Applied Analytical Data Science

Teil 10: Text Mining

Dr. Jörg-Uwe Kietz,
Vorlesung an der Univ. Zürich,

Mittwoch, 14:00-15:45 Uhr Vorlesung,
16:00-17:30 Uhr Übung

http://www.kietz.ch/AADS/

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung
- Taxonomien

Wozu Text Mining?

- Jeden Tag entsteht auf diesem Planeten neues Wissen, neue Informationen
- Dieses Wissen ist meistens in textueller Form abgelegt
- In Texten ist das Wissen unnütz, es sei den jemand liest es
- Niemand kann alle Texte lesen
- So bleiben Zusammenhänge verborgen

Was für Texte sind das?

- Grosse Sammlungen von Dokumenten, z.B.:
  Nachrichten, Forschungsberichte, Technische Papiere, Patente, Briefe, Bücher, Emails, SMS, Inter- und Intranet Web-Seiten, Memos, Gesetzestexte, ...
- Datenbanken mit grossen Freitextfeldern, z.B.:
  Problem & Lösungs Beschreibungen in „Help-Line“-DBs, Kundenkontakt Beschreibungen in CRM-DBs, …

Was ist Text Mining überhaupt?

„The discovery by computer of new, previously unknown information, by automatically extracting information from different written resources“

- Informationen aus Texten extrahieren
- Dabei geht es oft um große Mengen von Texten

Probleme:

- Computer verstehen die menschliche Sprache nicht
- Informationen sind oft implizit und ungenau
- Auch Informationen über die Texte, nicht nur die Informationen aus Texten müssen gesammelt werden
- Durch die große Menge an Text, oft sehr langsam

Wobei kann Text Mining helfen?

- Genau definierte Informationen extrahieren und so maschinenlesbar machen
  z.B. Telefonnummern aus SMS extrahieren
- Interaktives Suchen in Texten ermöglichen
  z.B. Indexierungen (Suchmaschinen)
- Charakterisierungen von Texten
  z.B. Ähnlichkeiten zwischen zwei Texten feststellen

Klassischer Umgang mit Text

- Computerlinguistik
  - syntaktische und semantische Analyse von Texten
  - Fokus auf einzelne Dokumente und Kontextinformation
  - Ziel: Verständnis und Verarbeitbarkeit
- Information Retrieval (IR)
  - als Forschungsgebiet parallel zu Datenbanken entwickelt
  - Sichtweise: Information ist als grosse Menge von Dokumenten organisiert
  - IR Problem: Lokalisieren relevanter Dokumente ausgehend von einer Benutzeranfrage
  - Typische IR Systeme
    - Online-Bibliothekskataloge
    - Online-Dokumentenverwaltungssysteme
    - WEB-Suchmachinen

Text Mining Aufgaben

- **Informations Extraktion**, z.B.: wer, was, wann, wo aus Terminvereinbarungen
- **Zusammenfassung**, z.B. Auswahl von Kernaussagen
- **Textklassifikation**, z.B.: Email filtering: SPAM vs non-SPAM
- **Textorganisation**, z.B.: Erstellung einer Email-Organisation
- **Topic Tracking**, z.B.: Zustellung „interesanter“ Nachrichten
- **Concept Linkage**, z.B.: Symptom, Ursache, Therapie Zusammenhänge, Zusammenhängende Personen/Organisationen
- **Informations Visualisierung**, z.B.: Themen und Publikationen zu einem Sachbereich
- **Frage Beantwortung**

Aufgaben & Anwendungen

	information extraction	topic tracking	summarization	categorization	clustering	concept linkage	information visualization	question answering
Medical:								
FAQ's	x			x		x		x
Drug design	x				x	x		
New treatment		x				x		
Business:								
Competitive Analysis		x	x					
Media impact / analysis		x						
Current Awareness		x						
Intellectual property infringement	x	x			x			
Customer support for FAQ's	x			x	x			x
Social network detection							x	
Content personalization		x			x			
Government:								
Homeland security: detecting terrorist networks	x	x			x	x	x	
Law enforcement: crime detection / prevention	x	x			x	x	x	
Education:								
Research on a topic		x	x	x				
Citation analysis	x				x		x	
FAQ's	x			x	x			x

