IFME-Intelligent Filter for the Mathematical Expression

Andri Rai and Deepti Malhotra

Abstract

Mathematical expression extraction is one of the most important challenges for decades, and hence, there is an extreme need to counter the issue of mathematical expression and concept retrieval from scientific documents. While there have been many attempts for mathematical expression (ME) retrieval by using diverse approaches like Symbol Layout Tree (SLT), DenseNet, convolution neural network (CNN), support vector machine (SVM) and many more. As a result, they lead to new implication and restrictions in precise ME similarity retrieval and its specific mathematical semantic. In order to analyze the mathematical document, the automatic detection and retrieval of similar recognized ME is a key task. The research paper presents the existing mathematical plagiarism detection techniques and mathematical expression extraction techniques proposed by different researchers. The prime objective of this research work is to propose an intelligent tool to filter the standard mathematical expression and notation from the scientific document.

Keywords

Mathematical expression extraction • Information retrieval (IR) • Plagiarism detection (PD) • OCR • Mathematical expression (ME) • Machine learning (ML) • Syntax similarity • Semantic similarity

1 Introduction

Mathematical expression extraction from scientific documents is an onerous area of research in academic improvement. Mathematical expression extraction is a complex yet essential task for academic plagiarism detection and information retrieval. Various scholars have attempted to extract the mathematical notations and expressions from documents, but precision and recall of these are relatively low at par with simple text retrieval. The compelling and completion for detecting the mathematical plagiarism and retrieval of the source document depend on the detection of ME. With the advancement in the digitalization of documents, it is becoming more and more difficult to detect the ME from documents. Although many techniques for the OCR based detection of ME give better performance for simple text documents, retrieving the ME from source with exact name is not accurate and effective. Mainly, there are two types of ME detection that is inline and embedded detection process that is implemented Zanibbi and Blostein (Zanibbi and Blostein 2012a).

The recent research for the ME detection is based on online and offline handwritten MEs, which still lack fully solving the problem. OCR-based ME detection usually has difficulties for recognizing the larger no of character and different types of symbols from the image documents. Traditional methods for ME detection focused on the displayed and inline detection of MEs by using the rule-based methods for detection by Lee and Wang (Lee and Wang 1997), and they employ the n-gram model for recognizing the ME from a large corpus. However, many different methods were given by Phong et al. (2017, 2019) for classifying the inline and displayed ME detection like based on SVM. There are also DNN-based methods for mathematical ME detection OCR that recognize the symbols from PDFs, handwritten documents, printed documents by using much deep learning-based ME detection methods proposed by Gao et al. (Gao et al. 2017) and Chan (Chan and Yeung Aug. 2000).

A. Rai (✉) · D. Malhotra
Department of Computer Science and IT, Central University of Jammu, Rahya Suchani, Samba District, Bagla, Jammu and Kashmir, India
e-mail: andri.cujammu@gmail.com

D. Malhotra
e-mail: deepti.csit@cujammu.ac.in

P. K. Singh et al. (eds.), *Innovations in Information and Communication Technologies (IICT-2020)*,
Advances in Science, Technology & Innovation,
https://doi.org/10.1007/978-3-030-66218-9_11

Further, the mathematical expression extraction's formulas and symbols detection is an important subset of the academic plagiarism detection, which cannot be ignored, although it is a relatively small part of the plagiarism detection. It is accountable in mathematical plagiarism detection and in the lack of math information retrieval. The novel method for a possible feature selection and feature comparison strategies for developing the mathematical-based plagiarism detection approaches are designed by Norman Meuschke et.al. (Meuschke et al. 2017), and the result shows that the mathematical expressions are promising text-independent features to identify academic plagiarism. Later, they also presented a prototype that implements a hybrid approach to academic plagiarism detection by analyzing the similarity of mathematical expressions, images, citation patterns, and text, and shows a result visualization approach by using HyPlag to analyze the confirmed cases of content reuse. Norman Meuschuke et.al. (Meuschke et al. 2018) analyzed the concept of mathematical content similarity in different types of STEM documents and its implication in academic plagiarism detection. In their research paper, they presented a two-stage detection that combines the similarity assessments of mathematical content, academic content, and text. They also compared the effectiveness of math-based, citation-based, and text-based approaches using confirmed cases of academic plagiarism.

