Discovering Mathematical Objects of Interest—A Study of Mathematical Notations

André Greiner-Petter
University of Wuppertal
Germany
andre.greiner-petter@zbmath.org

Moritz Schubotz
FIZ-Karlsruhe and
University of Wuppertal
Germany
moritz.schubotz@fiz-karlsruhe.de

Fabian Müller
FIZ-Karlsruhe
Germany
fabian.mueller@fiz-karlsruhe.de

Corinna Breitinger
University of Wuppertal and
University of Konstanz
Germany
corinna.breitinger@uni-konstanz.de

Howard S. Cohl
Applied and Computational
Mathematics Division, National
Institute of Standards and Technology,
Mission Viejo, California, U.S.A.
howard.cohl@nist.gov

Akiko Aizawa
National Institute of Informatics
Japan
aizawa@nii.ac.jp

Bela Gipp
University of Wuppertal and
University of Konstanz
Germany
gipp@uni-wuppertal.de

ABSTRACT

Mathematical notation, i.e., the writing system used to communicate concepts in mathematics, encodes valuable information for a variety of information search and retrieval systems. Yet, mathematical notations remain mostly unutilized by today's systems. In this paper, we present the first in-depth study on the distributions of mathematical notation in two large scientific corpora: the open access arXiv (2.5B mathematical objects) and the mathematical reviewing service for pure and applied mathematics zbMATH (61M mathematical objects). Our study lays a foundation for future research projects on mathematical information retrieval for large scientific corpora. Further, we demonstrate the relevance of our results to a variety of use-cases. For example, to assist semantic extraction systems, to improve scientific search engines, and to facilitate specialized math recommendation systems.

The contributions of our presented research are as follows: (1) we present the first distributional analysis of mathematical formulae on arXiv and zbMATH; (2) we retrieve relevant mathematical objects for given textual search queries (e.g., linking $P_n^{(\alpha,\beta)}(x)$ with 'Jacobi polynomial'); (3) we extend zbMATH's search engine by providing relevant mathematical formulae; and (4) we exemplify the applicability of the results by presenting auto-completion for math inputs as the first contribution to math recommendation systems. To expedite future research projects, we have made available our source code and data.

CCS CONCEPTS

• **Information systems** → **Mathematics retrieval;** • **Information systems** → **Novelty in information retrieval;** *Information extraction*; • **Information systems~Recommender systems;** • **Information systems~Near-duplicate and plagiarism detection**;

KEYWORDS

Mathematical Objects of Interest, Mathematical Information Retrieval, Distributions of Mathematical Objects, Term Frequency-Inverse Document Frequency, Mathematical Search Engine

1 INTRODUCTION

Taking into account mathematical notation in the literature leads to a better understanding of scientific literature on the Web and allows one to make use of semantic information in specialized Information Retrieval (IR) systems. Nowadays applications in Math Information Retrieval (MathIR) [15], such as search engines [7, 17, 18, 22, 24, 27, 31], semantic extraction systems [23, 36, 37], recent efforts in math embeddings [10, 13, 25, 44], and semantic tagging of math formulae [6, 43] either consider an entire equation as one entity or only focus on single symbols. Since math expressions often contain meaningful and important subexpressions, these applications could benefit from an approach that lies between the extremes of examining only individual symbols or considering an entire equation as one entity. Consider for example, the explicit definition for Jacobi polynomials [8, (18.5.7)]

$$P_n^{(\alpha,\beta)}(x) = \frac{\Gamma(\alpha+n+1)}{n!\,\Gamma(\alpha+\beta+n+1)} \sum_{m=0}^{n} \binom{n}{m} \frac{\Gamma(\alpha+\beta+n+m+1)}{\Gamma(\alpha+m+1)} \left(\frac{x-1}{2}\right)^m. \quad (1)$$

The *interesting* components in this equation are $P_n^{(\alpha,\beta)}(x)$ on the left-hand side, and the appearance of the gamma function $\Gamma(s)$ on the right-hand side, implying a direct relationship between Jacobi polynomials and the gamma function. Considering the entire expression as a single object misses this important relationship. On the other hand, focusing on single symbols can result in the misleading interpretation of Γ as a variable and $\Gamma(\alpha + n + 1)$ as a multiplication between Γ and $(\alpha + n + 1)$. A system capable of identifying the important components, such as $P_n^{(\alpha,\beta)}(x)$ or $\Gamma(\alpha + n + 1)$, is therefore desirable. Hereafter, we define these components as *Mathematical Objects of Interest* (MOIs) [13].

The *importance* of math objects is a somewhat imprecise description and thus difficult to measure. Currently, not much effort has been made in identifying meaningful subexpressions. Kristianto et al. [23] introduced dependency graphs between formulae. With this approach, they were able to build dependency graphs of mathematical expressions, but only if the expressions appeared as single expressions in the context. For example, if $\Gamma(\alpha + n + 1)$ appears as a stand-alone expression in the context, the algorithm will declare a dependency with Equation (1). However, it is more likely that different forms, such as $\Gamma(s)$, appear in the context. Since this expression does not match any subexpression in Equation (1), the approach cannot establish a connection with $\Gamma(s)$. Kohlhase et al. studied in [19–21] another approach to identify essential components in formulae. They performed eye-tracking studies to identify important areas in rendered mathematical formulae. While this is an interesting approach that allows one to learn more about the insights of human behaviors of reading and understanding math, it is inaccessible for extensive studies.

This paper presents the first extensive frequency distribution study of mathematical equations in two large scientific corpora, the e-Print archive arXiv.org (hereafter referred to as arXiv[1]) and the international reviewing service for pure and applied mathematics zbMATH[2]. We will show that math expressions, similar to words in natural language corpora, also obey Zipf's law [33], and therefore follows a *Zipfian* distribution. Related research projects observed a relation to Zipf's law for single math symbols [6, 36]. In the context of quantitative linguistics, Zipf's law states that given a text corpus, the frequency of any word is inversely proportional to its rank in the frequency table. Motivated by the similarity to linguistic properties, we will present a novel approach for ranking formulae by their relevance via a customized version of the ranking function BM25 [34]. We will present results that can be easily embedded in other systems in order to distinguish between common and uncommon notations within formulae. Our results lay a foundation for future research projects in MathIR.

Fundamental knowledge on frequency distributions of math formulae is beneficial for numerous applications in MathIR, ranging from educational purposes [40] to math recommendation systems, search engines [7, 31], and even automatic plagiarism detection systems [28, 29, 39]. For example, students can search for the conventions to write certain quantities in formulae; document preparation systems can integrate an auto-completion or auto-correction service for math inputs; search or recommendation engines can adjust their ranking scores with respect to standard notations; and plagiarism detection systems can estimate whether two identical formulae indicate potential plagiarism or are just using the conventional notations in a particular subject area. To exemplify the applicability of our findings, we present a textual search approach to retrieve mathematical formulae. Further, we will extend zbMATH's faceted search by providing facets of mathematical formulae according to a given textual search query. Lastly, we present a simple auto-completion system for math inputs as a contribution towards advancing mathematical recommendation systems. Further, we show that the results provide useful insights for plagiarism detection algorithms. We provide access to the source code, the results, and extended versions of all of the figures appearing in this paper at https://github.com/ag-gipp/FormulaCloudData.

Related Work: Today, mathematical search engines index formulae in a database. Much effort has been undertaken to make this process as efficient as possible in terms of precision and runtime performance [7, 17, 26, 27, 45]. The generated databases naturally contain the information required to examine the distributions of the indexed mathematical formulae. Yet, no in-depth studies of these distributions have been undertaken. Instead, math search engines focus on other aspects, such as devising novel similarity measures and improving runtime efficiency. This is because the goal of math search engines is to retrieve relevant (i.e., similar) formulae which correspond to a given search query that partially [24, 26, 31] or exclusively [7, 17, 18] contains formulae. However, for a fundamental study of distributions of mathematical expressions, no similarity measures nor efficient lookup or indexing is required. Thus, we use the general-purpose query language XQuery and employ the BaseX[3] implementation. BaseX is a free open-source XML database engine, which is fully compatible with the latest XQuery standard [14, 41]. Since our implementations rely on XQuery, we are able to switch to any other database which allows for processing via XQuery.

2 DATA PREPARATION

LATEX is the de facto standard for the preparation of academic manuscripts in the fields of mathematics and physics [11]. Since LATEX allows for advanced customizations and even computations, it is challenging to process. For this reason, LATEX expressions are unsuitable for an extensive distribution analysis of mathematical notations. For mathematical expressions on the web, the XML formatted MATHML[4] is the current standard, as specified by the World Wide Web Consortium (W3C). The tree structure and the fixed standard, i.e., MATHML tags, cannot be changed, thus making this data format reliable. Several available tools are able to convert from LATEX to MATHML [35] and various databases are able to index XML data. Thus, for this study, we have chosen to focus on MATHML. In the following, we investigate the databases arXMLiv (08/2018) [12] and zbMATH[5] [38].

The arXMLiv dataset ($\approx$1.2 million documents) contains HTML5 versions of the documents from the e-Print archive arXiv.org. The HTML5 documents were generated from the TEX sources

---

via LᴬTᴇxᴍʟ [30]. LᴬTᴇxᴍʟ converted all mathematical expressions into MᴀᴛʜML with parallel markup, i.e., presentation and content MᴀᴛʜML. In this study we only consider the subsets *no-problem* and *warning*, which generated no errors during the conversion process. Nonetheless, the MᴀᴛʜML data generated still contains some errors or falsely annotated math. For example, we discovered several instances of affiliation and footnotes, SVG[6] and other unknown tags, encoded in MᴀᴛʜML. Regarding the footnotes, we presumed that authors falsely used mathematical environments for generating footnote or affiliation marks. We used the TᴇX string, provided as an attribute in the MᴀᴛʜML data, to filter out expressions that match the string '{}^{*}', where '*' indicates any possible expression. In addition, we filtered out SVG and other unknown tags. We assume that these expressions were generated by mistake due to limitations of LᴬTᴇxᴍʟ. The final arXiv dataset consisted of 841,008 documents which contained at least one mathematical formula. The dataset contained a total of 294,151,288 mathematical expressions.

