Chapter 12
Mathematical Information Retrieval

Akiko Aizawa and Michael Kohlhase

Abstract We present an overview of the NTCIR Math Tasks organized during NTCIR-10, 11, and 12. These tasks are primarily dedicated to techniques for searching mathematical content with formula expressions. In this chapter, we first summarize the task design and introduce test collections generated in the tasks. We also describe the features and main challenges of mathematical information retrieval systems and discuss future perspectives in the field.

12.1 Introduction

The NTCIR Math Tasks are aimed at developing test collections for mathematical search in STEM (Science/Technology/Engineering/Mathematics) documents to facilitate and encourage research in mathematical information retrieval (MIR) (Liska et al. 2011) and its related fields (Guidi and Sacerdoti Coen 2016; Zanibbi and Blostein 2012).

Mathematical formulae are important for the dissemination and communication of scientific information. They are not only used for numerical calculation but also for clarifying definitions or disambiguating explanations that are written in natural language. Despite the importance of math in technical documents, most contemporary information retrieval systems do not support users' access to mathematical formulae in target documents. One major obstacle to MIR research is the lack of readily available large-scale datasets with structured mathematical formulae, carefully designed tasks, and established evaluation methods.

MIR involves searching for a particular mathematical concept, object, or result, often expressed using mathematical formulae, which—in their machine-readable

A. Aizawa (✉)
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
e-mail: aizawa@nii.ac.jp

M. Kohlhase
FAU Erlangen-Nürnberg, Martenstr. 3, 91058 Erlangen, Germany
e-mail: michael.kohlhase@fau.de

T. Sakai et al. (eds.), *Evaluating Information Retrieval and Access Tasks*,
The Information Retrieval Series 43,
https://doi.org/10.1007/978-981-15-5554-1_12

forms—are expressed as complex expression trees. To answer MIR queries, a search system should tackle at least two challenges: (1) tree structure search and (2) utilization of textual context information.

To understand the problem, consider an engineer who wants to prevent an electrical system from overheating, thus, needs a tight upper estimate for the energy term

$$\int_a^b |V(t)I(t)|dt$$

for all a, b, where V is voltage and I current. Search engines, such as Google, are restricted to word-based searches of mathematical articles, which barely helps with finding mathematical objects because there are no keywords to search for. Computer algebra systems cannot help either since they do not incorporate the necessary special knowledge. However, the required information is out there, e.g., in the form of

Theorem 17. (Hölder's Inequality)
If f and g are measurable real functions, $l, h \in \mathbb{R}$, and $p, q \in [0, \infty)$, such that $1/p + 1/q = 1$, then

$$\int_l^h |f(x)g(x)|\,dx \leq \left(\int_l^h |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_l^h |g(x)|^q\,dx\right)^{\frac{1}{q}}$$

For mathematical content (here the statement of Hölder's inequality) to be truly searchable, it must be in a form in which an MIR system can find it from a query

$$\int_{\boxed{\text{a}}}^{\boxed{\text{b}}} |V(t)I(t)|dt \leq \boxed{\text{R}}$$

the boxed identifiers are query variables (see Sect. 12.3.2)—and can even extend the calculation to

$$\int_a^b |V(t)I(t)|dt \leq \left(\int_a^b |V(x)|^2\,dx\right)^{\frac{1}{2}} \left(\int_a^b |I(x)|^2\,dx\right)^{\frac{1}{2}}$$

after the engineer chooses $p = q = 2$ (Cauchy–Schwarz inequality). Estimating the individual V and I values is now a much simpler problem.

Admittedly, Google would have found the information by querying for "Cauchy–Schwarz Hölder", but that keyword was the crucial information the engineer was missing in the first place. In fact, it is not unusual for mathematical document collections to be so large that determining the identifier of the sought-after object is harder than recreating the actual object.

In this example we see the effect of both (1) formula structure search and (2) context information as postulated above:

1. The formula structure is mapped by unification (finding a substitution for the boxed query variables to make the query and main formula of Hölder's inequality structurally identical or similar (see Sect. 12.3.2).

2. We have used the context information about the parameters of Hölder's inequality, e.g., that the identifiers f, g, p, and q are universal (thus can be substituted for); the first two are measurable functions and the last two are real numbers.

In the following sections, we summarize our attempts at NTCIR to develop datasets for MIR together with some future perspectives of the field.

