Semantische Auszeichnung von Textarchiven mit Data Mining-Methoden

Autor: Dipl.-Kfm. Karsten Winkler

Handelshochschule Leipzig

Lehrstuhl für Wirtschaftsinformatik des E-Business

Jahnallee 59

D-04109 Leipzig

Tel.: +49 (341) 9851 761

E-mail: kwinkler@ebusiness.hhl.de

http://ebusiness.hhl.de/staff/kwinkler

Betreuerin: Prof. Dr. Myra Spiliopoulou

Handelshochschule Leipzig

Lehrstuhl für Wirtschaftsinformatik des E-Business

Jahnallee 59

D-04109 Leipzig

Tel.: +49 (341) 9851 763

E-mail: myra@ebusiness.hhl.de

http://ebusiness.hhl.de/staff/myra

Semantische Auszeichnung von Textarchiven mit Data Mining-Methoden

Karsten Winkler [1]

Handelshochschule Leipzig (kwinkler@ebusiness.hhl.de)

Zusammenfassung: Unternehmen, staatliche und nichtstaatliche Organisationen verfügen meist über große, stetig wachsende Archive fachspezifischer Textdokumente, die eine wichtige Quelle expliziten Wissens darstellen. Ein erfolgreiches Wissensmanagement zur Schaffung und Sicherung nachhaltiger Wettbewerbsvorteile erfordert jedoch im Gegensatz zur konventionellen Volltextsuche fortgeschrittene Verfahren für Suche und Informationsintegration. Die semantische, d.h. inhaltsbezogene Auszeichnung mittels XML ist dabei ein Verfahren der Erschließung von Textarchiven. In diesem Artikel wird das DIAsDEM-Vorgehensmodell zur semiautomatischen semantischen Annotation fachspezifischer Textarchive vorgestellt. Das Vorgehensmodell beinhaltet einen komplexen Prozeß der Wissensentdeckung in textuellen Daten, um die fachspezifischen Archiven häufig inhärente, jedoch undokumentierte semantische Struktur zu entdecken. Ein iteratives Clustering-Verfahren gruppiert dabei semantisch ähnliche Textelemente (z.B. Sätze), ermittelt halbautomatisch Bezeichner für qualitativ hochwertige Cluster, annotiert die betreffenden Textelemente mit XML-Textmarken und leitet eine vorläufige, unstrukturierte XML-Dokumenttypdefinition ab. Die Textmarken werden durch Attribute ergänzt, die zuvor extrahierte benannte Entitäten (z.B. Personen oder Unternehmen) beinhalten. Das vorgestellte Verfahren wird abschließend in einer Fallstudie zum deutschen Handelsregister evaluiert.

Schlüsselworte: Wissensentdeckung, Data Mining, Text Mining, semantische Auszeichnung, XML, XML-Dokumenttypdefinition, textuelle Altlastdaten

1 Einleitung

Organisationen erzeugen und speichern derzeit tendenziell mehr elektronisch verfügbare Daten als sie zielorientiert verwenden können. Diese sprichwörtliche

[1] Diese Arbeit wird durch die Deutsche Forschungsgesellschaft im Rahmen des Projekts DIAsDEM - Datenintegration von Altlastdaten und semistrukturierten Dokumenten mit Mining-Verfahren unterstützt (DFG-Zuwendung SP 572/4-1).

Informationsflut resultierte in einer aktiven Forschung zu Fragestellungen der „Wissensentdeckung in Datenbanken“ mit Methoden des Data Mining. Ziel ist die Gewinnung von neuem, nicht trivialem, interessantem und vor allem ökonomisch umsetzbarem Wissen aus riesigen Datenbeständen. Eingesetzt wird dafür ein prozeßorientierter Rahmen, in dem Methoden der Statistik, des maschinellen Lernens, der künstlichen Intelligenz und der Datenbankforschung zusammengeführt werden [FPS96, S. 29-31].

Hinsichtlich des Grades an interner Struktur werden strukturierte, z.B. relationale oder objekt-relationale Daten von semistrukturierten [Bune97] 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 [Tan99], werden seit Mitte der neunziger Jahre auch Methoden der Wissensentdeckung eingesetzt, um Texte zu kategorisieren und inhaltlich ähnliche Dokumente zu erkennen. Feldman und Dagan prägten für die Forschungsrichtung den Ausdruck „Wissensentdeckung in textuellen Datenbanken“ [FeDa95].

Organisationen verfügen häufig über große, anwendungsspezifische Textarchive homogener Dokumente, die aber mangels fehlender Technologien meist nicht oder nur unzureichend (z.B. mittels Volltextsuche) im Wertschöpfungsprozeß eingesetzt werden. Insbesondere innovative Unternehmen sind aber derzeit bestrebt, ein innerbetriebliches Wissensmanagement umzusetzen. Dabei wird explizites, bereits in Dokumentsammlungen gespeichertes Wissen als eine wichtige, i.d.R. aber erst zu erschließende Quelle für nachhaltige, wissensbasierte Wettbewerbsvorteile angesehen. Ein Verfahren für die Erschließung von z.B. in Archiven gesammelten Projektberichten oder Memos ist die feinkörnige semantische, d.h. inhaltsbezogene Auszeichnung der Textdokumente. Nach erfolgter Auszeichnung eines Textarchivs mit z.B. der Textauszeichnungssprache XML können die semantischen Zusatzinformationen in Form der Dokumenttypdefinition (DTD) und der eingefügten XML-Textmarken bspw. für eine inhalts- und strukturbasierte Suche sowie für Zwecke der weiterführenden Wissensentdeckung, des Wissensmanagements und der Informationsintegration mit weiteren, inhaltlich relevanten Datenquellen genutzt werden.

