



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

**6. Konferenz der SAS Anwender in
Forschung und Entwicklung (KSFE 2002)**

**Myra Spiliopoulou und Karsten Winkler
Handelshochschule Leipzig**

- 1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse**
- 2. DIAsDEM-Vorgehensmodell im Überblick**
- 3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung**
- 4. Fallstudie I: Handelsregistereinträge**
- 5. Fallstudie II: Ad-hoc-Mitteilungen**
- 6. Ihre Fragen und Diskussion**

- 1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse**
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion



Wettbewerberanalyse

- **Engl.: Competitive Intelligence (CI)**
- **"Competitive analysis is a systematic program for gathering and analyzing information about your competitors' activities and general business trends to further your own company's goals" (Kahaner)**
- **Ziele der Wettbewerberanalyse:**
 - **Antizipation von Wettbewerberaktivitäten oder relevanten Marktveränderungen**
 - **Sammlung von Wissen über potentiell relevante Technologien, Produkte und Vorschriften**



Wettbewerberanalyse: Datenbasis

Patente

Ad-hoc-
Mitteilungen

Aktienkurse

Meinungs-
portale

Presse-
mitteilungen

Handels-
register

Branchen-
verbände

Jahresab-
schluß (XBRL)

Produkt-
kataloge

(...)

Bilanz-
kennzahlen

Testberichte
über Produkte

Website

Wettbewerberanalyse: Datenbasis (2)

Strukturierte Daten

Aktienkurse

Bilanz-
kennzahlen

Branchen-
verbände

Testberichte
über Produkte

Unstrukturierte Daten

Patente

Handels-
register

Semistrukturierte Daten

Website

Produkt-
kataloge

Jahresab-
schluß (XBRL)

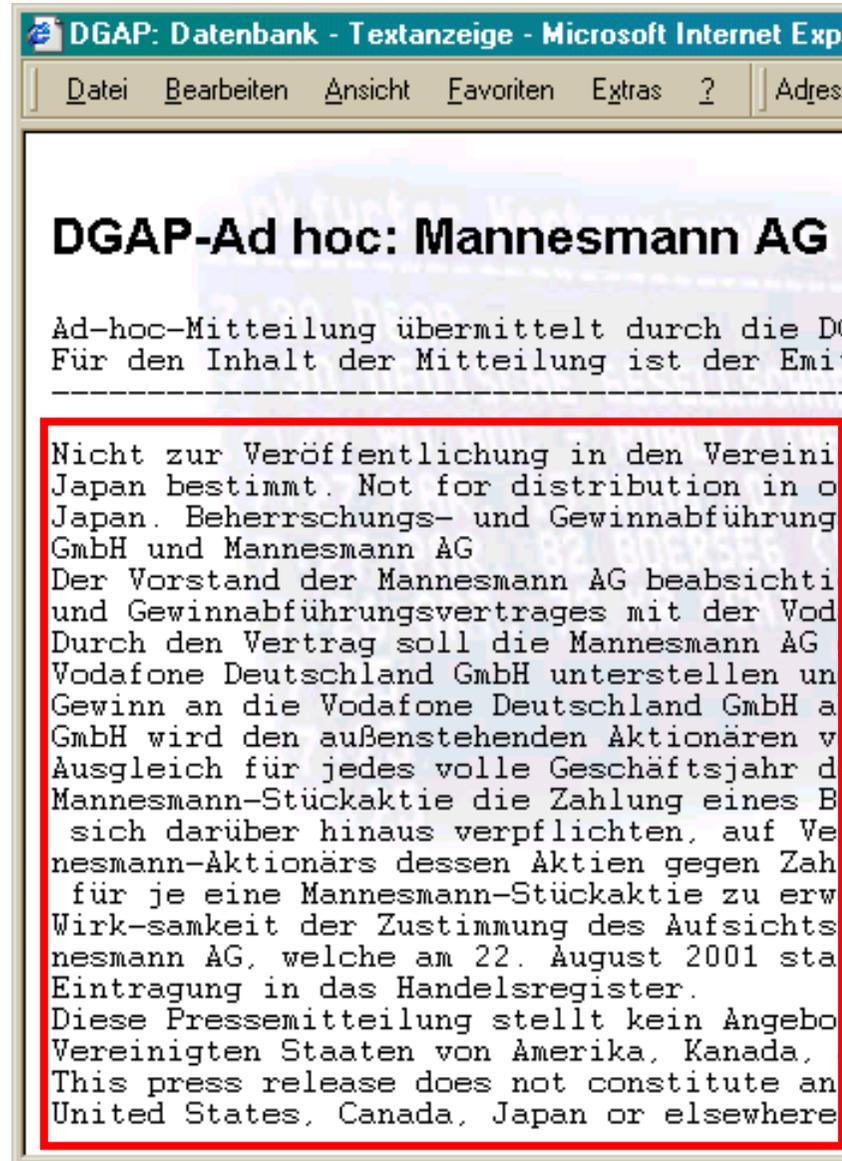
Presse-
mitteilungen

(...)

