TEIMMA: The First Content Reuse Annotator for Text, Images, and Math

Ankit Satpute*
Ankit.Satpute@fiz-karlsruhe.de
FIZ Karlsruhe Leibniz Institute for Information Infrastructure
Berlin, Germany

André Greiner-Petter
Greiner-Petter@gipplab.org
George August University of Göttingen
Göttingen, Germany

Moritz Schubotz
Moritz.Schubotz@fiz-karlsruhe.de
FIZ Karlsruhe Leibniz Institute for Information Infrastructure
Berlin, Germany

Norman Meuschke
meuschke@uni-goettingen.de
George August University of Göttingen
Göttingen, Germany

Akiko Aizawa
aizawa@nii.ac.jp
National Institute of Informatics
Tokyo, Japan

Bela Gipp
gipp@uni-goettingen.de
George August University of Göttingen
Göttingen, Germany

ABSTRACT

This demo paper presents the first tool to annotate the reuse of text, images, and mathematical formulae in a document pair—TEIMMA. Annotating content reuse is particularly useful to develop plagiarism detection algorithms. Real-world content reuse is often obfuscated, which makes it challenging to identify such cases. TEIMMA allows entering the obfuscation type to enable novel classifications for confirmed cases of plagiarism. It enables recording different reuse types for text, images, and mathematical formulae in HTML and supports users by visualizing the content reuse in a document pair using similarity detection methods for text and math.

CCS CONCEPTS

• **Information systems** → **Near-duplicate and plagiarism detection**; • **Applied computing** → *Digital libraries and archives.*

KEYWORDS

Reuse annotator, Offsets recording, Math annotator, Similarity visualization

1 INTRODUCTION

The effectiveness of an algorithm, particularly in Machine Learning, heavily depends on the quality of the dataset used to develop the algorithm. Accurate annotations of reused content in a document pair are crucial for developing systems to detect plagiarism, paraphrases, and summaries [3]. Existing plagiarism detection systems (PDS) can only identify copied and slightly altered text [11]. Developing advanced PDS capable of identifying strongly disguised cases of content reuse requires compiling a gold-standard dataset. To the best of our knowledge, no tool exists for annotating such cases of content reuse, which is required for creating a suitable dataset. A few tools allow annotating by selecting text from a single PDF, but none provide functionality for annotating PDF pairs [17]. This approach encounters issues such as varying encodings and text representation formats, incorrect character offsets, and undetected non-textual elements [16].

Bast et al. have shown the challenges of extracting text from PDFs and the shortcomings of tools supporting this task [4]. In particular, scientific documents in the STEM fields (Science, Technology, Engineering, Mathematics) often contain non-textual elements such as mathematical formulae, which are typically ignored during content annotation. Extracting text is easier than extracting mathematical expressions because the formula as presented in a PDF does not allow capturing the formula's structure or semantics, available in LaTeX or MathML[1] [5]. Annotating math is a highly complex task supported by specialized tools to enrich mathematical formulae, such as MioGatto [2], MathAlign [1], and AnnoMathTeX [22]. These tools allow to save math in its original form, such as LaTeX or MathML, but none support recording annotation on a document pair.

This paper introduces TEIMMA, the first tool that enables the annotation of reused text, images, and math in their original transcribed form. TEIMMA's **source code** is publicly available [21], and we provide a **live demonstration** of TEIMMA's features at https://teimma.gipplab.org.

2 ARCHITECTURE AND USE

TEIMMA (TExt, IMage, MAth) is a web-based tool to visualize and annotate similar content in document pairs using a machine-processable format. We refer to similar formulae, text, or a combination thereof as a case. The tool stores annotated cases of similar text as plain text, similar math as MathML, and similar images as IDs referring to the original images. Documents previously annotated with TEIMMA can be re-uploaded to modify and add annotations.

