ARQMath: A New Benchmark for Math-Aware CQA and Math Formula Retrieval

Richard Zanibbi, Behrooz Mansouri, and Anurag Agarwal
Rochester Institute of Technology (USA)
{*rxzvcs,bm3302,axasma*}*@rit.edu*

Douglas W. Oard
University of Maryland, College Park (USA)
oard@umd.edu

Abstract

The Answer Retrieval for Questions on Math (ARQMath) evaluation was run for the first time at CLEF 2020. ARQMath is the first Community Question Answering (CQA) shared task for math, retrieving existing answers from Math Stack Exchange (MSE) that can help to answer previously unseen math questions. ARQMath also introduces a new protocol for math formula search, where formulas are evaluated in context using a query formula's associated question post, and posts associated with each retrieved formula. Over 70 topics were annotated for each task by eight undergraduate students supervised by a professor of mathematics. A formula index is provided in three formats: LaTeX, Presentation MathML, and Content MathML, avoiding the need for participants to extract these themselves. In addition to detailed relevance judgments, tools are provided to parse MSE data, generate question threads in HTML, and evaluate retrieval results. To make comparisons with participating systems fairer, nDCG′ (i.e., nDCG for assessed hits only) is used to compare systems for each task. ARQMath will continue in CLEF 2021, with training data from 2020 and baseline systems for both tasks to reduce barriers to entry for this challenging problem domain.

1 Introduction

Math *is* hard. As a result, many people regularly consult Community Question Answering (CQA) sites such as Math Stack Exchange (MSE) and Math Overflow to find answers to their math questions. There is also evidence that for queries about math, users pose well formed questions to general-purpose search engines as much as ten times more often than they do for queries in general [Mansouri et al., 2019b]. This suggests an unmet user need for effective *math-aware* search engines that process both keywords and formulas effectively, and answering mathematical questions is a challenging AI task that has also begun to attract attention for its own sake, particularly from researchers in the Information Retrieval and Natural Language Processing communities.

Within the past five years, a number of shared tasks have been held to advance techniques for math question answering and math-aware search. One held at SemEval 2019 required answering

(primarily multiple-choice) questions from the MathSAT (Scholastic Achievement Test). Within the IR community, the first benchmarks for text + formula search and formula retrieval in document collections were developed through NTCIR using collections created from the arXiv and Wikipedia [Aizawa et al., 2013, 2014; Zanibbi et al., 2016a]. This helped to grow the community of researchers working on Mathematical Information Retrieval (MIR) [Zanibbi and Blostein, 2012; Aizawa and Kohlhase, 2020].

However, despite the popularity of services like Math Stack Exchange, there had not yet been a shared task on Community Question Answering (CQA) for mathematical content. To address this, we organized the first Answer Retrieval for Questions on Math (ARQMath[1]) lab at CLEF 2020 [Zanibbi et al., 2020] using MSE content. The primary task is to return a ranked list of potential answers to a given question post. A second task is formula search - given a formula, find formulas within question and answer posts that are relevant to the query formula in its *original context* (i.e., within the post where it appears). This has some similarities to the NTCIR-12 formula search task [Zanibbi et al., 2016a], but there formula context was not considered during assessment. Both ARQMath tasks are illustrated in Figure 1.

