*Erschienen in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval : SIGIR '18. - New York, USA : ACM Press, 2018. - S. 1321-1324. - ISBN 978-1-4503-5657-2*
*https://dx.doi.org/10.1145/3209978.3210177*

# HyPlag: A Hybrid Approach to Academic Plagiarism Detection

Norman Meuschke, Vincent Stange, Moritz Schubotz, Bela Gipp
Department of Computer and Information Science
University of Konstanz, Germany
{first.last}@uni-konstanz.de

## ABSTRACT

Current plagiarism detection systems reliably find instances of copied and moderately altered text, but often fail to detect strong paraphrases, translations, and the reuse of non-textual content and ideas. To improve upon the detection capabilities for such concealed content reuse in academic publications, we make four contributions: i) We present the first plagiarism detection approach that combines the analysis of mathematical expressions, images, citations and text. ii) We describe the implementation of this hybrid detection approach in the research prototype HyPlag. iii) We present novel visualization and interaction concepts to aid users in reviewing content similarities identified by the hybrid detection approach. iv) We demonstrate the usefulness of the hybrid detection and result visualization approaches by using HyPlag to analyze a confirmed case of content reuse present in a retracted research publication.

## KEYWORDS

Plagiarism Detection; Document Retrieval; Mathematical Information Retrieval; Citation Analysis; Image Retrieval

## 1 INTRODUCTION

Academic plagiarism (AP) has been defined as *'the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected'* [3]. Forms of AP range from copying content (copy&paste) to reusing slightly modified content, e.g., interweaving text from multiple sources, to heavily concealing content reuse, e.g., by paraphrasing or translating text, and lastly, reusing data or simply ideas without proper attribution [19]. The easily recognizable copy&paste-type AP is more prevalent among students [9], while concealed AP is more characteristic of researchers, who have strong incentives to avoid detection [2]. Plagiarized student assignments typically have no consequences for the public. However, plagiarized research publications can have a severe negative impact by distorting the mechanisms for tracing and correcting research results, and causing inefficient allocations of research funds. Therefore, detecting concealed AP in research publications is a pressing problem affecting many stakeholders, including academic publishers, research institutions, funding agencies, and of course other researchers.

## 2 RELATED WORK

Text retrieval research has yielded mature systems that reliably detect copied or moderately altered text in an input document and retrieve its source if the source is included in the system's reference collection. Such systems are well-suited to detect AP of the copy&paste type. Yet, they often fail to find concealed forms of AP, such as paraphrases, translations, or idea plagiarism [19].

Researchers have proposed numerous approaches to improve the text similarity assessment methods, e.g., semantic and syntactic analyses to better identify paraphrases, or cross-language retrieval to better detect translations [2, 10, 18].

Research also showed that hybrid approaches, i.e., the combined analysis of text and other content features, improve the retrieval effectiveness for plagiarism detection (PD) tasks. Alzahrani et al. combined an analysis of text similarity and structural similarity [1]. We showed that the combined analysis of citation patterns and text similarity improves the identification of concealed AP [5–7, 11]. Pertile et al. confirmed the positive effect of combining citation and text analysis and devised a hybrid approach using machine learning [15]. Recently, we demonstrated the potential of analyzing the similarity of mathematical expressions [13], semantic concept patterns [14], and images [12] for improving the detection of AP.

Concluding from prior research, we see a hybrid approach that analyses heterogeneous content features as most promising to prevent and detect the wide range of AP forms.

## 3 SYSTEM OVERVIEW

HyPlag is a research prototype that realizes a hybrid approach to plagiarism detection for academic documents. The system analyzes mathematical expressions, images, citations, and text to improve the identification of potentially suspicious content similarity, particularly in research publications, such as journal articles, PhD theses, and grant proposals. The target audience of our system are reviewers of such works, e.g., journal editors or PhD advisors.

Figure 1 gives an overview of the hybrid PD approach currently implemented in HyPlag. The approach follows the established design principle of a multi-stage detection process consisting of candidate retrieval, detailed comparison, and human inspection [17]. The following subsections present the analysis steps for each class of content features (math, images, citations, and text).