In addition to arXiv, we investigated zbMATH, an international reviewing service for pure and applied mathematics which contains abstracts and reviews of articles, hereafter uniformly called abstracts, mainly from the domains of pure and applied mathematics. The abstracts in zbMATH are formatted in TᴇX [38]. To be able to compare arXiv and zbMATH, we manually generated MᴀᴛʜML via LᴬTᴇxᴍʟ for each mathematical formula in zbMATH and performed the same filters as used for the arXiv documents. The zbMATH dataset contained 2,813,451 abstracts, of which 1,349,297 contained at least one formula. In total, the dataset contained 11,747,860 formulae. Even though the total number of formulae is smaller compared to arXiv, we hypothesize that math formulae in abstracts are particularly meaningful.

2.1 Data Wrangling

Since we focused on the frequency distributions of visual expressions, we only considered presentational MᴀᴛʜML (pMML). Rather than normalizing the pMML data, e.g., via MathMLCan [9], which would also change the tree structure and visual core elements in pMML, we only eliminated the attributes. These attributes are used for minor visual changes, e.g., stretched parentheses or inline limits of sums and integrals. Thus, for this first study, we preserved the core structure of the pMML data, which might provide insightful statistics for the MᴀᴛʜML community to further cultivate the standard. After extracting all MᴀᴛʜML expressions, filtering out falsely annotated math and SVG tags, and eliminating unnecessary attributes and annotations, the datasets required 83GB of disk space for arXiv and 6GB for zbMATH, respectively.

Listing 1: MathML representation of $P_n^{(\alpha,\beta)}(x)$.

```
1  <math><mrow>
2   <msubsup>
3    <mi>P</mi>
4    <mi>n</mi>
5    <mrow>
6     <mo>(</mo>
7     <mi>α</mi>
8     <mo>,</mo>
9     <mi>β</mi>
10    <mo>)</mo>
11    <mo> </mo>
12   </mrow>
13   </msubsup>
14   <mo> </mo>
15   <mrow>
16    <mo>(</mo>
17    <mi>x</mi>
18    <mo>)</mo>
19   </mrow>
20  </mrow></math>
```

In the following, we indexed the data via BaseX. The indexed datasets required a disk space of 143.9GB in total (140GB for arXiv and 3.9GB for zbMATH). Due to the limitations[7] of databases in BaseX, it was necessary to split our datasets into smaller subsets. We split the datasets according to the 20 major article categories of arXiv[8] and classifications of zbMATH. To increase performance, we use BaseX in a server-client environment. We experienced performance issues in BaseX when multiple clients repeatedly requested data from the same server in short intervals. We determined that the best workaround for this issue was to launch BaseX servers for each database, i.e., each category/classification.

Mathematical expressions often consist of multiple meaningful subexpressions, which we defined as MOIs. However, without further investigation of the context, it is impossible to determine meaningful subexpressions. As a consequence, every equation is a potential MOI on its own and potentially consists of multiple other MOIs. For an extensive frequency distributional analysis, we aim to discover all possible mathematical objects. Hence, we split every formula into its components. Since MᴀᴛʜML is an XML data format (essentially a tree-structured format), we define subexpressions of equations as subtrees of its MᴀᴛʜML format.

Listing 1 illustrates a Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$ in pMML. The <mo> element on line 14 contains the *invisible times* UTF-8 character. By definition, the <math> element is the root element of MᴀᴛʜML expressions. Since we cut off all other elements besides pMML nodes, each <math> element has one and only one child element[9]. Thus, we define the child element of the <math> element as the root of the expression. Starting from this root element, we explore all subexpressions. For this study, we presume that every meaningful mathematical object (i.e., MOI) must contain at least one identifier.

Hence, we only study subtrees which contain at least one <mi> node. Identifiers, in the sense of MᴀᴛʜML, are '*symbolic names or arbitrary text*' [10], e.g., single Latin or Greek letters. Identifiers do not contain special characters (other than Greek letters) or numbers. As a consequence, arithmetic expressions, such as $(1+2)^2$, or sequences of special characters and numbers, such as $\{1, 2, \dots\} \cap \{-1\}$, will not appear in our distributional analysis. However, if a sequence or arithmetic expression consists of an identifier somewhere in the pMML tree (such as in $\{1, 2, \dots\} \cap A$), the entire expression will be recognized. The Jacobi polynomial $P_n^{(\alpha,\beta)}(x)$, therefore, consists of the following subexpressions: $P_n^{(\alpha,\beta)}$, (α,β), (x), and the single identifiers P, n, α, β, and x. The entire expression is also a mathematical object. Hence, we take entire expressions with an identifier into account for our analysis. In the following, the set of subexpressions will be understood to include the expression itself.

For our experiments, we also generated a string representation of the MᴀᴛʜML data. The string is generated recursively by applying one of two rules for each node: (i) if the current node is a leaf, the node-tag and the content

---

[6]Scalable Vector Graphics

[7]A detailed overview of the limitations of BaseX databases can be found at http://docs.basex.org/wiki/Statistics [Accessed: Sep. 1, 2019].
[8]The arXiv categories *astro-ph* (astro physics), *cond-mat* (condensed matter), and *math* (mathematics) were still too large for a single database. Thus, we split those categories into two equally sized parts.
[9]Sequences are always nested in an <mrow> element.
[10]https://www.w3.org/TR/MathML3/chapter3.html [Accessed: Sep. 1, 2019]

will be merged by a colon, e.g., `<mi>x</mi>` will be converted to `mi:x`; (ii) otherwise the node-tag wraps parentheses around its content and separates the children by a comma, e.g., `<mrow><mo>(</mo><mi>x</mi><mo>)</mo></mrow>` will be converted to `mrow(mo:(,mi:x,mo:))`. Furthermore, the special UTF-8 characters for invisible times (U+2062) and function application (U+2061) are replaced by `ivt` and `fa`, respectively. For example, the gamma function with argument $x+1$, $\Gamma(x+1)$ would be represented by

$$mrow(mi{:}\Gamma,mo{:}ivt,mrow(mo{:}(,mrow(mi{:}x,mo{:}+,mn{:}1),mo{:})))). \quad (2)$$

Between Γ and $(x + 1)$, there would most likely be the special character for *invisible times* rather than for *function application*, because LaTeXML is not able to parse Γ as a function. Note that this string conversion is a bijective mapping. The string representation reduces the verbose XML format to a more concise presentation. Thus, an equivalence check between two expressions is more efficient.

2.2 Complexity of Math

Mathematical expressions can become complex and lengthy. The tree structure of MathML allows us to introduce a measure that reflects the complexity of mathematical expressions. More complex expressions usually consist of more extensively nested subtrees in the MathML data. Thus, we define the complexity of a mathematical expression by the maximum depth of the MathML tree. In XML the content of a node and its attributes are commonly interpreted as children of the node. Thus, we define the depth of a single node as 1 rather than 0, i.e., single identifiers, such as `<mi>P</mi>`, have a complexity of 1. The Jacobi polynomial from Listing 1 has a complexity of 4.

We perform the extraction of subexpressions from MathML in BaseX. The algorithm for the extraction process is written in XQuery. The algorithm traverses recursively downwards from the root to the leaves. In each iteration, it checks whether there is an identifier, i.e., `<mi>` element, among the descendants of the current node. If there is no such element, the subtree will be ignored. It seems counterintuitive to start from the root and check if an identifier is among the descendants rather than starting at each identifier and traversing upwards to the root. If an XQuery requests a node in BaseX, BaseX loads the entire subtree of the requested node into the cache (up to a specified size). If the algorithm traverses upwards through the MathML tree, the XQuery will trigger database requests in every iteration. Hence, the downwards implementation performs better, since there is only one database request for every expression rather than for every subexpression.

Since we only minimize the pMML data rather than normalizing it, two identically rendered expressions may have different complexities. For instance, `<mrow><mi>x</mi></mrow>` consists of two distinct subexpressions, but both of them are displayed the same. Another problem often appears for arrays or similar visually complicated structures. The extracted expressions are not necessarily logical subexpressions. We will consider applying more advanced embedding techniques such as special tokenizers [26], symbol layout trees [7, 45], and a MathML normalization via MathMLCan [9] in future research to overcome these issues.
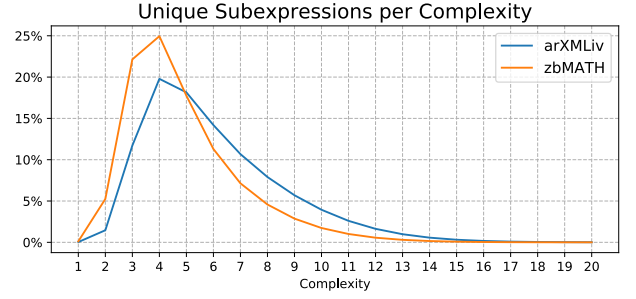

Figure 1: Unique subexpressions for each complexity in arXiv and zbMATH.

3 FREQUENCY DISTRIBUTIONS OF MATHEMATICAL FORMULAE

By splitting each formula into subexpressions, we generated longer documents and a bias towards low complexities. Note that, hereafter, we only refer to the mathematical content of documents. Thus, the length of a document refers to the number of math formulae—here the number of subexpressions—in the document. After splitting expressions into subexpressions, arXiv consists of 2.5B and zbMATH of 61M expressions, which raised the average document length to 2,982.87 for arXiv and 45.47 for zbMATH, respectively.