The rest of the paper is organized as follows. Section 2 presents the extent of work done in the research area. The proposed IFME framework has been illustrated and discussed in Sect. 3. The performance metrics that can be useful for our model results in future is discussed in the Sect. 5 and Finally, Sect. 6 finishes the research proposal by concluding and with some helpful future disclosures.

2 Background and Related Work

Mathematical Plagiarism Detection Techniques

Table 1 outlined the different methods of mathematical plagiarism detection techniques proposed by various researchers.

Mathematical Expression Extraction Techniques

Table 2 summarizes the various techniques of mathematical expression proposed by many researchers.

3 IFME-Intelligent Filter for Mathematical Expression

For the detection of standard mathematical expression and notation from the scientific document, the IFME framework is proposed which is presented in Fig. 1.

The description of the various components used in the IFME framework is discussed as follows:

Math Documents In this phase, the different mathematical documents are collected.

ME Extraction by Neural Network In this component, the mathematical expressions have been extracted from the mathematical document collected by the first component: using CNN and U-net framework for the extraction of in-line and embedding mathematical expressions.

Segmentation of ME Features The extracted ME features are then segmented in different sub-blocks for both the inline and embedded ME features.

Compute Cosine Similarity of ME The extracted features are created as vector for the computing the cosine similarity of MEs to identify the similarity between each detected MEs. It will use in improving the mathematical plagiarism detection.

ML Classification of ME After computing the similarity between the features of mathematical expressions, classification has been done by using the random forest algorithm to classify that whether the detected ME is a standard notations or it is identified as a new idea for detecting plagiarism. If it is new identified idea, then it will be manually validated.

Standard ME Database InftyProject databases called InftyCDB-1, InftyCDB-2 and the Marmot dataset, that contains characters, symbols and spatial features of mathematical documents, have been used as the standard mathematical expression databases.

Intelligent Filter for Mathematical Expression (IFME) Algorithm

This is the given pseudo-code of the algorithm for extraction of mathematical expression and detecting the plagiarism:

Input: *Mathematical documents*
Output: Standard Mathematical Notations
Step1: Take the mathematical document.
Step2: Extract the different mathematical expressions from the document.
Step3: Store the extracted Mathematical expressions in Vector A.
$[A] = A_{ME1} + A_{ME2} + A_{ME3} + \ldots A_{MEn}$

$$[A] = \sum_{i=1}^{n} A_{ME}$$

Step4: Take the ME dataset and store the different mathematical expressions in vector B

$[B] = B_{ME1} + B_{ME2} + B_{ME3} + \ldots B_{MEn}$

Step5: Calculate the Cosine Similarity between two Vectors A and B.

Step6::

$$Similarity = \cos\theta = \frac{A.B}{\|A\|\|B\|} \frac{\sum_{i=1}^{n} A_i \times B_i}{\sum_{i=1}^{n} (A_i)^2 \times \sum_{i=1}^{n} (B_i)^2}$$

Where, A_{MEi} and B_{MEi} are the components of vector A and B.

Step7:/*Classify the features of detected MEs*/

If " STANDARD NOTATION"

" NOT PLAGIARISM

Else

"MANUAL VALIDATION"

Step 8: END

4 Result and Discussion

To evaluate the proposed algorithms, simulation test bench for the IFME framework has been created with a Lenovo idea pad laptop, hardware configuration of 8 GB RAM, 2 TB Hard disk. The input for the proposed framework is mathematical document images that are collected from 400 different documents.

Mathematical Dataset

The training and testing of the model is done using the InftyMCCDB-2 which is the updated version of InftyCDB-2 dataset. It contains more than 30,000 expressions that are further grouped into 12,551 images for training and 6830 images for testing in the dataset (Fig. 2).

Evaluation of Framework

Recall $R_{ME,}$ Precision P_{ME} and F-measures F_{ME} have been used as the performance matrices to validate the identified mathematical expression. For the F_{ME} it is the average weighted score of recall R_{ME} and precision P_{ME} which measures how good is the designed framework works.

