Challenges of Mathematical Information Retrieval in the NTCIR-11 Math Wikipedia Task

Moritz Schubotz
Database Systems and Information Management Group, TU Berlin, Germany
schubotz@tu-berlin.de

Abdou Youssef
Department of Computer Science, The George Washington University, USA
ayoussef@gwu.edu

Volker Markl
Database Systems and Information Management Group, TU Berlin, Germany
volker.markl@tu-berlin.de

Howard S. Cohl
Applied and Computational Mathematics Div., National Inst. of Standards and Technology, USA
howard.cohl@nist.gov

ABSTRACT

Mathematical Information Retrieval concerns retrieving information related to a particular mathematical concept. The NTCIR-11 Math Task develops an evaluation test collection for document sections retrieval of scientific articles based on human generated topics. Those topics involve a combination of formula patterns and keywords. Another task in NTCIR-11 is the optional Wikipedia Task, which provides a test collection for retrieval of individual mathematical formula from Wikipedia based on search topics that contain exactly one formula pattern. We developed a framework for automatic query generation and immediate evaluation. This paper discusses our dataset preparation, our topic generation and evaluation methods, and summarizes the results of the participants, with a special focus on the Wikipedia Task.

1. INTRODUCTION

Math Information Retrieval (MIR) is a growing field. Recent publications (e.g., [4, 6, 10]) show that there is a significant demand for enhancement in Mathematical Knowledge Management. In order to compare different approaches and measure their performance, test collections are needed. At the CICM 2012 conference in Bremen, Germany, the first "MIR happening" took place with two participants, 10,000 arXiv documents and a dataset size of 293 MB. In 2013, the NTCIR-10 Math pilot task [2] for MIR attracted 6 participants and used 100,000 arXiv documents with a dataset size of 63GB. Based on the gathered experience, MIR qualified for a main task at NTCIR-11 [3] which took place in 2014 with 8 participants and 8 million document sections of 174GB in total. Additionally, the newly introduced Wikipedia Task was appreciated by the participants. We expect that the automated feedback and evaluation framework will lower the entrance level for participants and attract even more participants in the future. In this first section, we provide a general introduction to MIR, list some applications of MIR, and explain why MIR is fundamentally different from other IR tasks, such as text retrieval and XMLretrieval. In Section 2, we describe how the Wikipedia dataset was prepared and augmented. In Section 3 and 4, we describe the query design and evaluation process respectively. In Section 5, we present the participating teams and the performance their MIR systems. In Section 6 we give a future outlook for the Wikipedia Task.

Introduction to Math Information Retrieval

The use cases for MIR are diverse. They include applicable theorem search, plagiarism detection, related work search, patent search, and search in Excel spreadsheets [5]. Some of the fundamental concepts that relate back to tree structure search can be used for code-search or search for chemical formulae.

For the Wikipedia Task, the focus is on information needs that involve mathematics that is naturally expressed using mathematical expressions. With regard to the aforementioned applications, those information needs can be expressed as mathematical expressions, combination of expressions and keywords, or keywords only. While the main task uses all of these combinations to retrieve documents, the Wikipedia Task provides exactly one formula per topic to describe an information need.

2. DATASET & FEEDBACK TO WIKIPEDIA

The test collection used at NTCIR-10 in 2013 was based on a random selection of arXiv articles converted via LaTeXML to HTML5 [8]. The arXiv is a vast and expanding source of knowledge for researchers and experts in highly specialized domains. However, neither math search engine developers (participants) nor the assessors that evaluate the search engine results usually are domain experts in all the topics addressed in the arXiv publications. Some topics discussed in the research papers are so specialized that it becomes impossible for participants to get even a preliminary idea of the content and to decide on the relevance of a formula with respect to a topic (i.e., the underlying information need). This fact adds additional complexity to debugging and testing of the Math search engines. In contrast to the arXiv

SIGIR'15, August 09-13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08.
DOI: http://dx.doi.org/10.1145/2766462.2767787.

