativen Befragung und des Testens. Darüberhinaus sind „leider" die Mitbewerber permanent zu berücksichtigen, da sie selbst mit proaktiv an Optimierungen arbeiten. Um erfolgreich zu sein, bedarf es damit der Analyse/Berücksichtigung vieler Daten aus möglichst vielen unterschiedlichen Quellen.

Die Elemente und Denkweise der Webanalyse 2.0

Waren es zu Beginn die Klicks, die für die Webanalysten als alleiniger Indikator galten, sind es heute Erkenntnisse aus sowohl quantitativen als auch qualitativen Daten. In einem kontinuierlichen Prozess versuchen die Unternehmen, so viel wie möglich von der Zielgruppe, den Menschen, zu erfahren und ihre Aktivitäten auf deren Bedürfnisse auszurichten. Mittels klassischer Tracking-Tools wie Google Analytics, der Berücksichtigung vorhandener Kundendaten aus dem ERP bzw. Data-Warehouse sowie Befragungen und Webseiten-Tests wird versucht, tiefe Einblicke in den (potenziellen) Kunden zu bekommen. Idealerweise stehen dem Unternehmen Personas bzw. Erkenntnisse aus dem Neuromarketing als hilfreiche Hypothesengeneratoren für Test zur Verfügung. Tools wie bspw. Sistrix und Searchmetrics bieten eine hervorragende Datengrundlage, die Entwicklung des Marktes/der Mitbewerber

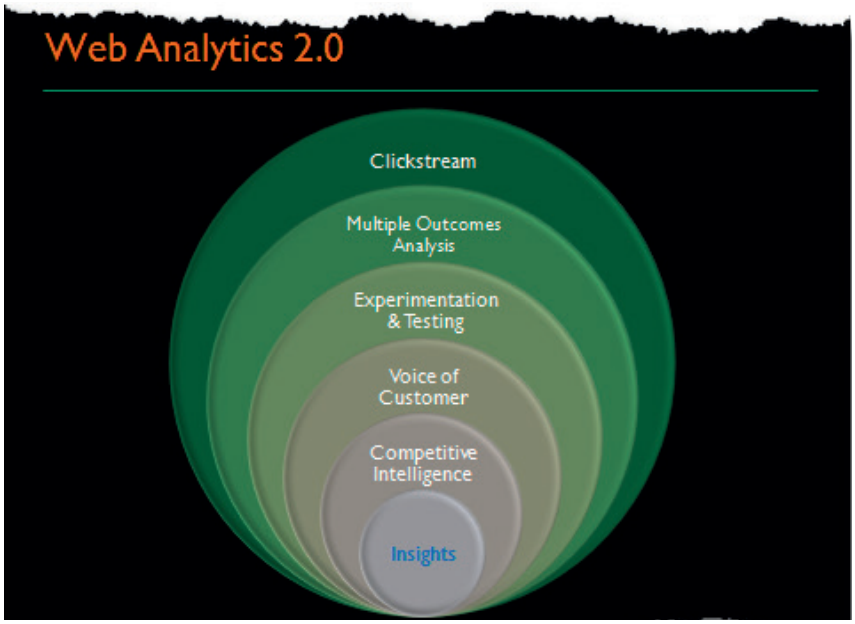

Abb. 2: Web Analytics 2.0 *(Bildquelle: www.kaushik.net)*

und damit den eigenen relativen Standort zu bestimmen. Kaushik spricht hier von Webanalye 2.0, der datengetriebenen, kontinuierlichen Weiterentwicklung der Unternehmensaktivitäten, bei der die Kunden den Ton angeben (siehe Abb. 2). Unternehmen, die den Sprung in die Denkweise von Webanalyse 2.0 noch nicht gewagt haben, verschenken mit großer Wahrscheinlichkeit ein enormes Potenzial.

Predictive Analytics – der Blick in die Zukunft

Die Vergangenheit kann gut durch die deskriptive Webanalyse dargestellt werden. Die Webanalyse 2.0 eröffnet durch die bessere Kenntnis der Zielgruppe und des Marktes Indizien, was auch in Zukunft bedeutsam wird. Schwierigkeiten bereitet immer der möglichst exakte Blick in die Zukunft. Im E-Commerce stellen sich hierzu bspw. Fragen nach der Kaufwahrscheinlichkeit neuer Artikel, der Zahlungswahrscheinlichkeit des Kunden, der möglichen Abwanderungsgefahr zum Mitbewerber. Neben der Zuverlässigkeit der Prognose werden diese Fragen im Idealfall in sehr schneller Zeit beantwortet. Wenn der Kunde am Telefon ist, sollte das System in Realtime Empfehlungen bzw. Angebote unterbreiten, um bspw. den Wechsel zum Mitbewerber zu verhindern oder ein Up-selling zu ermöglichen. Um Erfolg zu haben, ist Relevanz in kurzer Zeit gefragt.

