Text Mining auf Handelsregistereinträgen: Der *SAS Enterprise Miner* im Einsatz

Myra Spiliopoulou und Karsten Winkler[1]
Lehrstuhl für Wirtschaftsinformatik des E-Business
Handelshochschule Leipzig (HHL)

Abstract:

Text Mining bzw. Wissensentdeckung in textuellen Daten ist eine zielorientierte Form der Textanalyse, die eine effektive und effiziente Nutzung der in vielen Organisationen verfügbaren Textarchive im Wertschöpfungsprozess ermöglicht. Dieser Beitrag stellt die Grundlagen dieser Technologie vor und zeigt anhand von Fallstudien beispielhafte Einsatzmöglichkeiten des *SAS Enterprise Miner* zur Lösung Text-zentrierter Probleme.

Einführung

In der vorliegenden Fallstudie wird die Analyse eines fachspezifischen Textkorpus mit Text-Mining-Verfahren vorgestellt. Ein fachspezifischer Textkorpus ist ein großes Archiv, dessen Textdokumente in einer speziellen Fachterminologie verfasst sind. Dabei entsprechen Worthäufigkeiten, Syntax und grammatikalische Regeln nicht immer den linguistischen Gepflogenheiten der Alltagssprache.

Die Datenbasis dieser Fallstudie ist das über jeweils örtlich zuständige Amtsgerichte verteilte Archiv des deutschen Handelsregisters, dessen Inhalte von entscheidender Bedeutung für kaufmännische Transaktionen sind. Das Handelsregister enthält zudem aus makroökonomischer Sicht wertvolle Informationen über die Entwicklung deutscher Unternehmen. In diesem Beitrag werden Fragestellungen diskutiert, für deren Beantwortung eine Analyse der textuellen Handelsregistereinträge mit Text-Mining-Verfahren erforderlich ist.

Der nächste Abschnitt führt kurz in die Thematik des Text Mining ein. Nach einem Überblick über die Inhalte des Handelsregisters wird die erste Problemstellung in Zusammenhang mit diesem fachspezifischen Textkorpus diskutiert: Der Einsatz der prototypischen *DIAsDEM Workbench* und des *SAS Enterprise Miner* ermöglicht dabei effektive Anfragen auf unstrukturierten Inhalten. Anschließend werden zwei weitere Analyseszenarien unter Einsatz des neuen *SAS Enterprise Miner for Text Mining* vorgestellt. Der letzte Abschnitt schließt die Studie mit einer Zusammenfassung.

---

[1] Dieser Autor wird durch die Deutsche Forschungsgemeinschaft im Rahmen des Projekts DIAsDEM unterstützt (DFG-Zuwendung: SP 572/4-3).

Hinsichtlich des Grades an interner Struktur werden strukturierte, z.B. relationale oder objekt-relationale Daten von semistrukturierten Daten [ABS00, S. 11-13] wie beispielsweise HTML-Dateien und unstrukturierten Dokumenten (z.B. Texte) unterschieden. Die Mehrzahl der gegenwärtig am Markt verfügbaren Data-Mining-Produkte beschränkt sich jedoch auf die Unterstützung einer effizienten Wissensentdeckung in strukturierten Daten. Dieser Fakt ist mit der Erkenntnis zu kontrastieren, dass bis zu 80 Prozent der betrieblichen Informationen nicht etwa in strukturierter Form, sondern in unstrukturierten Textdokumenten abgelegt sind [Sull01, S. 56].

Unternehmen akkumulieren beispielsweise seit Jahrzehnten potenziell geschäftskritische Marktstudien, Geschäftsberichte, Kundenbefragungen oder Projektmemos meist in Datei-basierten Archiven. Neben diesen unternehmensinternen Dokumenten gibt es die scheinbar unendliche Menge potenziell entscheidungsrelevanter, kostenfrei oder gegen Entgelt abrufbarer Webseiten: Patentschriften, Branchennachrichten, Produktbewertungen oder Pressemitteilungen. Diese enthalten oft wertvolle Informationen für die Schaffung und Sicherung nachhaltiger Wettbewerbsvorteile.

Die sprichwörtliche Informationsflut erfordert aufgrund knapper Ressourcen intelligente Methoden der Datenanalyse, um weitgehend automatisiert entscheidungsrelevantes Wissen aus Textarchiven zu extrahieren. Analog zur Wissensentdeckung in Datenbanken (Data Mining) wurde dazu der Begriff Wissensentdeckung in textuellen Datenbanken (Text Mining) von Feldman und Dagan geprägt [FeDa95]. Das ultimative Ziel von Text Mining ist die Entdeckung von neuem, nicht-trivialem, interessantem und wirtschaftlich verwertbarem Wissen in großen Textbeständen.

