

Contents

Acknowledgements	V
Contents	VII
List of Tables	XI
List of Figures	XIII
Glossary	XVII
Abstract	XXIII
Kurzfassung	XXV
1 Introduction	1
1.1 Problem Setting	1
1.2 Motivation	3
1.3 Research Objective	4
1.4 Thesis Outline	6
2 Plagiarism Detection	9
2.1 Academic Plagiarism	9
2.1.1 Definition	10
2.1.2 Forms of Academic Plagiarism	11
2.1.3 Prevalence of Plagiarism in the Academic Environment	13
2.2 Plagiarism Detection Approaches	17
2.2.1 Generic Detection Approach	17
2.2.2 Overview of Plagiarism Detection Approaches	19
2.2.3 Fingerprinting	22
2.2.4 Term Occurrence Analysis	26
2.2.5 Stylometry	30
2.2.6 Cross-Language Plagiarism Detection	32
2.3 Plagiarism Detection Systems	33
2.3.1 Evaluations of PDS	34
2.3.2 Technical Weaknesses of PDS	39
2.4 Conclusion	40

3	Citation-based Document Similarity.....	43
3.1	Terminology.....	44
3.1.1	Citation vs. Reference.....	44
3.1.2	Similarity vs. Relatedness.....	45
3.1.3	Dimensions of Similarity: Lexical, Semantic, Structural.....	45
3.2	Citation-based Similarity Measures.....	47
3.2.1	Direct Citation.....	47
3.2.2	Bibliographic Coupling.....	48
3.2.3	Co-citation.....	50
3.2.4	Amsler.....	52
3.2.5	Co-citation Proximity-based Methods.....	52
3.3	Conclusion.....	54
4	Citation-based Plagiarism Detection.....	57
4.1	Concept.....	58
4.1.1	Citing Behavior.....	62
4.2	Citation Characteristics Considered.....	64
4.2.1	Bibliographic Coupling Strength.....	64
4.2.2	Probability of Citation Co-occurrence.....	64
4.2.3	Order and Proximity of Citations.....	65
4.3	Challenges to Citation Pattern Identification.....	66
4.3.1	Unknown Pattern Constituents.....	66
4.3.2	Transpositions.....	67
4.3.3	Scaling.....	67
4.3.4	Insertions or Substitutions of Citations.....	68
4.4	Design of Citation-based Detection Algorithms.....	68
4.4.1	Bibliographic Coupling (BC).....	69
4.4.2	Longest Common Citation Sequence (LCCS).....	70
4.4.3	Greedy Citation Tiling (GCT).....	70
4.4.4	Citation Chunking (Cit-Chunk).....	73
4.5	Projected Suitability of CbPD Algorithms for Plagiarism Forms.....	80
4.6	Assessment of Identified Citation Patterns.....	83
4.6.1	Citing Frequency-Score (CF-Score).....	83
4.6.2	Continuity-Score (Cont.-Score).....	85

4.7	Conclusion	87
5	Prototype: CitePlag.....	89
5.1	Document Parser	90
5.2	Database.....	93
5.2.1	Consolidation of Reference Identifiers	94
5.3	Detector.....	95
5.4	Frontend	96
5.5	Conclusion	99
6	Quantitative and Qualitative Evaluation.....	101
6.1	Methodology	102
6.1.1	Test Collection Requirements.....	104
6.1.2	Test Collection Challenges	106
6.1.3	GuttenPlag Wiki	108
6.1.4	VroniPlag Wiki.....	109
6.1.5	PubMed Central OAS	111
6.1.6	Summary and Comparison of Test Collections	113
6.2	Evaluation using GuttenPlag Wiki.....	116
6.3	Evaluation using VroniPlag Wiki	121
6.3.1	Evaluation: Random Sample of Sources.....	121
6.3.2	Evaluation: Translated Plagiarism	126
6.3.3	Evaluation: Plagiarism Case Heun.....	129
6.3.4	Conclusion VroniPlag Wiki.....	135
6.4	Evaluation using PubMed Central OAS.....	136
6.4.1	Methodology.....	139
6.4.2	Results	157
6.4.3	Conclusion of PMC OAS Evaluation	195
6.5	Conclusion of Evaluations	198
7	Summary & Future Work.....	203
7.1	Summary	203
7.2	Cóntributions	207
7.3	Future Work.....	211
7.3.1	General Research Need.....	212

7.3.2	Improvements to Detection Accuracy	213
7.3.3	Additional Applications	215
7.3.4	Further Evaluations	219
	References	223
	Appendix	265
A	Preliminary PMC OAS Corpus Analysis	266
A.1	Bibliographic Coupling	266
A.2	Longest Common Citation Sequence	273
A.3	Greedy Citation Tiling	278
A.4	Citation Chunking	286
A.5	Character-based PDS Sherlock	293
A.6	Character-based PDS Encoplot	294
B	Technical Details of the CitePlag Prototype	296
B.1	Sentence-Word-Tagger (SW-Tagger)	296
B.2	Data Parser	300
B.3	Consolidation of Reference Identifiers	302
B.4	Database Documentation	304
C	Data and Source-code Downloads	311
D	Related Publications	313
E	Patent Application	318
F	User Study Feedback	329
G	Reactions of Contacted Authors	331
H	Empirical Studies on Plagiarism Frequencies	336
I	Studies on Citation-based Similarity Measures	339
J	Overview of Selected PDS	343
	Index	347