Formula Retrieval Using Structural Similarity

Sandip Sarkar[1], Dipankar Das[2], Partha Pakray[3] and David Pinto[3]

[1]Computer Science and Application, Hijli College, Kharagpur
[2]Computer Science and Engineering, Jadavpur University, Kolkata
[3]Computer Science and Engineering, NIT Silchar, Silchar
[3]Benemerita Universidad Autonoma de Puebla, Facultad de Ciencias de la Computacion, Puebla, Mexico

Abstract

Nowadays, the amount of scientific documents which contain mathematical equations is gradually increasing. Retrieving specific documents related to mathematical equations is very challenging. Traditional IR systems cannot deal with scientific documents. For this reason, researchers give more focus to the Mathematical Information Retrieval system, which is one of the domain-specific applications of Information Retrieval (IR). The goal of MathIR is to retrieve documents that contain normal texts as well as mathematical equations. This paper describes the JU_NITS system which has been used for math type data of Math Stack Exchange of ARQMath-3 task. We have participated in Task 2: formula retrieval. For our task, we have created three inverted indexes using three representations of mathematical equations (i.e. Latex, Opt, and slt). For similarity calculation between query and inverted index, we have used cosine similarity. The efficiency of the system is represented in terms of nDCG', MAP' and Precision at 10 measures. For each year we have submitted three runs. Among them, formulaL run gives better results which are 0.238, 0.178, and 0.161 (in nDCG' measure) for ARQMath-1, ARQMath-2, and ARQMath-3 respectively.

Keywords

Mathematical Information Retrieval, Natural Language Processing, Question Answering, Semantic Similarity

1. Introduction

Mathematics is an essential part of many disciplines, like physics, computer science, chemistry, economics, and quantum mechanics. Similarly, mathematical equations and formulas play a crucial role in scientific documents. The researcher proposed different types of approaches to solving a math word problem [1, 2]. But they give less importance if we consider the MathIR system. To retrieve this mathematical information is very challenging because it is difficult to capture the similarity between mathematical equations. The traditional IR system is not sufficient for retrieving information from those documents because those IR systems only deal with text rather than mathematical equations. Similarly, the MathIR system can handle users'

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
sandipsarkar.ju@gmail.com (S. Sarkar); dipankar.dipnil2005@gmail.com (D. Das); parthapakray@gmail.com (P. Pakray); davideduardopinto@gmail.com (D. Pinto)
http://www.dasdipankar.com/ (D. Das); http://cs.nits.ac.in/partha/ (P. Pakray); https://lke.cs.buap.mx/english/index.php/david-eduardo-pinto-avendano/ (D. Pinto)
000-0002-0955-5091 (S. Sarkar); 0000-0002-8110-9344 (D. Das); 0000-0003-3834-5154 (P. Pakray); 0000-0002-8516-5925 (D. Pinto)

needs based on the formula-based query or text-based query, or both (formula+text). The main goal of the MathIR system is to retrieve documents that fill the needs according to the query which contains text as well as mathematical equations.

It is more challenging to find the semantic similarities of two mathematical equations although they are different if we consider the syntactic pattern. For example, $a_1^2 + a_2^2 = a_3^2$ and $x_1^2 + x_2^2 = x_3^2$ are different if we consider the variable name but those mathematical expressions are semantically similar. These types of problems are challenging issues for syntax-based indexing. These shortcomings can be resolved using different types of MathML representation.

In ARQMath-3, our team JU_NITS has participated in Task 2: Formula Retrieval [3]. In this task, participants were asked to build a system in which they have to return relevant formulas according to the query. Three representations of mathematical equations i.e. Latex, Symbol Layout Trees (SLTs), and Operator Trees (OPTs) were provided by the organizer. Besides, the organizer asked the participants to produce the result of ARQMath-1 and ARQMath-2. For MathIR system, we build three inverted indexes using mathematical representations (i.e. Latex, OPT SLT). For similarity checking, we have used cosine similarity.