12.2 NTCIR Math: Overview

Prior to the NTCIR Math Tasks, MIR had been mainly approached by researchers in digital mathematics libraries, and only a little attention has been paid by the information retrieval community. Unlike other scientific disciplines that require a search for specific types of named entities such as genes, diseases, and chemical compounds, mathematics is based on abstract concepts with many possible interpretations when mapped to a real-world phenomenon. This means that although their mathematical definitions are rigid, mathematical concepts are inherently ambiguous in their applications to the real world. Also, the representation of mathematical formulae can be highly complicated with diverse types of symbols including user-defined functions, constants, and free and bound variables. As such, MIR requires dedicated search techniques such as approximate tree matching or unification. To summarize, in the context of information retrieval, MIR is not only a challenge for novel retrieval targets but also featured as a testbed for (1) retrieval of non-textual objects in documents using their context information and (2) a large-scale complex tree structure search with a realistic application scenario.

The NTCIR Math tasks were the first trial to introduce an evaluation framework of information retrieval to mathematical formula search. NTCIR Math Tasks were organized three times during NTCIR-10, 11, and 12, i.e., the NTCIR-10 Math Pilot Task, NTCIR-11 Math-2 Task, and NTCIR-12 MathIR Task.

12.2.1 NTCIR-10 Math Pilot Task

The NTCIR-10 Math Pilot Task (Aizawa et al. 2013) was the first attempt to develop a common workbench for mathematical formula search. This task was organized as two independent subtasks:

1. The first was the Math Retrieval Subtask in which the objective was to retrieve relevant documents given a math query.
2. The second was the Math Understanding Subtask in which the objective was to identify textual spans that describe math formulae that appear in the document.

The corpus used for this task was based on 100,000 arXiv documents converted from LaTeX to XHTML by the arXMLiv project.[1]

Six teams participated in this task, all six contributing to the Math Retrieval Subtask and only one to the Math Understanding Subtask.

12.2.2 NTCIR-11 Math-2 Task

The NTCIR-10 Math Pilot Task showed that participants considered the Math Retrieval Subtask more important. Therefore, the succeeding two tasks focused only on this subtask and made it as compulsory for all participants. In the NTCIR-11 Math-2 Task (Aizawa et al. 2014), based on the feedback from the participants in the pilot task, both the arXiv corpus and topics were reconstructed. Apart from this main subtask using the arXiv corpus, the NTCIR-11 Math-2 Task also provided an open free subtask using math-related Wikipedia articles. This optional subtask required an exact formula search (without any keywords) and complements the main subtask with an automated performance evaluation.

The NTCIR-11 Math-2 Task had eight teams participating (two new teams joined), most contributing to both subtasks .

12.2.3 NTCIR-12 MathIR Task

For the NTCIR-12 MathIR Task (Zanibbi et al. 2016), we reused the arXiv corpus we prepared for the NTCIR-11 Math-2 Task but with new topics. This subtask introduced a new formula query operator, *simto region*, that explicitly requires an approximate matching function for math formulae. We also created a new corpus of Wikipedia articles to provide a use case of math retrieval by nonexperts. The design of the subtask for the Wikipedia corpus was similar to that in the NTCIR-11 Math-2 Task except that a topic includes not only exact formula search but also formula+keyword search (Table 12.1).

Six teams participated in the NTCIR-12 MathIR Task.

12.3 NTCIR Math Datasets

In this section, we mainly describe the two datasets, arXiv and Wikipedia, designed for the Math Retrieval Subtasks during NTCIR-12. Each dataset consists of a corpus with mathematical formulae, a set of topics in which each query is expressed as

[1] https://kwarc.info/projects/arXMLiv/.

Table 12.1 Summary of NTCIR math subtasks

Subtasks		NTCIR-10	NTCIR-11	NTCIR-12
Math Retrieval Subtask for the ArXiv cor pus	Formula search	○		
	Formula+keyword search	○	○	○
	Formula+keyword search with “simto”			○
	Free-form query search	○		
Math Retrieval Subtask for the Wikipedia cor pus	Formula search		○	○
	Formula+keyword search			○
	Formula+keyword search with ‘simto’			
Math understanding subtask		○		

a combination of mathematical formulae schemata and keywords, and relevance judgment results based on the submissions from participating teams.