Text Mining Aufgaben und DM-Methoden

Aufgabe	Methode
• Informations Extraktion	(Pattern Matching)
• Zusammenfassung	(Computerlinguistik)
• Textklassifikation	Klassifikation
• Textorganisation	Clustering
• Topic Tracking	Klassifikation
• Concept Linkage	Assoziationen
• Informations Visualisierung	Clustering, SOM
• Frage Beantwortung	(Q&A)

Problem des Text Minings

- Wie müssen wir **Texte aufbereiten**, so das **Data Mining** Verfahren benutzt werden können und die **Aufgaben des Text Minings** zu lösen?

Ein Beispiel Text

Anbei erhältst du den Stream der Gruppe X. Folgende Anmerkungen sind wichtig. Wir haben uns ziemlich genervt ab der Grösse des bisherigen Modellen, was sich in langen Wartezeiten in Clementine äusserte, oder man konnte den Stream nicht mehr öffnen. Deshalb haben wir einen Workaround gewählt. Wir haben die Übungen 1-3 in einem Stream, von welchem wir ganz am Schluss ein SPSS Output für eval und learn produzieren. Diese .sav-files sind viel kleiner als die Ausgangsdaten. Den Stream für die Übung 4 beginnt dann bei null und nimmt nicht die originalen .sav-files, sondern diejenigen, die von uns produziert wurden. Im Anhang findest du den Stream für die Übung 4, basierend auf den obigen Eingabedaten sowie unsere Präsentation.

Beispiele der Benutzung in Text Mining Aufgaben:

- Klassifikation: kein Spam
- Organisation: DM-Vorlesung SS05
- Informations Extraktion: Stream (Lösung) von Gruppe X für Übung 4

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung
- Taxonomien

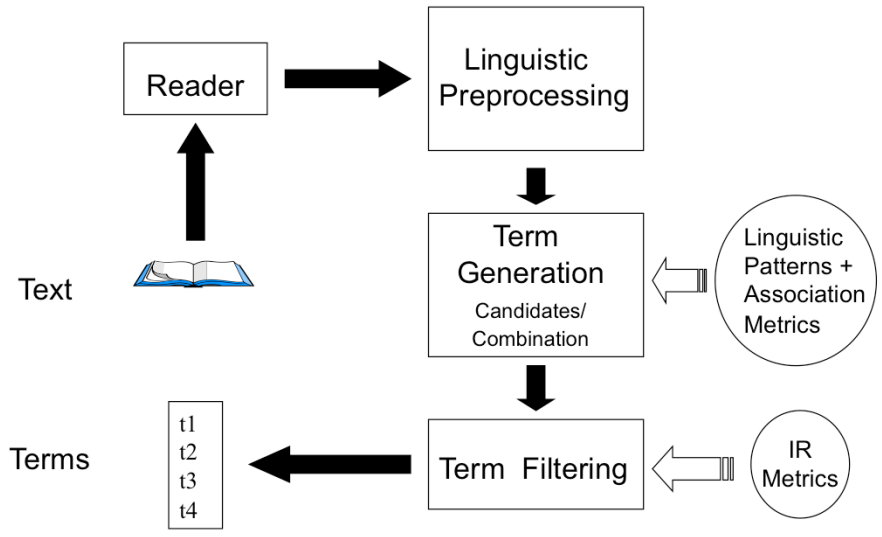
Texte Sammeln

- Wenn ich Informationen suche, muss ich zunächst Texte suchen, von denen es wahrscheinlich ist, dass sie die gewünschte Information enthalten
- Manchmal sehr einfach, oft sehr schwer, wegen Zugriffsbeschränkungen, etc.
- Alle Texte müssen in einem einheitlichen, lesbaren Format zugänglich sein.
  - Viele verschiedene Formate (Bitmap, ASCII, Word, PDF, HTML, …)
  - Was ist Text, was ist layout (oder Werbebanner)?

