Volltextsuche und Text Mining

Seminar: Einfuehrung in die Computerlinguistik

Dozentin: Wiebke Petersen

2oo5-Jan-2o

Gliederung

1. Volltextsuche
    1. Zweck
    2. Prinzip
        1. Index
        2. Retrieval
    3. Qualitaet
2. Text Mining
    1. Analyse von Einzeltexten
    2. Merkmalsextraktion
    3. Analyse von Textkollektionen

Problem: große Datenmengen (z.B. Internet)

Lösung: Volltextsuche

- kann auf sehr effiziente Weise Informationen aus einer beachtlichen Datenmenge aufspueren, was manuell aufgrund des Aufwands gar nicht denkbar waere

Prinzip, Arbeitsweise, Arbeitsbedingungen

- Vergleichbar mit Schlagwortregister in einem Buch:
  - schlage ein Thema im Index nach und folge dem Verweis
- Besser: ein Register einer Bibliothek, in dem alle Woerter aller Buecher aufgelistet sind

- Index-Konstruktion
- Anfrage wird dem Index gegenuebergestellt

- Arbeitsbedingungen:
  - Kollektion K ist sehr groß
  - viele Anfragen zwischen K_{alt} und K_{neu}
  - Retrieval muss in kurzer Zeit arbeiten (Sekundenarbeit)