Sullivan betrachtet Text Mining umfassender als den Prozess der Zusammenstellung, Organisation und Analyse großer Dokumentsammlungen [Sull01]. Primäre Ziele sind hier die bedarfsgerechte Distribution von Informationen an Entscheidungsträger sowie die Entdeckung versteckter Beziehungen zwischen Texten und Textfragmenten. Neben Methoden der Informatik, Statistik und des maschinellen Lernens bezieht dieser Ansatz zwei etablierte Forschungsbereiche explizit in den Methodenkanon des Text Mining ein: Während die Suche in natürlichsprachigen Dokumentarchiven Gegenstand des Information Retrieval ist [BaRi99], untersucht der Bereich Informationsextraktion die Entdeckung benannter Entitäten (z.B. Namen von Personen) in Texten.

Abbildung 1 illustriert in Anlehnung an [FPS96] den generischen Prozess der Wissensentdeckung in textuellen Datenbanken. Der Ausgangspunkt eines Text-Mining-Projekts ist die Zielstellung, die i.d.R. ein abgrenzbares betriebswirtschaftliches Problem sein sollte. Insbesondere tendenziell textorientierte betriebliche Aufgaben wie z.B.

Marktforschung, Kundenbeziehungsmanagement, Wettbewerberanalyse oder Wissensmanagement bieten vielfältige Einsatzmöglichkeiten für Text Mining.

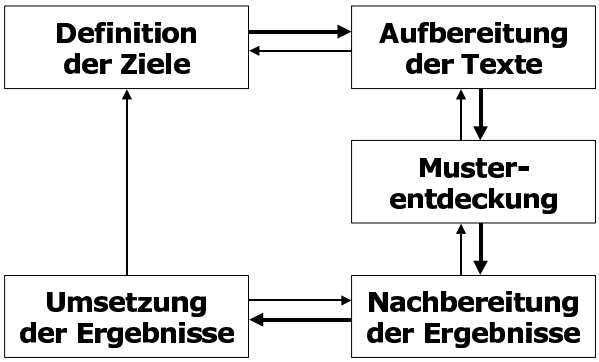

Abb. 1: Prozess der Wissensentdeckung in textuellen Datenbanken

Nach Auswahl inhaltlich relevanter Textarchive sind die Dokumente zu bereinigen und in eine für die Musterentdeckung geeignete Datenstruktur zu überführen. Natürliche Sprache ist naturgemäß hochdimensional, syntaktisch komplex sowie semantisch oft mehrdeutig. Deshalb umfasst die aufwendige Phase der Aufbereitung meist folgende Schritte: Sprachidentifikation, Extraktion der Texte aus multimedialen Dokumenten, Zerlegung der Texte in einzelne Wörter, Wortreduktion auf grammatische Grundformen sowie Entfernung sinnleerer, besonders seltener und sehr häufiger Wörter.
Für die Repräsentation textueller Daten wird häufig das Vektorraummodel des Information Retrieval verwendet [BaRi99, S. 27-30]. Dabei wird jedes Dokument in einen Vektor überführt, dessen Dimensionen den insgesamt in einem Textarchiv vorkommenden Wörtern entsprechen. Jede Dimension eines Dokumentvektors repräsentiert die Häufigkeit des entsprechendes Worts innerhalb des korrespondierenden Textes. Somit wird jedes Dokument als ein Vektor in einem hochdimensionalen Raum modelliert. Zwei Dokumente sind einander ähnlich, wenn ihre Vektoren topologisch nahe zueinander sind. Diese Modellierung ist aber suboptimal, weil die meisten Dokumente lediglich eine kleine Auswahl der Wörter des gesamten Wortschatzes enthalten und somit nur wenige vergleichbare Texte zu finden sind. Deshalb wird versucht, die Anzahl der Dimensionen zu reduzieren und sämtliche Dimensionen entsprechend ihrer Relevanz zu gewichten. Diese Gewichtung erfolgt meistens auf Basis der Worthäufigkeit, wobei das Gewicht für ein Wort des Dokumentvektors mit zunehmender Häufigkeit des gleichen Worts im gesamten Archiv abnimmt.
Die in der Phase der Musterentdeckung einzusetzenden Verfahren werden von den Anforderungen der Problemstellung diktiert. Klassifikationsverfahren werden z.B. benutzt, um Texte entsprechend ihres Inhalts in vorgegebenen Klassen einzuordnen. Diese Algorithmen sind oft nicht speziell für Textdaten optimiert. Viele Segmentie-

rungs- und Klassifikationsalgorithmen können beispielsweise nicht so viele Dimensionen effizient verarbeiten, wie sie typischerweise im Vektorraum von Dokumenten auftreten. In diesem Beitrag konzentrieren wir auf Segmentierungs- und Klassifikationsverfahren sowie auf Verfahren der Assoziationsregelentdeckung.
Nach der Musterentdeckung werden die Ergebnisse statistisch ausgewertet und aus betriebswirtschaftlicher Sicht interpretiert. In dieser Phase werden meist auch Visualisierungsverfahren eingesetzt, um den Experten zu unterstützen. Eine anspruchsvolle Aufgabe ist dabei die Ableitung von Handlungsempfehlungen, um das betriebliche Projektziel zu erreichen. Der gesamte Prozess der Wissensentdeckung ist iterativ, d.h. einzelne Phasen sollten bei Bedarf wiederholt werden. Zusätzlich ist die Umsetzung der Ergebnisse zu überwachen, um gegebenenfalls ein neues Projekt zu initiieren.