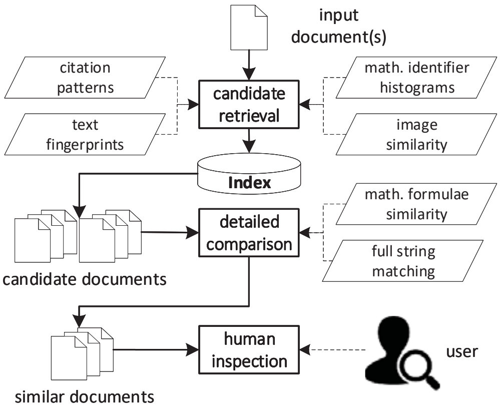


**Figure 1: Overview of the hybrid detection process.**

## 3.1 Math Similarity

To improve the detection of AP, primarily in STEM disciplines, we proposed including mathematical expressions into the similarity analysis of documents [13].

For candidate retrieval, our approach computes signatures ('fingerprints') for mathematical content contained in documents. The signatures represent the histograms of the frequencies of the basic components of mathematical expressions, i.e., identifiers, numbers, and operators. To determine the similarity between histograms, we use a relative distance measure as described in our paper [13].

In the detailed comparison stage, our approach performs a pairwise similarity assessment of formulae using three similarity measures introduced by Zhang and Youssef [21]: The *coverage* measure quantifies the number of matching tokens in two formulae. The *match depth* measure assigns higher weights to matching concepts in two formulae if the concepts occur at higher levels, i.e., closer to the root of the MathML expression tree. The idea is that higher level concepts are more significant for the nature of the expression. The *taxonomic distance* measure assigns a higher weight to elements from the same class in a content dictionary. For instance, two trigonometric functions, such as sin and cos, would receive a higher similarity score than sin and log. HyPlag uses the content dictionary of the MathML standard (see Section 3.5).

## 3.2 Image Similarity

Images in academic documents convey much information independent of the text, which makes them valuable features for assessing the semantic similarity present in such documents. To consider a wide range of image types commonly occurring in academic documents, e.g., charts or plots, schematic representations, and photographs, HyPlag includes both established and novel image similarity assessments. We will present a brief overview of the analysis steps HyPlag performs for candidate retrieval. Details on our image-based PD approach can be found in our paper [12].

We employ *perceptual hashing (pHash)* using a Discrete Cosine Transform (DCT) and comparing pHash values using their hamming distance as a well-established, fast and reliable method to find highly similar images of arbitrary image types.

As an initial approach to analyzing visually differing images, we include two methods that analyze the text, such as labels, extracted from the images using Optical Character Recognition (OCR). The first method performs basic set-based character 3-gram matching for all characters in an image. The second method performs position-aware character matching by using single characters as the center points around which a fixed-size circular proximity region is defined. The similarity measure to compare two images considers the number of position-aware text matches normalized by the number of characters in the longer of the two OCR texts.

As a first approach to identifying potential data reuse, i.e., representing (nearly) identical data in visually different charts, we employ *ratio hashing*. This novel algorithm finds semantically equivalent bar charts by computing a hash value from the relative heights of bars compared to the height of the largest bar. To determine the distance of two ratio hashes, we compare the components of the hash, i.e., the relative bar heights, in decreasing order and calculate the sum of the absolute differences of the bar heights.

## 3.3 Citation Similarity

For candidate retrieval, the hybrid approach employs four citation-based similarity measures, which prior research proved effective for discovering concealed AP [5, 6].

*Bibliographic Coupling (BC)*, quantifies the absolute number or fraction of shared references while ignoring the number, position, and order of citations in the text. We use BC as a basic filter.

*Longest Common Citation Sequence (LCCS)* is the maximum number of citations that match in both documents in the same order, but not necessarily in a contiguous block. We showed that LCCS achieves good results for retrieving longer passages of reused text, in which the sequence of ideas remained unchanged.

*Greedy Citation Tiling (GCT)* identifies all individually longest matching substrings of citations in two documents ('citation tiles'), i.e., all blocks of consecutive shared citations in identical order. Longer citation tiles are a strong indicator for high semantic similarity of text passages, even if the order of the passages was changed.

*Citation Chunking (CC)* is a class of heuristic measures to find variably-sized patterns of matching citations, in which the count and order of matching citations can differ.

## 3.4 Text Similarity

To find similar text, we rely on established text retrieval methods. For candidate retrieval, our approach employs a text fingerprinting method, which we realized by adapting the Sherlock tool[1]. The method performs text chunking using word 3-grams and probabilistically selects a subset of chunks for computing a digital signature of the input text. The mean probability for chunk retention is $\frac{1}{16}$.