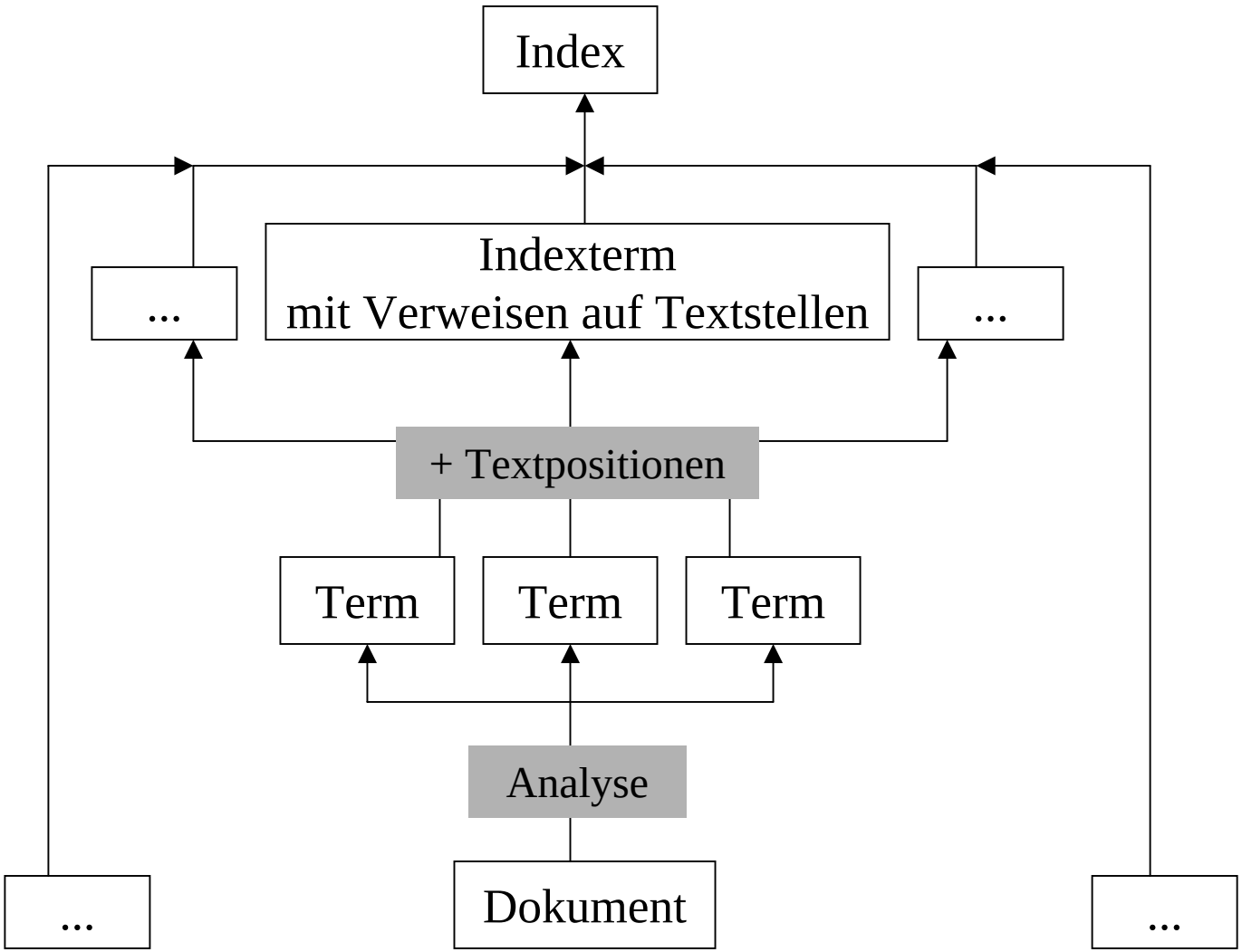
Volltextsuche in zwei Schritten

- Index-Konstruktion
  - findet schon bei der Eingabe in die Kollektion statt
  - speichert die Terme eines Dokuments mit den Textpositionen als **Indexterme**
  - Adressierung: Seitenzahlen, URLs, Buchstabenposition, ...

- Anfrage wird dem Index gegenuebergestellt
  - Ohne Zugriff auf die Originaltexte
  - abhaengig von dem Retrievalmodell

- Oft zusaetzlich: Sortierung des Retrievals nach unterschiedlichen Kriterien
  - am bedeutendsten: Relevanz fuer die Anfrage
  - man erhaelt eine Rangliste („Ranking“)

Index-Konstruktion



„Kostspieliges Unterfangen“

- Die Index-Konstruktion ist sehr kostspielig: $\text{Speicherbedarf}_{\text{Index}} = \text{Speicherbedarf}_{\text{Kollektion}}$ (oder knapp darunter)

- Allerdings werden die Kosten gerechtfertigt, wenn genug Anfragen gemacht wurden

Schritt 2: Retrieval

- Einfache Suche entspricht „Nachschlagen“ im Index
- Retrievalmodell bestimmt, welche Art von Verknuepfungen von Anfragen erlaubt sind
- Zwei Beispiele (fuer Retrievalmodelle):

  - Boolesches (oder mengentheoretisches) Retrievalmodell

  - Vektor(raum)modell

Qualitaet einer Suchmaschine

- Praezision & Vollstaendigkeit
- P: fuer die Anfrage tatsaechlich relevante Dokumente im Ergebnis / alle Dokumente im Ergebnis
- V: fuer die Anfrage tatsaechlich relevante Dokumente im Ergebnis / alle relevanten Dokumente in der Kollektion

- Bsp.: Anfrage „Kuchen“

  - 1oo Dokumente ueber Kuchen in der Kollektion
  - 1o Dokumente davon im Ergebnis
  - 1o andere Dokumente im Ergebnis

  P = 10 / 20 = 50 %

  V = 10 / 100 = 10 %

- Wird nur fuer den Anfang eines Retrievals errechnet (Ranking)

Text Mining

- aeusserst grosse / dramatisch wachsende Datenmengen
- Informationsaufbereitung: Mensch zu langsam
- Loesung: Automatisierung

Analyse von Einzeltexten

- stellt Datenformat, verwendete Zeichencodierung, Datenstruktur fest
- Quelle: Uebertragungsprotokoll, Markierungen im Text, etc.
- vergleiche haeufige kurze Woerter mit einem Profil typischer Zeichenfolgen einer Sprache
- bei fehlender Information → Annahmen

Merkmalsextraktion

- Normalisierung (Stammformbildung)
- Filterung von Stoppwoertern
- Normalisierung auch von Datumsangaben, Zahlenwerten, Abkuerzungen, etc. durch bestimmte Algorithmen

➔ Merkmalsmatrix

Analyse von Textkollektionen

- Ziel: Markierung signifikanter Unterschiede oder Zusammenhaenge zwischen Texten
- Zwei Beispiele:

  - Clustering

    dynamische Strukturierung der Dokumente

  - Textklassifikation

    Strukturierung der Dokumente nach vorgegebenem Schema

Bleiben Sie sitzen!

Das naechste Referat:

Textklassifikation

von und mit

Peter Bücker

Quellen

- K.-U. Carstensen et al., Computerlinguistik und Sprachtechnologie (2001)