Kapitel IR:II (Fortsetzung)

II. Grundlagen des Information Retrieval

- Retrieval-Evaluierung
- Indexterme

Indexterme

Die klassischen Dokumentmodelle abstrahieren ein Dokument auf eine Menge von sogenannten Indextermen oder Deskriptoren.

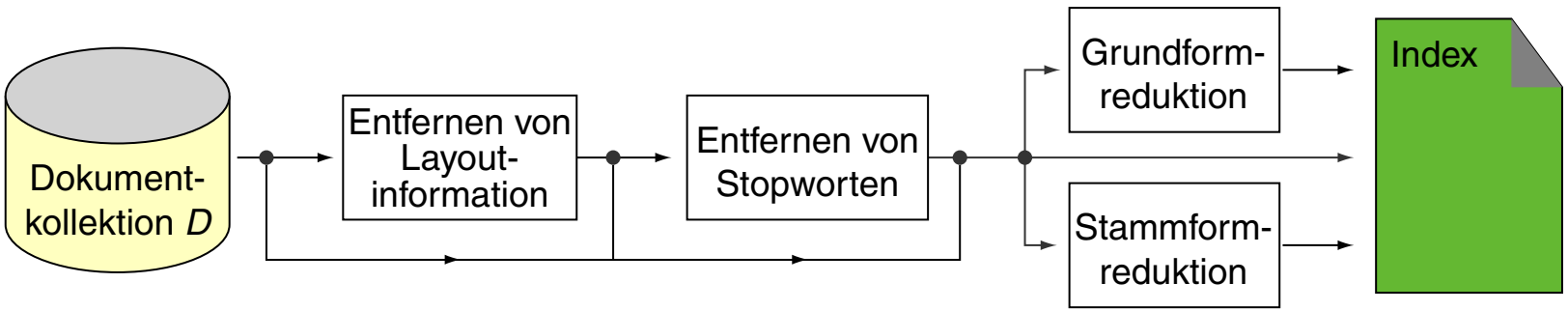
Idealerweise sollten Indexterme so gewählt sein, dass sie

1. den Inhalt der einzelnen Dokumente adäquat repräsentieren,
2. eine möglichst klare Abgrenzung der einzelnen Dokumente gewährleisten,
3. die Verknüpfung von thematisch ähnlichen Dokumenten ermöglichen.

Der Prozess der Auswahl von Indextermen heisst Indexierung.

Indexterme

Techniken, die bei der Indexierung zum Einsatz kommen:



- Unter Layout-Informationen *(Rendering Tags)* fallen insbesondere die Sprachen zur Textauszeichnung wie HTML, XML oder SGML.
- Als Stopworte werden häufige und gleichverteilt vorkommende Terme bezeichnet.
- Mit Grund- und Stammformreduktion *(Stemming)* versucht man, eine Generalisierung von Indextermen zu erzielen.
- Datenstrukturen zur Speicherung und zum effizienten Zugriff auf einen Index sind Hashtabellen, Bäume und Tries.

Bemerkungen:

- Auch Layout-Informationen und Stopworte können Retrieval-Modellen eine Rolle spielen. Ein Beispiel hierfür sind Retrieval-Modelle für Filteraufgaben wie die Genre-Klassifikation oder zur Erkennung der Sprache.
- Beachte die Leistungsfähigkeit eines menschlichen Editors bei der Indexierung von Dokumenten, z. B. bei der Vergabe von Stichworten für einen wissenschaftlichen Text.
- Die Dokumentrepräsentationen $\mathbf{d} \in \mathbf{D}$ *referenzieren* die Terme eines Index; d. h., sie verwenden nicht die Index-Terme selbst, sondern Zeiger hierauf.

Indexterme

Lexikalische Analyse

Konvertierung eines Zeichenstroms in einen Strom auf Worten.

Die Detektion von Leerzeichen ist nicht ausreichend – problematische Fälle:

1. Ziffern, Zahlen:
   `Airbus A360, 1984, RS232`
   ➜ Zahlen löschen, einzeln aufnehmen?

2. Trennstriche, Gedankenstriche, Minuszeichen:
   `State-of-Art, State of the Art, IR-Algorithmen`

3. Zeichen der Interpunktion:
   `510.B.C., n!`
   ➜ Interpunktion eliminieren, Worte aufteilen oder zusammenfassen?

4. Groß- und Kleinschreibung.
   ➜ Üblich ist eine einheitliche Konvertierung.

Indexterme

Stopwort-Elimination

Aufgrund ihres häufigen und gleichverteilten Auftretens sind Stopworte zur inhaltlichen Diskrimierung zwischen Dokumenten ungeeignet.

