

Märkte richtig einschätzen



Die Auswertung von Freitexten hilft Unternehmen dabei, potenzielle Kunden zu identifizieren, relevante Technologieänderungen aufzuspüren oder Produktentwicklungen gezielt an Kundenbedürfnisse anzupassen.

Unstrukturierte Daten – insbesondere in Form von Freitexten – wachsen nicht nur extrem schnell, sondern lassen sich mittlerweile auch sehr effizient auslesen. Eine solche Analyse mittels aktueller SAP-Technologie ist für Unternehmen höchst attraktiv: Denn Freitexte enthalten vielerlei Informationen, die z.B. als wichtige Grundlage für das erfolgreiche Einschätzen von Marktentwicklungen dienen und somit unmittelbar zur Gewinnsteigerung beitragen können.

* Michael P. May ist Geschäftsführer der Sybit & Peers AG.

** Thorsten Niehues ist SAP Consultant Business Analytics bei der Sybit & Peers AG.

Von Michael P. May* und Thorsten Niehues**

Neben der Marktanalyse und dem Erkennen von Trends bietet die Auswertung von Freitexten (u.a. Produktbewertungen, Preserveröffentlichungen und Einträge in Jobbörsen, Blogs oder sozialen Medien) noch viele weitere Möglichkeiten: Diese Maßnahme hilft Unternehmen unter anderem dabei, potenzielle Kunden zu identifizieren, relevante Technologieänderungen aufzuspüren oder Produktentwicklungen gezielt an Kundenbedürfnisse

anzupassen. Damit liefert die Freitextanalyse jeweils elementare Beiträge zur Umsatzsteigerung. Parallel kann sie jedoch auch die Kostenreduktion effektiv unterstützen und auf diese Weise den Unternehmensgewinn weiter erhöhen. Aktuell nehmen unstrukturierte Daten wie Freitexte nicht nur quantitativ stark zu, sondern sind auch in immer größerem Umfang verfügbar. Um diese höchst ergiebige Basis für Analysezwecke optimal nutzen zu können, müssen jedoch neue Wege beschritten werden: Traditionelle bzw. herkömmliche Analyseansätze für strukturierte Daten reichen an-

gesichts von unstrukturierten Big Data nicht mehr aus. In Bezug auf Freitexte bietet sich dabei der Einsatz von „SAP HANA Text Analysis“ als eine maximal leistungsfähige Ideallösung an, die heutige Herausforderungen im Big-Data-Format souverän meistert und Textdaten in vielen verschiedenen Sprachen gleichsam bestmöglich verarbeitet.

Technologische Umsetzung mit SAP HANA Text Analysis

Mithilfe von SAP HANA können Unternehmen zuverlässig aussagekräftige Ergebnisse aus unstrukturierten Textdaten gewinnen: Via „Text Analysis“ bietet die schnelle In-Memory-Plattform entsprechend vielfältige Optionen – vor allem in Sachen Suchfunktionen, Textanalyse und Text Mining. Diese drei Gebiete sind zwar in Anwendungshinsicht separat zu betrachten, auf technischer Ebene aber miteinander verwoben. Zudem verwenden sie größtenteils dieselbe Basistechnologie.

Präzise maschinelle Sprachverarbeitung (engl. Natural Language Processing; NLP) mittels SAP-Technologie ermöglicht komplexe linguistische Analysen, mit denen sich die In-Memory-Daten von Unternehmen wirkungsvoll anreichern lassen. Gleichzeitig besteht dabei die ebenso vorteilhafte Möglichkeit, mittels statistischer Algorithmen bestimmte Muster in großen Dokumentenmengen automatisiert zu erkennen (u.a. Identifizierung von Schlüsselbegriffen) und die analysierten Dokumente effizient zu kategorisieren. Angesichts dessen präsentiert sich SAP HANA jeweils als optimale Plattform: Dank der Vielseitigkeit von Text Analysis werden Freitexte zu einer niemals versiegenden und stetig wachsenden Informationsquelle, die wichtige Grundlagen für eine nachhaltige Steigerung des Unternehmensgewinns schafft.

Marktanalyse als wichtiger Erfolgsfaktor

Umfassende und exakte Marktanalysen zählen allgemein zu den wichtigsten Erfolgsfaktoren von Unternehmen. Innerhalb dieses Anwendungsszenarios eröffnet das Auswerten von Freitexten mit SAP HANA Text Analysis zahlreiche neue Chancen für noch präzisere Prognosen: Presseveröffentlichungen, Qualitätsbeschreibungen in Jobbörsen etc. sind wichtige Indikatoren für Marktentwicklungen. Möglichst umfangreiche Berücksichtigung solcher Texte hilft Unternehmen daher, sich optimal am jeweiligen Markt auszurichten und flexibel

reagieren zu können. Herkömmliches Lesen geht jedoch mit Ressourcenbindung, hohem Aufwand, potenziellen Fehlerquellen und Begrenzung in puncto Kapazität einher. Somit besteht die deutlich sinnvollere Alternative darin, vor allem extrem große Textmengen automatisiert zu analysieren und über entsprechende Algorithmen auszuwerten.