Table 1 Analysis of mathematical plagiarism detection techniques

S. No.	Author and year	Problem handled	Input	Analysis	Dataset
1	Phong et al. (2019)	Similarity detection of mathematical, text content	102 k STEM documents	Detection of similarity assessments for mathematical, academic and text content	PDF of AP and NTCIR-11 MathIR task
2	Isele et al. (2018)	Mathematical information retrieval for plagiarism detection	MIR documents	Analyzed the various math plagiarism approaches	MIR documents corpus
3	Meuschke et al. (2018)	Similarity of mathematical expressions	Plagiarized document	Analyzed the similarity of mathematical expressions, images, citation patterns, and text	NTCIR-11 MathIR task dataset
4	Iwatsuki et al. (2017)	Detected in-line mathematical expressions	Mathematical PDF document	A conditional random field (CFR) applied for math identification in layout and linguistic features	Manual corpus
5	Meuschke et al. (2017)	Mathematical-based plagiarism detection	Mathematical documents	Feature selection and feature comparison of math document	NTCIR-11 MathIR task dataset

Table 2 Analysis of mathematical expression extraction techniques

S. No.	Author and year	Problem handled	Input	Analysis	Dataset
1	Ohyama et al. (2019)	Expression detection from mathematical images	Document images	ME detection using U-NET framework	GTDB-1 and GTDB-2
2	Phong et al. (2019)	Variable detection using CNN and SVM	Math document image	CNN and SVM for analyzing image	Marmot dataset with 400 scientific document
3	Guangcun et al. (2019)	Recognize the 2-D structure existing in mathematical expressions	Image of expressions, symbols	A multi-scale CNN called DenseNet used for ME identification	CROHME 2014
4	Pathak et al. (2019)	LSTM-based math retrieval	LATEX document	LSTM neural network for detecting similarity in query search formula	NTCIR-12 MathIR task
5	Mahdavi et al. (2019)	Math formula recognition from document	Math image documents	ME extraction by weighted LOS graph using Edmond's algorithm	Infty MCCDB-2 and CROHME
6	Zhang et al. (2017)	Handwritten math expression recognition	Online Handwritten math document	GRU-RNN used for encoding and decoding of document	CROHME 2014 and CROHME 2013
7	Phong et al. (2017)	Displayed mathematical expression detection	Math document	ME detection using SVM and FFT	Harvard Mathematical Textbooks Dataset and InftyCDB-2
8	Gao et al. (2017)	Formula detection using deep learning	PDF Document	CNN and RNN for formula detection form features	Marmot and Created dataset from CiteSeer
9	Guidi and Coen (2016)	Study of math retrieval	MR retrieval documents	Identified several methods, dataset, systems designed for MR retrieval	ArXiv, DLMF, PlanetMath
10	Stathopoulos and Teufel (2016b)	Math information retrieval	Math Document	C-value algorithm for automatic extraction and detection of MR	18,730 math articles and MREC as a dictionary of 10,601 types
11	Asebriy et al. (2016)	Retrieval of expression	Math document	Encoding MathML and KNN for search algorithm	6925 ME using symbols from five languages
12	Zanibbi. et al. (2016)	Finding similarity for math formula	ME documents	Designed a tangent search engine for query search by expression that can be matched with the corpus	NTCIR-11 Wikipedia
13	Zanibbi et al. (2015b)	Math extraction from PDF	PDF Document	Glyph bounding box and syntax tree for converting and extraction of ME	MikTex and LATEX
14	Yu et al. (2014b)	Equation retrieval	MathML document	Feature retrieval form equations	2000 Math Equations
15	Kim et al. (2012b)	Retrieval of mathematical expression	Math document	Analyzed different ME recognition methods	ME research papers collection
16	Zanibbi and Blostein (2012c)	Formula identification from PDF	PDF document images	A combined rule-based (SVM) and learning-based methods for analyzing the features	Math textbooks in English and Chinese

