Text Mining in der Wettbewerberanalyse: Konvertierung von Textarchiven in XML-Dokumente

Karsten Winkler[*], **Myra Spiliopoulou**
Handelshochschule Leipzig (HHL)
Lehrstuhl für Wirtschaftsinformatik des E-Business
Jahnallee 59
04109 Leipzig
kwinkler@ebusiness.hhl.de
myra@ebusiness.hhl.de

Zusammenfassung

Methoden der Wettbewerberanalyse dienen der Erzielung und Verteidigung nachhaltiger Wettbewerbsvorteile. Im Gegensatz zu strukturierten, tendenziell unkompliziert auswertbaren Daten liegt eine Vielzahl an Informationen über Wettbewerber in textueller Form vor. Eine effiziente Wettbewerberanalyse erfordert daher weitgehend automatische Verfahren der Wissensentdeckung in textuellen Datenbanken. In diesem Beitrag wird das DIAsDEM-Vorgehensmodell zur semiautomatischen semantischen Annotation fachspezifischer Textarchive vorgestellt. Es beinhaltet einen komplexen Text Mining-Prozess, um die diesen Archiven häufig inhärente, jedoch undokumentierte semantische Struktur zu entdecken, eine XML DTD abzuleiten und die Texte in semantisch ausgezeichnete XML-Dokumente zu überführen. Dabei wird ein Clustering-Algorithmus des *SAS Enterprise Miner* im Rahmen eines Plug-In-Konzepts verwendet. Die Qualität der Auszeichnung wird im Rahmen einer Fallstudie evaluiert.

Keywords: Text Mining, Textstrukturierung, Clustering, XML.

[*]Dieser Autor wird durch die Deutsche Forschungsgesellschaft im Rahmen des Projekts DIAsDEM - Datenintegration von Altlastdaten und semistrukturierten Dokumenten mit Mining-Verfahren unterstützt (DFG-Zuwendung SP 572/4-3).

1 Einleitung

Erfolgreiche Unternehmen setzen Methoden der Wettbewerberanalyse (engl.: competitive intelligence) ein, um nachhaltige Wettbewerbsvorteile in volatilen Märkten zu erzielen und zu verteidigen. Kahaner definiert Wettbewerberanalyse als systematischen Ansatz zur Sammlung und Analyse von Informationen über Wettbewerber und wirtschaftliche Trends, die potentiell für die Erreichung unternehmenseigener Ziele förderlich sind ([10], S. 16). Die Antizipation relevanter Marktveränderungen, die Vorwegnahme von Mitbewerberaktivitäten oder auch die Entdeckung neuer potentieller Wettbewerber sind typische Aufgaben der Wettbewerberanalyse. Im dynamischen Umfeld von Wettbewerbern, Behörden, Finanzmärkten und öffentlicher Meinung dient das Wettbewerberinformationssystem eines Unternehmens auch der Sammlung von Wissen über potentiell relevante Technologien, Produkte, Gesetze und Verordnungen ([17], S. 3-7).

Hinsichtlich des Grades an interner Struktur werden strukturierte, z.B. relationale oder objekt-relationale Daten von semistrukturierten Daten [1] wie z.B. HTML-Seiten und unstrukturierten Dokumenten (z.B. Texte) unterschieden. Da bis zu 80 Prozent der betrieblichen Informationen in unstrukturierten Textdokumenten abgelegt sind [18], werden seit Mitte der neunziger Jahre auch Data Mining-Methoden eingesetzt, um Texte zu kategorisieren und inhaltlich ähnliche Dokumente zu erkennen. Feldman und Dagan prägten für diese Forschungsrichtung den Begriff „Wissensentdeckung in textuellen Datenbanken“ [7].

Im Gegensatz zu strukturierten, tendenziell unkompliziert auswertbaren Daten (z.B. Entwicklung des Aktienkurses) liegt eine Vielzahl an Informationen über Aktivitäten der Wettbewerber und Markttrends in textueller, d.h. unstrukturierter Form vor. Eine effiziente Wettbewerberanalyse erfordert daher, im Gegensatz zur konventionellen Volltextsuche, weitgehend automatische Text Mining-Verfahren. Die Entdeckung ökonomisch umsetzbaren Wissens in Textarchiven wird durch die Integration relevanter, meist heterogener Datenquellen im einem sog. Document Warehouse erleichtert ([17], S. 81-102). Die semantische, d.h. inhaltsbezogene Auszeichnung mittels XML ist ein Verfahren zur Erschließung von Textarchiven. Nach erfolgter Auszeichnung eines Textarchivs mit z.B. der Textauszeichnungssprache XML können die semantischen Metadaten in Form der Dokumenttypdefinition (DTD) und der eingefügten XML-Textmarken bspw. für eine inhalts- und strukturbasierte Suche im Rahmen von Wettbewerberinformationssystemen sowie für Zwecke weiterführender Wissensentdeckung und der Informationsintegration mit weiteren, inhaltlich relevanten Datenquellen in einem Document Warehouse genutzt werden.

Ziel des im Folgenden vorgestellten DIAsDEM-Vorgehensmodells ist die Teilstrukturierung großer, anwendungsspezifischer Archive homogener Textdoku-

mente durch eine qualitativ hochwertige semantische Auszeichnung. Der Begriff Teilstrukturierung steht hier für die Ableitung einer semantischen, die innere Struktur des Textarchivs widerspiegelnden DTD und die anschließende Auszeichnung struktureller Textelemente mit gültigen XML-Textmarken und Attributen. Hierzu müssen innerhalb des Archivs semantisch ähnliche Textelemente entdeckt, benannt und zu einer XML-Dokumenttypdefinition aggregiert werden. Um den menschlichen Arbeitsaufwand zu minimieren, wurde dazu ein komplexes Verfahren der Wissensentdeckung vorgeschlagen.

