

Inhaltsverzeichnis

1	Einführung	1
1.1	Was ist Semantische Suche?	3
1.2	Wozu benötigt man Semantische Suche?	4
1.3	Notwendigkeit von Vorwissen	6
1.4	Sprachlücke zwischen Autoren und Nutzern	7
1.5	Probleme der Verarbeitung deutschsprachiger Texte	9
1.6	Evolution der Schriftsprache	10
1.7	Suche – eine informationstheoretische Sicht	10
1.8	Suchanfragen	12
1.9	Nutzung von Hintergrundwissen	13
1.10	Relevanzbewertung und Treffersortierung	14
1.11	Sichtbarkeit des Hintergrundwissens	15
1.12	Zusammenfassung	15
	Literatur	16
2	Grundlagen der Textverarbeitung	17
2.1	Token, Wörter, Terme, Entitäten, Benannte Entitäten	19
2.2	Tokenisierung	20
2.3	Bedeutungstragende Bezeichnungen	23
2.4	Nominalkomposita	23
2.5	Schreibfehlererkennung und -korrektur	25
2.5.1	Schreibfehlerkorrektur	26
2.5.2	Schreibfehlerkorrekturvorschläge	30
2.6	N-Gramme	31
2.7	Kookkurrenzen und Kollokationen	32
2.7.1	Kookkurrenzen auf der Basis gegenseitiger Information	33
2.7.2	Kookkurrenzen auf der Basis statistischer Tests	34
2.7.3	Übertragung auf Kollokationen	35
2.7.4	Vergleich für Nomen-Nomen Bi-Gramme	35
2.8	Part-of-Speech-Tagging	36
2.9	Nominalphrasen	43
2.10	Erkennung benannter Entitäten	45
2.11	Erkennung von anwendungsgebiets-spezifischen Entitäten	46
2.12	Stoppwortentfernung	47
2.13	Stammformableitung	49
2.13.1	Lemmatisierung	49
2.13.1.1	Morphologielexikon	50
2.13.1.2	Lemmatisierung auf Basis von POS-Tags	50
2.13.1.3	Online-Services	51
2.13.2	Stemming	52
2.14	Normalisierung	53
2.15	Phonetische Kodierung	54

2.16	Und wie spielt das alles zusammen?	56
2.16.1	Basisarchitektur	57
2.16.2	Tippfehlertolerante Architektur	58
2.16.3	Standardarchitektur für Volltextsuche	59
2.16.4	Substandard-Architektur	60
2.17	Weiterführende Literatur	60
	Literatur	61
3	Grundlagen des Information Retrievals	63
3.1	Repräsentation von Dokumenten	65
3.1.1	Term-Dokument-Inzidenzmatrix	65
3.1.2	Einfacher Invertierter Index	67
3.1.3	Invertierter Index zur Umsetzung eines Wildcard-Operators	68
3.1.4	Term-Dokument-Inzidenzmatrix (die Zweite)	69
3.2	Interpretation von Suchanfragen	72
3.3	Anfrage-Operatoren	73
3.4	Boolesche Anfragen an einen invertierten Index	75
3.4.1	AND-Operator über einem invertierten Index	75
3.4.2	Weitere Boolesche Operatoren im Selbstbau	77
3.5	Erweiterte Anfragen an einen positionellen invertierten Index	77
3.5.1	NEAR-Operator über positionellen invertierten Index	79
3.5.2	Komplexere Anfragen	81
3.6	Ranking der Ergebnisse	81
3.6.1	Dokumentfaktoren	82
3.6.1.1	Unterscheidung und Gewichtung von Inhalten	82
3.6.1.2	Termfrequenz	83
3.6.1.3	Inverse Dokumentfrequenz	84
3.6.1.4	Termfrequenz und Inverse Dokumentfrequenz kombiniert	85
3.6.2	Netzwerkfaktoren	86
3.6.2.1	PageRank	86
3.6.2.2	Distanz- und Ähnlichkeitsmaße auf Graphen	87
3.6.2.3	Pfadlänge	87
3.6.2.4	Gewichtung nach Tiefe	88
3.6.2.5	Gewichtung nach Kantentyp	88
3.6.3	Anfragefaktoren	89
3.6.3.1	Reihenfolge der Anfrageterme	89
3.6.3.2	Termhäufigkeit in Anfragen	90
3.6.4	Vektorraummodell	90
3.6.5	Kosinus-Ähnlichkeit	91
3.6.5.1	Vereinfachte Kosinus-Ähnlichkeit	92
3.6.6	Soft-Kosinus-Maß	93
3.6.6.1	Boolesches Synonym-Retrieval mit Soft-Kosinus-Ranking	94
3.7	Probleme der Vektorraumrepräsentation	95
3.7.1	Effekt der beharrlichen Dokumente	95
3.7.2	Kollabierende Bedeutungen	96
3.7.3	Verlust von Begriffsabhängigkeiten	98

