Kapitel 6 – Suchen und Gewinnen von Informationen

Prof. Dr.-Ing. Stefan Deßloch
AG Heterogene Informationssysteme
Geb. 36, Raum 329
Tel. 0631/205 3275
dessloch@informatik.uni-kl.de

Suchen und Gewinnen von Informationen

- Überblick über Such- und Anfragetechniken
- Werkzeuge für eine Digitale Bibliothek
- Text Information Retrieval
- Inhaltsbasierte Suche in Multimedia-Daten

Überblick über Such- und Anfragetechniken

- Anfragen: an Metadaten; strukturiert, mächtige Operationen
- Suche (Retrieval): in Metadaten als unstrukturierte Informationen oder in Volltexten/ Multimedia-Inhalten („inhaltsbasiert")
- Kombination von Anfrage und Suche
- Informationsgewinnung
- Formulierung der Anfrage / Suche; Auswahl der Werkzeuge
  - Iterativer Prozeß
  - Ranking / Relevance Feedback
  - Verwendung der Ergebnisse
- Alternativen zu Anfrage / Suche: Navigation, Browsing, Agententechnik

Anfrage- und Retrievalsprachen

- SQL
- in Bibliotheken: teilweise Spezialsprachen, einfacher und älter als SQL
- oft keine relationalen Datenbanken, sondern invertierte Dateien
- Retrieval-Sprachen
  - NEAR
  - Phrasen
  - Wortstammreduktion
  - Soundex-Funktion
  - Ranking, Relevance Feedback

Hilfsmittel zur Suche

- Lexikon
  - Tippfehlerkorrektur
  - Standardbegriffe
  - Mehrsprachigkeit
  - ...
- Thesaurus
  - Unter- und Oberbegriffe
  - Synonyme
  - in Beziehung stehende Begriffe
- Klassifikationssysteme
  - Informatik: ACM-Klassifikationssystem

Werkzeuge für eine Digitale Bibliothek

- Anfragen an Metadaten
- Suche in Metadaten
- Finden: Ort des Dokuments oder Dokument selbst
- Inhaltsbasierte Suche (etwa über Index, der oft doppelt so groß ist wie der Dokumentbestand)
- Fuzzy-Suche, konzeptbasierte Suche (Thesaurus, Klassifikationssysteme)
- Anfragen an Fakten-Datenbanken (Biotechnologie, Medizin, Chemie)

Alternativen zum Suchen

- Navigation über Klassifikationssystem (in Verbindung mit Suchfunktionen auf jeder Ebene)
- Browsing über Web-Browser, Hypertext-Werkzeuge (Dokument zu zitiertem Dokument, Band zu Artikel, Konferenz zu Artikel, ...)
- Software-Agenten

Suchen im Internet

- Suchmaschinen (Gatherer / Broker)
- Meta-Suchmaschinen
  - Parallele Suche
  - Mischen von Ergebnissen
  - Duplikateliminierung
  - kein Informationsverlust
  - Kapselung der Teilsysteme
- Internet-Kataloge
  - Ariadne von FIZ Karlsruhe für Informatik-Literatur
  - Yahoo allgemein
- Portale
  - diverse Daten
  - diverse Dienste
  - Personalisierung
  - News-Ticker / Delta-Funktion

Kombination von Suchangeboten in einer Digitalen Bibliothek

- Suche und Anfragen
  - in Bestandskatalogen (OPAC)
  - in lokalen und entfernten Nachweissystemen (DBLP)
  - mit lokalen und globalen Suchmaschinen
  - Internet-Kataloge und -Portale
- nötig: Transformation von Anfragen
- Sortieren und Mischen von Ergebnissen
- iteratives Suchen
- Navigieren in Ergebnisdokumenten

Selektion im Überangebot

- Lese-, Hör- und Sehproben
- Kommentare / Gutachten
- Browsen in Umgebung von (engen) Suchergebnissen

