NTCIR-11 Math-2 Task Overview

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

Michael Kohlhase
Jacobs University Bremen
m.kohlhase@jacobs-university.de

Iadh Ounis
University of Glasgow
Iadh.Ounis@glasgow.ac.uk

Moritz Schubotz
Technische Universität Berlin
schubotz@tu-berlin.de

ABSTRACT

This paper presents an overview of the NTCIR-11 Math-2 Task, which is specifically dedicated to information access to mathematical content. In particular, the paper summarizes the task design, analysis of the submitted runs, and the main approaches deployed by the participating groups. It also contains an introduction to the optional free Wikipedia subtask, a newly introduced mathematical retrieval task using Wikipedia articles.

Team Name

MATH-2

Subtasks

Math-2 Main Task (English)
Optional Math-2 Wikipedia Subtask (English)

Keywords

information access to mathematical content, MathML

1. INTRODUCTION

The NTCIR-11 Math-2 Task develops an evaluation test collection for mathematical formula/keyword search with the aim of facilitating and encouraging research in mathematical information retrieval (MIR) and its related fields.

MIR is search for a particular mathematical concept, object, or result, often expressed using mathematical formulae, which – in their machine readable forms – are expressed as complex expression trees. To answer MIR queries, a search system should tackle at least (either of the) two challenges: (1) tree structure search, and (2) utilization of textual context information.

Mathematical formulae are important means for dissemination and communication of scientific information. They are not only used for numerical calculation but also for clarifying definitions or disambiguating explanations that are written in natural language.

Despite the importance of Math in technical documents, most of the contemporary information retrieval systems do not support users' access to mathematical formulae in target documents. One major obstacle for the research is the lack of readily available large-scale datasets with structured mathematical formulae, carefully designed tasks, and established evaluation methods.

Motivated by the current situation, the NTCIR-10 Math Pilot Task [2] was our initial attempt to develop a common workbench for mathematical formula search. The task was completed successfully as an initial pilot task with 15 registrations and 6 submissions, showing a clear interest in the task. The pilot task featured a “formula search task” with pure formula search, formula/keyword search, and free-form question search tasks as well as a “math understanding task” in anticipation that semantics-oriented MIR systems would have to extract semantics from the formula context and use that in query answering.

In the NTCIR-11 Math-2 Task, we continue to pursue our initial goal of creating a shared evaluation platform for an active and emerging community in Math IR. We incorporate the experiences from NTCIR-10 Pilot task into the design of the NTCIR-11 Math-2 task, which is a traditional adhoc retrieval task with formula/keyword queries, as simple formula queries simple for expression-oriented approaches and free-form queries “math understanding” were deemed infeasible for current systems.

Given a query representing a target made of mathematical formulae schemata and keywords, participating systems return ranked lists of relevant retrieval units containing a formula matching the query. Apart from the compulsory main task, Math-2 also provides an open free subtask using Wikipedia math-related articles. The optional Wikipedia subtask complements the main MIR task with an automated performance evaluation.

In the following sections, we describe in details the task documents, the topic development, and the pooling and assessment procedures for the main task. We also briefly describe the analysis of the submitted runs and the main approaches deployed by the participating groups. We also introduce the details of optional free Wikipedia subtask, including the participating runs and the result summary.

2. PARTICIPATION

Table 1 shows the eight groups that submitted their results to the NTCIR-11 Math-2 Task. KWARC and MCAT are the organizers' groups.

All eight participating teams contributed to the Math Retrieval main task. Each group could submit up to four runs. Table 2 shows the number of runs submitted by each group. Out of the eight groups, 4 teams joined the optional Wikipedia subtask.

Compared to the NTCIR-10 Math Pilot task, all participating teams submitted to the main task. We attribute

Table 1: NTCIR-11 Math-2 Task Participants.

#	Group ID	Organization in English	from
1	ICST	Peking University	CN
2	IFISB	TU Braunschweig	DE
3	FSE	TU Berlin	DE
4	KWARC	Jacobs University Bremen	DE
5	MCAT	National Inst. of Informatics	JP
6	MIRMU	Masaryk University	CZ
7	RIT	Rochester Inst. of Technology	US
8	TUW-IMP	Vienna Univ. of Technology	AT

this to the fact participants were much better prepared to handling math content in this year an had early access to the data set. It is worth noting that NTCIR-11 attracted four new teams, which shows that a Math Task fills a need otherwise unaddressed.

Table 2: Number of runs for each subtask category.

Group ID	Main task	Wikipedia subtask
FSE	1	5
ICST	1	0
IFISB	1	0
KWARC	1	1
MCAT	4	0
MIRMU	4	1
RIT	4	1
TUW-IMP	4	0
Total	20	8

3. MAIN TASK DESIGN

3.1 Document Set

The source dataset of NTCIR-11 Math-2 consists of about 105,120 scientific articles (in English) that were converted from LaTeX to an HTML+MathML-based format by the KWARC project (http://kwarc.info/). We selected articles from the arXiv categories math, cs, physics:math-ph, stat, physics:hep-th, physics:nlin to get a varied sample of mathematical/technical documents.

Initially, we planned to re-use the NTCIR-10 dataset, except for the changes affecting the new retrieval units. However, we decided to regenerate the dataset using the latest version of the conversion tool (LaTeXML [11]) to improve the consistency and the quality of math formulae in the dataset. Based on this, Math-2 dataset contains documents resampled from the same arXiv dataset.

For NTCIR-11 Math-2, each document in the corpus is divided into paragraphs. The NTCIR-10 pilot task showed that retrieving full – 15-page on average – documents made the evaluation of results very difficult without hit identifiers. Therefore we decided to segment the papers into paragraphs and use those as return units. This resulted in a total of 8,301,578 search units with about 60 million math formulae including monomial expressions. Each search unit is stored independently as a file, in both HTML5 and XHTML5 formats to cater to system preferences. The dataset is about 174GB uncompressed.

3.2 Topic Development

For the NTCIR-11 Math-2 Task, we collected 50 formula/keyword queries and distributed the set to the participants in a custom XML format. Given a set of queries, systems returned a ranked list of search results.

A Math-2 query contains: (1) a Topic ID, (2) a Query (formula + key words), but no natural language description. Formula queries are encoded in presentation and content MathML as well as the LaTeX source. A query also contains narrative field, which is a precise description of the user situation and information need and relevancy criteria. This is used only for assessment and was not be included in the query set delivered to participants. The Math-2 task is designed so that all the topics include multiple relevant documents at least a single relevant document in the data set. Details of topics format can be found in [4].

Formula queries may contain named query variables that act as wildcards.

Query variables were included in the Math-2 task even though they are non-standard for information retrieval because they were determined to be an important feature for adequately expressing information needs by the NTCIR-10 pilot and other user studies. We will give an overview to make this paper self-contained.

A query variable with name foo is represented by the XML element <mws:qvar name="foo"/>; we write it as ?foo in LaTeX and presented formulae. $\frac{\mathtt{?f(?v+?d)-?f(?v)}}{\mathtt{?d}}$ is a typical example for a formula query with query variables ?f, ?v, and ?i. It matches the definition

$$g'(cx) = \lim_{h\to 0} \frac{g(cx+h)-g(cx)}{h} \tag{1}$$

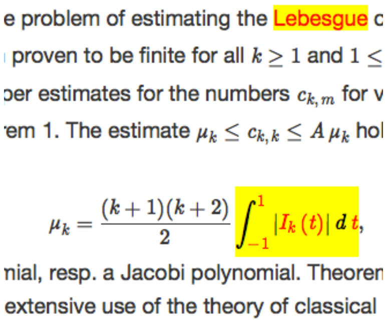
since we can substitute g for ?f, cx for ?v, and h for ?i to obtain the subformula $\frac{g(cx+h)-g(cx)}{h}$ of (1). The subformula matching the query and the substitutions form a "justification" of the match, i.e. information that can be used to verify its adequacy and should be provided as part of the result.

