Analyzing Mathematical Content for Plagiarism and Recommendations

Ankit Satpute[0000−0003−3219−026X]

Georg August University of Göttingen, Göttingen, Germany,
FIZ Karlsruhe Leibniz Institute for Information Infrastructure, Berlin, Germany
Ankit.Satpute@fiz-karlsruhe.de

1 Motivation

Defined as "the use of ideas, concepts, words, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected" [6], plagiarism poses a severe concern in the rapidly increasing number of scientific publications. The Vroniplag has documented plagiarism in 212 dissertations [19], and zbMATH Open pointed plagiarised research papers in mathematics [17]. The easily recognizable copy-paste type plagiarism [12] will likely diminish due to the accessibility of AI-powered models like ChatGPT. Plagiarism among researchers is more concealed, which is challenging for existing Plagiarism Detection Systems (PDS) [1]. Scientific research often builds upon the foundations laid by existing literature, discovering similar works as an integral part of research. Recommender systems (RS) assist users in coping with many scientific documents by showing similar ones the user might be interested in. RS has become a crucial filtering and discovery tool that many users of digital libraries rely on.

Even though both PDS and RS have different final objectives, they share the goal of finding similar content. However, existing PDS and RS focus primarily on textual content and do not utilize non-textual elements, specifically mathematical content, to their full potential [7,16]. The significance of mathematical content is much higher and valued in scientific documents from STEM (Science, Technology, Engineering, and Mathematics). Despite this, efforts to utilize mathematical content for document similarity remain in infancy. This thesis addresses this gap by analyzing and utilizing mathematical content for content similarity.

2 Backrgound and Related Work

There are two main reasons for the lack of methods utilizing mathematical content to find similar scientific documents. *Reason 1* is the unavailability of large-scale, annotated datasets of similar mathematical content in a machine-processable format like LaTeX or MathML. PAN Datasets [18] are a frequently used resource to develop and evaluate PDS [15,2,20,5]. Gienapp et al. [8] presented the Webis-STEREO-21 dataset containing reused text passages (without math) from scientific publications. For plagiarised mathematical content, resources as comprehensive as those for text reuse are missing. There are Math

Information Retrieval (MIR) datasets such as NTCIR [21] and ARQMath [11], but they have very limited similar math content pairs. No existing RS datasets consider mathematical contents.

Reason 2 is the sole analysis of the presentational similarity of mathematical content, such as matching math symbol occurrences [13,14]. Considering presentational mathematical similarity at all is a valuable starting point, but there is a need to develop advanced methods analyzing the semantics of mathematical formulae [17]. Moreover, current approaches analyze text and mathematical formulae separately. However, identifying semantic similarity of mathematical content requires a combined analysis because mathematical formulae are mostly context-dependent [9]. Only two works analyzed mathematical content in scientific documents to identify plagiarism [14,13]. Both studied primary mathematical symbol occurrences and used a small evaluation dataset of 10 document pairs. No PDS thus far considers semantic textual and non-textual content similarity [10,7]. Search engines, such as Searchonmath [4], Approach0 [24], etc., are tailored towards mathematical formulae. Even though they allow textual and mathematical terms for searching, the content is not considered semantically. Language models are considered for finding math similarities [3,23,22]. Most of these language models are not trained on mathematical content, thus questioning the extensibility of these approaches to consider math semantic similarity.

3 Proposed Research

The objective of this doctoral thesis is to:

Conceive, devise, and evaluate robust approaches for math content similarity capable of identifying obfuscated plagiarism and generating relevant recommendations for scientific documents.

To achieve this objective, we will perform the following research tasks:

T1: Investigate the strengths and limitations of state-of-the-art mathematical content similarity detection approaches.
T2: Formulate features of mathematical contents, develop and evaluate detection approaches for locating semantically and syntactically similar math.
T3: Devise a similarity assessment that combines text, math, and citations to detect similar scientific documents.
T4: Implement a PDS and an RS with the best-performing developed approach to demonstrate its applicability in a real-world document collection.
T5: Evaluate the proposed approach by assessing the implemented PDS and RS's effectiveness, computational efficiency, and usability.

Acknowledgements

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 437179652 and the Deutscher Akademischer Austauschdienst (DAAD, German Academic Exchange Service - 57515245).

