

3 Prinzipien des Information Retrieval

Einführung
Information-Retrieval-Modelle
Relevance Feedback
Bewertung von Retrieval-Systemen
Nutzerprofile
Literaturempfehlungen



Hinweise:

Andreas Henrich
Information Retrieval 1
Grundlagen, Modelle und Anwendungen
Version: 1.0 (Rev: 5697, Stand: 12. Dezember 2007)
Otto-Friedrich-Universität Bamberg

Bachelor-Vorlesung
„Information Retrieval“
Jetzt im Sommersemester
von Dr. Dr. Sergej Sizov

notwendig zur Suche von Multimedia-Objekten in Datenbanken
z.B. Bild mit Flusslandschaft
Besonderheit: Verwaltung von Daten anhand ihrer
Interpretation
Suche von relevanten Dokumenten
Information Retrieval versus Daten Retrieval
Informations-Retrieval-System (IRS)

Zugriff auf Daten – 2 Varianten:

1. Datenbankabfragen anhand exakt formulierter Bedingungen
SELECT isbn FROM Buch WHERE Titel = „Multimedia-Datenbanken“.
2. unscharfe Formulierung in IRS
Finde alle Text-Dokumente, die sich mit dem Thema „Multimedia-Datenbank“ beschäftigen.

Information-Retrieval-Systemen

Einsatz in Bibliotheken seit 70er Jahren
Stärke bei Verwaltung schwachstrukturierter Daten,
etwa Text-Dokumente
aufgrund Datenflut:
Problem des Findens geeigneter Informationen
Internet-Suchmaschinen

Grundproblem:

Inhalt von Dokumenten und Medien-Objekten oft nur
schwer anhand Datenbankschema strukturier- und
erschließbar

Lösungsansatz:

Ähnlichkeitssuche mittels IR-Techniken

Formulierung eines Suchbedarfs...

als Dokument (Ähnlichkeitssuche oder query by example)

- **Liefere alle Text-Dokumente, die ähnlich zum Text-Dokument #0815 sind**
- **Liefere alle Text-Dokumente die ähnlich zum Dokument „Urlaub Sommer Mittelmeer“ sind**

Formulierung eines Suchbedarfs...

- *als Anfrage (Eingrenzung durch Bedingung)*
Datenbank and (Bild or Video)

Im Folgenden umfasst der Begriff Anfrage beide Varianten

Information Retrieval
(außer Boole'sches Retrieval)

Daten sind unstrukturiert
Implizit formulierte Information
(erfordert Interpretation)

Suche nach Dokumenten, die ausreichend **wahrscheinlich relevant** bzgl. Anfrage sind

- ◆ Beispiel:
Suche von Text-Dokumenten anhand Text
→ Toleranz bzgl. Fehler bei
Anfrageformulierung

Daten versus Information Retrieval (3)

auch irrelevante Ergebnisse möglich
→ Anfrage-Iteration oft hilfreich

Ergebnisreihenfolge ist wesentlich

Einschränkung der Ergebnisgröße durch Schwellwert oder
Ergebnisanzahl

Daten versus Information Retrieval (4)

Merkmal	Daten Retrieval	Information Retrieval
Information	explizit	implizit
Ergebnisse	exakt	unscharf
Anfrage	einmalig	iterativ verfeinernd
Fehlertoleranz	keine	vorhanden
Ergebniskollektion	Menge	Liste

NB: kein direkter Vergleich zwischen Anfrage und Dokumenten

1. Überführung Anfrage und Dokumente in interne Darstellung

Extraktion von Daten, welche Semantik der Dokumente beschreiben

relevante Informationen explizit und kompakt z.B.:

- ♦ Zusammenfassungen von Texten

2. Vergleich der internen Darstellung durch Ähnlichkeitsfunktion

numerischer Relevanzwert drückt Ähnlichkeit aus

häufig Distanzfunktion

(das ist eine Unähnlichkeitsfunktion)