For calculating frequency distributions, we merged two subexpressions if their string representations were identical. Remember, the string representation is unique for each MathML tree. After merging, arXiv consisted of 350,206,974 unique mathematical subexpressions with a maximum complexity of 218 and an average complexity of 5.01. For high complexities over 70, the formulae show some erroneous structures that might be generated from LaTeXML by mistake. For example, the expression with the highest complexity is a long sequence of a polynomial starting with '$P_4(t_1, t_3, t_7, t_{11}) =$' followed by 690 summands. The complexity is caused by a high number of unnecessarily deeply nested `<mrow>` nodes. The highest complexity with a minimum document frequency of two is 39, which is a continued fraction. Since continued fractions are nested fractions, they naturally have a large complexity. One of the most complex expressions (complexity 20) with a minimum document frequency of three was the formula

$$\left(\sum_{j_1=1}^{n} \left(\sum_{j_2=1}^{n} \left(\cdots \left(\sum_{j_m=1}^{n} \left| T\left(e_{j_1}, \ldots, e_{j_m}\right) \right|^{q_m} \right)^{\frac{q_{m-1}}{q_m}} \cdots \right)^{\frac{q_2}{q_3}} \right)^{\frac{q_1}{q_2}} \right)^{\frac{1}{q_1}} \leq C_{m,p,\mathbf{q}}^{\mathbb{K}} \|T\|. $$

$$(3)$$

In contrast, zbMATH only consisted of 8,450,496 unique expressions with a maximum complexity of 26 and an average complexity of 3.89. One of the most complex expressions in zbMATH with a minimum document frequency of three was

$$M_p(r, f) = \left(\frac{1}{2\pi} \int_0^{2\pi} \left| f\left(re^{i\theta}\right) \right|^p d\theta \right)^{1/p}. \quad (4)$$

As we expected, reviews and abstracts in zbMATH were generally shorter and consisted of less complex mathematical formulae. The

Category	arXiv	zbMATH
Documents	841,008	1,349,297
Formulae	294,151,288	11,747,860
Subexpressions	2,508,620,512	61,355,307
Unique Subexpressions	350,206,974	8,450,496
Average Document Length	2,982.87	45.47
Average Complexity	5.01	3.89
Maximum Complexity	218	26

Table 1: Dataset overview. Average Document Length is defined as the average number of subexpressions per document.

dataset also appeared to contain fewer erroneous expressions, since expressions of complexity 25 are still readable and meaningful.

Figure 1 shows the ratio of unique subexpressions for each complexity in both datasets. The figure illustrates that both datasets share a peak at complexity four. Compared to zbMATH, the arXiv expressions are slightly more evenly distributed over the different levels of complexities. Interestingly, complexities one and two are not dominant in either of the two datasets. Single identifiers only make up 0.03% in arXiv and 0.12% in zbMATH, which is comparable to expressions of complexity 19 and 14, respectively. This finding illustrates the problem of capturing semantic meanings for single identifiers rather than for more complex expressions [37]. It also substantiates that entire expressions, if too complex, are not suitable either for capturing the semantic meanings [23]. Instead, a middle ground is desirable, since the most unique expressions in both datasets have a complexity between 3 and 5. Table 1 summarizes the statistics of the examined datasets.

3.1 Zipf's Law

In linguistics, it is well known that word distributions follow Zipf's Law [33], i.e., the r-th most frequent word has a frequency that scales to

$$f(r) \propto \frac{1}{r^\alpha} \qquad (5)$$

with $\alpha \approx 1$. A better approximation can be applied by a shifted distribution

$$f(r) \propto \frac{1}{(r + \beta)^\alpha}, \qquad (6)$$

where $\alpha \approx 1$ and $\beta \approx 2.7$. In a study on Zipf's law, Piantadosi [33] illustrated that not only words in natural language corpora follow this law surprisingly accurately, but also many other human-created sets. For instance, in programming languages, in biological systems, and even in music. Since mathematical communication has derived as the result of centuries of research, it would not be surprising if mathematical notations would also follow Zipf's law. The primary conclusion of the law illustrates that there are some very common tokens against a large number of symbols which are not used frequently. Based on this assumption, we can postulate that a score based on frequencies might be able to measure the peculiarity of a token. The infamous TF-IDF ranking functions and their derivatives [1, 34] have performed well in linguistics for many years and are still widely used in retrieval systems [3]. However, since we split every expression into its subexpressions, we generated an anomalous bias towards shorter, i.e., less complex, formulae. Hence, distributions of subexpressions may not obey Zipf's law.

(a) Frequency Distributions   (b) Complexity Distributions
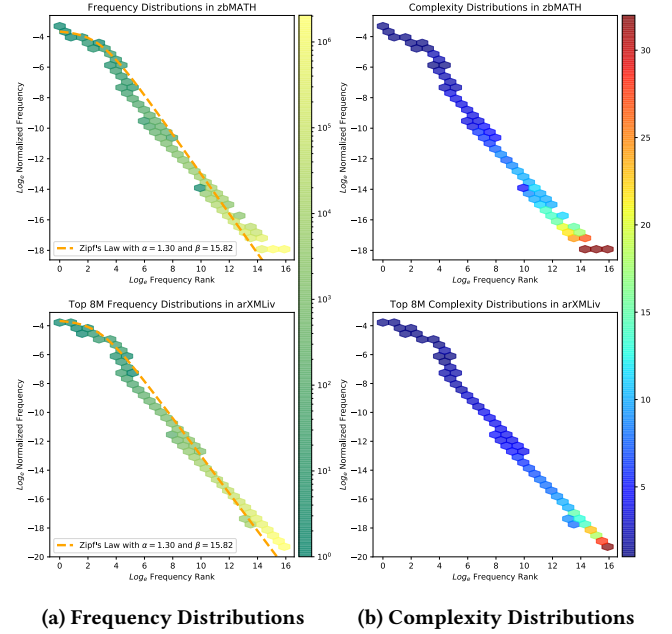
Figure 2: Each figure illustrates the relationship between the frequency ranks (x-axis) and the normalized frequency (y-axis) in zbMATH (top) and arXiv (bottom). For arXiv, only the first 8 million entries are plotted to be comparable with zbMATH (≈ 8.5 million entries). Subfigure (a) shades the hexagonal bins from green to yellow using a logarithmic scale according to the number of math expressions that fall into a bin. The dashed orange line represents Zipf's distribution (6). The values for α and β are provided in the plots. Subfigure (b) shades the bins from blue to red according to the maximum complexity in each bin.

Figure 2 visualizes a comparison between Zipf's law and the frequency distributions of mathematical subexpressions in arXiv and zbMATH. The dashed orange line visualizes the power law (6). The plots demonstrate that the distributions in both datasets obey this power law. Interestingly, there is not much difference in the distributions between both datasets. Both distributions seem to follow the same power law, with $\alpha = 1.3$ and $\beta = 15.82$. Moreover, we can observe that the developed complexity measure seems to be appropriate, since the complexity distributions for formulae are similar to the distributions for the length of words [33]. In other words, more complex formulae, as well as long words in natural languages, are generally more specialized and thus appear less frequent throughout the corpus. Note that colors of the bins for complexities fluctuate for rare expressions because the color represents the maximum rather than the average complexity in each bin.

3.2 Analyzing and Comparing Frequencies

Figure 3 shows in detail the most frequently used mathematical expressions in arXiv for the complexities 1 to 5. The orange dashed line visible in all graphs represents the normal Zipf's law distribution from Equation (5). We explore the total frequency values without
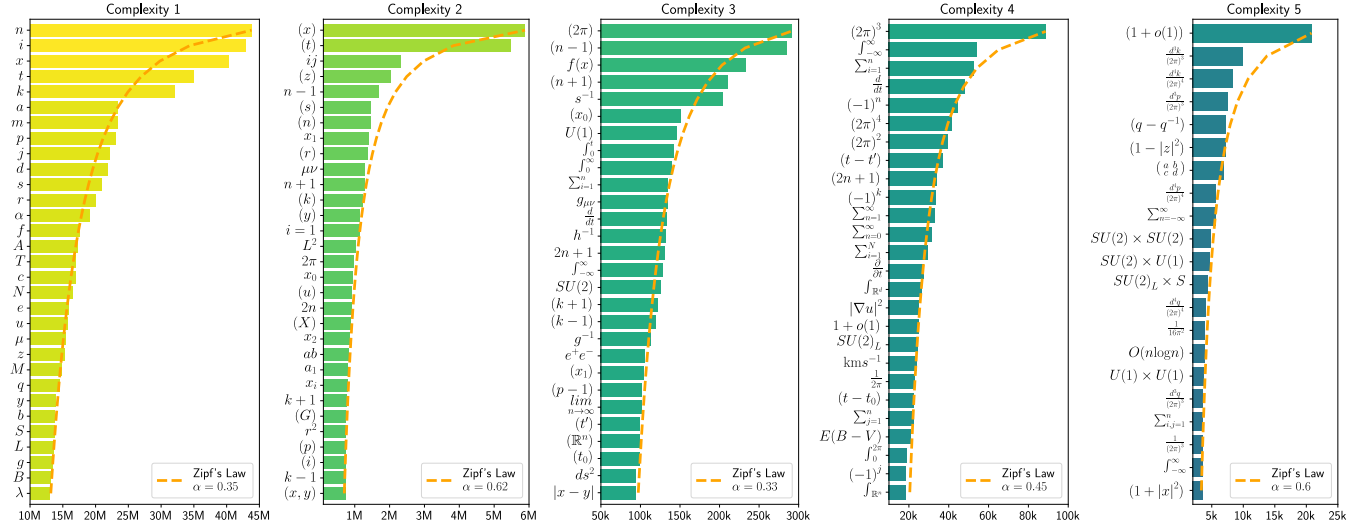
Figure 3: Overview of the most frequent mathematical expressions in arXiv for complexities 1-5. The color gradient from yellow to blue represents the frequency in the dataset. Zipf's law (5) is represented by a dashed orange line.

any normalization. Thus, Equation (5) was multiplied by the highest frequency for each complexity level to fit the distribution. The plots in Figure 3 demonstrate that even though the parameter α varies between 0.35 and 0.62, the distributions in each complexity class also obey Zipf's law.

The plots for each complexity class contain some interesting fluctuations. We can spot a set of five single identifiers that are most frequently used throughout arXiv: n, i, x, t, and k. Even though the distributions follow Zipf's law accurately, we can explore that these five identifiers are proportionally more frequently used than other identifiers and clearly separate themselves above the rest (notice the large gap from k to a). All of the five identifiers are known to be used in a large variety of scenarios. Surprisingly, one might expect that common pairs of identifiers would share comparable frequencies in the plots. However, typical pairs, such as x and y, or α and β, possess a large discrepancy.