Note that depending on whether we express the query in content or presentation MathML we may obtain different results: presentation MathML distinguishes the variants $\frac{n}{d}$, $n:d$ and n/d of a fraction, while content MathML only sees them as applications of the division function to n and d.

3.3 Submissions

Given a query, participant systems estimate the relevance of the paragraphs in the dataset to the query and return a ranked lists of the retrieval units of the documents containing a formula and/nor keywords matching the query. Each participant could submit up to four runs with 1,000 results per query and run. The results include the score of the returned retrieval units and optionally include supporting evidence (e.g. the formula identifier or the substitution terms for the query variables). Submissions can be either in a plain text format specified by trec_eval evaluation tool or a custom XML format. The former only specifies the result unit and its score, the XML form also allows a full specification of the hit justifications, which assist the evaluators in their decision by highlighting and justifying the results; see Figure 1. To assist result reporting, a submission validation script was distributed to the participants.

Details of results format can be found in [4].



e problem of estimating the Lebesgue c

proven to be finite for all $k \ge 1$ and $1 \le$

er estimates for the numbers $c_{k,m}$ for v

em 1. The estimate $\mu_k \le c_{k,k} \le A\,\mu_k$ hol

$$\mu_k = \frac{(k+1)(k+2)}{2}\int_{-1}^{1}|I_k(t)|\,d\,t,$$

nial, resp. a Jacobi polynomial. Theoren

extensive use of the theory of classical

Figure 1: Justifying and Highlighting Hits

4. MAIN TASK EVALUATION

4.1 Document and Query Statistics

Fig. 2 shows the distribution of the number of math formulae contained in each retrieval unit. The distribution shows that 95% of the retrieval units have 23 math formulae or less, which is sufficiently small for document-based relevance judgment by human reviewers. By introducing relatively small size of retrieval units, Math-2 makes the task feasible for both formula-based search systems and document-based retrieval engines. Fig. 3 summarizes basic statistics for the math formula trees in our dataset. Figs. 3-(a)∼(d) correspond to the distributions of total number of nodes, maximum tree depth, average number of child nodes, total number of leaf nodes in each math formula, respectively. These statistics show that the math trees in the NTCIR-2 Math-2 Dataset approximately follow the pow-law distribution in their size. While there exist a vast amount of relatively simple trees, there also exist a non-negligible number of highly complex trees. This clearly shows that, as a benchmark for tree structure search, Math-2 is characterized by its large scale as well as the heterogeneity of the trees in the dataset.

Table 3 shows the basic statistics for the 50 topics used in the task. Each topic includes at least one keyword and one formula. The number of query variables per topic varies from 0 (none) to 8. There are relatively small number of topics without any query variables in the NTCIR-2 Math-2 Dataset.

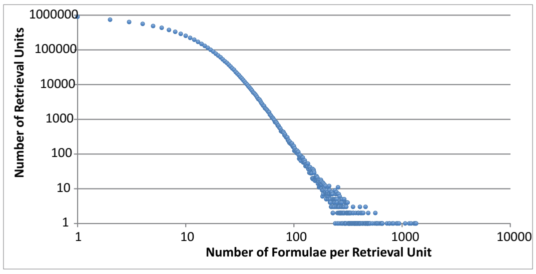


Figure 2: Distribution of the total number of formulae per retrieval unit.

4.2 Pooling

Eight participating groups submitted their results. Out of the total 20 submitted runs, 4 runs include additional “justification information” as was defined in 3.3.

The evaluation of Math-2 main task was pooling-based: First, all the submitted results were converted into a trec_eval result file format. Next, for each topic, 50 retrieval units were selected from the union of the returned results. Then, the selected units were assessed by human reviewers.

Considering the limited resources for human evaluation, we selected formulae for the assessment as evenly as possible from all the runs based on the ranking orders in the individual submitted files. For each iteration of the sampling, the current top-ranked formulae were taken from all the ranking lists, and added to the pool if they were not already on it. This process was repeated until the total size of the pool becomes equal to 50. The order of sampling in each iteration was decided based on the total assessment counts previously assigned to each group so far in order to put higher priority to groups with less number of runs.

Fig. 4 shows the statistics for the pooling. Fig. 4-(a) is the number of assessments assigned to each run. Fig. 4-(b) is the coverage of human assessment of top 5, 10 and 15 ranked retrieval units for each run. These values were low for FSE team since the results were not included in the pool. For other runs, the assessment coverage of top 5 ranked units was almost 100%, while the value was about 60∼70% and 30∼45% for top 10 and top 15, respectively. Based on the statistics, we included top 5 and 10 precision as accuracy measures in our evaluation.

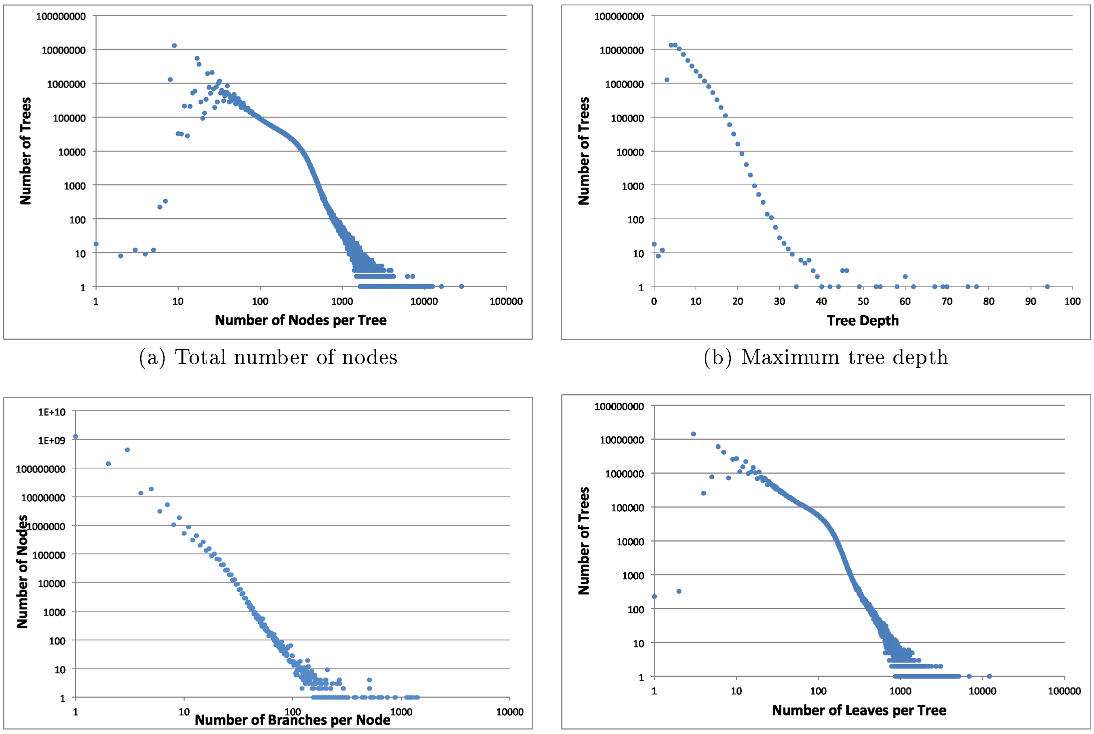
4.3 Human Assessment

After the pooling process, the selected 50 retrieval units per topic were fed into the SEPIA system [14] with MathML extensions developed by the organizers. Fig. 5 is a screenshot of the SEPIA actually used for the evaluation.

