In: Zimmermann, Harald H.; Schramm, Volker (Hg.): Knowledge Management und Kommunikationssysteme, Workflow Management, Multimedia, Knowledge Transfer. Proceedings des 6. Internationalen Symposiums für Informationswissenschaft (ISI 1998), Prag, 3. – 7. November 1998. Konstanz: UVK Verlagsgesellschaft mbH, 1998. S. 397 – 403

Selix im DFG-Projekt Kascade

Hubert Hüther

SOFTEX GmbH
Hubert Hüther
Schmollerstr. 31
66111 Saarbrücken
Email: hhuether@softex.de

Inhalt

1 Einleitung

Mit KASCADE wird der Aufbau einer Datenbasis aus inhaltlich angereicherten und maschinell tiefer gehend erschlossenen Dokumenten (über Titel und bibliographische Daten hinaus) durchgeführt.
Mit dem Indexierungssystem MILOS/IDX (LZ 97) werden die Daten zunächst erschlossen.
Mit dem System SELIX werden, ausgehend von diesen Ergebnissen von MILOS, Parameter für einzelne Dokumente und auch die gesamte Kollektion der Dokumente ermittelt. Mit Hilfe dieser dann vorhandenen Parameter werden weitgehend automatisch Entscheidungen gefällt, ob ein von MILOS erschlossener Indexterm für ein Dokument zur Indexierung benutzt werden soll. Gleichzeitig steht dem Benutzer ein breites Spektrum zur Steuerung zur Verfügung.

2 Der Rahmen von SELIX

2.1 KASCADE

Seit dem 1. April 1997 wird mit Förderung der Deutschen Forschungsgemeinschaft an der Universitäts- und Landesbibliothek Düsseldorf das Projekt KASCADE durchgeführt. Gemeinsam mit der Fachrichtung Informationswissenschaft der Universität des Saarlandes sollen in dem auf eine Laufzeit von 21 Monaten (ursprünglich 18 Monate) angelegten Projekt neue Möglichkeiten der Inhaltserschließung und des Information Retrieval entwickelt und im bibliographischen Einsatz getestet werden.
Ziel von KASCADE ist der Aufbau einer Datenbasis aus inhaltlich angereicherten und maschinell tiefer gehend erschlossenen Dokumenten (über Titel und bibliographische Daten hinaus). Die Welt des klassischen bibliothekarischen Online-Katalogs (OPAC) wird dabei also verlassen.
Grundlage für die Datenbasis von KASCADE und den entsprechenden Demonstrator ist das Fachgebiet Jura (Zeitschriften und Monographien) aus dem Datenbestand der Universitäts- und Landesbibliothek Düsseldorf, das hinsichtlich Größe (ca. 30.000 Titel) und sprachlicher Zusammensetzung (großer deutschsprachiger Anteil) für die Zielsetzung gut geeignet ist.
Für die Anreicherung der Erschließungsdaten wurden für etwa 3.000 Dokumente die Inhaltsverzeichnisse gewählt (da diese im Unterschied zu Abstracts und Registern fast immer vorhanden waren) (siehe auch (LZ 97)).

2.2 MILOS

Mit dem Indexierungssystem MILOS/IDX (LZ 97) werden die Daten zunächst erschlossen. Das Eingabeformat für dieses Indexierungssystem ist das ASCII-Format. Dabei können Zusatzinformationen in Form von Kommentaren mitgegeben werden. Für die Auswertung der Indexierungsergebnisse (ebenfalls ASCII-Format) ist für SELIX die Dokumentnummer als Zusatzinformation erforderlich. Bei der Indexierung werden Grundform, Wortklasse, Teilwörter usw. (siehe (LZ 97)) gewonnen (dabei steht ein Relationenwörterbuch (Thesaurus) zur Verfügung). Die dann vorhandenen Indexierungselemente für ein Dokument können sehr zahlreich sein.