Für statistisch valide Vorhersagen zukünftiger Entwicklungen werden meist Algorithmen/Methoden aus dem Bereich des Data-Minings eingesetzt. Klassische Beispiele sind Regressionen, Entscheidungsbäume oder neuronale Netze. Gemeinsam ist diesen Methoden, dass die Maschine mittels historischer Daten versucht, ein Muster zu erkennen, welches bspw. einen Käufer vom Nichtkäufer unterscheidet/diskriminiert. Diese Muster münden in ein Modell. Das Modell bekommt in Trainingsdaten Informationen zu Kunden und jeweils die Information, ob der Kauf stattfand oder nicht. Die Maschine lernt selbstständig und entwickelt schlussendlich das Modell sukzessive weiter.

TIPP

Die kostenlosen Open-Source-Softwares R *(http://www.r-project.org/)* oder RapidMiner *(http://rapid-i.com/)* können interessante Alternativen zu kostenpflichtigen statistischen Programmen wie SPSS sein.
Des Weiteren stellt die berühmte Universität von Stanford ein kostenloses umfassendes E-Book „Mining of massive Datasets" zur Verfügung *(http://einfach.st/ilab5)*.

Neuronale Netze als Möglichkeit der Prognostizierung

Unternehmen setzen neuronale Netze bspw. zur Absatzprognose je Artikel, Adressselektion für Print-Anstöße oder Personaleinsatzplanung im Call-Center bzw. in der Filiale ein. Neuronale Netze bilden die Funktionsweise des menschlichen Gehirns nach und suchen nach Verbindungen, die für einen Kauf/Nichtkauf sprechen. Hierbei wird das System mit einer Vielzahl historischer Daten (Variablen) gespeist. Über ein Training ergibt sich eine modelltechnisch optimale Topologie, d. h. eine Struktur des Netzwerkes mit einer oder mehreren verdeckten Schichten und den jeweiligen Gewichten der Einflussvariablen (siehe Abb. 3). Wendet man anschließend neue, unbekannte Daten an, kann eine statistische Zielprognose über Kauf/Nichtkauf abgegeben werden.

Was ist Big Data und welche Möglichkeiten ergeben sich daraus?

Wie bereits angesprochen, entwickelt sich die Webanalyse kontinuierlich weiter und bekommt durch Big Data kompetente Unterstützung, insbesondere für den Bereich der Prognosen. Leider nutzen nur wenige Firmen intensiv die Möglichkeiten von Big Data. Doch was bedeutet Big Data, das derzeit euphorisch als „das neue Öl" (Clive Humby) bezeichnet wird? „Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it …" (Dan Ariely).

Big Data ist dadurch gekennzeichnet, dass es sich um sehr große Datenmengen handelt, die aus vielen Quellen in strukturierter und unstrukturierter Weise vorliegen und schnell erhoben bzw. verarbeitet werden. Der Begriff ist damit relativ gemeint, da es keine definierte Größe an Tera- oder Peta-

byte gibt, ab wann von Big Data gesprochen wird. Gekennzeichnet ist Big Data durch die 3 V – Volume, Variety und Velocity. Die Herausforderung ist dabei nicht die Speicherung der Vielzahl von Daten, welche in den letzten Jahren deutlich kostengünstiger möglich wurde, sondern Methoden zu finden, welche schwache Signale aus der Vielzahl an Daten extrahieren. Bestehende Daten werden darüber hinaus sukzessive erweitert (bspw. Wetter, Kaufkraft, Social Media). Dieser Prozess sollte dabei möglichst in Echtzeit stattfinden.

Kontinuierliche Erweiterung der Datenbasis, insbesondere durch unstrukturierte Daten

Unternehmen haben in Ihren ERP bzw. Shopdatenbanken meist sehr strukturierte Daten vorliegen, auf die in den Analysen zurückgegriffen wird. Strukturierte Daten folgen einem definierten Format, wohingegen unstrukturierte Informationen inhaltlich nicht klassifiziert sind. Dies bedeutet, dass zusätzlich zur eigenen Webseite einem Unternehmen sehr viele Informationen vorliegen (bspw. aus Social Media, Kundenbriefen, Befragungen, statistischen Ämtern, Public Data), deren Inhalt bislang nicht in der klassischen Analyse berücksichtigt wurde. Es muss im ersten Schritt die technische Grundlage bei den Unternehmen geschaffen werden, damit auch sehr große Mengen von unstrukturierten Daten sinnvoll verarbeitet werden können. Gleichzeitig müssen umfassende Berechnungen zeitkritisch durchführbar sein, was im Normalfall durch eine parallele Verarbeitung und Virtualisierung der technischen Infrastruktur gewährleistet wird. Im Hinblick auf (Web-)Analyse hat sich die Technologie in den letzten Jahren massiv geändert. Folgende Entwicklungen begünstigen den derzeitigen Fortschritt mittels Big Data:

1. **Cloud Computing und günstig werdende Hardware**. Unternehmen können eine flexible Infrastruktur kostengünstig mit hoher Ausfallsicherheit nutzen und dabei nahezu beliebig skalieren. Amazon und Google bieten hier umfassende Services an.
2. **Nicht-relationale Datenbanken.** Diese Datenbanken folgen nicht der klassischen Tabellen und Schlüssel-Logik und verarbeiten damit die Daten schneller und skalieren besser. Beispiele sind MongoDB und CouchDB (jeweils Open Source) sowie HBase.
3. **Hadoop.** Hadoop ist ein Open-Source-Framework *(http://hadoop.apache.org/)*, welches große Datenmengen (strukturiert und unstrukturiert) über verteilte Cluster mittels des MapReduce-Algorithmus berechnen kann. Server können durch Redundanzen flexibel hinzugefügt bzw. entfernt werden.
4. **Open Source**. Es gibt sowohl für Datenbanken/Systeme als auch für Analysezwecke den Trend zur „kostenlosen" Software, die von einer riesigen Anzahl von Entwicklern, der Community, permanent erweitert und verbessert wird.

Welche Vorteile bietet Big Data für die (Web-)Analyse

1. Schnelle Verarbeitung mit Rohdaten/beliebigen Segmenten

Ad-hoc-Auswertungen und spezielle Fragestellungen gehören zur täglichen Arbeit von Webanalysen. Um sinnvolle Aussagen (über die Zukunft) treffen zu können, sollten zum einen viele Datenquellen und zum anderen immer die jeweilige Nutzungssituation des Besuchers berücksichtigt werden (bspw. mobil, wiederkehrend). Segmente bzw. Filter führen zu sehr performanceintensiven Auswertungen bzw. mit der großen Anzahl an Daten kommen viele Analysesysteme nicht zurande und berechnen entweder über Stichproben

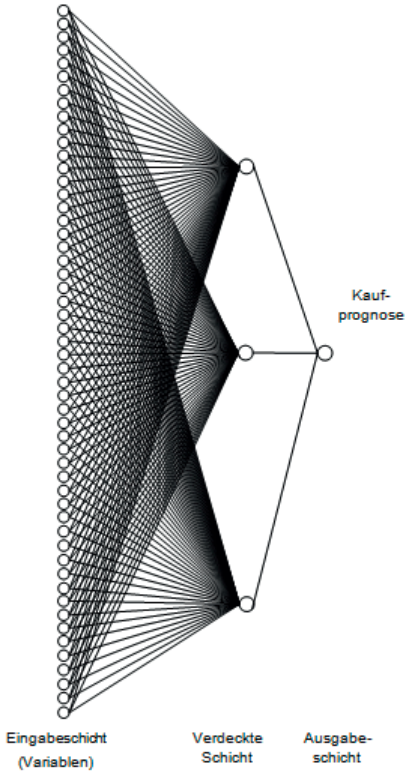

Abb. 3: Schematischer Aufbau eines neuronalen Netzes

bzw. mit extrem langen Verarbeitungszeiten. Aufgrund der technologischen und algorithmischen Weiterentwicklungen kann durch Big Data auf alle Daten zurückgegriffen werden, was zu tieferen Einblicken führt. Gleichzeitig erhält man breitere Einsicht, da neue Datenquellen in Betracht gezogen werden können.

2. Generierung von Wissen durch inhaltliche Textanalyse/Muster

Ein großer Teil der gespeicherten Daten liegt unstrukturiert in textueller Form vor (Social-Media-Daten, Kundenbewertungen, Briefe etc.). Mittels Algorithmen können die relevanten Informationen extrahiert und inhaltlich bewertet werden. Den Unternehmen steht damit eine Vielzahl zusätzlicher Informationsquellen zur Verfügung, die bislang noch nicht bzw. ungenügend in die Analysen einflossen. Insbesondere die Daten und die Verarbeitungsgeschwindigkeit von Social Media („Wer spricht wie über mich?") sind für ein Unternehmen von großer Bedeutung.

3. Frühwarnsystem

Durch die Vielzahl zur Verfügung stehender Daten können schwache Signale aus den Daten besser abgeleitet bzw. überhaupt entdeckt werden. Teilweise sind Veränderungen erst sichtbar, wenn sie über a) einen langen Zeitraum und b) aus verschiedenen Perspektiven betrachtet werden. Durch die Möglichkeit, auf eine weite Zeitstrecke in Kombination mit Visualisierungstechniken von Big Data zurückzugreifen, werden positive und negative Entwicklungen sichtbar, die mit der herkömmlichen Technologie nicht erkannt würden.