The rest of the paper is structured as follows. Section 2 talks about related work. The detailed description of ARQMath-3 Task 2 is given in Section 3. Section 4 shows the statistics of the dataset of ARQMath-3 Task 2. The model description is given in Section 5. The comparison and observation of our research work are given in Section 6. Finally, we conclude our task in Section 7.

2. Related Work

For the last two decades, the number of scientific publications has been gradually increasing. For this reason, the importance of mathematical information retrieval systems has increased. To encourage the researchers, many organizers organized a series of MathIR evaluation workshops to participate with the researchers. Similarly, NTCIR [4] organized MathIR workshops from 2010 to 2012. In this task, participants retrieve documents from a corpus of arXiv or Wikipedia articles according to the queries consisting of mathematical equations. ARQMath Task 2, Formula Retrieval is similar if we consider the design with NTCIR-12 Wikipedia Formula Browsing task, but it differs if we consider the query formation and evaluation method.

Similarly, the SemEval 2019 question-answering task is based on Community-question-answering [5]. The main aim of this task is to build a Math QA system that evaluates high school students: The Math SAT (short for Scholastic Achievement Test). In [6] they proposed a substitution-tree-based indexing method to solve the MathIR problem. First, Graf introduced Substitution tree indexing in which each node of the index tree represents predicates. Non-leaf nodes represent generalized substitution variables, whereas leaf nodes represent specific ones.

To find the similarity between mathematical equations, researchers use three types of mathematical representations i.e. Latex, presentation MathML, and Content MathML. The difference between those mathematical representations is that Latex is a linearized encoding that reflects the syntax of the formula while the other two forms of MathML are tree-based encoding.

To find the similarity of mathematical expressions in terms of semantically, researchers proposed variable ordering methods. In this approach, expressions are converted into a general

expression in which the name of the variable is given in common space. Expressions, "a+b^a" and "x+y^x", are semantically same but differ in terms of variable names. In this approach variable name is converted into "ids". For this those expressions are normalized into "id_1+id_2^{id1}".

To build the MathIR system researcher use different types of approaches. Those approaches can be categorized into three category a) structure-based searching via trees for math equations (tree search), b) text-based retrieval model (text search), c) both text-math search (hybrid) [7, 8]. In the structure-based approach, formula matching is performed based on tree matching (Partially or completely). As text-based search is a well-known field for researchers, they proposed different types of approaches like the bag-of-words model, TF-IDF, LSTM model, or bi-directional LSTM model[9, 10, 11] . The main challenging approach is text-math search.

3. ARQMath Task 2: Formula Retrieval

As discussed before, we have participated in ARQMath-3 Task 2: Formula Retrieval[1] which is part of CLEF 2022 (Conference and Labs of the Evaluation Forum)[2]. ARQMath has organized their workshop since 2020. The Lab uses a collection of questions and answers from the Math Stack Exchange. For training, they provided dataset[3] [4] from Math Stack Exchange[5] from 2010-2018 [12, 13, 14].

For ARQMath Task 2: Formula Retrieval, the title and the body of the questions were given to the participants. For more information, comments, answers and links to the related questions are also provided to the participants [15]. In this task, participants are asked to rank 1,000 formula according to the topic provided by the organizer. Three representations of mathematical formula (i.e. LATEX, Presentation MathML, and Content MathML format.) are provided by the organizer.

There are some differences between Task 1 and Task 2 of ARQMath. Task 1 only return answer posts while for Task 2 the answer post, as well as question post, may contain formulae. For Task 2, participants distinguish visually distinct formulae from instances of those formulae. Finally, the evaluation is done based on the ranking of the visually distinct formulae that they return. The main aim of Task 2 is not to find the answer to the question whereas the main aim of Task 1 is to find the answer to the question.

In Task 2, topic ids are written as "B.x" where x is the topic number. The next field Formula_Id is for formula instance. Another formula might be present in the title and body of the same question post. All formulas are in the latex format. For the MathIR system, queries are generated from the formulas which are provided in three TSV files (i.e. Latex, SLT, and OPT).

