Inhaltsverzeichnis

I Grundlagen und klassische IR-Methoden 1

1 Einführende Beispiele 3
1.1 Literatursuche 5
1.2 Recherche in einer Literaturdatenbank 7
1.3 Faktendatenbanken und -retrieval 10
1.4 Hypertext-Informationssysteme 11
1.5 Expertensysteme 12
1.6 Management-Informationssysteme 13
1.7 Data Mining 14
1.8 Kategorisierung mit einem Data-Mining-System 15
1.9 Assoziative Regeln und der Warenkorb 17
1.10 Wissensgewinnung und Information Retrieval 18

2 Grundlagen 21
2.1 Informationsübertragung 21
2.1.1 Datenübertragung 21
2.1.2 Komplexere Übertragungsbeispiele 22
2.2 Dialoge 24
2.3 Information Retrieval 26
2.3.1 Daten, Wissen, Information 26
2.3.2 Struktur eines Information-Retrieval-Systems 28
2.3.3 Information Retrieval: Definition und Abgrenzung 29

3 Klassische Information-Retrieval-Verfahren 33
3.1 Boolesches Retrieval 33
3.1.1 Logik des booleschen Retrieval 34
3.1.2 Boolesches Retrieval für Textdokumente 34
3.1.3 Implementierung mit invertierten Listen 36
3.1.4 Erweiterungen 38
3.2 Zeichenketten, Wörter und Konzepte 39
3.2.1 Reduktion von Wörtern auf ihre Grundformen 40
3.2.2 Lexikografische Grundformenreduktion nach Kuhlen 42
3.2.3 Lexikonbasierte Morphologie-Analyse 44
3.2.4 Auflösen von Mehrdeutigkeiten 46
3.3 Klassifikationen 47

3.3.1 Internationale Dezimalklassifikation ... 50
3.3.2 Erweiterte Klassifikationssysteme ... 52
3.4 Thesauren ... 54
3.5 Semantische Netze ... 59
3.6 Das Vektorraummodell ... 61
3.6.1 Das Modell ... 62
3.6.2 Vektorraummodell und boolesches Retrieval ... 64
3.6.3 Gewichtungsmethoden ... 66
3.6.4 Globale Gewichtungseinflüsse ... 67
3.6.5 Lokale Gewichtungseinflüsse ... 70
3.6.6 Relevance Feedback ... 72
3.6.7 Ähnlichkeitsfunktionen ... 72
3.6.8 Das Retrieval-System SMART ... 80
3.7 Bewertung und Vergleich von IR-Systemen ... 83
3.7.1 Einflussfaktoren ... 84
3.7.2 Relevanz ... 85
3.7.3 Precision und Recall ... 86
3.7.4 Mittelwertbildungen ... 90
3.7.5 Testkollektionen ... 91
3.7.6 Die TREC-Experimente ... 94

II Wissensgewinnung mit Data-Mining-Methoden 101

4 Lernen ... 103
4.1 Lernen als Informationsverarbeitung ... 106
4.2 Automatisches Lernen aus Beispielen ... 108
4.2.1 Faktendatenbanken ... 108

5 Kategorisieren ... 111
5.1 Attribute und Kategorien ... 111
5.2 Trainings- und Testmenge ... 113
5.3 Lernparadigmen ... 114
5.4 Der ID3-Algorithmus ... 115
5.4.1 Formale Beschreibung des ID3-Algorithmus ... 116
5.4.2 Kategorisieren mit dem ID3-Algorithmus ... 119
5.5 Rahmenbedingungen für Lernalgorithmen ... 119
5.5.1 Konsistenz ... 119
5.5.2 Größe von Entscheidungsbäumen ... 120
5.5.3 Wertebereiche der Attribute ... 121
5.5.4 Bewertung von Kategorisierungsergebnissen ... 123
5.5.5 Inkonsistente Trainingsdaten ... 124
5.5.6 Unvollständige Beispiele ... 126
5.5.7 Größe und Repräsentativität der Trainingsmenge ... 127
5.5.8 Inkrementelles Lernen ... 128

5.5.9 Overfitting ... 129
5.5.10 Suchstrategien ... 129
5.6 Einfache Regelsysteme ... 131
5.6.1 Entscheidungslisten ... 133
5.6.2 Ripple-down-Regelmengen ... 134
5.6.3 Top-down- und Bottom-up-Methoden ... 135
5.7 Der AQ-Algorithmus ... 137
5.7.1 Generalisierungsoperationen ... 143
5.8 Regelsysteme mit zusammengesetzten Attributen ... 143
5.9 Multivariate Entscheidungsbäume ... 145
5.9.1 Attributauswahl ... 147
5.9.2 Sequenzielle Elimination und Auswahl ... 148
5.9.3 Verteilungsbasiertes Eliminationsverfahren ... 148
5.9.4 Das CART-Verfahren ... 149
5.9.5 Koeffizientenbestimmung ... 149
5.9.6 Evaluierung ... 151

6 Cluster und unscharfe Mengen ... 153
6.1 Cluster ... 153
6.2 Unscharfe Mengen ... 155

7 Assoziative Regeln ... 163
7.1 Warenkorbmodell ... 164
7.2 DBLearn/DBMiner ... 166

8 Ein komplexeres Beispiel ... 173
8.1 Problemstellung ... 173
8.2 Lösungsansätze ... 174
8.3 Verfahren ... 174
8.4 Durchführung und Bewertung ... 176