12.3.1 Corpora

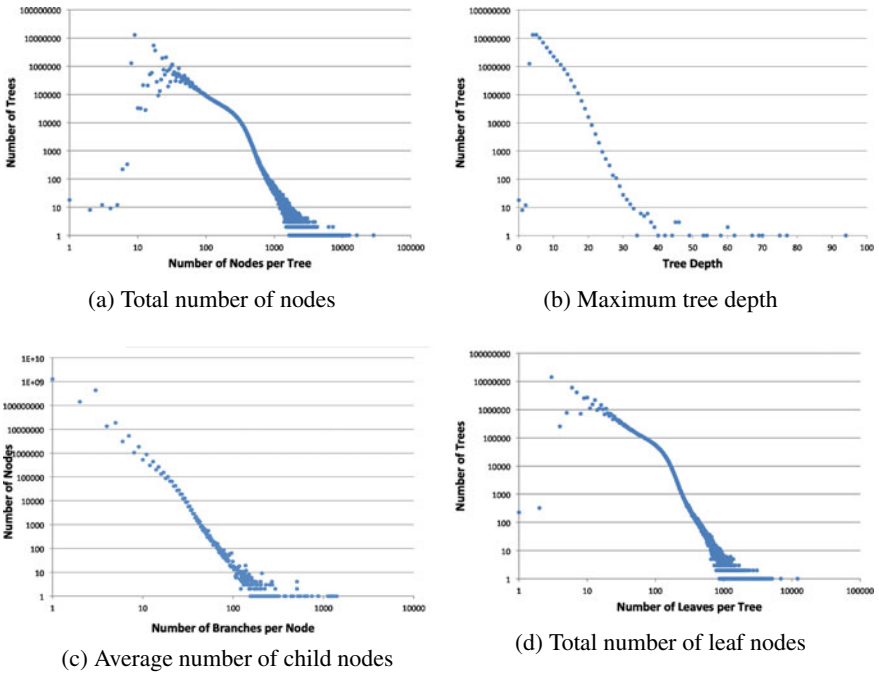
The arXiv corpus contains paragraphs from technical articles in the arXiv,[2] while the Wikipedia corpus contains complete articles from Wikipedia. Generally speaking, the arXiv articles (preprints of research articles) were written by technical experts for technical experts assuming a high level of mathematical sophistication from readers. In contrast, many Wikipedia articles on mathematics were written to be accessible for nonexperts at least in part.

12.3.1.1 ArXiv Corpus

The arXiv corpus consists of 105,120 scientific articles in English. These articles were converted from LaTeX sources available at http://arxiv.org to HTML5+MathML using the LaTeXML system[3] and include the arXiv categories math, cs, physics:math-ph, stat, physics:hep-th, and physics:nlin to obtain a varied sample of technical documents containing mathematics.

[2] http://www.arxiv.org.

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



(a) Total number of nodes

(b) Maximum tree depth

(c) Average number of child nodes

(d) Total number of leaf nodes

Fig. 12.1 Math formulae statistics for the arXiv corpus

This subtask was designed for both formula-based search systems and document-based retrieval systems. In document-wise evaluation, human evaluators need to check all math formulae in the document. To reduce the cost of relevance judgment, we divided each document into paragraphs and used them as the search units ("documents") for the subtask. This produced 8,301,578 search units with roughly 60 million math formulae (including isolated symbols) encoded using LaTeX, Presentation MathML, and Content MathML Formulae[4]; 95% of the retrieval units had 23 or fewer math formulae, which is sufficiently small for document-based relevance judgment by human reviewers. Excerpts are stored independently in separate files, in both HTML5 and XHTML5 formats.

Figure 12.1 summarizes the basic statistics for the math formula trees in the ArXiv corpus. Figure 12.1a–d correspond to the distributions of the total number of nodes, maximum tree depth, average number of child nodes, and total number of leaf nodes in each math formula, respectively. These statistics show that the math trees in the arXiv corpus approximately follow the power-law distribution in their size. While there exists a vast amount of relatively simple trees, there also exists a non-negligible number of highly complex trees. This clearly shows that, as a benchmark for tree

[4]MathML (Ausbrooks et al. 2010) supplies two sub-languages: Presentation MathML encodes the visual (and possibly aural) appearance of the formulae in terms of a tree of layout primitives and Content MathML encodes the functional structure of formulae in terms of an operator tree.

structure search, the corpus is characterized by its large scale as well as the heterogeneity of the trees in it.

12.3.1.2 Wikipedia Corpus

The Wikipedia corpus contains 319,689 articles from English Wikipedia converted into a simpler XHTML format with images removed (5.15 GB uncompressed).[5] Unlike the arXiv corpus, articles were not split into smaller documents since they were simple/small enough for human annotation. Only 10% of the articles of the Wikipedia corpus contain explicit `<math>` tags that demarcate LaTeX, reflecting the small proportion of articles related to math in Wikipedia, while keeping the corpus size manageable for participants. All articles with a `<math>` tag were included in the corpus and the remaining 90% were sampled from articles that do not contain any `<math>` tag. These "text" articles act as distractors for keyword matching. There are over 590,000 formulae in the corpus with the same format as the arXiv corpus, i.e., encoded using LaTeX, Presentation MathML, and Content MathML. Note that untagged formulae frequently appear directly in HTML text (e.g. `'where x <sup> 2 ...'`). We made no attempt to detect or label these formulae embedded in the main text.

