Grundbegriffe des Information Retrieval

Alexandra Bünzli

11.04.2001

1 Allgemeines

1.1 Motivation

- Datenmenge wächst
- Immer mehr Menschen haben Zugang zu diesen Daten
- Nutzen der Daten ist nur gewährleistet, wenn sie erschlossen sind

1.2 Anwendung

- schon länger im Literatur- und Bibliothekswesen (Ursprung)
- heute auch vor allem im WWW

1.3 Probleme

- Nutzer können ihren Informationsbedarf meist nur vage formulieren
- der Informationsbedarf kann sich im Verlauf der Anfrage auch verändern

2 Konzepte und Wissensrepräsentation

2.1 Zwei grundsätzliche Konzepte

2.1.1 Freitextsuche

Keine neue Repräsentation, sondern Verbesserung der Suche im Text der Dokumente: Vorbereitung des Textes durch Zerlegung in einzelne Wörter, Stoppworteliminierung (Stoppwörter sind häufige, nicht sinntragende Wörter wie Artikel, Präpositionen etc.) und Satzenderkennung. Im so reduzierten Text stellen sich vor allem Probleme mit Homographen, Polysemen, Flexions- und Derivationsformen sowie Komposita. Man versucht sie mit folgenden Ansätzen in den Griff zu bekommen:

- informatischer Ansatz (Funktionen auf Zeichenkettenebene)
- computerlinguistischer Ansatz (morphologische und syntaktische Verfahren → Suche nach Wörtern)

2.1.2 Semantischer Ansatz

Neue Repräsentation in einer Dokumentationssprache: Es wird weitgehend unabhängig vom Text formuliert. Beispiele sind Klassifikationen und Thesauri.

2.2 Erschliessungsmethoden

2.2.1 Grundformenreduktion

Informatischer Ansatz
In den folgenden Beispielen wird für beschränkte Trunkation (das heisst ein Operatorsymbol steht für genau ein Zeichen im Wort) das Symbol $, für die unbeschränkte Trunkation (das Operatorsymbol steht für eine beliebig lange Zeichenkette) das Symbol # verwendet.

- Trunkierung (truncation):

  schreib#: schreiben, schreibt, schreibst, schreibe
  schreib$$: schreiben, schreibst
  #schreiben: schreiben, beschreiben, anschreiben, verschreiben
  $$schreiben: beschreiben, anschreiben

- Maskierung

  schr$$b#: schreiben, schrieb / schrauben
  h$$s#: Haus, Häuser / Hanse, Hass, hassen, hausen

- Kontextoperatoren

  - genauer Wortabstand ($):
    retrieval $ information: retrieval of information, retrieval with information loss
  - maximaler Wortabstand (#):
    information # # retrieval: information retrieval, information storage and retrieval
  - Wortreihenfolge (,):
    information #, retrieval: informaton retrieval, retrieval of information
  - gleicher Satz (.):
    information # retrieval.: matcht weder '...this information. Retrieval of data...' noch '...storage of information. Its retrieval...'.

CL-Ansatz

- Wortstamm oder Grundform als Indexterm (durch stemming, Lemmatisierung)
- Musterbasierte Ersetzungsregeln: Ende eines Wortes wird nacheinander mit verschiedenen Zeichenmustern verglichen (Flexion, Derivation) → sobald eine Übereinstimmung stattfindet, wird es durch eine in der Regel angegebene Zeichenkette ersetzt. (Erfolgreich bei Sprachen, in denen sich die Worte bei der Bildung der verschiedenen Formen nur wenig ändern, z.B. Englisch, nicht aber Deutsch.)
- Lexikonbasierte Reduktionsmethoden: Die notwendige Information zur Bildung der verschiedenen Formen wird in Wörterbüchern gespeichert. Syntaktische Verfahren dienen v.a. zur Identifikation von Komposita (robuste und wörterbuchunabhängige Parser ideal) Dokumentationssprachen

