NTCIR-10 Math Pilot Task Overview

Akiko Aizawa & Michael Kohlhase & Iadh Ounis

http://kwarc.info/kohlhase
Center for Advanced Systems Engineering
Jacobs University Bremen, Germany

NTCIR-10, June 19. 2013

Introduction & Motivation for a Math Pilot Task

Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!

Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- There is a lot of documents with maths
  - there are 120.000 journal articles per year in pure/applied math, 3.5 Million overall
  - 50 million science articles in 2010 [Jin10] with a doubling time of 8-15 years [LvI10]

  And this excludes gray literature, engineering, and school textbooks.
  - Even in the Renaissance, polymaths like Leonardo de Vinci were a rare exception.

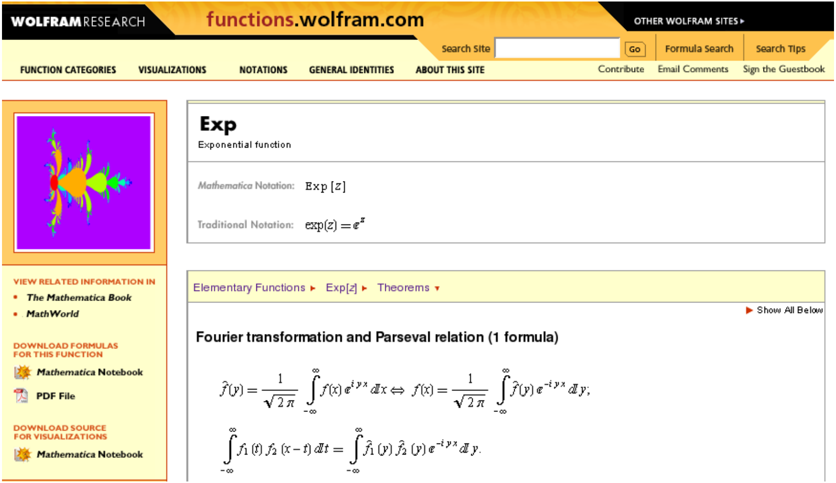
Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- There is a lot of documents with maths
  - there are 120.000 journal articles per year in pure/applied math, 3.5 Million overall
  - 50 million science articles in 2010 [Jin10] with a doubling time of 8-15 years [LvI10]

  And this excludes gray literature, engineering, and school textbooks.
  - Even in the Renaissance, polymaths like Leonardo de Vinci were a rare exception.
- We need IR support to deal with this! ($\rightsquigarrow$ NTCIR-10 Math Pilot Task)

Mathematics Resources on the Web

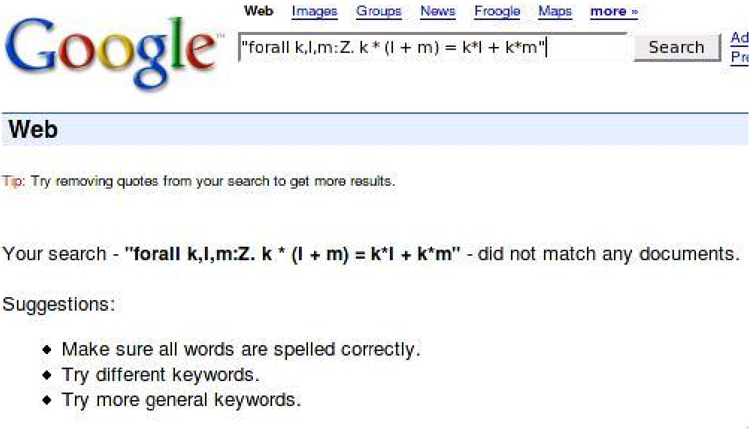
- **Example 1 (The Wolfram Functions Site)** contains $\geq$ 307k Formuae



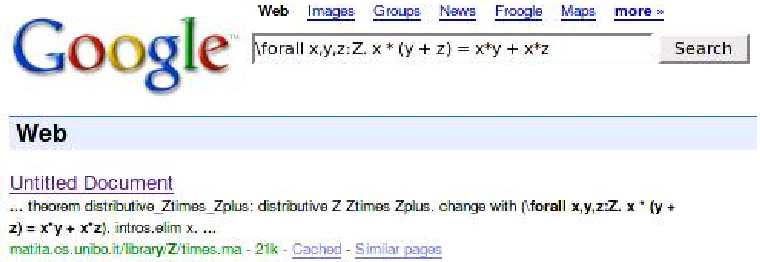
More Mathematics on the Web

- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- ... Engineering Company Intranets, ...
- Question: How will we find content that is relevant to our needs
- Idea: try Google (like we always do)
- Scenario: Try finding the distributivity property for $\mathbb{Z}$ $(\forall k, l, m \in \mathbb{Z}. k \cdot (l + m) = (k \cdot l) + (k + m))$

Searching for Distributivity



Searching for Distributivity



Searching for Distributivity

Of course Google cannot work out of the box

- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables. (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?