Bestellen und Lokalisieren von digitalen Dokumenten

- in lokaler Bibliothek (OPAC)
- Fernleihe (oft OPAC)
- Verlag (Nachweis-Datenbank)
- elektronisch (DOI, URL) oder Papier
- Aufnahme in digitale Handbibliothek

Text Information Retrieval

- Dokumente in bestimmten Text-Format: Information Retrieval muß diese Formate „verstehen"
- Arten von Formaten
  - Low-level, Druckseiten-orientiert, schwer weiterverarbeitbar (Postscript, PDF, PCL)
  - Middle-level, Layout-orientiert, aber WYSIWYG (Word, andere Textverarbeitungen und DTP-Programme)
  - High-level, logische Struktur, Markup-Sprache (LaTeX, SGML, XML)
- Arten von Zeichen
  - ASCII (7-bit)
  - EBCDIC (8-bit)
  - Unicode (16-bit)
- Textsuche
  - Verwendete Techniken: reguläre Ausdrücke, Boyer-Moore, Invertierte Dateien (Index mit Position), Hashen, Tries, Signatur-Indizes, Thesauri, OCR: Bild-nach-Text-Wandlung)

Information-Retrieval-Systeme

- Grundtechniken
  - Deskribierung
  - Recherche
  - Bewertung
- Systeme
  - STAIRS (IBM)
  - PASSAT/GOLEM2 (Siemens)
  - Fulcrum
  - ORDBMS mit Text-Erweiterung
  - SQL-99/MM/Text

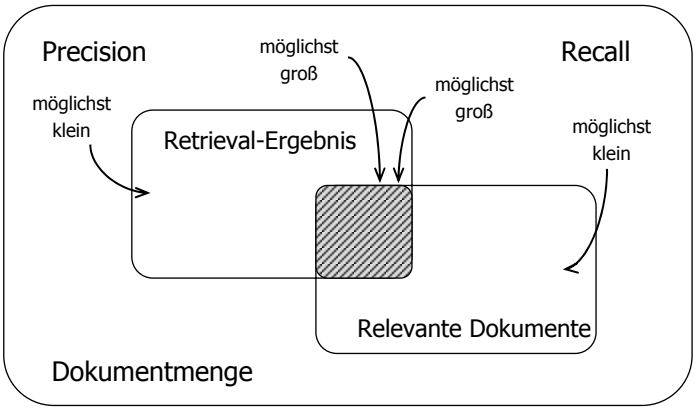
Historie, Motivation

- Information-Retrieval-Systeme für Texte oder unstrukturierte Daten
- älter als Datenbanksysteme!
- Grundtechniken: schon bis Anfang der siebziger Jahre bekannt und teilweise eingesetzt
- Techniken, die etwa in GOLEM2 benutzt werden,
  - werden heutzutage für Suchmaschinen im WWW wieder neu erfunden
  - wurden in relationalen Datenbanken (leider) zunächst vergessen (1NF)
  - werden heutzutage in objektrelationalen Systemen für den Datentyp „Volltext" wieder neu erfunden

Grundprinzipien und Ziel

- Recall = $\frac{\text{Anzahl gefundener relevanter Dokumente}}{\text{Gesamtanzahl relevanter Dokumente}}$
- Precision = $\frac{\text{Anzahl gefundener relevanter Dokumente}}{\text{Gesamtanzahl gefundener Dokumente}}$
- Fallout = $\frac{\text{Anzahl gefundener irrelevanter Dokumente}}{\text{Gesamtanzahl irrelevanter Dokumente}}$

Precision und Recall



Art der Verfahren

alle sprachabhängig!