Aktuelle Verfahren der Wissensentdeckung in textuellen Datenbanken analysieren Dokumente entweder nur in ihrer Gesamtheit oder aber mit Techniken der Sprachverarbeitung auf Wortebene. Beide Ansätze sind für die hier betrachteten Textarchive homogener Dokumente nur bedingt geeignet, da einerseits der grobe Inhalt dieser Texte ohnehin bekannt ist. Andererseits weichen sprachlich komplexe Dokumente wie z.B. Veröffentlichungen von Amtsgerichten, Produktbeschreibungen auf elektronischen Marktplätzen oder Geschäftsberichte börsennotierter Unternehmen stark von der Alltagssprache ab, so daß die Einbeziehung von Fachwissen in den Prozeß der Wissensentdeckung dringend erforderlich erscheint. Aus diesen Gründen wird in diesem Artikel eine prozeßorientierte Methodologie zur feinkörnigen Wissensentdeckung auf der Ebene struktureller Textelemente wie z.B. Sätze oder Absätze vorgestellt und im Rahmen einer Fallstudie evaluiert.

Das Ziel des DIAsDEM-Vorgehensmodells ist die Teilstrukturierung großer, anwendungsspezifischer Archive homogener Textdokumente 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 gegenwärtig noch unstrukturierten XML DTD aggregiert werden. Um den menschlichen Arbeitsaufwand zu minimieren, wurde dazu ein komplexes Verfahren der Wissensentdeckung in textuellen Datenbanken vorgeschlagen.

Die vorliegende Arbeit ist in das Forschungsprojekt DIAsDEM eingebunden, dessen deutsches Akronym für „Datenintegration von Altlastdaten und semistrukturierten Dokumenten mit Mining-Verfahren“ steht. In diesem Projekt werden Verfahren der Wissensentdeckung eingesetzt, um die Integration von Texten mit relevanten strukturierten und semistrukturierten Daten vorzubereiten. Im Rahmen des Projekts, das durch die Deutsche Forschungsgemeinschaft finanziert wird, kooperieren die Gruppen um Prof. Stefan Conrad an der Ludwig-Maximilians-Universität München (Institut für Informatik) und um Prof. Dr. Myra Spiliopoulou an der Handelshochschule Leipzig (Lehrstuhl für Wirtschaftsinformatik des E-Business). Das ultimative Ziel von DIAsDEM ist die Zusammenführung von Altlastdaten und semistrukturierten Dokumenten in einem integrierten Informationssystem, das für Zwecke der Entscheidungsunterstützung mit einer geeigneten Anfragesprache abgefragt werden kann.

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 des Artikels erörtert nach einer Zusammenfassung einzelne Aspekte der künftigen Forschung im Rahmen des Projekts.

2 Relevante Literatur im Überblick

Die für das DIAsDEM-Vorgehensmodell relevanten Arbeiten können wie folgt kategorisiert werden: Wissensentdeckung in textuellen Datenbanken, Forschung zu semistrukturierten Daten und Forschungsprojekte mit ähnlichen Zielsetzungen.

Tan faßt kurz den aktuellen Stand und künftige Herausforderungen des Forschungsgebietes Wissensentdeckung in textuellen Datenbanken zusammen

[Tan99]. Der Autor stellt ein allgemeingültiges, zwei Phasen umfassendes Vorgehensmodell für Text Mining-Aktivitäten vor: In der ersten Textaufbereitungsphase werden unstrukturierte Textdokumente zunächst in eine intermediäre Datenstruktur überführt, die anschließend in der Phase der Wissensentdeckung genutzt wird. Dieses allgemeine Modell ist ebenfalls Basis des DIAsDEM-Vorgehensmodells, das feinkörnige semantische Analysen unterstützt und Fachwissen einbettet. Diese beiden Aspekte sind nach Tan offene Forschungsprobleme des Text Mining.

Nahm und Mooney schlagen die Kombination von Methoden der Wissensentdeckung in Datenbanken und der Informationsextraktion vor, um Text Mining-Aufgaben zu lösen [NaMo00]. Die Autoren wenden Standardverfahren der Wissensentdeckung auf strukturierte Datensätze an, die zuvor aus Texten extrahierte, anwendungsspezifische benannte Entitäten (z.B. Personen) enthalten. Feldman et al. schlagen vor, Text Mining auf Begriffsebene und nicht auf der Ebene linguistisch annotierter Wörter auszuführen [FFK+98]. Die Autoren repräsentieren jedes Dokument durch eine Menge von Begriffen und konstruieren zusätzlich eine Begriffshierarchie. Diese Daten sind anschließend Eingabedaten für Algorithmen der Wissensentdeckung wie z.B. der Entdeckung von Assoziationsregeln. Das DIAsDEM-Vorgehensmodell repräsentiert ebenfalls Texte durch Begriffe und Konzepte. Im Gegensatz zu [FFK+98] ist das Ziel von DIAsDEM aber die semantische Auszeichnung von Textelementen sowie die Ableitung einer DTD und nicht die inhaltliche Charakterisierung eines gesamten Dokuments. Loh et al. empfehlen die Extraktion von Konzepten anstatt einzelner Wörter aus Texten für eine weitere Verwendung im Prozeß der Wissensentdeckung auf Dokumentebene [LWO00]. Ähnlich zum DIAsDEM-Vorgehensmodell empfehlen die Autoren die Nutzung eines kontrollierten Vokabulars (z.B. eines Thesaurus) für die Konzeptextraktion aus Texten. Mikheev und Finch beschreiben eine Umgebung zur Gewinnung von Fachwissen aus Texten [MiFi95]. Analog zu DIAsDEM kombinieren die Autoren Methoden aus verschiedenen Forschungsgebieten in einem einheitlichen Vorgehensmodell.