Die Dokumente der Fallstudie: Handelsregistereinträge

Amtsgerichte führen in Deutschland ein öffentlich zugängliches Handelsregister, das wirtschaftlich relevante Informationen über die Unternehmen des jeweiligen Einzugsbereichs enthält. Entsprechend den Vorschriften des deutschen Handelsrechts müssen (mit wenigen Ausnahmen) alle Unternehmen gesetzlich geregelte Vorgänge wie z.B. Gründung, Erteilung von Prokura, Eröffnung einer Zweigniederlassung oder Liquidation einer Gesellschaft zur Eintragung in das zuständige Handelsregister anmelden. Die Kenntnis der Handelsregistereintragungen ist für gegenwärtige und potenzielle Geschäftspartner von besonderem Interesse, da diese Eintragungen gemäß § 15 HGB (Publizität des Handelsregisters) weitreichende juristische Konsequenzen haben.

HRB 12990 30.09.1999	**Behrens & Klein Oberbausysteme GmbH (Seeblickstraße 26, 15758 Zernsdorf)**	**publiziert am 09.10.1999**
Vertrieb und Entwicklung von Gleisoberbautechnik. Stammkapital: 50.000 DM. Gesellschaft mit beschränkter Haftung. Der Gesellschaftsvertrag ist am 2. Dezember 1994 abgeschlossen. Durch Beschluss der Gesellschafterversammlung vom 7. April 1999 ist der Sitz der Gesellschaft von Berlin nach Zernsdorf verlegt und der Gesellschaftsvertrag geändert in § 1 (Sitz). Ist nur ein Geschäftsführer bestellt, so vertritt er die Gesellschaft einzeln. Sind mehrere Geschäftsführer bestellt, so wird die Gesellschaft durch zwei Geschäftsführer oder durch einen Geschäftsführer in Gemeinschaft mit einem Prokuristen vertreten. Einzelvertretungsbefugnis kann erteilt werden. Hendrik Klein, 16.02.1967, Zernsdorf, und Klaus Behrens, 04.01.1958, Braunschweig, sind zu Geschäftsführern bestellt. Sie vertreten die Gesellschaft stets einzeln und sind befugt, Rechtsgeschäfte mit sich im eigenen Name oder als Vertreter eines Dritten abzuschließen. Nicht eingetragen: Die Bekanntmachungen der Gesellschaft erfolgen im Bundesanzeiger.		

Abb. 2: Handelsregistereintrag des Amtsgerichts Potsdam

Abbildung 2 zeigt exemplarisch einen Handelsregistereintrag, der die Gründung einer GmbH anzeigt. Der unstrukturierte Textabschnitt enthält die wesentlichen, gesetzlich geforderten Informationen in Form des von Justizangestellten erfassten Eintragungstextes. Aufgrund des regen Interesses der Wirtschaft an den Einträgen im Handelsregister gibt es bereits Dienstleister, die online oder offline Handelsregisterauskünfte

anbieten. Gegenwärtig unterstützen diese Informationsanbieter jedoch nur SQL-Anfragen auf den strukturierten Handelsregisterdaten (z.B. Firma, Adresse oder Publikationsdatum) und konventionelle Volltextrecherchen in den Eintragungstexten.

Strukturierung von Handelsregistereinträgen: Die *DIAsDEM Workbench*

Der wichtigste Bestandteil eines Handelregistereintrags ist der Eintragungstext. Eine Strukturierung dieser Texte würde SQL-Anfragen auf diesen textuellen Inhalten und somit eine zielgerichtete Suche ermöglichen. In diesem Kontext verfolgt das von der Deutschen Forschungsgemeinschaft geförderte Projekt DIAsDEM („Datenintegration von Altlastdaten und semistrukturierten Dokumenten mit Mining-Verfahren") folgende Ziele: Für ein Textarchiv ist erstens eine geeignete DTD abzuleiten und zweitens die semantische Auszeichnung sämtlicher Texte des Archivs durchzuführen.
Zur Erreichung der Ziele wurde der in Abbildung 3 zusammengefasste Prozess der Entdeckung von Wissen in textuellen Datenbanken vorgeschlagen [GSW01, WiSp01]. Dabei werden Cluster inhaltlich ähnlicher Textelemente (z.B. Sätze) entdeckt, halbautomatisch benannt und danach die Textelemente mit den Cluster-Namen auszeichnet. Die prototypische *DIAsDEM Workbench* ermöglicht die Annotation fachspezifischer Textarchive mit der Textauszeichnungssprache XML. Abbildung 4 zeigt den im Rahmen dieser Fallstudie erzeugten, semantisch ausgezeichneten Eintragungstext der Abbildung 2. Dieses XML-Dokument ist ein Ergebnis der Anwendung des Text-Mining-Prozesses auf ein Archiv von 1.145 im Internet veröffentlichten Neueintragungen des Amtsgerichts Potsdam im Jahr 1999.