2.2.2 Klassifikation

Strukturieren der Themen und Objekte nach einem Schema

- Hierarchische Klassifikation (Baumstrukturen)
- Internationale Dezimalklassifikation: sachgebietsorientierter Zugang zu den Wissensgebieten. Es sind immer maximal zehn Verzweigungen möglich, ausgehend von folgenden zehn Hauptabteilungen:

0	Allgemeines
1	Philosophie
2	Religion, Theologie
3	Sozialwissenschaften, Recht, Verwaltung
4	(Zur Zeit nicht belegt)
5	Mathematik, Naturwissenschaften
6	Angewandte Wissenschaften, Medizin, Technik
7	Kunst, Kunstgewerbe, Photographie, Musik, Spiel, Sport
8	Sprachwissenschaft, Philologie, Schöne Literatur, Literaturwissenschaft
9	Heimatkunde, Geographie, Biographien, Geschichte

Nachteile

- frühe Entschiedungen beim Suchprozess nötig
- lassen sich nur schwer an neue Entwicklungen anpassen, nicht flexibel → z.T. darum zusätzliche Verweise zwischen den Klassen möglich (Doppelstellen und Polyhierarchien) oder/und Sachverhalte können zum Zeitpunkt der Klassifizierung von Objekten zusätzlich genauer beschrieben werden (postkoordinierende Verfahren)

2.2.3 Thesauri

Thesauri sind geordnete Zusammenstellungen von Begriffen eines Fachgebiets mit ihren (natürlichsprachlichen) Beziehungen (Relationen, relations). Hierarchische Strukturen werden durch Beziehungen erzeugt, Beziehungen wie allgemeiner (Oberbegriff, broader Term), spezieller (Unterbegriff, narrower Term), synonym, verwandt (related), Gegensatz (antonym) etc.
Terminologische Kontrolle durch:

- Erfassung von Synonymen
- Kennzeichnung von Homographen und Polysemen
- Festlegung von Vorzugsbenennungen → Deskriptoren

Thesauri werden verwendet um die Dokumente einer Sammlung möglichst eindeutig mit Fachbegriffen zu indexieren. Kernstück ist ein kontrolliertes Vokabular, das die sogenannten Deskriptoren bilden. Deskriptoren sind die im Thesaurus zugelassenen ausgewählten Terme, die eine klar definierte Bedeutung im zu beschreibenden Sachverhalt haben. Es ist auch eine Menge von Synonymen definiert, die zulässt, dass auch Terme, die im Thesaurus nicht zugelassen sind, aber eine gleiche oder ähnliche Bedeutung haben wie ein Deskriptor, gefunden werden können.

3 Suchmethoden

3.1 Nicht-probabilistische IR-Modelle

3.1.1 Boolesches Retrieval (boolesche Suche, boolean retrieval)

Idee Mengenoperationen auf Mengen von Dokumenten anwenden, die durch das Auftreten von bestimmten Wörtern (Termen) charakterisiert sind

Methoden

- Elimination von Stoppwörtern (häufige, nicht sinntragende Wörter wie Artikel, Präpositionen...→ in Stoppwortliste gespeichert)
- Indexieren des Dokuments mit intellektuell vergebenen Termen oder automatischen Verfahren
- Anwendung von Booleschen Operatoren (AND, OR, NOT)

- Erweiterungen mit Operatoren wie NEAR (Suche nach Wortgruppen, deren Wörter aufeinander folgen müssen oder einen angegebenen Abstand maximal einhalten müssen) und Platzhaltern (wild cards) für Buchstaben oder Zeichenketten, sowie mit speziellen Feldern für Zahlen und entsprechende Vergleichsoperatoren ($>, <$)

Nachteile

- keine Gewichtung, keine Rangordnung
- Ergebnismenge schwer zu kontrollieren
- Umständliche Frageformulierung → Überforderung der Benutzer
- Schlechteste Retrievalqualität aller bekannten Verfahren (v.a. wegen der Nichtberücksichtigung der Unsicherheit der Textrepräsentation und Vagheit der Anfragen), trotzdem am häufigsten gebraucht!