The upper light red box contains information on the target topic, including the keywords and the formulae with query variables (marked as red). Also, the title of the topic, the relevance description, and the example hit (if any) are displayed as supplementary information. Note that the supplementary information was not disclosed to the participants during the task period. The lower-left green box lists the 50 documents selected by the pooling process. Finally, the lower-right white box shows the target retrieval unit with the URL of the original arXiv article.

Evaluators judged relevance of the hit to the query by comparing it to the formulae and their contexts in the retrieval unit. Note that relevance is assessed not on formula basis, which was the case in NTCIR-10, but on retrieval unit basis. When evaluators judged the relevance of each retrieval unit to the query, the keywords, as well as the formulae included as justification in the submission files, were highlighted on the screen to assist the judgment.

To ensure sufficient familiarity with mathematical documents, the evaluators were chosen from third-year and graduate students of (pure) mathematics. For each retrieval unit, the evaluators were asked to select either relevant (R), partially-relevant (PR), or not-relevant (N). Each retrieval unit was assessed by two assessors. Since trec_eval only accepts binary relevance judgment, the scores of the two judges were converted into an overall relevance score using a mapping table shown in Table 4. Apart from these guidelines, no specific instructions were given as to how the relevance was to be judged, so that the evaluators had to rely on their mathematical intuition, the described information need, and the query itself. Due to the wide variety of domains covered



(a) Total number of nodes

(b) Maximum tree depth

(c) Average number of child nodes

(d) Total number of leaf nodes

Figure 3: Math formulae statistics for the Math-2 dataset.

in the data set, the evaluators were often not directly familiar with the specific content of the documents. They self-reported being relatively lenient with formula hits judging them to be partially relevant if there was a considerable overlap in symbols or if the respective keywords were found in the result.

Table 4: Relevance score assignment.

Scores of two judges	Relevant	Partially relevant
R/R	Yes	Yes
R/PR	Yes	Yes
PR/PR, R/N	No	Yes
PR/N	No	Yes
N/N	No	No

The distribution of relevance score for each topic is summarized in Table 5. In the table, total hit is a sum of the number of judged documents for all submitted runs, and uniq ratio is the ratio of documents supported by only a single run in the total 50 judged documents. Based on the judgment statistics, we decided to use all the topics in our evaluation.

4.4 Evaluation Measure

Evaluation measures used in the task were as follows; see [15] for reference:

- MAP: Mean average precision over judgment groups.
- P-5: Precision at rank 5.
- P-10: Precision at rank 10.
- Bpref: Preference-based information retrieval measure for incomplete relevance judgment.

These values were obtained from the output file produced by trec_eval version 9.0 and were labeled as map_avgjg, P_avgjg_5, P_avgjg_10, and R_bpref_avgig, respectively.

To check that our assessment guidelines were sufficiently clear to the assessors, we also calculated the inter-assessor agreement using Fleiss' Kappa Agreement and Pearson Correlation value. The agreement scores, shown in Table 6, showed that the assessment was moderately consistent [7].

Table 6: Inter-annotator agreement.

	Fleiss Kappa agreement	Pearson Correlation value
Relevant	0.544	0.548
Partially relevant	0.578	0.591

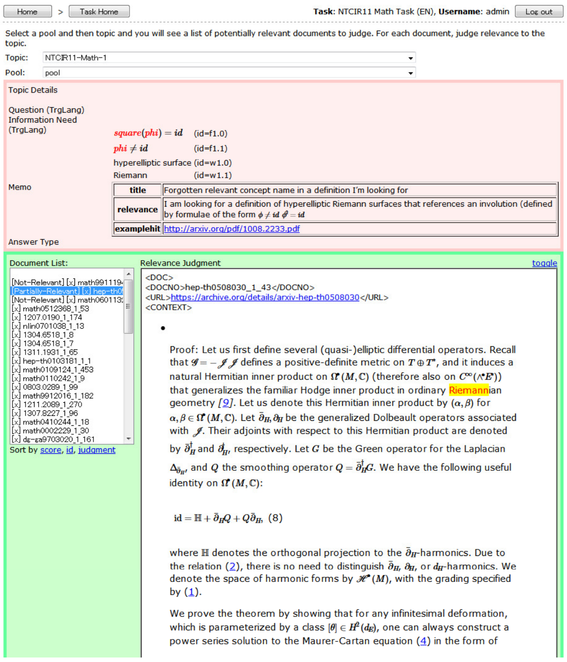
5. MAIN TASK RESULTS

5.1 Outline of the Systems

In this section, we briefly describe the salient features of the approaches deployed by the participating groups in NTCIR-11. These descriptions were contributed by the participating groups. Further details about the deployed approaches could be found in the cited papers below.

Table 3: Query statistics for the Math-2 dataset.

Topic ID	Num of keywords	Num of formulae	Sum of nodes	Max depth	Num of qvar	Topic ID	Num of keywords	Num of formulae	Sum of nodes	Max depth	Num of qvar
NTCIR11-Math-1	2	2	62	7	6	NTCIR11-Math-26	1	1	93	10	3
NTCIR11-Math-2	2	1	73	7	1	NTCIR11-Math-27	4	1	66	8	0
NTCIR11-Math-3	2	1	42	7	2	NTCIR11-Math-28	3	1	17	5	0
NTCIR11-Math-4	3	1	45	7	1	NTCIR11-Math-29	3	1	69	7	2
NTCIR11-Math-5	3	1	105	10	1	NTCIR11-Math-30	3	1	24	6	1
NTCIR11-Math-6	3	1	66	11	6	NTCIR11-Math-31	3	1	44	9	2
NTCIR11-Math-7	4	1	38	6	6	NTCIR11-Math-32	2	1	52	7	2
NTCIR11-Math-8	2	1	47	8	2	NTCIR11-Math-33	2	1	76	8	2
NTCIR11-Math-9	2	1	83	9	6	NTCIR11-Math-34	2	1	126	9	3
NTCIR11-Math-10	4	1	65	10	2	NTCIR11-Math-35	2	1	57	9	2
NTCIR11-Math-11	1	1	65	9	0	NTCIR11-Math-36	2	1	85	9	1
NTCIR11-Math-12	2	1	16	5	0	NTCIR11-Math-37	2	1	33	7	4
NTCIR11-Math-13	2	1	14	5	0	NTCIR11-Math-38	2	1	70	8	3
NTCIR11-Math-14	2	1	51	8	1	NTCIR11-Math-39	1	1	58	8	2
NTCIR11-Math-15	2	1	22	6	1	NTCIR11-Math-40	2	1	98	7	3
NTCIR11-Math-16	1	1	26	7	2	NTCIR11-Math-41	2	1	84	8	2
NTCIR11-Math-17	2	1	41	7	3	NTCIR11-Math-42	1	1	47	8	2
NTCIR11-Math-18	3	1	48	7	1	NTCIR11-Math-43	2	1	68	8	0
NTCIR11-Math-19	2	1	34	7	3	NTCIR11-Math-44	1	2	70	6	8
NTCIR11-Math-20	2	1	37	7	1	NTCIR11-Math-45	1	1	36	8	0
NTCIR11-Math-21	2	1	94	9	3	NTCIR11-Math-46	2	1	258	15	0
NTCIR11-Math-22	2	1	33	6	2	NTCIR11-Math-47	2	1	78	9	0
NTCIR11-Math-23	3	1	29	5	0	NTCIR11-Math-48	3	4	177	8	4
NTCIR11-Math-24	3	1	40	8	3	NTCIR11-Math-49	2	1	150	12	3
NTCIR11-Math-25	4	1	107	9	2	NTCIR11-Math-50	3	1	66	8	1



Task: NTCIR11 Math Task (EN), **Username**: admin

Select a pool and then topic and you will see a list of potentially relevant documents to judge. For each document, judge relevance to the topic.