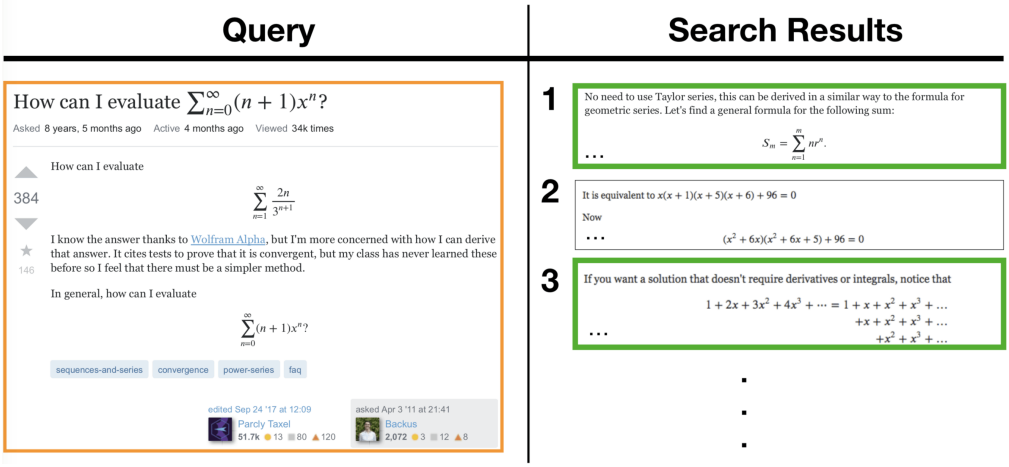
As a very brief introduction, existing formula retrieval models may be categorized as *text-based*, *tree-based* or *embedding-based*. Text-based models use string representations for formula indexing and retrieval. For example, the MIaS system [Sojka and Líška, 2011] linearized MathML representations of formula appearance, and then incorporated text tokens generated from formulas in a TF-IDF retrieval model. Tree-based models use hierarchical representation(s) of formula appearance and operation syntax, with similarity computed using paths and/or subtrees [Kristianto et al., 2016; Zhong et al., 2020; Davila and Zanibbi, 2017], or entire trees (e.g., using tree edit distances [Kamali and Tompa, 2013]). Embedding models project formulas onto points in Euclidean space, with retrieval performed by identifying neighboring formulas. Early embedding models [Thanda et al., 2016; Gao et al., 2017] made use of text models such as the distributed bag of words (PV-DBOW) [Le and Mikolov, 2014], trying to embed equations similar to text sentences, with the 'words' comprised of variables, operators and other symbols in the formula. Tangent-CFT [Mansouri et al., 2019a] uses fastText [Bojanowski et al., 2017] to produce an n-gram embedding model for linearized formula tree edges. Recently [Pfahler and Morik, 2020] proposed embedding formulas using graph convolutional neural networks. Methods that jointly embed text and formulas include TopicEQ [Yasunaga and Lafferty, 2019], which uses topic models for text and topic-dependent RNNs for formulas.

In the remainder of this paper we summarize the tasks, collection, and available tools and other resources for ARQMath. Our goal in sharing this is to let others know of the test collections and tools that ARQMath is making available, and to invite participation in future shared task evaluations.

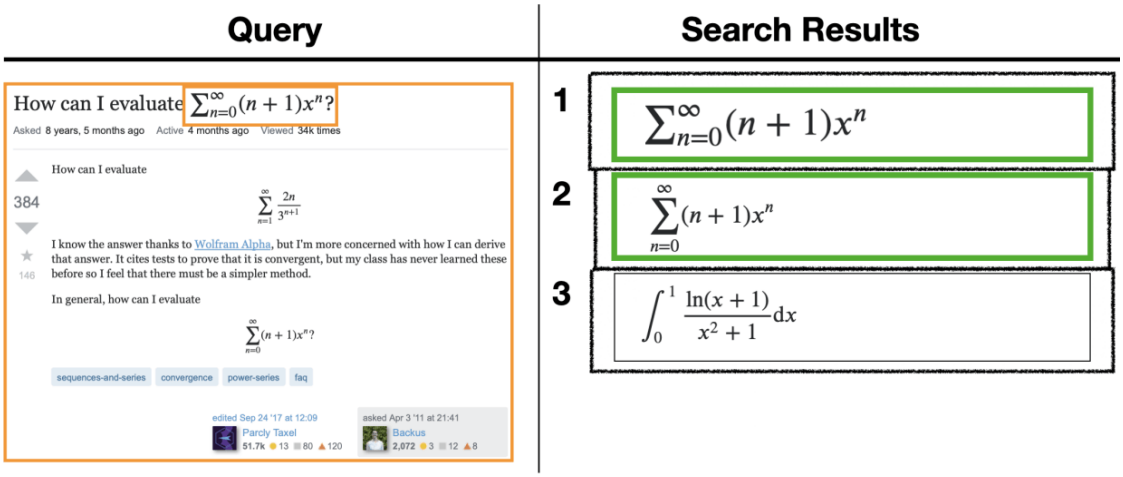
2 ARQMath Tasks

Figure 1 illustrates the two tasks in ARQMath 2020: (1) Finding Answers to Math Questions, and (2) Formula Search. ARQMath 2020's primary task was answer retrieval for math questions. The participants were presented with a question posted in 2019 on MSE, and were asked to return up