Ein automatisierter Ansatz bietet zudem den Vorteil, dass die Ergebnisse visuell dargestellt werden können: Aussagekräftige Diagramme erleichtern die Interpretation durch Entscheidungsträger sehr stark. Wenn die Ergebnisdaten zusätzlich über einen längeren Zeitraum gesammelt werden, sind Trends über Liniendiagramme auf einen Blick erkennbar. Trends wiederum sind sehr starke Indikatoren für zukünftige Marktentwicklungen. Somit wird die automatisierte Textanalyse zu einem wirkungsvollen Werkzeug für qualifizierte Prognosen, die sich unter anderem für folgende Zwecke nutzen lassen:

- Entscheidungen bezüglich Erschließung neuer Märkte (Umsatzsteigerung) oder sinnvollen Rückzugs aus bestimmten Märkten bzw. Geschäftsfeldern (Kostenreduktion).
- Planung und Ausbildung im Personalbereich: Rechtzeitiges Erkennen von Technologietrends beeinflusst im Idealfall die Unternehmensstrategie und erleichtert die HR-Planung (u.a. Neueinstellungen oder Schulungen in wachsenden Technologien).
- Produktmanagement: Freitexte in Form von Online-Bewertungen (gut, schlecht, neutral) geben Aufschluss darüber, wie Kunden bestimmte Produkteigenschaften beurteilen. Eine entsprechende Auswertung kann die Entwicklung bzw. Produktion daher ganz gezielt und sehr positiv beeinflussen.

Beispielszenario: SAP-Marktindex durch Auslesen von Jobbörsen

Als äußerst flexible Lösung deckt SAP HANA Text Analysis alle erdenklichen Arten von Freitexten ab. Das folgende Beispiel aus der erfolgreichen Praxis erläutert stellvertretend die konkrete Umsetzung dieses automatisierten und höchst effizienten Ansatzes:

Stellenanzeigen repräsentieren unstrukturierte und frei verfügbare Textdaten, die aussagekräftige Informationen zu SAP-Marktrends beinhalten. Das Erkennen

solcher Trends ist unter anderem für Beratungsunternehmen von zentraler Bedeutung – was wiederum Jobbörsen (z.B. im Rahmen des Businessportals XING) zur idealen Basis für entsprechende Maßnahmen mittels aktueller IT-Technologie macht. Die Qualifikationsangaben in Stellenanzeigen sind diesbezüglich elementar, da sie Kompetenzen üblicherweise mittels Produktbezeichnungen beschreiben. Über gezielte Analyse und Auswertung lassen sich somit aktuelle Tendenzen bei Angebot und Nachfrage in bestimmten SAP-Anwendungsbereichen ermitteln. Zu diesem Zweck werden zunächst alle relevanten Textdaten per Java API aus der jeweiligen Jobbörse extrahiert und in SAP HANA gespeichert.

SAP HANA als optimale Plattform

Anschließend erweist sich SAP HANA als extrem leistungsfähige Plattform für die Weiterverarbeitung der gewonnenen Textdaten: Die Funktion „Text Analysis“ profitiert direkt von der schnellen In-Memory-Performance und meistert somit problemlos auch sehr große Datenmengen in sehr kurzer Zeit.

Wie bei allen anderen NLP-Konzepten stellt sich dabei bislang ein computerlinguistisches Grundproblem im Bereich der Tokenisierung (Textsegmentierung auf Wortebene): Bestimmte Wörter und Wortkombinationen werden nicht immer vollständig bzw. korrekt erkannt. Eine zuverlässige Tokenisierung (im konkreten Beispiel von SAP-Produktbezeichnungen) ist jedoch obligatorisch für das erfolgreiche Durchführen aller Folgeprozesse. Somit besteht Bedarf, die genannten Erkennungsdefizite mittels eines benutzerdefinierten Wörterbuchs (Dictionary) zu beheben – wobei rein manuelle Wörterbucheinträge aufgrund der Datenmenge einen kaum vertretbaren Aufwand darstellen. Aufgrund dessen empfiehlt sich zumindest eine teilautomatisierte Wörterbucheinstellung.