[1] https://www.cs.rit.edu/ dprl/ARQMath/task2-formulas.html
[2] https://clef2022.clef-initiative.eu/
[3] https://www.cs.rit.edu/ dprl/ARQMath/arqmath-resources.html
[4] https://drive.google.com/drive/u/0/folders/1ZPKIWDnhMGRaPNVLi1reQxZWTfH2R4u3
[5] https://math.stackexchange.com

4. Dataset

The organizer of ARQMath provides the dataset from the knowledge-sharing platform (i.e. Math Stack Exchange). As discussed before, the mathematical equations of the provided dataset (i.e. in question, answer, and comment post) are presented in three representations (i.e. LATEX, Presentation MathML, and Content MathML format). Figure 2 shows those three representation of $\|A\|_2 = \sqrt{\rho(A^T A)}$ equation.

The organizer used LateXML to generate the presentation MathML and Content MathML from Latex to reduce the effort of participants. For some limitations, some of the Latex formulas can not be converted into those MathML formats. The number of formula comprised in LATEX, Content MathML and Presentation MathML formats are 2,83,20,920 and 2,83,20,920 and 2,74,32,848 respectively. Table 1 gives a details description of the ARQMath Dataset.

The formula provided by the ARQMath is in Tab Separated files. Each formula is represented by each line of a TSV file which contains formula_id, thread_id, post type, and the formula. The format of each dataset is given in Figure 1.

```
<Topic number="B.309">
  <Formula_Id>q_97</Formula_Id>
  <Latex>a\in \mathbb{F}_p</Latex>
  <Title>Number of solutions of equation over a finite field</Title>
  <Question><p>I have a question regarding the number of solutions of a equation over a finite field <span class="math-
  container" id="q_91">$\mathbb{F}_p$</span>. First of all, consider the equation <span class="math-container"
  id="q_92">$x^3=a$</span> over <span class="math-container" id="q_93">$\mathbb{F}_p$</span>, where <span class="math-
  container" id="q_94">$p$</span> is a prime such that <span class="math-container" id="q_95">$p\equiv 2 (\text{mod
  }3)$</span>. The book that I'm currently reading says that this equation has exactly one solution in <span class="math-
  container" id="q_96">$\mathbb{F}_p$</span> for every <span class="math-container" id="q_97">$a\in \mathbb{F}_p$</span>,
  because <span class="math-container" id="q_98">$\gcd(3,p-1)=1$</span>, but the book does not prove this. Unfortunately,
  this doesn't convince me enough. Is there is a convincing elementary straightforward proof justifying why is this true?
  </p> </Question>
  <Tags>number-theory,elementary-number-theory,finite-fields</Tags>
</Topic>
```

Figure 1: Format of ARQMath Task 2 Dataset

Table 1
Statistics of ARQMath dataset

Dataset	No. of Formulas	Dataset Size	Index Size
Latex	2,83,20,920	1.6 GB	3.2 GB
Content MathML	2,83,20,920	9.17 GB	4.57 GB
Presentation MathML	2,74,32,848	9.84 GB	7.05 GB

5. System Architecture

For our experiments, we use Apache Solr[6] which is built on Apache Lucene, a popular, Java-based, open-source, information retrieval library. To deal with Information Retrieval System four major steps are required. Those steps are extraction, enrichment, analysis, and indexing. Extraction is the process in which documents are retrieved from the source. Enrichment is the optional step that adds information to the documents for making the documents useful for