III Erweiterte Retrieval-Ansätze 179

9 Das Vektorraummodell als Fuzzy-Set-Ansatz ... 181
9.1 Verallgemeinerte boolesche Verfahren ... 181
9.1.1 Das MMM-Modell ... 182
9.1.2 Das Paice-Modell ... 183
9.1.3 Das P-Norm-Modell ... 184

10 Der probabilistische Retrieval-Ansatz ... 185
10.1 Wahrscheinlichkeiten in endlichen Mengen ... 185
10.1.1 Beispiel: Würfel ... 186
10.2 Abschätzung des Retrieval-Status-Werts ... 188
10.3 Die Robertson-Sparck-Jones-Formel ... 192

11 Logikbasierte Modelle des Information Retrieval ... 195
11.1 Imaging ... 197
11.2 Bayessche Inferenznetze ... 200
11.3 Abduktive Anfrageoptimierung ... 205

12 Erfolgreiche TREC-Systeme ... 207
12.1 Die TREC-3-Ergebnisse von SMART ... 208
12.2 Die TREC-4-Ergebnisse von SMART ... 210
12.3 Ein Spreading-Activation-Modell ... 215
12.4 INQUERY in TREC-4 ... 217
12.5 Das Okapi-System ... 219
12.6 Spezialaufgaben (TREC Tracks) ... 221

13 Korpusbasierte Verfahren ... 223
13.1 Der assoziative Ansatz im IR ... 223
13.2 Kookurrenzverfahren ... 226
13.2.1 Ein Machine-Learning-Ansatz ... 226
13.2.2 Term-Term-Matrizen ... 227
13.2.3 Anwendung im IR ... 228
13.2.4 Häufigkeit der Terme ... 228
13.2.5 Expansion von Termen oder Anfragen ... 231
13.2.6 Größe der Dokumentensammlung ... 231
13.2.7 Eine Untersuchung zur Bestimmung von Suchtermen ... 231
13.2.8 Komplexere Kookurrenzverfahren ... 232
13.3 Anwendung im mehrsprachigen Retrieval ... 233
13.4 Deskriptoren bestimmen ... 235
13.5 Latent Semantic Indexing ... 239
13.6 Gewichtungsmethoden Lernen ... 239
13.7 Social oder Collaborative Filtering ... 241

IV Information Retrieval und das Web 245

14 Explizit strukturierte Dokumente ... 247
14.1 Standard Generalized Markup Language (SGML) ... 248
14.1.1 SGML-Elemente ... 248
14.1.2 Elementattribute ... 250
14.1.3 SGML-Entities ... 252
14.2 HTML ... 253
14.3 XML ... 254
14.3.1 Verweise: XPointer und XLink ... 255
14.3.2 XML Schema ... 255
14.3.3 XPath, XQuery ... 256
14.4 Suche nach und in XML-Dokumenten ... 258
14.4.1 Anwendungen von XML bei der Suche ... 258
14.4.2 Indexierungsmethoden ... 259

14.4.3 Modelle für die Suche in XML-Dokumenten . . . 261
14.4.4 Ein Vektorraummodell für strukturierte Anfragen an Sammlungen von XML-Dokumenten . . . 262
14.4.5 Suche bei unterschiedlichen DTDs . . . 265

15 Metadaten . . . 267
15.1 Dublin-Core-Metadaten . . . 268
15.2 Hierarchisch strukturierte Metadaten . . . 272
15.3 PICS . . . 275
15.4 RDF und das Semantische Web . . . 276
15.4.1 Resource Description Framework . . . 276
15.4.2 Pläne für ein Semantisches Web . . . 281

16 Suche im World Wide Web . . . 285
16.1 Das Web als Dokumentensammlung . . . 285
16.1.1 Medienarten . . . 286
16.1.2 Sprache . . . 287
16.1.3 Länge und Granularität . . . 287
16.1.4 Dynamik und Alter von Web-Seiten . . . 288
16.1.5 Anbieter und ihre Ziele . . . 289
16.1.6 Zielgruppen . . . 290
16.1.7 Inhalte . . . 291
16.1.8 Spamming . . . 291
16.2 Suchmechanismen der Web-Protokolle . . . 292
16.3 Hierarchische Verzeichnisse oder Web Directories . . . 295
16.3.1 Klassifikation des Open Directory Project . . . 296
16.4 Web-Suchmaschinen . . . 299
16.4.1 Web-Roboter, Crawler oder Spider . . . 300
16.4.2 Ranking-Strategien . . . 302
16.4.3 Ranking nach externen Daten . . . 303
16.4.4 Metasuchdienste . . . 306
16.5 Spezialisierte und verteilte Sammlungen . . . 308
16.5.1 Der Z39.50-Standard . . . 309
16.5.2 Beispiele verteilter Sammlungen . . . 310
16.5.3 Peer-to-Peer-Netze . . . 313
16.6 Digitale Bibliotheken . . . 316
16.6.1 Inhalte einer digitalen Bibliothek . . . 318
16.6.2 Dienste . . . 319
16.6.3 Archivierung . . . 320

Literaturverzeichnis . . . 323

Index . . . 333