Ad-hoc-
Mitteilungen

Meinungs-
portale

Text Mining: Fokus Dokument



DGAP: Datenbank - Textanzeige - Microsoft Internet Exp

Datei Bearbeiten Ansicht Favoriten Extras ? Adres

DGAP-Ad hoc: Mannesmann AG

Ad-hoc-Mitteilung übermittelt durch die DGAP
Für den Inhalt der Mitteilung ist der Emittent verantwortlich

Nicht zur Veröffentlichung in den Vereinigten Staaten von Amerika.
Japan bestimmt. Not for distribution in the United States or elsewhere.
Japan. Beherrschungs- und Gewinnabführungsvertrag zwischen Vodafone
GmbH und Mannesmann AG
Der Vorstand der Mannesmann AG beabsichtigt, einen Beherrschungs- und Gewinnabführungsvertrag mit der Vodafone Deutschland GmbH abzuschließen.
Durch den Vertrag soll die Mannesmann AG der Vodafone Deutschland GmbH unterstellt und der Gewinn an die Vodafone Deutschland GmbH abgeführt werden.
Die Vodafone Deutschland GmbH wird den außenstehenden Aktionären von Mannesmann AG für jedes volle Geschäftsjahr den Ausgleich für Mannesmann-Stückaktie die Zahlung eines Barbetrags vorzunehmen.
Die Mannesmann-Aktionäre sind verpflichtet, auf Verlangen der Vodafone Deutschland GmbH für je eine Mannesmann-Stückaktie zu erwirken.
Die Wirksamkeit der Zustimmung des Aufsichtsrates der Mannesmann AG, welche am 22. August 2001 stattgefunden hat, ist in das Handelsregister eingetragen.
Diese Pressemitteilung stellt kein Angebot dar.
This press release does not constitute an offer in the United States, Canada, Japan or elsewhere.

- Clustering von Dokumenten nach Inhalt
- Klassifizierung von Dokumenten
- Entdeckung von Themen in Dokumenten
- Zusammenfassung von Dokumenten

Text Mining: Fokus Terme

DGAP: Datenbank - Textanzeige - Microsoft Internet Exp

Datei Bearbeiten Ansicht Favoriten Extras ? Adres

DGAP-Ad hoc: Mannesmann AG

Ad-hoc-Mitteilung übermittelt durch die DGAP
Für den Inhalt der Mitteilung ist der Emittent verantwortlich

Nicht zur Veröffentlichung in den Vereinigten Staaten von Amerika bestimmt. Not for distribution in the United States of America. Japan. Beherrschungs- und Gewinnabführungsvereinbarung zwischen Mannesmann AG und Vodafone Deutschland GmbH. Der Vorstand der Mannesmann AG beabsichtigt, einen Beherrschungs- und Gewinnabführungsvertrag mit der Vodafone Deutschland GmbH zu schließen. Durch den Vertrag soll die Mannesmann AG die Vodafone Deutschland GmbH unterstützen und Gewinne an die Vodafone Deutschland GmbH abführen. Die Vodafone Deutschland GmbH wird den außenstehenden Aktionären von Mannesmann AG einen Ausgleich für jedes volle Geschäftsjahr der Mannesmann AG leisten. Mannesmann-Aktionäre sind verpflichtet, auf Verlangen der Mannesmann AG, die Zahlung eines Betrags für je eine Mannesmann-Stückaktie zu erwirken. Die Mannesmann AG, welche am 22. August 2001 die Eintragung in das Handelsregister beantragt hat, stellt kein Angebot in den Vereinigten Staaten von Amerika, Kanada, Japan oder anderswo dar. This press release does not constitute an offer in the United States, Canada, Japan or elsewhere.

- Extraktion wichtiger benannter Entitäten
- Entdeckung von Ontologien oder Thesauri
- Computerlinguistik: Grammatische Auszeichnung von Termen

Text Mining: Fokus Textelemente (2)

DGAP: Datenbank - Textanzeige - Microsoft Internet Exp

Datei Bearbeiten Ansicht Favoriten Extras ? Adres

DGAP-Ad hoc: Mannesmann AG

Ad-hoc-Mitteilung übermittelt durch die DGAP
Für den Inhalt der Mitteilung ist der Emittent verantwortlich

Nicht zur Veröffentlichung in den Vereinigten Staaten
Japan bestimmt. Not for distribution in the United States
Japan. Beherrschungs- und Gewinnabführungsvertrag
GmbH und Mannesmann AG

Der Vorstand der Mannesmann AG beabsichtigt den Abschluß eines Beherrschungs- und Gewinnabführungsvertrages mit der Vodafone Deutschland GmbH, Düsseldorf.

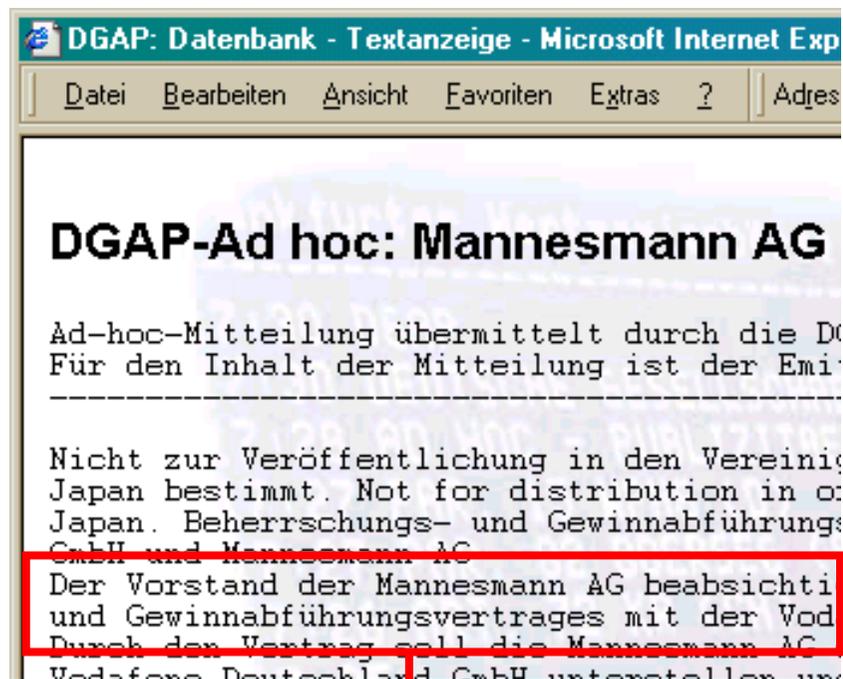
Vodafone Deutschland GmbH unterstellen und den Gewinn an die Vodafone Deutschland GmbH abzuführen. Die Vodafone Deutschland GmbH wird den aufstehenden Aktionären in der Regel eine Dividende ausschütten.