The plot of complexity two also reveals that two expressions are proportionally more often used than others: (x) and (t). These two expressions appear more than three times as often in the corpus than any other expression of the same complexity. On the other hand, the quantitative difference between (x) and (t) is negligible. We may assume that arXiv's primary domain, physics, causes the quantitative disparity between (x), (t), and the other tokens. The primary domain of the dataset becomes more clearly visible for higher complexities, such as $SU(2)$ (C3[11]) or kms^{-1} (C4).

Another surprising property of arXiv is that symmetry groups, such as $SU(2)$, appear to play an essential role in the majority of articles on arXiv, see $SU(2)$ (C3), $SU(2)_L$ (C4), and $SU(2) \times SU(2)$ (C5), among others. The plots of higher complexities[12], which we do not show here, made this even more noticeable. Given a

complexity of six, for example, the most frequently used expression was $SU(2)_L \times SU(2)_R$, and for a complexity of seven it was $SU(3) \times SU(2) \times U(1)$. Given a complexity of eight, ten out of the top-12 expressions were from symmetry group calculations.

It is also worthwhile to compare expressions among different levels of complexities. For instance, (x) and (t) appeared almost six million times in the corpus, but $f(x)$ (at position three in C3) was the only expression which contained one of these most common expressions. Note that subexpressions of variations, such as (x_0), (t_0), or $(t - t')$, do not match the expression of complexity two. This may imply that (x), and especially (t), appear in many different scenarios. Further, we can examine that even though (x) is a part of $f(x)$ in only approximately 3% of all cases, it is still the most likely combination. These results are especially useful for recommendation systems that make use of math as input. Moreover, plagiarism detection systems may also benefit from such a knowledge base. For instance, it might be evident that $f(x)$ is a very common expression, but for automatic systems that work on a large scale, it is not clear whether duplicate occurrences of $f(x)$ or $\Xi(x)$ should be scored differently, e.g., in the case of plagiarism detection.

Figure 3 shows only the most frequently occurring expressions in arXiv. Since we already explored a bias towards physics formulae in arXiv, it is worth comparing the expressions present within both datasets. Figure 4 compares the 25-top expressions for the complexities one to four. In zbMATH, we discovered that computer science and graph theory appeared as popular topics, see for example $G = (V, E)$ (in C3 at position 20) and the Bachmann-Landau notations in $O(\log n)$, $O(n^2)$, and $O(n^3)$ (C4 positions 2, 3, and 19).

From Figure 4, we can also deduce useful information for MathIR tasks which focus on semantic information. Current semantic extraction tools [37] or LATEX parsers [35] still have difficulties distinguishing *multiplications* from *function calls*. For example as mentioned before, LATEXML [30] adds an *invisible times* character

---

[11]We refer to a given complexity n with Cn, i.e., C3 refers to complexity 3.
[12]More plots showing higher complexities are available at https://github.com/ag-gipp/FormulaCloudData
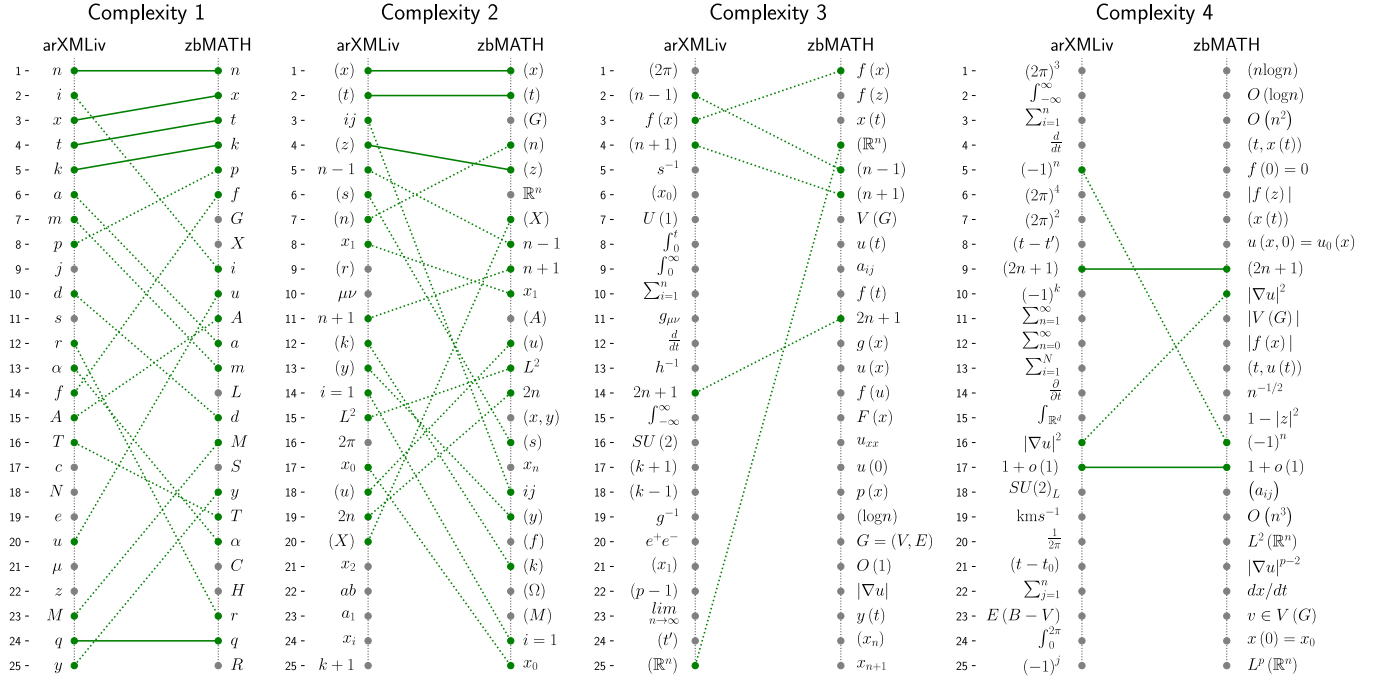
Figure 4: The top-25 most frequent expressions in arXiv (left) and zbMATH (right) for complexities 1-4. A line between both sets indicates a matching set. Bold lines indicate that the matches share a similar rank (distance of 0 or 1).

	Complexity 1		Complexity 2		Complexity 3		Complexity 4	
	arXMLiv	zbMATH	arXMLiv	zbMATH	arXMLiv	zbMATH	arXMLiv	zbMATH
1	n	n	(x)	(x)	(2π)	$f(x)$	$(2\pi)^3$	$(n\log n)$
2	i	x	(t)	(t)	$(n-1)$	$f(z)$	$\int_{-\infty}^\infty$	$O(\log n)$
3	x	t	ij	(G)	$f(x)$	$x(t)$	$\sum_{i=1}^n$	$O(n^2)$
4	t	k	(z)	(n)	$(n+1)$	$(\mathbb{R}^n)$	$\frac{d}{dt}$	$(t, x(t))$
5	k	p	$n-1$	(z)	s^{-1}	$(n-1)$	$(-1)^n$	$f(0)=0$
6	a	f	(s)	$\mathbb{R}^n$	(x_0)	$(n+1)$	$(2\pi)^4$	$|f(z)|$
7	m	G	(n)	(X)	$U(1)$	$V(G)$	$(2\pi)^2$	$(x(t))$
8	p	X	x_1	$n-1$	$\int_0^t$	$u(t)$	$(t-t')$	$u(x,0)=u_0(x)$
9	j	i	(r)	$n+1$	$\int_0^\infty$	a_{ij}	$(2n+1)$	$(2n+1)$
10	d	u	$\mu\nu$	x_1	$\sum_{i=1}^n$	$f(t)$	$(-1)^k$	$|\nabla u|^2$
11	s	A	$n+1$	(A)	$g_{\mu\nu}$	$2n+1$	$\sum_{n=1}^\infty$	$|V(G)|$
12	r	a	(k)	(u)	$\frac{d}{dt}$	$g(x)$	$\sum_{n=0}^\infty$	$|f(x)|$
13	α	m	(y)	L^2	h^{-1}	$u(x)$	$\sum_{i=1}^N$	$(t, u(t))$
14	f	L	$i=1$	$2n$	$2n+1$	$f(u)$	$\frac{\partial}{\partial t}$	$n^{-1/2}$
15	A	d	L^2	(x,y)	$\int_{-\infty}^\infty$	$F(x)$	$\int_{\mathbb{R}^d}$	$1-|z|^2$
16	T	M	2π	(s)	$SU(2)$	u_{xx}	$|\nabla u|^2$	$(-1)^n$
17	c	S	x_0	x_n	$(k+1)$	$u(0)$	$1+o(1)$	$1+o(1)$
18	N	y	(u)	ij	$(k-1)$	$p(x)$	$SU(2)_L$	(a_{ij})
19	e	T	$2n$	(y)	g^{-1}	$(\log n)$	$\mathrm{km}s^{-1}$	$O(n^3)$
20	u	α	(X)	(f)	e^+e^-	$G=(V,E)$	$\frac{1}{2\pi}$	$L^2(\mathbb{R}^n)$
21	μ	C	x_2	(k)	$O(1)$	$O(1)$	$(t-t_0)$	$|\nabla u|^{p-2}$
22	z	H	ab	(Ω)	$(p-1)$	$|\nabla u|$	$\sum_{j=1}^n$	dx/dt
23	M	r	a_1	(M)	$\lim_{n\to\infty}$	$y(t)$	$E(B-V)$	$v \in V(G)$
24	q	q	x_i	$i=1$	(t')	(x_n)	$\int_0^{2\pi}$	$x(0)=x_0$
25	y	R	$k+1$	x_0	$(\mathbb{R}^n)$	x_{n+1}	$(-1)^j$	$L^p(\mathbb{R}^n)$

| C3		C4		C5		C6		C7	
114.84	$(n!)$	129.44	$i,j=1,\dots,n$	119.21	$\mathrm{Gal}\!\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$	110.83	$(1+|z|^2)^\alpha$	98.72	$\mathrm{div}\!\left(|\nabla u|^{p-2}\,\nabla u\right)$
108.85	ϕ^{-1}	108.52	x_{ij}	112.55	$|f(z)|^p$	105.69	$f\!\left(re^{i\theta}\right)$		–
100.19	z^{n-1}	108.50	$\dot{x}=A(t)x$	110.52	$\left(1+|x|^2\right)$	94.14	$f(z)=z+\sum_{n=2}^\infty a_n z^n$		–
100.06	(c_n)	106.66	$|x-x_0|$	109.19	$|f(x)|^p$	92.33	$\left(|\nabla u|^{p-2}\,\nabla u\right)$		–
100.05	$B(G)$	105.52	S^{2n+1}	106.22	$|\nabla u|^2 dx$	87.27	$(\log n/\log\log n)$		–
99.87	$\log_2 n$	104.91	$L^2\!\left(\mathbb{R}^2\right)$	102.86	$n(n-1)/2$	78.54	$O(n\log^2 n)$		–
99.65	$\xi(x)$	103.70	$\dot{x}=Ax+Bu$	101.40	$O(n^{-1})$		–		–