For the detailed comparison, we offer users a choice between full string matching and the Encoplot algorithm. We adapted the Boyer-Moore algorithm to match all strings (including repetitions) with 12 or more identical words. Encoplot, developed by Grozea et al. [8], is an efficient character 16-gram comparison that achieves a time-complexity of $O(n)$ by ignoring repeated matches.

[1] http://www.cs.usyd.edu.au/~scilect/sherlock/

### 3.5 Implementation

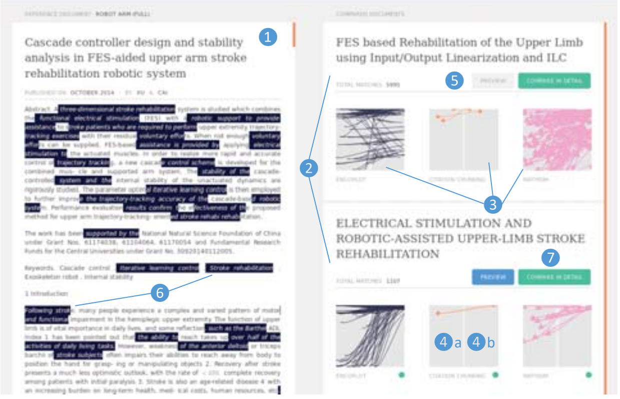
The HyPlag prototype consists of a backend server application and a web-based fronted application, which are loosely coupled via a REST web service interface. The *backend* is realized in Java using the Spring Boot framework. We use an Elasticsearch index as the main data storage for content features. To extract text, header metadata, citations, and references from PDF, we integrated the GROBID[2] and ParsCit[3] parsers. We combine the result sets of both parsers to increase precision and recall of the extraction. HyPlag relies on MathML[4] to represent and process mathematical content. We use InftyReader[5] to convert PDF that include mathematical content to TeX. We then employ LaTeXML[6] to convert the TeX output of InftyReader to XHTML with embedded MathML. The *frontend* (see Section 4) is realized in Ruby on Rails.

Currently, HyPlag's *reference collection* includes 185K biomedical articles from the PubMed Central OA Subset[7] and 105K arXiv.org documents from the dataset of the NTCIR-11 MathIR Task[8].

## 4 DEMONSTRATION

HyPlag's user interface includes two main views to present the results of the hybrid detection approach: the *Results Overview* shown in Figure 2 and the *Detailed Comparison View* shown in Figure 3. HyPlag also features a dashboard area that allows users to upload and manage files as well as to configure, start and track analyses.

We explain the functionality of the analysis views using a retracted journal article from bioengineering [20]. The retraction note[9] explains that the journal retracted the article, because it reused a three-page mathematical analysis without attribution from a paper by Freeman et al. [4]. We used HyPlag to compare the retracted article with the source indicated in the retraction note and with other publications by Freeman, the source paper's first author.



Cascade controller design and stability analysis in FES-aided upper arm stroke rehabilitation robotic system
1
FES based Rehabilitation of the Upper Limb using Input/Output Linearization and ILC
5
2
3
ELECTRICAL STIMULATION AND ROBOTIC-ASSISTED UPPER-LIMB STROKE REHABILITATION
7
6
4a
4b

**Figure 2: Results Overview.**

The *Results Overview* (Figure 2) is the first screen a user sees after selecting to view the results for an analysis. The left part of the screen shows the full text of the input document (see (1) in Figure 2). The right part shows a list of *result summaries* (2) for all documents, for which similarities to the input document have been identified. Each result summary includes one or more *match views* (3). Each match view has two panels and represents the similarities that an analysis method identified, e.g., matching citations or similar formulae. The left panel (4a) represents the input document and the right panel (4b) the comparison document. Matching features appear in the match views connected by lines. The positions of features in the match views reflect their relative positions in the documents. Therefore, similar features in the same order yield parallel lines. Such patterns are a strong indicator for potentially undue content similarity. Features in each match view have a unique color. The user can activate the preview (5) of matches for one comparison document at a time. All features in the input document that match features in the currently active comparison document are highlighted in the full text of the input document using the unique color of the feature (6). The results overview enables users to quickly browse all identified similarities and check which parts of the input document are affected. By clicking a button (7), a user can then switch to the detailed comparison.