Aus dem Bericht der Mannesmann AG vom 17. März 2000
si
nes
fü
Wir
nes
Ein
Die
Ver
This press release does not constitute an offer in the United States, Canada, Japan or elsewhere

- Entdeckung von Beziehungen zwischen benannten Entitäten (Strukturierung)
- Ableitung einer XML DTD und semantische Textauszeichnung

Der Vorstand der Mannesmann AG beabsichtigt den Abschluß eines Beherrschungs- und Gewinnabführungsvertrages mit der Vodafone Deutschland GmbH, Düsseldorf.

Semantische Textauszeichnung



- Entdeckung von Beziehungen zwischen benannten Entitäten (Strukturierung)
- Ableitung einer XML DTD und semantische Textauszeichnung

**<AbschlußBeherrschungsvertrag Unternehmen="Mannesmann AG && Vodafone Deutschland GmbH">
Der Vorstand der Mannesmann AG beabsichtigt den Abschluß eines Beherrschungs- und Gewinnabführungsvertrages mit der Vodafone Deutschland GmbH, Düsseldorf. </AbschlußBeherrschungsvertrag>**

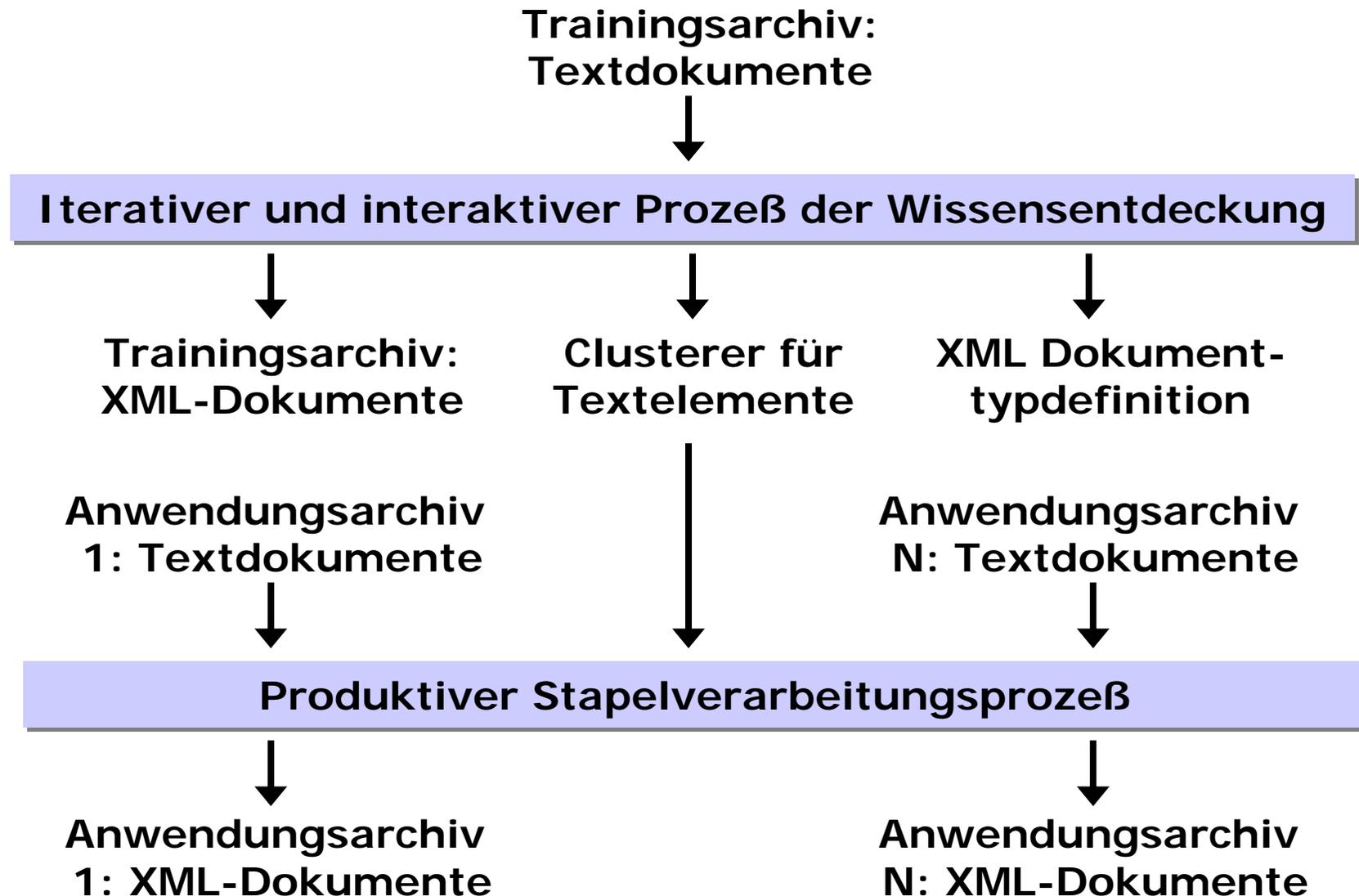
1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
- 2. DIAsDEM-Vorgehensmodell im Überblick**
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion



DIAsDEM-Vorgehensmodell

- **Ziel 1: Semantische Auszeichnung großer, anwendungsspezifischer Textarchive mit der Textauszeichnungssprache XML**
 - Überführung des Textarchivs in ein Archiv semistrukturierter XML-Dokumente
- **Ziel 2: Ableitung einer möglichst strukturierten XML Dokumenttypdefinition für das Archiv**
 - Nutzung einer XML-Anfragesprache für inhalts- und strukturbasierte Suche in XML-Archiven
- **Methodik: Wissensentdeckung in Datenbanken**

DIAsDEM-Vorgehensmodell (2)



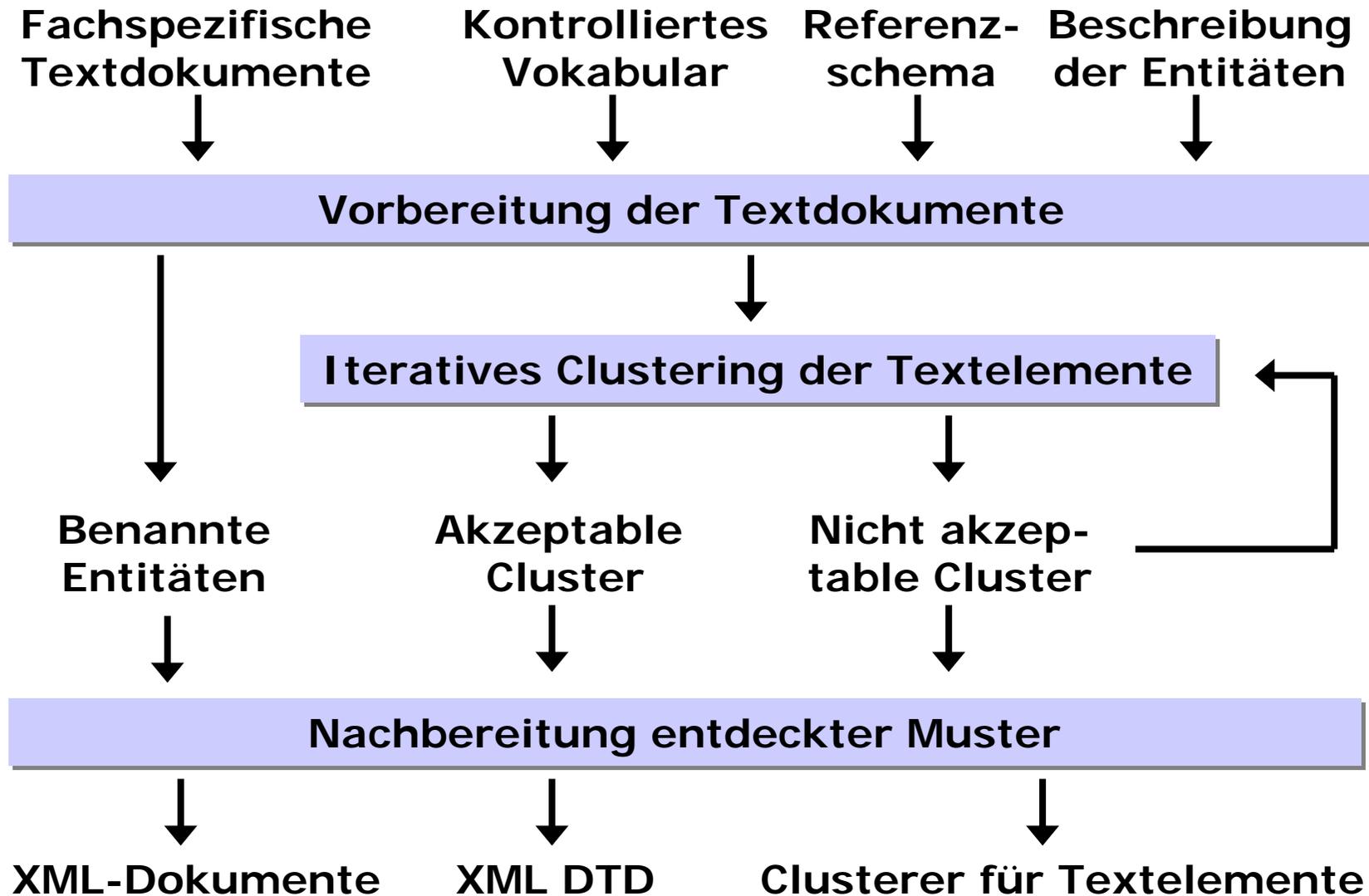
1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
- 3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung**
4. Fallstudie I: Handelsregistereinträge
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion



Prozeß der Wissensentdeckung

- 1. Inhaltsbasierte Gruppierung struktureller Textelemente (z.B. Sätze oder Absätze) mit Clustering-Algorithmus**
- 2. Inhaltliche Charakterisierung der entdeckten Gruppen von Textelementen: Benennung qualitativ hochwertiger Cluster**
- 3. Identifizierung und Benennung benannter Entitäten (z.B. Personen) in Textelementen**
- 4. Zusammenführung von Cluster-Bezeichnern und Bezeichnern benannter Entitäten zu einer XML Dokumenttypdefinition**

Prozeß der Wissensentdeckung (2)



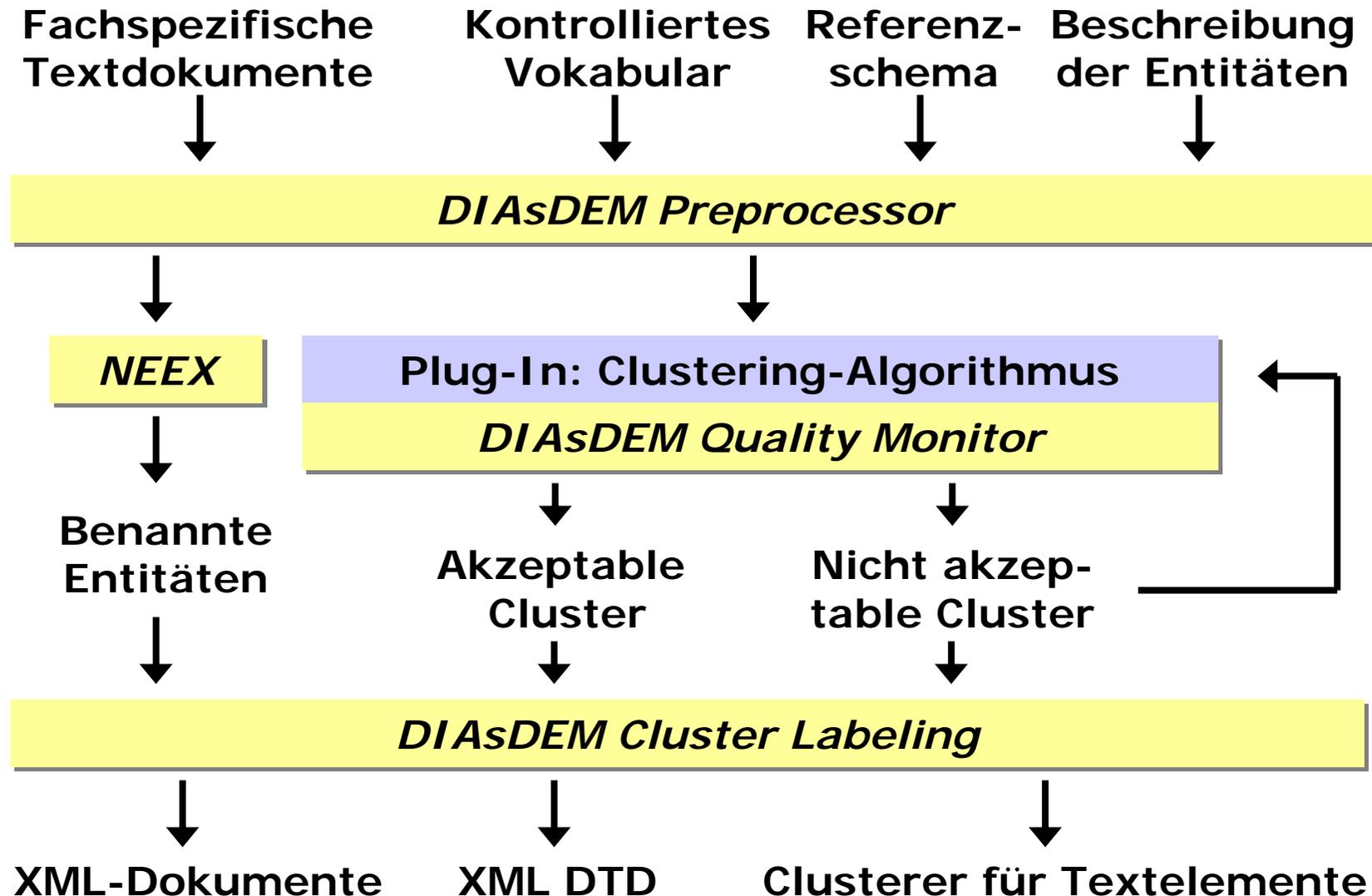


Iteratives Clustering

1. Festlegung der Parameter des Clustering-Algorithmus (Ähnlichkeit?)
2. Gruppierung der Vektoren entsprechend des gewählten Ähnlichkeitsmaßes
3. Evaluation der Qualität entdeckter Cluster mit Qualitätskriterien (Interessanz?)
4. Ausblendung aller Vektoren in "akzeptablen" Clustern für spätere semantische Benennung
5. Wiederholung des Gruppierungsvorgangs mit Vektoren in "nicht akzeptablen" Clustern, sofern die Stop-Bedingung nicht erfüllt ist



Die *DIAsDEM-Workbench*



1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
- 4. Fallstudie I: Handelsregistereinträge**
5. Fallstudie II: Ad-hoc-Mitteilungen
6. Ihre Fragen und Diskussion



Fallstudie I: Handelsregistereintrag

Daniel Spiel-Center GmbH	HRB 12576
Potsdamer Str. 94, 14513 Teltow	06.05.99

Der Betrieb von Spielhallen in Teltow und das Aufstellen von Geldspiel- und Unterhaltungsautomaten. Stammkapital: 25.000 EUR. Gesellschaft mit beschränkter Haftung. Der Gesellschaftsvertrag ist am 12. November 1998 abgeschlossen und am 19. April 1999 abgeändert. (...) Pawel Balski, 14.04.1965, Berlin, ist zum Geschäftsführer bestellt. Er vertritt die Gesellschaft stets einzeln und (...)