Das Ziel einer Extraktion semantischer Konzepte verfolgen sowohl diese Autoren als auch das DIAsDEM-Vorgehensmodell. Die von DIAsDEM zu extrahierenden Konzepte müssen jedoch als Elemente einer das Archiv beschreibenden XML DTD geeignet sein. So ist z.B. ein semantisches Konzept, das nur ein einzelnes Textelement (wenn auch perfekt) beschreibt, als Element einer globalen DTD ungeeignet. Um eine XML DTD abzuleiten ist es vielmehr erforderlich, Gruppen von Textelementen zu entdecken, die ähnliche semantische Konzepte beinhalten. Darüber hinaus beschränkt sich das DIAsDEM-Vorgehensmodell auf anwendungsspezifische Texte, die im Hinblick auf Worthäufigkeiten stark von der Alltagssprache abweichen. Diese Textarchive können nur schwer mit standardisierten Text Mining-Verfahren analysiert werden, da die Einbeziehung von Fachwissen hier eine Voraussetzung für eine erfolgreiche Wissensentdeckung ist.

Semistrukturierte Daten sind ein weiteres aktuelles und relevantes Forschungsgebiet innerhalb der Informatik [ABS00, Bune97]. Hier wurden leistungsfähige

Methoden entwickelt, die eine vorhandene Struktur in einer Menge ähnlicher semistrukturierter Dokumente erstens ableiten und zweitens diese entdeckte Struktur in Anlehnung an ein Datenbankschema, meist als gerichteten Graph repräsentieren [LMP00, NAM97, WaLi00]. Diese Ansätze leiten jedoch nur ein Schema für eine gegebene Menge semistrukturierter Dokumente ab. In Kontrast dazu sind im Rahmen des DIAsDEM-Vorgehensmodells jedoch simultan sowohl unstrukturierte Textdokumente teilzustrukturieren und es ist eine das Archiv beschreibende XML DTD abzuleiten. Auch Sengupta und Purao schlagen eine Methode vor, um eine XML DTD für bereits existierende XML-Dokumente abzuleiten [SePu00]. Im Gegensatz dazu dient das DIAsDEM-Vorgehensmodell jedoch der Annotation von Textdokumenten und leitet zusätzlich eine Dokumenttypdefinition ab. Inhaltlich näher zu DIAsDEM ist hingegen die Arbeit von Lumera, der Schlüsselwörter und Regeln verwendet, um Altlastdaten wie z.B. Fertigungshandbücher halbautomatisch in XML-Dokumente zu überführen [Lume00]. Dieser Ansatz basiert jedoch auf einer manuell erstellten Regelbasis, während DIAsDEM einen Prozeß der Wissensentdeckung zur Minimierung des menschlichen Aufwands in das Vorgehensmodell integriert.

Ähnliche Ziele verfolgen derzeit nur wenige folgende Forschungsgruppen: Bruder et al. stellen mit GETESS einen Anfrage- und Suchdienst für das Web vor, der Ontologie-basiert HTML-Dokumente in XML-Zusammenfassungen überführt [BDB+00]. 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 deren detaillierte Inhalte für weitere Analysen und eine Visualisierung zu erhalten. Moore und Berman präsentieren ein Verfahren zur Überführung textueller pathologischer Berichte, d.h. Untersuchungen von Gewebeproben, in XML-Dokumente [MoBe01]. 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.

Decker et al. verwenden das Ontologie-basierte System ONTOBROKER, um Metadaten aus Webseiten zu extrahieren [DEFS99]. Embley et al. nutzen ebenfalls Ontologien, um Informationen aus fachspezifischen, unstrukturierten Texten sowohl zu extrahieren als auch zu strukturieren [ECSL98]. Maedche und Staab stellen eine Architektur vor, die das halbautomatische Lernen von Ontologien aus HTML-Dokumenten unterstützt [MaSt01]. TURBIO ist ein System von Turmo et al. für die Informationsextraktion, das Methoden des maschinellen Lernens anwendet, um Extraktionsregeln aus anwendungsspezifischen Texten abzuleiten [TCR98]. Innerhalb des DIAsDEM-Vorgehensmodell 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 u.a. eine effiziente struktur- und inhaltsbasierte Suche zu ermöglichen.

3 Das DIAsDEM-Vorgehensmodell

Für ein gegebenes Archiv anwendungsspezifischer Textdokumente verfolgt das in [GSW01, GWS01] 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 annotierten Sätze eines Handelsregistereintrags verdeutlichen dieses Konzept der semantischen Auszeichnung, wobei ein Textelement in diesem Beispiel einem Satz entspricht:

> <BusinessPurpose> Die Planung, Projektierung und der Vertrieb von auch für den Einsatz in Bahnen geeigneten Telekommunikationsanlagen. </BusinessPurpose> <ShareCapital AmountOfMoney="25000 EUR"> Stammkapital: 25.000 EUR. </ShareCapital> <ConclusionArticles Date="22. 02. 1999"> Der Gesellschaftsvertrag wurde am 22. Februar 1999 geschlossen. </ConclusionArticles>

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 dient der semantischen Auszeichnung von Textelementen. Es sei deshalb noch einmal betont, daß die Menge aller Textelemente in sämtlichen Dokumenten die Datenbasis für den Prozeß der Wissensentdeckung darstellt. Eingabedaten sind aus diesen Grund also weder die Menge der Dokumente noch einzelne Textelemente.