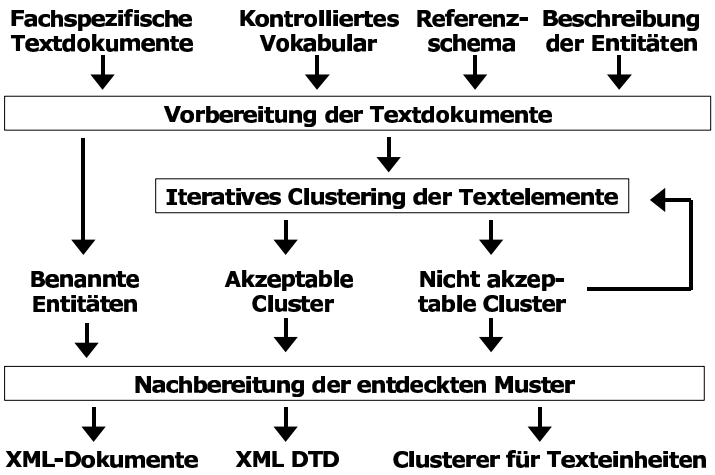

Abb. 3: Der DIAsDEM-Prozess der Wissensentdeckung

Für die kontrollierte Dimensionsreduzierung der Textdaten wird ein konzeptionelles Referenzschema und ein Thesaurus der Anwendungsdomäne eingesetzt: Sämtliche Textelemente des Archivs werden auf Vektoren abgebildet, deren Dimensionen max.

200 ausgewählten Konzepten bzw. Begriffen aus Schema und Thesaurus entsprechen. Diese Vektoren sind anschließend Eingabedaten für einen klassischen, d.h. nicht Text-optimierten Clustering-Algorithmus, der aber iterativ ausgeführt wird.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Handelsregistereintrag SYSTEM 'Handelsregistereintrag.dtd'>