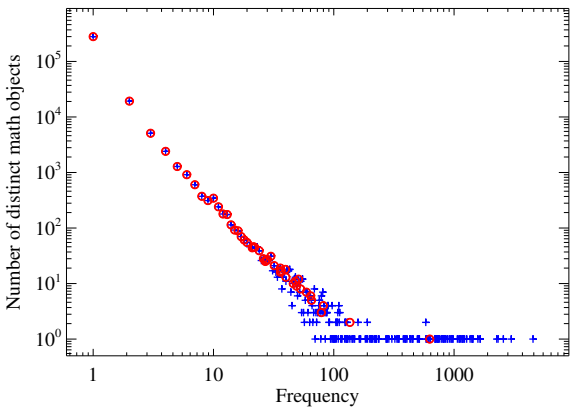

Figure 1: Distribution (blue "+") and mean (red "o") of the number of distinct formulae and their frequencies. For example the expression n occurs 2988 times and there are about 280 000 formulae that occur only once.

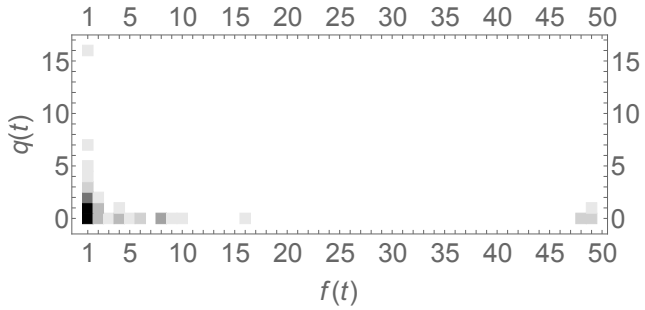

Figure 2: Density of the topics with regard to the frequency of seed formula $f(t)$ and the number of query variables in the topic $q(t)$. The blackest box corresponds to the highest count (up to 41 topics) and the lightest grey corresponds to only 1 topic.

dataset, the Wikipedia encyclopaedia contains most of the mathematical world knowledge explained in simple terms. While this knowledge is not sufficient for new research, it is perfectly suitable as a test corpus for math search competitions. This knowledge simplifies debugging and testing of the math search engines and enables the participants to test their systems on a dataset that is easier to understand and contains all formulae they are already familiar with. The English Wikipedia contains about 30k encyclopaedic articles with mathematical formula. Those are written using the TEX-like input format texvc. Even though the syntax in texvc is restricted and does not allow to write Turing complete programs, as it is possible with TEX, TEX is neither the optimal way to represent Mathematics on the web nor to search for formula. In contrast, MATHML was designed to serve the aforementioned purposes.

Schubotz and Wicke [7] compare different conversion methods and identify the LATEXML converter as the most reasonable solution with content MATHML support for Wikipedia. A majority of the participating systems use content for the search task. Therefore, both tasks (arXiv and Wikipedia) use LATEXML to convert the original user input to MATHML with parallel content and presentation markup and the original input as annotation.

In order to generate stable and unique references to each individual formula used in Wikipedia, we created an unique index, from which one can derive many interesting statistics about the usage of mathematics within Wikipedia. For example, Figure 1 shows the frequency distributions of the formulae. One use case we published at `formulasearchengine.com` is an auto-completion list for TEX commands based on the usage statistics that we gathered by tokenizing the frequency ordered formulae. People editing Wikipedia articles about math with mobile devices will benefit from this feature.

3. TOPIC DESIGN

From our experience with the Math pilot task at NTCIR-10, we draw the following conclusions: 1) for each query there should be at least one relevant hit in the dataset; 2) the semantics of query variables should be well-defined; 3) only the information that would be exposed to a MIR system should be communicated to the participants; 4) the relevance criteria for each topic should be independent of the topic author and assessor. While the main task addressed these issues with human intuition, we chose a different approach for the Wikipedia Task that does not involve humans.