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte Vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung
- Taxonomien

Texte Vorverarbeiten



Wortindex des Beispiel Textes

6 x die	5 x Stream	5 x den	4 x wir	3 x in
3 x haben	3 x für	2 x Übung 4	2 x von	2 x uns
2 x und	2 x sind	2 x sav-files	2 x nicht	2 x du
2 x der	1 x Übungen 1-3	1 x öffnen	1 x äusserte	1 x ziemlich
1 x wurden	1 x Workaround	1 x wichtig	1 x welchem	1 x was
1 x Wartezeiten	1 x viel	1 x unsere	1 x Spss	1 x sowie
1 x sondern	1 x sich	1 x Schluss	1 x Präsentation	1 x produziert
1 x produzieren	1 x Output	1 x originalen	1 x oder	1 x obigen
1 x null	1 x nimmt	1 x Modellen	1 x mehr	1 x man
1 x learn	1 x langen	1 x konnte	1 x kleiner	1 x im
1 x Grösse	1 x Gruppe X	1 x gewählt	1 x genervt	1 x ganz
1 x folgende	1 x findest	1 x eval	1 x erhältst	1 x Eingabedaten
1 x einen	1 x einem	1 x ein	1 x diese	1 x diejenigen
1 x deshalb	1 x des	1 x dann	1 x Clementine	1 x bisherigen
1 x bei	1 x beginnt	1 x basierend	1 x Ausgangsdaten	1 x auf
1 x Anmerkungen	1 x Anhang	1 x anbei	1 x am	1 x als
1 x ab				

Bestimmung von Schlüsselwörtern

- Vorgegebene Listen von irrelevanten Worten (stop-words), z.B.: Artikel (der, die, das, …), Präpositionen (mit, durch, …), …
- Einfache Heuristiken, z.B. alle Substantive
- Zusammenfassung syntaktischer Variationen z.B.: produziert, produzieren, …. (Wortstammbildung)
- Statistische Bestimmung von Schlüsselwörtern
- Komplexe Schlüsselwörter, z.B. Namen, Phrasen, … (Übung 4, Gruppe X)
- Vorgegebene Listen von Schlüsselwörtern
- Vorgegebene Terminologien/Ontologien

Einfache Heuristiken für Schlüsselwörter

Schlüsselwörter (z.B. Substantive):

Stream Gruppe X
Anmerkungen Grösse Modellen
Wartezeiten Clementine Stream
Workaround Übungen 1-3
Stream Schluss SPSS
Output .sav-files Ausgangsdaten
Stream Übung 4 .sav-files
Anhang Stream Übung 4
Eingabedaten Präsentation.

Bag of Words:

2 x .sav-files	1 x Anhang
1 x Anmerkungen	1 x Ausgangsdaten
1 x Clementine	1 x Eingabedaten
1 x Gruppe X	1 x Grösse
1 x Modellen	1 x Output
1 x Präsentation	1 x SPSS
1 x Schluss	5 x Stream
1 x Wartezeiten	1 x Workaround
2 x Übung 4	1 x Übungen 1-3

Statistische Bestimmung der Schlüsselwörter

- **TF (Term Frequency):**

$$tf(term_i) = \frac{|term_i|}{\sum_i |term_i|}$$

- **TFIDF (Term Frequency - Inverted Document Frequency):**

$$tf\ idf(term_i) = tf(term_i)\log(\frac{N}{n})$$

n = number of documents with word
N = total number of documents

Ergebnis der Datenaufbereitung

Dokument - Schlüsselwort Matrix:

ID	keyword1	keyword2	…	keywordn
doc1	1	0		4
doc2	3	1		0
doc3	2	1		1
doc4	0	0		0
…				
docn	1	2		1

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte Vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung
- Taxonomien

Datenanalyse: Klassifikation

Klassifikation:

- Eingabe
  - Dokument - Schlüsselwort Matrix
  - Ergänzt um (manuelle) Klassenzugehörigkeit, z.B. SPAM: Y/N
- Erstellen einer Klassifikationsfunktion (siehe Teil 6)

⇒ Klassifikationsfunktion zur Klassifikation neuer Dokumente (Filtering oder Topic Tracking)

Datenanalyse: Assoziationen

Assoziationen :

- Eingabe
  - Keywords als Items,
  - Dokumente als Warenkörbe
- (Sequence) Assoziationsregel Mining (siehe Teil 7)

⇒ Bessere Schlüsselworte, z.B.:
„Stadt Zürich“, „Kanton Zürich“, „Zürich Versicherung“
„Mineral Mining“, „Data Mining“

⇒ Endecken von Zusammenhängen, z.B.:
„Preprocessing“, „Data Mining“, „Evaluation“ → „KDD“

Datenanalyse: Clustering

Clustering:

- Eingabe
  - Dokument - Schlüsselwort Matrix
- Erstellen eines Clusterings (siehe Teil 8)
- Beschreibung der Cluster
- Benutzung des Clusterings zur Organisation der Dokumente

Heutiges Programm

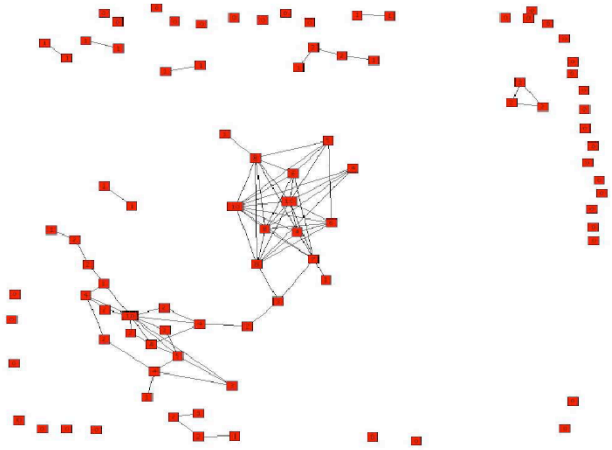
- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte Vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung (Teil II)
- Taxonomien (Teil II)

Visualisierung

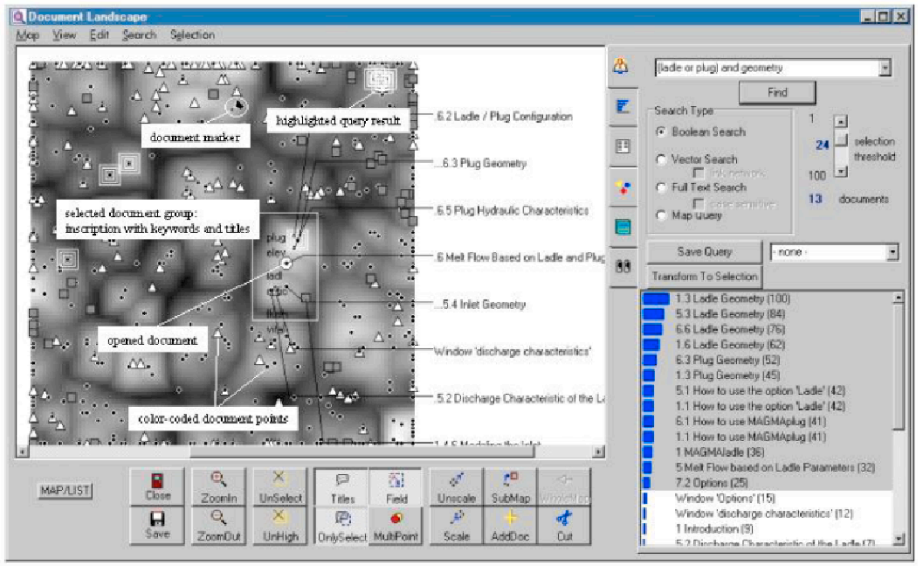
- Sorgt für die angemessene Präsentation der Ergebnisse
- Klarheit und Benutzerfreundlichkeit ist dabei wichtiger als bunte Bilder
- Je nach Aufgabe ist etwas anderes Angemessen