- statistisch und wortbasiert (häufig verwendet): Indizierung
- linguistisch: Satzstruktur, Worten werden Rollen im Satz zugeordnet, abhängig vom Anwendungsbereich
- wissensbasiert: Ontologie zur Strukturierung des Anwendungsgebietes

hier: statistisch und wortbasiert

Grundtechniken

- Deskribierung
  - (manuell) Klassifizierung
  - (manuell) Indizierung
  - (automatisch) Stichwortverfahren
  - (automatisch) Morphologische Reduktion
  - (automatisch) Inhaltserschließung
- Recherche
  - Grundmuster
  - Dokument- und Recherche-Deskriptoren
  - Retrieval-Sprachen
  - Retrieval-Modelle
  - Zugriffsstrukturen
- Bewertung
  - Ranking
  - Interaktion und Relevance Feedback

Deskribierung

- Transformation Dokument ⟶ Dokumentbeschreibung
  - dokumenttypspezifische (attributierte) Struktur (Metadaten): etwa bibliographische Angaben aus Zeitschriftenartikel (hier nicht betrachtet)
  - Deskriptoren
- Forderung: Dokumente gleichen Inhalts bekommen gleiche Deskriptoren
- manuelle und automatische Verfahren

Manuelle Deskribierung: Klassifizierung

- Dokumente ⟶ fest vorgegebenes System von Dokumentklassen
- hierarchische: etwa Dezimalklassifikation
- Facetten: aus n Deskriptormengen n Deskriptoren wählen und zur Klasse konkatenieren

Manuelle Deskribierung: Indizierung

- Deskriptoren sind Worte aus dem Text
  - Stichwortverfahren (Glossar, Positivliste)
- Deskriptoren sind nicht Worte aus dem Text
  - Schlagwortverfahren
- Deskriptoren häufig aus kontrolliertem Wortschatz
  - Thesaurus (Vorzugsbenennungen, Querverweise, Ober- und Unterbegriffe, Used For, Related Terms, Synonyme, Antonyme)
- Deskriptoren auch attributiert (Aspekte, Rollen)
  - Name: Müller, Beruf: Müller; Name: Bodden, Landschaft: Bodden
- Deskriptoren auch gewichtet

Automatische Deskribierung: Stichwortverfahren

- Alle Worte im Text, außer Elemente der Stopwortliste (Negativliste)
- häufigste Worte streichen
- seltenste Worte streichen (trotz maximaler Selektivität)

Automatische Deskribierung: Morphologische Reduktion

- Flexionsformen
  - Deklination Substantive, Adjektive
  - Konjugation Verben
  - Komposition
- als Deskriptor nur Grundform (Lexem und Hauptmorphem), davon getrennt weitere Flexionsformen (Morpheme)
- Flexionsklassen = Menge von Morphemen; dann Wörterbuch von Lexemen in Flexionsklassen zerlegen
- etwa PASSAT: > 100 Flexionsklassen
  - Klasse i: ␣ , S, ES, E, ER, ERN (GELD, ...)
  - Klasse j: ␣ , ES, S (HAUS, ...)
  - Klasse k: ␣ , N (HÄUSER, ...)
  - Klasse l: EN, E, T, EST, ET, TE, TEST, TET, TEN, END, ENDE, ... (GRÜSS, ...)

  erste Kombination ist Grundform
- etwa PASSAT: > 100 Kombinationsklassen
  - Klasse m: X---, X---S, X---EN (FAHRT, ...)

    Fahrtroute, Schifffahrtsgesellschaft, Fahrtenbuch

Automatische Deskribierung: Klassifizierung

- Ähnlichkeitsmaße definieren
- Cluster bilden, deren Elemente bestimmten Ähnlichkeitsgrad aufweisen
- Klassifikationssystem (Cluster) hier deshalb dynamisch
- Literatur: Salton (Information Retrieval)

Recherche: Grundmuster

- *Einfache Terme*:
  - `[ `Datenbanken'] **in** Text`
- *Konjunktive Anfragen*:
  - `[ `Datenbanken' **and** `Multimedia' ] **in** Text`

  auch Disjunktionen und Negationen
- Nachbarschaft von Suchworten:
  - `[ `Objekt' 1 **Wort vor** `Orientierung' ] **in** Text,`
  - `[ `Objekt' **im gleichen Satz mit** `Orientierung' ] **in** Text,`
  - `[ `Objekt' **innerhalb** 2 **Abschnitte mit** `Orientierung' ] **in** Text`

