

Inhalt

1. Einleitung	1
2. Ein konzeptionelles Modell für Information-Retrieval-Systeme	3
2.1. Das Modell	3
2.2. Klassifikation von Information-Retrieval-Modellen	5
3. Allgemeine Betrachtungen über Ranking	8
3.1. Wozu Ranking?	8
3.2. Die Beziehung zwischen Benutzerfrage und Antwortdokument	9
3.3. Optimales Ranking	10
3.4. Das probabilistische Ranking-Prinzip	11
3.5. Retrievalfunktionen	13
4. Probabilistische Information-Retrieval-Modelle	17
4.1. Indexierungsmodelle mit binärer Unabhängigkeit	17
4.1.1. Das Basismodell	17
4.1.2. Varianten des Basismodells	18
4.1.3. Schätzung der probabilistischen Parameter	20
4.2. Retrievalmodelle mit binärer Unabhängigkeit	21
4.2.1. Das Basismodell	21
4.2.2. Varianten des Basismodells	23
4.2.3. Schätzung der probabilistischen Parameter	25
4.3. Das Unified Model	26
4.3.1. Einführung	26
4.3.2. Grundlegende Annahmen	27
4.3.3. Herleitung des Unified Model	29
4.3.4. Diskussion des konzeptionellen Modells	31
4.3.5. Eine Anwendung des Unified Model	33
4.4. Indexierungsmodelle auf der Basis von Dokumentrepräsentationen	34
4.4.1. Konzeption des Darmstädter Indexierungsansatzes	34
4.4.2. Ein probabilistisches Modell für den DIA	36
4.4.3. Schätzung der probabilistischen Parameter für den DIA	39
4.4.4. Ein probabilistisches Indexierungsmodell für beliebige Retrievalfragen	40
4.4.5. Vergleich mit anderen probabilistischen Indexierungsmodellen	41
4.5. Retrievalmodelle für probabilistische Indexierung	43
4.5.1. Ein Retrievalmodell für das Zwei-Poisson-Modell	43
4.5.2. Ein allgemeines Retrievalmodell für probabilistische Indexierungen	46
4.5.3. Varianten und Anwendung des Modells	49
4.5.4. Korrektheit und Relevanz	50
4.5.5. Vergleich mit anderen probabilistischen IR-Modellen	52
4.6. Grundsätzliche Probleme bei probabilistischen Information-Retrieval-Modellen	55

4.6.1. Schätzung von Wahrscheinlichkeiten	55
4.6.2. Unabhängigkeitsannahmen	58
4.6.3. Das Prinzip der maximalen Entropie	60
4.7. Der Quadratmittel-Polynomansatz	61
4.7.1. Beschreibung des Polynomansatzes	61
4.7.2. Kombination mit probabilistischen Ansätzen	64
4.7.3. Belegung des Merkmalsvektors	66
4.7.4. Mehrstufige Relevanzskalen	67
5. Andere IR-Modelle	70
5.1. Vektorraummodelle	70
5.1.1. Vektorraum und Ähnlichkeitsmaße	70
5.1.2. Binäre Vektorelemente	75
5.1.3. Gewichtete Fragevektoren	77
5.1.4. Gewichtete Dokumentvektoren	80
5.2. Boolesche und Fuzzy-Modelle	83
5.2.1. Boolesches und Fuzzy Retrieval	83
5.2.2. Ansätze für zweistufiges Retrieval	85
5.2.3. Zweistufiges Retrieval mit vollständig-disjunktiver Normalform	88
5.2.4. Zweistufiges Retrieval mit partiell-disjunktiver Normalform	89
5.2.5. Zweistufiges Retrieval mit partiell-kojunktiver Normalform	90
5.2.6. Das „Extended Boolean Model“ von Salton	91
6. Bewertung von Retrievalverfahren	94
6.1. Das Bewertungsproblem	94
6.2. Distributionen	94
6.3. Definition der verwendeten Retrievalmaße	96
6.4. Methoden der Mittelwertbildung	99
6.5. Signifikanztests zum Vergleich von Retrievalergebnissen	102
6.6. Bewertung von Indexierungsverfahren	103
7. Vorbereitung der Experimente	105
7.1. Die Darmstädter Projekte zur automatischen Indexierung	105
7.2. Das Testmaterial für die Retrievalexperimente	107
7.3. Das Testmaterial für die Indexierungsexperimente	112
7.4. Die Systeme UNIDARES und ALIBABA	113
8. Indexierungsexperimente	118
8.1. Berechnung Deskriptor-spezifischer Angaben	118
8.2. Verbesserte Schätzungen für z-Werte	119
8.3. Weiterentwicklung von Indexierungsfunktionen mit dem Polynomansatz	122
8.4. Untersuchung theoretisch begründeter Indexierungsfunktionen	129
8.5. Zusammenfassende Bewertung der Indexierungsergebnisse	138
8.6. Indexierung auf der Basis von Retrievalergebnissen	146
9. Experimente mit booleschen, Fuzzy- und Vektorraum-Modellen	151
9.1. Retrievalfunktionen ohne Berücksichtigung der booleschen Suchlogik	151

9.2. Retrievalfunktionen mit Berücksichtigung der booleschen Suchlogik	154
10. Rankingexperimente mit dem Polynomansatz	157
10.1. Vorgehensweise	157
10.2. Schätzung des frageunabhängigen Dokumentgewichtes	158
10.3. Optimierung heuristischer Ansätze	160
10.4. Untersuchung log-linearer Ansätze	167
10.4.1. Heuristisch motivierte log-lineare Ansätze	167
10.4.2. Theoretisch begründete log-lineare Ansätze	171
10.5. Bewertung der verschiedenen Ansätze	174
10.6. Untersuchung einzelner Parameter	179
10.7. Experimente mit Hinweisen	188
11. Rankingexperimente mit probabilistischen Formeln	192
11.1. Zur Anwendung und Bewertung	192
11.2. Experimente mit dem Unified Model	193
11.3. Experimente mit dem RPI-Modell	197
11.4. Experimente mit Fragetermgewichtung	200
11.4.1. Kollektionsbezogene Gewichte	200
11.4.2. Fragetermgewichtung durch Relevance Feedback	203
12. Zusammenfassende Bewertung der untersuchten Retrievalfunktionen	208
12.1. Zur Vorgehensweise	208
12.2. Vergleich der Retrievalfunktionen für die Indexierung A1	209
12.3. Vergleich der Retrievalfunktionen für die Indexierung I1	213
12.4. Zusammenfassung der Ergebnisse	218
12.5. Wege zu besserem Retrieval	225
Literatur	230