4. Optimierung Predictive Analytics hin zu Prescriptive Analytics

Data-Mining und Predictive Analytics sollten in keiner Toolbox von (Web-)Analysten fehlen. Die Kombination mit Big Data bietet die Möglichkeit, noch bessere Ergebnisse zu erlangen, da deutlich mehr Daten im Modell eingesetzt werden können. Darüber hinaus können diverse analytische Modelle und Berechnungen kombiniert werden, was zu einer weiteren Steigerung der Qualität der Vorhersage führt. Der nächste große Schritt der Analyse, Prescriptive Analytics, wird maßgeblich durch Big Data eröffnet. Es geht hierbei nicht mehr nur darum, die Zukunft vorherzusagen, sondern die Ergebnisse der Vorhersage und deren Auswirkungen zu beurteilen. Im Idealfall stehen als Ergebnis der Vorhersage/Simulation mehrere Möglichkeiten zur Verfügung, die es inhaltlich zu bewerten gilt. Prescriptive Analytics liefert damit Antworten auf die Frage: Was ist die beste Lösung? Was wäre die beste Wahl unter den möglichen? Welche Vorteile und Risiken sind jeweils damit verbunden?

Umsetzung

Big Data und die erweiterten Möglichkeiten der Analyse können mittels einer Vielzahl von Open-Source-Systemen/-Software betrieben werden. Ein effizienter und effektiver Einsatz setzt umfassendes technisches und statistisches Wissen voraus. Das bedeutet, jedes Unternehmen muss für sich entscheiden, ob, wann und wie es sich mit den Möglichkeiten von Big Data beschäftigt. Neben unzähligen Quellen im Netz bieten diverse Anbieter die Möglichkeit, bei der individuellen Big-Data-Implementierung bzw. Strategiefindung zu beraten.

An dieser Stelle ein kurzer Hinweis zum Datenschutz: Nicht jede Datenquelle, die verarbeitet werden kann, darf nach deutschem Recht verarbeitet werden. Eine rechtliche Prüfung, welche Daten erhoben und wie sie verarbeitet werden sollen, ist damit unerlässlich.

TIPP

Der BITKOM-Arbeitskreis Big Data bietet als Ausgangspunkt einen kostenlosen Leitfaden mit Praxisbeispielen auf 82 Seiten an *(http://einfach.st/bitkom3)*.

War for Talents – der Data Scientist

Mit Big Data können Unternehmen sehr viel Geld einsparen bzw. verdienen, am Ende bzw. am Anfang sind es jedoch immer Menschen, die die Intelligenz in die Systeme einbringen bzw. umsetzen. Big Data bzw. moderne (Web-)Analyse benötigt Menschen, die einen bunten Strauß an Fähigkeiten mitbringen.

Das sind sowohl technische, betriebswirtschaftliche als auch mathematische Facetten, die ein idealer „Data Scientist" zeigen sollte, in dem lt. Harvard Business Review „sexiest job of the 21st century". Diese Spezialisten werden in der Zukunft sicher sehr gesucht und daher besteht dringender Handlungsbedarf in den Unternehmen, diese Menschen zielgerichtet zu suchen und zu entwickeln. Kaushik forderte bereits 2006 die 10/90-Regel. Für jeweils 10 €, die die eingesetzte Analysetechnik kostet, sollten zusätzlich 90 € in „intelligente Ressourcen/Analysten" investiert werden. Im Zeitalter von Big Data und Predictive Analytics mag dies zwar veraltet klingen, da viele Entscheidungen durch maschinelles Lernen vorbereitet werden. Schlussendlich sind es immer die Menschen, die Algorithmen aufstellen, modifizieren und die Analysen bewerten.

Fazit:

Durch die neuen Technologien besteht die Möglichkeit, umfassend Daten aller Art relativ kostengünstig und schnell zu generieren, auszuwerten und für die Unternehmensentwicklung einzusetzen. Chancen und Risiken können frühzeitig erkannt werden und damit kann ein enormer Wettbewerbsvorteil generiert werden. Die Kombination des Wissens und der Fähigkeiten der IT, des Analysebereiches und der Vertriebswelt kann Zusammenhänge offenbaren, was das Unternehmen nachhaltig positiv beeinflusst. Schafft es das Unternehmen, nicht nur die Zukunft abzusehen, sondern kundenindividuelle Strategien der Kundenüberzeugung zu schaffen (predictive persuasion), sieht es positiven Zeiten entgegen. Dass dies keine Fiktion ist, hat Obama in seinem Wahlkampf gezeigt, der maßgeblich durch Big Data beeinflusst wurde. Hier wurde nichts dem Zufall überlassen.

In diesem Sinne: Überlassen auch Sie nichts dem Zufall. Befassen Sie sich mit der Zukunft. Jetzt.

In der nächsten Ausgabe werden die umfassenden Möglichkeiten der Prognose mittels Excel mit praktischen „Hands-on"-Tipps aufgezeigt: „Predictive Analytics mit Excel – einfach wirkungsvoll."