Of course Google cannot work out of the box

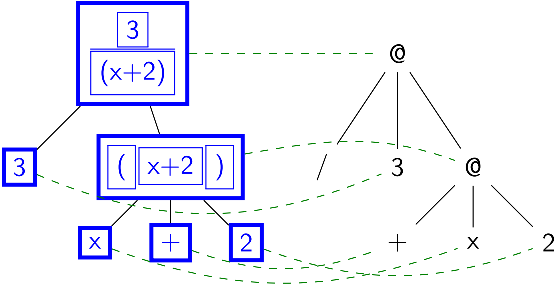
- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables. (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, MG11])
  . . . and works surprisingly well (using Lucene as an indexing engine)
- Idea: Use database techniques (extract metadata and index it)
  Indeed this is done for the Coq/HELM corpus ([AGC+06])
- Idea: Use Automated Reasoning Techniques (Term Indexing [Nor06, KŞ06, KMP12])
- Idea: Use standard IR techniques (Learn from the NTCIR crowd?)

Of course Google cannot work out of the box

- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables. (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, MG11])
  ...and works surprisingly well (using Lucene as an indexing engine)
- Idea: Use database techniques (extract metadata and index it)
  Indeed this is done for the Coq/HELM corpus ([AGC+06])
- Idea: Use Automated Reasoning Techniques (Term Indexing [Nor06, KŞ06, KMP12])
- Idea: Use standard IR techniques (Learn from the NTCIR crowd?)
- Which one is best?: We do not really know, evaluation is very difficult
- Future: maybe even mix/integrate the respective best features (once we know)

Markup Markup e.g. in MathML and LaTeX

- MathML3 is a W3C Recommendation for representing Formulae [ABC+10]
- Idea: Combine the presentation and content markup and cross-reference

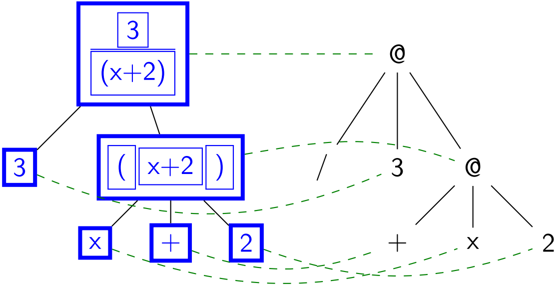


- use e.g. for semantic copy and paste.

  (click on presentation, follow link and copy content)

Markup Markup e.g. in MathML and LaTeX

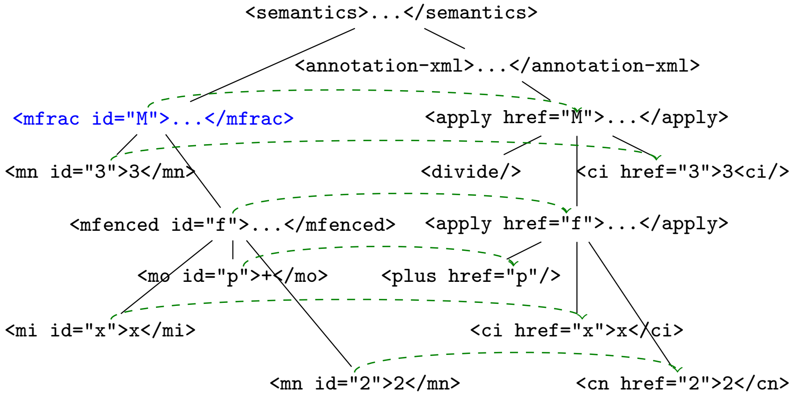
- MathML3 is a W3C Recommendation for representing Formulae [ABC+10]
- Idea: Combine the presentation and content markup and cross-reference



- use e.g. for semantic copy and paste.
  (click on presentation, follow link and copy content)
- But: Formulae are mostly written in LaTeX, e.g. `\frac{3}{(x+2)}`
- Solution: Write LaTeX, convert to HTML5 $\hat{=}$ HTML+MathML+SVG

Parallel Markup Markup in MathML

- Concrete Realization in MathML: `semantics` element with presentation as first child and content in `annotation-xml` child