[6]https://solr.apache.org/

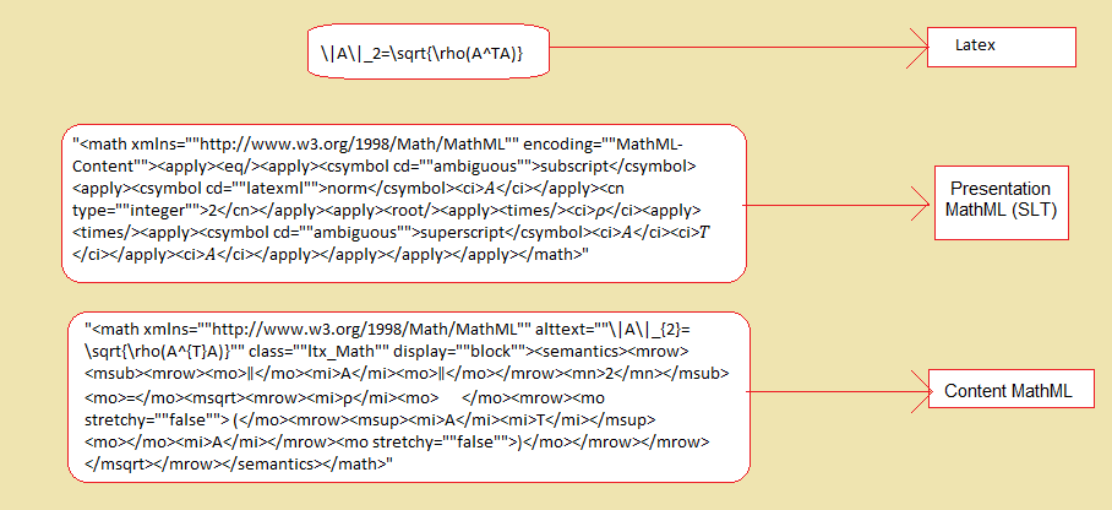

Figure 2: Representation of $\|A\|_2 = \sqrt{\rho(A^T A)}$ formula in Latex, Presentation MathML, Content MathML

relevant. For our work extraction and enrichment are done by the organizer. The next part is the analysis step in which documents are converted into tokens. Those tokens are useful for matching. For our work, we have used a standard tokenizer that split on word white space and punctuation. After analysis, Next step is indexing in which data are stored. As discussed before we have used Apache Solar in our work. Apache Solr use a specialized data structure which is known as the inverted index for matching queries with text-based documents. For this reason, it has fast searching capabilities as well as many other features. An inverted index is a dictionary of terms and a list of documents in which each term occurs. Using this index Solr finds the match between the documents to the query provided by the user. For multiple term we have used Boolean search operation (Mainly OR operation) and for ranking BM25 is used.

We have created three inverted indexes using Apache Solr from the collection dataset (i.e. from 2010-2018) based on three mathematical representations which are provided by the organizer. Next, we have extracted the formula from the topic title into a formula pool. The formula is extracted from the body if there is more than one formula is presented in the body. This formula pool is used for query processing. Next, we compute the cosine similarity between the query and inverted index which is produced from the collected dataset. Finally, our system produces 1000 lists according to the query as defined by the organizer.

6. Evaluation Measure

The primary evaluation measure for ranking in Task 2 is nDCG'. For more comparative analysis, the result is also computed in terms of MAP' and P@10. Participants have to submit 1000 ranked lists according to the query. For calculating the performance of the system trec_eval tool is used which compares the result produced by the system with the gold standard dataset(qrel file).

7. Result

The result of our system is described in Table 2 and Table 3. All participants produce the result for ARQMath-1 and ARQMath-2 by using the same system which is used in ARQMath-3. Similarly, we have produced output for ARQMath-1, ARQMath-2, and ARQMath-3 using the same system. The accuracy of the system is provided in nDCG', MAP', and P'@10 by the organizer. Table 4 shows that our approach gives satisfactory output according to the query formula.

For ARQMath-1, ARQMath-2 and ARQMath-3 Organizer consider 45 topics, 58 topics and 76 topics respectively. The result of our system(i.e. ARQMath-1 and ARQMath-2), as well as other top systems, are given in Table 3. Similarly, Table 2 shows the result of ARQMath-3 and the comparison with the winner team.