The tool accepts documents in PDF, LaTeX, and plain text (.txt)[2] format. TEIMMA performs a multi-step conversion of PDFs to HTML and MathML to ensure the accurate representation of text, images, and math. First, it uses the open-source Python package *pdftolatex* [12] to extract the positions of text, math, and images from the PDF. The package converts text to LaTeX and math to images. Second, TEIMMA employs the LaTeX OCR model *pix2tex* [7] to convert the images of math formulae returned by *pdftolatex* to LaTeX. The model uses a pre-trained Vision Transformer encoder [10] with a ResNet backbone and a Transformer decoder trained on the im2latex-100k dataset [9, 25]. In the third step, TEIMMA combines the extracted text, images, and math to create a complete LaTeX

*Also with George August University of Göttingen.

[1] https://www.w3.org/Math/

[2] Files in .txt format do not support image or math annotations

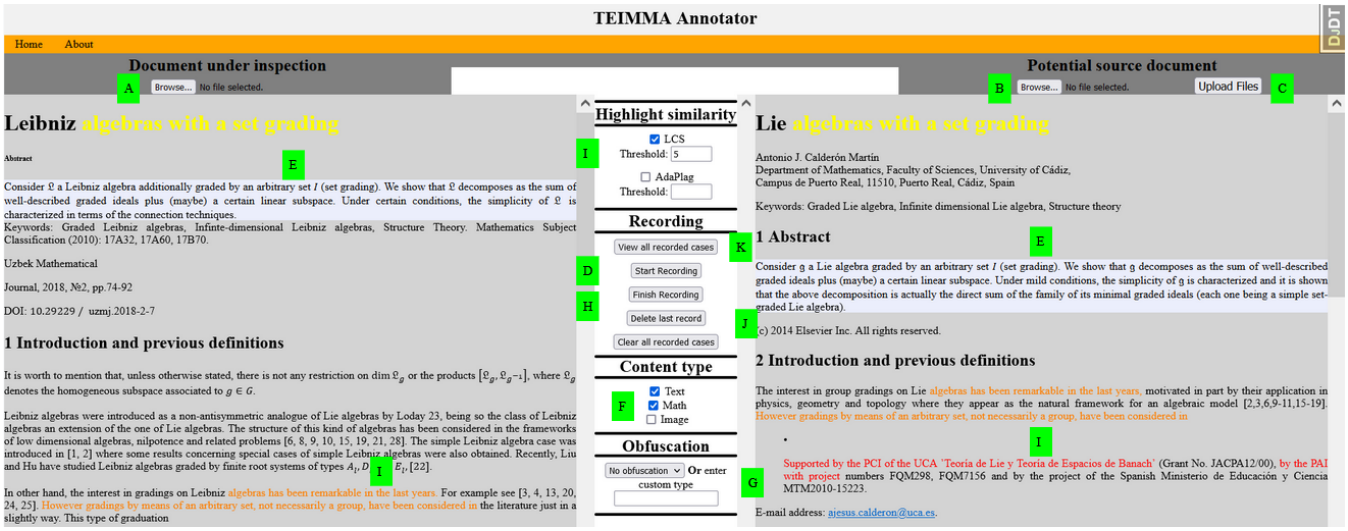

Figure 1: TEIMMA User interface. The left document [19] has been retracted for plagiarizing the right document [8].

representation of the PDF. If possible, we recommend using input documents in LaTeX because PDF to LaTeX, like any other conversion, entails the risk of errors. Thus far, no comprehensive evaluation of the conversion accuracy for math extraction from PDF exists [14, 24, 26]. Lastly, TEIMMA converts the LaTeX output to HTML and MathML for mathematical content using LaTeXML [18]—the best-performing tool for this task [23].

TEIMMA uses HTML tag names to extract text, images, and math [6] and records annotations in terms of the character positions of selected content in a plain text file. The tool replaces each math formula in MathML format with its assigned ID while maintaining its start character position in the plain text file. This allows separating formulae from the plain text to prevent the typically extensive MathML markup of formulae from distorting the character positions.