Der Artikel ist wie folgt strukturiert: Im folgenden Abschnitt wird die relevante Literatur kurz diskutiert. In Abschnitt 3 wird das DIAsDEM-Vorgehensmodell zur semantischen Auszeichnung anwendungsspezifischer Textarchive zusammenfassend dargestellt. Der vorgeschlagene Prozess der Wissensentdeckung zur Ableitung einer vorläufigen XML DTD wird anschließend in Abschnitt 4 detailliert erläutert. Die Anwendung des Vorgehensmodells in einer Fallstudie mit einem Archiv des deutschen Handelsregisters wird in Abschnitt 5 präsentiert. Der letzte Abschnitt erörtert nach einer Zusammenfassung einzelne Aspekte der künftigen Forschung im Rahmen des Projekts.

2 Relevante Literatur im Überblick

Die Forschung zur Wissensentdeckung in textuellen Datenbanken ist zugleich Basis und Inspiration für den vorliegenden Beitrag. Aktuelle Arbeiten auf dem Gebiet der semistrukturierten Daten zielen meist auf die Ableitung eines Schemas für gegebene Dokumente ab. Im Gegensatz dazu sind aber innerhalb des DIAsDEM-Vorgehensmodells simultan zwei Probleme zu lösen: Teilstrukturierung der Textdokumente und Ableitung einer XML-Dokumenttypdefinition. Eine ausführliche Diskussion relevanter Literatur in diesen zwei Forschungsgebieten findet sich in [8]. An dieser Stelle wird die Literaturdiskussion auf Projekte beschränkt, die ähnliche Ziele verfolgen:

Bruder et al. stellen mit GETESS einen Anfrage- und Suchdienst für das Web vor, der Ontologie-basiert HTML-Dokumente in XML-Zusammenfassungen überführt [3]. Diese Zusammenfassungen sind fachspezifisch annotierte XML-Dokumente und dienen als Datenquelle für Suchdienste. In Kontrast zu DIAsDEM ist hier einerseits die DTD durch eine Ontologie a priori festgelegt. Andererseits zielt DIAsDEM auf die semantische Annotation der ursprünglichen Texte, um die Inhalte für weitere Analysen und Visualisierungen zu erhalten.

Moore und Berman präsentieren ein Verfahren zur Überführung textueller pathologischer Berichte, d.h. Untersuchungen von Gewebeproben, in XML-Dokumente [14]. Die Autoren leiten im Gegensatz zu DIAsDEM jedoch weder

eine DTD ab, noch wenden sie Verfahren der Wissensentdeckung in Datenbanken an. Es werden lediglich Methoden der Sprachverarbeitung und ein Thesaurus genutzt, um Wörter bzw. Wortgruppen auf medizinische Konzepte abzubilden und mit diesen semantisch auszuzeichnen.

Inhaltlich näher zu DIAsDEM ist hingegen die Arbeit von Lumera, der Schlüsselwörter und Regeln verwendet, um textuelle Altlastdaten wie z.B. Fertigungshandbücher halbautomatisch in XML-Dokumente zu überführen [12]. Dieser Ansatz basiert jedoch auf einer manuell erstellten Regelbasis, während DIAsDEM einen Prozess der Wissensentdeckung zur Minimierung des menschlichen Aufwands in das Vorgehensmodell integriert.

Decker et al. verwenden das Ontologie-basierte System ONTOBROKER, um Metadaten aus Webseiten zu extrahieren [4]. Embley et al. nutzen ebenfalls Ontologien, um Informationen aus fachspezifischen, unstrukturierten Texten sowohl zu extrahieren als auch zu strukturieren [5]. Maedche und Staab stellen eine Architektur vor, die das halbautomatische Lernen von Ontologien aus HTML-Dokumenten unterstützt [13]. Innerhalb des DIAsDEM-Vorgehensmodells werden Metadaten jedoch nicht von den ursprünglichen Texten getrennt. Ziel ist vielmehr die semantische Auszeichnung der Textdokumente und die Ableitung einer spezifischen XML DTD, um z.B. eine effiziente struktur- und inhaltsbasierte Suche zu ermöglichen.

3 Das DIAsDEM-Vorgehensmodell

Für ein Archiv fachspezifischer Textdokumente verfolgt das in [8] vorgeschlagene DIAsDEM-Vorgehensmodell zwei Ziele: Erstens sind die Texte semantisch auszuzeichnen und zweitens ist eine das Archiv inhaltlich beschreibende, zunächst unstrukturierte XML-Dokumenttypdefinition abzuleiten. Dabei werden weder Dokumente in ihrer Gesamtheit klassifiziert noch individuelle Wörter annotiert. Das Ziel ist vielmehr die semantische Auszeichnung von strukturellen Komponenten der Textdokumente, die hier als Textelemente bezeichnet werden. Sinnvolle Textelemente sind z.B. Sätze, Absätze, Substantivgruppen oder sogar n-Gramme, die aus n aufeinanderfolgenden Wörtern bzw. Sätzen bestehen. Die drei folgenden, mit XML annotierten Sätze eines Handelsregistereintrags verdeutlichen dieses Konzept der semantischen Auszeichnung, wobei ein Textelement in diesem Beispiel jeweils einem Satz entspricht:

<Gegenstand> Die Planung, Projektierung und der Vertrieb von auch für den Einsatz in Bahnen geeigneten Telekommunikationsanlagen. </Gegenstand> <Stammkapital Geldbetrag="25000 EUR"> Stammkapital: 25.000 EUR. </Stammkapital> <AbschlussGesellschaftsvertrag Datum="22.02.1999"> Der Gesellschaftsvertrag wurde am 22. Februar 1999 geschlossen. </AbschlussGesellschaftsvertrag>

Die Semantik eines Textelements bzw. hier eines Satzes wird durch die jeweilige XML-Textmarke explizit und abfragbar widergespiegelt. XML-Textmarken enthalten zusätzlich Attribute, deren Namen und zugehörige Werte benannten Entitäten entsprechen, die im jeweiligen Anwendungsgebiet von besonderem Interesse sind. Neben Datumsangaben und Währungsbeträgen stellen z.B. auch Personen, Unternehmen oder Markenzeichen benannte Entitäten dar, die mit Methoden der Informationsextraktion identifizierbar sind.

