



# Document Warehousing und Text Mining in der Wettbewerberanalyse

Vorlesung IKS, 30. Januar 2002  
Humboldt-Universität zu Berlin

Karsten Winkler  
Handelshochschule Leipzig

Agenda

1. Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerberanalyse
2. Document Warehousing: Technologie im Überblick
3. Text Mining: Technologie im Überblick
4. Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text
5. Zusammenfassung und Literaturhinweise

**HHL**...

Agenda

1. Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerberanalyse
2. Document Warehousing: Technologie im Überblick
3. Text Mining: Technologie im Überblick
4. Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text
5. Zusammenfassung und Literaturhinweise

**HHL**...

Wettbewerberanalyse: Definition

- 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)
- "Competitive a systematic and ethical program for gathering, analyzing, and managing information that can affect your company's plans, decisions, and operations" (<http://www.scip.org>, 29.01.2002)

**HHL**...

Wettbewerberanalyse: Umfeld

(Sullivan)

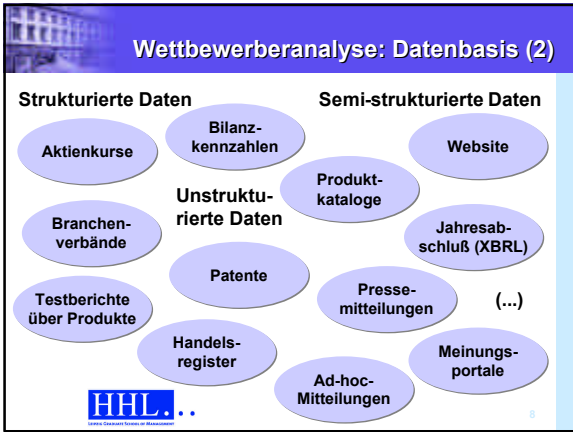
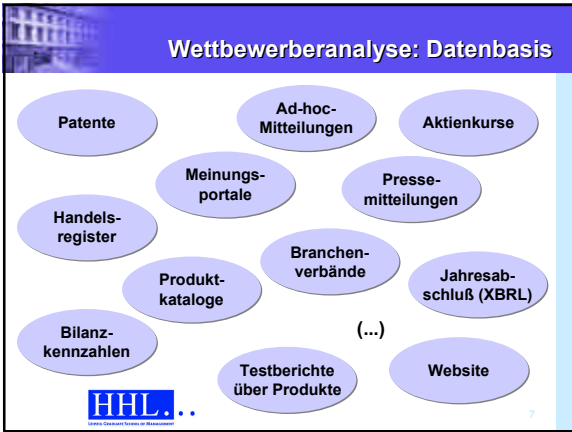
**HHL**...

Wettbewerberanalyse: Aufgaben

- Antizipation von relevanten Marktveränderungen
- Antizipation von Aktivitäten der Wettbewerber
- Entdeckung neuer, potentieller Wettbewerber
- Sammlung von Wissen über für das eigene Unternehmen potentiell relevante Technologien, Produkte, Gesetze und Verordnungen
- Auswertung der Erfolge und Mißerfolge anderer
- Verbesserung von Umfang und Qualität bei Unternehmenszusammenschlüssen

(Kahaner)

**HHL**...



### Herausforderung: Textuelle Daten

**DGAP-Ad hoc: Mannesmann AG**

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

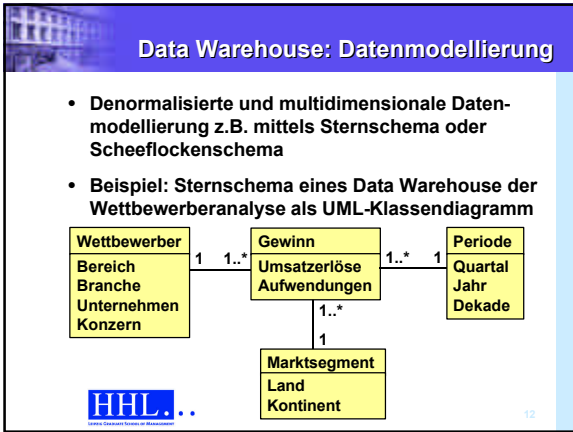
Nicht zur Veröffentlichung in den Vereinigten Staaten bestimmt. Not for distribution in the United States. This press release does not constitute an offer of securities in the United States, Canada, Japan or elsewhere.