Recall R_{ME} It is the ratio of correctly predicted positive mathematical expression from the actual class of mathematical expression, defined as:

$$R_{ME} = \frac{TP}{TP + FN} \tag{1}$$

Precision P_{ME} It is the ratio correctly predicted positive mathematical expression to the total predicted mathematical expression, defined as:

$$P_{ME} = \frac{TP}{TP + FP} \tag{2}$$

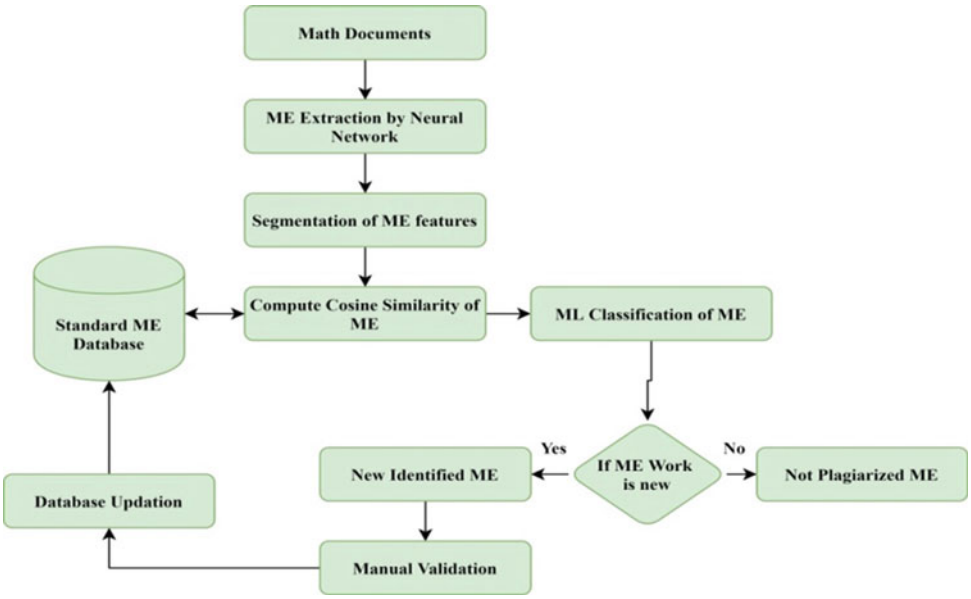

Fig. 1 IFME framework

F-measures F_{ME} The F-measure is the weighted average of recall and precision that is measured for predicted mathematical expression, because it takes both false negative and false positive values of predicted mathematical expression, it defined as:

$$F_{ME} = \frac{2 \times P_{ME} R_{ME}}{R_{ME} + P_{ME}} \tag{3}$$

where the *TP* stands for True Positive; it is for the number of truly predicted values, *FN* stands for False Negative that