Das DIAsDEM-Vorgehensmodell besteht aus zwei Phasen. Die erste Phase ist ein Prozess der Wissensentdeckung, der qualitativ hochwertige Segmente von Textelementen entdeckt, Textdokumente mit halbautomatisch ermittelten Segmentbezeichnern bzw. XML-Textmarken annotiert und schließlich eine zunächst unstrukturierte XML DTD für das Archiv ableitet. Dieser Text-Mining-Prozess ist „iterativ“, da der verwendete Clustering-Algorithmus mehrmals mit jeweils reduzierten Eingabedaten aufgerufen wird. Der hier verwendete Begriff des iterativen Clustering sollte aber nicht mit der Tatsache verwechselt werden, dass die Mehrzahl der Clustering-Algorithmen intern ebenfalls iterativ die Clusterzuordnungen von Datenpunkten bis zur Erreichung eines Konvergenzkriteriums verfeinern. Innerhalb von DIAsDEM werden jedoch in jeder Iteration die Parameter des Clustering-Algorithmus verändert. Der Prozess der Wissensentdeckung ist „interaktiv“, da ein Experte am Ende jeder Iteration für die endgültige Festlegung der Bezeichner qualitativ hochwertiger Segmente konsultiert wird. Die erste Phase konvertiert das initiale Trainingsarchiv in eine Sammlung semantisch ausgezeichneter XML-Dokumente sowie erzeugt abschließend eine XML DTD und eine Menge von Segmentbeschreibungen (bzw. den sog. „Clusterer für Texteinheiten“), deren Bezeichner als XML-Textmarken Elemente der abgeleiteten XML-Dokumenttypdefinition sind.

Die zweite Phase des DIAsDEM-Vorgehensmodells verwendet den zuvor erzeugten „Clusterer für Texteinheiten“, um in einem produktiven Stapelverarbeitungsprozess neue Textarchive des gleichen Anwendungsbereichs semantisch auszuzeichnen. Dabei wird der „Clusterer für Texteinheiten“ ebenfalls iterativ angewendet, um sämtliche Textelemente der neuen Archive den zuvor entdeckten Segmenten zuzuordnen. Textelemente, die Teil qualitativ hochwertiger Segmente sind, werden anschließend mit deren Bezeichnern als XML-Textmarken ausgezeichnet. Die in der ersten Phase abgeleitete XML DTD ist auch für alle in Phase 2 erzeugten XML-Dokumente gültig.

Das hier vorgestellte Vorgehensmodell ist nur für große, anwendungsspezifische Archive relativ homogener Textdokumente anwendbar, da der vorgeschlagene Text-Mining-Prozess Clustering-Algorithmen für die Entdeckung qualitativ hochwertiger Textelementsegmente nutzt. Die Anwendung der explorativen Datenanalyse mit Clustering-Algorithmen ist jedoch nur dann angemessen, wenn die Datenbasis zumindest eine Cluster-Tendenz aufweist ([9],

S. 267). Textelemente, die einer gemeinsamen Anwendungsdomäne entstammen, werden das Kriterium der Cluster-Tendenz eher erfüllen, als Textelemente aus Texten unterschiedlicher Fachbereiche. Die Ableitung einer XML-Dokumenttypdefinition ist zudem nur für Dokumente eines abgrenzbaren Anwendungsbereiches sinnvoll.

Trotz der Beschränkung auf große und fachspezifische Textarchive ist das hier vorgestellte Vorgehensmodell für die semantische Auszeichnung einer Vielzahl unterschiedlicher Dokumentsammlungen im Rahmen der Wettbewerberanalyse anwendbar: Beispielhaft sind Veröffentlichungen von administrativen Behörden und Gerichten, Geschäftsberichte wie Anhänge zu Jahresabschlüssen und Lageberichte börsennotierter Unternehmen, Ad-hoc-Mitteilungen und Unternehmensnachrichten sowie auf elektronischen Marktplätzen veröffentlichte Produkt- und Dienstleistungsbeschreibungen zu nennen.

4 Semantische Auszeichnung als Text Mining

Die erste Phase des DIAsDEM-Vorgehensmodells umfasst einen komplexen Prozess der Wissensentdeckung in Texten, der in diesem Abschnitt detailliert vorgestellt wird und in Abbildung 1 zusammengefasst dargestellt ist.

Das Trainingsarchiv fachspezifischer Textdokumente darf als Ausgangsbasis für den Text Mining-Prozess nur reine oder strukturell annotierte Texte enthalten. In letzterem Fall ist eine eindeutig definierte Textauszeichnungssprache (z.B. SGML) für die Kennzeichnung von Textstrukturen wie etwa Überschriften, Absätzen und Sätzen zu verwenden. Multimedia-Dokumente enthalten neben Texten vielfach Bilder, Grafiken oder auch Musik- und Videosequenzen. In diesen Fällen sind die relevanten Textabschnitte zuvor mittels individuell implementierter oder semiautomatischer Werkzeuge wie z.B. NoDoSE [2] aus den Dateien zu extrahieren.

Um das Ziel einer qualitativ hochwertigen semantischen Auszeichnung zu erreichen, wird in großem Umfang von Experten bereitgestelltes Fachwissen in den Text Mining-Prozess eingebunden: Ein kontrolliertes Vokabular enthält z.B. in Form eines Thesaurus anwendungsspezifische Begriffe und Konzepte, ein konzeptuelles Referenzschema spiegelt die Domäne wider und Beschreibungen der jeweilig interessanten benannten Entitäten (z.B. Personen oder Unternehmen) ermöglichen deren Identifizierung in Textelementen. Das vorläufige konzeptuelle Schema reflektiert die Anwendungsdomäne, deren Entitäten und ihre Beziehungen aus anfänglicher Expertensicht und dient später als Referenzschema für die abgeleitete XML DTD. Es ist jedoch nicht sichergestellt, dass diese DTD das Referenzschema enthält oder aber in diesem enthalten ist.