- Bis zu 80 % betrieblicher Informationen sind Texte
- Informationsüberfluß
- Fehlende (offensichtliche) Struktur in Texten
- Speicherung der Texte und Abfrage bzw. Suche in Textdokumenten
- Entscheidungsrelevanz von Textdokumenten
- Beurteilung der Qualität

HHL...

- ### Agenda
1. Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerberanalyse
  2. Document Warehousing: Technologie im Überblick
  3. Text Mining: Technologie im Überblick
  4. Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text
  5. Zusammenfassung und Literaturhinweise
- HHL...

- ### Data Warehouse: Definition
- Verschiedene Anforderungen an Datenhaltung
    - Online Transaction Processing (OLTP)
    - Online Analytical Processing (OLAP)
  - "a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making" (Inmon)
  - "a copy of transaction data specifically structured for query and analysis" (Kimball)
  - Ziel: Historische, verdichtete und aus mehreren Quellen konsolidierte Daten hoher Qualität
- HHL...



### Data Warehouse: OLAP

- OLAP-Operationen bei multidimensionaler Datenhaltung: Slice, Dice, Roll-Up, Drill-Down, Rotate

HHL...

### Data Warehouse: Beispielarchitektur

(Chaudhuri und Dayal)

HHL...

### Document Warehouse: Definition

- Relativ junger Begriff (Google: 1.400 Webseiten) im Gegensatz zu Data Warehouse (Google: 364.000 Webseiten)
- "document warehouse provides a repository for text and text metadata" (Sullivan)
- Analogie zu Definition des Data Warehouse:
  - "a subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making" (Inmon)
  - "a copy of transaction data specifically structured for query and analysis" (Kimball)

HHL...

### Document Warehouse: Ziele

- Unterstützung von Entscheidungsprozessen
  - Datenbasis für Text Mining-Aktivitäten
  - Ergänzung eines existierenden Data Warehouse
  - Beispiel: Signifikanter Gewinnanstieg eines Wettbewerbers in Deutschland in IV/2001

(Sullivan)

HHL...

### Document Warehouse: Ziele (2)

- Systematische Verwaltung von Dokumenten aus verschiedenen internen und externen Quellen

HHL...

### Document Warehouse: Ziele (3)

- Systematische Ablage unterschiedlicher Dokumenten ohne einheitliche Struktur

HHL...

### Document Warehouse: Ziele (4)

- Erzeugung und Speicherung von Metadaten
- Extraktion und Ablage entscheidungsrelevanter Informationen aus Dokumenten
  - Dominierende Themen in Dokumenten
  - Zusammenfassung von Dokumenten
  - Extraktion benannter Entitäten (z.B. Personen)
- Integration semantisch verwandter Dokumente
  - Segmentierung von Dokumenten (Clustering)
  - Indexierung von Dokumenten

HHL...

### Document Warehouse: Inhalt

(Sullivan)

HHL...

### Document Warehouse: Metadaten

- "Daten über Daten" (z.B. Meta-Tags in HTML)
- Inhaltsbezogene Metadaten
  - Autor, Titel, Beschreibung, Schlüsselwörter, ...
  - Doblin Core-Standard: dc.language=de
- Herkunftsbezogene Metadaten
  - Quelle, Suchanfrage, Nutzernamen, Passwort, ...
- Document Warehouse-bezogene Metadaten
  - Ladedatum, Übersetzung, Speicherung von Dokument vs. Zusammenfassung vs. URI, ...

HHL...

### Document Warehouse: Metadaten (2)

Umsetzung mit Oracle SQL:

```

CREATE TABLE Documents (
  Id INTEGER NOT NULL,
  Content_Metadata_Id INTEGER,
  DocSourceId INTEGER,
  Source_Id INTEGER,
  Storage_Metadata_Id INTEGER,
  Contents CLOB,
  Summary CLOB,
  Version INTEGER,
  Document_Expires_On DATE,
  Summary_Expires_On DATE,
  Date_Loaded DATE,
  Last_Verified DATE
);

CREATE TABLE Content_Metadata (
  Id INTEGER NOT NULL,
  Creator VARCHAR2(100),
  Subject VARCHAR2(100),
  Title VARCHAR2(100),
  Description VARCHAR2(100),
  Publisher VARCHAR2(100),
  Contributor VARCHAR2(100),
  Published DATE,
  Revised DATE,
  Type VARCHAR2(100),
  Format VARCHAR2(100),
  Language VARCHAR2(100),
  Rights VARCHAR2(100)
);
  
```

(ähnlich Sullivan)

HHL...

