Information Retrieval - Wiederholung

Norbert Fuhr

13. Juli 2006

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

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
- Neuere Ansätze
  - Language models
  - Graph-basierte Ansätze
- Evaluierung
  - Arten von Evaluierungen
  - Arten von Metriken/Kriterien
-

Informationsextraktion

- Arten von Textverarbeitung
- Aufgabentypen bei der Informationsextraktion (Identifikation von Entitäten, Eigenschaften von Entitäten, Beziehungen zwischen Entitäten)
- Zwei Ansätze zur Konstruktion von Systemen zur Informationsextraktion
  - Eigenschaften
  - Vor- und Nachteile
- Architektur eines IE-Systems
- Ansätze zur Entwicklung von Namenserkennern
- Parsing-Methoden für IE
- Koreferenzen:
  - Problemstellung
  - Lösungsansätze
- Domänenphase: Molekularer vs. atomarer Ansatz

Probabilistisches Ranking-Prinzip

- Probabilistische Interpretation von Recall, Precision und Fallout
- Optimierung der Retrievalqualität durch das PRP
  - Vorgabe von Fallout/Recall/# Dokumente: Auswirkung auf Recall und Fallout
  - Auswirkung auf Precision
- Entscheidungstheoretische Rechtfertigung für mehrstufige Relevanzskalen
- Unterschiede zur zweistufigen Skala
- Kombination von Fuzzy- und probabilistischem Retrieval

Allgemeine Konzpte probabilistischer Modelle

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

Optimale polynomiale Retrievalfunktionen

- 2 Phasen beim Beschreibungs-orientierten Ansatz
- Optimierungskriterium beim Entscheidungsschritt
- Vorgehensweise beim Entwickeln einer polynomialen Retrievalfunktion
- Eigenschaften der Zwischenlösungen
- Vergleich Modell- vs. Beschreibungs-orientierter Ansatz

Divergence from Randomness

- Grundlegende Annahmen
- Generelle Form der Termgewichtung
- Binomial-Modell
- Bose-Einstein-Modell
- Erste Normalisierung
- Dokument-Längen-Normalisierung

Beschreibungslogiken: OWL

- Aussagen-vs. Prädikatenlogik für IR
- Thesaurus vs. Beschreibungslogik
- OWL:
  - Objekte, Klassen, Literale und Datentypen
  - Beziehungen zwischen Klassen
  - Properties und restrictions
  - Schwächen von OWL

Probabilistisches Datalog

- Grundkonzepte von Datalog: gewichtete Fakten und Regeln
- Modellierung von Hypertext-Strukturen und Aggregationen
- Semantik von pD:
  - Ereignisschlüssel, - ausdrücke
  - possible worlds-Semantik
  - disjunkte Ergeignisse
- Abbildung von OWL auf pD
  - Klassen
  - Properties
  - Individuen
  - Literale

Bild-Retrieval: Farbsuche

- Ebenen der Bildersuche
- Suchkriterien auf der syntaktischen Ebene
- Perzeptive Aspekte der Farbähnlichkeitssuche
- Suche nach Farbhäufigkeit: zu lösende Probleme
- Suche nach Farblayout: zu lösende Probleme
- Ansätze zur Suche nach Farblayout

Implementierung von IR-Systemen

- Aufbau eines IRS
- Dokumentarchitekturen
  - Zusammenhang zwischen logischer und Layout-Struktur in ODA
  - Markup vs. Generalized Markup Languages
  - SGML-DTDs
  - HTML vs. SGML

Zugriffspfade: Scanning

- naiver Algorithmus
- Funktionsweise von Knuth-Morris-Pratt
- Funktionsweise von Boyer-Moore-Horespool
- Shift-Or
  - Funktionsweise
  - Erweiterungen
- Ähnlichkeit von Zeichenketten
  - Damerau-Levenshtein-Metrik
  - Trigramme

Zugriffspfade: invertierte Listen

- prinzipieller Aufbau
- Boolesches Retrieval
- Ranking
  - Naiver Algorithmus
  - Quit-Algorithmus
  - Continue-Algorithmus
- Komprimierung
  - Codes zur Lauflängencodierung
  - Sprunglisten

XML-Standards

- Grundlegende Ideen von XML
- Weiterführende XML-Standards und ihre Funktion aus IR-Sicht:
  - Namespaces
  - DTDs
  - Schema
- XML Schema vs. DTDs
- XPath
  - Arten von Bedingungen
  - Achsen
- XQuery
  - FLWOR-Ausdrücke
  - Erweiterungen gegenüber XPath

XMLInformation Retrieval und INEX

- Modelle und Methoden
  - Arten von Anfragen und zugehörige Ziele
  - zu lösende Probleme bei der Entwicklung von Retrievalmethoden für XML
  - Behandlung von überlappenden Ergebnissen
- Interaktives Retrieval
  - Gemeinsamkeiten mit / Unterschiede zu Web-Retrieval
  - Probleme bei einfachen Interfaces
- Evaluierung
  - Elemente eines Testbeds
  - Relevanz bei XML-Retrieval vs. herkömmlichem Retrieval
  - Berücksichtigung von Überlappungen