

1 Introduction

This doctoral thesis addresses an unsolved information retrieval problem: the automatic detection of disguised plagiarism forms, including paraphrases, translated plagiarism and structural and idea plagiarism.

Section 1.1 of this chapter introduces the problem setting of currently non-machine-detectable academic plagiarism. Section 1.2 describes my motivation for research, and Section 1.3 presents the resulting research objective pursued in this thesis. Section 1.4 provides an outline of the thesis.

1.1 Problem Setting

The problem of academic plagiarism¹ has been present for centuries. Yet the widespread dissemination of information and communication technology, including the Internet, has greatly contributed to the ease of plagiarizing. Many online services exist to facilitate student plagiarism, including essay databases, and text "synonymizer" tools, such as *synomizer.com*², which outputs input text with a list of synonyms for each word.

The most extensive study on plagiarism surveyed ~82,000 students at North American colleges. Approximately 40 % of the students admitted having plagiarized within the last year [220]. However, students are not the only group to plagiarize. In Germany, more than 30 prominent cases of academic dishonesty among politicians recently made headlines. The German politicians who plagiarized in their doctoral theses include former Minister of Defense, Karl-Theodor zu Guttenberg, and even the Federal Minister of Education and Research, Annette Schavan. The question arises why cases of plagiarism, which are apparent in hindsight, often remain undiscovered for so long. Why can academic misconduct not be caught much earlier using plagiarism detection software?

¹ Refer to Section 2.1.1, page 10, for a definition of plagiarism.

² <http://www.synomizer.com>

D. Weber-Wulff, who conducts regular performance evaluations for Plagiarism Detection Systems (PDS), gives a disillusioning summary regarding available systems:

“[...] Plagiarism Detection Systems find copies, not plagiarism.”
([357], p. 6)

Substantial research on the approaches and systems aiding in the detection of plagiarism has been performed for almost two decades. Currently available PDS use sophisticated and highly efficient character-based text comparisons. These approaches are capable of detecting verbatim and moderately disguised copies of text reliably. However, the cleverly veiled and re-structured real-world plagiarism more commonly found in research contains insufficient character-based similarities, making it undetectable by current PDS.

Today, manual inspection of suspicious documents by experts or through crowd-sourced projects, such as the VroniPlag Wiki [350], an online platform used to expose plagiarism cases, represents the only reliable method to detect more heavily disguised plagiarism. However, the time commitment required to examine plagiarism manually is significant. The 48 cases³ in the VroniPlag Wiki alone amounted to hundreds of hours, making manual inspection and crowd-sourced examination unfeasible for examining lower-profile plagiarism or for checking entire databases.

³ As of 2013-07-04. The VroniPlag Wiki is an ongoing project.

1.2 Motivation

My motivation to research new approaches to plagiarism detection grew out of my disillusionment with the state-of-the-art systems. Current software solutions label themselves "plagiarism detectors". This is a misnomer because it leads users to believe the software is indeed capable of detecting real-world plagiarism, including the disguised plagiarism more common to research. In reality, however, this is not the case.

While I believe that plagiarism should not be tolerated in student assignments, I find that plagiarism in research – and particularly in the medical field – has far more serious consequences to society. An example of a plagiarized medical study⁴ [165] in Table 1, illustrates this point. The plagiarism discusses the correct care for patients suffering from acute respiratory distress syndrome. The key difference between the plagiarism and the original study are the numbers stated in the results section. The excerpt from the medical study's results in Table 1 highlights the differences in reported values between the earlier and later publication in red. Both the original and the plagiarism were retrieved from an openly available subset of PubMed's medical publication database.

⁴ This study was identified because it was retrieved among the top results by the approach presented in this thesis. As I later discovered, the study had already been retracted by the journal, although at the time of evaluation it was still available in the database. Visit <http://citeplag.org/compare/5583/117324> for a visual comparison of the plagiarism and the original.

Table 1: Excerpt from a Plagiarized Section Describing Experimental Results

Original [48] PMCID: 1065018	Plagiarism [281] PMCID: 2772258
PEEP had no effect on CO2 gap (median [range], baseline: 19 [2–30] mmHg; PEEP 10: 19 [0–40] mmHg; PEEP 15: 18 [0–39] mmHg; PEEP 20: 17 [4–39] mmHg; ideal PEEP: 19 [9–39] mmHg; $P=0.18$). Cardiac index also remained unchanged (baseline: 4.6 [2.5–6.3] l min ⁻¹ m ⁻² ; PEEP 10: 4.5 [2.5–6.9] l min ⁻¹ m ⁻² ; PEEP 15: 4.3 [2–6.8] l min ⁻¹ m ⁻² ; PEEP 20: 4.7 [2.4–6.2] l min ⁻¹ m ⁻² ; ideal PEEP: 5.1 [2.1–6.3] l min ⁻¹ m ⁻² ; $P=0.08$).	PEEP had no effect on CO2 gap (median [range], baseline: 18 [2–30] mmHg; PEEP 10: 18 [0–40] mmHg; PEEP 15: 17 [0–39] mmHg; PEEP 20: 16 [4–39] mmHg; ideal PEEP: 19 [9–39] mmHg; $P=0.19$). Cardiac index also remained unchanged (baseline: 4.7 [2.6–6.2] l min ⁻¹ m ⁻² ; PEEP 10: 4.4 [2.5–7] l min ⁻¹ m ⁻² ; PEEP 15: 4.4 [2.2–6.8] l min ⁻¹ m ⁻² ; PEEP 20: 4.8 [2.4–6.3] l min ⁻¹ m ⁻² ; ideal PEEP: 4.9 [2.4–6.3] l min ⁻¹ m ⁻² ; $P=0.09$).