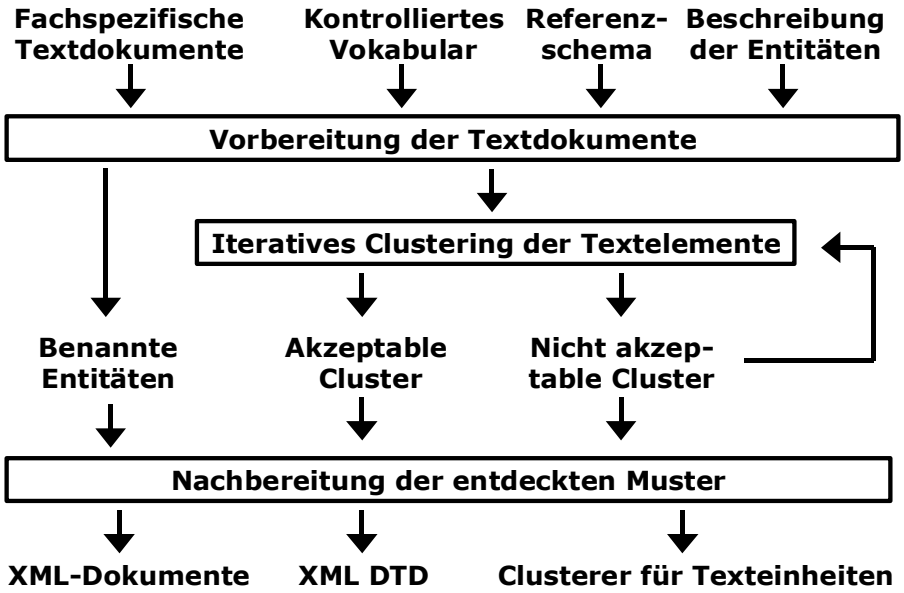

Abbildung 1: Der DIAsDEM-Prozess der Wissensentdeckung

4.1 Aufbereitung der Textdokumente

Analog zu einem konventionellen Data Mining-Prozess beginnt der hier dargestellte Prozess mit einer Phase, in dem ein reduzierter Merkmalsraum erzeugt wird. Innerhalb dieser Vorbereitungsphase ist zunächst die Einheit eines Textelements festzulegen. DIAsDEM betrachtet Textdokumente nicht in ihrer Gesamtheit, sondern als Menge struktureller Textelemente, deren Semantik durch XML-Textmarken explizit hervorgehoben werden soll. Diese Entscheidung hinsichtlich des Granularitätsgrades ist sehr bedeutsam, da im Verlauf des Vorgehensmodells nur die zu Beginn festgelegten Textelemente semantisch ausgezeichnet werden. In den aktuellen Fallstudien entspricht ein Satz jeweils einem Textelement.

Nachdem jedes Dokument in seine Textelemente zerlegt, benannte Entitäten mit einem speziellen Modul der *DIAsDEM Workbench* extrahiert und die grammatischen Grundformen sämtlicher Wörter ermittelt wurden, wird der Merkmalsraum festgelegt. In konventionellen Text Mining-Applikationen ist der Merkmalsraum meist durch die ggf. mit einem anwendungsspezifischen Vokabular begrenzte Menge aller Terme abzüglich bedeutungsloser, sog. Stopwörter festgelegt. Innerhalb des DIAsDEM-Vorgehensmodells wird der Merkmalsraum jedoch noch drastischer reduziert, um sowohl die Dimensionalität der Vektoren zu senken als auch die Ableitung einer das Archiv beschreibenden DTD zu erleichtern. Der Merkmalsraum setzt sich somit aus Begriffen und Konzepten zusammen, die (i) nicht selten innerhalb des Archivs auftreten und (ii) das Fachvokabular des jeweiligen Anwendungsbereichs reflektieren.

Die Bedingung (i) schließt alle Wörter mit einer sehr geringen Worthäufigkeit aus dem Merkmalsraum aus. Dazu zählen aber auch benannte Entitäten (z.B. Namen von Personen), die jedoch für spätere Anfragen relevant sind. Deswegen werden sie im Rahmen der Vorverarbeitung extrahiert und in der Nachbereitungsphase den XML-Textmarken als Attribute wieder zugeordnet. Bedingung (ii) schließt zusätzlich alle allgemeinsprachlichen Begriffe aus, die für die abzuleitende XML DTD nur sehr begrenzt nutzbar sind. Diese Bedingung kann letztendlich nur bei einer Merkmalauswahl durch Experten des Anwendungsgebiets erfüllt werden. Deshalb wird hier die konzeptuelle Modellierung der Domäne empfohlen. Die dabei für Entitäten, Attribute, Methoden oder Beziehungen gewählten Begriffe sind die Basis des fachspezifischen Merkmalsraums. Am Ende der Aufbereitungsphase wird der Merkmalsraum durch die Auswahl sog. Textelementdeskriptoren oder kurz Deskriptoren durch Experten festgelegt. Deskriptoren referenzieren jeweils einen Begriff, einen Oberbegriff für andere Begriffe oder ein Konzept, das verschiedene Begriffe umfasst.

Abschließend werden alle Textelemente des Archivs in Boolesche Vektoren des Merkmalsraums überführt. Der Wert 1 einer Dimension bedeutet hierbei, das der entsprechende Deskriptor Teil des Textelements ist. Das hier im Grundsatz für die Textrepräsentation verwendete Vektorraum-Modell wurde im Rahmen des Information Retrieval-Projekts Smart entwickelt [15]. Analog zu Smart wird außerdem eine sog. TF-IDF-Gewichtung der Deskriptoren vorgenommen. Dabei nimmt das Gewicht eines Textelementdeskriptors mit zunehmender Häufigkeit des gleichen Deskriptors im gesamten Archiv ab.