3.1.2 Vektorraummodell (Vector-Space-Model)

Idee Dokumente werden durch Vektoren repräsentiert, die aus den Termgewichten ("Wichtigkeit" eines Terms zur Beschreibung des jeweiligen Dokuments, damit es gut gefunden werden kann) hervorgehen. Anfragen werden auch durch einen Vektor definiert. Durch den Abstand dieser Vektoren kann ein Ähnlichkeitsmass (Berechnung durch Skalarprodukt oder Kosinusmass etc.) definiert werden, um die Dokumente zu vergleichen → eine Rangfolge wird hergestellt: Die ähnlichsten Dokumente zur Anfrage befinden sich ganz oben in der Liste.

Gewichtung

- lokale Einflüsse
  - Häufigkeit des Terms im Dokument
  - Häufigkeit des Terms in Relation zum häufigsten Term im Dokument
  - bei Dokumenten mit Struktur: Gewichtung nach Ort des Vorkommens (z.B. wird ein Wort im Titelfeld als wichtiger eingestuft)
  - Gewichtung nach Position: Wichtige Terme eher am Anfang des Dokuments (z.B. Nachrichtenmeldungen haben immer den gleichen Aufbau: 1. wesentliche Neuigkeit, 2. Hintergründe, 3. Details und Kommentare)
- globale Einflüsse
  - Dokumenthäufigkeit eines Terms (Terme, die in vielen Dokumenten vorkommen sind nicht geeignet zur Beschreibung → werden als Indexterme schwach gewertet)

  In der Regel werden lokale und globale Kriterien bei der Gewichtung kombiniert, woraus dann die verwendete Gewichtung des Terms resultiert.
- Relevanzfeedback (Relevanzrückmeldung, relevance feedback): In kurzen Anfragen sind Termhäufigkeiten meist wenig sinnvoll → in interaktiven Sitzungen muss der Benutzer nach der Anfrage die gefundenen Dokumente als relevant oder nicht relevant kennzeichnen, woraus ein neuer Anfrage-Vektor berechnet wird.

3.2 Probabilistische IR-Modelle

Idee Probabilistische Modelle versuchen das Problem der Unsicherheit in der Textrepäsentation besser in den Griff zu bekommen. Es wird ein Mass für die Bedeutsamkeit eines Begriffs verwendet, das der Wahrscheinlichkeit entspricht, mit der dieser Begriff in einer bestimmten Umgebung vorkommt, oder mit der dieser Begriff einem Dokument als Deskriptor zugeordnet werden sollte. Dieses Mass lässt sich als Deskriptorgewicht benutzen, so dass sich die Dokumente in einer Rangfolge nachweisen lassen, die der Wahrscheinlichkeit der Relevanz bzgl. bestimmter Suchanfragen entspricht.

3.3 Evaluierung

Ein ganzes System empirisch zu evaluieren ist nicht möglich, zu viele Faktoren müssten berücksichtigt werden → Teilaspekte werden untersucht. (Man muss sich im Klaren sein, welchen Standpunkt man einnehmen will und ob er geeignet ist für die jeweilige Anwendung.)

3.3.1 Standardevaluationsmethode

Evaluierung mit den Qualitätsmassen Precision (Präzision) und Recall (Abdeckung): Man untersucht den Teilaspekt der Anzahl der gefunden relevanten Dokumente

- Precision: Anteil der relevanten Dokumente in der gefundenen Ergebnismenge
- Recall: Anteil der gefundenen relevanten Dokumente von allen vorhandenen relevanten Dokumenten (Problem: Die Bestimmung aller im System vorhandenen Dokumente, die ja von Menschen gemacht werden muss → sehr aufwändig bei grossen Sammlungen (Die TREC stellt grosse Testsammlungen zur Evaluierung zur Verfügung)

Gewünscht wäre ein hoher Recall und eine gute Präzision. Die beiden Masse sind in der Regel jedoch gegenläufig: wächst der Recall, nimmt die Präzision ab und umgekehrt. Man muss daher oftmals entscheiden, welches Mass für die jeweilige Aufgabe wichtiger ist.