Dimensionen von Recherche-Operationen

- Abbildung der Deskriptoren in der Recherche auf die Dokumentdeskriptoren
- Retrieval-Sprachen (etwa Boolesche Recherchen, attributierte Recherchen, Recherchen mit Wichtungen, ...)
- Retrieval-Modelle (Art der Abarbeitung der Recherche; etwa Boolesches Retrieval, Abarbeitung nach dem Vektorraummodell, probabilistisches Retrieval, ...)
- Retrieval-Unterstützung durch Indexstrukturen

Recherche: Dokumentdeskriptoren

- Zuordnung Dokument- zu Frage-Deskriptoren
- Identische Abbildung
  - einfach, aber wenig attraktiv
  - geringe Wahrscheinlichkeit für Übereinstimmung von beiden Deskriptoren
- Einschränkung auf kontrollierten Wortschatz
  - Fragedeskriptoren auf Wortschatz der Dokumentdeskriptoren beschränken
- Erweiterung um verwandte Wörter
  - Fragedeskriptoren ergänzen um verwandte Deskriptoren
  - etwa unter Verwendung eines Thesaurus
  - verbreitet: zumindest Synonyme ergänzen

Recherche: Retrieval-Sprachen (1)

- Angabe eines Deskriptors
- Angabe einer Deskriptormenge: durch Komma getrennt
- Boolesche Recherchen: Deskriptoren durch **und**, **oder** und **nicht** verknüpft
- Kontext-Recherchen:
  - absolute Ortsangabe (in einem bestimmten Kapitel) oder relative Ortsangabe (dieser Deskriptor drei Abschnitte vor dem anderen Deskriptor);
    - Index muß Ortsinformationen (Konkordanzen) berücksichtigen.
  - Üblich: gleicher Satz, gleiche Struktureinheit (wie Abschnitt, Kapitel), Abstand zum Satz oder zur Struktureinheit}, vorgegebener Abstand zwischen Deskriptoren, etwa auch in Anzahl von Worten
  - einfachste Form: Phrase

    'Sein oder nicht Sein, das ist hier die Phrase'
- Gewichtete Recherchen: jedem Deskriptor einen Wichtungsfaktor mitgeben

  Hotel:0.8 **and** Ostsee:0.5 **and** Strandkorb:0.2
- Suche nach Mustern: exakt dieses Wort, nur Präfix, beliebige Teilzeichenkette, Groß- und Kleinschreibung unberücksichtigt, Zulassen von Fehlern in einem Suchbegriff, reguläre Ausdrücke

Recherche: Retrieval-Sprachen (2)

- Komplexere Recherchen: Strukturierung des Dokumentes berücksichtigen
  - Attributierung (falls diese in der Struktur des Dokumentes durch eine semantische Auszeichnung bekannt ist, wie etwa in XML)
    Bsp.: Lage: Ostsee und Ausstattung: Strandkorb
- im Textdokument solche Attribute nicht ausgezeichnet:
  - durch manuelle Deskribierung erfassen oder
  - durch automatische Deskribierungsverfahren ermitteln (etwa oben erwähnte Verfahren zur Inhaltserschließung oder „Structure Mining", Wrapper-Generatoren)

Recherche: Retrieval-Modelle

- Boolesches Retrieval:
  - true, wenn Deskriptoren im Dokument vorkommen, und false, wenn sie nicht vorkommen
  - alle Dokumente im Ergebnis mit gleicher Relevanz
- Vektorraummodell:
  - Recherche- und Dokument-Deskriptoren als Vektor im mehrdimensionalen Raum: mit Ähnlichkeitsmaß Dokumente mit „benachbarten," Vektoren als Ergebnis zurückgeben: vages Ergebnis
- Probabilistisches Modell:
  - Wahrscheinlichkeiten berechnen, ob Dokument relevant ist oder nicht