In der Phase der Wissensentdeckung werden die Textelemente entsprechend ihrer Inhalte segmentiert, um anschließend XML-Textmarken für qualitativ hochwertige Segmente und zusätzlich eine XML DTD abzuleiten. Diese Ziele des Vorgehensmodells sind eine besondere Herausforderung für die anzuwendende Clustering-Methodologie: Erstens kann nur semantisch homogenen Segmenten ein

sinnvoller Bezeichner zugeordnet werden und zweitens darf ein Segment von Textelementen nicht zu spezifisch sein, denn ein aus vielen semantischen Konzepten zusammengesetzter Bezeichner kann als DTD-Element nur bedingt für Anfragen genutzt werden. Drittens sollte die Kardinalität qualitativ hochwertiger Segmente groß sein, da viele und somit sehr spezielle XML-Textmarken ebenfalls nicht effizient für Anfragen an ein Informationssystem verwendet werden können.

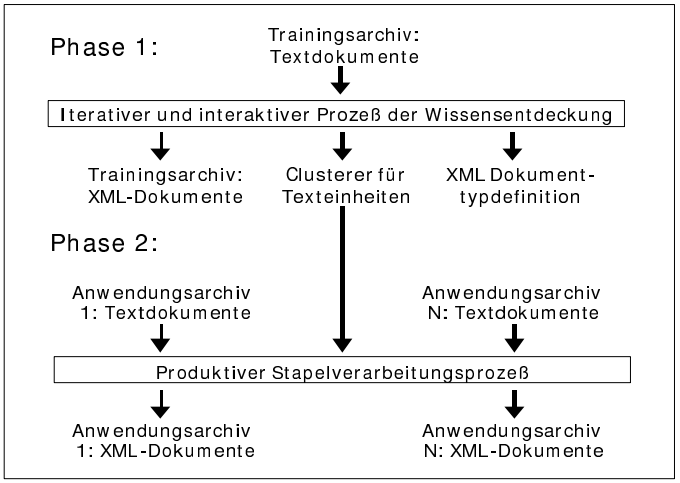

Abbildung 1: Das DIAsDEM-Vorgehensmodell im Überblick

Im Abbildung 1 ist das zweiphasige DIAsDEM-Vorgehensmodell illustriert. Die erste Phase ist ein Prozeß der Wissensentdeckung, der qualitativ hochwertige Segmente von Textelementen entsprechend der o.a. Kriterien entdeckt, Textdokumente mit halbautomatisch ermittelten Segmentbezeichnern bzw. XML-Textmarken annotiert und schließlich eine zunächst unstrukturierte XML DTD für das Archiv ableitet. Dieser 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 bzw. ein Wissensingenieur am Ende jeder Iteration zwecks endgültiger Festlegung der Bezeichner für qualitativ hochwertige 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 „Clusterer für Texteinheiten“), deren Bezeichner als XML-Textmarken Elemente der abgeleiteten XML DTD sind.

Die zweite Phase des DIAsDEM-Vorgehensmodells verwendet den in der ersten Phase 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, denn der vorgeschlagene Prozeß der Wissensentdeckung nutzt Clustering-Algorithmen für die Entdeckung von qualitativ hochwertigen Textelementsegmenten. Die Anwendung der explorativen Datenanalyse mit Clustering-Algorithmen ist jedoch nur dann angemessen, wenn die Datenbasis zumindest eine Cluster-Tendenz aufweist [JMF99, S. 267]. Textelemente, die einer gemeinsamen Anwendungsdomäne entstammen, werden das Kriterium der Cluster-Tendenz eher erfüllen, als Texteinheiten 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 DIAsDEM-Vorgehensmodell für die semantische Auszeichnung einer Vielzahl möglicher Dokumentsammlungen in unterschiedlichsten Bereichen anwendbar: Beispielhaft sind Veröffentlichungen von administrativen Behörden und Gerichten, Geschäftsberichte wie Anhänge zu Jahresabschlüssen und Lageberichte börsennotierter Unternehmen, Ad-hoc-Mitteilungen und Unternehmensnachrichten, textuelle Patientenakten sowie auf elektronischen Marktplätzen veröffentlichte Produkt- und Dienstleistungsbeschreibungen zu nennen.

4 Semantische Auszeichnung als Prozeß der Wissensentdeckung in Texten

Die erste Phase des DIAsDEM-Vorgehensmodells umfaßt einen komplexen Prozeß der Wissensentdeckung in Texten, der in diesem Abschnitt detailliert vorgestellt wird und in Abbildung 2 zusammengefaßt dargestellt ist.

Das Trainingsarchiv fachspezifischer Textdokumente darf als Ausgangsbasis für den Prozeß der Wissensentdeckung nur reine oder strukturell annotierte Texte enthalten. Im letzteren Fall muß eine eindeutig definierte Textauszeichnungsspra-

che (z.B. SGML) verwendet werden, um strukturelle Textkomponenten wie etwa Abschnitte, Überschriften, Absätze und Sätze innerhalb der Dokumente zu kennzeichnen. Multimedia-Dokumente enthalten neben Texten vielfach Bilder, Grafiken oder auch Musik- und Videosequenzen. In diesen Fällen müssen die hier relevanten Textabschnitte zuvor mittels individuell implementierter oder semiautomatischer Werkzeuge wie z.B. NoDoSE [Adel98] aus den proprietären Dateiformaten extrahiert werden.