### Document Warehouse: Thesaurus

- Kontrolliertes, meist fachspezifisches Vokabular für themenbezogene Indexierung und Suche, z.B.

HHL...

### Document Warehouse: Thesaurus (2)

- Thesaurus als Sammlung semantischer Konzepte
- Inhalt: Deskriptoren, Nicht-Deskriptoren mit Verweis zu Deskriptor, Definition, Synonyme sowie Beziehungen zu anderen Termen
- Auszug aus einem Thesaurus nach DIN 1463 / ISO 2788

```

TT Unternehmen
SYN Firma (Organisation)
NT Handelsgesellschaft
NT Kapitalgesellschaft
NT Eingetragener Kaufmann
Handelsgesellschaft
NT Kommanditgesellschaft
NT Offene Handelsgesellschaft
UF Kommanditgesellschaft
UF Offene Handelsgesellschaft
BT Unternehmen
Kommanditgesellschaft
USE Handelsgesellschaft
BT Handelsgesellschaft
SYN KG (...)
  
```

HHL...

### Document Warehouse: Indexierung

- **Volltextindex (keyword index)**
- **Inhalt: Alle Terme sämtlicher Texte, ggf.**
  - Bereinigung um sinnleere Worte (Stopworte)
  - Indexierung der grammatischen Grundformen

HHL...

### Document Warehouse: Indexierung (2)

- **Thematischer Index (thematic index)**
- **Inhalt: Themen und Konzepte sämtlicher Texte**
  - Basis: Kontrolliertes Vokabular (z.B. Thesaurus)
  - Einbettung von Semantik des Anwendungsgebiets

HHL...

### Document Warehouse: Indexierung (3)

- **Index benannter Entitäten (feature index)**
- **Inhalt: Benannte Entitäten sämtlicher Texte**
  - Kontrolliertes Vokabular nicht erforderlich
  - Identifikation wichtiger Terme und Entitäten

HHL...

### Document Warehouse: Beispielarchitektur

HHL... (in Anlehnung an Sullivan)

### Document Warehouse: Beispielanfragen mit Oracle SQL

```

SELECT Id FROM Documents WHERE
CONTAINS (Contents, 'Bestellung Geschäftsführer') > 0;

SELECT Id FROM Documents WHERE
CONTAINS (Contents, 'Gründung | $gründen') > 0;

SELECT Id FROM Documents WHERE
CONTAINS (Contents, NT('Kapitalgesellschaft')) > 0;

SELECT Id FROM Documents WHERE
CONTAINS (Contents, ABOUT('GmbH')) > 0;

SELECT Id FROM Documents WHERE
CONTAINS (Contents, '$Preis & ABOUT('Änderung')) > 0;

SELECT Id FROM Documents WHERE CONTAINS (Contents,
'NEAR(ABOUT('neu'), ABOUT('Produkt'), 20)' > 0;

```

HHL...

### Agenda

1. Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerbsanalyse
2. Document Warehousing: Technologie im Überblick
3. Text Mining: Technologie im Überblick
4. Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text
5. Zusammenfassung und Literaturhinweise

HHL...

### Data Mining: Definition

- Wissensentdeckung in Datenbanken
- Gewinnung von **neuem, nicht trivialem, interessantem und vor allem ökonomisch umsetzbarem Wissen** aus riesigen Datenbeständen (Fayyad et al.)
- Typische Fragestellungen:
  - Welche Kunden eines TK-Anbieters sind abwanderungsgefährdet?
  - Welche Kunden kaufen tendenziell Produkte eines Spezialkatalogs?

HHL...

### Data Mining: Vorgehensmodell

- Interaktiver und iterativer Prozeß (Mannila)

```

    graph TD
      A[Definition der Ziele] <--> B[Aufbereitung der Daten]
      B <--> C[Musterentdeckung]
      C <--> D[Nachbereitung der Ergebnisse]
      D <--> E[Umsetzung der Ergebnisse]
      E --> A
  
```

HHL...

### Data Mining: Methodik

- Musterentdeckung durch Methoden der Statistik, des maschinellen Lernens, der künstlichen Intelligenz und der Informatik
- Aufgaben vs. Methoden

Segmentierung	→	Clustering
Warenkorb-analyse	→	Assoziationsverfahren
Klassifikation (Steinicke)	→	Neuronale Netze
	→	Entscheidungsbäume

HHL...