Kandidaten für Stopworte sind Präpositionen, Konjunktionen, Artikel und spezielle, geschlossene Wortklassen:

- Zahlworte
- Monate
- Währungen
- Namen
- ...

Die Elimination von Stopworten ist nicht immer sinnvoll – Beispiel:

"to be or not to be"

Bemerkungen:

- Stopworte sind sprachspezifisch.
- Eine Stopwortliste einer Sprache enthält größenordnungsmäßig 200-1000 Worte.
- die Entfernung von Stopworten verbessert die Kompression bis zu 40%.

Indexterme

Beispiel

Text mit Markups [Quelle Reuters]:

```
<TEXT> <TITLE>CHRYSLER> DEAL LEAVES UNCERTAINTY FOR AMC
WORKERS</TITLE> <AUTHOR> By Richard Walker, Reuters</AUTHOR>
<DATELINE> DETROIT, March 11 - </DATELINE><BODY>Chrysler
Corp's 1.5 billion dlr bid to takeover American Motors Corp;
AMO> should help bolster the small automaker's sales, but it
leaves the future of its 19,000 employees in doubt, industry
analysts say. It was "business as usual"yesterday at the
American ...
```

Indexterme

Beispiel (Fortsetzung)

Rohtext:

```
chrysler deal leaves uncertainty for amc workers by richard
walker reuters detroit march 11 chrysler corp s 1 5 billion
dlr bid to takeover american motors corp should help bolster
the small automaker s sales but it leaves the future of its
19 000 employees in doubt industry analysts say it was
business as usual yesterday at the american
```

Indexterme

Beispiel (Fortsetzung)

Stopworte fett:

chrysler deal leaves uncertainty **for** amc workers **by** richard walker reuters detroit **march 11** chrysler **corp s 1 5 billion dlr** bid **to** takeover american motors **corp should** help bolster **the small** automaker **s** sales **but it** leaves **the** future **of its 19 000** employees **in** doubt industry analysts **say it was** business **as usual** yesterday **at the** american

Indexterme

Beispiel (Fortsetzung)

Nach der Wortstammreduktion:

```
chrysler deal leav uncertain amc work richard walk reut
detroit takeover american motor help bols automak sal leav
futur employ doubt industr analy business usual yesterday
```

Indexterme

Manuelle Indexierung

Ein oder mehrere Editoren ordnen jedem einzelnen Dokument Deskriptoren zu, die seiner/ihrer Meinung nach das Dokument inhaltlich gut beschreiben.

Automatische Indexierung

- statistisch:
  basiert auf der Häufigkeitsverteilung der Terme in einem Korpus
- probabilistisch:
  unterstellt auf Basis von Dokumentmodellen bestimmte Wahrscheinlichkeitsverteilungen für die Terme eines Korpus
- linguistisch:
  basiert auf der morphologischen, syntaktischen und semantischen Analyse der Dokumente

Semi-automatische Indexierung

- Relevanz-Feedback
- computerunterstütztes Indexing

Indexterme

Nachteile manueller Indexierung

- sehr großer Zeitaufwand:
  Aufgrund der stark zunehmenden Informationsmengen wird der Zeitabstand zwischen der Veröffentlichung eines Dokuments und der Verfügbarkeit seiner Indexierung in Datenbanken immer größer.

- Kontinuität bei der Vergabe von Deskriptoren ist nicht gewährleistet:
  1. Bei großen Dokumentenkollektionen geht der Überblick verloren und für gleiche Themen werden unterschiedliche Deskriptoren vergeben.
  2. Oft sind mehrere Personen mit der Indexierung betraut, und jeder Editor kann eine andere Intuition bzgl. der Deskriptorvergabe haben.

- hohe fachliche Anforderungen an die Editoren:
  Sie müssen den Inhalt eines Dokumentes schnell erfassen, das Vokabular gut kennen, geeignet abstrahieren, etc.