Topic: NTCIR11-Math-1

Pool: pool

Topic Details

Question (TrgLang)
Information Need (TrgLang)

$square(phi) = id$ (id=f1.0)

$phi \neq id$ (id=f1.1)

hyperelliptic surface (id=w1.0)

Riemann (id=w1.1)

Memo

title	Forgotten relevant concept name in a definition I'm looking for
relevance	I am looking for a definition of hyperelliptic Riemann surfaces that references an involution (defined by formulae of the form $\phi \neq id$ $\phi^2 = id$
examplehit	http://arxiv.org/pdf/1008.2233.pdf

Answer Type

Document List:

[Not-Relevant] [x] math991119
[Partially-Relevant] [x] hep-th0
[Not-Relevant] [x] math060113
[x] math0512368_1_53
[x] 1207.0190_1_174
[x] nlin0701038_1_13
[x] 1304.6518_1_8
[x] 1304.6518_1_7
[x] 1311.1931_1_65
[x] hep-th0103181_1_1
[x] math0109124_1_453
[x] math0110242_1_9
[x] 0803.0289_1_99
[x] math9912016_1_182
[x] 1211.2089_1_270
[x] 1307.8227_1_96
[x] math0410244_1_18
[x] math0002229_1_30
[x] dg-ga9703020_1_161

Sort by score, id, judgment

Relevance Judgment

```
<DOC>
<DOCNO>hep-th0508030_1_43</DOCNO>
<URL>https://archive.org/details/arxiv-hep-th0508030</URL>
<CONTEXT>
```

- Proof: Let us first define several (quasi-)elliptic differential operators. Recall that $\mathscr{G} = -\mathscr{J}\mathscr{J}$ defines a positive-definite metric on $T \oplus T^*$, and it induces a natural Hermitian inner product on $\Omega^\bullet(M,\mathbb{C})$ (therefore also on $C^\infty(\wedge^\bullet E^*)$) that generalizes the familiar Hodge inner product in ordinary Riemannian geometry [9]. Let us denote this Hermitian inner product by (α,β) for $\alpha,\beta \in \Omega^\bullet(M,\mathbb{C})$. Let $\bar{\partial}_H, \partial_H$ be the generalized Dolbeault operators associated with $\mathscr{J}$. Their adjoints with respect to this Hermitian product are denoted by $\bar{\partial}_H^\dagger$ and $\partial_H^\dagger$, respectively. Let G be the Green operator for the Laplacian $\Delta_{\bar{\partial}_H}$, and Q the smoothing operator $Q = \bar{\partial}_H^\dagger G$. We have the following useful identity on $\Omega^\bullet(M,\mathbb{C})$:

$$\mathrm{id} = \mathbb{H} + \bar{\partial}_H Q + Q\bar{\partial}_H, \quad (8)$$

where $\mathbb{H}$ denotes the orthogonal projection to the $\bar{\partial}_H$-harmonics. Due to the relation (2), there is no need to distinguish $\bar{\partial}_H$, ∂_H, or d_H-harmonics. We denote the space of harmonic forms by $\mathscr{H}^\bullet(M)$, with the grading specified by (1).

We prove the theorem by showing that for any infinitesimal deformation, which is parameterized by a class $[\theta] \in H^2(d_E)$, one can always construct a power series solution to the Maurer-Cartan equation (4) in the form of

Figure 5: Evaluation Screen Snapshot in SEPIA.

5.1.1 FSE (TU Berlin/NIST); see [13]

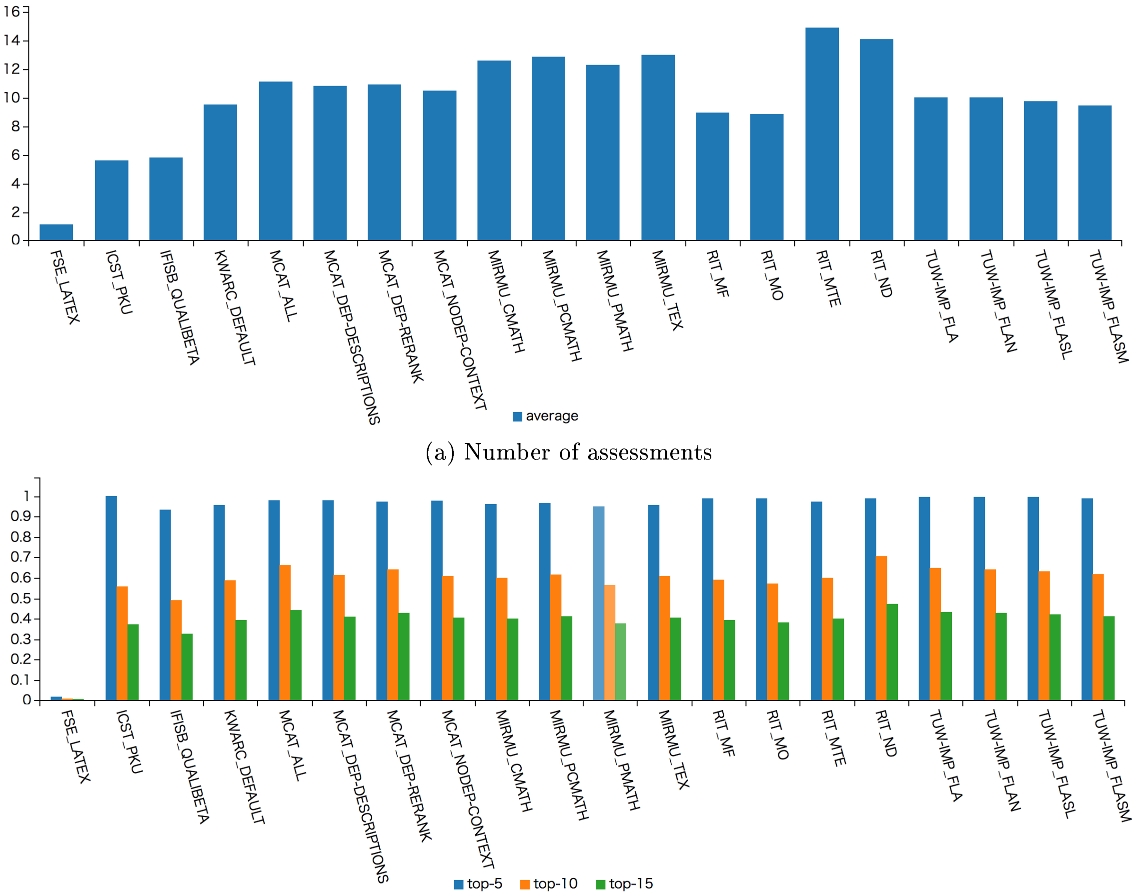
For NTCIR-11, the FSE team focused on the Wikipedia Subtask and the evaluation of Math Similarity factors based on the results from the pooling process. One sub-team (the XQueryBenchmark group) compared the result sets and execution times for different XQuery execution engines. The XQuery expressions were generated from the content MathML provided by the Wikipedia Subtask. The XQuery-Benchmark group noted significant differences in the execution times, but more or less similar result sets. The other sub-team (the Similarity Search group) focused on fundamental research in the area of Math Similarity search. The human annotated ground truth generated by the main task enabled us to verify the significance of fundamental math similarity factors.

