

Contents

1	Introduction	1
1.1	Problem	1
1.2	Research Gap	3
1.3	Research Objective	4
1.4	Thesis Outline and Prior Publications	5
1.4.1	Publications	6
2	Academic Plagiarism Detection	9
2.1	Definition and Typology of Plagiarism	9
2.1.1	Typologies of Academic Plagiarism	11
2.1.2	Our Typology of Academic Plagiarism	12
2.2	Research on Academic Plagiarism	14
2.3	Plagiarism Detection Paradigms	16
2.3.1	External Plagiarism Detection	16
2.3.2	Intrinsic Plagiarism Detection	19
2.4	Plagiarism Detection Methods	21
2.4.1	Preprocessing	21
2.4.2	Similarity Measures	22
2.4.3	Lexical Detection Methods	25
2.4.4	Syntax-based Detection Methods	33
2.4.5	Semantics-based Detection Methods	34
2.4.6	Idea-based Detection Methods	41
2.4.7	Hybrid Detection Methods	43
2.4.8	Plagiarism Detection Systems	48

2.5	Evaluation of PD Methods and Systems	50
2.5.1	Evaluation Datasets for Plagiarism Detection	50
2.5.2	Evaluation of Plagiarism Detection Methods	57
2.5.3	Evaluation of Plagiarism Detection Systems	69
2.6	Findings of the Literature Review	75
2.7	Research Approach	77
3	Citation-based Plagiarism Detection	79
3.1	Citation-based Document Similarity	79
3.1.1	Terminology	79
3.1.2	Related Work	80
3.2	Citation-based PD Concept	81
3.2.1	Conceptualization of the Detection Approach	82
3.2.2	Challenges to Citation Pattern Identification	85
3.3	Citation-based Detection Methods	86
3.3.1	Bibliographic Coupling	86
3.3.2	Longest Common Citation Sequence	87
3.3.3	Greedy Citation Tiling	87
3.3.4	Citation Chunking	88
3.4	Evaluation Methodology	94
3.4.1	Dataset	95
3.4.2	Evaluated Detection Methods	96
3.4.3	Corpus Preprocessing	98
3.4.4	Applying Detection Methods and Pooling	100
3.4.5	Addressing False Positives	102
3.4.6	Collecting Relevance Judgments	104
3.5	Results	106
3.5.1	Retrieval Effectiveness	106
3.5.2	Computational Efficiency	113
3.5.3	User Utility	116
3.6	Conclusion Citation-based PD	119
4	Image-based Plagiarism Detection	121
4.1	Related Work and Research Gap	122
4.2	Types of Image Similarity	124
4.2.1	Exact Copies	125
4.2.2	Near-Duplicate Images	125
4.2.3	Altered Images	126
4.2.4	Visualizing Reused Data	128
4.3	Requirements Analysis	131

4.4	Image-based Detection Process	132
4.4.1	Image Extraction and Decomposition	133
4.4.2	Image Classification	134
4.4.3	Perceptual Hashing	135
4.4.4	OCR Preprocessing	135
4.4.5	n -gram Text Matching	136
4.4.6	Positional Text Matching	137
4.4.7	Ratio Hashing	138
4.4.8	Scoring	140
4.5	Evaluation	143
4.5.1	Discussion of Results	146
4.6	Conclusion Image-based PD	148
5	Math-based Plagiarism Detection	149
5.1	Math-based PD Concept	149
5.1.1	Related Work and Research Gap	151
5.2	Conceptualization of Detection Methods	153
5.2.1	Investigation of Plagiarism Cases	153
5.2.2	Devising Mathematical Feature Comparisons	154
5.3	Evaluation Dataset	156
5.3.1	Document Preprocessing	158
5.3.2	Dataset Statistics	159
5.4	Preliminary Experiments	160
5.4.1	Performance Measures	160
5.4.2	Results	161
5.5	Math-based Detection Process	162
5.5.1	Indexing	163
5.5.2	Candidate Retrieval	164
5.5.3	Detailed Analysis	165
5.6	Evaluation Methodology	166
5.7	Results	167
5.7.1	Confirmed Cases of Plagiarism	167
5.7.2	Exploratory Study	173
5.8	Conclusion Math-based PD	176
6	Hybrid Plagiarism Detection System	179
6.1	System Overview	179
6.2	Backend	183
6.2.1	Supported Document Formats	183
6.2.2	Conversion and Parsing	183

6.2.3	Feature Extraction & Indexing	184
6.2.4	Detection Process	185
6.3	Frontend	185
6.3.1	Dashboard View	186
6.3.2	Results Overview	189
6.3.3	Detailed Analysis View	192
6.4	Conclusion Hybrid PD System	195
7	Conclusion and Future Work	197
7.1	Summary	197
7.2	Contributions of the Thesis	202
7.3	Future Work	205
7.3.1	Increase Detection Effectiveness	205
7.3.2	Improve Plagiarism Detection Systems	211
7.3.3	Other Applications	215
7.3.4	Plagiarism Policy Implications	217
	Glossary	219
	References	225