Recherche: Zugriffsstrukturen

- Invertierte Listen:
  - indizierte Worte bilden lexikographisch sortierte Liste
  - einzelner Eintrag besteht aus Wort und Liste von Dokument-Identifikatoren
  - zusätzlich: Position des (ersten Auftretens des) Wortes im Text, Häufigkeit des Wortes im Text
- Linguistischer Index:
  - durch Stemming entsteht Index, der unabhängig von der Wortform ist
- Konzept-Index:
  - Berücksichtigung von Synonymen, Einsatz von Thesauri, ...
- Signatur-Index:
  - Signatur eines Dokumentes basierend auf allen (relevanten) Worten (Hash-Wert)
  - mehrere Dokumente können gleiche Signatur haben

Bewertung

- gefundene Dokumente in der Reihenfolge ihrer Relevanz ausgeben
  - Relevanz ermittelt sich aufgrund *Ranking*-Funktion etwa nach Vektorraummodell oder probabilistischem Modell.
- danach Interaktion und bestimmte Dokumente als sehr relevant oder irrelevant markieren
  - dann Recherche-Deskriptoren so verändern, dass sich verbesserter Fragevektor im Vektorraummodell ergibt: *Relevance Feedback*

Bewertung: Ranking

- Gefundene Dokumente gemäß Relevanz absteigend sortieren
- Maße:
  - f_{in}: Häufigkeit Deskriptor T_i im Dokument D_n
  - t_n: Anzahl verschiedener Deskriptoren in Dokument D_n
  - d_m: Anzahl Dokumente in Datenbasis, in denen Deskriptor T_m auftritt
  - F_m: Auftretenshäufigkeit des Deskriptors T_m in gesamter Datenbasis
  - $Sf_n = \sum_{i=1}^{t_n} f_{in}$
  - w_{nm} Bewertung des Deskriptors T_m in Dokument D_n
  - Ranking-Funktionen: nächste Seite
- Man kann auch Fragedeskriptoren wichten
- alle 1: Gleichgewichtung
- Dann Korrelation zwischen Frage und Dokument berechnen

Bewertung: Ranking-Verfahren für einen Deskriptor

Ranking-Formel	Idee
$\frac{1}{d_m}$	Spezielle Begriffe, die nicht so häufig in der Datenbasis auftreten, sind wichtiger.
$\frac{1}{t_n}$	Ein einzelner Deskriptor ist umso unwichtiger, je mehr Deskriptoren insgesamt im Dokument auftreten.
f_{in}	Die Häufigkeit des Deskriptors in einem Dokument ist entscheidend.
$\frac{f_{in}}{Sf_n}$	Die Häufigkeit des Deskriptors in einem Dokument relativ zur Dokumentlänge ist entscheidend.
$\frac{f_{in}}{F_m}$	Die Ausschließlichkeit des Deskriptors in diesem Dokument ist entscheidend.

Bewertung: Ranking-Funktionen

- nur Boolesche Recherchen ohne Gewichtung der Recherche-Deskriptoren:

$$R_n = \frac{1}{M} \sum_{m=1}^{M} w_{nm}$$

- Gewichtung der Recherche-Deskriptoren vm :

$$R_n = \frac{1}{M} \sum_{m=1}^{M} v_m \bullet w_{nm}$$

Bewertung: Beispiel

```
Hotel:0.8 and Ostsee:0.5 and Strandkorb:0.2
```

mit Wichtungen

	Hotel	Ostsee	Strandkorb	Ranking
Gewichte →	0.8	0.5	0.2	
Dokument ↓				
D_1	0.7	0.9	0.3	(0.56+0.45+0.06)/3=0.36
D_2	0.3	1.0	1.0	(0.24+0.50+0.20)/3=0.31
D_3	0.9	0.4	0.9	(0.72+0.20+0.18)/3=0.37

Bewertung: Beispiel (2)

ohne Wichtungen:

	Hotel	Ostsee	Strandkorb	Ranking
Gewichte →	1.0	1.0	1.0	
Dokument ↓				
D_1	0.7	0.9	0.3	(0.7+0.9+0.3)/3=0.63
D_2	0.3	1.0	1.0	(0.3+1.0+1.0)/3=0.77
D_3	0.9	0.4	0.9	(0.9+0.4+0.9)/3=0.73