3 Die Aufgabe von SELIX

Mit dem System SELIX (Selektive automatische Indexierung) werden, ausgehend von den Ergebnissen von MILOS unter Benutzung eines Relationenwörterbuches, Häufigkeiten für Grundformen und Parameter für einzelne Dokumente und auch die gesamte Kollektion der Dokumente ermittelt (dabei können unter Benutzung des Relationenwörterbuches (Thesaurus) auch verschiedene Grundformen mit Hilfe von Synonym-, Ähnlichkeits-, Teilwort-, Vorzugsrelationen und/oder weiterer Relationen zusammengeführt werden). Mit Hilfe dieser dann vorhandenen Parameter sollen weitgehend automatisch Entscheidungen gefällt werden, ob ein von MILOS erschlossener Indexterm für ein Dokument zur Indexierung benutzt werden

soll. Hierzu wird für jede in einem Dokument (vorkommende oder zugeteilte Grundform) ein auf dieses Dokument bezogenes Gewicht ermittelt.
Im Unterschied zu aus der Literatur bekannten Gewichtsfunktionen (z.B. in HA 75, SA 89 oder HR 98) sollen auch bestimmte selten vorkommende Grundformen (komplexe Mehrwortgruppen oder Komposita) ‚statistisch gesehen' eine Chance erhalten. Ebenso soll der Benutzer sogenannte ‚Registerbegriffe' festlegen können (diese Begriffe werden einem Dokument zugeteilt, wenn sie im Dokument vorkommen oder mit einem im Dokument vorkommenden Begriff durch eine entsprechende Relation verbunden sind (unabhängig von seinem ‚Gewicht')). Das statistische Verfahren muss die Massenverarbeitung erledigen, aber der Benutzer soll ein umfangreiches Spektrum (Parameter, Relationenwörterbuch (Thesaurus)) zur Steuerung zur Verfügung haben.
Von SELIX werden hierzu drei Teilgewichte (nG1 – nG3) ermittelt. Die drei Teilgewichte sollen eine Maßzahl zu einen Begriff liefern:

(nG1) Eignung als Indexterm für die Dokumentkollektion (kollektionsbezogen)
(nG2) Wichtigkeit des Begriffes für ein Dokument (dokumentbezogen)
(nG3) Bedeutung des Terms für das Fachgebiet.

Die drei Gewichte ergeben additiv das Gesamtgewicht, wobei jedes Gewicht noch mit einem Faktor versehen werden kann.

4 Die Gewichtung nG von SELIX

Folgende Parameter werden mit SELIX ermittelt:

nColLen: Länge der Dokumentenkollektion bezogen auf die erfassten Wortformen.
nAnzDok: Anzahl der Dokumente in der Kollektion.
HfkImD(g,d): Häufigkeit einer Grundform ’g‘ in einem Dokument 'd'.
nDokLen(d): Dokumentlänge von ’d‘ bezogen auf die erfassten Wortformen.
nAnzGru(d): Anzahl der erfassten verschiedenen Grundformen im Dokument ’d‘.
nDok(g): Anzahl der Dokumente, die eine Grundform ’g‘ enthalten.
nColl(g): Anzahl der Vorkommen, der Grundform ’g‘ in der Dokumentenkollektion.
nGruLen(g): Länge der Grundform ’g‘.

Mit den oben definierten Parametern, die von SELIX für bestimmte Grundformen (Indexkandidaten) berechnet werden, lassen sich auch Gewichtungsfunktionen aus der Literatur benutzen. Für das System SELIX wurden hierzu zwei Funktionen ausgewählt:

Salton (SA 89)
HfkImD(g,d) * log (nDok(g) / nAnzDok)
Robertson (HR 98)
((K 1 + 1) * HfkImD(g,d) / (K + HfkImD(g,d))) * log ((nAnzDok – nDok(g) + 0.5) / (nDok(g) + 0.5))

Unabhängig von obigen beiden Gewichtungsfunktionen und weiteren Funktionen aus der Literatur wurden zwei eigene Gewichtungen entwickelt. Dabei haben (wie schon erwähnt) drei Gesichtspunkte eine wichtige Rolle gespielt. Für jeden Gesichtspunkt wurde ein Teilgewicht entwickelt.