3. Ergebnis

Dokumente mit höchsten Relevanzwerten absteigend sortiert

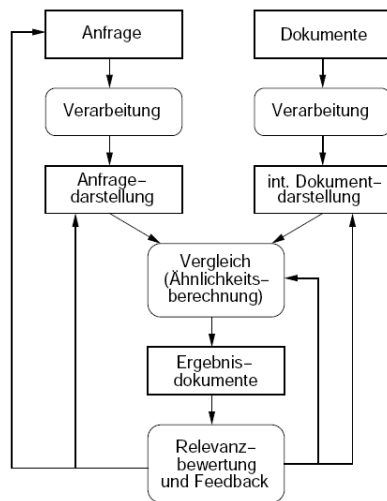
Einschränkung der Ergebnisliste durch Schwellwert bzw. Ergebnisanzahl

4. Relevanzbewertung und Feedback, wenn Ergebnis nicht zufriedenstellend

Anfrage-Iteration

Modifikation der Anfrage

Modifikation der internen Darstellung



Modifikation der Anfrage: Beispiel

Anfrage

Finde alle Text-Dokumente, die sich mit dem Thema „Multimedia-Datenbanken“ beschäftigen.

Modifikation der Anfrage: Beispiel (2)

Verfeinerung:

Finde alle Text-Dokumente, die sich mit dem Thema „Bild-Datenbanken“ beschäftigen.

Modifikation der Anfrage: Beispiel (3)

Verallgemeinerung:

Finde alle Text-Dokumente, die sich mit dem Thema „Datenbanken“ beschäftigen

automatische Anfragemodifikation
Nutzer bewertet Ergebnisdokumente bzgl. Relevanz zur
Anfrage
System nutzt Bewertungen zur automatischen
Anfragemodifikation
Anfrage-Iteration

Extraktionsverfahren erzeugen interne, kompakte
Dokumentdarstellung

Verfahren abhängig:

- ♦ vom Typ des Dokuments z.B.
Text-Dokument versus Audio-Dokument
- ♦ von der Art beabsichtigter Anfragen
in Bild-DB:
Ähnlichkeitssuche über Farbverteilung versus über Textur

3.2 Information-Retrieval-Modelle

IR-Modell definiert
interne Dokumentdarstellung,
Anfrageformulierung und interne Anfragedarstellung,
Vergleichsfunktion zwischen jeweils zwei Dokumenten
beziehungsweise zwischen Anfrage und jeweils einem
Dokument.

Text-Retrieval

Modelle ursprünglich für Text-Retrieval
Existenz einer vordefinierten Menge von Indextermen
(Indexierungsvokabular)
Darstellung eines Dokumentes über auftretende Indexterme
verschiedene Arten von Termgewichten
Modelle lassen sich auch auf andere Medien-Typen anwenden

Boole'sches Modell
Fuzzy-Modell
Vektorraummodell

basiert auf Mengentheorie und Boole'scher Algebra
sehr einfaches Modell mit klarer Semantik
Dokumente als Mengen von Indextermen → Termgewichte
sind binär: im Dokument enthalten oder nicht enthalten
Test auf Enthaltensein als Vergleichsfunktion

Verknüpfung von Enthaltenseinsbedingungen mittels
Boole'scher Junktoren

- ♦ *and* (Mengendurchschnitt)
- ♦ *or* (Mengenvereinigung)
- ♦ *not* (Mengendifferenz)