5.1.2 ICST (Peking University); see [8]

The ICST system aims at searching for mathematical formulae based on both textual and spatial similarities. The system consists of a tree constructor, a tokenizer, an indexer and ranker. Presentation MathML is used as the formula format here. A tree constructor extracts a layout presentation tree directly from Presentation MathML. The layout presentation is then converted into a corresponding semantic presentation, which is called semantic operator tree, by using the semantic enrichment technique.

The tokenization step includes normalization and term extraction. The goal of normalization is to convert different formulae with same meaning into one uniform format, so as to ensure the high recall of relevant formulae. Term extraction aims at extracting the index terms from the formula. It is well known that generalization is an important behavior when people understand formulae. Additionally, generalization of substructures in low level can avoid overrating for the substructures, which is a common problem occurring in tree-based indexing methods. In order to mimic the formula understanding process of users, a term extraction method with generalization is proposed towards the semantic operator tree of formulae.

Lastly, in indexing and raking, index files for terms and formulae are built on Lucene to calculate the similarity score between query and formulae. Due to the time limitation, the ICST system only dealt with four files of all so that the performance of it is not good. In the future, the team will improve their algorithm and handle all files.



(a) Number of assessments

(b) Coverage at top 5 and 10

Figure 4: Pooling statistics.

5.1.3 IFISB (TU Braunschweig); see [3]

IFIS_QUALIBETA combined a feature-extracted sequence mechanism of the formulae and a sentence level representation of the text describing the formulae to model the collection. The feature-extracted sequences used were: the category of the formulae, the sets of identifiers, constants, and operators. This representation with the text surrounding the formulae were indexed in Elastic Search for query processing.

5.1.4 KWARC (Jacobs University Bremen; see [5])

MathWebSearch (MWS) is a web application that provides low-latency answers to full-text queries which consist of keywords and formulae. MWS front-ends convert formula schemata (with query variables) into content MathML expressions, which the MWS formula indexer answers by unification and combines with keyword results from a text search engine – Elastic Search. The modular architecture and standardized formats makes MWS applicable to a wide range of querying tasks - all, where formulae can be transformed into content MathML. The low-latency characteristic makes MWS well-suited as a back-end for interactive applications, e.g. web-based formula search engines or editing support services. Unification queries form the basis of an expressive, query language with well-defined semantics. As substitution instances of the original query, MWS results are highly significant, if the encoding of data set and search query are adequate - i.e. do not forget or spuriously introduce salient semantic features.

MWS 1.0 focuses on scalability (memory footprint, index persistence), integration of keyword- and formula search, and hit presentation issues. It forms a stable basis for future research into extended query languages and user-interaction issues. The system has been integrated into high-profile math information systems like Zentralblatt Math.

5.1.5 MCAT (National Institute of Informatics); see [6]

MCAT group enabled mathematical expressions searching using queries which contain both formulae and keywords. This group implemented an indexing scheme for mathematical expressions within an Apache Solr (Lucene) database. Mathematical expressions were encoded as a series of factors reflecting both the Presentation MathML tree structure and specific symbols they uses. To capture the meaning of each mathematical expression in natural language, each expression was accompanied by two types of automatically extracted textual information, namely words in context window and descriptions. Subsequently, a dependency graph approach was utilized to tackle the encountered lack-of-descriptions issue. This group also proposed a reranking

Table 5: Relevance judgment statistics.

Query	Relevance score					Total	Total	Uniq
ID	4	3	2	1	0	judged	hit	ratio
NTCIR11-Math-1	1	8	10	7	24	50	204	0.16
NTCIR11-Math-2	2	0	1	17	30	50	163	0.30
NTCIR11-Math-3	0	1	5	6	38	50	119	0.32
NTCIR11-Math-4	1	1	6	6	36	50	222	0.14
NTCIR11-Math-5	3	6	18	11	12	50	178	0.28
NTCIR11-Math-6	8	9	5	11	17	50	164	0.26
NTCIR11-Math-7	0	9	1	7	33	50	145	0.30
NTCIR11-Math-8	0	5	4	4	37	50	152	0.24
NTCIR11-Math-9	0	0	2	15	33	50	137	0.26
NTCIR11-Math-10	6	1	7	13	23	50	145	0.26
NTCIR11-Math-11	7	4	1	19	19	50	214	0.22
NTCIR11-Math-12	4	8	7	8	23	50	213	0.16
NTCIR11-Math-13	3	6	4	3	34	50	196	0.18
NTCIR11-Math-14	5	5	4	7	29	50	196	0.20
NTCIR11-Math-15	14	0	2	2	32	50	199	0.16
NTCIR11-Math-16	3	7	2	5	33	50	135	0.24
NTCIR11-Math-17	8	7	1	2	32	50	199	0.18
NTCIR11-Math-18	24	4	0	4	18	50	289	0.16
NTCIR11-Math-19	16	8	5	4	17	50	209	0.18
NTCIR11-Math-20	8	5	3	2	32	50	263	0.12
NTCIR11-Math-21	2	5	7	11	25	50	155	0.22
NTCIR11-Math-22	21	3	7	9	10	50	201	0.22
NTCIR11-Math-23	9	17	14	3	7	50	280	0.12
NTCIR11-Math-24	9	11	2	6	22	50	238	0.24
NTCIR11-Math-25	2	3	3	3	39	50	182	0.24
NTCIR11-Math-26	0	12	3	8	27	50	210	0.18
NTCIR11-Math-27	6	2	3	15	24	50	194	0.30
NTCIR11-Math-28	16	8	5	3	18	50	358	0.18
NTCIR11-Math-29	2	10	5	15	18	50	194	0.22
NTCIR11-Math-30	31	4	0	5	10	50	258	0.18
NTCIR11-Math-31	0	5	7	4	34	50	154	0.22
NTCIR11-Math-32	0	2	13	5	30	50	203	0.22
NTCIR11-Math-33	4	1	4	5	36	50	193	0.28
NTCIR11-Math-34	0	7	8	5	30	50	209	0.22
NTCIR11-Math-35	1	2	6	4	37	50	193	0.22
NTCIR11-Math-36	10	4	4	5	27	50	248	0.22
NTCIR11-Math-37	9	5	3	10	23	50	224	0.16
NTCIR11-Math-38	3	10	4	16	17	50	223	0.16
NTCIR11-Math-39	2	6	11	11	20	50	176	0.24
NTCIR11-Math-40	3	3	13	9	22	50	229	0.22
NTCIR11-Math-41	1	1	26	10	12	50	251	0.18
NTCIR11-Math-42	3	2	15	16	14	50	229	0.20
NTCIR11-Math-43	2	3	5	15	25	50	183	0.24
NTCIR11-Math-44	4	4	5	5	32	50	151	0.22
NTCIR11-Math-45	7	5	6	6	26	50	227	0.16
NTCIR11-Math-46	2	9	2	4	33	50	174	0.22
NTCIR11-Math-47	8	5	2	8	27	50	227	0.26
NTCIR11-Math-48	2	1	4	32	11	50	215	0.24
NTCIR11-Math-49	0	0	5	11	34	50	151	0.20
NTCIR11-Math-50	12	5	13	1	19	50	213	0.18
Total	301	233	304	379	1,283	2,500	10,085	0.21

method which was applied to the initially retrieved expressions. The experiments were performed to examine the effectiveness of the two types of textual information (words in context window and descriptions), dependency graph, and reranking method. The results showed that the use of descriptions and dependency graph together gave higher ranking performances (statistically significant) than the use of context window. Furthermore, it was also shown that the use of descriptions and dependency graph together with the context window delivered even better results. Finally, the results also indicated that the reranking method effectively improved the ranking performances.