RapidMiner –

Data-Mining ohne Programmierkenntnisse

… so beschreibt sich die Open-Source-Software selbst, was nach einiger Übung auch tatsächlich möglich ist.

Neben dem Download der Software gibt es unter *http://rapid-i.com* viele Informationen zum Programm (Handbücher, E-Books, Videos), welche den Einstieg erleichtern. Nach dem Programmstart und der einmaligen Anlage eines Repository, d. h. einer Arbeitsumgebung, kann es mit der Analyse schon losgehen. Die Funktionsweise von RapidMiner ist zweistufig aufgebaut. Im ersten Schritt wird der Analyseprozess vom Anwender designt, d. h. die logische Abfolge wird festgelegt. Auf dieser Basis wird dann im zweiten Schritt das berechnete Analyseergebnis in diversen Formaten ausgegeben. Diese beiden Schritte sowie der Startbildschirm werden über die drei Schaltflächen im Bereich Perspektive erreicht (siehe Abb. 1).

Design der Analyse

Das zentrale Element von RapidMiner ist die Gestaltung des Analyseprozesses. Hierzu bedarf es der logischen Aneinanderreihung von Operatoren (Arbeitsschritten). Diese Operatoren haben jeweils eine spezifische Funktion wie bspw. einen Entscheidungsbaum abzuarbeiten oder zwei Excel-Dateien zu verbinden (Join). In Abb. 1 wurde bspw. nach der Zusammenführung von zwei Dateien ein neues Attribut ermittelt (Generate Attributes). Dieses neue Feld (Attribut) kann bspw. das Ergebnis der Multiplikation von Preis aus Excel-Datei 1 und Mengeninformation aus Datei 2 sein. Anschließend wurde über Select Attributes eine Auswahl an Feldern getroffen, welche für den Entscheidungsbaum verwendet werden sollen. Der Entscheidungsbaum ist schlussendlich das optimierte Modell. Die Operatoren erhalten über Verbindungslinien die Information, woher die Daten kommen (links) und an wen die Ergebnisse abgegeben werden (rechts). Jeder Operator hat zur inhaltlichen Definition diverse Parameter, die definiert werden müssen. Dies kann bspw. der Speicherort der Excel-Datei sein oder das gewünschte Konfidenzlevel des Entscheidungsbaumes (siehe Bereich Parameter in Abb. 1). Die anschließende Berechnung des Modells erfolgt über das große blaue Play-Icon.

Tipp: Sofern ein komplexer Prozess definiert wird, kann mittels Breakpoints das Ergebnis Schritt für Schritt ermittelt werden. Besonders bei der Fehlersuche ist dies absolut hilfreich.