4.2 Iteratives Clustering der Textelemente

Kern des DIAsDEM-Vorgehensmodells zur semantischen Auszeichnung von Textarchiven ist das Clustering bzw. die Segmentierung der Textelementvektoren in Gruppen mit sehr ähnlichem Inhalt. Der Inhalt eines Clusters wird dabei durch die in diesem Segment dominierenden Deskriptoren (d.h. die Dimensionen des Merkmalsraums) reflektiert. Die dominierenden Deskriptoren werden anschließend verwendet, um Bezeichner für Cluster abzuleiten, die für eine semantische Auszeichnung der jeweiligen Textelemente verwendet werden.

Cluster-Analyse wird häufig als die Kunst bezeichnet, Gruppen innerhalb von Daten zu finden ([11], S. 1). Jain et al. definieren Clustering informell als die auf Ähnlichkeit basierende, unüberwachte Klassifikation von Objekten in Gruppen ([9], S. 265.) In den vergangenen Jahrzehnten ist eine Vielzahl erfolgreicher Clustering-Algorithmen aus verschiedenen Forschungsbereichen wie z.B. Statistik, Information Retrieval oder Informatik hervorgegangen. Diese Algorithmen spiegeln jeweils unterschiedliche Konzepte und Methodologien wider und sind meist für bestimmte Datentypen oder Anwendungsbereiche optimiert. Vor diesem Hintergrund basiert der hier beschriebene Prozess der

Wissensentdeckung auf einem Plug-In-Konzept, das die Nutzung verschiedener Clustering-Algorithmen innerhalb der *DIAsDEM Workbench* ermöglicht.

Wie in Abbildung 1 angedeutet, wird der jeweils gewählte Clustering-Algorithmus iterativ ausgeführt. Alle innerhalb einer Iteration entdeckten Segmente werden anhand der drei unten beschriebenen DIAsDEM-Qualitätskriterien evaluiert. Für sämtliche Cluster, die später entsprechend dieser Kriterien qualitativ hochwertig bzw. akzeptabel sind, wird wie in Abschnitt 4.3 beschrieben ein semantischer Bezeichner halbautomatisch abgeleitet. Die Textelementvektoren in akzeptablen Segmenten werden anschließend aus dem Datensatz entfernt, während die verbleibenden Vektoren erneut Eingabedaten für den Clustering-Algorithmus in der nächsten Iteration sind. In jeder Iteration werden zusätzlich restriktive Parameter des jeweiligen Algorithmus gelockert, um weitere Cluster zu entdecken. Dabei werden akzeptable Cluster inhaltlich allmählich weniger spezifisch. Das iterative Clustering-Verfahren setzt die Zielsetzungen von DIAsDEM bei der semantischen Auszeichnung von Texten um: Zunächst müssen primär XML-Textmarken entdeckt werden, welche die Semantik ihrer Textelemente möglichst präzise und spezifisch beschreiben. Werden jedoch keine weiteren präzisen Inhaltsbeschreibungen gefunden, so ist auch die Entdeckung semantisch allgemeinerer XML-Textmarken zu ermöglichen.

Nur qualitativ akzeptable Segmente werden semantisch bezeichnet und ermöglichen somit die Annotation der darin enthaltenen Textelemente mit XML-Textmarken. Ein entdeckter Cluster ist qualitativ hochwertig bzw. akzeptabel, wenn (i) die darin enthaltenen Vektoren semantisch homogen sind, (ii) die Kardinalität des Segments groß ist und (iii) der Inhalt des Segments durch eine geringe Anzahl dominierender Deskriptoren beschrieben werden kann. Alle Parameter dieser drei DIAsDEM-Qualitätskriterien werden interaktiv durch den Wissensingenieur festgelegt. Die Umsetzung der Bedingung (i) stellt der ähnlichkeitsbasierte Clustering-Algorithmus sicher. Die geforderte Homogenität der gefundenen Segmente wird jedoch wie oben beschrieben schrittweise gesenkt, um die Maximierung der beiden weiteren Bedingungen zu ermöglichen. Die Bedingungen (ii) und (iii) regeln die erforderliche Segmentgröße und die geforderte semantische Reinheit der Cluster-Inhalte mittels einstellbarer Schwellenwerte. Die dritte Bedingung soll für die spätere Ableitung sinnvoller semantischer XML-Textmarken sicherstellen, dass akzeptable Cluster nur Textelemente beinhalten, die alle durch wenige, dafür aber in fast allen Textelementen enthaltenen Deskriptoren beschrieben werden.

4.3 Nachbereitung entdeckter Muster

Ergebnis der Phase des iterativen Clustering ist eine Menge qualitativ akzeptabler Segmente. Die *DIAsDEM Workbench* annotiert alle akzeptablen Cluster mit Statistiken des Clustering-Algorithmus und verbalen Cluster-Beschrei-

bungen, die aus den in einem Segment dominierenden Deskriptoren gebildet werden. Diese Cluster-Beschreibungen werden zusammen mit den Namen von zuvor extrahierten benannten Entitäten verwendet, um semantische Cluster-Bezeichner bzw. XML-Textmarken abzuleiten. Die endgültigen Namen der XML-Textmarken werden jedoch durch einen fachkundigen Experten festgelegt.

Der Experte wird bei dieser Tätigkeit durch automatisch generierte Cluster-Beschreibungen und entsprechend vorgeschlagene Standardbezeichner für jeden akzeptablen Cluster unterstützt. Die Beschreibung eines Segments enthält die in diesem Cluster dominierenden Deskriptoren, die absteigend nach der relativen Häufigkeit ihres Auftretens innerhalb des Clusters geordnet sind. Das Visualisierungsmodul der *DIAsDEM Workbench* unterstützt den Experten bei der Wahl der XML-Textmarken, indem die jeweiligen Textelemente sowie weitere, häufig auftretende Begriffe und die berechneten Statistiken angezeigt werden.

