Suchtechnologien und Information Retrieval

Dr. Christian Herta

April, 2009

Outline

Lernziele

- Inhalt der Vorlesungsreihe und Zielsetzung sind klar
- Problemstellung des Information Retrieval
- einige grundlegende Begriffe werden geklärt

Outline

- Themen die im Rahmen der Vorlesung behandelt werden
- Welche Mittel gibt es um mit der Informationsüberflutung zurechtzukommen?
- Bessere Suche
- Automatisches Filtern der Information, z.B. über Personalisierung (APML)
- Technische Apekte: Datenformate, Software, Algorithmen, Informationsarchitektur

Outline

Problemstellung oder warum braucht man Suche

- Der Großteil der zugängliche Information liegt unstrukturiert vor:
  - Über 80% der in Firmen vorliegenden Information liegt in unstrukturierter, textueller Form vor (siehe z.B. [2])
  - World Wide Web ist eine heterogene Sammlung von mit Hyperlinks verknüpften Dokumenten. Das vorherschende Format ist dabei HTML, d.h. Text inkl. Markups für die Darstellung.
  - Informationsexplosion: Digital verfügbare Datenmenge ist in letzter Zeit dramatisch gewachsen
- Nutzer haben Informationsbedürfnis (Information Need)
- Problemstellung: Informationsbedürfnis des Nutzers zu stillen ohne ihn mit zu Information zu konfrontieren (Information Overload vermeiden)
  - z.B. die nur passenden Dokumente (bzw. Links) oder Fakten zu liefern.

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

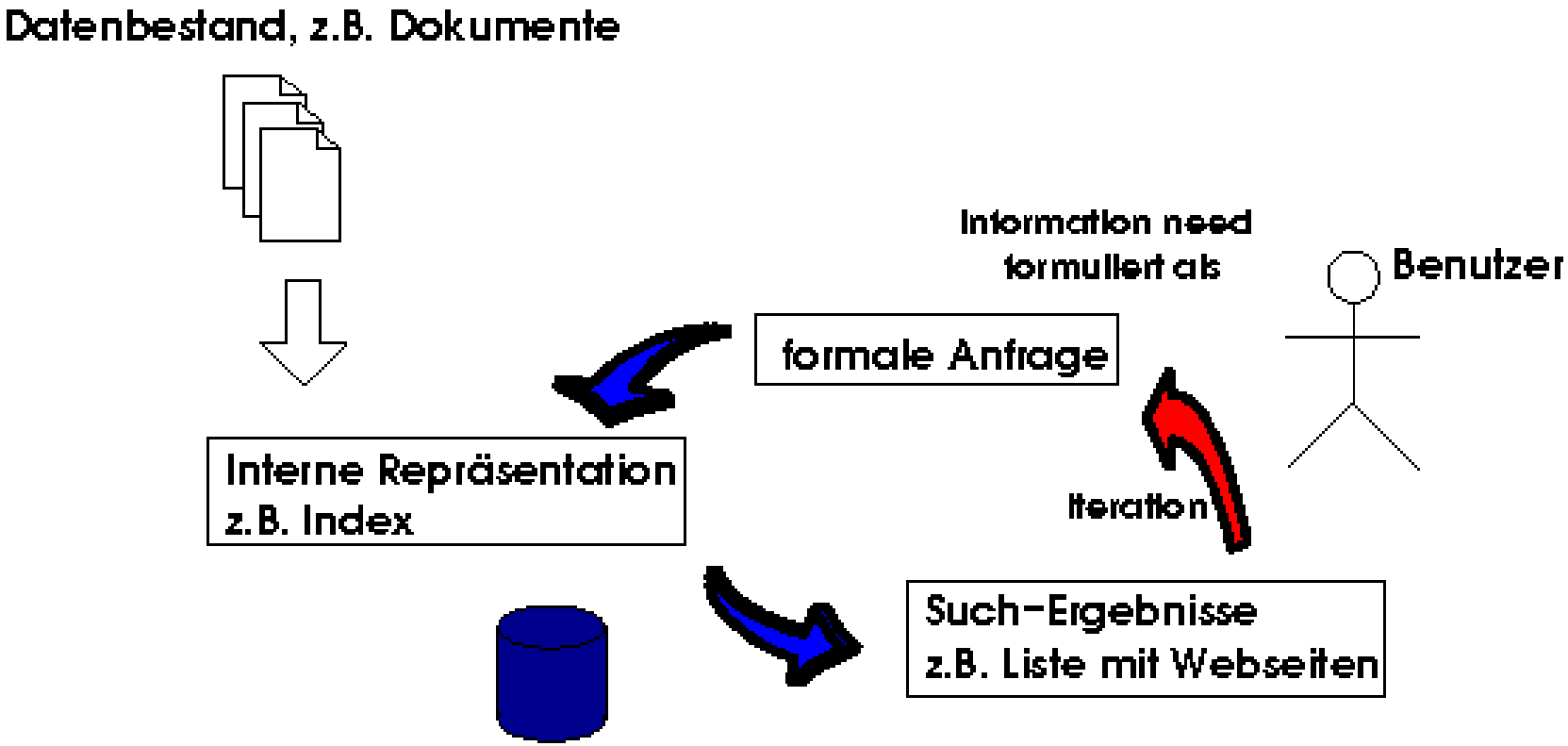
Definition von Information Retrieval (nach [1])

