Inhaltsverzeichnis

Geleitwort i

Vorwort iii

Abbildungsverzeichnis xi

Tabellenverzeichnis xiii

Abkürzungsverzeichnis xv

Symbolverzeichnis xvii

1 Einleitung 1

1.1 Problemstellung 1

1.2 Zielsetzung der Arbeit 3

1.3 Aufbau der Arbeit 4

2 Grundlegende Definitionen und Methoden 7

2.1 Information-Filtering und -Retrieval 7

2.1.1 Information-Retrieval 7

2.1.2 Information-Filtering 10

2.1.3 Gemeinsamkeiten und Unterschiede 11

2.1.4 Modell der Repräsentation von Dokumenten 12

2.1.5 Modell der Interaktion mit dem Benutzer 13

2.2 Datenmodelle 14

2.2.1 Entity-Relationship-Modelle 16

2.2.2 Relationale Datenbanken und SQL 19

2.3 Computerlinguistik 20
2.3.1 Phonologie 20
2.3.2 Morphologie 21
2.3.2.1 Flexion, Komposition und Derivation 21
2.3.2.2 Stemming (Normalisierung) 21
2.3.3 Syntax 23
2.3.3.1 Syntaktische Strukturen und formale Grammatiken 23
2.3.3.2 Automatisierte Analyse syntaktischer Strukturen 26
2.3.4 Semantik 29
2.3.4.1 Satz- und Diskurssemantik 29
2.3.4.2 Lexikalische Semantik 30
2.3.5 Pragmatik 33
2.3.6 Bedeutung für IF und IR 35
2.4 Ontologien 35
2.4.1 Ontologie-Modellierungssprachen 38
2.4.2 Anwendungsmöglichkeiten für IF- und IR-Systeme 41

3 Gängige IF/IR-Modelle **43**
3.1 Fundamentale Konzepte 45
3.2 Modelle ohne Terminterdependenzen 49
3.2.1 Standard Boolean Model (SBM) 49
3.2.2 Vector Space Model (VSM) 50
3.2.3 Extended Boolean Model (EBM) 52
3.2.4 Binary Independence Retrieval (BIR) 54
3.2.5 Inference Network Model (INM) 56
3.2.6 Belief Network Model (BNM) 59
3.2.7 Language Model (LM) 60
3.3 Modelle mit immanenten Terminterdependenzen 63
3.3.1 Generalized Vector Space Model (GVSM) 68
3.3.2 Latent Semantic Index (LSI) 69
3.3.3 Spreading Activation Neuronal Network (SANN) 70
3.4 Modelle mit transzendenten Terminterdependenzen 72
3.4.1 Fuzzy Set Model (FSM) 73
3.4.2 Retrieval by Logical Imaging (RbLI) 76
3.4.3 Backpropagation Neuronal Network (BNN) 79

3.5 Bewertung der gängigen Modelle 81

4 Topic-based Vector Space Model (TVSM) 87

4.1 Motivation 87
4.2 Konzept 88
4.2.1 Vektorraum, Terme und Dokumente 89
4.2.2 Dokumente und Dokumentenähnlichkeiten 90
4.2.3 Berechnung der Dokumentenähnlichkeiten 93
4.2.4 Implementierung mit einer relationalen Datenbank 94
4.2.5 Einstellen neuer Dokumente / Durchführen von Anfragen 98
4.3 Stoppwort-Lemma 99
4.4 Stemming-Lemma 101
4.5 Synonym-Lemma 103
4.6 Vergleich mit anderen Modellen 104
4.7 Kritik am TVSM 107