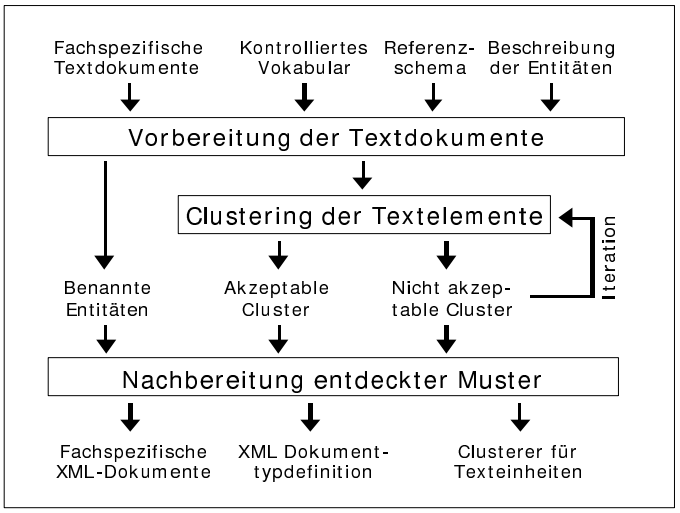

Abbildung 2: Der iterative und interaktive Prozeß der Wissensentdeckung im Überblick

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

Analog zu einem konventionellen Prozeß der Wissensentdeckung in Datenbanken [Mann97, S. 42] beginnt der hier dargestellte Prozeß mit einer Aufbereitungsphase, in dem ein reduzierter Merkmalsraum erzeugt wird. Sämtliche Textelemente werden anschließend in Vektoren des Merkmalsraums überführt. Zusätzlich

werden benannte Entitäten aus Textelementen extrahiert, damit diese später als Attribute den XML-Textmarken zugeordnet werden können. In der Phase der Musterentdeckung werden Textelementvektoren nach semantischer Ähnlichkeit segmentiert. Das Ziel ist dabei die Entdeckung von dicht besetzten und homogenen Segmenten. Dazu wird der Clustering-Algorithmnus mehrfach iterativ ausgeführt. Jede Iteration erzeugt eine Segmentmenge, die entsprechend der DIAsDEM-Qualitätskriterien in „akzeptable“ und „nicht akzeptable“ Cluster aufgeteilt wird. In der Nachbereitungsphase werden akzeptable Segmente halbautomatisch benannt, wobei die Standardbezeichner von den in einem Cluster dominierenden Dimensionen des Merkmalsraums abgeleitet werden. Der Bezeichner eines akzeptablen Segments dient gleichzeitig als XML-Textmarke für alle darin enthaltenen Textelemente. Abschließend werden sämtliche Dokumente mit gültigen XML-Textmarken annotiert und es wird eine vorläufige DTD abgeleitet.

4.1 Aufbereitung der Textdokumente

Innerhalb der 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 diese 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* aus den Textelementen extrahiert und die grammatischen Grundformen für sämtliche 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 einerseits die Dimensionalität der Vektoren zu senken und andererseits 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 und deswegen im Rahmen der Vorverarbeitung extrahiert und in der Nachbereitungsphase den XML-Textmarken als Attribute wieder zugeordnet werden. Bedingung (ii) schließt zusätzlich alle allgemeinsprachlichen Begriffe aus, die für die abzuleitende XML DTD nur sehr begrenzt nutzbar sind. Diese Bedingung kann letztend-

lich 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 dienen als Grundlage des fachspezifischen Merkmalsraums. Am Ende der Aufbereitungsphase wird der Merkmalsraum durch die Auswahl sog. Textelementdeskriptoren oder kurz Deskriptoren festgelegt. Diese Deskriptoren werden von Experten bestimmt und referenzieren 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 verwendete Vektorraum-Modell zur Repräsentation von Texten wurde im Rahmen des Information Retrieval-Projekts Smart entwickelt [SaBu88, SAB94]. Im Gegensatz zu Smart werden aber innerhalb des DIAsDEM-Vorgehensmodells gegenwärtig noch keine Gewichtungen einzelner Deskriptoren vorgenommen.

4.2 Iteratives Clustering der Textelemente

Der 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 Cluster wird dabei durch die in diesem Segment dominierenden Deskriptoren (d.h. die Dimensionen des Merkmalsraums) reflektiert. Die dominierenden Deskriptoren werden in der letzten Phase verwendet, um Bezeichner für Cluster abzuleiten und diese für die semantische Auszeichnung der jeweiligen Textelemente zu verwenden.

Cluster-Analyse wird häufig als die Kunst bezeichnet, Gruppen innerhalb von Daten zu finden [KaRo90, S. 1]. Jain et al. definieren Clustering informell als die auf Ähnlichkeit basierende, unüberwachte Klassifikation von Objekten in Gruppen [JMF99, 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* gestattet.

Gegenwärtig wird für die Gruppierung der Textelementvektoren der *Demographic Clustering Algorithm* des *IBM DB2 Intelligent Miner for Data* [IBM01] verwendet, der das Condorcet-Kriterium maximiert [Mich97]. Dieses Kriterium kann informell als die Differenz zwischen Intrasegment- und Intersegment-Ähnlichkeit betrachtet werden. Der Wert des Condorcet-Kriteriums ist die Differenz zwischen der Summe aller paarweisen Ähnlichkeiten von Vektoren in gleichen Segmenten

und der Summe aller paarweisen Ähnlichkeiten zwischen Vektoren in unterschiedlichen Segmenten. Parameter des *Demographic Clustering Algorithm* sind die maximale Anzahl von Segmenten sowie ein Ähnlichkeitsschwellenwert, der die Zuordnung von Vektoren in gleiche Segmente steuert und hier als „Schwellenwert für die Intrasegment-Ähnlichkeit" bezeichnet wird.