Figure 1 shows TEIMMA's user interface for visualizing and annotating similar content. The buttons A, B, and C allow uploading the two documents for investigation. TEIMMA converts both documents to HTML and saves the extracted plain text, math formulae, and images in the database. After clicking the *Start Recording* button D, users select a span in both documents. TEIMMA extracts the text from the span and matches it to the text in the plain text file to obtain the span's start and end character positions. The selected span is highlighted by assigning it a unique background color E. The checkboxes in the *Content type* section F allow configuring the type of annotations to be performed, e.g., only annotations of similar text. The section *Obfuscation* G allows users to enter the obfuscation type, e.g., paraphrase or summary, they think has been used to obfuscate the content. Users can activate one of four algorithms[3] I to receive support with annotating by viewing similar text and math content in the uploaded document pair. To view similar text, users can choose between the longest common substring (LCS) or AdaPlag, the winning method in the latest PAN plagiarism competition [20]. For similar math tokens, the longest common identifier sequence (LCIS) or greedy identifier tiling (GIT) [13, 15]. Moreover, users can specify the minimum length required for displaying the matches that each algorithm identified. For the text-based algorithms LCS and AdaPlag, the length threshold represents the number of words, and for the math-based algorithms LCIS and GIT, the number of math symbols in the match. The *Finish Recording* H button saves the recorded span in the database along with the data entered by users for describing the identified similarity. The *Delete the last record* J button deletes the previously recorded annotation. Saved annotations in the database can be viewed and downloaded as a JSON Lines (.jsonl) file by clicking on *View all recorded cases* K button. Users must create annotations for each document pair separately if a document shares content with multiple other documents. However, TEIMMA keeps track of overlapping annotations by checking if previous annotations for re-uploaded documents exist in the database.

The final annotation stored in the database in JSON format contains document names, the character offsets for the start and end of text spans, and the IDs for images and formulae. The original images and formulae in MathML are also stored in the database. Additionally, the annotation contains the content and obfuscation type if users entered them.

[3]Note: Only two algorithms are visible in Figure 1 due to space limitations.

ACKNOWLEDGMENT

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 437179652, the German Academic Exchange Service (DAAD) - 57515245, and the Lower Saxony Ministry of Science and Culture and the VW Foundation.

REFERENCES

[1] Maria Alexeeva, Rebecca Sharp, Marco A Valenzuela-Escárcega, Jennifer Kadowaki, Adarsh Pyarelal, and Clayton Morrison. 2020. MathAlign: Linking formula identifiers to their contextual natural language descriptions. In *Proceedings of the 12th language resources and evaluation conference*. 2204–2212.

[2] Takuto Asakura, André Greiner-Petter, Akiko Aizawa, and Yusuke Miyao. 2020. Towards Grounding of Formulae. In *Proceedings of the First Workshop on Scholarly Document Processing*. 138–147.

[3] Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. Text reuse detection using a composition of text similarity measures. In *Proceedings of COLING 2012*. 167–184.

[4] Hannah Bast and Claudius Korzen. 2017. A benchmark and evaluation for text extraction from PDF. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 1–10.

[5] Marco Beck, Isabel Beckenbach, Thomas Hartmann, Moritz Schubotz, and Olaf Teschke. 2020. Transforming scanned zbMATH volumes to LaTeX: planning the next level digitisation. *European Mathematical Society Magazine* 117 (2020), 49–52.

[6] Marco Beck, Moritz Schubotz, Vincent Stange, Norman Meuschke, and Bela Gipp. 2021. Recognize, Annotate and Visualize Parallel Structures in XML Documents. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, Illinois, USA, 258–261. https://doi.org/10.1109/JCDL52503.2021.00078

[7] Lukas Blecher. 2022. LaTeX-OCR: pix2tex: Using a ViT to convert images of equations into LaTeX code. https://github.com/lukas-blecher/LaTeX-OCR SWHID: swh:1:dir:6affa30af9a3e35dfc8a9e4175647e2f95e9033c. [Software: Accessed 21-Jan-2023].

