

## Inhalt

Zusammenfassung .....	9
Abstract.....	11
1. Vom Textvergleich zur Wiedergewinnung von Information.....	12
1.1 Informationssysteme und natürlichsprachliche Information .....	13
1.2 Nicht-grammatikalische Textvergleiche .....	15
1.3 Beiträge der Arbeit.....	16
2. Sprachliche Einheiten, Indexierung und Ähnlichkeit .....	19
2.1 Natürliche Sprache und natürlichsprachliche Texte.....	19
2.2 Wortfragmente und Informationsspuren .....	21
2.3 Automatische vs. intellektuelle Indexierung .....	23
2.4 Ähnlichkeitsfunktionen.....	26
3. Text und Sprache: beschreibende Statistik vs. Wahrscheinlichkeiten .....	29
3.1 Statistische Betrachtungen.....	29
3.1.1 Textstatistik.....	29
3.1.2 Häufigkeitwörterbücher und Worthäufigkeitsmodelle.....	34
3.1.3 Das Gesetz von Zipf .....	36
3.2 Auf Wahrscheinlichkeiten beruhende Betrachtungen.....	38
3.2.1 Information und Entropie.....	38
3.2.2 Redundanz.....	39
3.2.3 Markov-Ketten .....	40
3.3 Informationsspuren und Markov-Ketten .....	41
3.3.1 Das Modell.....	42
3.3.2 Bedeutung für n-Gramm-basierte Textvergleichssysteme.....	46
4. Reduktionsmechanismen zur Filterung von Texten.....	49
4.1 Stoppwortlisten .....	49
4.2 Wortreduktion .....	50
4.2.1 Klassifikation von Reduktionsalgorithmen.....	51
4.2.2 Wortreduktion im Englischen.....	52
4.3 Methoden zur Wortreduktion .....	54
4.3.1 Der T- und der S-Algorithmus .....	54
4.3.2 Der Algorithmus von Lovins.....	54
4.3.3 Der Algorithmus von Porter.....	55
5. Textvergleich mittels Informationsspuren .....	59

5.1	Wortfragmente als operationale Grundeinheiten .....	59
5.2	Trigramme .....	61
5.2.1	Praktikabilität .....	61
5.2.2	Statistische Eigenschaften von Trigrammen .....	63
5.2.3	Selektivität und Rekonstruierbarkeit .....	64
5.3	Der de Heer'sche Ansatz .....	66
5.3.1	Homeosemie .....	66
5.3.2	Direkte und indirekte Ähnlichkeit .....	67
6.	n-Gramm-Retrieval auf reduzierten Texten .....	71
6.1	Probleme des Ansatzes von de Heer .....	71
6.2	Semantische Ergänzung der syntaktischen Methode von de Heer .....	72
6.2.1	Reduzierte Texte .....	73
6.2.2	Erweiterter Porter-Algorithmus .....	74
6.2.3	Gestrafte Spuren machen Vergleiche sicherer .....	76
6.3	Erweiterte Informationsspuren .....	79
6.4	Ähnlichkeitsfunktionen und Antwortmengen .....	83
6.4.1	Direkte Ähnlichkeitsfunktion .....	84
6.4.2	Indirekte Ähnlichkeitsfunktion .....	86
6.4.3	Bestimmung eines "Cutoff-Punktes" .....	87
6.5	Das System VISIR .....	88
7.	Visualisierung von Informationsspuren .....	90
7.1	Dokumente als Punkte in mehrdimensionalen Räumen .....	90
7.2	Warum Visualisierung von Daten .....	92
7.2.1	Erkundende Datenanalyse .....	92
7.2.2	Daten über natürlichsprachliche Dokumente .....	93
7.3	Methoden zur Visualisierung multivariater Daten .....	94
7.3.1	Methoden zur graphischen Darstellung von Punkten aus mehrdimensionalen Räumen .....	95
7.3.2	Parallele Koordinaten .....	99
7.4	Methoden zur Visualisierung von Informationsspuren .....	101
7.4.1	Übergangsmatrizen .....	104
7.4.2	Trigrammdiagramme .....	105
7.4.3	Übergangsdigramme .....	107
7.5	Beispiele .....	108
7.6	Visualisierungen ergänzen numerische Resultate .....	114
8.	Implementation eines Prototypen und Testergebnisse .....	116
8.1	Systemumgebung und Zielsetzung .....	116
8.2	Informationsstruktur .....	118

8.3	Modulhierarchie.....	121
8.3.1	Übersicht.....	121
8.3.2	Basismodule.....	123
8.4	Vorbereitungsphase.....	124
8.5	Indexierung.....	127
8.6	Auswertung.....	128
8.7	Visualisierung.....	129
8.8	Testergebnisse.....	131
8.8.1	Die Dokumentenkollektion.....	131
8.8.2	Der Testrahmen.....	132
8.8.3	Die Resultate.....	133
9.	Bewertungen und Ausblick.....	141
	Glossar.....	144
	Literatur.....	147