Indexvokabular {Korsika, Sardinien, Strand,
Ferienwohnung, Gebirge}

Dokument d1:{Sardinien, Strand, Ferienwohnung}

Dokument d2:{Korsika, Strand, Ferienwohnung}

Dokument d3:{Korsika, Gebirge}

Anfrage

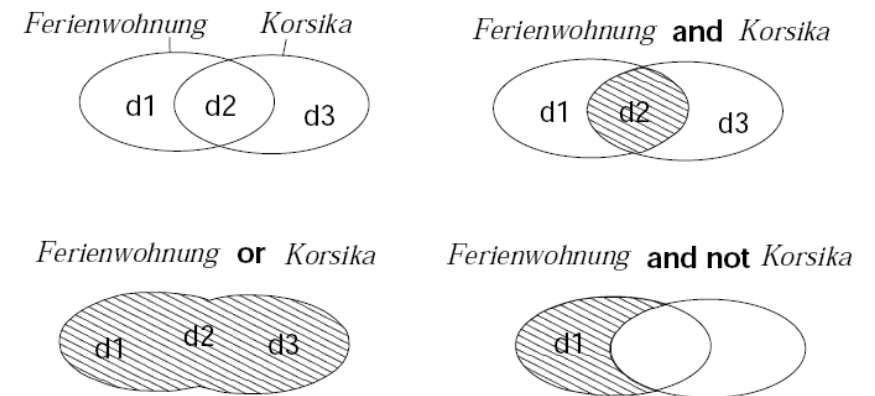
Korsika
liefert {d2, d3}

Ferienwohnung
liefert {d1, d2}

Ferienwohnung *and* Korsika liefert {d2}

Ferienwohnung *or* Korsika liefert {d1, d2, d3}

Ferienwohnung *and not* Korsika liefert {d1}



but-Junktor

Anfrage: *not* Korsika

liefert alle Dokumente, die „Korsika“ nicht enthalten, u.U. die ganze Dokumentkollektion

statt dessen Verwendung but-Junktor

(entspricht and not, also Komplement bzgl. Vorauswahl)

Beispiel: Ferienwohnung but Korsika

of-Konstrukt

Suche nach Dokumenten, die m von n ($m < n$) Termen enthalten

2 of (Korsika, Strand, Ferienwohnung)

ersetzt komplexen Boole'schen Ausdruck

(Korsika *and* Strand) *or* (Korsika *and* Ferienwohnung) *or* (Strand *and* Ferienwohnung)

$$\left(\frac{n!}{(n-m)! * m!} \text{ and-Kombinationen} \right)$$

Anfragenormalisierung in DNF bzw. KNF

Anfrage:

Ferienwohnung and ((Sardinien and Strand) or Korsika)

in disjunktiver Normalform (DNF):

(Ferienwohnung and Sardinien and Strand) or
(Ferienwohnung and Korsika)

in konjunktiver Normalform (KNF):

Ferienwohnung and (Sardinien or Korsika) and (Strand or
Korsika)

Anfrageauswertung

jeder Term liefert Menge von Dokumenten, die diesen
Term enthalten

komplexe Anfrage: Kombination der Dokumentenmengen
durch entsprechende Mengenoperationen

kleine Zwischenergebnisse durch DNF (zuerst
Durchschnitt, dann Vereinigung)

Nachteile des Boole'schen Modells

exaktes Modell aufgrund binärer Gewichte

- ◆ eher Daten-Retrieval
- ◆ keine Ähnlichkeitssuche durch zu scharfe Suche

Größe des Ergebnisses

- ◆ alle Dokumente sind bzgl. Anfrage gleichrangig →
Präsentation der gesamten Ergebnismenge
- ◆ Ergebnismenge in Abhängigkeit von Anfrage oft zu
groß oder zu klein bzw. leer

Nachteile des Boole'schen Modells (2)

Boole'sche Junktoren

- ◆ schwierige Anwendung Boole'scher Junktoren
- ◆ Verwechslung mit „und“, „oder“ und „nicht“

⇒ Impliziter weiterer Gegensatz:
IR wird stärker von Laien benutzt als SQL

verschiedene Stufen der Relevanz durch Überführung
Konjunktion in Disjunktion
Präsentation der Ergebnisse sortiert nach Relevanzstufen
Beispiel: Umwandlung von
„Korsika and Strand“ → „Korsika or Strand“

Ergebnis: zuerst die and-({d2}), dann restliche or-
Dokumente ({d1, d3})

faceted query

zweistufiges Suchverfahren

1. Formulierung und Verfeinerung der Anfrage anhand
benannter Anfragen und Ergebnisanzahl
2. Ergebnis zur finalen Anfrage anzeigen

Beispiel:

- 1 Korsika liefert Q1: 1345
- 1 Q1 and Strand liefert Q2: 13
- 2 Anzeige Q2 liefert die 13 Ergebnisdokumente