[1] https://www.cs.rit.edu/~dprl/ARQMath



Task 1: Answer Retrieval



Task 2: Formula Retrieval

Figure 1: ARQMath Tasks. The main task requires retrieving 'old' answers to 'new' questions within the 2010-2018 MSE collection, for question posts taken from 2019. The second task is formula retrieval, where query formulas are taken from the question posts for Task 1, and relevant formulas are returned along with their associated posts (i.e., with their context)

to 1,000 relevant answers from answers published in prior years (2010-2018). For the evaluation of participating systems, 77 topics were assessed.

In the formula retrieval task, participants were presented with one formula from a 2019 question used in Task 1, and asked to return a ranked list of up to 1,000 formula instances from questions or answers from the collection. Formulas were returned by their identifiers within the collection along with their associated post identifiers. For the evaluation of participating systems, 45 topics were assessed, and an additional 27 topics were annotated after system scores were calculated.

3 ARQMath Collection (Math Stack Exchange 2010-2018)

The ARQMath collection contains question and answer posts from Math Stack Exchange (MSE), which are freely available from the Internet Archive. The collection contains posts published from 2010 to 2018, with a total of 1 million questions and 1.4 million answers. In ARQMath 2020, posts from 2019 were used for topic construction. The collection contains the following:

- **Posts**: Each MSE post has a unique identifier, and a field indicating whether it is a question or answer post. Questions have a title and a body (content of the question), while answers only have a body. Each answer has a '`parent id`' indicating which question it was posted in answer to. Additional information such as the score, the post owner id and creation date are also available.
- **Comments**: MSE users can comment on posts. Each comment has a unique identifier and a '`post id`' indicating which post the comment is written for.
- **Post links**: Moderators sometimes identify duplicate or related questions that have been previously asked. A '`post link type id`' of value 1 indicates related posts, while value 3 indicates duplicates.
- **Tags**: Questions can have one or more tags describing the subject matter of the question (e.g., 'real-analysis').
- **Votes**: The post score is the difference between up and down votes, and there are also other vote types such as '`offensive`' or '`spam.`' Each vote has a '`vote type id`' for the vote type and a '`post id`' for the associated post.
- **Users**: Registered MSE users have a unique id. Each user has a reputation score, which may be increased through activities such as posting a high quality answer, or posting a question that receives up votes. Registered MSE users can also receive one of three badges: bronze, silver or gold.

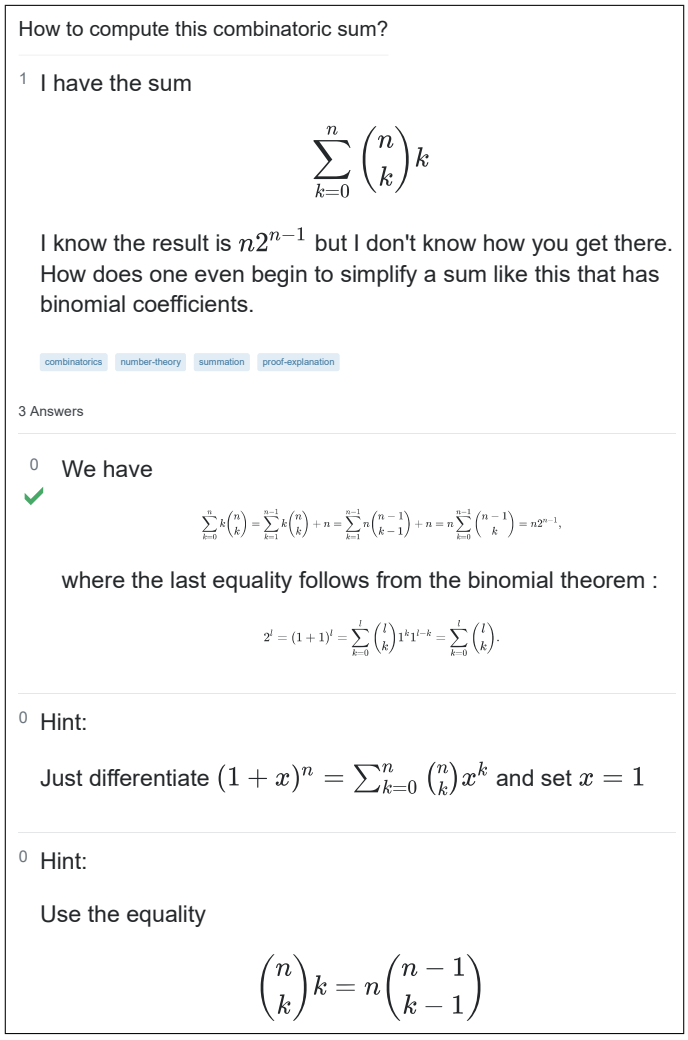
Edit history for posts and comments is also available from the Internet Archive, but for simplicity and to reduce the collection size we have not included history information in the ARQMath collection.