Table 2: Top $s(t,D)$ scores, where D is the set of all zbMATH documents with a minimum document frequency of 200, maximum document frequency of 500k, and a minimum complexity of 3.

between $f(x)$ rather than a *function application*. Investigating the most frequently used terms in zbMATH in Table 4 reveals that u is most likely considered to be a function in the dataset: $u(t)$ (rank 8), $u(x)$ (rank 13), u_{xx} (rank 16), $u(0)$ (rank 17), $|\nabla u|$ (rank 22). Manual investigations of extended lists reveal even more hits: $u_0(x)$ (rank 30), $-\Delta u$ (rank 32), and $u(x,t)$ (rank 33). Since all eight terms are among the most frequent 35 entries in zbMATH, it implies that u can most likely be considered to imply a function in zbMATH. Of course, this does not imply that u must always be a function in zbMATH (see $f(u)$ on rank 14 in C3), but this allows us to exploit probabilities for improving MathIR performance. For instance, if not stated otherwise, u could be interpreted as a function by default, which could help increase the precision of the aforementioned tools.

Figure 4 also demonstrates that our two datasets diverge for increasing complexities. Hence, we can assume that frequencies of less complex formulae are more topic-independent. Conversely, the more complex a math formula is, the more context-specific it is. In the following, we will further investigate this assumption by applying TF-IDF rankings on the distributions.

4 RELEVANCE RANKING FOR FORMULAE

Zipf's law encourages the idea of scoring the relevance of words according to their number of occurrences in the corpus and in the documents. The family of BM25 ranking functions based on TF-IDF scores are still widely used in several retrieval systems [3, 34]. Since we demonstrated that mathematical formulae (and their subexpressions) obey Zipf's law in large scientific corpora, it appears intuitive

A. Greiner-Petter, et al.
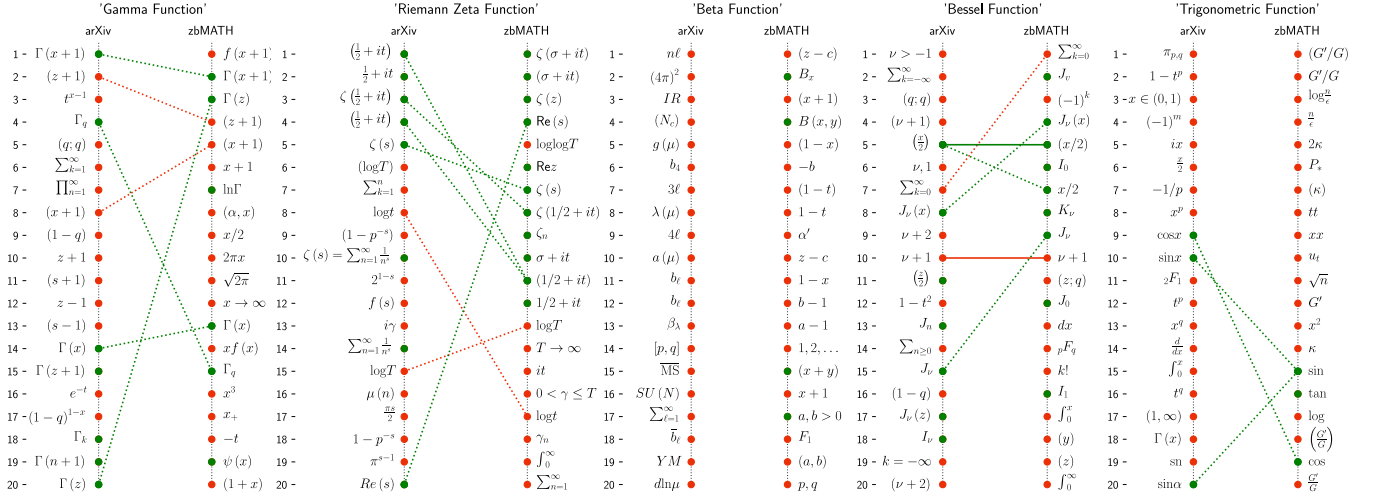
Figure 5: Top-20 ranked expressions retrieved from a topic-specific subset of documents D_q. The search query q is given above the plots. Retrieved formulae are annotated by a domain expert with green dots for relevant and red dots for non-relevant hits. A line is drawn if a hit appears in both result sets. The line is colored in green when the hit was marked as relevant.

to also use TF-IDF rankings, such as a variant of BM25, to calculate their relevance. In its original form [34], *Okapi BM25* was calculated as follows

$$\text{bm25}(t, d) := \frac{(k+1)\,\text{IDF}(t)\,\text{TF}(t, d)}{\text{TF}(t, d) + k\left(1 - b + \frac{b|d|}{\text{AVG}_{\text{DL}}}\right)}, \tag{7}$$

where $\text{TF}(t, d)$ is the term frequency of t in the document d, $|d|$ the length of the document d (in our case, the number of subexpressions), AVG_{DL} the average length of the documents in the corpus (see Table 1), and $\text{IDF}(t)$ is the inverse document frequency of t, defined as

$$\text{IDF}(t) := \log \frac{N - n(t) + \frac{1}{2}}{n(t) + \frac{1}{2}}, \tag{8}$$

where N is the number of documents in the corpus and $n(t)$ the number of documents which contain the term t. By adding $\frac{1}{2}$, we avoid $\log 0$ and division by 0. The parameters k and b are free, with b controlling the influence of the normalized document length and k controlling the influence of the term frequency on the final score. For our experiments, we chose the standard value $k = 1.2$ and a high impact factor of the normalized document length via $b = 0.95$.

As a result of our subexpression extraction algorithm, we generated a bias towards low complexities. Moreover, longer documents generally consist of more complex expressions. As demonstrated in Section 2.1, a document that only consists of the single expression $P_n^{(\alpha, \beta)}(x)$, i.e., the document had a length of one, would generate eight subexpressions, i.e., it results in a document length of eight. Thus, we modify the BM25 score in Equation (7) to emphasize higher complexities and longer documents. First, the average document length is divided by the average complexity AVG_C in the corpus that is used (see Table 1), and we calculate the reciprocal of the document length normalization to emphasize longer documents.

Moreover, in the scope of a single document, we want to emphasize expressions that do not appear frequently in this document, but are the most frequent among their level of complexity. Thus, less

	arXiv	zbMATH
Retrieved Doc.	40	200
Min. Hit Freq.	7	7
Min. DF	50	10
Max. DF	10k	10k

Table 3: Settings for the retrieval experiments.

complex expressions are ranked more highly if the document overall is not very complex. To achieve this weighting, we normalize the term frequency of an expression t according to its complexity $c(t)$ and introduce an inverse term frequency according to all expressions in the document

$$\text{ITF}(t, d) := \log \frac{|d| - \text{TF}(t, d) + \frac{1}{2}}{\text{TF}(t, d) + \frac{1}{2}}. \tag{9}$$

Finally, we define the score $\text{s}(t, d)$ of a term t in a document d as

$$s(t, d) := \frac{(k+1)\,\text{IDF}(t)\,\text{ITF}(t, d)\,\text{TF}(t, d)}{\max_{t' \in d|_{c(t)}} \text{TF}(t', d) + k\left(1 - b + \frac{b\,\text{AVG}_{\text{DL}}}{|d|\,\text{AVG}_C}\right)}. \tag{10}$$

The TF-IDF ranking functions and the introduced $\text{s}(t, d)$ are used to retrieve relevant documents for a given search query. However, we want to retrieve relevant subexpressions over a set of documents. Thus, we define the score of a formula (mBM25) over a set of documents as the maximum score over all documents

$$\text{mBM25}(t, D) := \max_{d \in D} \text{s}(t, d), \tag{11}$$

where D is a set of documents. We used *Apache Flink* [16] to count the expressions and process the calculations. Thus, our implemented system scales well for large corpora.

