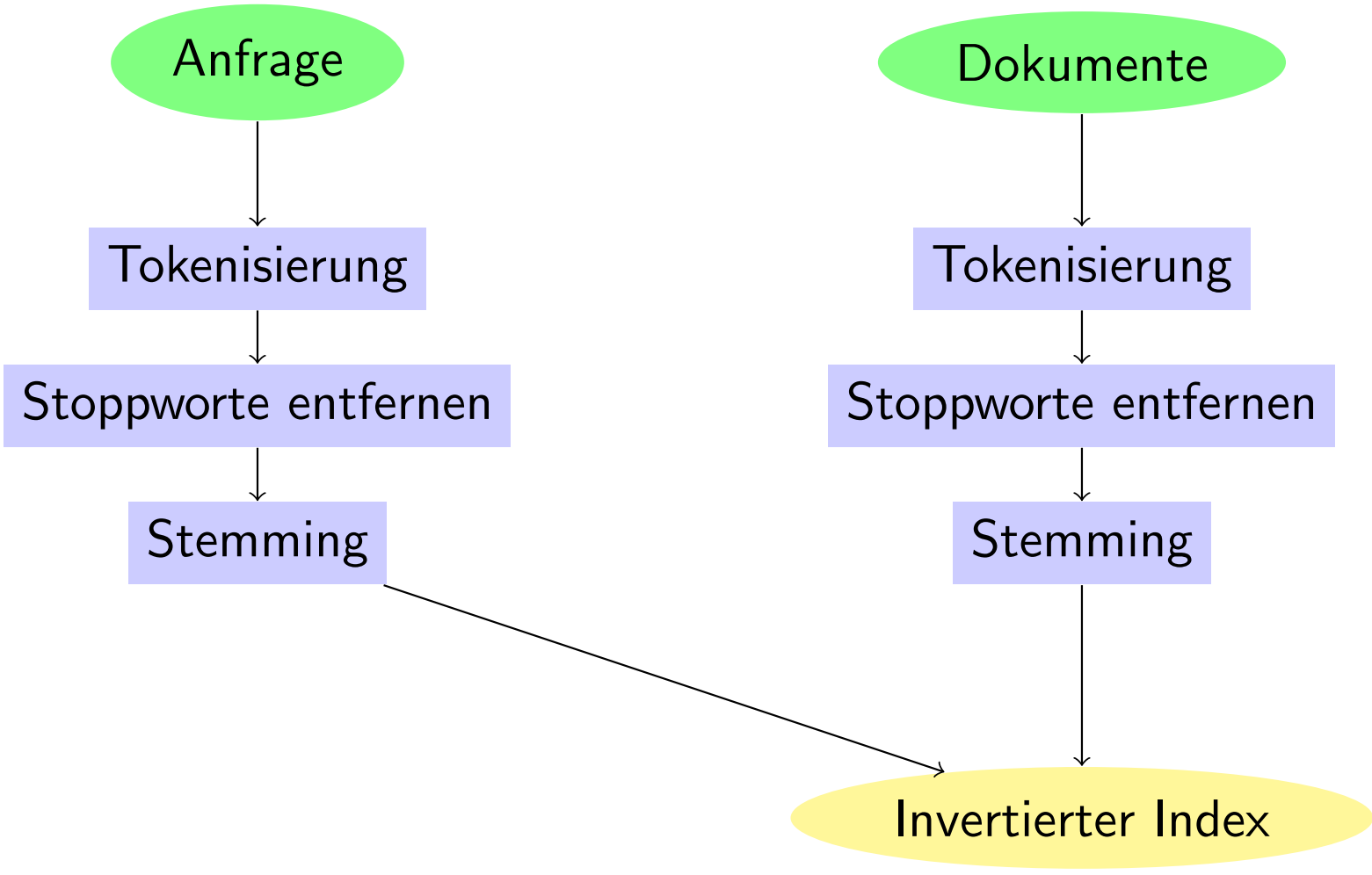
Information Retrieval

Invertierter Index, Boolesches Retrieval

Stefan Birkner

7. September 2010

Bisheriger Stand



Index

Index

Indizes werden in der Informatik verwendet, um einen *schnellen Zugriff* auf Daten in *umfangreichen Datensammlungen* zu gewährleisten.

Invertierter Index

Term	Dokumente (IDs)
Haus	1, 2

Invertierter Index

Beispieldokumente (Quelle: Wikipedia)

1. Unter den Linden ist die zentrale Prachtstraße Berlins und verläuft durch die Dorotheenstadt und den Friedrichswerder im Ortsteil Mitte. Sie führt vom Pariser Platz an der Ostseite des Brandenburger Tors bis zur Schloßbrücke, die die Verbindung zur Museumsinsel herstellt.
2. Der Berliner Dom ist eine evangelische Kirche auf dem nördlichen Teil der Spreeinsel, der hier Museumsinsel genannt wird, im Ortsteil Mitte von Berlin.