Provided with the collection are the question threads, containing each question along with its associated answer posts in the collection (see Figure 2). These threads can be regenerated from the ARQMath collection using a provided Python script. Threads are provided to make studying the collection easier for participants, and for use in relevance assessment (discussed further below). The organizers have also provided Python scripts for loading and parsing MSE data.[2]

3.1 Formula Representations and Formula Index

In the Internet Archive version of the collection, formulas are located between two '$' or '$$' signs, or inside a 'math-container' tag. For ARQMath, all formulas in posts (and all MSE comments on those posts) have been tagged, with a unique identifier assigned to each formula instance.

[2] https://github.com/ARQMath/ARQMathCode



How to compute this combinatoric sum?

1 I have the sum

$$\sum_{k=0}^{n} \binom{n}{k} k$$

I know the result is $n2^{n-1}$ but I don't know how you get there. How does one even begin to simplify a sum like this that has binomial coefficients.

combinatorics number-theory summation proof-explanation

3 Answers

0 We have

$$\sum_{k=0}^{n} k\binom{n}{k} = \sum_{k=1}^{n-1} k\binom{n}{k} + n = \sum_{k=1}^{n-1} n\binom{n-1}{k-1} + n = n\sum_{k=0}^{n-1} \binom{n-1}{k} = n2^{n-1},$$

where the last equality follows from the binomial theorem :

$$2^l = (1+1)^l = \sum_{k=0}^{l} \binom{l}{k} 1^k 1^{l-k} = \sum_{k=0}^{l} \binom{l}{k}.$$

0 Hint:

Just differentiate $(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k$ and set $x = 1$

0 Hint:

Use the equality

$$\binom{n}{k} k = n\binom{n-1}{k-1}$$

Figure 2: Example question thread generated from MSE content using an ARQMath Python script. The amount of text in posts varies greatly; this thread has relatively little text.

Each formula is represented three ways to facilitate participation by teams without specialized expertise in mathematical notation processing: (a) as LaTeX strings, (b) as (appearance-based) Presentation MathML, and (c) as (operator tree) Content MathML. Appearance is represented by the spatial arrangement of symbols on writing lines (in Symbol Layout Trees (SLTs)), and mathematical syntax is represented using a hierarchy of operators and arguments (in Operator Trees (OPTs)). Figure 3 shows the SLT and OPT representations for the formula $x^2 = 4$. In the SLT representation, edge labels show the spatial relationship between the formula elements. For instance, the edge label 'a' shows that the number '2' is located above the variable 'x', while the edge label 'n' shows operator '=' is located next after 'x'. In the OPT representation, the edge labels for non-commutative operators indicate argument position. For more details, refer to [Zanibbi et al., 2016b].