12.3.2 Topics

The Math Retrieval Subtasks were designed so that all topics include at least a single relevant document in the corpus, and ideally multiple relevant documents. In some cases, this is not possible, for example, with navigational queries where a specific document is sought after.

12.3.2.1 Topic Format

Details about the topic format are available in the documentation provided by the organizers (Kohlhase 2015). For participants, a math retrieval topic contains a (1) topic ID and (2) query (formula + keywords), but no textual description. The description is omitted to avoid participants biasing their system design toward the specific information needs identified in the topics. For evaluators, each topic also contains a narrative field that describes a user situation, the user's information needs, and relevance criteria. Formula queries are encoded in LaTeX, Presentation MathML, and Content MathML. In addition to the standard MathML notations, the following two subtask-specific extensions are adopted : *formulae query variables* and *formula simto regions* (see below).

[5] http://www.cs.rit.edu/~rlaz/NTCIR12_MathIR_WikiCorpus_v2.1.0.tar.bz2.

Formulae Query Variables (Wildcards). Formulae may contain query variables that act as wildcards, which can be matched to arbitrary subexpressions on candidate formulae. Query variables were represented using two different representations for the arXiv and Wikipedia topics. For the arXiv topics, query variables are named and indicated by a question mark (e.g., *?v*) while for the Wikipedia topics, wildcards are numbered and appear between asterisks (e.g., **1**).

This is an example query formula with the three query variables ?f, ?v, and ?d.

$$\frac{?f(?v + ?d) - ?f(?v)}{?d} \tag{12.1}$$

This query matches the argument of the limit on the right side of the equation below, substituting g for ?f, cx for ?v, and h for ?d. Note that each repetition of a query variable matches the same subexpression.

$$g'(cx) = \lim_{h \to 0} \frac{g(cx + h) - g(cx)}{h} \tag{12.2}$$

Formula Simto Regions. *Similarity regions* modify our formula query language, distinguishing subexpressions that should be identical to the query from those that are similar to the query in some sense. Consider the query formula below, which contains a *similarity region* called "a."

$$\frac{[a|\ g(cx + h) - g(cx)}{h} \tag{12.3}$$

The fraction operator and numerator h should match exactly, while the numerator may be replaced by a "similar" subexpression. Depending on the notion of similarity we choose to adopt, simto region "a" might match "$g(cx + h)+g(cx)$", if addition is similar to subtraction, or "$g(cx + h) - g(\mathbf{d}x)$", if c is somehow similar to d. The simto regions may also contain exact match constraints (see Kohlhase 2015).

12.3.2.2 ArXiv Topics

A total of 50 and 37 topics were provided during NTCIR-11 and NTCIR-12, respectively. Many of the topics in the arXiv subtask are sophisticated, for example, seeking to determine whether a connection exists between a factorial product and products starting with one. Some queries are simpler, such as looking for applications of operators, or loss functions used in machine learning. Eight out of the 37 topics during NTCIR-12 contained simto regions.

12.3.2.3 Wikipedia Topics

Topics for the Wikipedia subtask were designed with a less expert user population in mind. We imagined undergraduate and graduate students searching Wikipedia to locate or remember and relocate specific articles (i.e. navigational queries), browse math articles, learn/review mathematical concepts and notation they come across in their studies, find applications of concepts, or find information to help solve particular mathematical problems (e.g., for homework). A total of 30 topics were provided during NTCIR-12.

12.3.3 Relevance Judgment

The evaluation of the Math Retrieval Subtasks was pooling-based. First, all submitted results were converted into a `trec_eval` result file format. Next, for each topic, the top-20 ranked documents were selected from each run. Then, the set of pooled hits were evaluated by human assessors. After the pooling process, the selected retrieval units were fed into the SEPIA system[6] with MathML extensions developed by the organizers. Evaluators judged the relevance of each retrieval unit by comparing it to the query formulae and keywords, along with the described scenario provided with the topic, and selected one of the judgments `relevant` (R), `partially relevant` (PR), or `not-relevant` (N). The retrieval units were documents except for Wikipedia formula-only subtask, where the evaluation was based on individual formulae.