Invertierter Index

Beispieldokumente (Quelle: Wikipedia)

3. Die Friedrichstraße liegt in den Berliner Ortsteilen Mitte und Kreuzberg. Sie ist eine der bekanntesten Straßen im östlichen Zentrum Berlins und wurde nach dem Kurfürsten Friedrich III. von Brandenburg benannt.
4. Die Berliner Museumsinsel ist die nördliche Spitze der Spreeinsel im Zentrum Berlins. Sie ist historisch die Keimzelle der Berliner Museumslandschaft und mit ihren Museen ein vielbesuchter touristischer Anlaufpunkt und einer der wichtigsten Museumskomplexe der Welt.

Invertierter Index

Beispieldokumente nach Tokenisierung, Entfernung der Stoppwörter und Stemming

1. Linde, Zentrum, Prachtstraße, Berlin, Verlauf, Dorotheenstadt, Friedrichswerder, Ort, Teil, Mitte, führen, Paris, Platz, Osten, Seite, Brandenburg, Tor, Osten, Schloß, Brücke, Verbindung, Museum, Insel, herstellen
2. Berlin, Dom, evangelisch, Kirche, Norden, Spree, Insel, Museum, nennen, Ort, Teil, Mitte

Invertierter Index

Beispieldokumente nach Tokenisierung, Entfernung der Stoppwörter und Stemming

3. Friedrichstraße, Berlin, Ort, Teil, Mitte, Kreuzberg, bekannt, Straße, Osten, Zentrum, Kurfürst, Friedrich, Brandenburg, nennen
4. Berlin, Museum, Insel, Norden, Spitze, Spree, Zentrum, Geschichte, Keim, Zelle, Landschaft, Besuch, Tourismus, Anlauf, Punkt, Komplex, Welt

Invertierter Index

Term	Dokumente (IDs)
Linde	1
Zentrum	1, 3, 4
Prachtstraße	1
Berlin	1, 2, 3, 4
Verlauf	1
Dorotheenstadt	1
Friedrichswerder	1
Ort	1, 2, 3
Teil	1, 2, 3
Mitte	1, 2, 3
führen	1
Paris	1
Platz	1
Osten	1

Invertierter Index

Term	Dokumente (IDs)
Seite	1
Brandenburg	1, 3
Tor	1
Osten	1, 3
Schloß	1
Brücke	1
Verbindung	1
Museum	1, 2, 4
Insel	1, 2, 4
herstellen	1
Dom	2
evangelisch	2
Kirche	2

Invertierter Index

Term	Dokumente (IDs)
Norden	2, 4
Spree	2, 4
nennen	2, 3
Friedrichstraße	3
Kreuzberg	3
bekannt	3
Straße	3
Kurfürst	3
Friedrich	3
Spitze	4
Geschichte	4
Keim	4
Zelle	4
Landschaft	4
Besuch	4

Invertierter Index

Term	Dokumente (IDs)
Tourismus	4
Anlauf	4
Punkt	4
Komplex	4
Welt	4

Invertierter Index (Übung)

Beispieldokumente (Quelle: Wikipedia)

1. Teotihuacán ist eine ehemalige Stadt im mexikanischen Bundesstaat México. Die Azteken nannten sie mit dem bis heute fortlebenden Namen Teotihuacán.
2. Die Sonnenpyramide ist das zweitgrößte Bauwerk im vorspanischen Mittelamerika. Sie befindet sich in Teotihuacán an der Straße der Toten zwischen der Mondpyramide und der Ciudadela.
3. Plazuelas ist eine Ausgrabungsstätte in Mexiko. Sie besteht aus drei Pyramiden und einem Ballspielplatz.

Invertierter Index

Erweiterung der Dokumenteninformation

- Häufigkeit der Terme
- Position der Terme
- ursprüngliches Wort

Invertierter Index

Dokumente um die Häufigkeit der Terme erweitern

Term	Dokumente (ID, Anzahl)
Linde	1 (1)
Berlin	1 (1), 2 (2) , 3 (2), 4 (3)
Insel	1 (1), 2 (2), 4 (2)

Invertierter Index

Dokumente um die Häufigkeit der Terme erweitern

Term	Dokumente (ID, Positionen)
Linde	1 (3)
Berlin	1 (8), 2 (2, 23) , 3 (6,20), 4 (2,12,19)
Insel	1 (38), 2 (13,16), 4 (3,9)

Invertierter Index

Dokumente um die ursprünglichen Wörter erweitern

Term	Dokumente (ID, ursprüngliche Wörter)
Linde	1 (Linden)
Berlin	1 (Berlins), 2 (Berliner, Berlin) , 3 (Berliner, Berlins), 4 (Berlin
Insel	1 (Museumsinsel), 2 (Spreeinsel, Museumsinsel), 4 (Museumsins

Zipfsches Gesetz

Das Produkt aus der Häufigkeit eines Terms und seinem Rang ist konstant.