Im Booleschen Modell mit Deskribierung-Schwellwert 0.5 bei

```
Hotel and Ostsee and Strandkorb
```

kein Dokument im Ergebnis und bei

```
Hotel or Ostsee or Strandkorb
```

alle Dokumente im Ergebnis

Bewertung: Interaktion und Relevance Feedback

- stufenweise Einschränkung (unter gefundenen Dokumenten weitere Recherchen, System nennt jeweils Umfang der Zielpunktliste)
- Browsing (Dokumente ansehen, eventuell „gutes" Dokument als Ausgangspunkt für Recherche nehmen)
- Thesaurus einbeziehen (Anzeige für jeden Deskriptor, in die Anfrage übernehmen)
- Relevance Feedback (Relevanzrückkopplung): Nutzer bewertet Dokumente auf Relevanz, diese Bewertung erhöht / verringert Relevanzmaße in den Ranking-Funktionen (pro Benutzer)

Selektionsgüte

- Relevanzquote (Precision)
- Nachweisquote (Recall)

Systeme

STAIRS und GOLEM2: Deskriptoren, boolesche Suche

- STAIRS (IBM)
  - Freitextverfahren
  - bibliographische Angaben
  - automatische Deskribierung: Stichwortverfahren (ohne morphologische Reduktion)
- PASSAT/GOLEM2 (Siemens)
  - PASSAT: automatische Deskribierung (Grundformen, Inhaltserschließung)
  - GOLEM2: erweitertes Schlagwortverfahren, Thesauri
- auch datenbankbasierte (etwa ADABAS C)

beide Systeme Ende der 60er Jahre

STAIRS (IBM)

- STorage And Information Retrieval System
- Deskribierung und Speicherung
  - Speicherung von Dokumentbeschreibungen und Dokumenten
  - Stichwortverfahren: frei definierbare Stopwortliste, keine morphologische Reduktion, Textteile können von Deskribierung ausgenommen werden
  - schlechte Updatemöglichkeiten für Datenbasis (nur INSERT, DELETE)
  - Boolesche Suche mit invertierter Datei
  - Minimal-Thesaurus: Synonyme

STAIRS: Recherche

- Ergebnis jeder Anfrage bekommt ID (ist damit wiederverwendbar)
- Phasen einer Recherche
  - Grobrecherche mit Deskriptoren (SEARCH)
    Boolesche Suche, Grundform mit Wildcards, metrische Operatoren, verschiedene Ranking-Formeln
  - Grobrecherche mit bibliographischen Angaben (SELECT; wie SQL-WHERE-Klausel)
  - Feinrecherche mit Dokumenttext (RANK und BROWSE)
- Beispiele für Operatoren (SEARCH)
  - AND, OR, NOT, XOR
  - ADJ: nebeneinander
  - WITH: selber Satz
  - SAME: selbes Segment
  - Vorgabe der zu durchsuchenden Teile der Datenbasis
- Feinrecherche
  - RANK ermöglicht Auswahl von Ranking-Funktionen
  - BROWSE ermöglicht das Laden von Segmenten oder Texten

STAIRS: Interne Ebene

- STAIRS-Datenbasis: vier Dateien
  - Deskriptorliste
  - Zielpunktliste (mit Konkordanzen)
  - Text-Index-Datei (bibliographische Angaben)
  - Text-Datei (Dokumenttexte)
- einige implementierungstechnische Einschränkungen: Segmente maximal 54 Zeilen, darum auch Segmentspannen

PASSAT/GOLEM2 (Siemens)

- GOLEM2 (Großspeicher-Orientierte Listenorganisierte Ermittlungs-Methode 2. Entwicklungsstufe)
- Frei wählbarer Thesaurus
- Beziehungen zwischen Deskriptoren (15 verschiedene, u.a. Synonyme) können berücksichtigt werden
- Speicherung und Deskribierung
  - Zielinformation: Dokumenttext und Beschreibung (Deskriptoren)
  - Deskriptoren: Stichworte, Schlagworte (oder bibliographische Angaben auch als attributierte Stichworte = Aspekte)
  - Deskribierung selbst nicht in GOLEM2, sondern in PASSAT
    - entnimmt Dokument Stichwörter, reduziert auf Grundform
    - dazu benötigt: Vergleichswortliste
    - Inhaltserschließung durch Assoziationsmatrix