Evaluators had to rely on their mathematical intuition, the described information needs, and actual query to determine judgments. For the arXiv dataset, to ensure sufficient familiarity with mathematical documents, three evaluators were chosen from third-year and graduate students of (pure) mathematics. Each topic was evaluated by at least two evaluators. For the Wikipedia dataset, intended to represent mathematical information needs for nonexperts, ten students were recruited for evaluation: five undergraduates and five graduate (MSc) students. The Fleiss' κ values were 0.5615 and 0.5380 for the arXiv dataset and 0.3546 and 0.2619 for the Wikipedia dataset. Agreement between evaluators for the arXiv dataset was higher. This may be because of the greater mathematical expertise and shared background by these evaluators.

[6] https://code.google.com/p/sepia/.

12.4 Task Results and Discussion

12.4.1 Evaluation Metrics

In our evaluation, the judgment of each evaluator was converted into a relevance score using the mappings "Relevant" → 2, "Partially Relevant" → 1, and "Not Relevant" → 0. Then, the average score was binarized as follows:

- For "relevance" evaluation, the overall judgment is considered `relevant` if the average score is equal or greater than 1.5, and `not relevant` otherwise.
- For "partial relevance" evaluation, the overall judgment is considered `relevant` if the average score is equal or greater than 0.5, and `not relevant` otherwise.

Precision@k for $k = \{5, 10, 15, 20\}$ was used to evaluate participating systems. We chose these measures because they are simple to understand and characterize retrieval behavior as the number of hits increases. Precision@k values were obtained from `trec_eval` version 9.0, with which they were labeled `P_avgjg_5`, `P_avgjg_10`, `P_avgjg_15`, and `P_avgjg_20`, respectively.

12.4.2 MIR Systems

The numbers of participating teams were 6, 8, 6 for the NTCIR 10, 11, 12 Math Tasks. Three teams participated in all three tasks. For NTCIR 11 and 12, there were one or two new participating teams. The architectures of the participating systems were quite diverse. For formula encodings, all the LaTeX, MathML Presentation Markup, MathML Content Markup formats were used by at least one system; Presentation Markup was the most popular notation. Also, the majority of systems used a general-purpose search engine for indexing.

The following common technical decisions should be considered in designing MIR systems.

12.4.2.1 How to Index Math Formulae?

Mathematical formulae are expressed as XML tree structures, which often become very complex. However, the search sometimes requires approximate matching to guarantee certain flexibility. There are two strategies for indexing math formulae: token-based and subtree-based. While token-based indexing takes into account math tokens, the same as words in a text, subtree-based indexing decomposes the XML structure into smaller fragments, i.e.,subtrees, and treats them as indexing units. In the NTCIR Math Tasks, the majority of systems took into account structural information for formulae.

12.4.2.2 How to Deal with Query Variables?

One of the prominent features of MIR is that a query formula can contain "variables", i.e., symbols that can serve as named wildcards. Since the unification operation is expensive, most participating systems used a re-ranking step, wherein one or more initial rankings are merged and/or reordered. This approach of obtaining an initial candidate ranking followed by a refined ranking is a common and effective strategy. To locate strong partial matches, all the automated systems used unification, whether for variables (e.g., "$x^2 + y^2 = z^2$" unifies with "$a^2 + b^2 = c^2$"), constants, or entire subexpressions (e.g., via structural unification or indirectly through *generalized terms* with wildcards for operator arguments).

12.4.2.3 Other Technical Decisions

Other issues include how to identify the importance of the keywords/math formulae in queries and documents; exploit context information; normalize math formulae with possibly many notation variations; deal with ambiguity in the original LaTeX notation; combine keyword-based search with math formula search; and deal with "simto"-type queries. To summarize, there can be many options for MIR system design, and they should be balanced with computation cost.

12.5 Further Trials

The NTCIR Math Tasks also contain several important trials that lead to further exploration in succeeding research, as detailed below.

12.5.1 ArXiv Free-Form Query Search at NTCIR-10

The NTCIR-10 Math Pilot Task contained 19 *open* queries from mathematicians expressed as free descriptions with natural language text and formulae. Here is an example (NTCIR10-OMIR-19):

> *Let X_n be a decreasing sequence of nonempty closed sets in a Banach space such that their diameter tends to 0. Is their intersection nonempty?*

These topics were collected from questions asked by mathematicians in related forums, which makes the task settings more realistic and general. Since converting the textual descriptions into "keyword+formula" queries requires deep natural language comprehension, we did not pursue this direction further in this task. However, real queries in forums are an important resource for analyzing user information needs in their retrieval (Mansouri et al. 2019; Stathopoulos and Teufel 2015).

The Answer Retrieval for Questions on Math (ARQMath) is a newly launched task for the 11th Conference and Labs of the Evaluation Forum (CLEF 2020).[7] Data from Math Stack Exchange,[8] a mathematics-dedicated question answering forum, are expected to be used for ARQMath. Such explorations are expected to give further insights into realistic information needs.