Mit Hilfe obiger Zahlen werden zu jedem Indextermkandidat die drei Teilgewichte

nG1 (kollektionsbezogen (für alle Dokumente (einer Kollektion) gleich),
nG2 (dokumentbezogen) und
nG3 (unabhängig von einem Dokument und einer Kollektion)

ermittelt und additiv miteinander zu dem Gesamtgewicht nG verknüpft. Dabei können die Gewichte noch mit einem Faktor versehen werden (F1 – F3).

nG = F 1 * nG1 + F 2 * nG2 + F 3 * nG3

Da aus der Literatur kein Ansatz bekannt war, der diese drei Gesichtspunkte berücksichtigt, wurde diese Vorgehensweise gewählt. Es ist denkbar, dass für die Teilgewichte nach den Retrievaltests auch Teile der Funktionen bei Salton, Robertson oder anderen aus der Literatur bekannten Funktionen genommen werden, bzw. dass die ’eigenen‘ verbessert werden.

Jedes Teilgewicht (nG1 – nG3) soll in dem Intervall (0,1) liegen, damit die Benutzung verschiedener Faktoren (F 1 – F 3) übersichtlicher wird.

4.1 Das Kollektionsgewicht nG1

In dem Teilgewicht nG1 soll sich widerspiegeln, ob eine Grundform für eine Dokumentenkollektion als Indexterm geeignet ist. Hierbei wird sich an einer Zufallsverteilung orientiert (siehe hierzu neben (SA 89), (HR 98) auch (HA 75)).
Aus der Überlegung ’wenn sich viele Vorkommen einer Grundform auf relativ wenige Dokumente konzentrieren, kann keine zufällige Verteilung vorliegen‘ wurde für **KasEinf** folgende Funktion gewählt.

KasEinf (KASCADE einfach):

nG1 = 1 – nDok(g) / nColl(g)

Bei KasKomp **wurde für nDok(g) als Vergleichsgröße der Erwartungswert E(nDok(g)) für die Anzahl der Dokumente, in denen die Grundform 'g' vorkommt, genommen.**
KasKomp (KASCADE komplex):

nG1 = 1 – nDok(g) / E(nDok(g)) (für: nDok(g) < E(nDok(g)); 0 sonst)
mit
E(nDok(g)) = nAnzDok * (1 – exp(-λ))
wenn eine Poisson Zufallsverteilung angenommen wird:
P(i) = exp (-λ) * (λ^i/i!) λ = nColl(g) / nAnzDok

4.2 Das Dokumentgewicht nG2

In dem Teilgewicht nG2 soll sich widerspiegeln, ob eine Grundform für eine Dokument als Indexterm wichtig ist.
Aus der Überlegung 'wenn die Häufigkeit einer Grundform 'g' innerhalb eines Dokumentes 'd' im Vergleich zur durchschnittlichen Häufigkeit groß ist, muss diese Grundform für dieses Dokument wichtig sein' wurde für **KasEinf** folgende Funktion gewählt.
KasEinf:

nG2 = 1 – ((nDokLen(d) / nColLen) / (HfkImD(g,d) / nColl(g)))
(falls ≥ 0; sonst 0)

Bei KasKomp **wurde als Gewicht der Erwartungswert, dass der Indexterm 'g' in dem Dokument 'd' 1 bis HfkImD(g,d) mal vorkommt, in Relation zum Erwartungswert gewählt.**
KasKomp:

nG2 = (p(1) * 1 + ... + p(HfkImD(g,d)) * HfkImD(g,d)) / λ
mit
P(i) = exp (-λ) * (λ^i/i!)
λ = nColl * (nDokLen/nColLen)

4.3 Das Längengewicht nG3:

Betrachtet man Begriffe wie

EU-Denkmalförderungsrichtlinie,
Elektrizitätsbinnenmarktrichtlinie,
Urheberrechtsschiedsstellenverordnung oder
Schiedsgerichtsvereinbarungen in Rückversicherungsverträgen,

dann sind diese Terme sicherlich wichtig für ein Dokument, in denen sie vorkommen, ebenso stehen sie kaum zufällig in einem Dokument, etwa im Vergleich zu 'Richtlinie'. Wenn jedoch keine sehr große Dokumentenkollektion vorliegt, und

solche Terme selten vorkommen, haben diese Terme bei einem statistischen Ansatz kaum eine Chance. Deshalb wurde bei dem dritten Gewicht nG3 die Länge miteinbezogen.

nG3 = log (nGruLen(g)) / 4

5 Das Programmsystem von SELIX

Durch Auswahl bestimmter Wortklassen, 'Relationen' und 'Semantiken' (siehe (LZ 97)) können bestimmte Grundformen von der Häufigkeitszählung ausgeschlossen werden, bzw. können verschiedene Grundformen auf eine Vorzugsgrundform bei der Zählung zusammengeführt werden.

