

Inhaltsverzeichnis

Kurzfassung-----	iii
Summary-----	v
Deskriptoren-----	vii
Danksagung-----	ix
Kapitel 1: Einleitung -----	1
1.1 Motivation und Zielsetzung der Arbeit-----	1
1.2 Neuere Arbeiten zum Themenfeld-----	4
1.3 Aufbau der Arbeit-----	8
1.4 Allgemeine Definitionen-----	8
1.5 Begriffserklärung aus der Computerlinguistik und Sprachtechnologie-----	10
Kapitel 2: Linguistische Analyseverfahren -----	13
2.1 Automatische Termextraktion-----	13
2.1.1 Morphologische Analyse (Das MPRO-Programm)-----	15
2.1.2 Grammatische Analyse-----	20
2.1.3 MPRO-Anwendungen-----	25
2.1.4 Termextraktion und Gewichtung-----	25
2.2 Indexierung-----	26
2.2.1 Indexierungsverfahren:-----	26
2.2.2 Automatische Indexierung-----	27
2.2.3 AUTINDEX System-----	29
2.3 Zusammenfassung der Aufbereitung von Termextraktionsergebnissen-----	36
Kapitel 3: Termgewichtung und Termoptimierung -----	37
3.1 Grundlagen der Wahrscheinlichkeitstheorie und Statistik-----	38
3.1.1 Bedingte Wahrscheinlichkeit-----	38
3.1.2 Bayes'sche Klassifikation (Bayesian classification)-----	38
3.1.3 Bayes-Entscheidungsregel-----	39
3.1.4 Bayes-Klassifikator für die Dokumentklassifizierung-----	39
3.1.5 Definition (3.1): Der Mittelwert (Arithmetisches Mittel)-----	39
3.1.6 Standardabweichung-----	40
3.1.7 Varianz-----	40
3.1.8 Kovarianz-----	40
3.1.9 Korrelation (Pearson- Korrelationskoeffizient)-----	41

3.2	Evaluierung und Relevanz-----	42
3.3	Die grundlegenden Termgewichtungsfaktoren-----	42
3.3.1	Termhäufigkeit (term frequency tf)-----	42
3.3.2	Dokumenthäufigkeit (document frequency df)-----	43
3.3.3	Inverse Dokumentenhäufigkeit (Inverse document frequency idf)-----	43
3.4	Gewichtungsverfahren -----	43
3.4.1	Lokale Verfahren-----	43
3.4.2	Lokale und globale Gewichtungsverfahren -----	44
3.4.3	Globale Verfahren -----	44
3.5	Gewichtung und statistische Filterung der Textterme -----	47
3.5.1	Meine Sortier- und Gewichtungsregeln (Sort Values)-----	47
3.6	Optimierung von Klassentermen-----	53
3.6.1	Methode der mittleren Gewichte-----	54
3.6.2	Methode der optimierten Klassenterme-----	55
3.6.3	Vorgehensweisen bei der Entwicklung der Sortier- bzw. Gewichtungsregel-----	55
3.7	Die DATEV-Fallstudie (Klassentermoptimierung)-----	58
3.7.1	Anwendung der Sortierregeln an den Klassen von DATEV-----	58
3.7.2	Fazit der Evaluierungsergebnisse der Sortierregel-----	59
3.8	Fallstudie -WISSMER (automatische Thesauruserweiterung) -----	61
3.8.1	Methode 1: Statistische Filterung:-----	62
3.8.2	Methode 2: Nur wahrscheinliche Optimierung und Filterung (idf>0) -----	62
3.8.3	Methode 3: Statistische und wahrscheinliche Optimierung und Filterung-----	63
Kapitel 4: Dokumentenähnlichkeit -----	65	
4.1	Einleitung-----	65
4.2	Ähnlichkeitsmaße (Distanzfunktionen)-----	67
4.2.1	Das Kosinus-Ähnlichkeitsmaß-----	68
4.2.2	Das Korrelationsähnlichkeitsmaß-----	69
4.2.3	Vergleich der Ähnlichkeitsmaße -----	71
4.2.4	Anwendungsmöglichkeiten der Ähnlichkeitsmatrizen-----	73
4.3	Dokumentenähnlichkeit -----	74
4.3.1	Ermittlung der Dokumentenähnlichkeit (Prozessphasen) -----	74
4.3.2	Die Methode der erweiterten Korrelation-----	83
4.3.3	Ergebnisse und Verbesserungsvorschläge -----	84
4.4	Siemens-Fallstudie-----	85