Beispiel: Suchergebnisse (Summarisation)

Beispiel: Concept Linkage



Beispiel: Dokument SOM



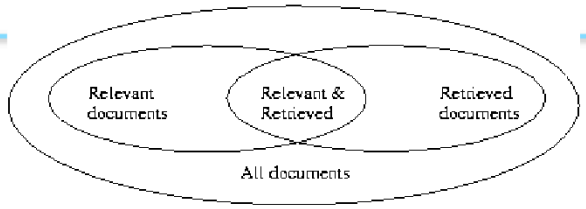
Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
  - Texte sammeln
  - Texte Vorverarbeiten
  - Datenanalyse
  - Visualisierung
  - Evaluation
- Linguistische Vorverarbeiung (Teil II)
- Taxonomien (Teil II)

Evaluation

- Überprüfen der Ergebnisse
- Automatische Evaluation
  - Messen der Laufzeit
  - Zählen der „korrekten“ Ergebnisse
  - Vergleich mit anderen
- Subjektive Evaluation
  - Direkte Bewertung durch Benutzer
  - Indirekte Bewertung, z.B. durch Webloganalyse

Evaluation von Text Retrieval/Klassifikation



- Precision: Anteil der gefundenen Dokumente, die tatsächlich relevant für eine Anfrage sind ("korrekte Antworten")

$$precision = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Retrieved\}|}$$

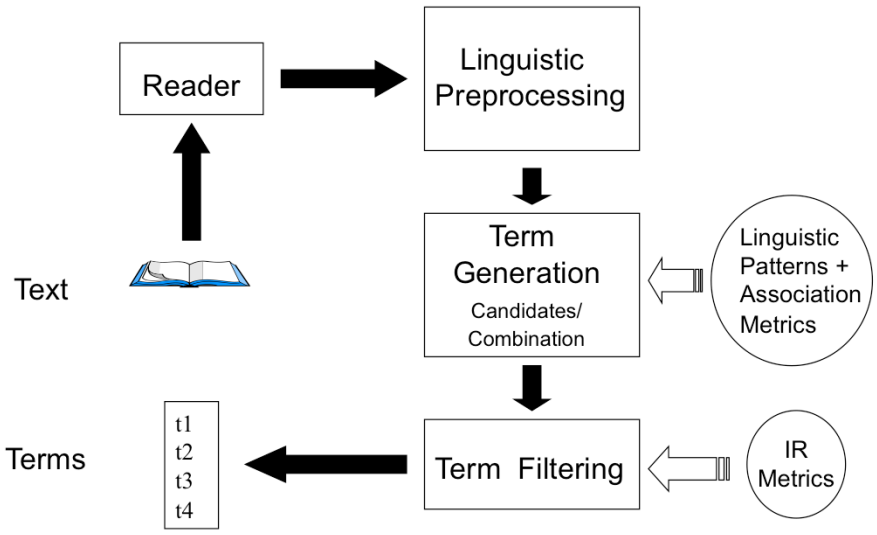
- Recall: Anteil der relevanten Dokumente, die tatsächlich gefunden wurden

$$recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}$$

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text
- Linguistische Vorverarbeitung von Texten
  - Zur Verbesserung der Schlüsselwortextraktion
  - Zur Verbesserung der Informations Extraktion
- Text Mining und Taxonomien
  - Was ist eine Taxonomie
  - Benutzung von Taxonomien

Texte Vorverarbeiten



Linguistischer Vorverarbeitung von Texten

- Verbesserung der Schlüsselwortextraktion
  - Spracherkennung und Tokenesierung
  - Lexikalische Analyse
  - Eigennamenerkennung

- Verbesserung der Informations Extraktion
  - Phrasenerkennung
  - Kasusrahmenzuordnung
  - Koreferenzauflösung

Ebenen linguistischer Vorverarbeitung

- Spracherkennung: z.B.: englisch oder deutsch?
  „Wir emailen den Output für die Miningdaten.“
  „Kindergarten is a place for children.“
- Tokenesierung: Erkennung von Zeichenketten
  - Worte und Wordgrenzen: Übung, 3, Gruppe, X, .sav-files, …
  - Datums- und Zeitangaben: 12:10, 23.4.2005, 11. Nov, 8am, …
  - Abkürzungen: AG, GmbH, bzw., z.B., …
  - Interpunktionszeichen: . ? ! : ; , „ “ - …
  - Satzgrenzen (nicht jeder Punkt ist ein Satzende!)

