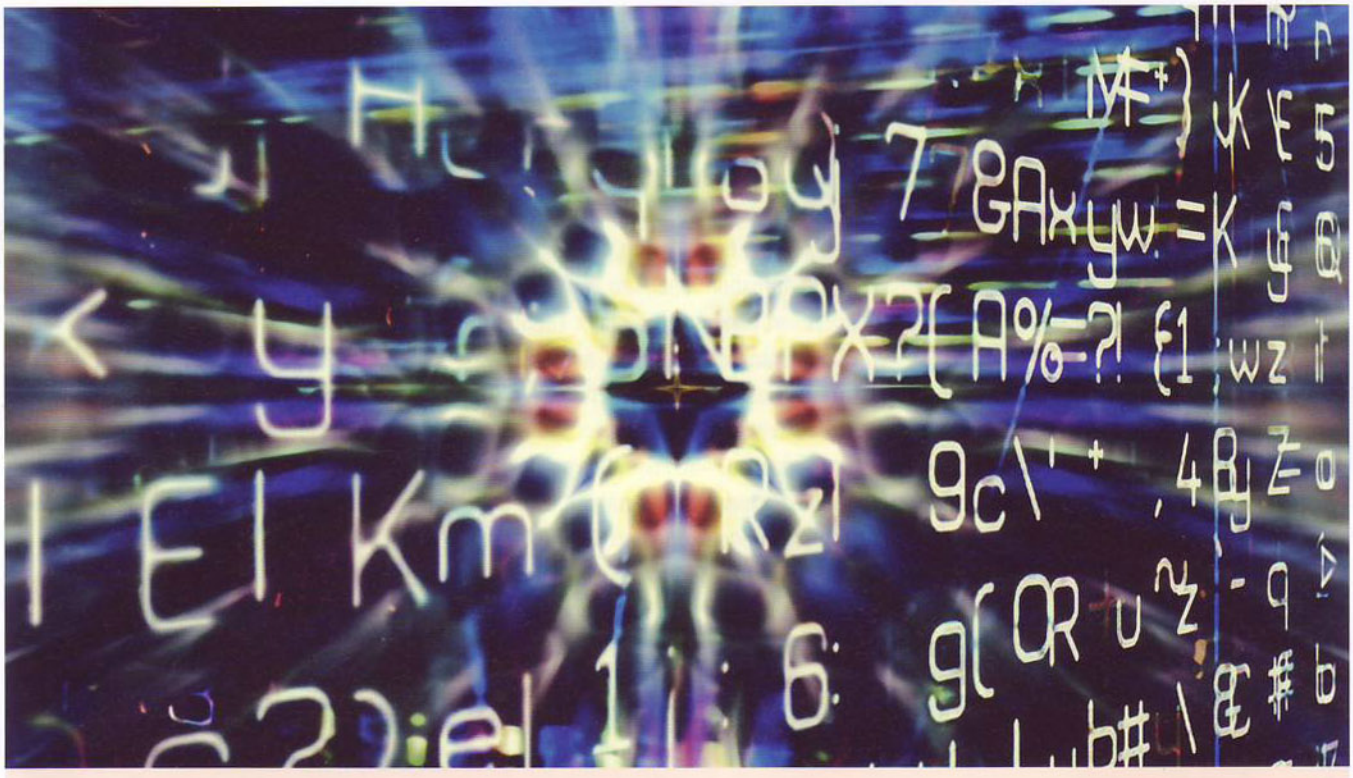

Semantische Analyse und Natural Language Processing – die Big-Data-Analyse-Instrumente der Zukunft

von Frank Zscheile

In allen Wirtschaftsbereichen wachsen im Zuge von Big Data die Informationsmengen und **kommen immer neue Datenformate hinzu.** Dies erfordert neue Analyse-Methoden, um aus der Datenflut relevante Ergebnisse zu generieren. **Semantische Analyse und Natural Language Processing (NLP)** sind in jüngster Zeit als **zukunftsträchtigste Analysemethoden** neben die bisherigen Business Intelligence (BI)-Verfahren getreten, die in der Regel nur numerische Informationen einbeziehen. Controlling-Abteilungen können sich dies zunutze machen.