5 Enhanced TVSM (eTVSM) 109

5.1 Konzept 110
5.1.1 Paarweise Themen-Ähnlichkeiten 115
5.1.1.1 Problemstellung 115
5.1.1.2 Repräsentationsform für Themenstrukturen 116
5.1.1.3 Herleitung der Themen-Ähnlichkeiten 120
5.1.1.4 Eigenschaften der Vektoren und Ähnlichkeiten 123
5.1.2 Interpretationen und ihre Beziehungen 128
5.1.2.1 Herleitung der Interpretations-Ähnlichkeiten 129
5.1.2.2 Repräsentation der (Totalen) Synonymie 131
5.1.2.3 Repräsentation der Homographie 132
5.1.2.4 Repräsentation der Partiellen Synonymie 134
5.1.2.5 Repräsentation der Metonymie 136
5.1.2.6 Definition der Dokumenten-Ähnlichkeiten 137
5.1.3 Wortstamm-Term-Zuordnung 139
5.1.4 Wort-Wortstamm-Zuordnung 140
5.2 Das eTVSM und der Ontologie-Begriff 143
5.2.1 Eine grafische Ontologie-Repräsentation für das eTVSM 143
5.2.2 Eine Beispiel-Ontologie 144

5.3 Umsetzung mit einer relationalen Datenbank 148
5.3.1 Datenmodell 148
5.3.2 Initialisierung 151
5.3.2.1 Initialisierung der Themenblätter 151
5.3.2.2 Initialisierung der Themenknoten 155
5.3.2.3 Ähnlichkeiten zwischen Themen 157
5.3.2.4 Skalarprodukte zwischen Interpretationen 158
5.3.3 Einstellen neuer Dokumente 159
5.3.3.1 Parsing 160
5.3.3.2 Zuordnung zu Interpretationen 162
5.3.3.3 Berechnung der Dokumentenbeträge 166
5.3.4 Anfrageausführung 166
5.4 Vergleich mit anderen Modellen / Kritik 168

6 Anwendung des eTVSM in der Praxis **171**
6.1 Ontologien für das eTVSM 171
6.1.1 Erstellung einer Ontologie 171
6.1.2 Nutzung vorhandener Ontologien 172
6.1.2.1 Wortschatz-Lexikon 172
6.1.2.2 WordNet und GermaNet 174
6.2 Anwendung für das Information-Retrieval 182
6.3 Anwendung für das Information-Filtering 183
6.4 Quantitative Evaluierung 187
6.4.1 Evaluationsmaße 188
6.4.2 Evaluation von IR-Systemen 190
6.4.3 Evaluation von IF-Systemen 191

7 Zusammenfassung **193**

A Datenbankeinträge der Beispiel-Ontologie **195**
A.1 Vor der Initialisierung bekannt 195
A.1.1 Tabelle `Thema` 195
A.1.2 Tabelle `Themenstruktur` 196
A.1.3 Tabelle `Interpretation` 197
A.1.4 Tabelle `IT_Zuo` 197
A.1.5 Tabelle `Term` 198

A.1.6 Tabelle `TI_Zuo` 199
A.1.7 Tabelle `Supportterm` 200
A.1.8 Tabelle `Wortstamm` 200
A.1.9 Tabelle `WT_Zuo` 201
A.1.10 Tabelle `Wort` 202
A.2 Nach der Initialisierung 203
A.2.1 Tabelle `Themavektor` 203
A.2.2 View `ThemenAehnlichkeit` 205
A.2.3 Tabelle `Aehnlichkeit` 210
A.3 Dokumente einfügen 215
A.3.1 Tabelle `Dokument` vor der Betragsberechnung 215
A.3.2 Tabelle `DW_Zuo` 216
A.3.3 Tabelle `DI_Zuo` 216
A.3.4 Tabelle `Dokument` nach der Betragsberechnung 217
A.4 Dokumentenähnlichkeit 217
A.4.1 Ergebnisse der View `DokAehn` 217

B VSM: simuliert mit den eTVSM-Tabellen **219**
B.1 View-Definitionen 219
B.1.1 Anzahl der Terme pro Dokument 219
B.1.2 Dokumentabhängige Termgewichte 220
B.1.3 Berechnung der Dokumentenähnlichkeit 221
B.2 View-Ergebnisse 221
B.2.1 View `VSM_a` 221
B.2.2 View `VSM_w` 222
B.2.3 View `VSM_dokaehn` 222

Literaturverzeichnis **225**

Index **239**