

Inhaltsverzeichnis

1 Einleitung	1
1.1 Information Retrieval	2
1.1.1 Indexierung von Textdokumenten	2
1.1.2 Thesauruskonstruktion	3
1.1.3 Unterstützung bei der Formulierung von Suchfragen	3
1.2 Retrieval-Anforderungen an Dokumentenmanagementsysteme	4
2 Information-Retrieval-Verfahren in Dokumentenmanagementsystemen	7
2.1 Indexierung	7
2.1.1 Stemming-Verfahren	9
2.1.2 Einsatz von Thesauren	11
2.1.3 Termgewichtung	11
2.2 Modelle zur Berechnung von Dokumentenähnlichkeiten	12
2.2.1 Vektorraummodell	12
2.2.2 Latent Semantische Indexierung	15
2.2.3 Probabilistische Modelle	16
2.3 IR-Verfahren in Dokumentenmanagementsystemen	16
2.3.1 Volltextrecherche	17
2.3.2 Ähnlichkeitssuche	18
2.3.3 Push-Technologien	18
2.3.4 Textkategorisierung	19
2.3.5 Text-Clustering	20
2.3.6 Automatic Abstracting	23
2.3.7 Recherche nach relevanten Dokumenten in unstrukturierten Datenbanken	24
3 Typografische Termgewichtung	29
3.1 Stand der Technik	29
3.2 HTML-Tag-Gewichtung	31
3.2.1 Absolute Gewichtung von HTML-Tags	31
3.2.2 Relative Gewichtung von HTML-Tags	34
3.3 Allgemeine typografische Termgewichtung	35
3.3.1 Richtlinien zur Typografie von Textdokumenten	35
3.3.2 Entwickeltes Verfahren zur typografischen Termgewichtung	36
3.4 Feature-Selektion und Typografiegewichtung	39
3.4.1 Bekannte binäre Feature-Bewertungsverfahren	40
3.4.2 Relative Feature-Bewertung	43

3.4.3 Globale Auswahl der Kategorie-Features	45
4 Evaluierung der typografischen Termgewichtung	47
4.1 Bewertungsverfahren für Klassifikationsprobleme	47
4.1.1 Fehlerrate	48
4.1.2 Precision und Recall	48
4.1.3 Precision-Recall-Breakeven-Punkt	49
4.1.4 F-Maß	49
4.1.5 Mikro- und Makro-Bewertung	49
4.2 Klassifikationsverfahren für Textdokumente	50
4.2.1 k-NN-Klassifizierer	51
4.2.2 SVM-Klassifizierer	52
4.3 Beschreibung der Testkollektionen	55
4.3.1 Anforderungen an die Testkollektionen	55
4.3.2 Die WebKB-Testkollektionen	56
4.3.3 Die ACM-Testkollektion	56
4.3.4 Die C-LAB Marketing, Sales & PR Testkollektion	57
4.3.5 Vergleich der Testkollektionen	57
4.4 Messungen	58
4.4.1 Evaluierung des relativen k-NN-Verfahrens	60
4.4.2 Feature-Selektionsverfahren für k-NN-Klassifizierer	60
4.4.3 Vergleich der Typografiegewichtungsverfahren	67
4.4.4 Vergleich von SVM- und k-NN-Klassifizierern	70
4.4.5 Gesamtverbesserung der vorgestellten Verfahren	72
4.5 Typografische Termgewichtung in Cluster-Verfahren	76
4.5.1 Cluster-Qualität	77
4.5.2 Evaluierung	77
4.6 Zusammenfassung der Evaluierungsergebnisse	80
5 Das VKC-System	83
5.1 Geschichte des VKC-Systems	83
5.2 Die Datenschicht des VKC-Systems	85
5.2.1 Die Datenbank-Persistenzschicht	85
5.2.2 Die Archivschicht	85
5.2.3 Die Indexschicht	86
5.3 Die Logikschicht des VKC-Systems	86
5.3.1 Information-Retrieval-Komponenten	86
5.3.2 Dokumentenmanagement-Komponente	87
5.3.3 Benutzermanagement-Komponente	88
5.3.4 Projektmanagement-Komponente	89
5.3.5 Messenger-Komponente	91
5.3.6 Systemverwaltung	91
5.4 Die Präsentationsschicht des VKC-Systems	93
5.4.1 HTML Model-View-Controller 2	93
5.4.2 WebDAV-Schnittstelle	94
5.4.3 Web-Services-Schnittstelle	96
5.5 VKC-Konverter-Server	97
5.6 Die VKC-Workflow-Komponente	98
5.7 Verteilung der VKC-Komponenten zur Leistungssteigerung	99

5.7.1 Die Verteilung auf mehrere Servlet-Container	99
5.7.2 Der parallele Einsatz mehrerer Konverter-Server	100
6 Die Information-Retrieval-Funktionen des VKC-Systems	103
6.1 Typografiegewichtung	103
6.1.1 Das Typography Description Format	103
6.1.2 Berechnung der TDF-Gewichte	105
6.2 Realisierung des IR-Index	106
6.2.1 Indexierung der TDF-Dateien	107
6.2.2 VKC-Suchfunktionen	108
6.3 Klassifikationsfunktion	109
6.3.1 Die praktische Bedeutung der Trefferrate in DMS	111
6.3.2 Bestimmung der optimalen k-NN-Parameter	112
6.3.3 Bestimmung der Vorhersagequalität	115
6.4 Clustering-Funktion	116
6.5 Analysefunktion	117
6.6 Der C-LAB-Retriever	118
7 Zusammenfassung und Ausblick	121
Literaturverzeichnis	123
Index	129