The open source LaTeXML tool[3] is used to convert LaTeX to MathML. For ARQMath 2020,

[3] https://dlmf.nist.gov/LaTeXML

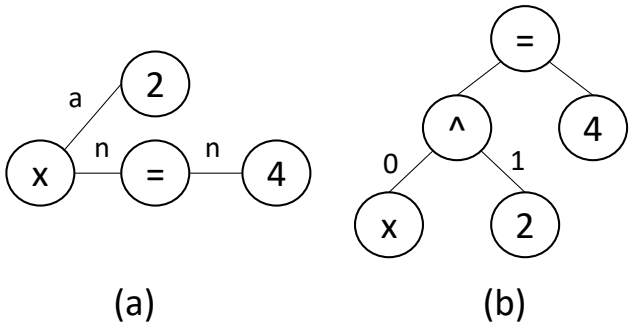


Figure 3: Formula $x^2 = 4$ represented as (a) Symbol Layout Tree (SLT) and (b) Operator Tree (OPT). SLTs represent formula appearance by the position of symbols on writing lines, while operator trees represent the hierarchy of operations and (where pertinent) argument order.

92% of indexed LaTeX formulas were successfully converted to Presentation MathML (SLT), and 90% to Content MathML (OPT). LaTeXML was recently updated to cover more notation, so we expect to increase the number of successful MathML conversions very substantially for ARQMath 2021.

4 Evaluation and Results

One risk when performing pooling for a new task for which rich training data is not yet available is that an unusually large number of relevant answers may remain unjudged. Measures which treat unjudged documents as not relevant can be used when directly comparing systems that contributed to the judgment pools, but subsequent use of such a first-year test collection (e.g., to train new systems for the second year of the lab) can be disadvantaged by treating unjudged documents (which as systems improve might actually be relevant) as not relevant. We therefore chose nDCG′ (read as "nDCG-prime") [Sakai and Kando, 2008] as the primary measure for both tasks as a simple way of maximizing the comparability of *post hoc* experiments.

Task 1. A total of 77 questions were assessed for Task 1 by 8 paid upper-year or recently graduated undergraduate math students who worked for over a month, and underwent training in multiple phases. For each question, 4 relevance degrees were considered, as shown in Table 1. For these questions, an average of 508.5 answer posts were judged for each question, with an average assessment time of 63.1 seconds per answer post. Among the 5 participating teams and the baseline systems, the MathDowsers team [Ng et al., 2020] obtained the highest nDCG′.

Task 2. A total of 45 topics were assessed for Task 2 with an average of 125.0 formulas assessed per topic, with an average assessment time of 38.1 seconds per formula. The assessment was done by 3 assessors, each of whom had done some initial assessment work for Task 1. As with the first task, 4 relevance levels were used for task 2, as defined in Table 1. Assessors were presented with formula instances, and asked to decide their relevance by considering whether retrieving that instance of that formula would have been useful, assigning each formula instance in the judgment pool one of four scores. Instance assessments were then max-aggregated to establish the maximum relevance degree for each visually distinct formula, and nDCG′ was then computed over visually

Table 1: Relevance Scores, Ratings, and Definitions for Tasks 1 and 2.

Task 1: Question Answering		
Score	Rating	Definition
3	High	Sufficient to answer the complete question on its own
2	Medium	Provides some path towards the solution. This path might come from clarifying the question, or identifying steps towards a solution
1	Low	Provides information that could be useful for finding or interpreting an answer, or interpreting the question
0	Not Relevant	Provides no information pertinent to the question or its answers. A post that restates the question without providing any new information is considered non-relevant

Task 2: Formula Retrieval		
Score	Rating	Definition
3	High	Just as good as finding an exact match to the query formula would be
2	Medium	Useful but not as good as the original formula would be
1	Low	There is some chance of finding something useful
0	Not Relevant	Not expected to be useful

distinct formulas. After results were reported, an additional 27 topics were annotated, resulting in a total of 74 topics that are now available for Task 2 training. Among the 3 participating teams and the baseline system, a Tangent-S baseline [Davila and Zanibbi, 2017] achieved the highest nDCG′; although a participating system [Mansouri et al., 2020] did achieve the best P@10.

5 ARQMath 2021

ARQMath 2021 will continue with the same two tasks. For ARQMath 2020, the selection of question posts for topic construction was restricted to those with at least one related post link to a question in the collection to be searched. We did this to minimize the risk of investing assessment effort on topics that yielded no relevant documents. For ARQMath 2021 we plan to remove this restriction, and instead guard against wasted assessment effort by doing a limited amount of pre-assessment for the results of an ARQMath 2020 system.