Wie in Abbildung 2 angedeutet, wird der jeweils gewählte Clustering-Algorithmus innerhalb des DIAsDEM-Vorgehensmodells iterativ ausgeführt. Alle innerhalb einer Iteration entdeckten Segmente werden anhand der unten beschriebenen DIAsDEM-Qualitätskriterien evaluiert. Für sämtliche Cluster, die entsprechend dieser Kriterien qualitativ hochwertig bzw. akzeptabel sind, wird ein semantischer Bezeichner halbautomatisch abgeleitet. Das Vorgehen wird in Abschnitt 4.3 beschrieben. 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 wird der Schwellenwert für die Intrasegment-Ähnlichkeit schrittweise gesenkt, so daß akzeptable Cluster inhaltlich allmählich weniger spezifisch werden. 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 mehr 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 innerhalb des DIAsDEM-Vorgehensmodells 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 wird. 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 von qualitativ akzeptablen Segmenten. Die *DIAsDEM Workbench* annotiert alle akzeptablen Cluster mit Statistiken des Clustering-Algorithmus und mit verbalen Cluster-Beschreibungen, 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 den fachkundigen Experten festgelegt.

Der Experte wird bei dieser Tätigkeit durch die automatisch generierten 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 Cluster 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 noch Teil 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 Folge 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. Die weitere Strukturierung der vorläufigen XML DTD ist Teil der zukünftigen 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 verschiedenen Clustering-Algorithmen gekoppelt werden kann. Die *DIAsDEM Workbench* unterstützt sämtliche Phasen des DIAsDEM-Vorgehensmodells. In dieser Fallstudie, die vollständig in [WiSp01c] vorgestellt ist, wurde das hier beschriebene DIAsDEM-Vorgehensmodell angewendet, um ein Archiv deutscher Handelsregistereinträge semantisch zu annotieren.

In Deutschland führt jedes Amtsgericht ein öffentlich zugängliches Handelsregister, das wirtschaftlich relevante Informationen über die Unternehmen im Einzugsbereich des jeweiligen Gerichts enthält. Entsprechend des deutschen Handelsrechts müssen (mit wenigen Ausnahmen) alle Unternehmen gesetzlich geregelte Vorgänge wie z.B. Gründung, Veränderung des Stammkapitals einer GmbH, Gründung einer Zweigniederlassung, Erteilung der Prokura oder Erlöschung der Firma zur Eintragung in das zuständige Handelsregister anmelden. Die Kenntnis der Eintragungen im Handelsregister sind für gegenwärtige und potentielle Geschäftspartner von besonderem Interesse, da diese Eintragungen weitreichende juristische Konsequenzen haben. Aufgrund des regen Interesses der Wirtschaft an den Handelsregistereintragungen gibt es bereits Informationsdienstleister, die Auskünfte online oder offline anbieten. Gegenwärtig unterstützen die Anbieter jedoch nur SQL-Anfragen an strukturierte Handelsregisterdaten und einfache Volltextrecherchen in den Eintragungstexten. Vor diesem Hintergrund wurde die Domäne „Handelsregistereinträge“ ausgewählt, um das DIAsDEM-Vorgehensmodell in einer ersten Fallstudie zu evaluieren.

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 02. Dezember 1994 abgeschlossen. Durch Beschluss der Gesellschafterversammlung vom 07. 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 3: Exemplarischer Handelsregistereintrag des Amtsgerichts Potsdam

In Abbildung 3 ist beispielhaft ein Handelsregistereintrag dargestellt, in dem die Gründung einer Gesellschaft mit beschränkter Haftung angezeigt wird. Jeder Eintrag besteht aus einem strukturierten Teil und einem unstrukturierten Text. Ersterer enthält neben der Firma, der Adresse und der Handelsregisternummer als Identifikator vor allem die rechtlich relevanten Daten der Eintragung und Veröffentlichung. Diese Informationen können problemlos mit entsprechenden Werkzeugen extrahiert, in relationalen DBMS gespeichert und anschließend mit SQL abgefragt werden. Der unstrukturierte Textabschnitt enthält die wesentlichen Informationen in Form des von Justizangestellten erfaßten Texts. Im Rahmen dieser Fallstudie wurde ein Archiv mit 1.145 Handelsregistereintragungen des Amtsgerichts Potsdam semantisch ausgezeichnet. Das sind sämtliche Potsdamer Eintragungen des Jahres 1999, die Unternehmensneugründungen anzeigen.

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

<CommercialRegisterEntry> <BusinessPurpose> Vertrieb und Entwicklung von
Gleisoberbautechnik. </BusinessPurpose> <ShareCapital AmoutOfMoney="50000 DM">
Stammkapital: 50.000 DM. </ShareCapital> <LimitedLiabilityCompany> Gesellschaft
mit beschränkter Haftung. </LimitedLiabilityCompany> <ConclusionArticles Date=
"02.12.1994"> Der Gesellschaftsvertrag ist am 02. Dezember 1994 abgeschlossen.
<ConclusionArticles> <ModificationArticles_MainOffice Date="07.04.1999"
Paragraph="§ 1 (Sitz)"> Durch Beschluss der Gesellschafterversammlung vom 07.
April 1999 ist der Sitz der Gesellschaft von Berlin nach Zernsdorf verlegt und der
Gesellschaftsvertrag geändert in § 1 (Sitz). </ModificationArticles_MainOffice>
(...) Einzelvertretungsbefugnis kann erteilt werden. <AppointmentManagingDirector
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,
Braunschweig, sind zu Geschäftsführern bestellt. </AppointmentManagingDirector>
(...) <PublicationMedia> Nicht eingetragen: Die Bekanntmachungen der Gesellschaft
erfolgen im Bundesanzeiger. </PublicationMedia> </CommercialRegisterEntry>
```

Abbildung 4: Semantisch annotierte Handelsregistereintragung als XML-Dokument

Die Details dieser Fallstudie, der Leser möge sie bitte [WiSp01c] 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 insgesamt 1.145 Handelsregistereintragungen wurden in ihre 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* [Schm94] wurde danach eingesetzt, um durch Ermittlung der grammatikalischen Grundformen sämtlicher Wörter 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. Nach drei Iterationen wurden insgesamt 73 qualitativ akzeptable Segmente entdeckt und halbautomatisch semantisch benannt. Diese 73 akzeptablen Cluster enthielten etwa 85 Prozent aller Texteinheiten. Abbildung 4 zeigt einen Auszug des semantisch annotierten Handelsregistereintrags aus Abbildung 3, der nach Auszeichnung des Archiv von der *DIAsDEM Workbench* erzeugt wurde. Abbildung 5 enthält einen Auszug der ebenfalls abgeleiteten, unstrukturierten XML-Dokumenttypdefinition. Diese vorläufige XML DTD beschreibt grob die semantische Struktur des erzeugten Archivs semantisch annotierter Handelsregistereintragungen.

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

<!ELEMENT CommercialRegisterEntry ( #PCDATA | BusinessPurpose |
ResolutionByShareholders_ShareCapital | ShareCapital | ModificationMainOffice |
FullyLiablePartner | AppointmentManagingDirector | GeneralPartnership |
InitialShareholders | NonCashCapitalContribution | LimitedLiabilityCompany |
ConclusionArticles | DivisionCapitalStock | (...) | FoundationPartnership )* >

<!ELEMENT BusinessPurpose (#PCDATA)>
<!ELEMENT ShareCapital (#PCDATA)> (...)
<!ELEMENT FoundationPartnership (#PCDATA)>

<!ATTLIST ShareCapital AmountOfMoney CDATA #IMPLIED> (...)
<!ATTLIST AppointmentManagingDirector Person CDATA #IMPLIED> (...)
<!ATTLIST ConclusionArticles Date CDATA #IMPLIED> (...)
```