<Handelsregistereintrag>
<Gegenstand> Vertrieb und Entwicklung von Gleisoberbautechnik. </Gegenstand>
<Stammkapital Geldbetrag="50000 DM"> Stammkapital: 50.000 DM.
</Stammkapital> <GmbH> Gesellschaft mit beschränkter Haftung. </GmbH>
<AbschlussGesellschaftsvertrag Datum= "02.12.1994"> Der Gesellschaftsvertrag
ist am 2. Dezember 1994 abgeschlossen. </AbschlussGesellschaftsvertrag>
<AenderungGesellschaftsvertrag Datum="07.04.1999" Paragraph="§ 1 (Sitz)">
Durch Beschluss der Gesellschafterversammlung vom 7. April 1999 ist der Sitz
der Gesellschaft von Berlin nach Zernsdorf verlegt und der Gesellschaftsvertrag
geändert in § 1 (Sitz). </AenderungGesellschaftsvertrag> (...) Einzelvertretungs-
befugnis kann erteilt werden. <BestellungGeschaeftsfuehrer Person= "Klein;
Hendrik; Zernsdorf; 16.02.1967 && Behrens; Klaus; Braunschweig; 04.01.1958">
Hendrik Klein, 16.02.1967, Zernsdorf, und Klaus Behrens, 04.01.1958, Braun-
schweig, sind zu Geschäftsführern bestellt. </BestellungGeschaeftsfuehrer>
(...) <Bekanntmachungen> Nicht eingetragen: Die Bekanntmachungen der Gesell-
schaft erfolgen im Bundesanzeiger. </Bekanntmachungen>
</Handelsregistereintrag>
```

Abb. 4: Semantisch ausgezeichneter Handelsregistereintrag

Nach jeder Iteration werden qualitativ akzeptable Cluster anhand von Qualitätskriterien ermittelt, benannt und dem Wissensingenieur zur endgültigen semantischen Namensgebung präsentiert. Vektoren in qualitativ nicht akzeptablen Clustern sind wiederum Eingabedaten für die nächste Iteration. In dieser werden zusätzlich restriktive Parameter des jeweiligen Algorithmus gelockert, um weitere Cluster zu entdecken. Nach der letzten Iteration werden Textelemente, die auf Vektoren in akzeptablen Clustern abgebildet wurden, mit den Cluster-Namen entsprechenden XML-Textmarken ausgezeichnet. Zusätzlich wird eine das Archiv beschreibende DTD abgeleitet.

In dieser Fallstudie wurde die Java-basierte *DIAsDEM Workbench* für die Textaufbereitung, die Steuerung des iterativen Clustering, die halbautomatische Benennung der Cluster sowie für die Ableitung einer DTD und die Erzeugung von XML-Dokumenten eingesetzt. Für das dreimalige iterative Clustering der insgesamt 10.785 Textelementvektoren wurde eine Clustering-Funktion des *SAS Enterprise Miner* verwendet.

Die dabei eingesetzten selbstorganisierenden Karten (engl.: Kohonen maps) sind spezielle neuronale Netze, die n-dimensionale Vektoren entsprechend einer Ähnlichkeitsfunktion auf ein zweidimensionales Gitter abbilden [EsSa00, S. 271-272]. Abbildung 5 zeigt links das im *SAS Enterprise Miner* ausgeführte Data-Mining-Diagramm. Die *DIAsDEM Workbench* generierte sämtliche Eingabedaten für die drei Iterationen und verarbeitete die jeweils mit einer SAS-Prozedur exportierten Clustering-Ergebnisse. Rechts ist das in der ersten Iteration erzeugte 10x10-Gitter sowie die Visualisierung eines Clusters abgebildet, in dessen 882 Vektoren die Konzepte „bestellen" und „Ge-

schäftsführer" dominieren. Deshalb wurden diese Textelemente anschließend mit der XML-Textmarke „BestellungGeschäftsführer" ausgezeichnet.

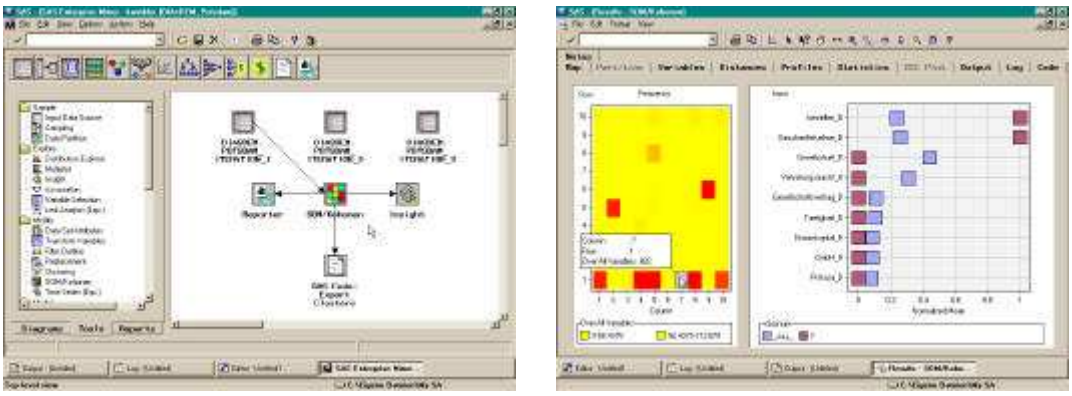

Abb. 5: Clustering-Diagramm und Visualisierung des Ergebnisses

Insgesamt entdeckte die *DIAsDEM Workbench* 68 qualitativ akzeptable Cluster, die etwa 95 Prozent der Textelemente enthielten. Die Qualität der Textauszeichnung wurde anschließend mangels korrekt annotierter Trainingsdaten durch manuelle Sichtung einer 5-Prozent-Zufallsstichprobe der Textelemente des Archivs geprüft: Der Gesamtfehler der Textauszeichnung lag mit 0,95-Konfidenz im Intervall [2,4%; 4,2%]. Für eine erste Fallstudie ist das ein vielversprechendes Ergebnis.

Der *SAS Enterprise Miner for Text Mining*