GOLEM2: Recherche

- Grobrecherche mit Deskriptoren
  - Stichwort- und Schlagwortsuche mit booleschen Operatoren und Aspekten
  - Synonyme automatisch berücksichtigt (mit V), andere Beziehungen auf Anforderung
  - Kein Ranking(aber Vorkommenshäufigkeit von Deskriptoren angezeigt)
- Feinrecherche mit Deskriptoren
- Automatische Feinrecherche in Textabschnitten
  - Teilstringsuche in Textabschnitten (mit Art und Zahl von Zwischenzeichen), diese verknüpft auch über metrische Operatoren (SATZ)
- Intellektuelle Feinrecherche (Browsing)

DB2: Text Extender

Deskribierungswerkzeug zur Erstellung einer invertierten Liste mit oder ohne Stammwortreduktion

vier verschiedene Index-Arten:

- *Linguistischer Index*: mit Stammwortreduktion, Stoppwortliste, *Feature-Extraktion* zur Erkennung von Eigennamen und Abkürzungen
- *Präziser Index*: invertierte Liste, Stoppwortliste
- *Dualer Index*: Index kombiniert linguistischen und präzisen Index
- *N-gram Index*: N-gram ist jede Teilzeichenkette der Länge N

Indexdateien außerhalb des Datenbanksystems

bis DB2 Version 5.2: nur ein Index pro Attribut, Retrieval-Funktionen abhängig vom Index

DB2: Recherche-Funktionen

- **contains**: Suche nach Deskriptoren in Texten
- **no_of_matches**: Auftretenshäufigkeit des Deskriptors im Dokument
- **rank**: einfachste Form: Auftretenshäufigkeit
- **search_result**: kombiniert obige Funktionen in temporäre Tabelle (Dokumentidentifikator, Auftretenshäufigkeit, Ranking-Wert)
- **contains** oder **search_result**: Boolesche Recherchen
- **precise form of**: exakte Recherche (nicht mit linguistischem Index)
- **stemmed form of**: lingustische Recherche (nicht mit präzisem Index und N-gram Index)
- **fuzzy form of**: Muster in Deskriptoren (Match aber wenigstens auf den ersten drei Positionen; N-gram Index erforderlich)
- Kontext-Suche: Umgebungen „Satz" (**in same sentence as**) oder „Absatz" (**in same paragraph as**)
- Anfragen können um Synonyme, Unterbegriffe oder assoziierte Begriffe aus einem Thesaurus ergänzt werden

DB2 Text Extender: Beispiel

mit **contains** in Dokumentmenge „Datenbankliteratur" nach Synonymen (**syn**) von „Anfragesprache" suchen, die im Fachthesaurus „Datenbanklexikon" verzeichnet sind (**contains** simuliert im alten SQL-Stil den Wert **true** mit dem Zahlenwert 1 simuliert)

```
select Titel
  from Datenbankliteratur
  where contains(Dokument-Handle,
   'thesaurus "Datenbanklexikon"
   expand "Syn" term of "Anfragesprache"')=1
```


Informix Excalibur Text Search Data Blade

- im Gegensatz zum Text Extender etwas eingeschränkte Fähigkeiten
- im Gegensatz zum Text Extender jedoch Index mit SSL-Klausel **create index**

using kann Form der Indizierung vorgeben

```
create index Textindex
on Datenbankliteratur (Dokument, Operatorklasse)
using etx (word_support = 'exact')in LiteraturSpace
```

Index über Volltext-Attribut „Dokument" der Tabelle „Datenbankliteratur" für exakte Suche