[8] Antonio J. Calderón Martín. 2014. Lie algebras with a set grading. *Linear Algebra Appl.* 452 (2014), 7–20. https://doi.org/10.1016/j.laa.2014.03.031

[9] Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. 2016. Image-to-Markup Generation with Coarse-to-Fine Attention. https://doi.org/10.48550/ARXIV.1609.04938

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://doi.org/10.48550/ARXIV.2010.11929

[11] Tomas Foltynek, Norman Meuschke, and Bela Gipp. 2019. Academic Plagiarism Detection: A Systematic Literature Review. *Comput. Surveys* 52, 6 (Oct. 2019), 112:1–112:42. https://doi.org/10.1145/3345317

[12] Vinay Kanigicherla. 2021. pdftolatex: Python tool for generation of latex code from PDF files. https://github.com/vinaykanigicherla/pdftolatex SWHID: swh:1:dir:713a4905fcc2c65d5618a226b2d67019451e7dda. [Software: Accessed 21-Jan-2023].

[13] Norman Meuschke. 2021. *Analyzing Non-Textual Content Elements to Detect Academic Plagiarism*. Doctoral Thesis. University of Konstanz, Dept. of Computer and Information Science, Konstanz, Germany. https://doi.org/10.5281/zenodo.4913345

[14] Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. 2023. A Benchmark of PDF Information Extraction Tools Using a Multi-task and Multi-domain Evaluation Framework for Academic Documents. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*. LNCS, Vol. 13972. Springer Nature Switzerland, Cham, 383–405. https://doi.org/10.1007/978-3-031-28032-0_31

[15] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 2019. Improving Academic Plagiarism Detection for STEM Documents by Analyzing Mathematical Content and Citations. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (Urbana-Champaign, IL, USA). https://doi.org/10.1109/JCDL.2019.00026

[16] Rishabh Mittal and Anchal Garg. 2020. Text extraction using OCR: a systematic review. In *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. IEEE, 357–362.

[17] Mark Neumann, Zejiang Shen, and Sam Skjonsberg. 2021. PAWLS: PDF annotation with labels and structure. *arXiv preprint arXiv:2101.10281* (2021).

[18] NIST. 2007. LATExml: A LATEX to XML/HTML/MathML Converter — math.nist.gov. https://math.nist.gov/~BMiller/LaTeXML/. [Accessed 21-Jan-2023].

[19] José M. Sánchez. 2018. Leibniz algebras with a set grading. RETRACTED. *Uzb. Math. J.* 2018, 2 (2018), 74–92. https://doi.org/10.29229/uzmj.2018-2-7

[20] Miguel A Sanchez-Perez, Alexander Gelbukh, and Grigori Sidorov. 2015. Adaptive algorithm for plagiarism detection: The best-performing approach at PAN 2014 text alignment competition. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 402–413.

[21] Ankit Satpute, André Greiner-Petter, Moritz Schubotz, Norman Meuschke, Akiko Aizawa, and Bela Gipp. 2023. TEIMMA: The First Content Reuse Annotator for Text, Images, and Math. https://github.com/gipplab/TEIMMA-Reuse-Annotator SWHID: swh:1:dir:a3b95e4ce8893030696393525c1d5a71d27aa303. [Software: Accessed 21-Jan-2023].

[22] Philipp Scharpf, Ian Mackerracher, Moritz Schubotz, Joeran Beel, Corinna Breitinger, and Bela Gipp. 2019. AnnoMathTeX - a Formula Identifier Annotation Recommender System for STEM Documents. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys 2019)*. ACM, Copenhagen, Denmark. https://doi.org/10.1145/3298689.3347042

[23] Moritz Schubotz, Andre Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard Cohl, and Bela Gipp. 2018. Improving the Representation and Conversion of Mathematical Formulae by Considering their Textual Context. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*. Fort Worth, USA. https://doi.org/10.1145/3197026.3197058

[24] Tanuj Sur, Aaditree Jaisswal, and Venkatesh Vinayakarao. 2023. Mathematical Expressions in Software Engineering Artifacts. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*. 238–242.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/ARXIV.1706.03762

[26] Zelun Wang and Jyh-Charn Liu. 2020. PDF2LaTeX: A Deep Learning System to Convert Mathematical Documents from PDF to LaTeX. In *Proceedings of the ACM Symposium on Document Engineering 2020*. 1–10.