4.4.1	Ziel der Fallstudie	85
4.4.2	Die Prozessphasen.....	85
4.4.3	Stärken und Schwächen der Ähnlichkeitsergebnisse.....	91
4.4.4	Endgültige Struktur der Ähnlichkeitsausgabe.....	93
4.4.5	Die Ähnlichkeitsergebnisse im Siemens-Portal	94
4.4.6	Ähnlichkeitsauswertung	96
4.4.7	Verbesserungsvorschläge.....	96
Kapitel 5: Dokumentklassifizierung		97
5.1	Automatische Klassifizierung.....	97
5.2	Das automatische statistische Klassifizierungssystem	98
5.2.1	Aufgabenstellung.....	99
5.2.2	Prozesskette des Klassifizierungsvorgangs.....	99
5.2.3	Bewertung und Evaluierung des Klassifikationssystems	105
5.2.4	Verbesserung der Klassifizierungsergebnisse.....	106
5.2.5	Durchführung der Klassifizierungsverfahren in Fallstudien.....	108
5.2.6	Zusammenfassung des Klassifizierungsmechanismus	108
5.3	FIZ-Fallstudie(2007-2009)	109
5.3.1	Der Klassifizierungsvorgang mit AUTINDEX.....	109
5.3.2	Die Klassifizierungsprozesskette in der FIZ-Fallstudie.....	110
5.3.3	Bewertung der Klassifizierungsergebnisse durch verschiedene Kriterien	115
5.3.4	Evaluierung der Klassifizierungsergebnisse	116
5.3.5	Zusammenfassung der Ergebnisse	117
5.3.6	Ausblick.....	118
5.4	Wolters Kluwer-Fallstudie	119
5.4.1	Prozesskette in WK-Fallstudie.....	120
5.4.2	Die Klassifizierungsergebnisse	123
5.5	Optimierung der aufgebauten Klassenterme.....	126
5.5.1	Anwendung der Sortierregeln bei den Klassen der FIZ-Fallstudie	126
5.5.2	Fazit der automatischen Evaluierung des Klassifizierungssystems nach der Klassentermoptimierung (2009)	133
5.5.3	Fazit der Evaluierung der Sortierregeln:	136
5.6	SIEMENS-Fallstudie (2009)	137
5.6.1	Durchführung des Projekts:	137
5.6.2	Systemaufbau.....	137
5.6.3	Ergebnisse der automatischen Klassifizierung.....	140

5.6.4	Bewertung der automatischen Klassifikation	141
5.6.5	Weitere Verbesserungsvorschläge	143
5.6.6	Endgültige Klassifizierungsergebnisse nach Klassenoptimierung	143
5.6.7	Endgültiges Fazit der Siemens-Fallstudie	144
5.7	Fazit und Ausblick	145
5.7.1	Zusammenfassung der Dokumentklassifizierung (Alle Fallstudien)	145
5.7.2	Ausblick	146
Kapitel 6 :Wortwolkenermittlung		147
6.1	Einführung	147
6.1.1	Zielsetzung der Wortwolkenermittlung	148
6.1.2	Verfahren zur Ermittlung von Wortwolkenbegriffen	148
6.2	Die allgemeinen Schritte der Wortwolkenermittlung	149
6.2.1	Aufbereitung der Dokumentensets	149
6.2.2	Bearbeitung der Stichwörter bzw. Suchanfragen und der Stoppwortliste	149
6.2.3	Linguistische Termextraktion	150
6.2.4	Generierung der Wortwolkenrelationen	150
6.2.5	Optimierung der Wortwolkenergebnisse	156
6.2.6	Abschließende Arbeiten	163
6.3	Arten der Ermittlung von Wortwolken	164
6.3.1	Das bestimmte wahrscheinlichste Verfahren	164
6.3.2	Das adaptive Verfahren	164
6.4	Verwendung der Wortwolkenenergebnisse	172
6.5	Anwendung der Wortwolkenermittlung in Fallstudien	172
6.6	Die DATEV-Fallstudie	173
6.6.1	Einführung	173
6.6.2	Prozessphasen	173
6.7	Die Wien-Fallstudie	204
6.8	Auflösung von Akronymen auf Basis der Wortwolken	208
6.9	Zusammenfassung und Ausblick	210
Kapitel 7: Disambiguierung auf Basis der Wortwolken		213
7.1	Automatische Disambiguierung von mehrdeutigen Begriffen	213
7.2	Kernidee	213
7.3	Abkürzungsverzeichnis	214
7.4	Arten der Mehrdeutigkeit	214

7.5	Disambiguierungsalgorithmen und -methoden	214
7.5.1	Überwachtes Lernen	214
7.5.2	Unüberwachtes Lernen	215
7.6	Meine Ansätze für Disambiguierung	216
7.6.1	Mein unüberwachtes Auflösungsverfahren	216
7.6.2	Mein überwachtes Auflösungsverfahren:	217
7.7	Aufbau meines Disambiguierungssystems	217
7.7.1	Aufbereitung des Trainingsmaterials	217
7.7.2	Trainieren des Systems	219
7.7.3	Termextraktion und Gewichtung der extrahierten Texte	223
7.7.4	Ermittlung der Wortwolken	226
7.7.5	Disambiguierung der Mehrdeutigkeiten	241
7.7.6	Evaluierung und Bewertung des Disambiguierungssystems	248
7.7.7	Bewertung aller Methoden der Auflösung	257
7.7.8	Evaluierungsfazit	257
7.7.9	Der endgültige Aufbau des Disambiguierungssystems	258
7.8	Das endgültige Fazit	259
	Kapitel 8: Zusammenfassung und Ausblick	261
	Kapitel 9: Anhänge	263
	Abbildungsverzeichnis	265
	Tabellenverzeichnis	269
	Literaturverzeichnis	273
	Erklärung	279
	Lebenslauf des Promovenden	281