Informix: Index-Optionen

- **word_support**: exakte Suche (**exact**) und Mustersuche (**pattern**)
- **phrase_support**: Suche in Kontexten oder nach Phrasen in verschiedenen Genauigkeiten
- **char_set**: zugrundeliegender Zeichensatz
- **stopword_list**}: Eliminierung von Stoppwörtern
- **include_stopwords**: Deskriptoren der Stoppwortliste werden indiziert, aber nur auf Anforderung in der Recherche berücksichtigt

keinerlei Index mit
Stammwortreduktion (keine linguistischen Suche)

Informix: Recherche-Funktion

- Funktion **etx_contains**
- exakte Suche: in **etx_contains** mit **search_type** = **word**
- Boolesche Suche: **search_type** = **boolean_search**
- Phrasensuche und Suche innerhalb von Kontexten
- Kontext: nur Abstand zwischen Deskriptoren in Anzahl Worten
- Muster: Transpositionen (Vertauschung benachbarter Zeichen) oder Substitutionen (Ersetzung eines Zeichens) erlaubt?
- **match_synonym**: Synonymliste

```
select Titel
from Datenbankliteratur
where etx_contains(Dokument,
  (row (''Anfragesprache'', match_synonym = Datenbanklexikon))
```

Ranking: nur Güte der Übereinstimmung des Recherche-Deskriptors mit dem Dokument-Deskriptor (exakte Übereinstimmung besser als Übereinstimmung über Muster)

SQL-99/MM

- Multimedia-Datentypen im Teilstandard SQL/MM
- räumlichen Daten, Bilder, Texte, ...
- Kontexte: Wörter, Sätze, Absätze
- Boolesche Suche, Suche nach Mustern, Ranking-Funktion, Suche mit Thesaurus

SQL/MM Volltext

- Suche sprachabhängig (englisch, deutsch, ...)
- Phrasensuche: in **contains** ohne zusätzlichen Parameter

```
select Titel
from Datenbankliteratur
where contains(Dokument,
 ''Sein oder nicht Sein, das ist hier die Phrase'') = 1
```

- Kontextsuche: **in same sentence as** oder andere Kontexte
- Stoppwörter automatisch berücksichtigt, falls Stoppwortliste im System definiert
- linguistische Suche mit **stemmed form of**
- Ranking-Funktion implementierungsabhängig, liefert numerischen Wert

Boolesche Recherche: **contains** (Text, Wort_1 **and** ... **and** Wort_n)

Fulcrum

- Fulcrum SearchServer: Volltext-Datenbanksystem (Volltext-Dokumente nicht im System gespeichert, externe Dokumente und integrierte Indexdaten)
- Deskribierungsverfahren für diverse Text-Dokumenttypen
- Extrahierte, attributierte Daten können neben den Referenzen in Tabellen gehalten werden
- SearchSQL: Anfragen an attributierte Daten und Volltexte
- SearchSQL nur geringe Teilmenge von SQL-92 (keine Verbundoperationen, nur Ein-Variablen-Anfragen, keine Schachtelung von Anfragen, keine Attributselektionen, ...)
- SearchSQL: gute IR-Fähigkeiten wie Mustersuche (mit Wildcards), Phrasen, Thesaurussuche, Stammwortreduktionen, gewichtete Recherche-Deskriptoren, Retrieval-Modell von Boolescher Suche bis Vektorraummodell und Ranking-Algorithmus frei einstellbar, Relevance Feedback durch Angabe eines Vergleichsdokuments, Kontext-Suche
- nur Abstand in Zeichen

Literatur

- Zu allgemeinen Techniken:
  Baeza-Yates, R.; Ribeiro-Neto, B.: *Modern Information Retrieval*. Addison-Wesley, Harlow, England. 1999
- Zu Altsystemen:
  Lockemann, P.C.; Mayr, H.C.: *Rechnergestützte Informationssysteme*. Springer, Heidelberg. 1978
- zu allgemeinen Techniken, Datenbanklösungen:
  Heuer, A.; Saake, G.: *Datenbanken - Konzepte und Sprachen*. MITP-Verlag, Bonn. 2. Auflage 2000 (Abschnitt 10.3)