Nach der endgültigen Festlegung der Namen von XML-Textmarken werden die Textdokumente des Trainingsarchivs abschließend in semantisch annotierte XML-Dokumente überführt: Alle Textelemente, deren Vektoren Teil akzeptabler Cluster sind, werden mit der entsprechenden XML-Textmarke ausgezeichnet. Textmarken werden außerdem um Attribute ergänzt, die den in der Aufbereitungsphase extrahierten benannten Entitäten entsprechen. Die ebenfalls vorgesehene semantische Benennung dieser Attribute ist jedoch Bestandteil künftiger Arbeit. Textelemente, deren Vektoren entweder Teil nicht akzeptabler Cluster oder die keinem Cluster zugeordnet sind, werden nicht semantisch annotiert.

Im letzten Schritt wird eine XML-Dokumenttypdefinition für das Archiv abgeleitet, welche die Sammlung von XML-Dokumenten inhaltlich als Enumeration zulässiger XML-Textmarken charakterisiert und als sehr einfaches, vorläufiges und Datenbank-ähnliches Schema angesehen werden kann. Bei einer Implementierung des Quasi-Schemas in einem DBMS entsprechen die XML-Textmarken den Relationen und die Attributnamen einer XML-Textmarke den Attributen der jeweiligen Relation. Die gegenwärtig abgeleitete, eher vorläufige XML DTD ist unstrukturiert und enthält somit keine Aussagen über die Ordnung von Textmarken oder die Existenz geschachtelter Textmarken. Eine weitere Strukturierung der vorläufigen XML DTD ist Teil künftiger Forschungsarbeit.

5 Fallstudie zur semantischen Auszeichnung

DIAsDEM ist ein allgemeingültiges Vorgehensmodell, dessen in Java und z.T. in Perl prototypisch implementierte *DIAsDEM Workbench* mit anwendungsspezifischen Thesauri und Regeln für die Identifikation benannter Entitäten

sowie mit verschiedenen Clustering-Algorithmen gekoppelt werden kann. Sie unterstützt sämtliche Phasen des DIAsDEM-Vorgehensmodells. In dieser Fallstudie wurde das Vorgehensmodell angewendet, um ein Archiv deutscher Handelsregistereinträge semantisch zu annotieren.

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 potentielle 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.		

Abbildung 2: Exemplarischer Handelsregistereintrag

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. Im Rahmen dieser Fallstudie wurde ein Archiv von 1.145 im Internet veröffentlichten Handelsregistereintragungen des Amtsgerichts Potsdam semantisch ausgezeichnet. Das sind sämtliche Eintragungen des Jahres 1999, die Unternehmensneugründungen anzeigen.