Beispiel Brown-und-Lob-Texkorpus

Term	Rang	Anzahl	Rang · Anzahl
the	1	138.323	138.323
of	2	72.259	144.318
and	3	56.750	170.250
to	4	52.941	211.764
a	5	46.523	232.615
in	6	42.603	255.618
that	7	22.177	155.239

Zipfsches Gesetz

Beispiel:
Das zweithäufigste Wort kommt in einem Textkorpus 100.000 mal vor. Wie oft kommt dann das vierthäufigste Wort ungefähr vor?

$$
\begin{aligned}
4 \cdot x &= 2 \cdot 100000 \\
x &= \frac{2 \cdot 100000}{4} \\
x &= 50000
\end{aligned}
$$

Das vierthäufigste Wort kommt ungefähr 50.000 Mal vor.

Boolesches Retrieval

Der Suchende kann nach Dokumenten suche, indem er in der Abfrage Wörter mit Hilfe der boooleschen Operatoren UND, ODER und NICHT verknüpft.

Boolesches Retrieval

Beispieldokumente

- Rotkäppchen spricht im Wald mit dem Wolf.
- Die sieben Geißlein lassen den Wolf ins Haus.
- Hänsel und Gretel verirren sich im Wald.

Boolesches Retrieval

Anfrage: Wald Wolf
Dokumente: Rotkäppchen

Boolesches Retrieval

Anfrage: Wolf UND (Wald ODER Haus)
Dokumente: Rotkäppchen, Die sieben Geißlein

Boolesches Retrieval

Anfrage: Wald UND NICHT Wolf
Dokumente: Hänsel und Gretel

Boolesches Retrieval - Invertierter Index

- UND-Verknüpfung mit Hilfe eines invertierten Index
- Anfrage: Term1 UND Term2
- Dokumente für Term1: 1, 3, 5, 7
- Dokumente für Term2: 2, 3, 4, 5
- Dokumente für Term1 UND Term2: 3, 5

Boolesches Retrieval - Invertierter Index

- ODER-Verknüpfung mit Hilfe eines invertierten Index
- Anfrage: Term1 ODER Term2
- Dokumente für Term1: 3, 5, 7
- Dokumente für Term2: 2, 3, 5
- Dokumente für Term1 ODER Term2: 2, 3, 5, 7

Boolesches Retrieval - Invertierter Index

- Negation mit Hilfe eines invertierten Index
- Anfrage: Term1 UND NICHT Term2
- Dokumente für Term1: 1, 3, 5, 7
- Dokumente für Term2: 2, 3, 4, 5
- Dokumente für Term1 UND NICHT Term2: 1, 7

Boolesches Retrieval - Unscharfe Suche

Levenstheinabstand

- Wieviel Buchstaben muss ich ersetzen, löschen oder hinzufügen.
- Levstheinabstand für Tier und Tor ist 2:
  i einfügen und e durch o ersetzen
- Nur gleiche Worte haben einen Levenstheinabstand von 0.
- Zur Korrektur von Rechtschreibfehlern geeignet.

Boolesches Retrieval - Unscharfe Suche

Soundex

- Welche Wörter klingen ähnlich?
- Code aus dem ersten Buchstaben und Zahlen zu den folgenden drei Konsonanten

B, F, P, V	1
C, G, J, K, Q, S, X, Z	2
D, T	3
L	4
M, N	5
R	6

- Birkner, Bergen: B-625
- Meier, Maier, Meyer, Mair; M-6
- Besser für die deutsche Sprache: Kölner Phonetik

Boolesches Retrieval - Unscharfe Suche

Dice-Koeffizient

- Wie unterschiedlich sind zwei Wörter bezüglich ihrer N-Gramme?
- $T_N(x)$: N-Gramm-Zerlegung des Terms x
- $T_3(Autokran) = \{Aut, uto, tok, okr, kra, ran\}$
- $T_3(Autobahn) = \{Aut, uto, tob, oba, bah, ahn\}$
- Dice-Koeffizient: $D_N(x, y) = \frac{2 \cdot |T_N(a) \cap T_N(b)|}{|T_N(a)| + |T_N(b)|}$
- $T_3(Autokran) \cap T_3(Autobahn) = \{Aut, uto\}$
- $D_3(Autokran, Autobahn) = \frac{2 \cdot 2}{6+6} \approx 33\%$
- Der Dice-Koeffizient hat immer einen Wert zwischen 0 und 1.

Literatur

- Andreas Henrich: Information Retrieval 1
- `http://www.uni-bamberg.de/minf/ir1_buch/`
- Kapitel 4.2.1, 4.3