Table 2 shows the top-7 scored expressions, where D is the entire zbMATH dataset. The retrieved expressions can be considered as meaningful and real-world examples of MOIs, since most expressions are known for specific mathematical concepts, such as

Riemann Zeta Function

C1		C2		C3		C4		C5		C6		TF-IDF	mBM25
15,051	n	4,663	(s)	1,456	$\zeta(s)$	349	$(\frac{1}{2}+it)$	203	$\zeta(\frac{1}{2}+it)$	105	$\|\zeta(1/2+it)\|$	$\zeta(s)$	$\zeta(1/2+it)$
11,709	s	2,460	(x)	340	$\sigma+it$	232	$(1/2+it)$	166	$\zeta(1/2+it)$	88	$\|\zeta(\frac{1}{2}+it)\|$	$\zeta(1/2+it)$	$(1/2+it)$
9,768	x	2,163	(n)	310	$\sum_{n=1}^{\infty}$	195	$(\sigma+it)$	124	$\zeta(\sigma+it)$	81	$\|\zeta(\sigma+it)\|$	$(1/2+it)$	$(\frac{1}{2}+it)$
8,913	k	1,485	(t)	275	$(\log T)$	136	$\frac{1}{2}+it$	54	$\zeta(1+it)$	32	$\|\zeta(1+it)\|$	$\frac{1}{2}+it$	$\zeta(\frac{1}{2}+it)$
8,634	T	1,415	it	264	$1/2+it$	97	$s=\sigma+it$	44	$\zeta(2n+1)$	22	$\|\zeta(+it)\|$	$(\frac{1}{2}+it)$	$(\sigma+it)$

Eigenvalue

C1		C2		C3		C4		C5		C6		TF-IDF	mBM25
45,488	n	12,515	(x)	686	$-\Delta u$	218	$\|\nabla u\|^{p-2}$	139	$\|\nabla u\|^{p-2}\nabla u$	137	$\left(\|\nabla u\|^{p-2}\nabla u\right)$	$Ax=\lambda Bx$	$-\operatorname{div}\left(\|\nabla u\|^{p-2}\nabla u\right)$
43,090	x	6,598	(t)	555	$(n-1)$	218	$-\Delta_p u$	68	$-d^2/dx^2$	35	$-(py')'$	$-\Delta p$	$\operatorname{div}\left(\|\nabla u\|^{p-2}\nabla u\right)$
37,434	λ	4,377	λ_1	521	$\|\nabla u\|$	133	$W_0^{1,p}(\Omega)$	51	$A=(a_{ij})$	26	$(\|u'\|^{p-2}u')$	$P(\lambda)$	$p=\frac{N+2}{N-2}$
35,302	u	2,787	(Ω)	512	a_{ij}	127	$\|\nabla u\|^2$	46	$-\frac{d^2}{dx^2}$	18	$(\phi_p(u'))'$	λ_{k+1}	$(\phi_p(u'))'$
22,460	t	2,725	$\mathbb{R}^n$	495	$u(x)$	97	(a_{ij})	45	$u\in W_0^{1,p}(\Omega)$	18	$\int_\Omega \|\nabla u\|^2\,dx$	$\lambda_1>0$	$\lambda\in(0,\lambda^*)$

Table 4: The top-5 frequent mathematical expressions in the result set of zbMATH for the search queries 'Riemann Zeta Function' (top) and 'Eigenvalue' (bottom) grouped by their complexities (left) and the hits reordered according to their relevance scores (right). The TF-IDF score was calculated with normalized term frequencies.

$\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, which refers to the Galois group of $\overline{\mathbb{Q}}$ over $\mathbb{Q}$, or $L^2(\mathbb{R}^2)$, which refers to the L^2-space (also known as *Lebesgue space*) over $\mathbb{R}^2$. However, a more topic-specific retrieval algorithm is desirable. To achieve this goal, we (i) retrieved a topic-specific subset of documents $D_q \subset D$ for a given textual search query q, and (ii) calculated the scores of all expressions in the retrieved documents. To generate D_q, we indexed the text sources of the documents from arXiv and zbMATH via elasticsearch (ES)[13] and performed the pre-processing steps: filtering stop words, stemming, and ASCII-folding[14]. Table 3 summarizes the settings we used to retrieve MOIs from a topic-specific subset of documents D_q. We also set a minimum hit frequency according to the number of retrieved documents an expression appears in. This requirement filters out uncommon notations.

Figure 5 shows the results for five search queries. We asked a domain expert from the National Institute of Standards and Technology (NIST) to annotate the results as related (shown as green dots in Figure 5) or non-related (red dots). We found that the results range from good performances (e.g., for the Riemann zeta function) to bad performances (e.g., beta function). For instance, the results for the Riemann zeta function are surprisingly accurate, since we could discover that parts of Riemann's hypothesis[15] were ranked highly throughout the results (e.g., $\zeta(\frac{1}{2}+it)$). On the other hand, for the beta function, we retrieved only a few related hits, of which only one had a strong connection to the beta function $B(x,y)$. We observed that the results were quite sensitive to the chosen settings (see Table 3). For instance, according to the beta function, the minimum hit frequency has a strong effect on the results, since many expressions are shared among multiple documents. For arXiv, the

expressions $B(\alpha,\beta)$ and $B(x,y)$ only appear in one document of the retrieved 40. However, decreasing the minimum hit frequency would increase noise in the results.

Even though we asked a domain expert to annotate the results as relevant or not, there is still plenty of room for discussion. For instance, $(x+y)$ (rank 15 in zbMATH, 'Beta Function') is the argument of the gamma function $\Gamma(x+y)$ that appears in the definition of the beta function [8, (5.12.1)] $B(x,y) := \Gamma(x)\Gamma(y)/\Gamma(x+y)$. However, this relation is weak at best, and thus might be considered as not related. Other examples are $\mathrm{Re}z$ and $\mathrm{Re}(s)$, which play a crucial role in the scenario of the Riemann hypothesis (all non-trivial zeroes have $\mathrm{Re}(s)=\frac{1}{2}$). Again, this connection is not obvious, and these expressions are often used in multiple scenarios. Thus, the domain expert did not mark the expressions as being related.

Considering the differences in the documents, it is promising to have observed a relatively high number of shared hits in the results. Further, we were able to retrieve some surprisingly good insights from the results, such as extracting the full definition of the Riemann zeta function [8, (25.2.1)] $\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}$. Even though a high number of shared hits seem to substantiate the reliability of the system, there were several aspects that affected the outcome negatively, from the exact definition of the search queries to retrieve documents via ES, to the number of retrieved documents, the minimum hit frequency, and the parameters in mBM25.

5 APPLICATIONS

The presented results are beneficial for a variety of use-cases. In the following, we will demonstrate and discuss several of the applications that we propose.

Extension of zbMATH's Search Engine: Formula search engines are often counterintuitive when compared to textual search, since the user must know how the system operates to enter a search query properly (e.g., does the system supports LaTeX inputs?). Additionally, mathematical concepts can be difficult to capture using only mathematical expressions. Consider, for example, someone

---

[13] https://github.com/elastic/elasticsearch [Accessed Sep. 2019]. We used version 7.0.0
[14] This means that non-ASCII characters are replaced by their ASCII counterparts or will be ignored if no such counterpart exists.
[15] Riemann proposed that the real part of every non-trivial zero of the Riemann zeta function is $1/2$. If this hypothesis is correct, all the non-trivial zeros lie on the critical line consisting of the complex numbers $1/2+it$.

Auto-completion for '$E = m$'			Suggestions for '$E = \{m, c\}$'		
Sug. Expression	TF	DF	Sug. Expression	TF	DF
$E = mc^2$	558	376	$E = mc^2$	558	376
$E = m \cosh \theta$	23	23	$E = \gamma mc^2$	39	38
$E = mv_0$	7	7	$E = \gamma m_e c^2$	41	36
$E = m/\sqrt{1 - \dot{q}^2}$	12	6	$E = m \cosh \theta$	23	23
$E = m/\sqrt{1 - \beta^2}$	10	6	$E = -mc^2$	35	17
$E = mc^2 \gamma$	6	6	$E = \sqrt{m^2 c^4 + p^2 c^2}$	10	8

Table 5: Suggestions to complete '$E = m$' and '$E = \{m, c\}$' (the right-hand side contains m and c) with term and document frequency based on the distributions of formulae in arXiv.

who wants to search for mathematical expressions that are related to eigenvalues. A textual search query would only retrieve entire documents that require further investigation to find related expressions. A mathematical search engine, on the other hand, is impractical since it is not clear what would be a fitting search query (e.g., $Av = \lambda v$?). Moreover, formula and textual search systems for scientific corpora are separated from each other. Thus, a textual search engine capable of retrieving mathematical formulae can be beneficial. Also, many search engines allow for narrowing down relevant hits by suggesting filters based on the retrieved results. This technique is known as faceted search. The zbMATH search engine also provides faceted search, e.g., by authors, or year. Adding facets for mathematical expressions allows users to narrow down the results more precisely to arrive at specific documents.

Our proposed system for extracting relevant expressions from scientific corpora via mBM25 scores can be used to search for formulae even with textual search queries, and to add more filters for faceted search implementations. Table 4 shows two examples of such an extension for zbMATH's search engine. Searching for 'Riemann Zeta Function' and 'Eigenvalue' retrieved 4,739 and 25,248 documents from zbMATH, respectively. Table 4 shows the most frequently used mathematical expressions in the set of retrieved documents. It also shows the reordered formulae according to a default TF-IDF score (with normalized term frequencies) and our proposed mBM25 score. The results can be used to add filters for faceted search, e.g., show only the documents which contain $u \in W_0^{1,p}(\Omega)$. Additionally, the search system now provides more intuitive textual inputs even for retrieving mathematical formulae. The retrieved formulae are also interesting by themselves, since they provide insightful information on the retrieved publications. As already explored with our custom document search system in Figure 5, the Riemann hypothesis is also prominent in these retrieved documents.

The differences between TF-IDF and mBM25 ranking illustrates the problem of an extensive evaluation of our system. From a broader perspective, the hit $Ax = \lambda Bx$ is highly correlated with the input query 'Eigenvalue'. On the other hand, the raw frequencies revealed a prominent role of $\text{div}(|\nabla u|^{p-2} \nabla u)$. Therefore, the top results of the mBM25 ranking can also be considered as relevant.

