

Forscherguppe der RWTH Aachen evaluiert Text Mining mit dem *STATISTICA Text Miner*

Hintergrund: In einem Forschungsprojekt hat der Lehrstuhl für Betriebswirtschaftslehre (Schwerpunkt Technologie- und Innovationsmanagement) an der RWTH Aachen die Eignung von Text-Mining-basierten Verfahren für betriebliche Informationssysteme zur Technologiefrühaufklärung untersucht. Im Vordergrund stand die Frage, wie neue technologische Trends rechtzeitig erkannt und ihre weitere Entwicklung prognostiziert werden können. Für Unternehmen in Technologie-Branchen kann dieses Wissen einen entscheidenden Wettbewerbsvorteil bedeuten.

Um zu klären, welchen Beitrag statistische Text-Mining-Verfahren hier zur Entscheidungsunterstützung im Technologie- und Innovationsmanagement leisten können, entschied sich das Forscherteam der RWTH Aachen für den *STATISTICA Text Miner*.

Eine wichtige Aufgabe des im Herbst 2006 / Frühjahr 2007 durchgeführten Projekts bestand darin, anhand von Textsammlungen über aus der Vergangenheit bekannten technologischen Entwicklungen zu prüfen, ob und unter welchen Bedingungen relevante Themen und Trends gefunden werden. Die Stabilität der Ergebnisse war über Änderungen der Bedingungen (Textdatenbank, Einstellungen zum Auslesen der Textinformationen) und die Wahl verschiedener Anschlussmethoden zu testen.

Einsatz des *STATISTICA Text Miner* von StatSoft: Die genutzten Funktionen des *STATISTICA Text Miner* lagen in der Vorbereitung, dem Import und der Modellierung der zu untersuchenden Texte sowie der Anwendung von Clustering- und Association-Rule-Mining-Methoden, mit denen Muster in den Texten erkannt werden können. Insbesondere in der Textvorbereitung (dem so genannten Preprocessing) wurde die volle Breite der von der Software zur Verfügung gestellten Möglichkeiten genutzt.

Über Stopwörter-Listen wurden alle Begriffe, deren Informationsgehalt über den Inhalt der Texte als gering eingeschätzt wurde, vor dem Import der Texte herausgefiltert und Synonyme (durch Verwendung von Synonym-Listen) sowie Wörter mit gleichem Wortstamm unter einem Begriff vereinigt. Alternativ zur Verwendung von Stopwörter-Listen wurde der Einsatz einer Liste einzuschließender Wörter getestet.

„Das integrierte VBA-Modul verleiht dem *STATISTICA Text Miner* eine hohe Flexibilität und ermöglicht aufgrund vielfältiger Automatisierungsmöglichkeiten eine große Zeitersparnis.“

Jens Völler, wiss. Angestellter, Lehrstuhl für Technologie- und Innovationsmanagement, RWTH Aachen

Die in reinem Textformat vorliegenden Dokumente wurden mit dem *STATISTICA Text Miner* zunächst in Tabellen mit Worthäufigkeiten modelliert. Die enthaltenen Optionen zur Transformation der Häufigkeiten erlaubten dabei vielfältige Anschlussanalysen.

Die Durchführung der Vorbereitung, des Imports und der Modellierung der Texte sowie die Anwendung weiterer Analysen und das Reporting wurden mit Hilfe des integrierten *STATISTICA Visual Basic* automatisiert. Die Ergebnisse verschiedener Szenarien (etwa zur Variation der Ausleseinstellungen) konnten so bequem verglichen werden.

Ergebnisse: Die Forschergruppe der RWTH Aachen sieht für Text Mining in der Technologieförderung Zukunftspotenzial. Besondere Bedeutung messen die Wissenschaftler der Berücksichtigung dynamischer Aspekte bei. Enthalten Textdokumente beispielsweise Datumsinformationen (z.B. das Veröffentlichungsdatum), können Zeitintervalle vorgegeben werden, um Teilmengen von Dokumenten zu bilden. Technologische Trends dürften sich so wesentlich effektiver aufspüren lassen.

Beeindruckt zeigten sich die Wissenschaftler von den Möglichkeiten des *STATISTICA Text Miner*. Dies gilt für die Textaufbereitung, die sie als Voraussetzung für ein erfolgreiches Text Mining ansehen, wie auch für das methodische Spektrum der angebotenen Anschlussanalysen. So erlaubt es der *STATISTICA Text Miner* unter anderem, die genannten Zeitinformationen in Auswertungen einfließen zu lassen, beispielsweise beim Association Rule Mining.

Als äußerst wertvoll erwies sich für das Forscherteam von der RWTH Aachen das vom *STATISTICA Text Miner* unterstützte Visual Basic (VBA). Durch seine hohe Flexibilität können selbst spezielle Anforderungen durch den Benutzer programmiert und integriert werden. Bei Bedarf lässt sich zudem eine leistungsfähige Schnittstelle zu anderen Software-Programmen herstellen.