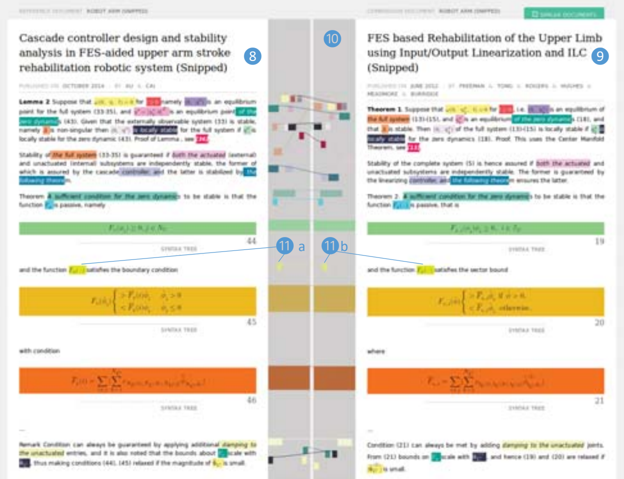
For the example, the match views in Figure 2 show the similarity of text (left), citations (middle) and mathematical content (right) in the retracted article by Xu et al. and two papers by Freeman et al. The upper result summary represents the source paper named in the retraction note. The match views for text indicate moderate similarity of the retracted article, particularly in the introduction, to both comparison documents. This similarity is largely due to overlap in keywords and general scientific phrases and likely would not have caused suspicion for either of the two comparison documents. However, the match view for mathematical content (right) in the upper result summary shows a clearly suspicious similarity that should prompt a user to review the documents in detail.

Figure 3 shows the *Detailed Comparison View*, which displays the full text of the input document (8) and a selected comparison document (9) side-by-side. Between the full texts, a match view (10) similar to the match views in the *Results Overview* highlights all matching features in both documents. However, in this view, each feature match (11a,b) is assigned a separate color. Clicking on any highlight in the full text panels or the central match view aligns the respective feature matches. Since the central match view represents the entire document, the current view port, i.e., the segment of text visible in the adjacent full text panel and the position of the text segment in the document, is indicated using a darker shade.

To improve the legibility of the screen capture, we manually selected a passage with high math similarity that does not exceed the screen. For the example, the combined visualization of similar content features shows that in addition to dispersed keyword matches, especially the mathematical formulae in both documents exhibit a high similarity and occur in nearly identical order. Also, the only source cited in the shown segments (reference 36 on the left and 13 on the right) is identical.

To enable users to review why HyPlag flagged mathematical formulae as similar, clicking on a highlighted formula match opens the interactive visualization of pairwise formula similarity as proposed in our paper [16] and shown in Figure 4. The visualization shows the MathML expression tree for the formula in the input

[2] https://github.com/kermitt2/grobid
[3] https://github.com/knmnyn/ParsCit
[4] https://www.w3.org/Math/
[5] http://www.inftyreader.org/
[6] http://dlmf.nist.gov/LaTeXML/
[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[8] http://ntcir-math.nii.ac.jp/data/
[9] https://doi.org/10.1007/s11071-015-2422-5



REFERENCE DOCUMENT ROBOT ARM (SNIPPED)

Cascade controller design and stability analysis in FES-aided upper arm stroke rehabilitation robotic system (Snipped)

PUBLISHED ON OCTOBER 2014 BY XU & CAI

Lemma 2 Suppose that ... is an equilibrium point for the full system (33-35), and ... is an equilibrium point of the zero dynamics (43). Given that the externally observable system (33) is stable, namely ... is non-singular then ... is locally stable for the full system if ... is locally stable for the zero dynamic (43). Proof of Lemma... see [24].

Stability of the full system (33-35) is guaranteed if both the actuated (external) and unactuated (internal) subsystems are independently stable, the former of which is assured by the cascade controller and the latter is stabilized by the following theorem.

Theorem A sufficient condition for the zero dynamics to be stable is that the function ... is passive, namely

44

and the function ... satisfies the boundary condition

45

with condition

46

Remark Condition can always be guaranteed by applying additional damping to the unactuated entries, and it is also noted that the bounds about ... scale with ..., thus making conditions (44), (45) relaxed if the magnitude of ... is small.