Abbildung 5: Abgeleitete, unstrukturierte XML-Dokumenttypdefinition (Auszug)

Im Gegensatz zur automatischen Textklassifikation [Seba99] gibt es in dieser Domäne keine vorklassifizierten Trainingsdokumente, mit deren Hilfe die Genauigkeit der erfolgten semantischen Auszeichnung überprüft werden kann. Aus diesem Grund wurde eine Zufallsstichprobe gezogen, die 5 Prozent aller Textelemente enthielt, und ein Experte 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.

Innerhalb der gezogenen Zufallsstichprobe betrug die Fehlerrate für den Fehlertyp I (Fehlertyp II) 0,375 Prozent (3,565 Prozent). Die Fehlerrate für den ersten Fehlertyp ist sehr niedrig, d.h. ein semantisch ausgezeichnetes Textelement ist mit hoher Wahrscheinlichkeit richtig annotiert. Die Fehlerrate für den zweiten Fehlertyp ist hingegen bedeutend höher. Diese Fehler deuten an, daß diese Textelemente nicht den Segmenten zugeordnet wurden, zu denen sie semantisch gehören. Eine

vorläufige Erklärung für die höhere Fehlerrate des Fehlertyps II ist, das die fälschlicherweise nicht ausgezeichneten Textelemente spezielle oder weniger häufige Begriffe enthalten, die nicht Teil des verwendeten Thesaurus sind. Eine Erweiterung des Thesaurus z.B. durch die Aufnahme weiterer Synonyme könnte somit eventuell zu einer Vermeidung von Fehlern des Typs II beitragen.

Die Gesamtfehlerrate innerhalb der gezogenen Stichprobe betrug 3,940 Prozent. Auf einem Konfidenzniveau von 0,95 liegt die Gesamtfehlerrate in dem Intervall [2,591 Prozent; 5,948 Prozent]. Das ist ein vielversprechendes Ergebnis für die 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, die zuvor extrahierte benannte Entitäten enthalten. Das vorgestellte Vorgehensmodell wurde in einer Fallstudie zu einem Archiv des deutschen Handelsregisters erfolgreich evaluiert.

Sowohl der hohe Anteil semantisch ausgezeichneter Sätze als auch die niedrigen Fehlerraten im Rahmen der ersten 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 der Qualität der semantischen Auszeichnung sind jedoch ebenfalls vielversprechend.

Offene Forschungsaspekte sind derzeit die Bewertung klassischer Clustering-Algorithmen und Ähnlichkeitsmetriken im Hinblick auf die Verfahrensziele, die automatisierte Erstellung des Thesaurus, die halbautomatische Auswahl der Vektordimensionen sowie die Ableitung strukturierter Dokumenttypdefinitionen oder XML-Schemata. Diese Schemata sollen (ggf. probabilistische) Informationen zu Reihenfolge und Verschachtelung der XML-Textmarken enthalten. Erste Ansätze zur Strukturierung der vorläufigen XML DTD werden in [WiSp01a,

WiSp01b] 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.

Danksagungen

Der Deutschen Forschungsgemeinschaft möchte der Autor für die Finanzierung des Forschungsprojekts DIAsDEM und der Bundesanzeiger Verlagsgesellschaft mbH für die Bereitstellung von Dokumenten danken. Der *IBM Intelligent Miner for Data* wurde freundlicherweise im Rahmen des *IBM DB2 Scholars Program* kostenfrei für Lehr- und Forschungszwecke bereitgestellt.

Literatur

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

[Adel98] Adelberg, B.: 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, June 1998, S. 283-294.

[BDB+00] Bruder, I.; Düsterhöft, A.; Becker, M.; Bedersdorfer, J.; Neumann, G.: 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, June 2000, S. 227-238.

[Bune97] Buneman, P.: Semistuctured Data. In: Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Tucson, AZ, USA, May 1997, S. 117-121.

[DEFS99] Decker, S.; Erdmann, M.; Fensel, D.; Studer, R.: ONTOBROKER: Ontology Based Access to Distributed and Semi-Structured Information. In: Meersman, R. et al. (Hrsg.): Database Semantics: Semantic Issues in Multimedia System, Kluwer Academic Publisher, Boston 1999, S. 351-369.