### Data Mining: Beispiel

- Klassifikation profitabler Kunden mit Entscheidungsbaumverfahren: Was charakterisiert profitable Kunden?

- Aktion: Kaufanreize in Echtzeit schaffen!

HHL...

### Text Mining: Definition

- Wissensentdeckung in textuellen Datenbanken (Feldman und Dagan)
- Eher umfassender: "Text mining is the **process of compiling, organizing, and analyzing large document collections**"
  - to support the **delivery of targeted information** to analysts and decision makers and
  - to **discover relationships between related facts** that span wide domains of inquiry." (Sullivan)
- Methoden des Data Mining, des Information Retrieval und der Information Extraction

HHL...

### Text Mining: Unstrukturierter Text?

- Verteilung der Wort-häufigkeiten in einem Textarchiv (**Zipf's Gesetz**):
- Größe des Vokabulars in Abhängigkeit von der Textgröße (**Heap's Gesetz**):

HHL... (Baeza-Yates und Ribeiro-Neto)

### Text Mining: Unstrukturierter Text? (2)

- Textsstruktur aus Sicht der Linguistik, z.B.
- Struktur und Form der Worte (**Morphologie**)
  - Präfix - Wortstamm - Suffix, flektierte Wortformen
  - Basis für Ermittlung grammatischer Grundformen
- Bildung von Wortgruppen und Sätzen (**Syntax**)
  - Substantivgruppen, Subjekt - Prädikat - Objekt
- Bedeutung der Wörter und Aussagen (**Semantik**)
  - Synonyme: Unternehmen, Firma (Organisation)
  - Homonym: Firma (Organisation, registrierter Name)

**HHL**...

### Text Mining: Textrepräsentation

- Vektorraummodell des Information Retrieval (Salton et al.)
  - Extraktion der Merkmale für jedes Dokument (z.B. sämtliche Terme oder auch nur bestimmte Konzepte)
  - Transformation aller Dokumente in einen i.d.R. hochdimensionalen Vektor

	Doku- ment 1	(...)	Doku- ment M
Merkm- al 1	H <sub>1,1</sub>	(...)	H <sub>1,M</sub>
Merkm- al 2	H <sub>2,1</sub>	(...)	H <sub>2,M</sub>
(...)	(...)	(...)	(...)
Merkm- al N	H <sub>N,1</sub>	(...)	H <sub>N,M</sub>

H<sub>n,m</sub>: Absolute Häufigkeit des Merkmals n in Dokument m

**HHL**...

### Text Mining: Textrepräsentation (2)

- Bestimmung des Gewichts der Terme in Dokumenten z.B. als Produkt aus
  - Absoluter Häufigkeit des Terms n in Dokument m und
  - Inverser Häufigkeit des Terms n in allen Dokumenten.
- Dimensionsreduktion!

	Doku- ment 1	(...)	Doku- ment M
Merkm- al 1	G <sub>1,1</sub>	(...)	G <sub>1,M</sub>
Merkm- al 2	G <sub>2,1</sub>	(...)	G <sub>2,M</sub>
(...)	(...)	(...)	(...)
Merkm- al N	G <sub>N,1</sub>	(...)	G <sub>N,M</sub>

G<sub>n,m</sub>: Gewicht des Merkmals n in Dokument m

**HHL**...

### Text Mining: Textrepräsentation (3)

Dokument 1:  
Pawel Balski, 14.04.1965, Berlin, ist zum Geschäftsführer bestellt.

Term	Dokument 1	Archiv
Geschäftsführer	H = 1	H = +++
bestellen	H = 1	H = +

Term	Dokument 1
Geschäftsführer	G = 0,9010
bestellen	G = 1,4283

**HHL**...

### Text Mining: Vorbereitung der Daten (Beispiel)

1. Zerlegung der Dokumente, um Terme zu isolieren  
(...) 1999 abgeändert . Pawel Balski , 14.04.1965 , Berlin, ist zum Geschäftsführer bestellt . Er vertritt die (...)
2. Extrahierung und Ersetzung benannter Entitäten  
(...) PERSON , ist zum Geschäftsführer bestellt . Er (...)
3. Bestimmung der grammatischen Grundformen  
(...) PERSON , sein zu Geschäftsführer bestellen . Er (...)
4. Festlegung der Dimensionen des Vektorraums  
(bestellen, ..., Geschäftsführer, Gründung, Gesellschaft)
5. Abbildung der Dokumente auf den Vektorraum  
(1, ..., 1, 0, 0 )
6. Bestimmung der Gewichte je Term und Dokument  
(1.4283, ..., 0.9010, 0, 0 )