12.5.2 Wikipedia Formula Search at NTCIR-11

The NTCIR-11 Math-2 Task provided the first open platform for comparing formula search engines, based upon their ability to retrieve specific formula in Wikipedia articles (Schubotz et al. 2015). By using formula-only queries that require an exact match of the math tree structure, the platform enables automatic evaluation without any human intervention. Regardless of the simplicity of the task, the automatic evaluation framework was useful in verifying and tuning the formula search function of math search engines. This will enable us to establish leaderboard-style comparison of different strategies for complicated large-scale formula searches.

12.5.3 Math Understanding Subtask at NTCIR-10

The goal of the Math Understanding Subtask was to extract natural language definitions of mathematical formulae in a document for their semantic interpretation. The dataset for this subtask contains 10 manually annotated articles used in a dry run and an additional 35 used in a formal run.

A description is obtained from a continuous text region or concatenation of some discontinuous text regions. Shorter descriptions may also be obtained from a longer one. For instance, in the text "$log(x)$ is a function that computes the natural logarithm of the value x", the complete description of "$log(x)$" is "a function that computes the natural logarithm of the value x". Moreover, the shorter descriptions "a function" and "a function that computes the natural logarithm" can be obtained from the previous one. This corpus defines two types of possible descriptions of mathematical expressions, namely full description (contains the complete type) and short description (contains the short type). Participants could extract any type of description in their submission.

The training and test set consists of 35 and 10 annotated papers selected from the arXiv copus, respectively. Inter-annotator agreement was tested for the five papers taken from the corpus. There are three measurements to test the reliability of annotation: F1-score, Cohen's kappa, and Krippendorff's alpha. To compute the F1-score, the position of the annotated descriptions from two annotators is strictly matched.

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

[8] https://math.stackexchange.com/.

The F1-score was 0.8670, Cohen's kappa was 0.8993, and Krippendorff's alpha was 0.7630 for full descriptions, and F1-score was 0.9014 for full and short descriptions). The evaluation was conducted by matching the position of the extracted descriptions against the positions of gold-standard descriptions, and precision, recall, and F1-score were used.

Math-description extraction is considered important to combine mathematical formulae with their textual descriptions for their interpretation. For example, Kristianto et al. (2017) combined the description extraction with formula dependency extraction and obtained consistent improvement in the Math Retrieval Subtasks in the succeeding NTCIR Math Tasks.

12.6 Further Impact of NTCIR Math Tasks

Several years after these NTCIR Math Tasks, we witnessed a number of valuable developments in mathematical content access studies. This section provides a brief introduction to some of these activities, although it is far less comprehensive.

12.6.1 Math Information Retrieval

Since these NTCIR Math Tasks, increasing attention has been paid to semantic retrieval of mathematical formulae. NLP techniques often play a critical role in bridging the gap between presentation and semantic representations of math formulae. Recent studies on this topic include variable typing (Stathopoulos et al. 2018), using the textual context for transformation from a presentation level to semantic level (Schubotz et al. 2018), and identifying declarations of mathematical objects (Lin et al. 2019).

Overall, there are several valuable approaches to MIR, including those we could not introduce in this book chapter. According to the number of citations on Semantic-Scholar,[9] the overview papers of the Math Tasks during NTCIR-10, 11, and 12 have 39, 39, 33 citations, respectively, as of December 2019. MIR is also characterized by the diversity of the conferences and journals of the related papers, including such fields as mathematics, information retrieval, image recognition, NLP, knowledge management, and document processing.

[9] https://www.semanticscholar.org.

12.6.2 Semantics Extraction in Mathematical Documents

Noteworthy recent work includes a general-purpose part-of-math tagger that performs semantic disambiguation and parsing of math formulae (Youssef 2017) and embeddings of math symbols (Mansouri et al. 2019; Youssef and Miller 2019). It has also been reported that image-based math-formula search is also capable of capturing semantic similarity without unification (Davila et al. 2019). Other related topics that were not addressed during the NTCIR Math Tasks include math document categorization (Barthel et al. 2013) using formulae information (Suzuki and Fujii 2017).

12.6.3 Corpora for Math Linguistics

The development work for the arXiv corpus (and the subsequent requests by the community) made it very clear that work on document understanding and information in Mathematics and STEM can only succeed based on large and shared document corpora. A single conversion run over the arXiv corpus (over 1.5 Million documents) is a multi-processor-year enterprise generating $10^8 - 10^9$ error reports in gigabytes of log files.