[ECSL98] Embley, D. W.; Cambell, D. M.; Smith, R.D.; Liddle, S. W.: Ontology-Based Extraction and Structuring of Information from Data-Rich Unstructured Documents. In:

Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management, Bethesda, MD, USA, 1998, S. 52-59.

[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.

[FFK+98] Feldman, R.; Fresko, M.; Kinar, Y.; Lindell, Y.; Liphstat, O.; Rajman, M.; Schler, Y.; Zamir, O.: Text Mining at the Term Level. In: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery, Nantes, France, September 1998, S. 65-73.

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

[GWS01] Graubitz, H.; Winkler, K.; Spiliopoulou, M.: Semantic Tagging of Domain-Specific Text Documents with DIAsDEM. In: Proceedings of the 1st International Workshop on Databases, Documents and Information Fusion (DBFusion 2001), Magdeburg, Germany, May 2001, S. 61-72.

[IBM01] IBM Corporation: IBM DB2 Intelligent Miner for Data. http://www.ibm.com/software/data/iminer, Abruf am 2001-08-31.

[JMF99] Jain, A. K.; Murty, M. N.; Flynn, P. J.: Data Clustering: A Review. In: ACM Computing Surveys 31 (1999) 3, S. 264-323.

[KaRo90] Kaufman, L.; Rousseeuw, P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, New York 1990.

[LMP00] Laur, P. A.; Masseglia, F.; Poncelet, P.: Schema Mining: Finding Regularity among Semistructured Data. In: Proceedings of the 4th European Conference on the Principles of Data Mining and Knowledge Discovery, Lyon, France, September 2000, S. 498-503.

[LWO00] Loh, S.; Wives, L. K.; Oliveira, J. P. M. d.: Concept-Based Knowledge Discovery in Texts Extracted from the Web. In: ACM SIGKDD Explorations 2 (2000) 1, S. 29-39.

[Lume00] Lumera, J.: Große Mengen an Altlastdaten stehen XML-Umstieg im Weg. In: Computerwoche 27 (2000) 16, S. 52-53.

[Mann97] Mannila, H.: Methods and Problems in Data Mining. In: Proceedings of the International Conference on Database Theory, Delphi, Greece, January 1997, S. 41-55.

[MaSt01] Maedche, A.; Staab, S.: Learning Ontologies for the Semantic Web. In: IEEE Intelligent Systems 16 (2001) 2, Special Issue on the Semantic Web, S. 72-79.

[Mich97] Michaud, P.: Clustering Techniques. In: Future Generation Computer Systems 13 (1997) 2-3, S. 135-147.

[MiFi95] Mikheev, A.; Finch, S.: A Workbench for Acquisition of Ontological Knowledge from Natural Language. In: Proceedings of the 7th Conference of the European Chapter for Computational Linguistics, Dublin, Ireland, March 1995, S. 194-201.

[MoBe01] Moore, G. W.; Berman, J. J.: Anatomic Pathology Data Mining. In: Cios, K. J. (Hrsg.): Medical Data Mining and Knowledge Discovery. Physica-Verlag, Heidelberg New York 2001. S. 72-117.

[NaMo00] Nahm, U. Y.; Mooney, R. J.: Using Information Extraction to Aid the Discovery of Prediction Rules from Text. In: Proceedings of the KDD-2000 Workshop on Text Mining, Boston, MA, USA, August 2000, S. 51-58.

[NAM97] Nestrov, S.; Abiteboul, S.; Motwani, R.: Inferring Structure in Semi-Structured-Data. SIGMOD Record 26 (1997) 4, S. 39-43.

[SaBu88] Salton, G.; Buckley, C.: Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management 24 (1988) 5, S. 513-523.

[SAB94] Salton, G.; Allan, J.; Buckley, C.: Automatic Structuring and Retrieval of Large Text Files. Communications of the ACM 37 (1994) 2, S. 97-108.

[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.

[Seba99] Sebastiani, F.: A Tutorial on Automated Text Categorization. In: Proceedings of ASAI-99, 1st Argentinian Symposium on Artificial Intelligence, Buenos Aires, Argentina, 1999, S. 7-35.

[SePu00] Sengupta, A.; Purao, S.: Transitioning Existing Content: Inferring Organization-Specific Document Structures. In: Tagungsband der 1. Deutschen Tagung XML 2000: XML Meets Business, Heidelberg, Germany, May 2000, S. 130-135.

[Tan99] Tan, A.-H.: 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, April 1999, S. 65-70.

[TCR98] Turmo, J.; Català, N.; Rodriguez, H.: TURBIO: A System for Extracting Information from Restricted-Domain Texts. In: Lecture Notes in Artificial Intelligence, Volume 1415. Springer-Verlag, 1998, S. 708-721.

[WaLi00] Wang, K.; Liu, H.: Discovering Structural Association of Semistruktured Data. IEEE Transactions on Knowledge and Data Engineering 12 (2000) 3, S. 353-371.

[WiSp01a] Winkler, K.; Spiliopoulou, M.: Integrating Data and Probabilistically Structured Text Documents. Akzeptiert für Publikation in: Tagungsband des 5. Workshops „Föderierte Datenbanken“ FDBS 2001, Berlin, Deutschland, Oktober 2001.

[WiSp01b] Winkler, K.; Spiliopoulou, M.: Extraction of Semantic XML DTDs from Texts Using Data Mining Techniques. Submitted for Publication.

[WiSp01c] Winkler, K.; Spiliopoulou, M.: Semi-Automated XML Tagging of Public Text Archives: A Case Study. Submitted for Publication.