Fallstudie I: Vorbereitung

1. Zerlegung der Texte in Textelemente und Terme

(...) 1999 abgeändert | Pawel Balski , 14.04.1965 , Berlin, ist zum Geschäftsführer bestellt . | Er vertritt die (...)

2. Ausblendung identifizierter benannter Entitäten

PERSON , DATE , PLACE , ist zum Geschäftsführer bestellt .

3. Ermittlung der Wortstämme sämtlicher Terme

PERSON , DATE , PLACE , sein zu Geschäftsführer bestellen .

4. Auswahl von Deskriptoren aus dem Thesaurus

(Bestellung, ..., Geschäftsführer, Gründung, Gesellschaft)

5. Abbildung der Textelemente auf Boolesche Vektoren

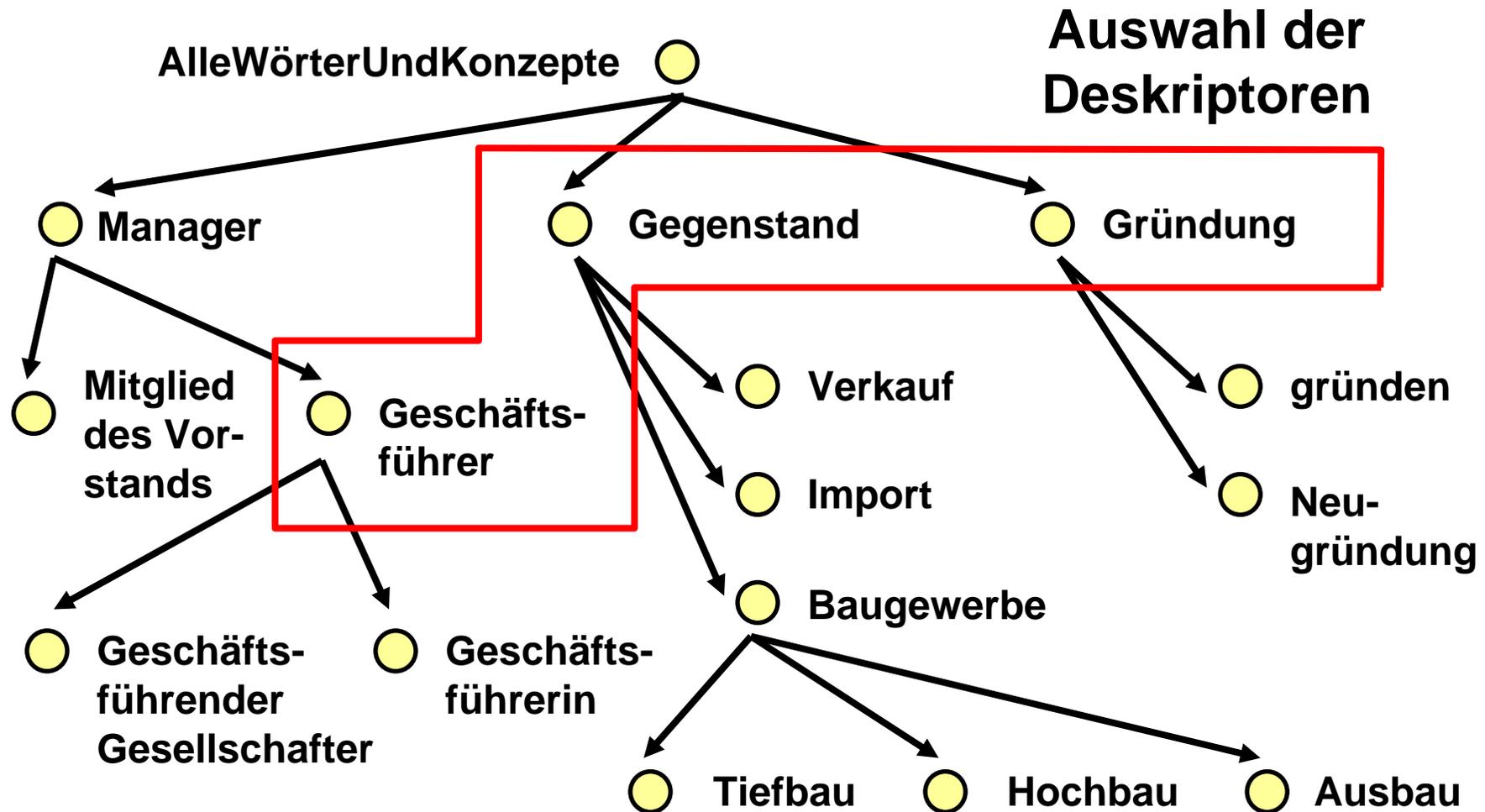
(1, ..., 1, 0, 0)

