

Inhalt

Vorwort	9
Suchmaschinen im Internet - informationswissenschaftlich betrachtet	9
1 Einleitung	13
2 Forschungsumfeld	21
2.1 Suchmaschinen-Markt	21
2.2 Formen der Suche im WWW	24
2.3 Aufbau von algorithmischen Suchmaschinen	26
2.4 Abfragesprachen	30
2.5 Arten von Suchanfragen	33
2.6 Nutzerstudien	34
2.6.1 Methoden der Nutzerforschung	35
2.6.2 Nutzergruppen	36
2.6.3 Recherchekennnisse und -strategien	36
2.6.4 Themen und Auswahl der Suchbegriffe	37
2.6.5 Sichten der Treffer	38
2.6.6 Wissen über Suchmaschinen und deren Finanzierung	38
2.7 Forschungsbereiche	39
3 Die Größe des Web und seine Abdeckung durch Suchmaschinen	41
3.1 Die Größe des indexierbaren Web	42
3.2 Struktur	45
3.3 Crawling	48
3.4 Aktualität	50
3.5 Invisible Web	51
4 Strukturinformationen	59
4.1 Strukturierungsgrad von Dokumenten	59
4.2 Strukturinformationen in den im Web gängigen Dokumenten	60
4.2.1 HTML	61
4.2.2 Word-Dokumente	65
4.2.3 PDF	66
4.3 Trennung von Navigation, Layout und Inhalt	67
4.4 Repräsentation der Dokumente in den Datenbanken der Suchmaschinen	68

5	Klassische Verfahren des Information Retrieval und ihre Anwendung bei WWW-Suchmaschinen	71
5.1	Unterschiede zwischen „klassischem“ Information Retrieval und Web Information Retrieval	71
5.2	Kontrolliertes Vokabular	77
5.3	Kriterien für die Aufnahme in den Datenbestand	78
5.4	Modelle des Information Retrieval	80
5.4.1	Boolesches Modell	80
5.4.2	Vektorraummodell	83
5.4.3	Probabilistisches Modell	86
6	Ranking	89
6.1	Rankingfaktoren	90
6.2	Messbarkeit von Relevanz	95
6.3	Grundsätzliche Probleme des Relevance Ranking in Suchmaschinen	97
7	Informationsstatistische und informationslinguistische Verfahren	99
7.1	Textstatistische Verfahren	99
7.2	Nutzungsstatistische Verfahren	101
7.3	Informationslinguistische Verfahren	104
7.3.1	Stemming	106
7.3.2	Phrasenerkennung	109
7.3.3	Synonyme, Homonyme, Akronyme	111
7.3.4	Rechtschreibkontrolle	113
8	Linktopologische Rankingverfahren	117
8.1	Grundlagen: Science Citation Indexing	118
8.2	PageRank	120
8.2.1	Der klassische PageRank-Algorithmus	120
8.2.2	Weiterentwicklungen: Reranking	123
8.3	HITS	126
8.4	Hilltop	130
8.5	Evaluierung der linktopologischen Verfahren	132
8.6	Problembereiche linktopologischer Rankingverfahren	134
8.7	Fazit linktopologische Verfahren	137

9 Retrievaltests	139
9.1 Aufbau und Nutzen von Retrievaltests.....	139
9.2 Aufbau und Ergebnisse ausgewählter Retrievaltests.....	142
9.3 Kritik	145
10 Verfahren der intuitiven Benutzerführung	149
10.1 Relevance Feedback	151
10.2 Vorschläge zur Erweiterung und Einschränkung der Suchanfrage	154
10.3 Klassifikation und Thesaurus	159
10.4 Clusterbildung	161
10.5 Graphische Ansätze der Ergebnispräsentation	165
11 Aktualität	169
11.1 Bedeutung der Beschränkung nach der Aktualität der Dokumente	169
11.2 Funktionsfähigkeit der Datumsbeschränkung in Suchmaschinen	170
11.2.1 Methodik.....	171
11.2.2 Ergebnisse	174
11.3 Möglichkeiten der Ermittlung von Datumsangaben in Web-Dokumenten...180	
11.4 Aktualitätsfaktoren im Ranking	182
11.5 Spezialisierte Suchmaschinen für Nachrichten.....	187
11.6 Auswahl der gewünschten Aktualität durch den Nutzer	188
12 Qualität	191
12.1 Bedeutung der Beschränkung nach der Qualität der Dokumente.....	192
12.2 Qualitätsbeschränkungen bei der Recherche in Datenbank-Hosts.....	192
12.3 Identifizierung von Top-Quellen im WWW.....	194
12.4 Manuelle Einbindung von Top-Quellen	195
12.5 Automatisierte Einbindung von Invisible-Web-Quellen	198
12.6 Einbindung von Web-Verzeichnissen in Suchmaschinen	200
12.6.1 Erschließung des Web mittels Suchmaschinen und Verzeichnissen.201	
12.6.2 Web-Verzeichnisse und ihre Integration in Suchmaschinen	203
12.6.3 Erschließung der Sites in Web-Verzeichnissen	204
12.6.4 Einbindung der Verzeichnisdaten in Suchmaschinen	206

13 Verbesserung der Dokumentrepräsentation	217
13.1 Beschränkung auf den Inhaltsteil der Dokumente	217
13.2 Erweiterungen der Dokumentrepräsentation	221
13.2.1 Strukturinformationen.....	221
13.2.2 Größenangaben.....	222
13.2.3 Abbildungen und Tabellen	223
13.3 Ersatz für die Nicht-Verwendbarkeit generischer Top-Level-Domains	224
13.4 Aufbereitung der Suchergebnisse in den Trefferlisten	224
14 Fazit und Ausblick	227
Literatur	231
Register	243