### Text Mining: Clustering

- "... the art of finding groups in data." (Kaufman, Rousseeuw)
- Einteilung der Daten in a priori unbekannte Kategorien, Klassen oder Gruppen, so daß
  - Objekte im gleichen Cluster möglichst ähnlich und
  - Objekte aus verschiedenen Clustern möglichst unähnlich zueinander sind (Ester, Sander)
- Bestimmung der Ähnlichkeit von Texten?
  - Euklidische Distanz der Dokumentvektoren
  - Kosinus des Winkels zwischen Dokumentvektoren

**HHL**...

### Text Mining: Clustering (2)

- Hierarchische Verfahren (z.B. Cobweb)**
- Partitionierende Verfahren (z.B. k-Means)**

- Beispiel: Entdeckung einer Taxonomie von Dokumenten**
- Beispiel: Entdeckung von Dokumentklassen für Klassifikation wie etwa Ereignisse**

**HHL...**

### Text Mining: Klassifikation

- Klassen der Dokumente sind a priori gegeben**
- Aufgabe ist die Zuordnung von Dokumenten aufgrund ihrer Attributwerte zu einer von n gegebenen Klassen, Teilaufgaben:**
  - Generierung von Klassifikationswissen auf Trainingsdaten mit bekannter Klassezugehörigkeit
  - Anwendung des Klassifikationswissen auf Dokumente mit unbekannter Klassezugehörigkeit (Ester, Sander)
- Nutzung verschiedener Methoden: Entscheidungsbaumverfahren, Neuronale Netze, ...**

**HHL...**

### Text Mining: Klassifikation(2)

- Generierung von Klassifikationswissen:**
- Anwendung von Klassifikationswissen:**

- Beispiel: Annotation des Trainingsarchivs mit n Ereignisklassen**
- Beispiel: Entdeckung von Ereignissen in neuen Texten wie etwa Presseerklärungen**

**HHL...**

### Agenda

- Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerbsanalyse
- Document Warehousing: Technologie im Überblick
- Text Mining: Technologie im Überblick
- Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text**
- Zusammenfassung und Literaturhinweise

**HHL...**

### Fallstudie: Handelsregistereintrag

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

**HHL...**

### Fallstudie: Anwendungsgebiet

**HHL...**

**Ziel: Klassifikation von Handelsregistereinträgen (Neueintragung, Veränderung, Löschung)**

**Software: SAS Enterprise Miner for Text**

**Diagramm**

**Ergebnis: e = 10.7%**

**HHL...**

## Agenda

1. Textuelle Daten als Herausforderung für die IT-Unterstützung der Wettbewerberanalyse
2. Document Warehousing: Technologie im Überblick
3. Text Mining: Technologie im Überblick
4. Software-Demo: Text Mining mit dem SAS Enterprise Miner for Text
5. Zusammenfassung und Literaturhinweise

**HHL...**

## Zusammenfassung und Ausblick

- Document Warehousing und Text Mining als komplementäre, zukunftsweisende Technologien
- Technologie in 1/02 m.E. im Early Adopters-Stadium
- Vielfältige Anwendungsgebiete:
  - Business Intelligence
  - Marketing
  - CRM und SCM
  - Produktion

**http://www.kdnuggets.com**

**HHL...**

## Literaturhinweise

- R. Baeza-Yates and B. Ribeiro-Neto: Modern Information Retrieval. Addison Wesley, 1999.
- G. Chang et al.: Mining the World Wide Web. Kluwer Academic Publishers, 2001.
- L. Kahaner: Competitive Intelligence. Touchstone Books, 1998.
- M. Mulhaupt: Data Mining und Text Mining im strategischen Controlling. Shaker Verlag, 2000.
- D. Sullivan: Document Warehousing and Text Mining. Wiley & Sons, 2001.

**HHL...**

## Vielen Dank

an die **DFG** !  
und den **SAS Academic Club** !

# Fragen ?

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

**HHL...**