5.1.6 MIRMU (Masaryk University); see [10]

Math Information Retrieval group (MIRMU) at the Masaryk University in Brno, http://mir.fi.muni.cz has used second generation of scalable full text search engine Math Indexer and Searcher (MIaS)with attested state-of-the-art information retrieval techniques implemented as MIaSMath extension above standard Lucene/Solr engine. This allows smooth integration of textual and math queries. An approach is the similarity search – as opposed to exact, strictly unification-based search – based on MathML Canonicalization and structural and semantical similarity approximations computed at indexing time. The system is complemented with novel WebMIaS interfaceand query expansion strategies.The system ranked first in four of six Math task metrics, and second in two metrics. The analysis of the evaluation results shows that the system performs best using TEX queries that are translated and canonicalized to Content MathML.

5.1.7 RIT (Rochester Institute of Technology); see [12]

The Tangent system created at RIT (USA) uses two indices: 1) a Solr/Lucene-based index for document text , and 2) a MySQL index for math expressions. When indexing a math expression, its Presentation-MathML representation is converted to a tree which is then tokenized into a set of symbol pairs. Given a query containing math and text, the respective indices are searched and rank scores are combined using a weighted linear combination. For the main task, four runs were submitted, where each run contained different text and math weights (0%/100%, 5%/95%, 25%/75%, and 50%/50%). As text weight increased, Precision@k values increased. A possible reason for the improvement is that the system has limited support for wildcards, and so queries that contain an expression with many wildcards may have benefited from a higher text weight. Additional factors are that hits pertinent to the query that do not contain the query keywords are ignored or ranked lower for the "math only" and "math emphasized" weight combinations, and sometimes irrelevant hits may contain a similar or identical expression to that in the query (especially in the presence of many wildcard symbols). For the Wikipedia subtask, a single run with a text weight of 0 was submitted. Since queries in the Wikipedia task have fewer wildcards and the rank at which the original query document was returned was used for evaluation, this run seems to have performed better than those submitted for the main task. Areas for improvement include reducing index size and retrieval time for the math index. In the main task, Tangent produced the highest Precision@5 for relevant and partially relevant results in the main task, and the strongest specific-item-recall rates for the Wikipedia subtask.

5.1.8 TUW-IMP (Vienna University of Technology); see [9]

The TUW-IMP team created four indices: one for the text, and three for the formulas in the NTCIR-Math 2 test collection. The text in the collection was processed using a non-aggressive stemmer, the English Minimal Analyser, built on top of Lucene's standard tokenisers and filters. To extract and process the formulas in the Math collection regular expressions were used, as the XML parser tried gave many parsing errors. Three types of tokens were extracted from the formulas: the literal tokens (like 'L', 'γ'), and subformula tokens, obtained from a linearized form of the formula tree structure. From the topics given by the track organizers we created several queries, one for each of the indices we created. The query using the text keywords was enriched with hyponyms previously extracted from the collection. The topic formulas were processed in the same manner as the formulas in the collection, where the 'qvar' tag was replaced, sequentially, by semantic MathML symbols (e.g. 'apply', 'ci'). The retrieval model used was Lucene's BM25.

5.2 Evaluation Results

Table 7 summarizes the configuration of participating systems. First, all the systems utilized math formulae information in the topic while some systems do not consider keywords. Second, the representation of math formula used as input varied across the systems, and all the format types, LaTeX, Presentation MathML, and Content MathML, were used in the task. Consideration of tree structure as well as query variables of math formulae depends on the design of the system. One group did not directly consider the structural information and two groups did not apply variable match in their search. As for search engine, five groups used general purpose search engine while three constructed of their own.

Table 8 and Figure 6 show the results of all participating runs. The performance values are averaged over all the queries for MAP, P-5, P-10, and Bpref .

6. WIKIPEDIA OPEN SUBTASK

In addition to the regular Math-2 task, there is an optional free Wikipedia subtask that uses the same topic and submission format as the main task. The goal of this task is to develop an evaluation test collection for mathematical formulae search. In contrast to the main task that uses the arXiv dataset, this task uses the English Wikipedia as a test collection. By free subtask, we mean that the submitted results will not be formally evaluated. However, a final judgment, based on oral presentations by the participants during the final event, will be presented. In contrast to arXiv, which provides knowledge for researchers and experts in highly specialized domains, the Wikipedia encyclopedia contains much of the mathematical folklore explained in simple terms. Therefore, the Wikipedia dataset is easier to understand than the arXiv dataset, which may simplify debugging and testing of the participating math search systems. The Wikipedia Subtask is ongoing and is currently experimental. We expect that this task will continuously evolve and improve. For instance in the future we will utilize formulae as retrieval units. See http://ntcir11-wmc.nii.ac.jp for an overview of the current results and also for participation opportunities. In this section we describe dataset development, query generation, and automatic evaluation methodology.

6.1 Query Generation

The Wikipedia Subtask included 100 performance queries. These queries were generated in the following way. Let $1 \leq i \leq 100$.

1. A random English Wikipedia article A_i containing math formulae is chosen.
2. A formula f_i is chosen randomly (with n_i variables) from the article A_i.
3. For every variable v_j with $1 \leq j \leq n_i$, occurring in the formula f_i, a uniform random number $0 < y_j < 1$ is calculated.
4. Given a constant threshold z and the nesting level of the variable l_j, if $y_j > zl_j$, the variable is replaced using the qvar concept.
5. In the formula f_i, K is the set of indices of selected variables, and $k \in K$ with $1 \leq k \leq \#(K)$ gives the element number of variables are selected. Using the qvar concept, the kth replaced variable $u_k := v_{K(k)}$ is expressed as "$?\{xk\}$".

6.2 Automated Evaluation Process

Participants are able to upload their results via a web form. The newly submitted run is evaluated using the algorithm described below and the result are immediately displayed to the participant.

For n_r submitted runs, each run r $(1 \leq r \leq n_r)$ with query i, selects $P(r,i)$ articles and is evaluated in the following way. The function $S(r,i,p(r,i))$ gives the pth ranked article $(1 \leq p(r,i) \leq P(r,i))$. If A_i is an element of $B(r,i) := \{S(r,i,x) : x \in \{1,\ldots,P(r,i)\}\}$, then $X(r,i) = 1$ otherwise $X(r,i) = 0$. The numbers of correct results for each run is given by $R(r) = \sum_{i=1}^{100} X(r,i)$.

