Volltextsuche und Text Mining

Datum: 22.12.2009

Seminar: Einführung in die Computerlinguistik

Referenten: Cornelia Baldauf, Valentin Heinz, Adriana Kosior

Agenda

1. Einführung
    a) Volltextsuche
    b) Text Mining
2. Volltextsuche
3. Text Mining
4. Quellen

1. Einführung: Volltextsuche

Problem: Wie finde ich das Dokument, das mir bei meinem Informationsbedürfnis weiterhilft?

Volltextsuche

Volltextsuche ermöglicht das Auffinden von Texten in einer Vielzahl von Dateien im Internet, auf dem Computer, usw.

1. Einführung: Volltextsuche

Vorkommende Termini und ihre Positionen im Text werden ermittelt und in einem Index als Indexterme abgespeichert

→ Dokumente werden auf diese Weise suchbar gemacht

Definition Index:
(alphabetisch) sortiertes Verzeichnis bestimmter Wörter oder Begriffe – Indexterme – unter denen Verweise auf Textstellen aufgelistet sind.

1. Einführung: Volltextsuche

Indexterme werden automatisch aus dem Text extrahiert und in eine sortierte Reihenfolge gebracht

- Meist Normalisierung der Worte auf Stammformen
- Meist Ausfilterung der Stoppworte

1. Einführung: Text Mining

~ Wissensgewinnung aus Texten

→ Automatischer Prozess, enthaltenes Wissen in Textdokumenten strukturiert aufzubereiten

→ Neues, interessantes und verwertbares Wissen, sowie Beziehungen zwischen Texten entdecken

→ Sammlung von Techniken und Algorithmen zur automatischen Analyse von unstrukturierten Daten

  → Methoden: Informationsextraktion, Clusteranalyse, Mustererkennung

1. Einführung: Text Mining

Abgrenzung zum Data Mining:

- Prozess zur Extraktion von impliziten, bislang unbekannten Informationen aus großen Datenbanken
- Prozess der Identifizierung neuer, potentiell nützlicher Muster in großen Datenbanken

→ Text Mining arbeitet nicht auf den strukturierten Daten einer Datenbank, sondern versucht unstrukturierte Daten (Textdokumente) in eine Struktur zu überführen

1. Einführung: Text Mining

Prozess des Text Mining:

2. Volltextsuche
Linguistische Ebenen

Phonetik:	Nein
Phonologie:	Nein
Morphologie:	Ja
Syntax:	Ja
Semantik:	Nein
Pragmatik:	Nein

2. Volltextsuche
Generelle Schwierigkeiten

- Textmenge meist sehr groß
- Suchanfragen während die Textmenge sich laufend verändert
- Suchergebnisse sollen schnell ausgegeben werden

2. Volltextsuche
Linguistische Schwierigkeiten

Indexerstellung: *(Morphologie + Syntax)*

- Erkennen der Wörter(Indexterme), Wortgrenzen etc.
  - z.B. Hunde-Kuchen, Sindbad der Seefahrer

Normalisierung: *(Morphologie + Syntax)*

- (sprachabhängige) Reduktion der ermittelten Indexterme auf Stammformen
  - z.B. *Hund-Kuchen, Kätzchen, leitete her

2. Volltextsuche
Linguistische Schwierigkeiten

Retrieval: *(Semantik + Pragmatik)*

- Recherchieren / Stellen der Suchanfrage *(teilweise lösbar durch Trunkierungen)*

  z.B. Hundekuchen | Hunde-Kuchen | Kuchen für Hunde,
  Schmidt | Schmid | Schmitt

- Erkennen der Bedeutung der Suchterme

  z.B. Golf (Automarke, Sportart, Bucht),
  Ring (Schmuck, Augenring, Boxring, Jahresring v. Bäumen)

3. Text Mining

Sammlung von Techniken und Algorithmen zur automatischen Analyse von unstrukturierten Daten

Unstrukturierter Text soll maschinell verarbeitet werden
→ strukturelle linguistische Aufbereitung

Arbeitsgegenstand → annotierter Textkorpus (Textsammlung)

3. Text Mining – Vorverarbeitung
Linguistische Bestandteile

Bereich:		**Methode:**
Phonetik/Phonologie	Nein	
Morphologie	Ja	Stemming Lemmatisierung Kompositaanalyse
Syntax	Ja	Parsing
Semantik	Ja	Wortsinndisambiguierung
Pragmatik	Nein	
Grammatik	Ja	POS Tagging

3. Text Mining
Linguistische Schwierigkeiten

- Wörter auf **Grundformen** reduzieren:

  Lauf aus dem über-laufen-den Staubecken.