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Information Retrieval

- Information Retrieval - Informationswiedergewinnung
- Informationen sind in großen Datenbeständen vorhanden, sie müssen zu einem Informationsbedürfnis gefunden werden.
- Beispiele
  - Text oder Document Retrieval: Finden von Text-Dokumenten zu Anfragen; wie Websuche oder Literatursuche
  - Multimedia Retrieval: Finden von Filme, Musik zu Anfragen

Abstraktes Modell des Information Retrieval



Unterschiede zu klassischen Datenbanksystemen

- Formulierung der Anfrage zum Informationsbedürfnis schwierig, z.B. bei der Schlagwortsuche Homonym- und Synonym-Problem
  - Homonyme
  - Synonyme
- Anfrage liefert sehr viele Treffer, aber nur die wenigsten sind wirklich relevant
- Interne Repräsentation der durchsuchten Daten ist meist nicht optimal

Auftretende Probleme

- Relevanzbewertungen und Sortierung unterstützt den Nutzer
  - Beispiel: googles Erfolg durch relevante Such-Treffer an den ersten Stellen
- Formulierung der Anfrage wird typischerweise in einem iterativen Prozess verbessert
  - Nutzer erhält (schlechte) Ergebnisse und formuliert daher seine Frage neu
  - Nutzer lernt aus den Treffen mehr über sein Informationsbedürfnis
  - Interaktive Unterstützt durch das Informationsystem (wie Search Result Clustering, Faceted Search)

Beispiele für Suchanwendung

- allg. Internetsuche, wie google
- Spezialsuchen, wie Nachrichtensuche, Rezeptsuche, Kleinanzeigen, Produktsuche
- Unternehmenssuche

Angrenzende und verwandte Fachgebiete

- Computerlinguistik und Sprachtechnologien
- Wissensrepräsentation, symbolische KI, Logik
- Maschinelles Lernen - Data-Mining
- Softwareengineering, Verteilte Systeme
- Datenbanken

Outline

Daten, Information, Wissen

- Daten
  - unstrukturierte Daten (bits, bytes, chars)
  - strukturierte Daten - Syntax
- Information
  - Bedeutung verschiedener Datenelemente - Semantik
  - d.h. Interpretation der Daten
- Wissen
  - kontextspezifische Anwendung von Information - Pragmatik

Metainformation zur Strukturierung der Datenbestände

- Metainformation erleichtern das Finden der passenden Dokumente
- Metainformation zu Datenbeständen: Anzahl der Dokumente in einem Korpus, etc.
- Metainformation zu Dokumenten
  - Vorkommende Personen, Themen etc.

Verschiedene Retrievalmodelle

- Mengentheoretische Modelle
  - Boolean Retrieval
  - Fuzzy Retrieval
- Vektorraum-basierte Modelle
  - Boolean Retrieval
  - Fuzzy Retrieval
- Probabilistisches Retrieval
  - Boolean Retrieval
  - Fuzzy Retrieval


H. S. Christopher Manning, P. Raghavan.
Introduction to Information Retrieval.
Cambridge, 2008.


C. C. Shilakes and J. Tylman.
Enterprise information portals.
Merrill Lynch, 1998.