In other words, $R(r)$ represents the number of topics for each run r, that contain the initial Wikipedia article A_i at any position in the result set. In addition to that, the evaluation system returns a list of results which take into account the top k returns hits for different values of k. The participants resubmitted their results several times and were able to improve their scores during this process. The final results for the Wikipedia Subtask are to be presented at NTCIR-11.

7. CONCLUSIONS

We have presented an overview of the Math-2 task at NTCIR-11, the first full Math Information Retrieval (MIR) task at an international IR evaluation forum. The Math-2 task was designed building on the experiences from the Math Pilot Task at NTCIR-10. We have developed a new test collection of more than 8 million paragraphs from more than 100.000 mathematics-heavy preprints from http://arxiv.org and a set of 50 mathematical formula/keyword queries. Eight teams have submitted a total of 20 runs with 5000 search results each, from which 2500 were collected for manual assessment by at least two evaluators. Data set, queries, and evaluated results together form a first comprehensive test collection which can be used to evaluate, test, and optimize MIR systems. The test collection will be released by NTCIR in 2015.

The Math-2 task has been very successful in facilitating the formation of a pluri-disciplinary community of researchers interested in the challenging problems underlying Math IR. The increased participation – four new teams participated in NTCIR-11 – shows that interest in MIR systems

Table 7: Summary of configuration of participating systems.

RunID	keywords	math formulae	format			tree structure	query variables	search engine
			LaTeX	Presentation	Content			
FSE_{latex}	yes	yes	yes	yes	yes	yes	yes	no
ICST_{pku}	no	yes	no	yes	no	yes	no	yes
$\text{IFIS}_{QUALIBETA}$	yes	yes	no	no	yes	no	yes	yes
$\text{KWARC}_{default}$	yes	yes	no	no	yes	yes	yes	yes
MCAT	yes	yes	no	yes	no	yes	no	yes
MIRMU_{cmath}	yes	yes	no	no	yes	yes	no[1]	no[2]
MIRMU_{pcmath}			no	yes	yes			
MIRMU_{pmath}			no	yes	no			
MIRMU_{tex}			yes	no	no			
RIT_{mf}	yes	yes	no	yes	no	yes	yes	yes[3]
RIT_{mo}	no							no
RIT_{mte}	yes							yes[3]
RIT_{nd}	yes							yes[3]
TUW-IMP	yes	yes	no	no	yes	yes	yes	yes

*1 Not explicitly; but qvar names were shortened to the unique one letter variables.
*2 MIaS extension was used to general search engine Lucene.
*3 Solr was used just for the text index; a custom mysql based index was used solely for math expressions.

is high, and the diversity of approaches reported at NTCIR shows that research in this field is active. It can be stated with confidence that the systems have progressed considerably since the NTCIR-10 pilot task; adding e.g. full text search capabilities, improving scalability or addressing result ranking in new ways. For new systems, the sheer size of the data set presented quite a challenge, which all systems have mastered eventually.

The design decision of the Math-2 task to exclusively concentrate on formula/keyword queries and use paragraphs as retrieval units made the retrieval task manageable, but has also focused research away from questions like result presentation and user interaction. In particular, few of the systems has invested into further semantics extraction from the data set, and used that in the search process to further address information needs. We feel that this direction should be addressed more in future challenges.

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

Acknowledgments

The work reported here here has been partially supported by the Leibniz association under grant SAW-2012-FIZ_KA-2 and JSPS KAKENHI Grant Numbers 2430062, 25245084, and 14J09896.

We are indebted Deyan Ginev for generating the dataset for the Math Retrieval Subtask and to Deyan Ginev, Radu Hambasan, Fulya Horozal, Mihnea Iancu, Constantin Jucovschi, Andrea Kohlhase, Sentao Miao, Mihai Munteanu, Corneliu Prodescu, Florian Rabe, Jürgen Schefter, Wolfram Sperber, and Adelin Travers for providing formula search queries. We gratefully We gratefully acknowledge the contribution of Goran Topić in administrating a backend system for the entire task and many useful suggestions to the task, and Giovanni Yoko Kristianto, Shunsuke Ohashi, Kazuhiro Hattori and TriAx co. for assisting the task organization. Finally, the outlines of Systems in Section 5.1 were supplied by the authors of the respective papers.

8. REFERENCES

[1] Proceedings of the 11th NTCIR Conference, Tokyo, Japan, 2014.

[2] A. Aizawa, M. Kohlhase, and I. Ounis. NTCIR-10 Math pilot task overview. In N. Kando and K. Kishida, editors, NTCIR Workshop 10 Meeting, pages 1–8, Tokyo, Japan, 2013.

[3] J. M. Gonzalez Pinto, S. Barthel, and W.-T. Balke. QUALIBETA at the NTCIR-11 Math 2 task: An attempt to query math collections. [1].

[4] M. Kohlhase. Formats for topics and submissions for the Math2 task at NTCIR-11. Technical report, NTCIR, 2014.

[5] M. Kohlhase, R. Hambasan, and C.-C. Prodescu. MathWebSearch at NTCIR-11. [1].

[6] G. Y. Kristianto, G. Topić, F. Ho, and A. Aizawa. The MCAT math retrieval system for NTCIR-11 Math track. [1].

[7] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. Biometrics, 33:159–174, 1977.

[8] L. H. Liangcai Gao, Yuehan Wang and Z. Tang. Icst math retrieval system for ntcir-11 math-2 task. [1].

[9] A. Lipani, L. Andersson, F. Piroi, M. Lupu, and A. Hanbury. TUW-IMP at the NTCIR-11 Math-2. [1].

[10] P. S. Michal Růžička and M. Líška. Math indexer and searcher under the hood: History and development of a winning strategy. [1].

[11] B. Miller. LaTeXML: A LaTeX to XML converter. Web Manual at http://dlmf.nist.gov/LaTeXML/.

[12] N. Pattaniyil and R. Zanibbi. Combining tf-idf text retrieval with an inverted index over symbol pairs in

Table 8: Summary of the retrieval performance

	Relevant						
	FSE$_{latex}$	ICST$_{pku}$	IFISB$_{qualibeta}$	KWARC$_{default}$	MCAT$_{all}$	MCAT$_{depdesc}$	MCAT$_{deprerank}$
MAP avg	0.002	0.007	0.028	0.285	0.071	0.064	0.074
P-5 avg	0.012	0.036	0.108	0.500	0.212	0.192	0.208
P-10 avg	0.006	0.018	0.062	0.306	0.124	0.116	0.130
Bpref avg	0.056	0.009	0.058	0.380	0.153	0.147	0.153

	Relevant (continued)						
	MCAT$_{nodepctxt}$	MIRMU$_{cmath}$	MIRMU$_{pcmath}$	MIRMU$_{pmath}$	MIRMU$_{tex}$	RIT$_{mf}$	RIT$_{mo}$
MAP avg	0.051	0.363	0.359	0.307	0.335	0.128	0.103
P-5 avg	0.164	0.568	0.556	0.512	0.540	0.244	0.200
P-10 avg	0.096	0.352	0.348	0.304	0.338	0.142	0.116
Bpref avg	0.133	0.513	0.500	0.462	0.478	0.171	0.153