  → Stemming.

- Wörter auf gemeinsame **Stämme** zurückführen:

  Lauf aus dem überlaufenden Staubecken.

  → Lemmatisierung.

3. Text Mining
Komposita auflösen

Lauf aus dem überlaufenden Staubecken.

→ **heuristische Kompositaanalyse**

→ vorkommende Begriffe ermöglichen die Angabe einer **Wahrscheinlichkeit**, welche Bedeutung gemeint ist (Affix undoder Kontextebene)

3. Text Mining
Part of Speech Tagging

Wortart herausfinden: **Lauf = Verb/Nomen?**

Lauf aus dem überlaufenden Staubecken.

→ Part of Speech tagging

linguistisch: (Constraint Grammar) → unsere QDATR Regeln
statistisch:

→ Markov Model
→ Hidden Markov Model
→ Maximum Entropie Modell

3. Text Mining
IR-Methoden und Verfahren zu Dokumenten

Wortebene:

- Stoppwortliste und Ngramme

Satzebene:

- Parsing

Verfahren bzgl. Dokumente:

- Clustering
- Klassifizieren
- Vektorraummodell

3. Text Mining
Anwendungsbeispiele:

Texte einem (pseudonymen) Autor zuordnen
Automatische Emailbeantwortung

Trends erkennen
[Cuba → Cigar vs. Cuba → Rocket]
Neues Wissen entdecken
[Migräne → Magnesium]

Texte in Kategorien einteilen [news.google.com]
(Nachrichten/Sport/Reportage/...)

3. Text Mining
Beispiel:

Don Swanson (1994):
Titles of articles in the biomedical literature

stress is associated with migraines
stress can lead to loss of magnesium
calcium channel blockers prevent some migraines
magnesium is a natural calcium channel blocker

Neue medizinische Hypothese :

→ Magnesiummangel spielt bei Migräne manchmal eine Rolle.

3. Text Mining
Probleme, Probleme, Probleme, ...

- Abgrenzungsproblem: Begriff Textmining

- Komplexe Programme (Bsp: RapidMiner)
- Komplizierte Bedienung
- Viele Algorithmen/Abstandsmaße/...
- Anfangsproblem: Generierung durch Filterung
- Kriterienproblem:

There is nothing either good or bad,
But thinking makes it so.
(William Shakespeare, Hamlet, II:2)

Quellen

Carstensen et al. (2001). *Computerlinguistik und Sprachtechnologie. Eine Einführung.* Heidelberg, Berlin: Spektrum, Akad. Verl.

Wissensexploration.de. Retrieved December, 17, 2009 from http://wissensexploration.de/textmining.php

Mailvaganam, Hari. Text Mining for Fraud Detection. Retrieved December, 19, 2009 from http://www.dwreview.com/Data_mining/Effective_Text_Mining.html

Hearst, Marti A.. Untangling Text Data Mining. Retrieved December, 19, 2009 from http://people.ischool.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

Text Mining. Retrieved December, 19, 2009 from: http://en.wikipedia.org/wiki/Textmining

Übung Text Mining:

Rufe i:/EinfCl/textmining-email/ auf und öffne die .txt Datei.

Überlege dir Kriterien anhand derer man den Namen des Email-Absenders herausfinden kann.

Übung Volltextsuche:

Sindbad der Seefahrer traf Marco Polo im Polo-Shirt zum Polo spielen.

Er fütterte das Ross mit einem Pferde-Leckerli.

Polo wollte sich in den Sattel schwingen, hatte zuviel Schwung und landete auf dem Hinterteil.

Der Andere sagte lachend:“Geschickt, wirklich sehr geschickt...“ und schickte sich an ebenfalls aufzusteigen.