COMPARISON DOCUMENT ROBOT ARM (SNIPPED)

SIMILAR DOCUMENTS

FES based Rehabilitation of the Upper Limb using Input/Output Linearization and ILC (Snipped)

PUBLISHED ON JUNE 2012 BY FREEMAN & TONG & ROGERS & HUGHES & MEADMORE & BURRIDGE

Theorem 1 Suppose that ... for ..., i.e. ... is an equilibrium of the full system (13)-(15), and ... is an equilibrium of the zero dynamics (18), and that ... is stable. Then ... of the full system (13)-(15) is locally stable if ... is locally stable for the zero dynamics (18). Proof: This uses the Center Manifold Theorem, see [11].

Stability of the complete system (5) is hence assured if both the actuated and unactuated subsystems are independently stable. The former is guaranteed by the linearizing controller, and the following theorem ensures the latter.

Theorem 2. A sufficient condition for the zero dynamics to be stable is that the function ... is passive, that is

19

and the function ... satisfies the sector bound

20

where

21

Condition (21) can always be met by adding damping to the unactuated joints. From (21) bounds on ... scale with ..., and hence (19) and (20) are relaxed if ... is small.

**Figure 3: Detailed Comparison View.**

document in light blue shading (here equation 46) and the formula in the comparison document in light green shading (here equation 21). Identical and similar leaf nodes are highlighted and a layout algorithm that minimizes edge crossings aligns the formulae to emphasize structural similarity. To facilitate the structure analysis, the user can collapse and freely arrange nodes. The formulae shown in Figure 4 exhibit a nearly identical structure, although the retracted document partially uses different identifiers, e.g., $N_C$ vs. $I_C$, and a different notation for the parameter, respectively argument $i$ in $\bar{F}$. Since in both equations, the inner sum uses different identifiers, the central leaf nodes for $N_C$ are denoted as similar and not identical.

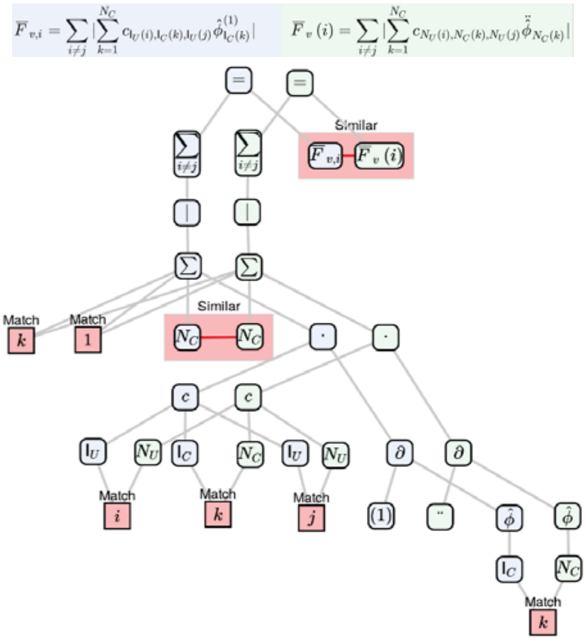

$\bar{F}_{v,i} = \sum_{i \neq j} \left| \sum_{k=1}^{N_C} c_{I_U(i),I_C(k),I_U(j)} \hat{\phi}^{(1)}_{I_C(k)} \right|$
$\bar{F}_v(i) = \sum_{i \neq j} \left| \sum_{k=1}^{N_C} c_{N_U(i),N_C(k),N_U(j)} \ddot{\hat{\phi}}_{N_C(k)} \right|$
Similar
Match
Match
Similar
Match
Match
Match
Match

**Figure 4: Visualization of pairwise formula similarity.**

## 5 CONCLUSION

We presented a prototype that implements a hybrid approach to academic plagiarism detection by analyzing the similarity of mathematical expressions, images, citation patterns, and text. Using a retracted journal article, we demonstrated that the hybrid analysis of these content features can improve the retrieval of potential source documents for cases of concealed content reuse. We also showcased how our interactive visualizations of these content features can aid reviewers in assessing the legitimacy of content similarity in academic manuscripts. Our code is available as open source from:

*http://purl.org/hyplag*

## ACKNOWLEDGMENTS

This work was supported by DFG grant no.: GI 1259/1.

## REFERENCES

[1] Salha Alzahrani, Vasile Palade, Naomie Salim, and Ajith Abraham. 2011. Using Structural Information and Citation Evidence to Detect Significant Plagiarism Cases in Scientific Publications. *JASIST* 63(2) (2011), 286–312.

[2] Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods. In *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, Vol. 42. 133–149.