To support and manage this computational task, the corTEXsystem[10] has been developed as a general-purpose processing framework for corpora of scientific documents. The licensing issues involved in distributing the ensuing corpora have led to the recent establishment of *Special Interest group for Math Linguistics (SIGMathLing)*, [11] a forum and resource cooperative for the linguistics of mathematical and technical Documents. The problem is that many of the mathematical corpora (e.g., the arXiv corpus or the 3 Million abstracts of zbMATH[12]) are not available under a license that allows republishing. While the copyright owners are open towards research, they cannot afford to make the corpora public. SIGMathLing hosts such data sets in corpus cooperative: Researchers in mathematical semantics extraction and information retrieval sign a cooperative non-disclosure agreement, get access to the data sets and can deposit derived data sets in the cooperative. Data sets have dedicated landing pages so that they can be cited. A prime example of a data set is the XHTML5+MathML version of the arXiv corpus up to August 2019.[13]

[10] https://github.com/dginev/CorTeX.

[11] https://sigmathling.kwarc.info/.

[12] http://zbmath.org.

[13] The landing page is at https://sigmathling.kwarc.info/resources/arxmliv-dataset-082019/.

12.7 Conclusion

The NTCIR Math Tasks were an initial attempt in facilitating the formation of an interdisciplinary community of researchers interested in the challenging problems underlying MIR. The diversity of approaches reported at NTCIR shows that research in this field is active. We witnessed the progress of participating systems since the NTCIR-10 Pilot Task; improving scalability or addressing result ranking in new ways.

The design decision of the arXiv subask to exclusively concentrate on formula/keyword queries and use paragraphs as retrieval units made the retrieval task manageable but has also focused research away from questions such as result presentation and user interaction. In particular, few systems have invested in further semantics extraction from a corpus and used that in the search process to further address information needs. We feel that this direction should be further addressed in future tasks.

Ultimately, the success of MIR systems will be determined by how well they are able to accommodate user needs in terms of the adequacy of the query language, trade-off between query language expressiveness/flexibility, and answer latency on the one hand and learnability on the other. Similarly, the result ranking and monetization strategies for MIR are still a largely uncharted territory; we hope that future MIR tasks can help make progress on this front.

Acknowledgements The work reported in this chapter was partially supported by the Leibniz Association under grant SAW-2012-FIZ_KA-2, JSPS KAKENHI grant numbers 2430062 and 16H01756, and JST CREST number JPMJCR1513 and the National Science Foundation (USA) under grant no. HCC-1218801. We especially thank NTCIR Math co-organizers and collaborators, Iadh Ounis, Richard Zanibbi, Moritz Schubotz, and Goran Topić. We are also grateful to Kazuki Hayakawa and Takeshi Sagara for assisting with the task organization, Deyan Ginev who did most of the actual work in preparing the arXiv corpus for the NTCIR Math Tasks, Michal Růžička for generating XHTML files for the Wikipedia corpus, and Anurag Agarwal for assistance with designing the Wikipedia queries. Finally, we thank the students who evaluated the search hit pools at Jacobs University (arXiv) and Rochester Institute of Technology (Wikipedia).

References

Aizawa A, Kohlhase M, Ounis I (2013) NTCIR-10 Math pilot task overview. In: Kando N, Kato T (eds) Proceedings of the 10th NTCIR conference on evaluation of information access technologies, NTCIR-10, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 18–21 June 2013

Aizawa A, Kohlhase M, Ounis I, Schubotz M (2014) NTCIR-11 Math-2 task overview. In: Kando N, Joho H, Kishida K (eds) Proceedings of the 11th NTCIR conference on evaluation of information access technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 9–12 Dec 2014

Ausbrooks R, Buswell S, Carlisle D, Chavchanidze G, Dalmas S, Devitt S, Diaz A, Dooley S, Hunter R, Ion P, Kohlhase M, Lazrek A, Libbrecht P, Miller B, Miner R, Sargent M, Smith B, Soiffer N, Sutor R, Watt S (2010) Mathematical markup language (MathML) version 3.0. W3C Recommendation, World Wide Web Consortium (W3C)

Barthel S, Tönnies S, Balke WT (2013) Large-scale experiments for mathematical document classification. In: Urs SR, Na JC, Buchanan G (eds) Digital libraries: social media and community networks. Springer International Publishing, Cham, pp 83–92

Davila K, Joshi R, Setlur S, Govindaraju V, Zanibbi R (2019) Tangent-V: Math formula image search using line-of-sight graphs. In: Advances in information retrieval - 41st European conference on IR research, ECIR 2019, Cologne, Germany, Proceedings, Part I, pp 681–695. https://doi.org/10.1007/978-3-030-15712-8_44. Accessed 14–18 Apr 2019