RapidMiner bietet in seinen Opera-

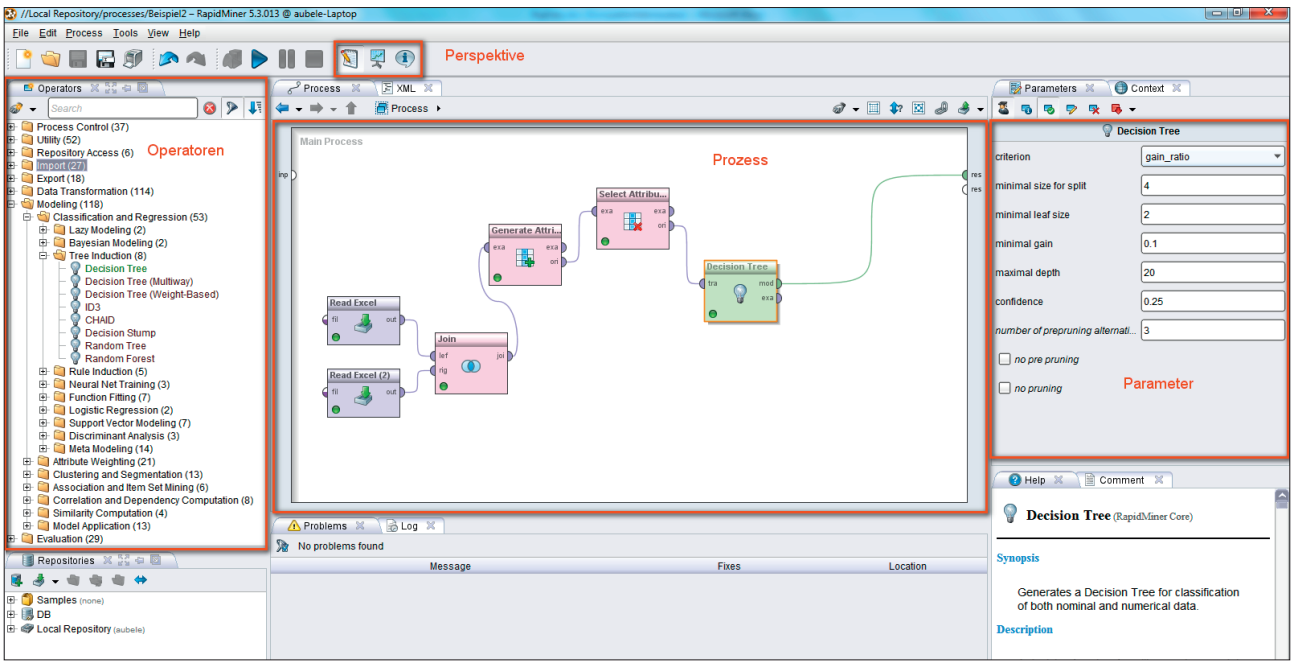

Abb. 1: Definition des Analyseprozesses in RapidMiner

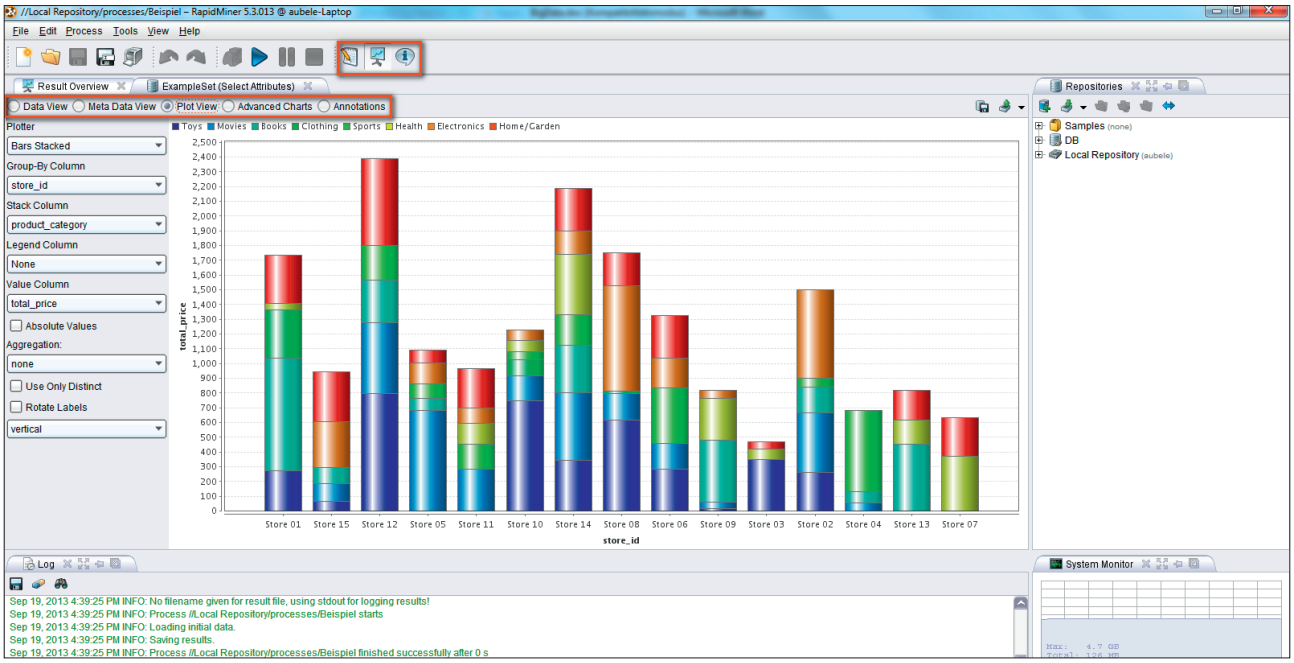

Abb. 2: Beispiel einer Umsatzanalyse nach Ladengeschäft

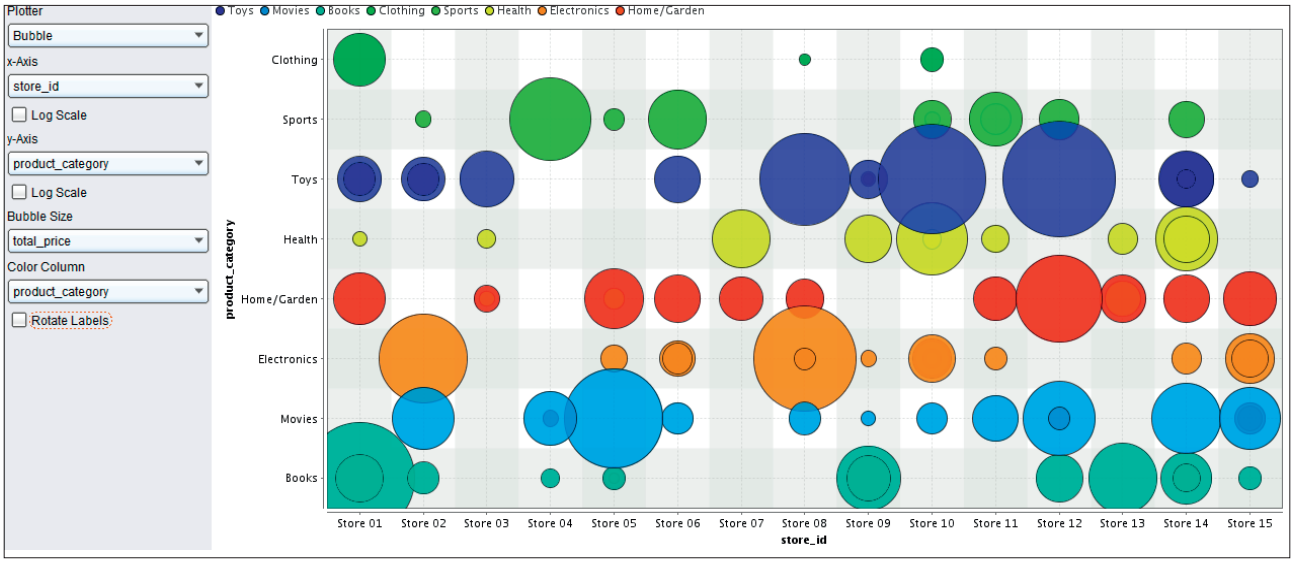

Abb. 3: Beispiel einer Umsatzanalyse nach Segment

toren eine Vielzahl von Modellen an, die statistische Vorhersagen treffen können. Neben den Entscheidungsbäumen sind dies bspw. Regressionen, Diskriminanzanalysen und neuronale Netze. Über die Hilfe werden die jeweiligen Methoden und Parametereinstellungen sehr gut erläutert. Einfach mit einigen Testdaten ausprobieren, kann sehr verblüffende Ergebnisse liefern.

Umfassende Ausgabemöglichkeiten der Ergebnisse

Nach der Berechnung stehen die Ergebnisse entweder als (interaktives) Modell, als Diagramm oder in Form von Zahlen zur Verfügung. Selbstverständlich kann auch eine Datei, bspw. im Format .csv, als Ausgabe definiert werden und von anderen Systemen anschließend weiterverarbeitet werden.

Besonders hervorzuheben sind die Möglichkeiten, die Ergebnisse in unterschiedlichsten Diagrammen darzustellen. Die einzelnen Achsen bzw. Diagrammelemente sind per Drop-down-Felder sehr einfach festzulegen (Abb. 2).

Weitere Diagrammtypen eröffnen die Möglichkeit, sehr umfassend Daten zu analysieren und dennoch die Zusammenhänge stark reduziert visuell darzustellen (Abb. 3).

RapidMiner ist ein hervorragendes Programm, um Daten mittels diverser Data-Mining-Techniken granular zu analysieren. Die grafischen Ausgaben sind ebenfalls sehr weitreichend und helfen, das große Bild aus den Daten aufzubauen. Durch die intuitive Bedienerführung können nach etwas Übung bereits einfache Analysen erfolgen. Hier gilt es, beharrlich zu bleiben und kontinuierlich die eigenen Analysen zu verfeinern. Es lohnt sich.¶

Das erste Mal mit R wird nicht einfach, Ausdauer wird sich langfristig bezahlt machen