Entsprechend einer SAS-Ankündigung wird ab Mitte 2002 eine leistungsstarke Text-Mining-Komponente für den *SAS Enterprise Miner* produktiv verfügbar sein [Bole01, S. 31]. Im Rahmen der Kooperation zwischen dem SAS Academic Club und der Handelshochschule Leipzig wurde aber bereits eine experimentelle Vorabversion des neuen *SAS Enterprise Miner for Text Mining* innerhalb des Projekts DIAsDEM eingesetzt.
Die neue Text-Mining-Komponente ist fest in die grafische Oberfläche sowie in die bereits vorhandene Funktionalität des *SAS Enterprise Miner* integriert und fügt sich zudem nahtlos in die hauseigene Data-Mining-Methodologie SEMMA ein. Dieses englische Akronym steht für „sample, explore, modify, model and assess" und fasst die von SAS entwickelte, prozessorientierte Data-Mining-Methodik zusammen. Der integrative Text-Mining-Ansatz wird auch durch die folgenden, von SAS empfohlenen Phasen zur Wissensentdeckung in textuellen Datenbanken verdeutlicht [Bole01, S. 32]:

1. Aufbereitung der Textdaten, d.h. Überführung der unstrukturierten Textdokumente in ein strukturiertes und für die Musterentdeckung geeignetes Format,
2. Dimensionsreduzierung, d.h. Reduktion der zuvor erzeugten und i.d.R. sehr hochdimensionalen Wort-je-Dokument-Matrix auf eine praktikable Größe, sowie

3. Musterentdeckung, d.h. Anwendung traditioneller Data-Mining-Verfahren auf den zuvor strukturierten und dimensionsreduzierten Textdaten.

In Anlehnung an den generischen Text-Mining-Prozess (vgl. Abbildung 1) sollten die Projektziele selbstverständlich vor der Textaufbereitung definiert werden. Im Anschluss an die Musterentdeckung müssen die Ergebnisse dann interpretiert und umgesetzt werden. Dieser Text-Mining-Ansatz könnte zusammenfassend mit „Text Mining ist Data Mining!" charakterisiert werden, da lediglich in der Aufbereitungsphase spezielle, Text-optimierte Verfahren erforderlich sind. Der *SAS Enterprise Miner for Text Mining* umfasst in der Vorabversion drei für die Textanalyse relevante Funktionen, von denen zwei im Folgenden kurz vorgestellt werden.

Die für Text-Mining-Aktivitäten elementare Textaufbereitungsfunktion (engl.: text parsing) zerlegt Textdokumente zuerst in individuelle, sämtlich kleingeschriebene Wörter sowie entfernt Interpunktionszeichen und sinnleere Wörter anhand einer editierbaren sog. Stopwort-Liste. Jedes Textarchiv wird danach in eine Wort-je-Dokument-Matrix überführt. Diese Matrix enthält die Häufigkeiten sämtlicher Terme je Dokument und ist die strukturierte Basis für weitere Analysen. Die Textaufbereitungsfunktion kann Texte verarbeiten, die entweder direkt als Attributwerte in einer Relation gespeichert sind oder indirekt mit Dateinamen referenziert werden. Die endgültige Version wird außerdem die Wortstammermittlung, die Identifikation mehrere Wörter umfassender Konzepte (z.B. Handelshochschule Leipzig) und die Entdeckung benannter Entitäten wie z.B. Personen, Unternehmen oder Adressen unterstützen [Bole01, S. 31].

Innerhalb des *SAS Enterprise Miner for Text Mining* wird ein mathematisches Verfahren der latent-semantischen Analyse (engl.: latent semantic analysis) eingesetzt, um die Dimension der Wort-je-Dokument-Matrix zu reduzieren [Sull01, S. 337-341]. Die verwendete Singulärwertzerlegung (engl.: singular value decomposition, SVD) vermindert die Anzahl der Dimensionen je Dokument von oft mehreren tausend Wörtern auf ein- bis zweihundert numerische Attribute [Bole01, S. 32]. Für diese unkontrollierte Dimensionsreduzierung wird eine Methode der linearen Algebra eingesetzt, um die Relevanz bedeutungsloser Wörter zu mindern und zugleich die relevanten Terme zu identifizieren. Die Ergebnismatrix ist eine semantisch komprimierte und auf statistisch bedeutsame Inhalte beschränkte Repräsentation des Textarchivs, die nachfolgend für Segmentierungs- und Klassifikationsaufgaben verwendet werden kann.