Math Notation Analysis: A faceted search system allows us to analyze mathematical notations in more detail. For instance, we can retrieve documents from a specific time period. This allows one to study the evolution of mathematical notation over time [4], or for
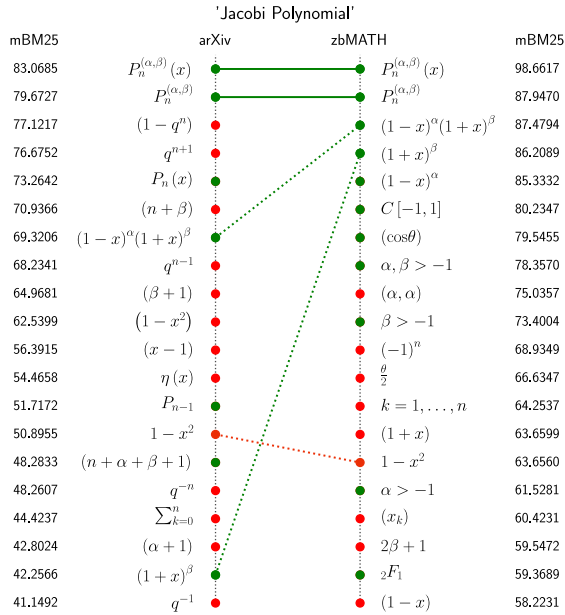
identifying trends in specific fields. Also, we can analyze standard notations for specific authors since it is often assumed that authors prefer a specific notation style which may vary from the standard notation in a field.

Math Recommendation Systems: The frequency distributions of formulae can be used to realize effective math recommendation tasks, such as type hinting or error-corrections. These approaches require long training on large datasets, but may still generate meaningless results, such as $G_i = \{(x, y) \in \mathbb{R}^n : x_i = x_i\}$ [42]. We propose a simpler system which takes advantage of our frequency distributions. We retrieve entries from our result database, which contain all unique expressions and their frequencies. We implemented a simple prototype that retrieves the entries via pattern matching. Table 5 shows two examples. The left side of the table shows suggested autocompleted expressions for the query '$E = m$'. The right side shows suggestions for '$E =$', where the right-hand side of the equation should contain m and c in any order. A combination using more advanced retrieval techniques, such as similarity measures based on symbol layout trees [7, 45], would enlarge the number of suggestions. This kind of autocomplete and error-correction type-hinting system would be beneficial for various use-cases, e.g., in educational software or for search engines as a pre-processing step of the input.

Plagiarism Detection Systems: As previously mentioned, plagiarism detection systems [28, 29, 39] would benefit from a system capable of distinguishing conventional from uncommon notations. The approaches described by Meuschke et al. [29] outperform existing approaches by considering frequency distributions of single identifiers (expressions of complexity one). Considering that single identifiers make up only 0.03% of all unique expressions in arXiv, we presume that better performance can be achieved by considering more complex expressions. The conferred string representation also provides a simple format to embed complex expressions in existing learning algorithms.

Expressions with high complexities that are shared among multiple documents may provide further hints to investigate potential plagiarisms. For instance, the most complex expression that was shared among three documents in arXiv was Equation (3). A complex expression being identical in multiple documents could indicate a higher likelihood of plagiarism. Further investigation revealed that similar expressions, e.g., with infinite sums, are frequently used among a larger set of documents. Thus, the expression seems to be a part of a standard notation that is commonly shared, rather than a good candidate for plagiarism detection. Resulting from manual investigations, we could identify the equation as part of a concept called *generalized Hardy-Littlewood inequality* and Equation (3) appears in the three documents [2, 5, 32]. All three documents shared one author in common. Thus, this case also demonstrates a correlation between complex mathematical notations and authorship.

Semantic Taggers and Extraction Systems: We previously mentioned that semantic extraction systems [23, 36, 37] and semantic math taggers [6, 43] have difficulties in extracting the essential components (MOIs) from complex expressions. Considering the definition of the Jacobi polynomial in Equation (1), it would be beneficial to extract the groups of tokens that belong together, such as $P_n^{(\alpha, \beta)}(x)$ or $\Gamma(\alpha + m + 1)$. With our proposed search engine

'Jacobi Polynomial'

mBM25	arXiv	zbMATH	mBM25
83.0685	$P_n^{(\alpha,\beta)}(x)$	$P_n^{(\alpha,\beta)}(x)$	98.6617
79.6727	$P_n^{(\alpha,\beta)}$	$P_n^{(\alpha,\beta)}$	87.9470
77.1217	$(1-q^n)$	$(1-x)^\alpha(1+x)^\beta$	87.4794
76.6752	q^{n+1}	$(1+x)^\beta$	86.2089
73.2642	$P_n(x)$	$(1-x)^\alpha$	85.3332
70.9366	$(n+\beta)$	$C[-1,1]$	80.2347
69.3206	$(1-x)^\alpha(1+x)^\beta$	$(\cos\theta)$	79.5455
68.2341	q^{n-1}	$\alpha,\beta>-1$	78.3570
64.9681	$(\beta+1)$	(α,α)	75.0357
62.5399	$(1-x^2)$	$\beta>-1$	73.4004
56.3915	$(x-1)$	$(-1)^n$	68.9349
54.4658	$\eta(x)$	$\frac{\theta}{2}$	66.6347
51.7172	P_{n-1}	$k=1,\ldots,n$	64.2537
50.8955	$1-x^2$	$(1+x)$	63.6599
48.2833	$(n+\alpha+\beta+1)$	$1-x^2$	63.6560
48.2607	q^{-n}	$\alpha>-1$	61.5281
44.4237	$\sum_{k=0}^n$	(x_k)	60.4231
42.8024	$(\alpha+1)$	$2\beta+1$	59.5472
42.2566	$(1+x)^\beta$	${}_2F_1$	59.3689
41.1492	q^{-1}	$(1-x)$	58.2231

Figure 6: The top ranked expression for '*Jacobi polynomial*' in arXiv and zbMATH. For arXiv, 30 documents were retrieved with a minimum hit frequency of 7.

for retrieving MOIs, we are able to facilitate semantic extraction systems and semantic math taggers. Imagine such a system being capable of identifying the term 'Jacobi polynomial' from the textual context. Figure 6 shows the top relevant hits for the search query 'Jacobi polynomial' retrieved from zbMATH and arXiv. The results contain several relevant and related expressions, such as the constraints $\alpha,\beta>-1$ and the weight function for the Jacobi polynomial $(1-x)^\alpha(1+x)^\beta$, which are essential properties of this orthogonal polynomial. Based on these retrieved MOIs, the extraction systems can adjust its retrieved math elements to improve precision, and semantic taggers or a tokenizer could re-organize parse trees to more closely resemble expression trees.

6 CONCLUSION & FUTURE WORK

In this study we showed that analyzing the frequency distributions of mathematical expressions in large scientific datasets can provide useful insights for a variety of applications. We demonstrated the versatility of our results by implementing prototypes of a type-hinting system for math recommendations, an extension of zbMATH's search engine, and a mathematical retrieval system to search for topic-specific MOIs. Additionally, we discussed the potential impact and suitability in other applications, such as math search engines, plagiarism detection systems, and semantic extraction approaches. We are confident that this project lays a foundation for future research in the field of MathIR.

We plan on developing a web application which would provide easy access to our frequency distributions, the MOI search engine, and the type-hinting recommendation system. We hope that this will further expedite related future research projects. Moreover, we will use this web application for an online evaluation of our MOI retrieval system. Since the level of agreement among annotators

will be predictably low, an evaluation by a large community is desired.

In this first study, we preserved the core structure of the MathML data which provided insightful information for the MathML community. However, this makes it difficult to properly merge formulae. In future studies, we will normalize the MathML data via MathML-Can [9]. In addition to this normalization, we will include wildcards for investigating distributions of formula patterns rather than exact expressions. This will allow us to study connections between math objects, e.g., between $\Gamma(z)$ and $\Gamma(x+1)$. This would further improve our recommendation system and would allow for the identification of regions for parameters and variables in complex expressions.

ACKNOWLEDGMENTS

REFERENCES

[1] Akiko N. Aizawa. 2003. An information-theoretic perspective of tf-idf measures. *Inf. Process. Manage.* Vol. 39, Issue 1, 45–65. https://doi.org/10.1016/S0306-4573(02)00021-3
[2] Gustavo Araujo and Daniel Pellegrino. 2014. On the constants of the Bohnenblust-Hille inequality and Hardy–Littlewood inequalities. In *Computing Research Repository (CoRR)* https://arxiv.org/abs/1407.7120.
[3] Jöran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *Int. J. on Digital Libraries* Vol. 17, Issue 4, 305–338. https://doi.org/10.1007/s00799-015-0156-0
[4] Florian Cajori. 1929. *A History of Mathematical Notations.* The Open Court Company, London, UK. Vol. 1 & 2.
[5] Jamilson R. Campos, Wasthenny Cavalcante, Vinícius V. Fávaro, Daniel Nu nez Alarcón, Daniel Pellegrino, and Diana M. Serrano-Rodríguez. 2015. Polynomial and multilinear Hardy–Littlewood inequalities: analytical and numerical approaches. In *Computing Research Repository (CoRR)* https://arxiv.org/abs/1503.00618.
[6] Pao-Yu Chien and Pu-Jen Cheng. 2015. Semantic Tagging of Mathematical Expressions. In *Proc. WWW'2015.* ACM, 195–204. https://doi.org/10.1145/2736277.2741108
[7] Kenny Davila and Richard Zanibbi. 2017. Layout and Semantics: Combining Representations for Mathematical Formula Search. In *Proc. ACM SIGIR.* ACM, 1165–1168. https://doi.org/10.1145/3077136.3080748
[8] DLMF 2019. *NIST Digital Library of Mathematical Functions.* http://dlmf.nist.gov/, Release 1.0.25 of 2019-12-15. http://dlmf.nist.gov F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.
[9] David Formánek, Martin Líška, Michal Růžička, and Petr Sojka. 2012. Normalization of Digital Mathematics Library Content. In *Proc. of OpenMath/ Math-UI/ CICM-WiP (CEUR Workshop Proceedings).* 91–103. http://ceur-ws.org/Vol-921/wip-05.pdf
[10] Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang. 2017. Preliminary Exploration of Formula Embedding for Mathematical Information Retrieval: can mathematical formulae be embedded like a natural language? In *Computing Research Repository (CoRR)* http://arxiv.org/abs/1707.05154.
[11] Alex Gaudeul. 2007. Do Open Source Developers Respond to Competition?: The LATEX Case Study. In *Review of Network Economics* https://doi.org/10.2202/1446-9022.1119
[12] Deyan Ginev. 2018. *arXMLiv:08.2018 dataset, an HTML5 conversion of arXiv.org.* https://sigmathling.kwarc.info/resources/arxmliv/ SIGMathLing – Special Interest Group on Math Linguistics.
[13] André Greiner-Petter, Terry Ruas, Moritz Schubotz, Akiko Aizawa, William I. Grosky, and Bela Gipp. 2019. Why Machines Cannot Learn Mathematics, Yet. In *Proceedings of the 4th Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL 2019) co-located with the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019), Paris, France, July 25, 2019* Vol. 2414. CEUR-WS.org, 130–137. http://ceur-ws.org/Vol-2414/paper14.pdf
[14] Christian Grün, Sebastian Gath, Alexander Holupirek, and Marc Scholl. 2009. XQuery Full Text Implementation in BaseX. In *Database and XML Technologies.* Springer Berlin, 114–128.
[15] Ferruccio Guidi and Claudio Sacerdoti Coen. 2016. A Survey on Retrieval of Mathematical Knowledge. *Mathematics in Computer Science* Vol. 10, Issue 4, 409–427. https://doi.org/10.1007/s11786-016-0274-0