- hohe Personalkosten

Indexterme

Statistische Indexierung: Zipf'sches Gesetz

Das am meisten verwandte Konzept bei statistischen Indexierungsverfahren ist die Häufigkeit eines Terms in einer Dokumentenkollektion. Grundlage hierfür bildet das Zipf'sche „Gesetz“ [George Kingsley Zipf, 1902-1950] :

> Die empirisch gefundenen Häufigkeit P eines Wortes w in einem ausreichend langen Textes korreliert mit seinem Häufigkeitsrang $r(w)$ in einem Skalengesetz:
>
> $$P(w) = \frac{c}{(r(w))^a}$$

Indexterme

Statistische Indexierung: Zipf'sches Gesetz

Das am meisten verwandte Konzept bei statistischen Indexierungsverfahren ist die Häufigkeit eines Terms in einer Dokumentenkollektion. Grundlage hierfür bildet das Zipf'sche „Gesetz" [George Kingsley Zipf, 1902-1950] :

> Die empirisch gefundenen Häufigkeit P eines Wortes w in einem ausreichend langen Textes korreliert mit seinem Häufigkeitsrang $r(w)$ in einem Skalengesetz:
>
> $$P(w) = \frac{c}{(r(w))^a}$$

Der Exponent a ist etwas größer als Eins; vereinfachend kann man schreiben:

$$P(w) \cdot r(w) = const$$

Bei der Berechung relativer Häufigkeiten wird c so gewählt, dass gilt:

$$\sum_{1}^{n} P(w) = \sum_{1}^{n} \frac{c}{r(w)} = 1$$

Indexterme

Statistische Indexierung: Zipf'sches Gesetz

$$P(w) = \frac{c}{(r(w))^a}$$

Durch Logarithmierung beider Skalen lässt sich die Gleichung in lineare Form bringen, so dass sie sich im Diagramm als Gerade darstellen lässt.

$$\log(P(w)) = \log(c) - a \cdot \log(r(w))$$

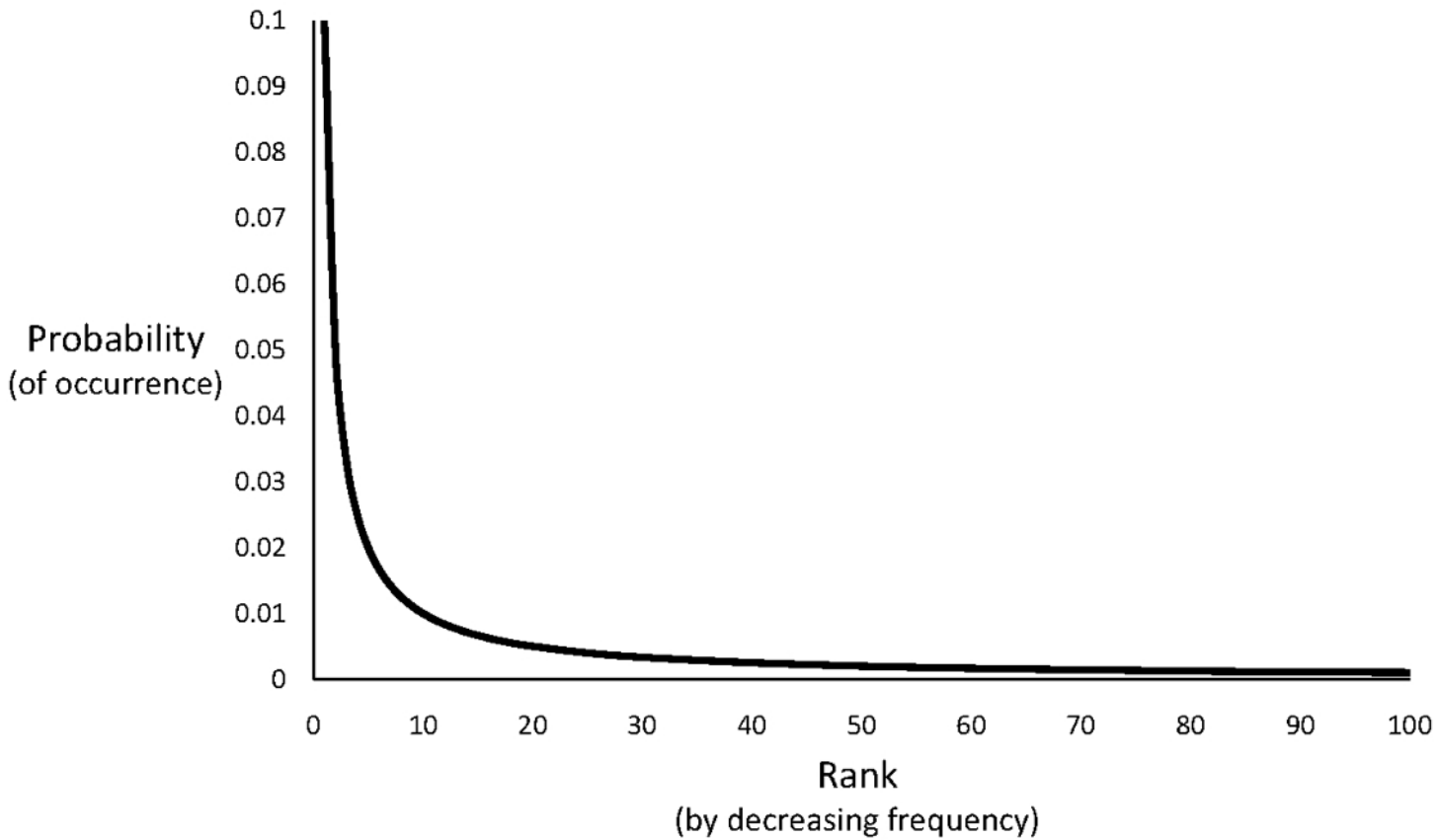
Als Erweiterung hat Mandelbrot folgende Form vorgeschlagen, auch Zipf-Mandelbrot-Gesetz genannt:

$$P(w) = \frac{c}{(i+b)^a}$$

D. h., für das einfache Gesetz von Zipf ist $b > 0$ und $a = 1$.

Indexterme

Statistische Indexierung: Zipf'sches Gesetz



Bemerkungen:

- Wie jedes empirische „Gesetz“ ist auch das Zipf’sche Gesetz nur näherungsweise gültig. Während es in dem mittleren Bereich die Häufigkeitsverteilung sehr gut wiedergibt, ist die Übereinstimmung bei sehr häufigen Worten (Stopworten) und sehr seltenen Worten geringer.
- Interessanterweise gilt diese Beziehung nicht nur für Worte und Buchstaben in Texten menschlicher Sprachen oder Noten in der Musik, sondern für so gut wie alle natürlichen Symbolsequenzen mit weitreichenden Korrelationen, wie z. B. der DNA.
- Eine Interpretation des Zipf’schen Gesetzes als Wahrscheinlichkeitsverteilung ist die Zeta-Verteilung, die deshalb auch Zipf-Verteilung genannt wird. Das Gegenstück für kontinuierliche Werte ist die Pareto-Verteilung.
- Ein unabhängig vom Zipf’schen Gesetz festgestellter Spezialfall ist das Gesetz von Benford bzgl. der Häufigkeit von Anfangsziffern.

Indexterme

Statistische Indexierung: Stopworte

Die 10 häufigsten Terme aus einer Stichprobe von 1.000.000 Worten aus dem englischen Brown-Korpus (links) und aus einer Stichprobe von 3.000.000 Worten aus dem Korpus der Stuttgarter Zeitung (rechts):

Rang	Term	Häufigkeit
1	the	69971
2	of	36411
3	and	28852
4	to	26852
5	a	23237
6	in	21347
7	that	10099
8	is	10019
9	was	9816
10	he	9543

Indexterme

Statistische Indexierung: Stopworte

Die 10 häufigsten Terme aus einer Stichprobe von 1.000.000 Worten aus dem englischen Brown-Korpus (links) und aus einer Stichprobe von 3.000.000 Worten aus dem Korpus der Stuttgarter Zeitung (rechts):

Rang	Term	Häufigkeit
1	the	69971
2	of	36411
3	and	28852
4	to	26852
5	a	23237
6	in	21347
7	that	10099
8	is	10019
9	was	9816
10	he	9543

Rang	Term	Häufigkeit
1	ein	710719
2	und	708531
3	in	613869
4	sein	534056
5	werden	400264
6	haben	340313
7	von	333335
8	zu	290911
9	mit	286015
10	im	278227