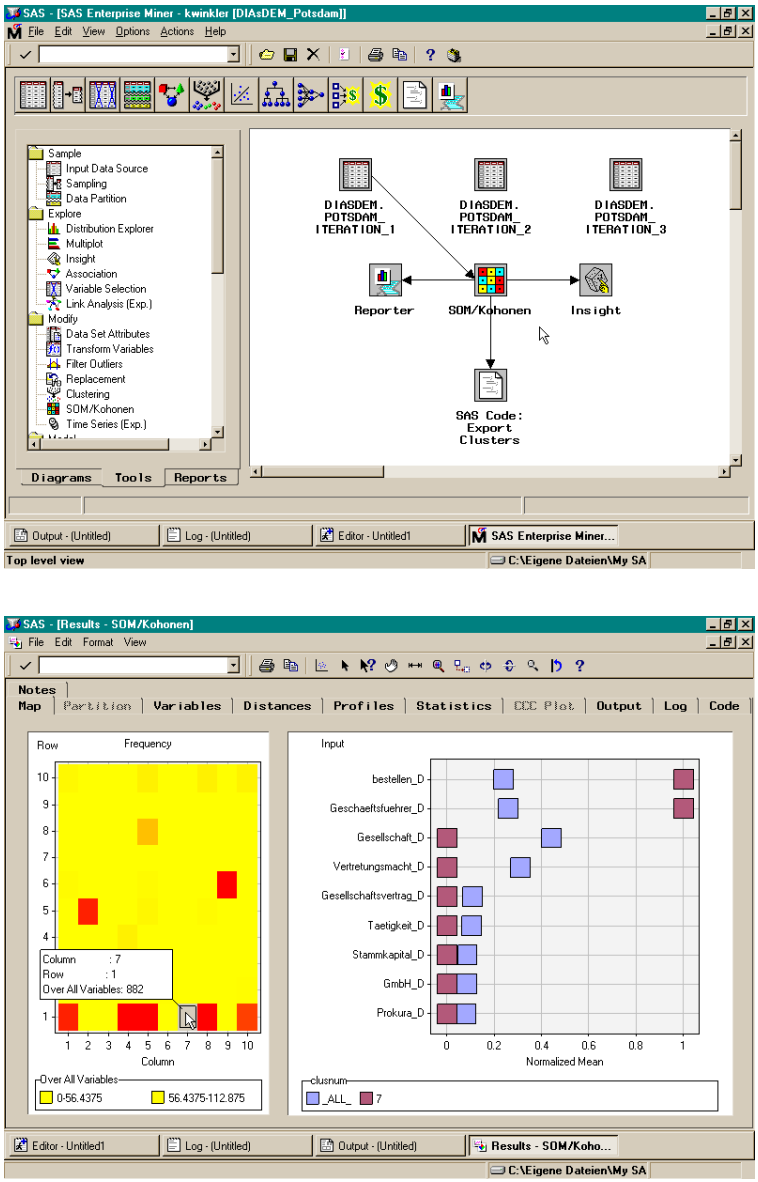

Abbildung 3: Einsatz des SAS Enterprise Miner in der Fallstudie

Die Details dieser Fallstudie, der Leser möge sie bitte [20] entnehmen, können hier nur zusammenfassend dargestellt werden: Da die Sätze der Eintragungen offensichtlich eine Cluster-Tendenz aufwiesen, wurde als Textelement der grammatikalische Satz festgelegt. Die Handelsregistereintragungen wurden in insgesamt 10.785 Textelemente zerlegt. Anschließend wurde *NEEX*, der auf Basis von Regeln und Heuristiken funktionierende *Named Entity Extractor* der *DIAsDEM Workbench* verwendet, um die benannten Entitäten „Person“, „Unternehmen“, „Datum“, „Geldbetrag“ und „Paragraph“ in den Textelementen zu identifizieren. Der mehrsprachige *Part-of-Speech-Tagger TreeTagger* [16] wurde danach eingesetzt, um durch Ermittlung der grammatikalischen Wort-

grundformen die Dimensionalität bereits von 10.613 auf etwa 5.400 verschiedene Wortformen zu senken. Nach der sich anschließenden konzeptuellen Modellierung des Anwendungsgebiets mittels UML-Klassendiagrammen wurde ein spezieller Thesaurus mit 85 Deskriptoren und 109 Nicht-Deskriptoren, die auf gültige Deskriptoren verweisen, erstellt.

Die *DIAsDEM Workbench* erzeugte die Textelementvektoren und steuerte danach den Prozess des iterativen Clustering der Vektoren. 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 ([6], S. 271-272). Abbildung 4 zeigt oben 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. Unten ist das in der ersten Iteration erzeugte 10x10-Gitter sowie die Visualisierung eines Clusters abgebildet, in dessen 882 Vektoren die Konzepte „bestellen“ und „Geschäftsführer“ dominieren. Deshalb wurden diese Textelemente anschließend von der *DIAsDEM Workbench* mit der XML-Textmarke „BestellungGeschäftsführer“ ausgezeichnet. Nach drei Iterationen wurden insgesamt 68 qualitativ akzeptable Segmente entdeckt und halbautomatisch semantisch benannt. Diese 68 akzeptablen Cluster enthielten etwa 95 Prozent aller Texteinheiten.

```
<?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. </Bestellung Geschaeftsfuehrer>
(...) <Bekanntmachungen> Nicht eingetragen: Die Bekanntmachungen der Gesell-
schaft erfolgen im Bundesanzeiger. </Bekanntmachungen>
</Handelsregistereintrag >
```

Abbildung 4: Semantisch annotierter Handelsregistereintrag

Abbildung 4 zeigt einen Auszug des semantisch annotierten Handelsregistereintrags aus Abbildung 2, der nach Auszeichnung des Archiv von der *DIAsDEM Workbench* erzeugt wurde. Abbildung 5 enthält einen Auszug der abgeleiteten, unstrukturierten XML-Dokumenttypdefinition. Diese vorläufige

XML DTD beschreibt grob die semantische Struktur des erzeugten Archivs annotierter Handelsregistereintragungen.

Im Gegensatz zur automatischen Textklassifikation gibt es in dieser Domäne keine zuvor annotierten Trainingsdokumente, mit deren Hilfe die Genauigkeit der semantischen Auszeichnung überprüft werden kann. Aus diesem Grund wurde eine Zufallsstichprobe gezogen, die 5 Prozent aller Textelemente enthielt. Ein Experte wurde gebeten, die Qualität der semantischen Auszeichnung im Hinblick auf zwei Fehlerarten zu analysieren: Fehlertyp I tritt auf, wenn eine XML-Textmarke nicht den genauen Inhalt des Textelements reflektiert. Fehlertyp II tritt hingegen auf, wenn ein nicht ausgezeichnetes Textelement ein semantisches Konzept beinhaltet, das aber Teil der abgeleiteten XML DTD ist.

```
<?xml version="1.0" encoding="ISO-8859-1"?>

<!ELEMENT Handelsregistereintrag (#PCDATA | Gegenstand | Bekanntmachungen |
BeschlussGesellschafter_Stammkaptital | Stammkapital | VerlegungSitz |
PersoenlichHaftendeGesellschafter | BestellungGeschaeftsfuehrer | GmbH |
OffeneHandelsgesellschaft | MitgliederAufsichtsrat | AufteilungGrundkapital |
AbschlussGesellschaftsvertrag | (...) | GruendungGesellschaft )* >

<!ELEMENT Gegenstand (#PCDATA)>
<!ELEMENT Bekanntmachungen (#PCDATA)> (...)
<!ELEMENT GruendungGesellschaft (#PCDATA)>

<!ATTLIST Stammkapital Geldbetrag CDATA #IMPLIED> (...)
<!ATTLIST BestellungGeschaeftsfuehrer Person CDATA #IMPLIED> (...)
<!ATTLIST GruendungGesellschaft Datum CDATA #IMPLIED>
```

Abbildung 5: Abgeleitete XML-Dokumenttypdefinition (Auszug)

Innerhalb der gezogenen Zufallsstichprobe betrug die Fehlerrate für den Fehlertyp I (Fehlertyp II) 1,5 Prozent (1,8 Prozent). Die Gesamtfehlerrate innerhalb der gezogenen Stichprobe betrug somit 3,2 Prozent. Auf einem Konfidenzniveau von 0,95 liegt die Gesamtfehlerrate im Intervall [2,4 Prozent; 4,2 Prozent]. Das ist ein vielversprechendes Ergebnis für eine erste Evaluation des DIAsDEM-Vorgehensmodells im Rahmen einer Fallstudie.