Ebenen linguistischer Vorverarbeitung

- Lexikalische Analyse
  - Wortart- und Stammerkennung (Morphologie), z.B.:
    ging => Verb geh
    Häusern => Substantiv Haus
    der => Artikel der
  - Flexionsformerkennung, z.B.:

ging =>	Person: 1., 3.;	Number: S,	Time: Imperfekt,	
Häusern =>	Person: 3,	Gender: N,	Number: P,	Case: Dat
der =>	Person: 3,	Gender : M;	Number: S;	Case: Nom
	Person: 3,	Gender : M, F, N;	Number: P;	Case: Gen
	…			

  - Kompositaerkennung, z.B.: Wartezeiten => Verb wart - Substantiv Zeit
  - Hypenkoordination, z.B.: An- und Verkauf

Ebenen linguistischer Vorverarbeitung

- Eigennamenerkennung und Koreferenzauflösung
  - Personennamen, z.B:
    Bundespräsident Schmidt, Samuel Schmid, S. Schmidt
  - Firmennamen und Produktnamen, z.B.:
    Zürich Financial Services, ZFS, Zürich, Zürich Versicherung
    Clementine, SPSS Clementine
  - Komplexe Datums, Zeit und Massausdrücke, z.B.:
    15. Juni 14:30, 2h 30min, € 4‘000.-

Ebenen linguistischer Vorverarbeitung

- Phrasenerkennung
  - NominalPhrasen,
    die langen Wartezeiten (NP: 3. Person, Nom oder Akk, plural),
    Wir (NP: 1. Person, Nom, plural)
  - Präpositionalphrasen
    mit den Übungen 1-3 (PP: mit)

⇒ Phrasen denotieren etwas (Extensionale Semantik)

⇒ Phrasen beschreiben etwas (Intensionale Semantik)

Ebenen linguistischer Vorverarbeitung

Kasusrahmenzuordnung

- Verben haben erforderte Ergänzungen, z.B:

lernen	Verb Activ	(Action)
die Studenten	NP Nom	(Agens)

- Verben können freie Ergänzungen haben, z.B:

Data Mining	NP Akk	(Object)
an der Uni	PP an	(Ort)
im Sommersemenster 2005	PP im	(Zeit)
von dem Dozenten	PP von	(Quelle)

⇒ Verben charakterisieren Vorgänge und Beschreibungen

⇒ Die Phrasen spielen darin verschiedene (gem. Kasus und PP) Rolen

Ebenen linguistischer Vorverarbeitung

- Koreferenzauflösung:, z.B.:
  Samuel Schmid ist neuer Bundespräsident der Schweiz geworden. Er besetzt das Amt in 2005. Schmids Stellvertreter ist Moritz Leuenberger, der Bundespräsident des Jahres 2001. Er war auch schon mal 2000 Vize-Präsident. Der neue Präsident ist seit Dez. 2000 als Vertreter des Kantons Bern im Bundesrat.
- Aufzulösende Koreferezen:
  - Samuel Schmid –Er – Schmids – Vertreter des Kantons Bern seit 2000 – der neue Präsident – Bundespräsident des Jahres 2005
  - Moritz Leuenberger – Bundespräsident des Jahres 2001 – Er – Vize-Präsident 2000 – (Vize-Präsident 2005)
  - Das Amt – Bundespräsident der Schweiz

⇒ Referenzauflösung ist schwierig und erfordert viel Hintergrundwissen.