[16] Fabian Hueske and Timo Walther. 2019. Apache Flink. In *Encyclopedia of Big Data Technologies.*, Sherif Sakr and Albert Y. Zomaya (Eds.). Springer. https://doi.org/10.1007/978-3-319-63962-8_303-1

[17] Shahab Kamali and Frank Wm. Tompa. 2010. A new mathematics retrieval system. In *Proc. ACM CIKM.* ACM, 1413–1416. https://doi.org/10.1145/1871437.1871635

[18] Shahab Kamali and Frank Wm. Tompa. 2013. Retrieving documents with mathematical content. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013.* ACM, 353–362. https://doi.org/10.1145/2484028.2484083

[19] Andrea Kohlhase. 2018. Factors for Reading Mathematical Expressions. In *Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", LWDA 2018, Mannheim, Germany, August 22-24, 2018. (CEUR Workshop Proceedings)*, Vol. 2191. CEUR-WS.org, 195–202. http://ceur-ws.org/Vol-2191/paper24.pdf

[20] Andrea Kohlhase, Michael Kohlhase, and Michael Fürsich. 2017. Visual Structure in Mathematical Expressions. In *Intelligent Computer Mathematics - 10th International Conference, CICM 2017, Edinburgh, UK, July 17-21, 2017, Proceedings (Lecture Notes in Computer Science)*, Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke (Eds.). Springer, 208–223. https://doi.org/10.1007/978-3-319-62075-6_15

[21] Andrea Kohlhase, Michael Kohlhase, and Taweechai Ouypornkochagorn. 2018. Discourse Phenomena in Mathematical Documents. In *Intelligent Computer Mathematics - 11th International Conference, CICM 2018, Hagenberg, Austria, August 13-17, 2018, Proceedings (Lecture Notes in Computer Science)*, Vol. 11006. Springer, 147–163. https://doi.org/10.1007/978-3-319-96812-4_14

[22] Michael Kohlhase, Bogdan A. Matican, and Corneliu-Claudiu Prodescu. 2012. MathWebSearch 0.5: Scaling an Open Formula Search Engine. In *Intelligent Computer Mathematics - 11th International Conference, AISC 2012, 19th Symposium, Calculemus 2012, 5th International Workshop, DML 2012, 11th International Conference, MKM 2012, Systems and Projects, Held as Part of CICM 2012, Bremen, Germany, July 8-13, 2012. Proceedings.* Springer Berlin Heidelberg, 342–357. https://doi.org/10.1007/978-3-642-31374-5_23

[23] Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2017. Utilizing dependency relationships between math expressions in math IR. *Information Retrieval Journal* Vol. 20, Issue 2, 132–167. https://doi.org/10.1007/s10791-017-9296-8

[24] Giovanni Yoko Kristianto, Goran Topic, Florence Ho, and Akiko Aizawa. 2014. The MCAT Math Retrieval System for NTCIR-11 Math Track. In *Proc. 11th NTCIR Conference on Evaluation of Information Access Technologies, National Center of Sciences,*, Noriko Kando, Hideo Joho, and Kazuaki Kishida (Eds.). National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/06-NTCIR11-MATH-KristiantoGY.pdf

[25] Kriste Krstovski and David M. Blei. 2018. Equation Embeddings. In *Computing Research Repository (CoRR)* http://arxiv.org/abs/1803.09123.

[26] Aldo Lipani, Linda Andersson, Florina Piroi, Mihai Lupu, and Allan Hanbury. 2014. TUW-IMP at the NTCIR-11 Math-2. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, NTCIR-11, National Center of Sciences, Tokyo, Japan, December 9-12, 2014.* National Institute of Informatics (NII). http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/09-NTCIR11-MATH-LipaniA.pdf

[27] Ashish Lohia, Kirti Sinha, Soujanya Vadapalli, and Kamalakar Karlapalem. 2005. An Architecture for Searching and Indexing Latex Equations in Scientific Literature. In *Proc. COMAD.* Computer Society of India, 122–130. http://comad2005.persistent.co.in/COMAD2005Proc/pages122-130.pdf

[28] Norman Meuschke, Moritz Schubotz, Felix Hamborg, Tomás Skopal, and Bela Gipp. 2017. Analyzing Mathematical Content to Detect Academic Plagiarism. In *Proc. ACM CIKM.* ACM, 2211–2214. https://doi.org/10.1145/3132847.3133144

[29] Norman Meuschke, Vincent Stange, Moritz Schubotz, Michael Kramer, and Bela Gipp. 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 (JCDL).* 120–129. https://doi.org/10.1109/JCDL.2019.00026

[30] Bruce R. Miller. 2019. LaTeXML A LATEX to XML/HTML/MathML Converter. http://dlmf.nist.gov/LaTeXML/. http://dlmf.nist.gov/LaTeXML/ Accessed: 2019-09-01.

[31] Shunsuke Ohashi, Giovanni Yoko Kristianto, Goran Topic, and Akiko Aizawa. 2016. Efficient Algorithm for Math Formula Semantic Search. *IEICE Transactions* Vol. E99.D, Issue 4, 979–988. https://doi.org/10.1587/transinf.2015DAP0023

[32] Daniel Pellegrino. 2015. A short communication on the constants of the multilinear Hardy–Littlewood inequality. In *Computing Research Repository (CoRR)* https://arxiv.org/abs/1510.00367.

[33] Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review* Vol. 21, Issue 5 (March 2014), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6

[34] Stephen E. Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* Vol. 3, Issue 4, 333–389. https://doi.org/10.1561/1500000019

[35] Moritz Schubotz, André Greiner-Petter, Philipp Scharpf, Norman Meuschke, Howard S. Cohl, and Bela Gipp. 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, June 03-07, 2018*, Jiangping Chen, Marcos André Gonçalves, Jeff M. Allen, Edward A. Fox, Min-Yen Kan, and Vivien Petras (Eds.). ACM, 233–242. https://doi.org/10.1145/3197026.3197058

[36] Moritz Schubotz, Alexey Grigorev, Marcus Leich, Howard S. Cohl, Norman Meuschke, Bela Gipp, Abdou S. Youssef, and Volker Markl. 2016. Semantification of Identifiers in Mathematics for Better Math Information Retrieval. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16).* ACM, 135–144. https://doi.org/10.1145/2911451.2911503 Full Paper.

[37] Moritz Schubotz, Leonard Krämer, Norman Meuschke, Felix Hamborg, and Bela Gipp. 2017. Evaluating and Improving the Extraction of Mathematical Identifier Definitions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings (Lecture Notes in Computer Science)*, Gareth J. F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, and Nicola Ferro (Eds.), Vol. 10456. Springer, 82–94. https://doi.org/10.1007/978-3-319-65813-1_7

[38] Moritz Schubotz and Olaf Teschke. 2019. Four decades of TEX at zbMATH. *Newsletter of the European Mathematical Society (EMS)* 6, 50–52. https://doi.org/10.4171/NEWS/112/15

[39] Moritz Schubotz, Olaf Teschke, Vincent Stange, Norman Meuschke, and Bela Gipp. 2019. Forms of Plagiarism in Digital Mathematical Libraries. In *Intelligent Computer Mathematics - 12th International Conference, CICM 2019, Prague, Czech Republic, July 8-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Vol. 11617. Springer, 258–274. https://doi.org/10.1007/978-3-030-23250-4_18

[40] Glenn Gordon Smith and David Ferguson. 2004. Diagrams and math notation in e-learning: growing pains of a new generation. *International Journal of Mathematical Education in Science and Technology* Vol. 35, 681–695. Issue 5. https://doi.org/10.1080/0020739042000232583

[41] Leonard Wörteler, Michael Grossniklaus, Christian Grün, and Marc Scholl. 2015. Function inlining in XQuery 3.0 optimization. In *Proc. 15th DBLP.* ACM, 45–48. https://doi.org/10.1145/2815072.2815079

[42] Michihiro Yasunaga and John Lafferty. 2019. TopicEq: A Joint Topic and Mathematical Equation Model for Scientific Texts. In *Computing Research Repository (CoRR)* http://arxiv.org/abs/1902.06034.

[43] Abdou Youssef. 2017. Part-of-Math Tagging and Applications. In *Intelligent Computer Mathematics*, Herman Geuvers, Matthew England, Osman Hasan, Florian Rabe, and Olaf Teschke (Eds.). Springer International Publishing, Cham, 356–374.

[44] Abdou Youssef and Bruce R. Miller. 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, July 8-12, 2019, Proceedings (Lecture Notes in Computer Science)*, Vol. 11617. Springer, 291–305. https://doi.org/10.1007/978-3-030-23250-4_20

[45] Richard Zanibbi, Kenny Davila, Andrew Kane, and Frank Wm. Tompa. 2016. 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 (SIGIR '16).* ACM, New York, NY, USA, 145–154. https://doi.org/10.1145/2911451.2911512