We developed a program that generates queries in three simple steps. At first, the so called seed formula is chosen based on random selection from our mathindex. In a second step, we inject query variables. In contrast to the main task, where humans chose a meaningful name for the query variable, we call our query variables $x_0, x_1, \ldots$. Finally, we generate the NTCIR-11 XML Topics [1] using LATEXML. Our relevance criterion is to find a formula similar to the seed. By our naming convention for the query variable and the absence of a topic title we ensure that no information is exposed to the participants that is not intended to be used according to the topic specification, i.e., MIR systems cannot use the name of query variables for relevance ranking. Our method generates two meta-information pieces $f(t)$ and $q(t)$ for each topic t. Here, $f(t)$ is the frequency, that indicates how many exact matches (based on exact matches on the original TEX input) for each seed are contained in the dataset, and $q(t)$ is the number of query variables used. This allows for an *a priori* classification of the search topics based on f and q. For simplicity, we partition the set of generated topics T (Figure 2) into the 4 following groups: *Easy topics* without query variables and exactly one precise match $E = \{t \in T : f(t) = q(t) - 1 = 1\}$; *variable topics* with query variables but only one exact match for the underlying seed $V = \{t \in T : f(t) = 1 \wedge q(t) > 0\}$; *frequent topics* without query variables but with non-unique seeds $F = \{t \in T : q(t) = 0 \wedge f(t) > 1\}$ and *hard topics* that contain query variables and non-unique seeds $H = \{t \in T : f(t) > 1 \wedge q(t) > 0\}$. For the set of queries used in NTCIR-11 the following cardinalities were given: $|E| = 41, |V| = 27, |F| = 24, |H| = 8 \Rightarrow |T| = 100$.

4. EVALUATION PROCESS

For retrieval tasks with one known good result, a typical evaluation measure is the mean reciprocal rank (mrr) [9]. In addition, our automatic evaluation software calculates mean average precision in the first k hits for different levels of k, and counts the *number of found seeds* referred to as *success*

Participant	runs		page					formula				
	total	distinct	**total**	easy	frequent	variable	hard	**total**	easy	frequent	variable	hard
TUB Technische Universität Berlin (Germany)	4	4	**91**	100	96	74	87	**87**	100	92	70	63
			73	94	30	90	46	**68**	87	25	86	30
KWARC Jacobs University Bremen (Germnay)	1	1	**75**	83	75	70	50	-	-	-	-	-
			82	95	44	97	67	-	-	-	-	-
RMHS Richard Montgomery High School (USA)	1	1	**48**	54	46	40	50	-	-	-	-	-
			02	01	00	03	01	-	-	-	-	-
RIT Rochester Institute of Technology (USA)	17	4	**88**	98	79	89	63	**78**	95	50	81	63
			80	96	31	92	83	**86**	94	47	96	83
MIAS Masaryk University (Czech republic)	19	4	**65**	97	92	15	-	**63**	95	83	15	13
			76	93	46	83	-	**81**	91	71	83	01
TUW Vienna University of Technology (Austria)	5	3	**97**	100	100	93	88	**93**	100	96	89	63
			82	97	50	96	54	**88**	96	72	94	71
NII National Institute of Informatics (Japan)	9	4	**97**	98	100	93	100	**94**	98	96	89	88
			76	99	49	82	67	**77**	89	92	78	48

Table 1: This table lists the best performing runs, with regard to (success upper number in %) and mrr (lower number in %) with page and formula-centric evaluation methods. Furthermore, we have displayed the individual results with respect to the query categories *easy*, *frequent*, *variable* and *hard*.

in the rest of this paper, which corresponds to the recall for $k \to \infty$. The evaluation tool performs two types of evaluations, a *page-centric* evaluation that regards a hit as correct if the seeding page was found, and the *formula-centric* evaluation, which assumes that a hit is correct, if a formula with exactly the same TEX input was found. To avoid over-fitting, only aggregated results are displayed as feedback to the participants. Thus the participants get feedback on how their systems performed on average for all topics, but they do not know how the systems performed on an individual topic or on a topic category. We observed that the intermediate feedback feature was highly appreciated by the participants, because it helped them to identify and fix bugs in their software. We observed that participants submitted 3 to 5 times until they were satisfied with the results. Some participants submitted subsequent runs under different names. This justifies the reports by the participants that the submission system helped to improve MIR systems. Some teams have improved their mrr by 50% or more.

5. PARTICIPANTS AND EVALUATION RESULTS

We had seven participants from five countries and in total 56 runs with about 2 million hits. The results of the evaluation described in the former section are listed in Table 1. A detailed overview of the participants and their MIR approaches can be found in [3]. The best result with regard to success was submitted by TUW and NII. Both runs found 97% of the topics according to the page-centric evaluation. With mrr = 82% the ranking of TUW is slightly better compared to Nii (74.5%). For the formula-centric evaluation Nii has the best success with 94% and mrr = 82%. The best TUW run achieves a mrr value of 88% at a sucess rate of 93%. Table 1 shows that there is a high correlation between the topic category and the system performance. Furthermore, all teams that submitted more than one run got very good results for the easy topics. The difference between page and formula exact evaluation is not significant.