Plagiarized studies typically do not only copy text, but are also more likely to contain fictitious evaluations and results. Such fake medical studies jeopardize the quality of medical research and can prevent patients from receiving optimal treatment⁵. Furthermore, for the progression of scientific disciplines it is crucial that researchers can trust the outcomes of past research. This motivated me to develop a plagiarism detection approach better capable of detecting disguised plagiarism as it occurs in higher education and in scientific research.

1.3 Research Objective

Motivated by the limitations of existing plagiarism detection systems, the following research objective was defined:

⁵ For examples of harmful studies, refer to Section 7.3.4.

Propose, implement, and evaluate a plagiarism detection approach capable of detecting non-machine-identifiable plagiarism forms, such as paraphrases, translated plagiarism, and idea plagiarism.

To achieve this objective the following research tasks were derived:

- Task 1:** *Perform a comprehensive analysis of the individual strengths and weaknesses of state-of-the-art plagiarism detection approaches and systems.*
- Task 2:** *Develop a plagiarism detection concept that addresses the identified weaknesses of current plagiarism detection approaches.*
- Task 3:** *Design detection algorithms that employ the theoretical concept introduced and are fitted to detect the plagiarism forms currently not machine-detectable.*
- Task 4:** *Implement a prototype of a plagiarism detection system that employs the developed algorithms to demonstrate the applicability of the approach in real-world scientific document collections.*
- Task 5:** *Evaluate the proposed concept in identifying strongly disguised plagiarism forms by comparing detection performance, user utility, and computational efficiency to state-of-the-art systems. As proof of concept, identify unknown and currently non-machine-detectable plagiarism instances.*

1.4 Thesis Outline

Chapter 1 describes the problem setting, the research motivation, and the corresponding research objective. The research objective is divided into five research tasks pursued in this thesis.

Chapter 2 introduces the reader to the problem of academic plagiarism and the existing research on plagiarism detection. Following a definition of what constitutes plagiarism and the prevalent forms of plagiarism, the scope of plagiarism in the academic and scientific environments is discussed. A detailed examination of current plagiarism detection approaches is given, and the challenges of detecting disguised and translated plagiarism are explained. This chapter addresses Research Task 1 by reviewing and exposing strengths and weaknesses of available plagiarism detection approaches.

Chapter 3 provides background information on citation-based document similarity measures. After introducing relevant terminology, a review of the literature introduces important measures, including Bibliographic Coupling and Co-citation Analysis.

Chapter 4 presents the novel detection approach proposed in this thesis. I coined this approach Citation-based Plagiarism Detection (CbPD). CbPD addresses weaknesses of current plagiarism detection approaches. By analyzing citation similarities within documents, CbPD can machine-detect currently non-automatically detectable disguised forms of plagiarism. Chapter 4 addresses Research Task 2 and Task 3 by proposing CbPD as a plagiarism detection approach and designing detection algorithms using the introduced concept.

Chapter 5 describes the implementation of the Citation-based Plagiarism Detection approach in a prototype, thus addressing Research Task 4.

Chapter 6 describes the CbPD evaluation framework and presents the evaluation results. In the methodology section potential test collections, ground truths and limitations of the evaluation are discussed. Chapter 6 addresses Research Task 5 by evaluating the effectiveness of the proposed approach for both known and yet unknown plagiarism cases.

Chapter 7 provides a summary, discusses research contributions, and gives an outlook on future work. The appendix includes a list of related publications, the preliminary corpus analysis, the CPA/CbPD patent application, material related to the prototype, and other resources as listed below.

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

I will use "we" rather than "I" in the subsequent chapters of this thesis, since I published and discussed my ideas with others including my advisor and fellow researchers. For more information on joint projects and publications, please refer to the acknowledgements in Appendix D.

Citation-based Plagiarism Detection
Detecting Disguised and Cross-language Plagiarism
using Citation Pattern Analysis

Gipp, B.

2014, XXVI, 350 p. 70 illus., Softcover

ISBN: 978-3-658-06393-1