Guidi F, Sacerdoti Coen C (2016) A survey on retrieval of mathematical knowledge. Math Comput Sci 10(4):409–427. https://doi.org/10.1007/s11786-016-0274-0

Kohlhase M (2015) Formats for topics and submissions for the Math-2 task at NTCIR-12. Technical report, NTCIR. http://ntcir-math.nii.ac.jp/wp-content/blogs.dir/13/files/2014/05/NTCIR11-Math-topics.pdf

Kristianto GY, Topić G, Aizawa A (2017) Utilizing dependency relationships between math expressions in math IR. Inf Retriev J 20:132–167. https://doi.org/10.1007/s10791-017-9296-8

Lin J, Wang X, Wang Z, Beyette D, Liu JC (2019) Prediction of mathematical expression declarations based on spatial, semantic, and syntactic analysis. In: Proceedings of the ACM symposium on document engineering 2019. ACM, New York, NY, USA, DocEng '19, pp 15:1–15:10. https://doi.org/10.1145/3342558.3345399

Liska M, Sojka P, Ruzcka M, Mravec P (2011) Web interface and collection for mathematical retrieval: WebMIaS and MREC. In: Sojka P (ed) Towards Digital Mathematics Library, DML workshop. Masaryk University, Brno

Mansouri B, Rohatgi S, Oard DW, Wu J, Giles CL, Zanibbi R (2019) Tangent-CFT: An embedding model for mathematical formulas. In: Proceedings of the 2019 ACM SIGIR international conference on theory of information retrieval, ICTIR 2019, Santa Clara, CA, USA, pp 11–18. https://doi.org/10.1145/3341981.3344235. Accessed 2–5 Oct 2019

Mansouri B, Zanibbi R, Oard DW (2019) Characterizing searches for mathematical concepts. In: 2019 ACM/IEEE joint conference on digital libraries (JCDL), pp 57–66. https://doi.org/10.1109/JCDL.2019.00019

Schubotz M, Youssef A, Markl V, Cohl HS (2015) Challenges of mathematical information retrieval in the NTCIR-11 Math Wikipedia Task. In: Proceedings of the 38th International ACM SIGIR conference on research and development in information retrieval. ACM, New York, NY, USA, SIGIR '15, pp 951–954. https://doi.org/10.1145/2766462.2767787

Schubotz M, Greiner-Petter A, Scharpf P, Meuschke N, Cohl HS, Gipp B (2018) Improving the representation and conversion of mathematical formulae by considering their textual context. In: Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, JCDL 2018, Fort Worth, TX, USA, pp 233–242. https://doi.org/10.1145/3197026.3197058. Accessed 03–07 June 2018

Stathopoulos Y, Teufel S (2015) Retrieval of research-level mathematical information needs: A test collection and technical terminology experiment. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (vol 2: Short Papers), Association for Computational Linguistics, Beijing, China, pp 334–340. https://doi.org/10.3115/v1/P15-2055

Stathopoulos Y, Baker S, Rei M, Teufel S (2018) Variable typing: Assigning meaning to variables in mathematical text. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, pp 303–312. https://doi.org/10.18653/v1/N18-1028

Suzuki T, Fujii A (2017) Mathematical document categorization with structure of mathematical expressions. In: 2017 ACM/IEEE joint conference on digital libraries (JCDL), pp 1–10. https://doi.org/10.1109/JCDL.2017.7991566

Youssef A (2017) Part-of-math tagging and applications. In: Geuvers H, England M, Hasan O, Rabe F, Teschke O (eds) Intelligent computer mathematics. Springer International Publishing, Cham, pp 356–374

Youssef A, Miller BR (2019) Explorations into the use of word embedding in math search and math semantics. In: Intelligent computer mathematics - 12th international conference, CICM 2019, Prague, Czech Republic, Proceedings, pp 291–305. https://doi.org/10.1007/978-3-030-23250-4_20. Accessed 8–12 July 2019

Zanibbi R, Blostein D (2012) Recognition and retrieval of mathematical expressions. Int J Doc Anal Recognit (IJDAR) 15(4):331–357. https://doi.org/10.1007/s10032-011-0174-4

Zanibbi R, Aizawa A, Kohlhase M, Ounis I, Topic G, Davila K (2016) NTCIR-12 MathIR task overview. In: Kando N, Sakai T, Sanderson M (eds) Proceedings of the 12th NTCIR conference on evaluation of information access technologies, National Center of Sciences, Tokyo, Japan, National Institute of Informatics (NII). Accessed 7–10 June 2016