In Verbindung mit gezielten Modifikationen bewältigt SAP HANA Text Analysis diese Herausforderungen jedoch äußerst zufriedenstellend. So bietet bereits die Basiskonfiguration eine sehr hohe Trefferquote in puncto korrekte Tokenisierung und Namenserkennung (Named Entity Recognition and Classification; NERC). Die Zahl der fehlerhaft erkannten Tokens bzw. Begriffskombinationen ist daher von vornherein vergleichsweise gering und lässt sich ergänzend durch automatisierte Wörterbucheinträge weiter drastisch reduzieren:

Innerhalb des Beispielszenarios enthalten potenziell inkorrekte Tokens den Begriff „SAP“ oder werden fälschlicherweise als SAP-Produkte eingestuft. Um solche Tokens automatisiert zu identifizieren, kommt eine SQL-Anweisung im Rahmen einer eigens entwickelten Java-Anwendung zum Einsatz. Anschließend werden spezielle Muster auf die identifizierten Textelemente angewandt. Bei Musterübereinstimmung (Pattern Match) entsteht jeweils ein neuer Wörterbucheintrag, durch den das betreffende Token zukünftig korrekt als Produkt erkannt wird (NERC). Durch Normalisierung im Rahmen der Informationsgewinnung (Information Retrieval) werden parallel alle Schreibvarianten mit derselben Entität verknüpft. Dieses kombinierte Gesamtprinzip lässt sich zudem beliebig auf andere gewünschte Kategorien anwenden und ausweiten.

Bezüglich der automatisch erstellten Wörterbucheinträge reicht eine Plausibilitätskontrolle durch IT-Fachpersonal aus; lediglich kritische Einzelfälle sind noch manuell zu überprüfen. Allerdings hält sich dieser Restaufwand sehr stark in Grenzen – insbesondere verglichen mit rein manuellen Wörterbucheinträgen oder herkömmlichem Lesen von Freitexten. Die Kombination aus Automatisierung und manueller Endkontrolle resultiert in einer extrem niedrigen Fehlerquote beim Erkennen und Analysieren von erfassten Daten, die SAP-Produkte in Jobbörsen repräsentieren. Die verbleibende Fehlerwahrscheinlichkeit ist so gering, dass sie bezüglich einzuschätzender SAP-Markt-

trends als nicht relevant betrachtet werden kann.

Reporting und visuelle Darstellung

Da ausschließlich als SAP-Produkte erkannte Elemente für das Reporting verwendet werden, generiert die Textanalyse eine optimale Datenbasis für aussagekräftige Berichte zu Markttrends. Die Ergebnisse der Analyse werden mittels Liniendiagrammen leicht verständlich visualisiert sowie mit einer Produkt- und Zeithierarchie kombiniert. Diese Hierarchien erlauben Einzelbetrachtungen oder Aggregationen hinsichtlich Zeiträumen und Produktgruppen, wobei die oberste Ebene in einer Aggregation aller SAP-Produkte besteht. Die erstellten Diagramme geben Analytikern somit detaillierten Aufschluss über Marktentwicklungen (Angebot und Nachfrage) auf dem Gebiet von SAP-Kompetenzen bzw. -Qualifikationen.

Hohe Flexibilität und multiple Möglichkeiten

SAP HANA Text Analysis eignet sich nicht nur optimal für die genannten Analyseaufgaben, sondern lässt sich mittels Anpassung von wenigen Parametern auch flexibel für das Erkennen anderer Trends verwenden. Zudem hat die Implementierung einen großen und positiven Einfluss auf die gesamte Softwarearchitektur: Durch das Output-Format der maschinellen Textverarbeitung (NLP) können Anwender bei beliebigen anderen Datenverarbeitungsprozessen auf dieselbe Datenstruktur zu-

rückgreifen – unabhängig von der jeweiligen Datenquelle und Art der Einspeisung in SAP HANA. Ebenso vorteilhaft ist die Möglichkeit, Hierarchien innerhalb des Datenmodells zu definieren und auf diese Weise das Standarddatenformat MDX zu generieren: Der entsprechende MDX-Cube kann sehr effizient von diversen Reporting-Werkzeugen genutzt werden.

Fazit

SAP HANA erweist sich als optimale Plattform für maschinelle Sprachverarbeitung (NLP) und Reporting-Zwecke. Das extrem leistungsfähige In-Memory-Konzept geht einher mit hoher Flexibilität und Erweiterbarkeit in Form von zahlreichen Zusatzmöglichkeiten. All dies macht SAP HANA und SAP HANA Text Analysis zu überzeugenden BI-Werkzeugen, die den Unternehmensgewinn dank folgender Vorteile erhöhen:

- Präzisere Prognosen, bessere strategische Entscheidungen und flexibles Reagieren durch Erkennen von Markttrends;
- aussagekräftigeres und leichter verständliches Reporting;
- wirkungsvollere Unterstützung operativer Prozesse;
- erleichterte Planung (u.a. im HR-Bereich);
- gesteigertes Kundenverständnis;
- gezieltere Produktentwicklung;
- effizientere Kostenreduktion.

(ap) ©