References

1. Alzahrani, S.M., Salim, N., Abraham, A.: Understanding plagiarism linguistic patterns, textual features, and detection methods. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **42**(2), 133–149 (2011)
2. Arabi, H., Akbari, M.: Improving plagiarism detection in text document using hybrid weighted similarity. Expert Systems with Applications **207**, 118034 (2022)
3. Dadure, P., Pakray, P., Bandyopadhyay, S.: Bert-based embedding model for formula retrieval. In: CLEF (Working Notes). pp. 36–46 (2021)
4. Diaz, Y., Nishizawa, G., Mansouri, B., Davila, K., Zanibbi, R.: The mathdeck formula editor: Interactive formula entry combining latex, structure editing, and search. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–5 (2021)
5. El-Rashidy, M.A., Mohamed, R.G., El-Fishawy, N.A., Shouman, M.A.: Reliable plagiarism detection system based on deep learning approaches. Neural Computing and Applications **34**(21), 18837–18858 (2022)
6. Fishman, T.: "we know it when we see it" is not good enough: Toward a standard definition of plagiarism that transcends theft, fraud, and copyright (2009)
7. Foltynek, T., Meuschke, N., Gipp, B.: Academic Plagiarism Detection: A Systematic Literature Review. ACM Computing Surveys **52**(6), 112:1–112:42 (Oct 2019). https://doi.org/10.1145/3345317
8. Gienapp, L., Kircheis, W., Sievers, B., Stein, B., Potthast, M.: A large dataset of scientific text reuse in Open-Access publications. Scientific Data **10**(1), 58 (Jan 2023). https://doi.org/10.1038/s41597-022-01908-z
9. Greiner-Petter, A., Schubotz, M., Breitinger, C., Scharpf, P., Aizawa, A., Gipp, B.: Do the math: Making mathematics in wikipedia computable. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–12 (2022). https://doi.org/10.1109/TPAMI.2022.3195261
10. Lovepreet, V.G., Kumar, R.: Survey on plagiarism detection systems and their comparison. In: Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018. vol. 990, p. 27. Springer (2019)
11. Mansouri, B., Agarwal, A., Oard, D.W., Zanibbi, R.: Advancing math-aware search: The arqmath-3 lab at clef 2022. In: European Conference on Information Retrieval. pp. 408–415. Springer (2022)
12. McCabe, D.L.: Cheating among college and university students: A north american perspective. International Journal for Educational Integrity **1**(1) (2005)
13. Meuschke, N., Schubotz, M., Hamborg, F., Skopal, T., Gipp, B.: Analyzing mathematical content to detect academic plagiarism. In: Proceedings of the International Conference on Information and Knowledge Management (CIKM) (2017). https://doi.org/10.1145/3132847.3133144
14. Meuschke, N., Stange, V., Schubotz, M., Kramer, M., Gipp, B.: 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) (Jun 2019). https://doi.org/10.1109/JCDL.2019.00026
15. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Pan plagiarism corpus 2011 (pan-pc-11) (Jun 2011). https://doi.org/10.5281/zenodo.3250095, https://doi.org/10.5281/zenodo.3250095
16. Scharpf, P., Mackerracher, I., Schubotz, M., Beel, J., Breitinger, C., Gipp, B.: 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 (Sept 2019). https://doi.org/10.1145/3298689.3347042
17. Schubotz, M., Teschke, O., Stange, V., Meuschke, N., Gipp, B.: Forms of plagiarism in digital mathematical libraries. In: Intelligent Computer Mathematics - 12th International Conference, CICM 2019, Prague, Czech Republic, July 8-12, 2019, Proceedings (2019). https://doi.org/10.1007/978-3-030-23250-4_18
18. Stein, B., Koppel, M., Stamatatos, E.: Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection PAN'07. ACM SIGIR Forum **41**(2), 68–71 (Dec 2007). https://doi.org/10.1145/1328964.1328976
19. Weber-Wulff, D.: Talking to a wall: The response of german universities to documentations of plagiarism in doctoral theses. In: Academic Integrity: Broadening Practices, Technologies, and the Role of Students: Proceedings from the European Conference on Academic Integrity and Plagiarism 2021. pp. 363–371. Springer (2023)
20. Yu, W., Pang, L., Xu, J., Su, B., Dong, Z., Wen, J.R.: Optimal partial transport based sentence selection for long-form document matching. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 2363–2373 (2022)
21. Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., Davila, K.: Ntcir-12 mathir task overview. In: NTCIR (2016)
22. Zhong, W., Xie, Y., Lin, J.: Applying structural and dense semantic matching for the arqmath lab 2022, clef. Proceedings of the Working Notes of CLEF 2022 pp. 5–8 (2022)
23. Zhong, W., Yang, J.H., Lin, J.: Evaluating token-level and passage-level dense retrieval models for math information retrieval. arXiv preprint arXiv:2203.11163 (2022)
24. Zhong, W., Zhang, X., Xin, J., Zanibbi, R., Lin, J.: Approach zero and anserini at the clef-2021 arqmath track: Applying substructure search and bm25 on operator tree path tokens. Proc. CLEF 2021 (CEUR Working Notes) (2021)