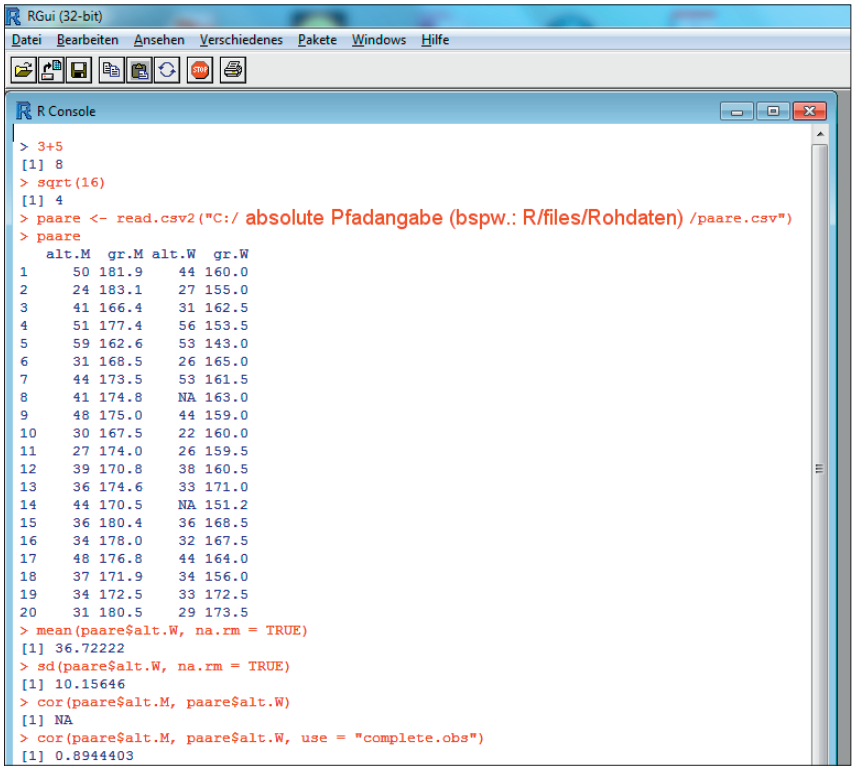

Abb. 1:Befehle für die Verarbeitung in R

R ist genau genommen eine Programmiersprache. Die Oberfläche ist für Programmierer intuitiv, für Nicht-Programmierer aus Usabilitysicht eine Katastrophe (Abb. 1). Gerade durch die Offenheit mittels Programmierung ist R unglaublich mächtig, was es zugleich komplex gestaltet. Unter *http://einfach.st/rload* kann die aktuelle Version 3.0.1 kostenfrei heruntergeladen werden und ist nach der Installation sofort einsatzbereit.

Über die Konsole müssen nun die gewünschten Befehle eingeben werden. R funktioniert u. a. wie ein Taschenrechner, d. h. die Berechnung von 3 + 5 oder die Wurzel aus 16 erfolgt einfach per Direkteingabe. Das Besondere an R ist die Möglichkeit der Nutzung von Objekten. Sobald einem Objekt ein Ergebnis zugewiesen wurde (über den Befehl <-), kann es im Folgenden immer wieder herangezogen werden. In Abb. 1 wurde bspw. dem Objekt „paare“ mittels der Funktion read.csv2 der Inhalt der CSV-Datei zugewiesen, welche auf „*C:/R/files/Rohdaten/paare.csv*“ liegt. In diesem Beispiel sind es das Alter und die Körpergröße von Ehepaaren. Sobald im weiteren Verlauf das Objekt ohne Parameter aufgerufen wird, erscheint der Inhalt dieses Objektes.

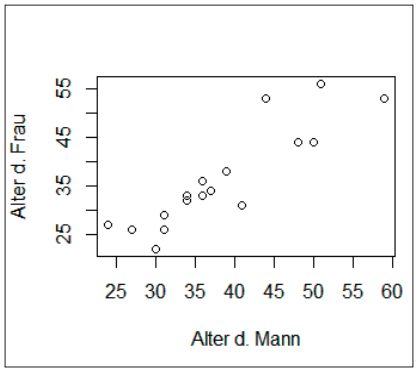

Abb. 2: Plot-Diagramm in R

Nachfolgend können dann sämtliche statistischen Auswertungen angestoßen werden. Die Funktion mean() liefert bspw. den Mittelwert, sd() die Standardabweichung von Datensätzen. Die zu analysierenden Felder werden mit „Objekt$Feldname“ angesprochen. Da in diesem Beispiel zwei Frauen kein Alter angaben, wird „NA“(not available) ausgegeben. Damit die Berechnung mit den restlichen Daten erfolgt, muss den beiden Funktionen der Parameter „na.rm = TRUE“ mitgegeben werden. Spannend ist bspw. die Frage, ob es beim Alter und der Größe einen Zusammenhang zwischen den Ehepaaren gibt (Korrelation). Die Funktion cor() ermittelt beim Alter einen Wert von 0,89 und bei der Größe von 0,37 (Attribut use = „complete.obs“ reduziert wieder auf vollständige Datensätze). Das heißt, es gibt einen sehr starken Zusammenhang zwischen dem Alter der Paare, bei der Größe nur einen schwachen. Ist demnach ein Alter bekannt, so kann das Alter des Partners gut vorhergesagt werden. Interessant, oder?

Dieses Ergebnis kann selbstverständlich auch grafisch mit einer Vielzahl von Darstellungsmöglichkeiten ausgegeben werden. Hierzu ist bspw. die Funktion plot() mit den gewünschten Feldern und Beschriftungen zu versehen (Ergebnis siehe Abb. 2).

Dieses Beispiel zeigt nur einen Bruchteil der Funktionen und Möglichkeiten, welche R bietet. Durch die Programmiersprache können eigene Funktionen, Schleifen und Diagramme definiert werden. Es gibt damit fast nichts, was in R unmöglich ist. Darüber hinaus existiert eine große Community, welche R kontinuierlich weiterentwickelt. Unter dem Menüpunkt Hilfe gibt es umfassende Handbücher. Neben vielen Fachbüchern gibt es auch ein umfassendes kostenloses E-Book als PDF in deutscher Sprache unter *http://einfach.st/rhandb*. Die Hauptseite *www.r-project.org* ist ebenfalls versehen mit Hinweisen, Leitfäden, Wikis und Ideen zu R. Der Start wird sicher holprig, die Ergebnisse/Erkenntnisse werden dafür belohnen. Halten Sie durch!¶