Citation-Based Plagiarism Detection: Practicability on a Large-Scale Scientific Corpus[1]

Bela Gipp
Department of Statistics, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: gipp@berkeley.edu

Norman Meuschke
Department of Statistics, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: meuschke@berkeley.edu

Corinna Breitinger
SciPlore Research Group, University of California, Berkeley, 493 Evans Hall, Berkeley, CA 94720.
E-mail: breitinger@berkeley.edu

The automated detection of plagiarism is an information retrieval task of increasing importance as the volume of readily accessible information on the web expands. A major shortcoming of current automated plagiarism detection approaches is their dependence on high character-based similarity. As a result, heavily disguised plagiarism forms, such as paraphrases, translated plagiarism, or structural and idea plagiarism, remain undetected. A recently proposed language-independent approach to plagiarism detection, Citation-based Plagiarism Detection (CbPD), allows the detection of semantic similarity even in the absence of text overlap by analyzing the citation placement in a document's full text to determine similarity. This article evaluates the performance of CbPD in detecting plagiarism with various degrees of disguise in a collection of 185,000 biomedical articles. We benchmark CbPD against two character-based detection approaches using a ground truth approximated in a user study. Our evaluation shows that the citation-based approach achieves superior ranking performance for heavily disguised plagiarism forms. Additionally, we demonstrate CbPD to be computationally more efficient than character-based approaches. Finally, upon combining the citation-based with the traditional character-based document similarity visualization methods in a hybrid detection prototype, we observe a reduction in the required user effort for document verification.

Introduction

Automated plagiarism detection (PD) is a task supported by specialized information retrieval systems termed *plagiarism detection systems* (PDS). PDS employ one of two detection approaches, intrinsic or extrinsic. Today's commercially available PDS rely exclusively on the extrinsic approach, meaning that they consult an external collection, typically a subset of the web, against which to compare suspicious text. The retrieval task is then to return from this collection all documents that contain text passages similar above a chosen threshold to segments in the suspicious document (Stein, Lipka, & Prettenhofer, 2011).

Intrinsic detection approaches statistically examine the linguistic characteristics of a text without comparisons with an external collection and have been explored less frequently (Meyer zu Eissen, Stein, & Kulig, 2007; Stein et al., 2011). Intrinsic approaches have not been commercially adopted, mainly because of the obstacles posed by the minimum required document length and the possibility of legitimate style differences through author collaboration, which can lead to false positives. In an evaluation of intrinsic approaches by Stein et al., documents under 35,000 words were excluded for not being reliably analyzable (Stein et al., 2011).

Extrinsic PDS typically follow a retrieval process that comprises several phases during which the systems successively narrow down the retrieval space to allow for increasingly fine-grained and computationally more expensive text comparisons. The initial phase typically involves some form of computationally moderate heuristic retrieval step, for example, using fingerprinting indices or vector space models

[1] This paper presents research results from the currently unpublished doctoral thesis of the primary author (Gipp, 2013).

Received September 25, 2013; revised December 19, 2013; accepted December 20, 2013

● Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23228

Citation data for this article

B. Gipp, N. Meuschke, and C. Breitinger. Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus. *Journal of the American Society for Information Science and Technology*, 65 (2): 1527–1540, 2014. doi: 10.1002/asi.23228.

RIS Format	**BibTeX Format**

TY - JOUR
AU - Gipp, Bela
AU - Meuschke, Norman
AU - Breitinger, Corinna
T1 - Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus
JO - Journal of the American Society for Information Science and Technology
Y1 - 2014
VL - 65
IS - 2
SP - 1527
EP - 1540
DO - 10.1002/asi.23228
	@ARTICLE{Gipp13b,
	author = {Gipp, Bela and Meuschke, Norman and Breitinger, Corinna},
	title = {Citation-based Plagiarism Detection: Practicability on a Large-scale Scientific Corpus},
	journal = {Journal of the American Society for Information Science and Technology},
	year = {2014},
	volume = {65},
	number = {2},
	pages = {1527--1540},
	doi = {10.1002/asi.23228}
	}