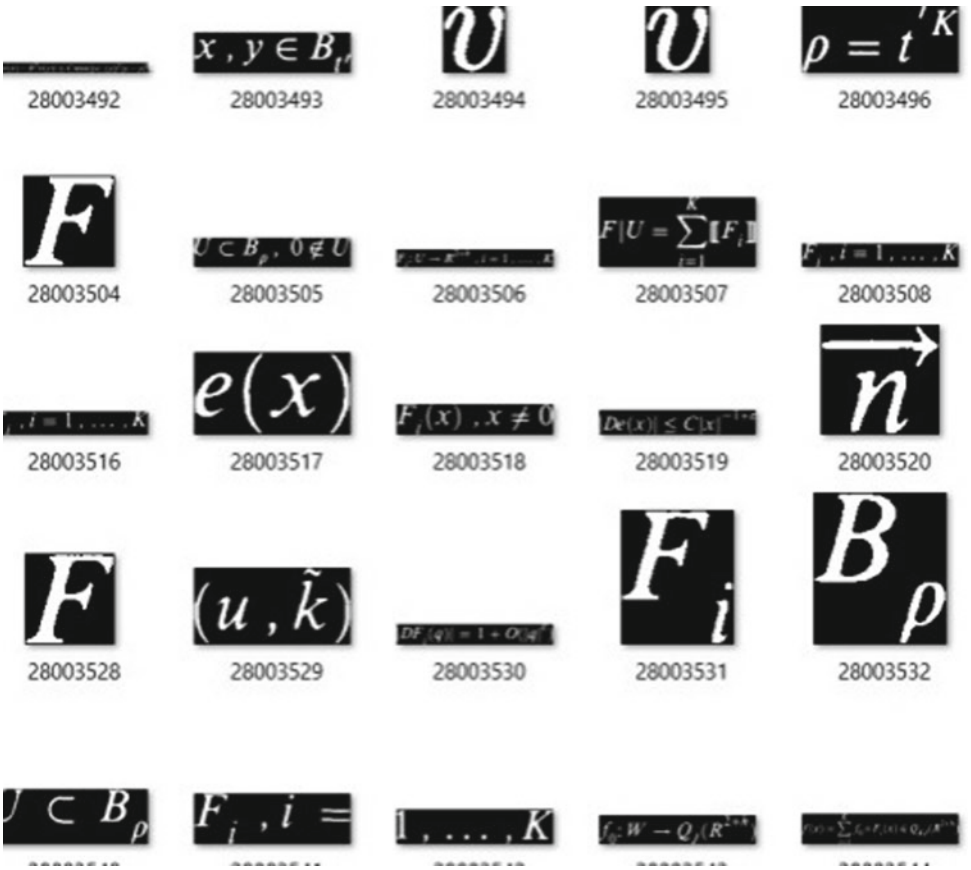

Fig. 2 InftyMCCDB-2 dataset

is the number of yes values predicted as false and *FP* represents the False Positive; it is the number of no values predicted as true.

The evaluation of classified class can be measured on by finding the accuracy (A_{ME}) of the model and A_{ME} is defined as

$$A_{ME} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

In this accuracy (A_{ME}) formula *TP*, *FN*, *FP* stand same as in the recall (R_{ME}), precision (P_{ME}) and F-measures (F_{ME}) and *TN* stands for the (True Negative); these are the number of values which original class is yes but predicted as the no value class. Accuracy (A_{ME}) shows the performance of framework on combining all the parameters taken in the system. Figure 3 shows achieved performance of each work carried out by the researchers based on some performance metrics:- recall, precision, F-measures and accuracy. By this we can conclude that some researcher achieved the best performance for the ME extraction that can be useful for using it to filtering out the ME for detecting the plagiarism in mathematical documents.

5 Conclusion and Future Work

This research paper presents the study of existing mathematical plagiarism detection techniques and mathematical expression extraction techniques proposed by different researchers. The proposed framework uses a convolution neural network and U-net framework for the extraction of in-line and embedding mathematical expressions. Cosine similarity algorithm has been used to find the similarity between the features of the mathematical expressions. After computing the similarity between the features of mathematical expressions, classification has been done by using the random forest algorithm to classify that whether the detected ME is a standard notations or it is identified as a new idea for detecting plagiarism. If it is newly identified idea, then it will be manually validated techniques. It has been analyzed that the convolution neural network and the U-net framework produce promising results in getting higher accuracy of (around 0.941, when compared to the machine learning-based framework). In the future, the framework can also be designed by using different kind of neural network for better performance, and it will also be useful for the information retrieval of the mathematical document.