As shown in Figure 3, the performance results from different teams vary more than different runs submitted by the same team. The MIaS team runs (circles) show the well known behavior that mrr (or precision) decays with growing success (or recall respectively).

We observe that all topics except one were found by two or more participants, even if the formula exact evaluation method is used. The known good result for query 99 $\frac{?x_0}{?x_1}$ that can be verbalized as "any fraction", was found only by one team at rank 8983. More interesting is that 4 teams assigned a high rank to the result $\frac{\frac{x}{y}}{\frac{y+z}{y}}$ from the Harmonic progression page in contrast to the first mathematical expression $\frac{1}{2}$ from the Wikipedia article titled fraction that was not ranked very high.

6. CONCLUSION

We discussed the NTCIR-11 Math Wikipedia Task that lowers the entrance barrier for new participants to Math Information Retrieval and broadens the scope of NTCIR Math tasks to encyclopedic applications. We presented three main technology contributions, integrated in the MediaWiki MathSearch extension. First, we developed methods to convert Wikipedia dumps (in any language) to the main task data format including content and presentation MathML. Second, we developed a method for automated search pattern generation with example hits. Third, our extension provides a fully automated evaluation framework with real time feedback at submission time and a comparative evaluation for multiple submissions including hit pooling.

For the future, we plan to continue development and improvement of the platform with regard to the following aspects. We will allow for user feedback for results with regard to relevance for the entries submitted by the systems. While it's questionable that volunteers can be found to evaluate the results, participants can evaluate their own results which will be helpful for system tuning. Furthermore, we will display some basic similarity scores for each hit and will allow users to create their own search topics. The queries used for NTCIR-11 will stay available for training and testing. A query exact feedback will be displayed for new submissions for the old topics. We intend to attract more participants and found a Math Search interest group made of mathematicians, scientists and people from the traditional information retrieval community. Due to the continuously available por-