Viel haben Unternehmen **in den vergangenen Jahren** in Business Warehouse und Business Intelligence investiert, um **strukturierte Daten zu erschließen**. Dabei handelt es sich um Daten, die erkennbar in Tabellen- oder Listenform aufgebaut sind und sich daher leicht mittels mathematischer Verfahren verarbeiten lassen. Finanz- und Controlling-Abteilungen fokussieren in ihren BI-Analysen und Reports traditionell sehr stark auf solch strukturierte Daten.

Jedoch liegt auch bei ihnen ein beachtlicher Teil geschäftlicher Informationen in unstrukturierter Form vor: Projektberichte und Veröffentlichungen als Office-Dokument oder E-Mail, neue Gesetzestexte in PDF-Form etc. In diesen Formaten stecken wichtige Informationen, die für das Controlling und Berichtswesen ebenso relevant sind wie numerische Daten aus BI-Systemen. Je unstrukturierter die Daten sind, desto höher sind die Anforderungen an die zugrundeliegenden Algorithmen zu deren Auswertung. An ihnen hängt demnach, ob dieser Wissensschatz gehoben und für das Controlling nutzbar gemacht werden kann.

Logische Zusammenhänge herstellen

Eine **moderne „Big Data-Initiative"** muss daher beide Datenwelten zusammenführen und neben den strukturierten **auch unstrukturierte Daten in die Suche und Analyse einschließen.** Semantische Analyse und Natural Language Processing (NLP) sind die Werkzeuge dafür. Sie werten Daten nicht nur statistisch aus, sondern bieten vielmehr eine vollständige

Sicht auf strukturierte, unstrukturierte und teilstrukturierte Daten, aus der das Unternehmen neue, wertvollere Erkenntnisse gewinnen kann. Über die reine Ermittlung von Unternehmens-Kennzahlen für eine Business-Scorecard geht Big-Data-Analyse somit heute hinaus. Controller werden in die Lage versetzt, Ergebnisse in einen logischen Zusammenhang mit allen zugehörigen Informationen zu bringen.

Die Auswertung von Text- bzw. von Menschen generierten Daten ist zweifelsohne die größte Herausforderung im Umfeld von Big Data. Hierzu bedarf es einer **tiefgehenden linguistischen und semantischen Analyse**. Erst dadurch lässt sich eine Suchanfrage wirklich verstehen und die Bedeutung eines Textes erfassen. So erhält der Suchende Ergebnisse, die über den Horizont seiner ursprünglichen Keyword-Abfrage inhaltlich weit hinausgehen. Gleichzeitig können **Informationen über geschäftsrelevante Filter kategorisiert** werden. Dies hilft dem Anwender, unter allen von der Suchmaschine als relevant angezeigten Ergebnissen die für ihn entscheidenden sofort zu erfassen. Ermöglicht wird dies durch die Technik des „Natural Language Processing" (NLP) oder auch Computerlinguistik. Such- und Analysewerkzeuge ohne NLP-Technologie werden den heutigen Anforderungen von Unternehmen an Enterprise Search und Big-Data-Analyse nicht mehr gerecht. Die Software des Herstellers Sinequa etwa beinhaltet NLP-Technologie für 20 verschiedene Sprachen, darunter solch „schwierige" wie Chinesisch, Japanisch, Koreanisch oder Arabisch.

Menschliche Sprache oft unpräzise und zweideutig

Unter **NLP versteht man die Fähigkeit eines Computerprogramms, menschliche Sprache so zu verstehen, wie sie gesprochen bzw. geschrieben wurde**. Traditionell versteht eine Software einen Menschen am besten, wenn dieser eine möglichst präzise, eindeutige und strukturierte Sprache verwendet. In der Realität aber ist die menschliche Sprache oft eben nicht eindeutig und genau, sondern hängt von komplexen Variablen ab (sozialer Kontext, regionale Spezifika …). Zum Einsatz kommen NLP-Technologien bevorzugt im Bereich des Enterprise Search, also der organisierten Suche in strukturierten und unstrukturierten Daten innerhalb einer Organisation.