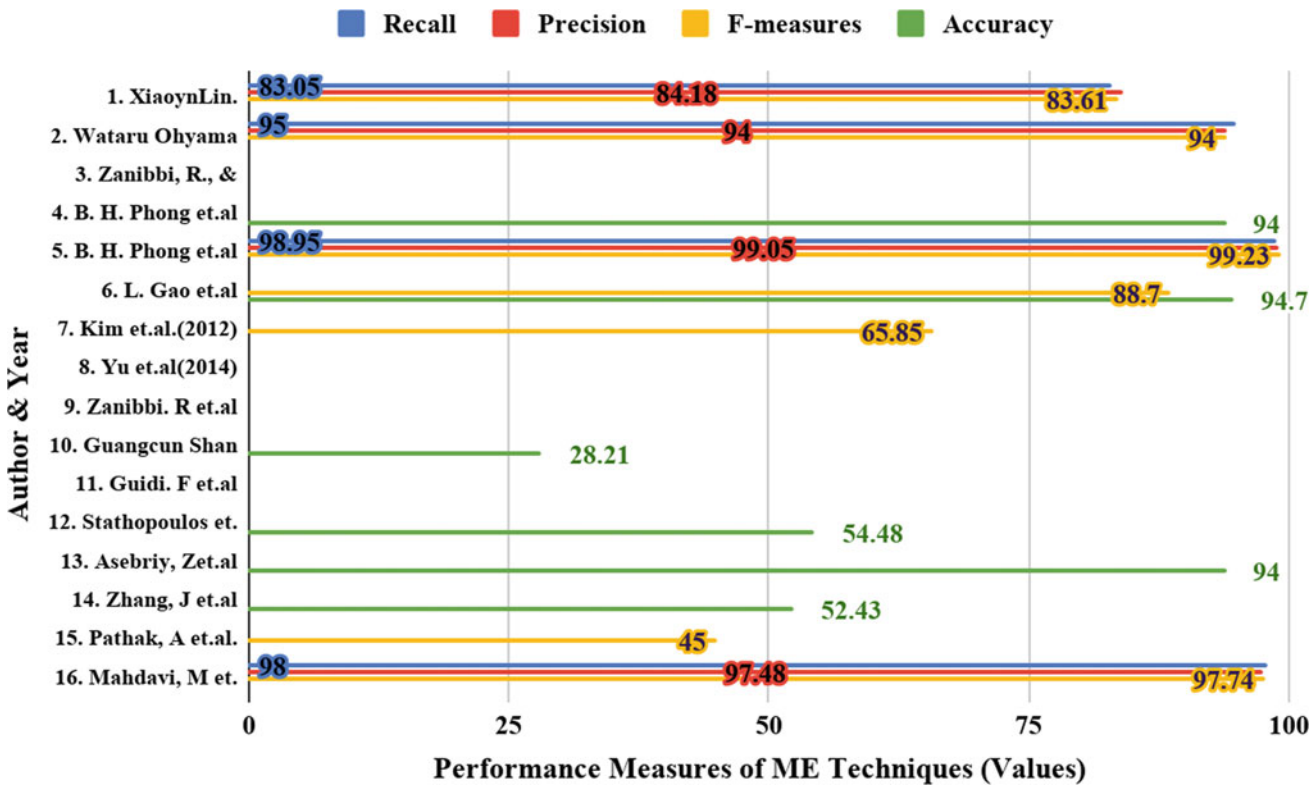

Fig. 3 Performance measures of ME techniques

References

Asebriy, Z., Raghay, S., Bencharef, O., & Kaloun, S. (2016). A semantic approach for mathematical expression retrieval. *IJACSA, 7,* 190–194.

Asebriy, Z., Raghay, S., Bencharef, O., & Kaloun. (2016). A semantic approach for mathematical expression retrieval. *IJACSA*, *7*, 190–194.

Chan, K.-F., & Yeung, D.-Y. (2000). Mathematical expression recognition: A survey. *International Journal of Document Analysis and Recognition, 3*(1), 3–15.

Foltýnek, T, Meuschke, N., Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. *ACM Computing Surveys (CSUR)*, *52*(6), 1–42.

Gao, L., Yi, X., Liao, Y., Jiang, Z., Yan, Z., & Tang, Z. (2017). A deep learning based formula detection method for PDF documents. In *Proceedings of 14th IAPR International Conference on Document Analysis Recognition (ICDAR)* (Vol. 1, pp. 553–558).

Guidi, F., & Coen, C. S. (2016). A survey on retrieval of mathematical knowledge. *Mathematics in Computer Science, 10*(4), 409–427.

Isele, M. R. (2018). Analyzing similarity in mathematical content to enhance the detection of academic plagiarism. ArXiv:1801.08439

Iwatsuki, K., Sagara, T., Hara, T., & Aizawa, A. (2017). Detecting in-line mathematical expressions in scientific documents. In *DOCENG 2017—Proceedings of the 2017 ACM Symposium on Document Engineering*. https://doi.org/10.1145/3103010.3121041

Kim, S., Yang, S., & Ko, Y. (2012a, October). Mathematical equation retrieval using plain words as a query. In Proceedings of the 21st ACM international conference on Information and knowledge management (pp. 2407–2410), (2012, October).

Kim, S., Yang, S., & Ko, Y. (2012, October). Mathematical equation retrieval using plain words as a query. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2407–2410).

Kristianto, G. Y., Goran Topic, & Aizawa, A. (2016). MCAT Math retrieval system for NTCIR-12 mathir task. In *NTCIR*.

Lee, H.-J., & Wang, J.-S. (1997). Design of a mathematical expression understanding system. *Pattern Recognition Letters, 18*(3), 289–298.