Anwender sucht Dokumente über Korsika und Dokumente
über Sardinien

"falsche" d.h. nicht intendierte Anfrage:

Korsika and Sardinien

liefert Dokumente, in denen beide Terme gemeinsam
auftreten

"richtige" bzw. intendierte Anfrage:

Korsika or Sardinien

Verwendung besserer Junktorenbegriffe, etwa Ersetzen von „and“ durch „all“
 „or“ durch „any“

Erweiterung des Boole'schen Modells um Unschärfe (fuzzy)
 Verallgemeinerung Boole'scher Junktoren
 Unschärfe durch graduelle Zugehörigkeit von Dokumenten zu Termen

Definition:

Eine Fuzzy-Menge $A = \{ \langle u; \mu_A(u) \rangle \}$ über einem Universum U ist durch eine Zugehörigkeitsfunktion $\mu_A : U \rightarrow [0, 1]$ charakterisiert, welche jedem Element u des Universums U einen Wert $\mu_A(u)$ aus dem Intervall $[0, 1]$ zuordnet.

Fuzzy-Mengen beim Information Retrieval

Universum ist Menge aller gespeicherten Dokumente
 Term definiert Fuzzy-Menge
 Zugehörigkeit (Fuzzy-Wert) des Dokuments d zu Term t durch Wert

- ♦ **0 für keine Relevanz**
- ♦ **1 für maximale Relevanz** $\mu_t(d)$
- ♦ **Wert zwischen 0 und 1 für graduelle Relevanz**

Beispiel

Universum umfasst 3 Dokumente {d1, d2, d3}
 Fuzzy-Mengen Korsika bzw. Strand drücken Zugehörigkeit zu Term „Korsika“ bzw. „Strand“ aus

$$\begin{aligned} \text{Korsika} &= \{ \langle d1; 0,1 \rangle, \langle d2; 0,6 \rangle, \langle d3; 1 \rangle \} \\ \text{Strand} &= \{ \langle d1; 0,3 \rangle, \langle d2; 0,2 \rangle, \langle d3; 0,8 \rangle \} \end{aligned}$$

μ	$d1$	$d2$	$d3$
μ_{Korsika}	0,1	0,6	1
μ_{Strand}	0,3	0,2	0,8

jedes Dokument in jeder Fuzzy-Menge
 → übliche Mengenoperationen nicht anwendbar
 Junktoren ermitteln neue Zugehörigkeitswerte
 and durch Min-Funktion
 or durch Max-Funktion

not durch Subtraktion von 1:

Konjunktion **and**: $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$
 Disjunktion **or**: $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
 Negation **not**: $\mu_{\bar{A}}(u) = 1 - \mu_A(u)$

Korsika and Strand
 Korsika or Strand
 not Korsika

Anfrage	μ	d1	d2	d3
1	μ_{Korsika}	0,1	0,6	1
	μ_{Strand}	0,3	0,2	0,8
2	$\mu_{\text{Korsika} \cap \text{Strand}}$	0,1	0,2	0,8
3	$\mu_{\text{Korsika} \cup \text{Strand}}$	0,3	0,6	1
	$\mu_{\overline{\text{Korsika}}}$	0,9	0,4	0

Fuzzy-Anfragen und Anfragebearbeitung

Überführung Anfrage in disjunktive Normalform
 jeder Suchterm induziert eine Fuzzy-Menge
 Anwendung entsprechender Fuzzy-Operationen auf Fuzzy-Mengen
 Ergebnis: Dokumente absteigend sortiert nach Zugehörigkeitsgrad ausgeben

Begrenzung Fuzzy-Ähnlichkeitsanfrage

Ergebnis umfasst alle Dokumente des Universums

Begrenzung der Anzahl durch

- ♦ Schwellwerte für Zugehörigkeitswerte
- ♦ vorgegebene Anzahl von Ergebnisdokumenten

Begrenzung Fuzzy-Ähnlichkeitsanfrage (2)

Beispiel:

Korsika and Strand

Schwellwert 0,5 liefert d3

Anzahl 2 liefert d2, d3