3.8	Und wie spielt dies jetzt alles zusammen?	101
3.9	Weiterführende Literatur	105
	Literatur	106
4	Grundlagen der Wissensrepräsentation	109
4.1	Begriffe und mehr	111
4.1.1	Begriffe	111
4.1.2	Klassen und Instanzen	112
4.1.3	Beziehungen	113
4.1.4	Schema und Fakten	113
4.2	Wissensorganisation: Vom Vokabular zur Ontologie	114
4.2.1	Kontrolliertes Vokabular	115
4.2.2	Taxonomie	116
4.2.3	Thesaurus	116
4.2.4	Wortnetze	118
4.2.5	Ontologie	118
4.2.6	Knowledge Graph	121
4.2.7	Und was benötigen wir davon für eine semantische Suche?	121
4.3	Wichtige Standards	121
4.3.1	RDF	122
4.3.2	RDFa	124
4.3.3	RDFS	124
4.3.4	SKOS	126
4.3.5	Schema.org	127
4.3.6	OWL	128
4.3.7	SPARQL	129
4.4	Linked Data	130
4.5	Technologien	131
4.6	Und woher kommt das Wissensmodell?	131
4.7	Weiterführende Literatur	133
	Literatur	133
5	Bausteine Semantischer Suche	135
5.1	Komponenten zur Semantifizierung konventioneller Suchfunktionen	138
5.1.1	Keyword-Tools	138
5.1.2	Anfrageerweiterung	140
5.1.3	Ergebnisfilterung	141
5.1.3.1	Boolesche Filterung	142
5.1.3.2	Semantische Filterung	144
5.2	Komponenten zur Textaufbereitung	145
5.2.1	Zusammenfassung zusammengesetzter Ausdrücke	146
5.2.2	Vereinheitlichung von Schreibweisen	146
5.2.2.1	Kompositazerlegung	148
5.2.3	Rechtschreibfehlerkorrektur	150
5.2.4	Phonetische Kodierung	152
5.3	Verschlagwortung von Dokumenten und Anfragen	153

5.3.1	Manuelle Verschlagwortung	156
5.3.2	Automatische Extraktion von Schlagwörtern und Phrasen	157
5.3.3	Linktext-Analyse	158
5.3.4	Halbautomatische Verschlagwortung anhand von kontrolliertem Vokabular	159
5.3.5	Automatische Verschlagwortung anhand von kontrolliertem Vokabular	161
5.3.5.1	Verschlagwortung mit einem allgemeinen Wissensgraphen	162
5.3.5.2	Verschlagwortung mit spezialisiertem Hintergrundwissen	164
5.3.5.3	Technische Realisierung der Verschlagwortung mit spezialisiertem Hintergrundwissen	168
5.3.6	Verschlagwortung von Anfragen	171
5.3.7	Exkurs: Disambiguierung	172
5.4	Indexierung von Annotationen.	174
5.4.1	Annotationen technisch betrachtet	174
5.4.1.1	Beziehungen zwischen Dokumenten, Annotationen und Wissensmodellen	174
5.4.1.2	Repräsentation von Annotationen	175
5.4.2	Invertierter Index über Annotationen	175
5.5	Benutzerschnittstellen-Komponenten	176
5.5.1	Semantische Auto-Vervollständigung	176
5.5.2	Facettierte Suche	181
5.5.3	Wissensbrowser	186
5.5.4	Erklärung von Suchanfrage-Erweiterungen	188
5.5.5	Hervorheben gefundener Begriffe in Snippets	190
5.5.5.1	Hervorheben gefundener Begriffe bei einer semantischen Suche	191
5.5.5.2	Technische Realisierung der Hervorhebung gefundener Begriffe	192
5.5.6	Hervorheben gefundener Begriffe in Dokumenten	195
5.6	Weiterführende Literatur	198
	Literatur	199
6	Konstruktionsprinzipien semantischer Suchverfahren	201
6.1	Definitionsansätze	202
6.2	Abgrenzung	203
6.2.1	Abgrenzung nach anderen Kriterien	206
6.2.2	Weitere Unterscheidungskriterien	207
6.3	Referenz-Architektur semantischer Anwendungen	207
6.4	Semantische Suche	209
6.4.1	Semantische Suche durch Anfrageerweiterung	209
6.4.2	Semantische Suche in Dokumentströmen	212
6.4.3	Semantische Suche über Annotationen	213
6.4.3.1	Triple-Store	215
6.4.3.2	Konventionelle relationale Datenbank	216
6.4.3.3	Konventionelle Volltextsuchmaschine	218
6.4.3.4	Beeinflussung des Rankings	219
6.4.3.5	Anreicherung von Annotationen um semantische Begriffsabstände	220
6.4.4	Hybride Semantische Suche über Annotationen und Volltext	224
6.5	Weiterführende Literatur	226
	Literatur	227

7	Lösungen	229
7.1	▶Kap. 2	230
7.2	▶Kap. 3	232
7.3	▶Kap. 4	236
7.4	▶Kap. 5	241
7.5	▶Kap. 6	244
	Literatur.....	246
	 Serviceteil	
	Web-Adressen	248
	Weiterführende Literatur	251
	Stichwortverzeichnis.....	253