6. Gewichtung der Textelementvektoren mit TFxIDF

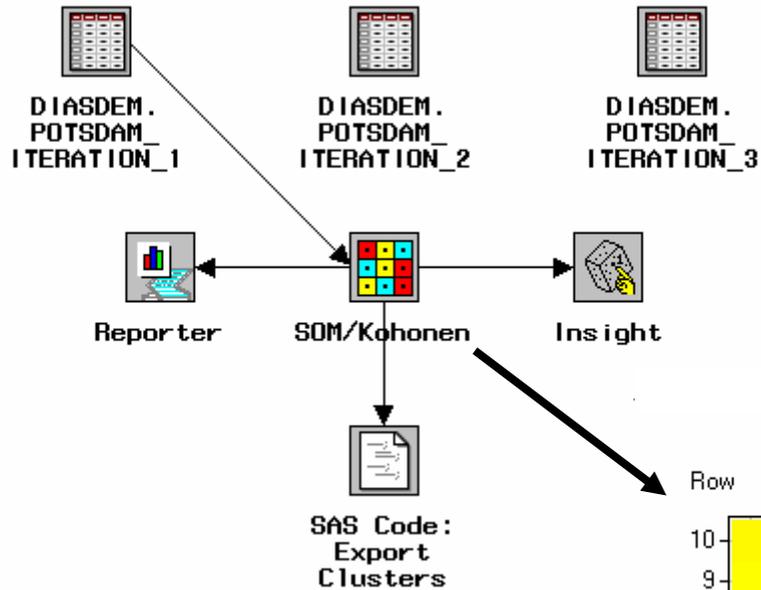
(1.4283, ..., 0.9010, 0, 0)



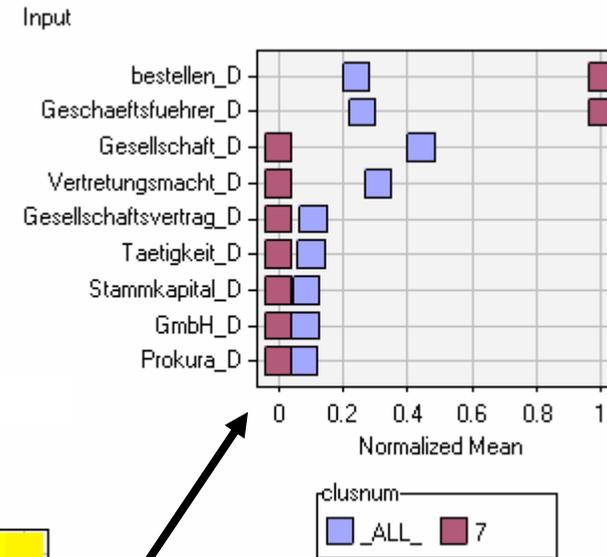
Fallstudie I: Vektorraum



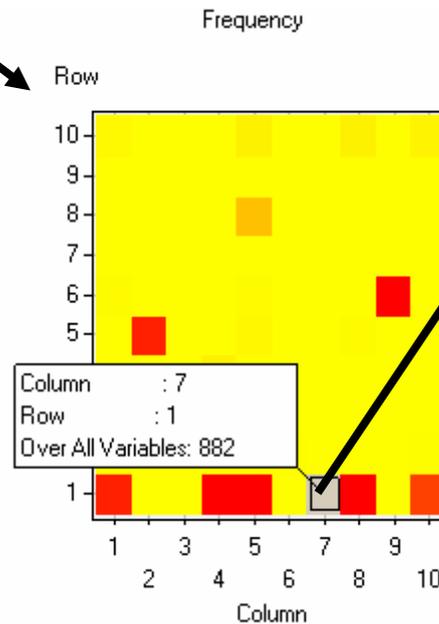
Fallstudie I: Data Mining mit SAS/EM



Data-Mining-Diagramm



Visualisierung des Cluster 7



Ergebnis der 1. Iteration



Fallstudie I: XML-Dokument

Daniel Spiel-Center GmbH
Potsdamer Str. 94, 14513 Teltow

HRB 12576
06.05.99

```
<Handelsregistereintrag> <Gegenstand>  
Der Betrieb von Spielhallen in Teltow  
und das Aufstellen von Geldspiel- und  
Unterhaltungsautomaten. </> (...)  
<GeschäftsführerBestellen Person="Balski,  
Pawel, 14.04.1965, Berlin, null"> Pawel  
Balski, 14.04.1965, Berlin, ist zum  
Geschäftsführer bestellt. </>  
(...) </Handelsregistereintrag>
```



Fallstudie I: XML DTD

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

<!ELEMENT Handelsregistereintrag ( #PCDATA |
Gegenstand | (...) | Stammkapital | Verän-
derungSitz | PersönlichHaftendeGesellschafter
| GeschäftsführerBestellen | AufteilungGrund-
kapital | (...) | GründungGesellschaft )* >

<!ELEMENT Gegenstand (#PCDATA)> (...)
<!ELEMENT Stammkapital (#PCDATA)> (...)
<!ELEMENT GründungGesellschaft (#PCDATA)>