6 Zusammenfassung und Ausblick

In diesem Artikel wurde das DIAsDEM-Vorgehensmodell zur halbautomatischen semantischen Auszeichnung anwendungsspezifischer Textarchive vorgestellt. Es beinhaltet einen Prozess der Wissensentdeckung, um in fachspezifischen Texten häufig vorhandene, aber i.d.R. undokumentierte semantische Strukturen zu entdecken. Dabei gruppiert ein iteratives Clustering-Verfahren semantisch ähnliche Textelemente, ermittelt halbautomatisch Bezeichner für qualitativ akzeptable Cluster, annotiert die zugehörigen Textelemente mit

XML-Textmarken und leitet eine vorläufige, unstrukturierte XML-Dokumenttypdefinition ab. Die Textmarken werden zusätzlich durch Attribute ergänzt, deren Werte zuvor extrahierte benannte Entitäten sind. Das Vorgehensmodell wurde in einer Fallstudie erfolgreich evaluiert.

Sowohl der hohe Anteil semantisch ausgezeichneter Sätze als auch die niedrigen Fehlerraten im Rahmen der Fallstudie könnten durch die sehr formalisierte und teilweise antiquierte juristische Sprache der Handelsregistereintragungen erklärbar sein. Aus diesem Grund wird das DIAsDEM-Vorgehensmodell gegenwärtig in einem sprachlich flexibleren und vielseitigeren Anwendungsgebiet evaluiert: Ad-hoc-Mitteilungen werden von börsennotierten Unternehmen herausgegeben und enthalten Nachrichten über aktuelle unternehmerische Entwicklungen, die potentiell den Aktienkurs beeinflussen können. Diese Fallstudie ist derzeit noch nicht abgeschlossen. Die ersten Ergebnisse einer Evaluation Auszeichnungsqualität sind jedoch vielversprechend.

Offene Forschungsaspekte sind derzeit die Bewertung klassischer Clustering-Algorithmen und Ähnlichkeitsmetriken im Hinblick auf die Verfahrensziele, die halbautomatische Auswahl der Vektordimensionen sowie die Ableitung strukturierter Dokumenttypdefinitionen. Diese Schemata sollen (ggf. probabilistische) Informationen zu Reihenfolge und Verschachtelung der XML-Textmarken enthalten. Erste Ansätze zur Strukturierung der vorläufigen XML DTD wurden in [19] vorgestellt. Zusätzlich sollen Attribute von Textmarken, die extrahierte Entitäten wie z.B. Namen von Personen und Unternehmen enthalten, semantisch benannt und im Typ festgelegt werden. Es ist ebenso geplant, ein Web-basiertes Informationssystem prototypisch zu implementieren, dass Anfragen an ein Archiv semantisch annotierter Handelsregistereintragungen ermöglicht. Im Gegensatz zur konventionellen Volltextsuche wird das System durch Auswertung von XML-Textmarken und deren Attributen auch strukturbasierte Anfragen unterstützen. Zu diesem Zweck sind aktuell verfügbare XML-Anfragesprachen zu evaluieren, die sowohl inhalts- als auch strukturbasierte Anfragen an XML-Archive ermöglichen.

Literatur

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

[2] Adelberg, B. (1998). NoDoSE - A Tool for Semi-Automatically Extracting Semi-Structured Data from Text Documents. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA, 283-294.

[3] Bruder, I., Düsterhöft, A., Becker, M., Bedersdorfer, J. und Neumann, G. (2000). GETESS: Constructing a Linguistic Search Index for an Internet Search Engine. In: Proceedings of the 5th International Conference on Applications of Natural Language to Information System, Versailles, France, 227-238.

[4] Decker, S., Erdmann, M., Fensel, D., Studer, R. (1999). ONTOBROKER: Ontology Based Access to Distributed and Semi-Structured Information. In: Database Semantics: Semantic Issues in Multimedia System, R. Meersman et al. (Hrsg.), Kluwer Academic Publisher, 351-369.

[5] Embley, D. W., Cambell, D. M., Smith, R.D. und Liddle, S. W. (1998). Ontology-Based Extraction and Structuring of Information from DataRich Unstructured Documents. In: Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management, Bethesda, MD, USA, 52-59.

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

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

[8] Graubitz, H., Spiliopoulou, M. und Winkler, K. (2001). 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, 171-178.

[9] Jain, A. K., Murty, M. N. und Flynn, P. J. (1999). Data Clustering: A Review. In: ACM Computing Surveys 31, **3**, 264-323.

[10] Kahaner, L. (1998). Competive Intelligence. Touchstone Books.

[11] Kaufman, L. und Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York.

[12] Lumera, J. (2000). Große Mengen an Altlastdaten stehen XML-Umstieg im Weg. In: Computerwoche 27, **16**, 52-53.

[13] Maedche, A. und Staab, S. (2001). Learning Ontologies for the Semantic Web. In: IEEE Intelligent Systems 16, **2**, Special Issue on the Semantic Web, 72-79.

[14] Moore, G. W. und Berman, J. J. (2001). Anatomic Pathology Data Mining. In: Medical Data Mining and Knowledge Discovery. Cios, K. J. (Hrsg.), Physica-Verlag, Heidelberg New York, 72-117.

[15] Salton, G. und Buckley, C. (1998). Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24, **5**, 513-523.

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

[17] Sullivan, D. (2001). Data Document Warehousing and Text Mining. John Wiley & Sons, New York.

[18] Tan, A.-H. (1999). Text Mining: The State of the Art and the Challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, Beijing, China, 65-70.

[19] Winkler, K. und Spiliopoulou, M. (2001). Extraction of Semantic XML DTDs from Texts Using Data Mining Techniques. In: Proceedings of the K-CAP 2001 Workshop on Knowledge Markup and Semantic Annotation, Victoria, BC, Canada, 59-68.

[20] Winkler, K. und Spiliopoulou, M. (2001). Semi-Automated XML Tagging of Public Text Archives: A Case Study. In: Proceedings of EuroWeb 2001 „The Web in Public Administration“. Pisa, Italy, 271-285.