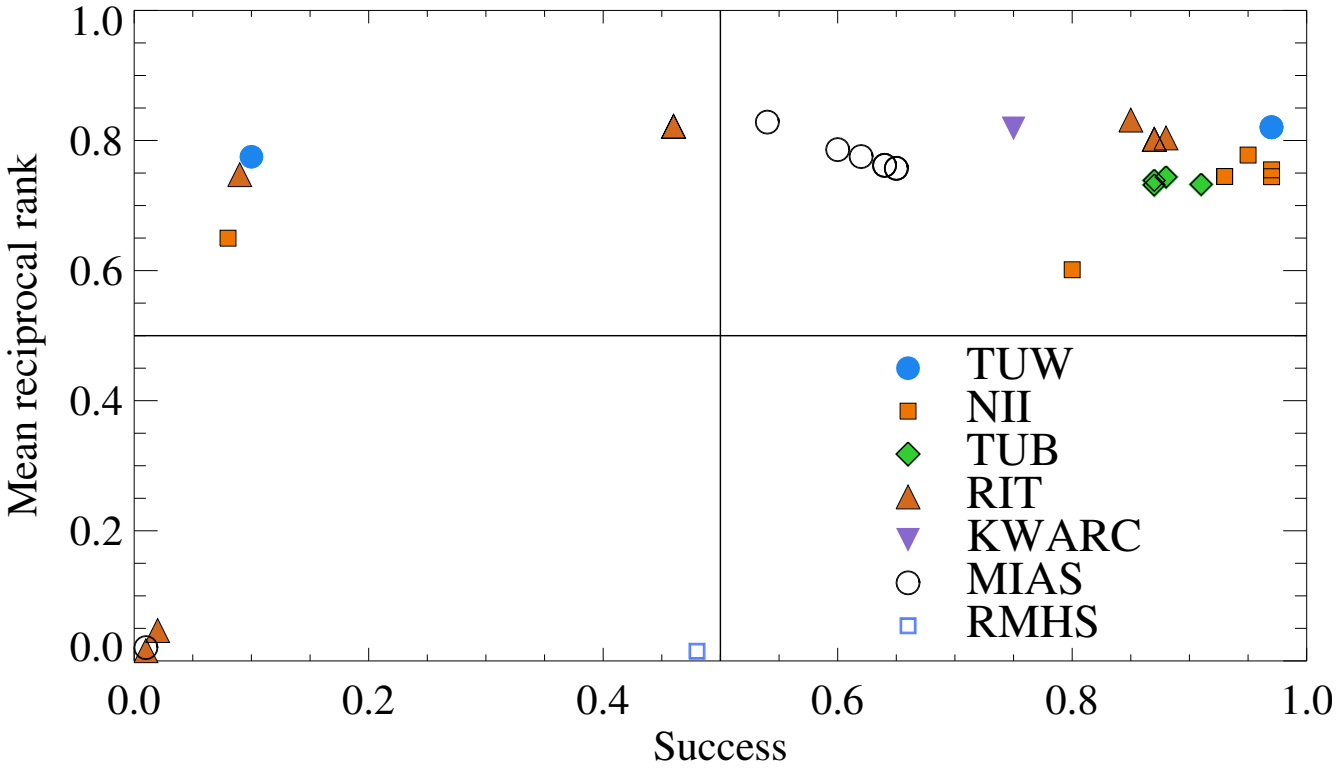

Figure 3: Page-centric evaluation: The Figure shows all runs with regard to mean reciprocal rank and sucess. For example, the runs of the MIAS team show a typical trade-off between mean reciprocal rank and success. Other teams reported that they used the automated feedback from the submission system to fix implementation problems. This increased mrr and sucess at the same time.

tal, participants will be able to test new features whenever they are ready. We will publish new queries on demand, and synchronize participants with the NTCIR events.

Acknowledgments. The authors acknowledge the help and support of Akiko Aizawa, Michael Kohlhase, Goran Topic and Marcus Leich. This work has been supported through grants by the German Science Foundation (DFG) as FOR 1306 Stratosphere and by the German Ministry for Education and Research as Berlin Big Data Center BBDC (funding mark 01IS14013A).

7. REFERENCES

[1] Formats for topics and submissions for the math2 task at ntcir-11. Technical report, NTCIR, 2014.

[2] Akiko Aizawa, Michael Kohlhase, and Iadh Ounis. NTCIR-10 Math Pilot Task Overview. In *Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies*, pages 654–661, Tokyo, Japan, 2013.

[3] Akiko Aizawa, Michael Kohlhase, Iadh Ounis, and Moritz Schubotz. NTCIR-11 Math-2 Task Overview. In *Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies*, pages 88–98, 2014.

[4] Michael Kohlhase, Helena Mihaljevic-Brandt, Wolfram Sperber, and Olaf Teschke. Mathematical Formula Search. pages 56–57, September 2013.

[5] Michael Kohlhase, Corneliu Prodescu, and Christian Liguda. Xlsearch: A search engine for spreadsheets. In Simon Thorne et. al, editor, *Proceedings of the EuSpRIG 2013 Conference "Spreadsheet Risk Management". July 4-5, London, United Kingdom*, pages 47–58. Five Star Printing Ldt, Claydon, 2013.

[6] Matthias S. Reichenbach, Anurag Agarwal, and Richard Zanibbi. Rendering expressions to improve accuracy of relevance assessment for math search. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, pages 851–854, 2014.

[7] Moritz Schubotz and Gabriel Wicke. Mathoid: Robust, scalable, fast and accessible math rendering for wikipedia. In Stephen Watt et al., editor, *Intelligent Computer Mathematics*, volume 8543 of *Lecture Notes in Computer Science*, pages 224–235. Springer International Publishing, 2014.

[8] Heinrich Stamerjohanns, Michael Kohlhase, Deyan Ginev, Catalin David, and Bruce Miller. Transforming large collections of scientific publications to xml. *Mathematics in Computer Science*, 3(3):299–307, 2010.

[9] Ellen M. Voorhees. The TREC-8 Question Answering Track Report. *TREC*, 1999.

[10] Keita Del Valle Wangari, Richard Zanibbi, and Anurag Agarwal. Discovering real-world use cases for a multimodal math search interface. *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval - SIGIR '14*, pages 947–950, 2014.