Finally, for ARQMath 2021 we will be making baseline systems available for both tasks (including the best-performing system for Task 2), so that people interested in the task can quickly run and study these systems as starting points. Registration for ARQMath 2021 will be open through April, 2021[4]. We warmly invite you to look at the ARQMath data and tools, and join the next edition of the lab - it should be interesting, challenging, and more than a bit of fun.

References

Akiko Aizawa and Michael Kohlhase. Mathematical information retrieval. In *Evaluating Information Retrieval and Access Tasks*, pages 169–185. Springer, Singapore, 2020.

[4] http://clef2021-labs-registration.dei.unipd.it

Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. NTCIR-10 math pilot task overview. In *NTCIR*, 2013.

Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. NTCIR-11 Math-2 task overview. In *NTCIR*, volume 11, pages 88–98, 2014.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguistics*, 5:135–146, 2017. URL https://transacl.org/ojs/index.php/tacl/article/view/999.

Kenny Davila and Richard Zanibbi. Layout and semantics: Combining representations for mathematical formula search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1165–1168, 2017.

Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang. Preliminary exploration of formula embedding for mathematical information retrieval: can mathematical formulae be embedded like a natural language? 2017.

Shahab Kamali and Frank Wm Tompa. Retrieving documents with mathematical content. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 353–362. ACM, 2013.

Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. MCAT math retrieval system for NTCIR-12 MathIR task. In *NTCIR*, 2016.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, 2014.

Behrooz Mansouri, Shaurya Rohatgi, Douglas W Oard, Jian Wu, C Lee Giles, and Richard Zanibbi. Tangent-CFT: An embedding model for mathematical formulas. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)*, pages 11–18, 2019a.

Behrooz Mansouri, Richard Zanibbi, and Douglas W. Oard. Characterizing searches for mathematical concepts. In *Joint Conference on Digital Libraries*, 2019b.

Behrooz Mansouri, Douglas W Oard, and Richard Zanibbi. DPRL systems in the CLEF 2020 arqmath lab. In *Working Notes of CLEF 2020-Conference and Labs of the Evaluation Forum*, 2020.

Yin Ki Ng, Dallas J. Fraser, Besat Kassaie, George Labahn, Mirette S. Marzouk, Frank Wm. Tompa, and Kevin Wang. Dowsing for math answers with Tangent-L. In *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum*, 2020.

Lukas Pfahler and Katharina Morik. Semantic search in millions of equations. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 135–143, 2020.

Tetsuya Sakai and Noriko Kando. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval*, 11(5):447–470, 2008.

Petr Sojka and Martin Líška. The art of mathematics retrieval. In *Proceedings of the 11th ACM Symposium on Document Engineering*, 2011.

Abhinav Thanda, Ankit Agarwal, Kushal Singla, Aditya Prakash, and Abhishek Gupta. A document retrieval system for math queries. In *NTCIR*, 2016.

Michihiro Yasunaga and John D Lafferty. Topiceq: A joint topic and mathematical equation model for scientific texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7394–7401, 2019.

Richard Zanibbi and Dorothea Blostein. Recognition and retrieval of mathematical expressions. *International Journal on Document Analysis and Recognition (IJDAR)*, 15(4):331–357, 2012.

Richard Zanibbi, Akiko Aizawa, Michael Kohlhase, Iadh Ounis, Goran Topic, and Kenny Davila. NTCIR-12 MathIR task overview. In *NTCIR*, 2016a.

Richard Zanibbi, Kenny Davila, Andrew Kane, and Frank Wm Tompa. Multi-stage math formula search: Using appearance-based similarity metrics at scale. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 145–154, 2016b.

Richard Zanibbi, Douglas W Oard, Anurag Agarwal, and Behrooz Mansouri. Overview of ARQMath 2020 (updated working notes version): CLEF lab on answer retrieval for questions on math. In *Working Notes of CLEF 2020-Conference and Labs of the Evaluation Forum*, 2020.

Wei Zhong, Shaurya Rohatgi, Jian Wu, C Lee Giles, and Richard Zanibbi. Accelerating substructure similarity search for formula retrieval. In *European Conference on Information Retrieval*, pages 714–727. Springer, 2020.