Bei Komposita und Mehrwortgruppen die Möglichkeit der Mehrfachzählungen (zusätzlich zum Teilgewicht nG3).

Zur Zeit werden die Gewichte 'nG' nur benutzt, um Indexterms für ein Dokument zu selektieren. Die Länge eines Dokumentes soll dabei auch für die Anzahl der Deskriptoren berücksichtigt werden. Hierzu gibt es folgende Parameter:

nGewMin: Minimales Gewicht.
nAnzMin: Minimale Anzahl von Deskriptoren.
nAnzMax: Maximale Anzahl von Deskriptoren.
nAnzMit: Mittlere maximale Anzahl von Deskriptoren. Diese Anzahl gilt für ein durchschnittlich langes Dokument (Länge: nColLen(d) / nAnzDok). Für jedes Dokument wird diese Anzahl (Mittlere maximale Anzahl von Deskriptoren) in Abhängigkeit von seiner Länge neu berechnet. Diese Zahl muss aber zwischen den Zahlen nAnzMin und nAnzMax liegen, die unabhängig von der Dokumentlänge sind.

Wenn eine Grundform als Registerstichwort gekennzeichnet ist, wird diese Grundform ausgegeben, unabhängig von seinem Gewicht. Zusätzlich wird bei dem Gewicht das Zeichen '*' vorangestellt.
Neben einem 'formalen Trivialwort' (z.B. 'der') gibt es noch ein 'Trivialwort (System)' (z.B. 'Kapitel') und ein 'Trivialwort (Nutzer)' (z.B. 'Recht' im Fachgebiet Jura). Ein formales Trivialwort wird als Indexterm nicht zugelassen. Die andern beiden werden bei der Gewichtsberechnung behandelt wie andere Indexkandidaten; bei der Ausgabe wird, falls sie die anderen Bedingungen erfüllen, dem Gewicht die Zeichenkette '**' vorangesetzt.

6 Ausblick

In dem System SELIX ist schon die Möglichkeit zur automatischen Gewinnung von Mehrwortgruppen (mit Hilfe eines MWG-Parsers) vorgesehen.

Beispiele aus ersten Tests:

Diskrepanz zwischen der ermittelten Blutalkoholkonzentration und Zeugenaussage

Ermittlung der Blutalkoholkonzentration aus der Blutprobe / ohne Blutuntersuchung

fahrlässige Herbeiführung der Beeinträchtigung der Schuldfähigkeit

Eine solche Mehrwortgruppe ist sicher nicht geeignet als Indextermkandidat im herkömmlichen Sinne. Diese Mehrwortgruppen sollen jedoch (in Zukunft) benutzt werden bei der Dialogführung mit dem Benutzer, wenn dieser mit bestimmten Teilkomponenten solcher Mehrwortgruppen in den Retrievaldialog einsteigt.

7 Literatur

[HA 75] Harter, S.D. (1975): A probabilistic approach to automatic keyword indexing (On the distribution of specialty words in a technical literature); Journal of ASIS, (1975), p. 197-206

[HR 98] Huang, Xiangji; Robertson, S.E. (1998): Okapi Chinese text retrieval experiments at TREC-6

[LZ 97] Lepsky, Klaus; Zimmermann, Harald H. (1997): Katalogerweiterung durch Scanning und Automatische Dokumenterschließung; Das DFG-Projekt KASCADE http://www.uni-duesseldorf.de/WWW/ulb/kas_home.htm

[SA 89] Salton, G. (1989): Automatic Text Processing; Addison Wesley