[3] Teddy Fishman. 2009. "We know it when we see it"? is not good enough: toward a standard definition of plagiarism that transcends theft, fraud, and copyright. In *Proc. Asia Pacific Conf. on Educational Integrity.*

[4] Christopher Freeman, Daisy Tong, Katie Meadmore, Ann-Marie Hughes, Eric Rogers, and Jane Burridge. 2012. FES based Rehabilitation of the Upper Limb using Input/Output Linearization and ILC. In *Proc. Amer. Control Conf.*

[5] Bela Gipp. 2014. *Citation-based Plagiarism Detection - Detecting Disguised and Cross-language Plagiarism using Citation Pattern Analysis.* Springer.

[6] Bela Gipp and Norman Meuschke. 2011. Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence. In *Proc. ACM Symp. on Doc. Eng.*

[7] Bela Gipp, Norman Meuschke, Corinna Breitinger, Jim Pitman, and Andreas Nuernberger. 2014. Web-based Demonstration of Semantic Similarity Detection using Citation Pattern Visualization for a Cross Language Plagiarism Case. In *Proc. Int. Conf. on Enterprise Information Systems.*

[8] Christian Grozea, Christian Gehl, and Marius Popescu. 2009. ENCOPLOT: Pairwise Sequence Matching in Linear Time Applied to Plagiarism Detection. In *Proc. 3$^{rd}$ PAN WS. Uncovering Plagiarism, Authorship and Social Software Misuse.*

[9] Donald L. McCabe. 2005. Cheating among College and University Students: A North American Perspective. *Int.J. for Academic Integrity* 1, 1 (2005), 1–11.

[10] Norman Meuschke and Bela Gipp. 2013. State-of-the-art in detecting academic plagiarism. *Int. J. of Educational Integrity* 9, 1 (2013).

[11] Norman Meuschke and Bela Gipp. 2014. Reducing Computational Effort for Plagiarism Detection by using Citation Characteristics to Limit Retrieval Space. In *Proc. ACM/IEEE Joint Conf. on Digital Libraries (JCDL).*

[12] Norman Meuschke, Christopher Gondek, Daniel Seebacher, Corinna Breitinger, Daniel Keim, and Bela Gipp. 2018. An Adaptive Image-based Plagiarism Detection Approach. In *Proc. ACM/IEEE Joint Conf. on Digital Libraries (JCDL).*

[13] Norman Meuschke, Moritz Schubotz, Felix Hamborg, Tomas Skopal, and Bela Gipp. 2017. Analyzing Mathematical Content to Detect Academic Plagiarism. In *Proc. Conf. on Inform. and Knowl. Manage. (CIKM).*

[14] Norman Meuschke, Nicolas Siebeck, Moritz Schubotz, and Bela Gipp. 2017. Analyzing Semantic Concept Patterns to Detect Academic Plagiarism. In *Proc. Int. WS on Mining Scientific Publ. (WOSP) at JCDL.*

[15] Solange de L. Pertile, Viviane P. Moreira, and Paolo Rosso. 2016. Comparing and combining Content- and Citation-based approaches for plagiarism detection. *JASIST* 67, 10 (2016), 2511–2526.

[16] Moritz Schubotz, Norman Meuschke, Thomas Hepp, Howard S. Cohl, and Bela Gipp. 2017. VMEXT: A Visualization Tool for Mathematical Expression Trees. In *Proc. Int. Conf. on Intelligent Computer Mathematics CICM.*

[17] Benno Stein, Sven Meyer zu Eissen, and Martin Potthast. 2007. Strategies for Retrieving Plagiarized Documents. In *Proc. ACM SIGIR Conf.*

[18] K. Vani and Deepa Gupta. 2016. Study on Extrinsic Text Plagiarism Detection Techniques and Tools. *J. Engin. Sc. & Techn. Review* 9, 5 (2016).

[19] Debora Weber-Wulff. 2014. *False Feathers: A Perspective on Academic Plagiarism.*

[20] Wenkang Xu, Chenxiao Cai, and Yun Zou. 2015. RETRACTED ARTICLE: Cascade controller design and stability analysis in FES-aided upper arm stroke rehabilitation robotic system. *Nonlinear Dynamics* 79, 2 (2015).

[21] Qun Zhang and Abdou Youssef. 2014. An Approach to Math-Similarity Search. In *Intelligent Computer Mathematics.*