NLP geht über bloße Sprachidentifikation, Worttrennung und Text-Extraktion, wie sie viele Suchmaschinen heute bieten, weit hinaus. Zu den NLP-Aufgaben innerhalb von Software-Programmen gehören zum einen **Techniken wie Satzsegmentierung und -analyse (Parsing)**, also das Aufteilen von Phrasen in verschiedene Teile, um Beziehungen und Bedeutung zu verstehen. Außerdem **Deep Analytics (Datensammlung und -Analyse aus sehr großen Datenpools), Named Entity-Extraktion (Erkennung und Klassifizierung von Bestandteilen eines natürlich-sprachlichen Textes)** und Co-Referenzauflösung.

Im Rahmen linguistischer Analysen ermöglichen diese Techniken eine automatische Extraktion von Begriffen und Navigation in begrifflich geordneten und nach Relevanz sortierten Informationen. **Text Mining** (ein Bündel von Algorithmus-basierten Analyseverfahren zur Entdeckung von Bedeutungsstrukturen aus un- oder schwachstrukturierten Textdaten) mit **Tagging einzelner Wörter und das Erkennen semantischer Zusammenhänge** (etwa bei gleichzeitigem Auftreten der Begriffe innerhalb eines Satzes) werden möglich. Das gewonnene „Unternehmens-Wissen" lässt sich in Form von Wörterbüchern, Taxonomien, Ontologien etc. aggregieren.

Gesetzestexte und Erlasse inhaltlich erschließen

Controlling-Abteilungen sind fast tagtäglich mit neuen Gesetzen, Regelwerken und Erlassen auf Landes-, Bundes und europäischer Ebene konfrontiert, deren Inhalte für ihre Arbeit eine Relevanz haben kann. Solche Daten enthalten üblicherweise eine Fülle von Informationen, die nicht „kodifiziert" sind und sich nicht in bloßen Zahlen ausdrücken lassen. Mit Hilfe der beschriebenen Werkzeuge für die Big Data Analyse erhält der Controller Suchresultate schnell und einfach innerhalb seiner täglichen Arbeitsumgebung – ohne wissen zu müssen, wo sie genau herkommen und welches Format sie haben.

In großen Mengen erzeugte BI-Reports, deren Existenz oder gar Inhalt die Controlling-Abteilung gar nicht im Einzelnen überblicken kann, lassen sich mittels Big-Data-Analyse so durchsuchen, dass der Anwender auch Fundstellen erhält, in denen der eigentliche Suchbegriff gar nicht vorkommt, hingegen Synonyme oder inhaltlich ähnliche Begriffe. Auch zur **Erkennung bestimmter Gefahren-Situationen und Verhaltensmuster** sind die beschriebenen Methoden der semantischen Analyse und NLP ideal geeignet. Im Versicherungsumfeld können sie somit signifikant **zur Risikominimierung** beitragen.

Einsatz im Einkauf

In Beschaffungsabteilungen lassen sich Textanalyse-Methoden für Controlling-Zwecke gut bei der Prüfung von Einkaufsverträgen einsetzen. Dort verklausuliert enthaltene intransparente Preise können damit sichtbar gemacht werden, Dashboards geben eine quantitative wie qualitative Übersicht über die Verträge wieder. So lassen sich auch Einkäufe aufspüren, die an der offiziellen Beschaffungspolitik des Unternehmens vorbeilaufen – ein in der Praxis häufig anzutreffender Fall. Durch Textanalyse der Verträge entdeckt das Controlling versteckte Unregelmäßigkeiten, und „Ausgabenausreißer" lassen sich schnell identifizieren. ■

Autor

■ Frank Zscheile
Presse- und Öffentlichkeitsarbeit
Bergmannstr. 26, 80339 München

Tel.: +49 89 5403 5114
E-Mail: zscheile@agentur-auftakt.de
www.agentur-auftakt.de