(...) <!ATTLIST GeschäftsführerBestellen
Person CDATA #IMPLIED> (...)
```



Fallstudie I: Zusammenfassung

1.145 Dokumente,
10.785 Textelemente Thesaurus:
85 Deskriptoren UML-
Schema Person, Geld,
Unternehmen, ...

DIAsDEM Preprocessor , Plug-In: *IMS Stuttgart TreeTagger*

NEEX

Plug-In: *SAS / Enterprise Miner*
DIAsDEM Quality Monitor

Benannte
Entitäten

68 akzeptable
Cluster (95%)

Nicht akzep-
table Cluster

3 Iterationen

DIAsDEM Cluster Labeling

1.145 XML-
Dokumente

Unstrukturierte
XML DTD (41 Tags)

Clusterer für Textelemente
weiterer ähnlicher Archive



Fallstudie I: Evaluation

- **Evaluation im Hinblick auf Auszeichnungsfehler:**
 - **Fehlertyp I: Ein XML-Tag reflektiert nicht den genauen Inhalt eines Textelements**
 - **Fehlertyp II: Ein Textelement ist nicht annotiert, obwohl es ein Konzept enthält, das Teil der abgeleiteten XML DTD ist**
- **Manuelle Überprüfung von 5% der Texteinheiten:**
 - **Fehlertyp I: 1,5% in der Stichprobe**
 - **Fehlertyp II: 1,8% in der Stichprobe**
 - **Gesamtfehler mit 0,95-Konfidenz: [2,35%; 4,24%]**

1. Semantische Auszeichnung fachspezifischer Textarchive in der Wettbewerberanalyse
2. DIAsDEM-Vorgehensmodell im Überblick
3. Prozeß der Wissensentdeckung zur semantischen Textauszeichnung
4. Fallstudie I: Handelsregistereinträge
- 5. Fallstudie II: Ad-hoc-Mitteilungen**
6. Ihre Fragen und Diskussion



Fallstudie II: XML-Dokument

```
<AdHocMitteilung>  
<NiederlegungMandatAufsichtsrat Company=  
"WESTGRUND AG" Person="Rurack; Klaus" Date=  
"04.01.2000"> Der Vorstand der WESTGRUND AG teilt  
gem. § 15 WpHG mit, dass das Mitglied des  
Aufsichtsrats der WESTGRUND AG, Herr Klaus  
Rurack (Remscheid), mit Wirkung zum 4.01.2000  
sein Amt als Mitglied des Aufsichtsrats der  
WESTGRUND AG niederlegt. </> <Hauptversammlung  
Company="WESTGRUND AG"> Da die nächste  
ordentliche Hauptversammlung der WESTGRUND AG  
voraussichtlich im Mai 2000 stattfinden wird,  
wird die WESTGRUND AG ein neues Aufsichtsrats-  
mitglied durch Gerichtsbeschluss bestellen  
lassen. </> (...) </AdHocMitteilung >
```



Fallstudie II: XML DTD

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

<!ELEMENT AdHocMitteilung ( #PCDATA |
PositivesQuartal | AuftragseingangImQuartal |
NeuesProdukt | AnstiegUmsatzImQuartal |
EntwicklungVonProdukten | ErwartungenUmsatz |
AkquisitionVonUnternehmen | NeuerKunde |
Grundkapital | AusbauVonBereichen | Planung |
Marktposition | Kapitalerhoehung | Vorstand |
ErfolgAmMarkt | NiederlegungMandatAufsichtsrat
| (...) | NeueTochtergesellschaft )* >

<!ELEMENT PositivesQuartal (#PCDATA)> (...)
<!ELEMENT NeueTochtergesellschaft (#PCDATA)>
```

Fallstudie II: Zusammenfassung

638 Dokumente,
11.539 Textelemente

Thesaurus:
129 Deskriptoren

UML-
Schema

Person, Datum,
Unternehmen, ...

DIAsDEM Preprocessor, Plug-In: *IMS Stuttgart TreeTagger*

NEEX

Plug-In: *SAS / Enterprise Miner*
DIAsDEM Quality Monitor

Benannte
Entitäten

127 akzeptable
Cluster (51%)

Nicht akzep-
table Cluster

3 Iterationen

DIAsDEM Cluster Labeling

638 XML-
Dokumente

Unstrukturierte
XML DTD (114 Tags)

Clusterer für Textelemente
weiterer ähnlicher Archive



Fallstudie II: Evaluation

- **Evaluation im Hinblick auf Auszeichnungsfehler:**
 - **Fehlertyp I: Ein XML-Tag reflektiert nicht den genauen Inhalt eines Textelements**
 - **Fehlertyp II: Ein Textelement ist nicht annotiert, obwohl es ein Konzept enthält, das Teil der abgeleiteten XML DTD ist**
- **Manuelle Überprüfung von 5% der Texteinheiten:**
 - **Fehlertyp I: 10,8% in der Stichprobe**
 - **Fehlertyp II: 6,4% in der Stichprobe**
 - **Gesamtfehler mit 0,95-Konfidenz: [16,29%; 18,11%]**



Zusammenfassung und Ausblick

- ***DIAsDEM Workbench + SAS / Enterprise Miner = Text Mining für die Strukturierung von Archiven***
- **Nächste Schritte im Forschungsprojekt DIAsDEM:**
 - **Evaluierung und Verbesserung des Vorgehensmodells in sprachlich komplexen Domänen**
 - **Zielorientierte Bewertung klassischer Clustering-Algorithmen und Ähnlichkeitsmetriken**
 - **Ableitung einer strukturierteren XML DTD**
 - **Integration von XML-Archiven mit relationalen Daten und Auswahl einer XML-Anfragesprache**



Vielen Dank

an die **DFG**

und den



Fragen



Karsten Winkler
kwinkler@ebusiness.hhl.de
<http://ebusiness.hhl.de>