# Information Retrieval - Wiederholung

Norbert Fuhr

31. Januar 2007

# Einführung

- Unterschiede zwischen Datenbank- und IR-Systemen
- Breite/enge Definition von IR:
  - Unsicherheit und Vagheit in Informationssystemen
  - inhaltsorientierte Suche
- Syntax, Semantik und Pragmatik von Objekten/Dokumenten

# IR-Konzepte

- Dimensionen zur Unterscheidung zwischen Datenbank- und IR-Systemen
- Daten, Information und Wissen
- Sichten auf Dokumente
- Anfragen und Sichten
  - Repräsenmtation vs. Beschreibung
  - Selektion und Projektion

# Evaluierung

- Grundlegende Anforderungen an eine Evaluierung
- Effektivität vs. Effizienz
- Arten von Relevanz
- Distributionen
- Benutzerstandpunkte
- Standardmaße: Recall und Precision
- Methoden zur näherungsweisen Bestimmung des Recalls
- Vergleich von Recall-Precision-Paaren
  - Wertepaare
  - F-Maß

# Rangordnungen

- Rangordnungen: lineare vs. schwache
- Berechnung von Recall und Precision bei linearen Ordnungen
  - direkte Berechnung
  - Treppenfunktion: zugrundeliegende Annahmen
- Schwache Rangordnungen
  - mögliche Standpunkte
  - expected search length
  - expected precision vs. PRR
  - Unterschied zur linearene Interpolation
- Evaluierungsinitiativen
  - Wesentliche Elemente einer Testkollektion: Doikumente, Topics, Relevanzurteile
  - Batch- vs. interaktives Retrieval

# Dokumentationssprachen

- Eigenschaften von Klassifikationssystemen
  - Mono- vs.Polyhierarchie
  - Mono- vs. Polydimensionalität
  - analytische vs. synthetische Klassifikation
  - Facettenklassifikation
- Elemente der Dezimalklassifikation:
  Hauptklassen, Facettierung, Verknüpfung

# Thesauri + RDF

- Terminologsche Kontrolle
  - Polyseme, Synonyme
  - Zerlegungskontrolle
  - Äquivalenzklasse Deskriptor
- Beziehungsgefüge
  - Äquivalenzrelation
  - Hierarchische Relation
  - Assoziationsrelation
- RDF
  - resource, literal, property, statement
  - RDF schemas

# Freitextsuche

- Probleme: Polyseme, Flexions- und Derivationformen, Komposita, Wortwahl
- Informatischer Ansatz:
  - Operatoren: Truncation, Maskierung, Kontextoperatoren
  - linguistische Interpretation, Recall/Precision
- Computerlinguistischer Ansatz
  - graphematische Verfahren: Grund- und Stammformreduktion
  - lexikalische Verfahene
  - syntaktische Verfahren: Wortklassenbestimmung, Parsing, Head-Modifier-Strukturen

# Nicht-probabilistische Retrievalmodelle

- Notationen
  - Informationsbedürdnis/Dokument
  - Repräsentationen
  - Beschreibungen
  - Retrievalfunktion
- Boolesches Retrieval
  - Fragebeschreibung, Retrievalfunktion
  - Mächtigkeit
  - Nachteile
- Fuzzy-Retrieval
  - Unterschiede zum booleschen Retrrieval
  - Retrievalfunktion / Alternativen

# Vektorraummodell

- Frage- und Dokumentbeschreibung
- Retrievalfunktionen
- Relevance Feedback
  - grundsätzliches Vorgehen
  - Optimierungsproblem/geometische Interpretation
  - Rocchio-Algorithmus
- Dokumenten-Indexierung: Heuristiken

# Clustering

- agglomeratives vs. partitionierendes Clustering
- agglomeratives Clustering
  - Vorgehen
  - Verfahren: single-link, complete link, average link
- partitionierendes Clustering: Vorgehensweise
- Probabilistisches Clustering
  - Ähnlichkeitsmaß
  - expectation maximization
- Scatter-Gather-Clustering

# Binary independence retrieval model

- Repräsentation von Fragen und Dokumenten
- Interpretation des Retrievalwertes im Ansatz
- Unabhängigkeitsannahme
- zu schätzende Parameter
- Parameterschätzung
  - $q_{ik}$
  - $p_{ik}$

# Probabilistisches Ranking-Prinzip

- perfektes vs. optimales Retrieval
- entscheidungstheoretische Rechtfertigung: Kostenfaktoren
- erwartete Kosten eines Dokumentes
- Optimierungsziel
- Rechtfertigung für Ordnung nach fallenden Relevanzwahrscheinlichkeiten

# Allgemeine Konzpte probabilistischer Modelle

- Repräsentationen und Beschreibungen im BIR-Modell
- Arten von Parameterlernen im IR
- Ereignisraum bei probabilistischer Modellen

# IR-Systeme

- Systemebenen
- Stufen der Systemunterstützung
- Ebenen von Suchaktivitäten

# Visualisierung und Benutzerschnittstellen

- Design-Prinzipien
- Prozessmodelle für die Informationssuche
  - klassisches Modell
  - Alternative Modelle
- Visualisierungen für die verschiedenen Schritte
  - Kollektionsauswahl
  - Anfrageformulierung
  - Ergebnisdarstellung
  - Relevance Feedback / Benutzerkontrolle
- Interfaces für den gesamten Suchprozess

# Summarization

- Arten von Zusammenfassungen
- Anwendungsbeispiele
- Ansätze:
  - lernende Verfahren
  - nichtlernende Verfahren
- Multi-Dosument summarization
  - Problemstellung
  - Verfahren: MEAD
- Wissensbasierte Ansätze: Radev & Mc Keown 98
- Evaluierung
  - Arten von Evaluierungen
  - Arten von Metriken