Assoziationsregelentdeckung in Handelsregistereinträgen

Es sind drei Typen von Handelsregistereinträgen zu unterscheiden: Neueintragungen gegründeter Unternehmen, Veränderungsmeldungen und Löschungseintragungen ein-

gestellter Unternehmen. Im Rahmen des Projekts DIAsDEM wurde die Text-Mining-Komponente des *SAS Enterprise Miner* zunächst eingesetzt, um Assoziationsregeln [EsSa00, S. 159-160] in den 2.294 Veränderungseintragungen des Amtgerichts Potsdam im Jahr 1999 zu finden. Ziel der Analyse war die Entdeckung häufiger Wortkombinationen in Veränderungseinträgen, um diese ggf. als Kandidaten für Cluster-Bezeichner bzw. XML-Textmarken berücksichtigen zu können.

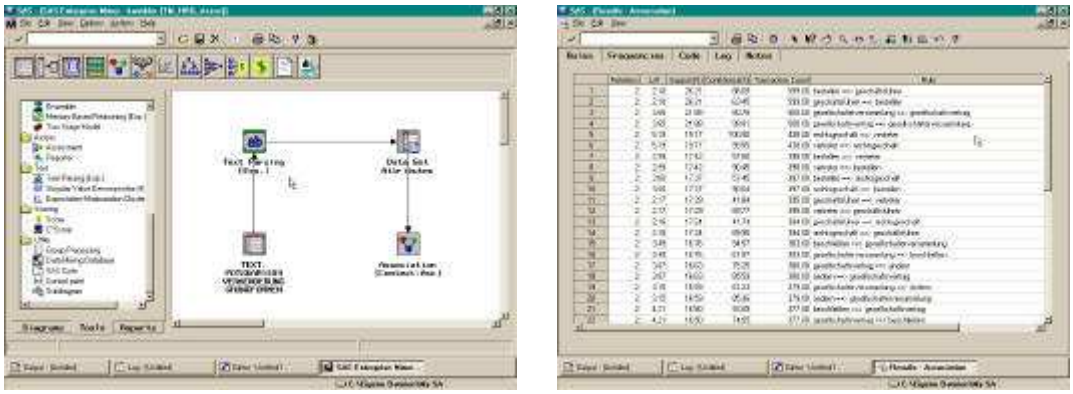

Abb. 6: Assoziationsdiagramm und Visualisierung entdeckter Regeln

Abbildung 6 zeigt links das innerhalb der Fallstudie eingesetzte Data-Mining-Diagramm. Die für eine erfolgreiche Assoziationsregelentdeckung notwendige Wortstammermittlung wurde zuvor durch die *DIAsDEM Workbench* und den als Plug-In eingebetteten *TreeTagger* [Schm94] ausgeführt. Zusätzlich wurde der Textaufbereitungsfunktion eine Liste von 541 deutschen Stopwörtern (z.B. „und", „die", „ein") übergeben. Die Wort-je-Dokument-Matrix war Basis für die Entdeckung von Assoziationsregeln. Für diese Form der Textanalyse darf keine unkontrollierte Dimensionsreduktion vorgenommen werden, um Assoziationen zwischen tatsächlich verwendeten Wörtern zu finden. Nach der Textaufbereitung wurde die traditionelle, d.h. nicht Text-optimierte Funktion zur Assoziationsregelentdeckung aufgerufen. Im Vokabular der Assoziationsanalyse entspricht hierbei jedes Dokument einer Transaktion, die sich aus den jeweilig benutzten Wörtern als Items zusammensetzt. Auf der rechten Seite der Abbildung 6 ist die Visualisierung entdeckter Assoziationsregeln dargestellt. Die Regel „Gesellschaftsvertrag -> ändern" wird z.B. von etwa 17 Prozent sämtlicher Veränderungseintragungen unterstützt und weist eine Konfidenz von ca. 75 Prozent auf.

Klassifikation von Handelsregistereinträgen

Nach der Assoziationsanalyse wurden insgesamt 3.608 Handelsregistereinträge des Potsdamer Amtsgerichts aus dem Jahr 1999 eingesetzt, um einen Klassifikator zur Vorhersage der Ausprägung der Zielvariablen „Eintragungstyp" zu trainieren. Ziel dieser Fallstudie war eine zuverlässige Identifizierung des Eintragungstyps deutscher

Handelsregistereintragungen. Die Zuordnung eines Dokuments in dessen Eintragungskategorie ist Voraussetzung für eine qualitativ hochwertige semantische Auszeichnung mit der *DIAsDEM Workbench*. Das Potsdamer Archiv konnte für diese Aufgabe als Trainings- und Testdaten genutzt werden, da dieses Amtsgericht die Veröffentlichungen im Gegensatz zu anderen Gerichten bereits kategorisiert publiziert. Deshalb ist die Ausprägung der Zielvariablen für alle Dokumente fehlerfrei gegeben.

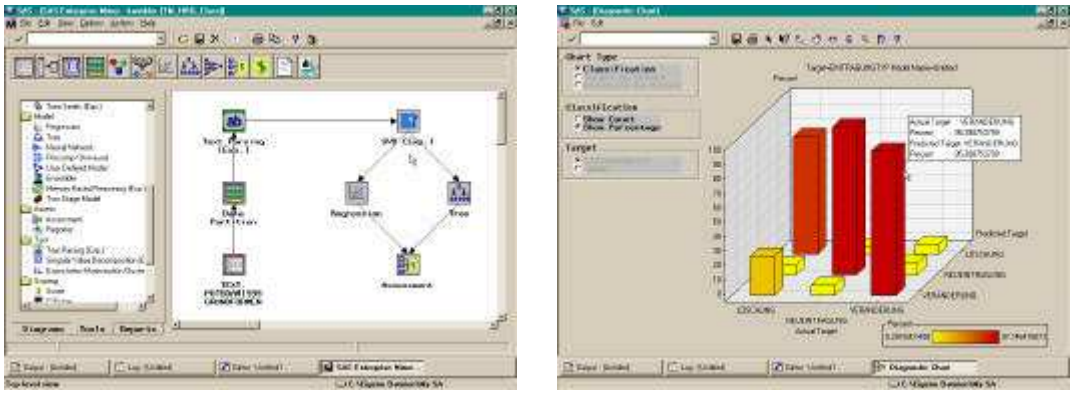

Abb. 7: Klassifikationsdiagramm und Visualisierung des Klassifikationsqualität

Abbildung 7 illustriert links das Data-Mining-Diagramm im *SAS Enterprise Miner*. Vor der Textaufbereitung wird das vorklassifizierte Archiv zunächst in einen Trainings- und einen Testdatensatz zerlegt. Bei dieser Klassifikationsaufgabe wurde eine Dimensionsreduktion mittels Singulärwertzerlegung auf 100 numerische Dimensionen je Textdokument durchgeführt, um anschließend traditionelle Klassifikationsalgorithmen einsetzen zu können. Auf der rechten Seite der Abbildung 7 ist die Klassifikationsmatrix des verwendeten Regressionsverfahrens abgebildet, das auf dem Testdatensatz eine Gesamtfehlerrate von etwa 5 Prozent aufweist und die Handelsregistereinträge somit besser als das Entscheidungsbaumverfahren (Fehlerrate: ca. 9 Prozent) klassifiziert.