Lin, X., Gao, L., Tang, Z., Lin, X., & Hu, X. (2011a, September). Mathematical formula identification in PDF documents. In *2011 International Conference on Document Analysis and Recognition* (pp. 1419–1423). IEEE.

Lin, X., Gao, L., Tang, Z., Lin, X., & Hu, X. (2011b, September). Mathematical formula identification in PDF documents. In *2011 International Conference on Document Analysis and Recognition* (pp. 1419–1423). IEEE.

Mahdavi, M., Condon, M., Davila, K., & Zanibbi, R. (2019, September). LPGA: Line-of-sight parsing with graph-based attention for math formula recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 647–654). IEEE.

Meuschke, N., Schubotz, M., Hamborg, F., Skopal, T., & Gipp, B. (2017). Analyzing mathematical content to detect academic plagiarism. In *International Conference on Information and Knowledge Management, Proceedings*. https://doi.org/10.1145/3132847.3133144

Meuschke, N., Stange, V., Schubotz, M., & Gipp, B., Hyplag, A. (2018). hybrid approach to academic plagiarism detection. In *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR2018*. https://doi.org/10.1145/3209978.3210177

Meuschke, N., Stange, V., Schubotz, M., Kramer, M., & Gipp, B. (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*. https://doi.org/10.1109/JCDL.2019.00026

Nishizawa, G., Liu, J., Diaz, Y., Dmello, A., Zhong, W., & Zanibbi, R. (2020, April) Mathseer: A math-aware search interface with intuitive formula editing, reuse, and lookup. In *European Conference on Information Retrieval* (pp. 470–475). Cham: Springer.

Ohyama, W., Suzuki, M., & Uchida, S. (2019). Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. *IEEE Access, 7,* 144030–144042.

Pathak, A., Pakray, P., & Das, R. (2019, February). LSTM neural network based math information retrieval. In *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)* (pp. 1–6). IEEE.

Phong, B. H., Hoang, T. M., & Le, T.-L. (2017). A new method for displayed mathematical expression detection based on FFT and SVM. In: *Proceedings of 4th NAFOSTED Conference on Information and Computer Science* (pp. 90–95).

Phong, B. H., Hoang, T. M., & Le, T.-L. (2019). Mathematical variable detection based on convolutional neural network and support vector machine. In *Proceedings of International Conference Multimedia Analysis and Pattern Recognition (MAPR)* (pp. 1–5).

Phong, B. H., Hoang, T. M., & Le, T. L. (2019, May). Mathematical variable detection based on convolutional neural network and support vector machine. In *2019 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)* (pp. 1–5). IEEE.

Stathopoulos, Y., Teufel, S. (2016, December). Mathematical information retrieval based on type embeddings and query expansion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2344–2355).

Stathopoulos, Y., Teufel, S. (2016a). Mathematical information retrieval based on type embeddings and query expansion. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2344–2355).

Yu, B., Tian, X., & Luo, W. (2014a). Extracting mathematical components directly from PDF documents for mathematical expression recognition and retrieval. In *International Conference in Swarm Intelligence* (pp. 170–179). Cham: Springer.

Yu, B., Tian, X., & Luo, W. (2014, October). Extracting mathematical components directly from PDF documents for mathematical expression recognition and retrieval. In *International Conference in Swarm Intelligence* (pp. 170–179). Cham: Springer.

Zanibbi, R., & Blostein, D. (2012a). Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR), 15*(4), 331–357. https://doi.org/10.1007/s10032-011-0174-4

Zanibbi, R., & Blostein, D. (2012b). Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR), 15*(4), 331–357.

Zanibbi, R., & Blostein, D. (2012c). Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)*, *15*(4), 331–357. https://doi.org/10.1007/s10032-011-0174-4

Zanibbi, R., Davila, K., Kane, A., & Tompa, F. (2015). The tangent search engine: Improved similarity metrics and scalability for math formula search. arxiv:1507.06235

Zanibbi, R., Davila, K., Kane, A., & Tompa, F. The tangent search engine: Improved similarity metrics and scalability for math formula search. arXiv:1507.06235

Zhang, J., Du, J., & Dai, L. (2017, November). A GRU-based encoder-decoder approach with attention for online handwritten mathematical expression recognition. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 902–907). IEEE.