	Relevant (continued)					
	RIT$_{mte}$	RIT$_{nd}$	TUW-IMP$_{fla}$	TUW-IMP$_{flan}$	TUW-IMP$_{flasl}$	TUW-IMP$_{flasm}$
MAP avg	0.272	0.229	0.036	0.038	0.043	0.035
P-5 avg	0.504	0.384	0.116	0.120	0.128	0.120
P-10 avg	0.286	0.268	0.080	0.074	0.084	0.078
Bpref avg	0.360	0.275	0.086	0.089	0.091	0.086

	Partially relevant						
	FSE$_{latex}$	ICST$_{pku}$	IFISB$_{qualibeta}$	KWARC$_{default}$	MCAT$_{all}$	MCAT$_{depdesc}$	MCAT$_{deprerank}$
MAP avg	0.002	0.020	0.091	0.250	0.092	0.086	0.093
P-5 avg	0.016	0.152	0.460	0.792	0.448	0.416	0.464
P-10 avg	0.008	0.078	0.246	0.530	0.282	0.262	0.282
Bpref avg	0.041	0.027	0.130	0.374	0.201	0.193	0.196

	Partially relevant (continued)						
	MCAT$_{nodepctxt}$	MIRMU$_{cmath}$	MIRMU$_{pcmath}$	MIRMU$_{pmath}$	MIRMU$_{tex}$	RIT$_{mf}$	RIT$_{mo}$
MAP avg	0.077	0.280	0.279	0.255	0.274	0.109	0.097
P-5 avg	0.404	0.872	0.864	0.844	0.848	0.484	0.464
P-10 avg	0.242	0.546	0.554	0.506	0.542	0.292	0.266
Bpref avg	0.179	0.473	0.474	0.454	0.472	0.179	0.174

	Partially relevant (continued)					
	RIT$_{mte}$	RIT$_{nd}$	TUW-IMP$_{fla}$	TUW-IMP$_{flan}$	TUW-IMP$_{flasl}$	TUW-IMP$_{flasm}$
MAP avg	0.286	0.205	0.057	0.061	0.067	0.058
P-5 avg	0.924	0.648	0.336	0.332	0.376	0.348
P-10 avg	0.556	0.460	0.220	0.216	0.238	0.216
Bpref avg	0.460	0.370	0.134	0.135	0.148	0.135

math expressions: The tangent math search engine at NTCIR 2014. [1].

[13] M. Schubotz, A. Youssef, V. Markl, H. Cohl, and J. Li. Evaluation of similarity-measure factors for formulae based on the NTCIR-11 Math task. [1].

[14] sepia: Standard evaluation package for information access systems. https://code.google.com/p/sepia/.

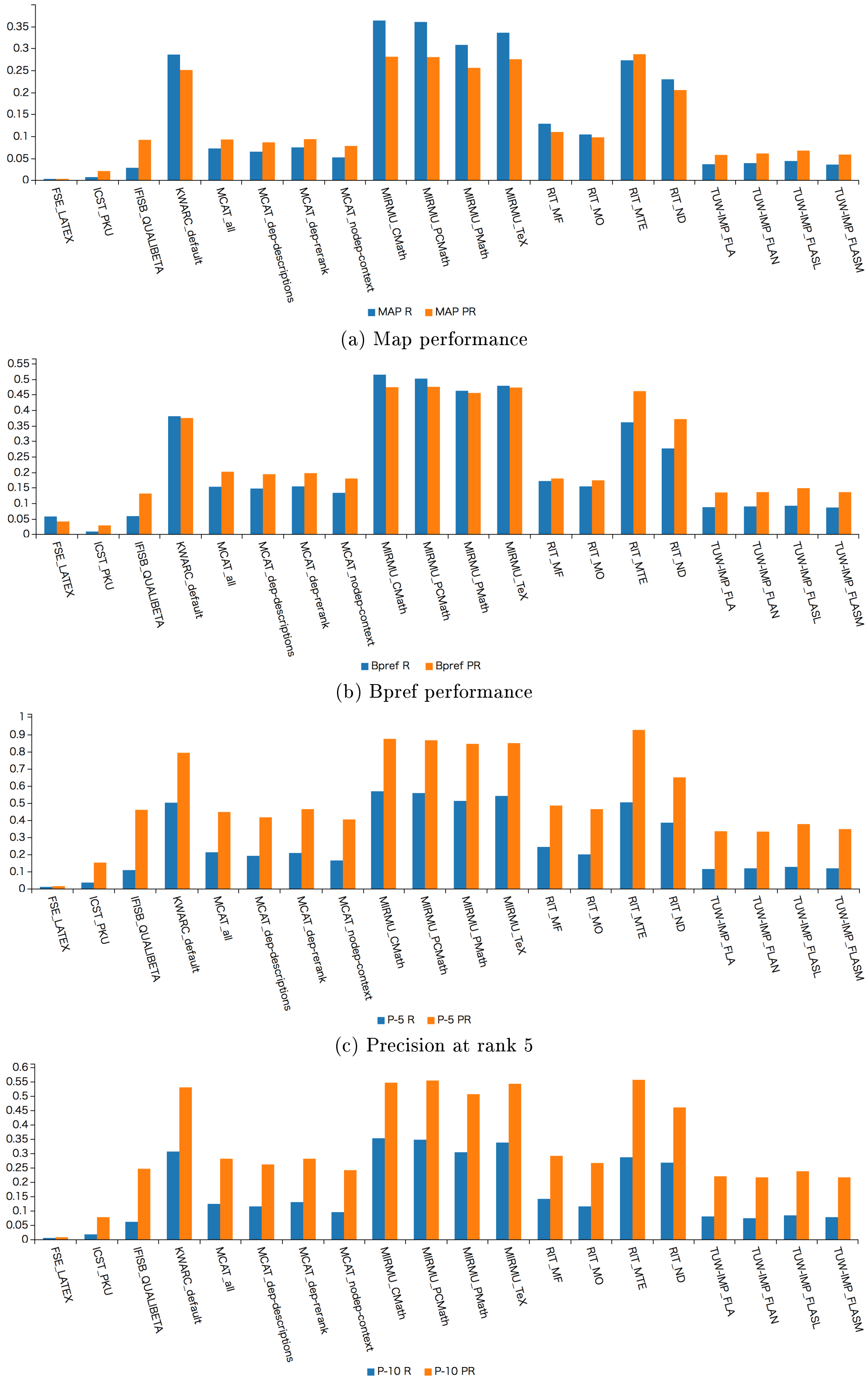
[15] Common evaluation measures. In E. M. Voorhees and L. P. Buckland, editors, The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings, number SP 500-274 in NIST Special Publication, 2007.

APPENDIX

A. TOOLS

We introduce here a list of tools that might be useful in the task. Note that the list is non-exhaustive. Any recommendation from participants is welcome.

- docs2harvest: Tool for parsing html / xhtml documents and generate harvest files with the Content Math data only.
  (https://github.com/KWARC/mws)
- mathml−converter: Tool for converting MathML into keywords.
  (http://code.google.com/p/mathml-converter/)
- MathJax : Javascript to render math formulae on a display.
  (http://www.mathjax.org/)
- LaTeXML: A LaTeX to MathML converter.
  (http://dlmf.nist.gov/LaTeXML/)
- SEPIA: Standard Evaluation Package for Information Access Systems. Used with MathML extension.
  (https://code.google.com/p/sepia/)
- trec_eval: A program to evaluate TREC results.
  (http://trec.nist.gov/trec_eval/)



(a) Map performance

(b) Bpref performance

(c) Precision at rank 5

(d) Precision at rank 10

Figure 6: Summary of NTCIR-11 Math-2 performance.