Table 2
Result of JU_NITS System and Comparison with winner team for ARQMath-3

RUN	Data	ARQMath-3 76 topics		
		nDCG'	MAP'	P'@10
JU_NITS				
formulaL	Math	0.161	0.059	0.125
formulaO	Math	0.016	0.008	0.001
formulaS	Math	0.000	0.000	0.000
Baseline				
Tangent-S	Math	0.540	0.336	0.511
approach0				
fusion_alph05	Math	0.720	0.568	0.688
fusion_alph03	Math	0.720	0.568	0.688
fusion_alph02	Math	0.715	0.558	0.659

8. Experimental Setup

Our experiments were conducted on a 4-core machine with Intel(R) i7 processor with 16 GB RAM. We have already discussed that to create an inverted index we have used Apache Solr. The version of Apache Solr is 8.11.1 with the default parameter.

9. Conclusion and Future Work

In this paper, we have built a MathIR system using Apache Solr. Information Retrieval deals with documents that are similar to the query. The difference between where traditional Information retrieval with MathIR is that in MathIR query contains both text and math equations. Normally researchers use Apache Solr for text-based search. Still, our system gives better results if we consider our basic approach with other teams. All experiments are done on the Math Stack Exchange corpus of the ARQMath task. To compare systems organizers use three different measurements. nDCG' metric is considered as the primary measure and they additionally use

Table 3
Result of JU_NITS System and Comparison with winner team for ARQMath-1 and ARQMath-2

Approach	Data	ARQMath-1 45 TOPICS			ARQMath-2 58 TOPICS		
		nDCG'	MAP'	P'@10	nDCG'	MAP'	P'@10
JU_NITS							
formulaL	Math	0.238	0.151	0.208	0.178	0.078	0.221
formulaO	Math	0.007	0.001	0.009	0.182	0.101	0.367
formulaS	Math	0.000	0.000	0.000	0.142	0.070	0.159
Baseline							
Tangent-S	Math	0.691	0.446	0.453	0.492	0.272	0.419
DPRL							
T-CFT2TED+MathAMR	Both	0.667	0.526	0.569	0.630	0.483	0.662
LTR	Both	0.733	0.532	0.518	0.550	0.333	0.491
approach0							
fusion_alph05	Math	0.647	0.507	0.529	0.652	0.471	0.612
fusion_alph03	Math	0.644	0.513	0.520	0.649	0.470	0.603
fusion_alph02	Math	0.633	0.502	0.513	0.646	0.469	0.597

Table 4
Retrieve formula of JU_NITS on ARQMath-3

Question_id	Query	Approach 1
q_11	$w = se^{i\phi}$	$w = se^{i\phi}$
		$se^{i\phi}$
		$z^n = w = se^{i\phi}$
q_6	$\|A\|_2 = \sqrt{\rho(A^T A)}$	$\|A\|_2 = \sqrt{\rho(A^T A)}$
		$\|A\|_2 = \sqrt{\rho(A^T A)} \leq \sqrt{\|A^T A\|_\infty}$
		$A \in \mathbb{R}^{m,n} \Rightarrow \ \|A\|_2 = \sqrt{\rho(A^T A)}$
q_29	$[x,[y,z]] + [y,[z,x]] + [z,[x,y]] = 0$	$[x,[y,z]] + [y,[z,x]] + [z,[x,y]] = 0$
		$[x,[y,z]] + [y,[z,x]] + [z,[x,y]] = ... = 0$
		$[[X,Y],Z] + [[Y,Z],X] + [[Z,X],Y] = 0.$
q_95	$p \equiv 2 (\mathrm{mod}\ 3)$	$p \equiv 2 (\mathrm{mod}\ 3)$
		$p \equiv 2 \bmod 3$
		$sum_{p \equiv 2 \bmod 3}$

MAP' and P'@10 measurements. We have shown that our system works well if we take the latex format of the mathematical equation as input. Our model only works on math equations. In the future, we want to build a system that can take different types of input (i.e. text-only, math-only, and math-text). Besides we want to reduce the searching and retrieval time. Index optimization is also our future task which helps with our retrieval and formula searching process.

References

[1] S. Sarkar, D. Das, P. Pakray, D. E. P. Avendano, Developing mcqa framework for basic science subjects using distributed similarity model and classification based approaches, International Journal of Asian Language Processing 30 (2020) 2050015. URL: https://doi.org/10.1142/S2717554520500150. doi:10.1142/S2717554520500150. arXiv:https://doi.org/10.1142/S2717554520500150.