Ebenen linguistischer Vorverarbeitung

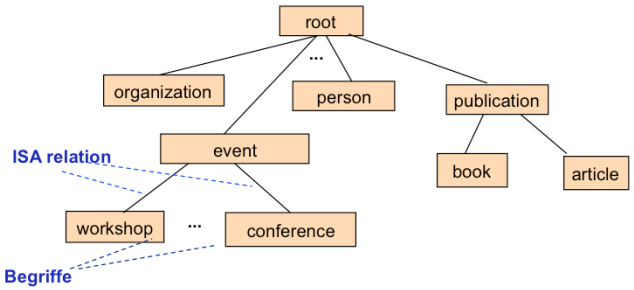
- Die Ebenen linguistischer Verarbeitung bauen aufeinander auf, d.h. die Erfolgsquote nimmt immer mehr ab.
  Z.B.: 5% Fehler/Verlust pro Ebene, ergibt 23% nach 5 Schritten
- Fehler- und Verlustrate basiert auf
  - (Un-) Vollständigkeit des Tools und
  - Textqualität der Eingabe
- Es gibt gibt kommerzielle und freie Tools die diese Linguistische Vorverarbeitung leisten, z.B.:
  - SMES (deutsch, frei):
    http://www.dfki.de/~neumann/pd-smes/pd-smes.html
  - Inxight LinguistX (> 30 Sprachen, kommerziel)
    http://www.inxight.com/products/sdks/lx/

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
- Linguistische Vorverarbeitung von Texten
  - Zur Verbesserung der Schlüsselwortextraktion
  - Zur Verbesserung der Informations Extraktion
- Text Mining und Taxonomien
  - Was ist eine Taxonomie
  - Benutzung von Taxonomien

Was ist eine Taxonomie

- Taxonomien dienen zur Repräsentation von
  - (Wort-) Bedeutungen und
  - Hintergrundwissen
- Die Basis ist eine (Ober-) Begriffs-Hierarchie, z.B.:



Was ist eine Taxonomie

Basis:

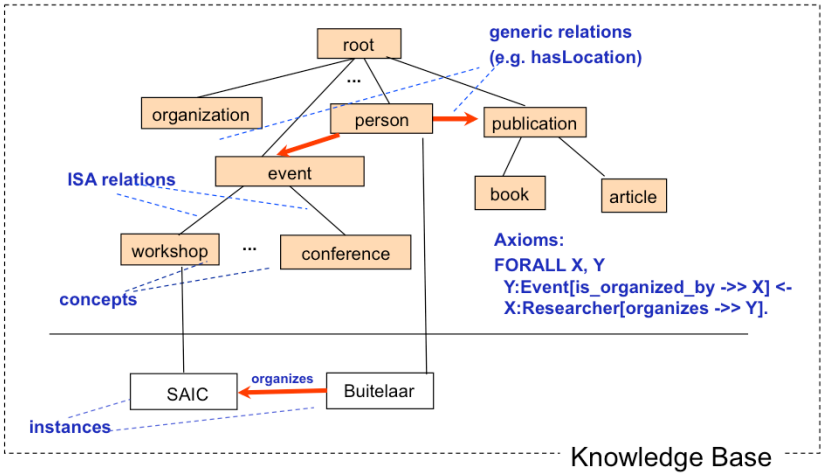
- Terminologie:
  kontrolliertes Vokabular
- Taxonomie:
  kontrolliertes Vokabular mit Ober- und Unterbegriffen

Erweiterungen:

- Thesaurus:
  kontrolliertes Vokabular mit Ober-, Unterbegriffen,
  Synonymen (USE ...), verwandten Begriffen
- Ontologie:
  Taxonomie von Begriffen, mit Relationen zwischen den Concepten und
  Festlegung der Begriffsbedeutungen

Was ist eine Ontology

Ontology = Taxonomie + Relationen + Bedeutungs Axiome



Was ist ein Thesaurus

Basis eines Thesaurus sind Wortbedeutungen (von Substantiven, Verben, Adjektiven und Adverbien) und (allgemeine, linguistische)
Relationen zwischen Ihnen:

• Polysemie	Schloss_1 (Schliessvorrichtung) Schloss_2 (Gebäude)
• Synonym	öffnen_3 **synonym** aufmachen_2
• Antonym	öffnen **versus** schliessen
• Hypernym / Hyponym	aufschliessen **isa** öffnen
• Meronym / Holonym	Tür **hasa** Schloss_1
• Reason / Cause	aufschliessen **erfordert** abgeschlossen öffnen **istGrund** offen
• Pointer (cf.)	schliessen **verweist** Schliessvorrichtung
• Related	finanziell **pertainsto** Finanzen

Vergleich Ontologie und Thesaurus

Ontology		Thesaurus
– Begriff (Concept)	~	Synonym-Menge
– Oberbegriff	~	Hyponym
– Gen. Relation	z.B.	Meronym, Holonym
Aber		
– Vererbung von Relationen	gilt nicht für	Antonym Reason / Cause Pointer (cf.) Related
– Gen. Relationen	nicht vorhanden	

Verfügbare Taxonomien

Thesaurus:

• WordNet	englisch	http://wordnet.princeton.edu/
• EuroWordNet inkl. GermanNet	multi deutsch	http://www.globalwordnet.org/
• OpenThesaurus	deutsch	http://www.openthesaurus.de/

Ontologie:

• CYC	-	http://www.cyc.com/

Heutiges Programm

- Einführung Text Mining
- Vorgehensweise Text Mining
- Linguistische Vorverarbeitung von Texten
  - Zur Verbesserung der Schlüsselwortextraktion
  - Zur Verbesserung der Informations Extraktion
- Text Mining und Taxonomien
  - Was ist eine Taxonomie
  - Benutzung von Taxonomien

Benutzung von Thesauri

- Die linguistischen Relationen im Thesaurus erlauben die auflösung von Mehrdeutigkeiten, z.B.:
  - *Bank* meint *Sitzmöbel*, weil *sitzen* und *Park* im Kontext vorkommen
  - *Bank* meint *Geldinstitut,* weil *Kredit* und *Konto* im Kontext vorkommen
- Synonym und Hypernym / Hyponym Relationen erlauben es nach Worten zu suchen/klassifizieren/clustern die gar nicht vorkommen.

Benutzung von Ontologien

Begriffe in Ontologien repräsentieren:
„Die Unterschiede die gemacht werden sollen“

- Begriffe eignen sich als die **keywords** für Text Mining und IR
  Aber die Begriffe müssen im Text identifiziert werden!
- Relationen eines Begriffs geben an was für solch einen Begriff wesentlich ist, z.B.:
  - Meeting

location	Place
startTime	TimePoint
duration	TimeInterval
participants	Person

Sie eignen sich als Pattern zur Informationsextraktion (IE)

Fazit

- Text Mining wird immer wichtiger, um große Mengen Text sinnvoll verarbeiten zu können
- Wichtiges Hilfsmittel bei Dokumentorganisation, Wissensmanagement, Literaturrecherche, etc.
- Eine gute Schlüsselwort Extraktion (Attribut Selektion) ist entscheidend für den Erfolg
- Linguitische Vorverarbeitung hilft bei der Schlüsselwort- und Informations-Extraktion.
- Taxonomien erlauben die Benutzung von Begriffen/Wörtern die nicht direkt im Text vorkommen für Text Mining und Retrival.

Literatur

W. Fan, L. Wallace, S. Rich, Z. Zhang: Tapping into the power of Text Mining, http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf

J. Dörre, P. Gerstl, R. Seiffert: Text Mining, Kapitel 12 In: Hippner; Küsters; Meyer; Wilde (Eds.): Handbuch Data Mining im Marketing,Vieweg, 2001.

K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, H. Langer (Eds.): Computerlinguistik und Sprachtechnologie: Eine Einführung, Spektrum Akademischer Verlag, Heidelberg, 2001