Zusammenfassung

Seit Jahrhunderten benutzen die Menschen ihre natürliche Sprache, um entscheidungsrelevante Informationen mündlich weiterzugeben oder schriftlich niederzulegen. Die beinahe flächendeckende Einführung der Informationstechnologie führte jedoch in den vergangenen Jahrzehnten zu einer beispiellosen Akkumulation elektronisch verfügbarer Textdokumente. Text Mining ist eine noch recht junge, prozessorientierte Methodologie für die Entdeckung neuen, interessanten, nicht-trivialen und wirtschaftlich verwertbaren Wissens in diesen großen Textarchiven. Technologische Herausforderungen für Text-Mining-Verfahren sind dabei insbesondere die Hochdimensionalität und die semantische Mehrdeutigkeit natürlicher Sprache.

In diesem Beitrag wurden Text-Mining-Verfahren vorgestellt, die Texte nach jeweils spezieller Aufbereitung und anschließender Dimensionsreduzierung mit konventionellen Data-Mining-Algorithmen analysierten. Dabei wurde einerseits der nicht Text-optimierte *SAS Enterprise Miner* erfolgreich als Plug-In in einem Forschungsprototyp zur Segmentierung von Textelementen eingesetzt. Die andererseits verwendete Vorabversion des *SAS Enterprise Miner for Text Mining* integriert Funktionen zur Textaufbereitung und Dimensionsreduzierung in die Umgebung des *SAS Enterprise Miner*. Durch die gelungene Integration dieser neuen Komponente wird Anwendern ein überaus breites Spektrum an Möglichkeiten in der weiten Welt des Text Mining eröffnet.

Literaturverzeichnis

[ABS00] Abiteboul, S.; Buneman, P.; Suciu, D.: Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufman Publishers, San Francisco 2000.

[BaRi99] Baeza-Yates, R.; Ribeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York 1999.

[Bole01] Bolen, A.: Data Mining for Text. In: sascom (2001) 6, S. 30-32.

[EsSa00] Ester, M.; Sander, J.: Knowledge Discovery in Databases: Techniken und Anwendungen. Springer-Verlag, Berlin Heidelberg 2000.

[FPS96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.: The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: Communications of the ACM 39 (1996) 11, S. 27-34.

[FeDa95] Feldman, R.; Dagan, I.: Knowledge Discovery in Textual Databases (KDT). In: Proceedings of the First International Conference on Knowledge Discovery and Data Mining, Montreal, Canada, August 1995, August 1995, S. 112-117.

[GSW01] Graubitz, H.; Spiliopoulou, M.; Winkler, K.: The DIAsDEM Framework for Converting Domain-Specific Texts into XML Documents with Data Mining Techniques. In: Proceedings of the First IEEE International Conference on Data Mining, San Jose, CA, USA, November/December 2001, S. 171-178.

[Schm94] Schmid, H.: Probabilistic Part-Of-Speech Tagging Using Decision Trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK, September 1994, S. 44-49.

[Sull01] Sullivan, D.: Data Document Warehousing and Text Mining. John Wiley & Sons, New York 2001.

[WiSp01] Winkler, K.; Spiliopoulou, M.: Semi-Automated XML Tagging of Public Text Archives: A Case Study. In: Proceedings of EuroWeb 2001 "The Web in Public Administration". Pisa, Italy, December 2001, S. 271–285.

Kontaktinformation und Kurzbiographie

Myra Spiliopoulou und Karsten Winkler
Lehrstuhl für Wirtschaftsinformatik des E-Business
Handelshochschule Leipzig (HHL)

Anschrift: Jahnallee 59, D-04109 Leipzig
Telefon: (03 41) 98 51 760
Telefax: (03 41) 98 51 764
E-Mail: {myra|kwinkler}@ebusiness.hhl.de
Homepage: http://ebusiness.hhl.de

Prof. Dr. Myra Spiliopoulou studierte Mathematik, promovierte in Informatik und ist seit April 2001 Inhaberin des Lehrstuhls für Wirtschaftsinformatik des E-Business an der HHL. Ihr wissenschaftlicher Mitarbeiter Dipl.-Kfm. Karsten Winkler ist Datenverarbeitungskaufmann und studierte BWL an der Humboldt-Universität zu Berlin.