[2] S. Sarkar, D. Das, P. Pakray, JU NITM at IJCNLP-2017 task 5: A classification approach for answer selection in multi-choice question answering system, in: Proceedings of the IJCNLP 2017, Shared Tasks, Asian Federation of Natural Language Processing, Taipei, Taiwan, 2017, pp. 213–216. URL: https://aclanthology.org/I17-4036.

[3] B. Mansouri, V. Novotný, A. Agarwal, D. W. Oard, R. Zanibbi, Overview of ARQMath-3 (2022): Third CLEF lab on Answer Retrieval for Questions on Math (Working Notes Version), in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, 2022.

[4] K. Kishida, M. P. Kato, Overview of NTCIR-12, in: N. Kando, T. Sakai, M. Sanderson (Eds.), Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences, Tokyo, Japan, June 7-10, 2016, National Institute of Informatics (NII), 2016. URL: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/pdf/ntcir/OVERVIEW/01-NTCIR12-OV-KishidaK.pdf.

[5] M. Hopkins, R. Le Bras, C. Petrescu-Prahova, G. Stanovsky, H. Hajishirzi, R. Koncel-Kedziorski, SemEval-2019 task 10: Math question answering, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 893–899. URL: https://aclanthology.org/S19-2153. doi:10.18653/v1/S19-2153.

[6] A. Pathak, P. Pakray, S. Sarkar, D. Das, A. Gelbukh, MathIRs: Retrieval System for Scientific Documents, ComputaciÃy Sistemas 21 (2017) 253 – 265. URL: http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-55462017000200253&nrm=iso.

[7] W. Zhong, X. Zhang, J. Xin, R. Zanibbi, J. Lin, Approach zero and anserini at the CLEF-2021 arqmath track: Applying substructure search and BM25 on operator tree path tokens, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 133–156. URL: http://ceur-ws.org/Vol-2936/paper-09.pdf.

[8] B. Mansouri, D. W. Oard, R. Zanibbi, DPRL systems in the CLEF 2021 arqmath lab: Sentence-bert for answer retrieval, learning-to-rank for formula retrieval, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 47–62. URL: http://ceur-ws.org/Vol-2936/paper-04.pdf.

[9] R. Avenoso, B. Mansouri, R. Zanibbi, XY-PHOC symbol location embeddings for math formula retrieval and autocompletion, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936

of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 25–35. URL: http://ceur-ws.org/Vol-2936/paper-02.pdf.

[10] P. Dadure, P. Pakray, S. Bandyopadhyay, Bert-based embedding model for formula retrieval, in: CLEF, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-03.pdf.

[11] S. Sarkar, P. Pakray, D. Das, A. Gelbukh, Regression based approaches for detecting and measuring textual similarity, in: R. Prasath, A. Gelbukh (Eds.), Mining Intelligence and Knowledge Exploration, Springer International Publishing, Cham, 2017, pp. 144–152.

[12] S. Rohatgi, J. Wu, C. L. Giles, Ranked list fusion and re-ranking with pre-trained transformers for arqmath lab, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 125–132. URL: http://ceur-ws.org/Vol-2936/paper-08.pdf.

[13] B. Mansouri, S. Rohatgi, D. Oard, J. Wu, C. Giles, R. Zanibbi, Tangent-cft: An embedding model for mathematical formulas, 2019. doi:10.1145/3341981.3344235.

[14] Y. K. Ng, D. J. Fraser, B. Kassaie, F. W. Tompa, Dowsing for answers to math questions: Ongoing viability of traditional mathir, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 63–81. URL: http://ceur-ws.org/Vol-2936/paper-05.pdf.

[15] B. Mansouri, R. Zanibbi, D. W. Oard, A. Agarwal, Overview of arqmath-2 (2021): Second clef lab on answer retrieval for questions on math, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2021, pp. 215–238.