Task Description

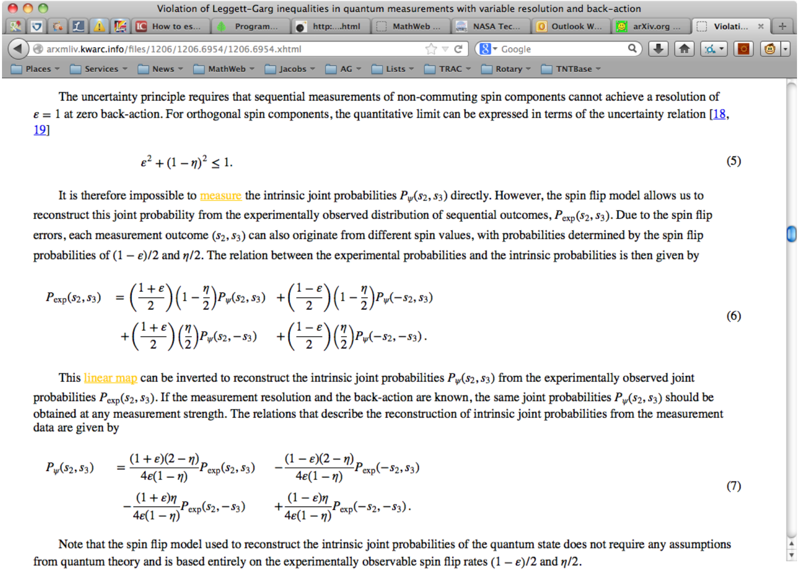
Task Overview

- Math Retrieval Subtask: Given a document collection, retrieve relevant mathematical formulae or documents for a given query.
- Math Understanding Subtask: Extract natural language definitions of mathematical expressions in a document for their semantic interpretation.

NTCIR-Math Pilot Task: Task Design & State

- NTCIR-10 Math Dataset: 100.000 Documents transformed to HTML5 from `http://arxiv.org` (10.000 for dry run)
  - 63 GiB overall size, 35 MFormulae, 297 MSubformulae (size challenge for systems)
  - every formula given in content MathML, presentation MathML, and LaTeX (23 GiB)

NTCIR-Math Pilot Task: Task Design & State



The uncertainty principle requires that sequential measurements of non-commuting spin components cannot achieve a resolution of $\varepsilon = 1$ at zero back-action. For orthogonal spin components, the quantitative limit can be expressed in terms of the uncertainty relation [18, 19]

$$\varepsilon^2 + (1-\eta)^2 \leq 1. \tag{5}$$

It is therefore impossible to measure the intrinsic joint probabilities $P_\psi(s_2, s_3)$ directly. However, the spin flip model allows us to reconstruct this joint probability from the experimentally observed distribution of sequential outcomes, $P_{\exp}(s_2, s_3)$. Due to the spin flip errors, each measurement outcome (s_2, s_3) can also originate from different spin values, with probabilities determined by the spin flip probabilities of $(1-\varepsilon)/2$ and $\eta/2$. The relation between the experimental probabilities and the intrinsic probabilities is then given by

$$\begin{aligned} P_{\exp}(s_2,s_3) &= \left(\frac{1+\varepsilon}{2}\right)\left(1-\frac{\eta}{2}\right)P_\psi(s_2,s_3) + \left(\frac{1-\varepsilon}{2}\right)\left(1-\frac{\eta}{2}\right)P_\psi(-s_2,s_3) \\ &+ \left(\frac{1+\varepsilon}{2}\right)\left(\frac{\eta}{2}\right)P_\psi(s_2,-s_3) + \left(\frac{1-\varepsilon}{2}\right)\left(\frac{\eta}{2}\right)P_\psi(-s_2,-s_3). \end{aligned} \tag{6}$$

This linear map can be inverted to reconstruct the intrinsic joint probabilities $P_\psi(s_2, s_3)$ from the experimentally observed joint probabilities $P_{\exp}(s_2, s_3)$. If the measurement resolution and the back-action are known, the same joint probabilities $P_\psi(s_2, s_3)$ should be obtained at any measurement strength. The relations that describe the reconstruction of intrinsic joint probabilities from the measurement data are given by

$$\begin{aligned} P_\psi(s_2,s_3) &= \frac{(1+\varepsilon)(2-\eta)}{4\varepsilon(1-\eta)}P_{\exp}(s_2,s_3) - \frac{(1-\varepsilon)(2-\eta)}{4\varepsilon(1-\eta)}P_{\exp}(-s_2,s_3) \\ &- \frac{(1+\varepsilon)\eta}{4\varepsilon(1-\eta)}P_{\exp}(s_2,-s_3) + \frac{(1-\varepsilon)\eta}{4\varepsilon(1-\eta)}P_{\exp}(-s_2,-s_3). \end{aligned} \tag{7}$$

Note that the spin flip model used to reconstruct the intrinsic joint probabilities of the quantum state does not require any assumptions from quantum theory and is based entirely on the experimentally observable spin flip rates $(1-\varepsilon)/2$ and $\eta/2$.

JACOBS UNIVERSITY

NTCIR-Math Pilot Task: Task Design & State

- NTCIR-10 Math Dataset: 100.000 Documents transformed to HTML5 from `http://arxiv.org` (10.000 for dry run)
  - 63 GiB overall size, 35 MFormulae, 297 MSubformulae (size challenge for systems)
  - every formula given in content MathML, presentation MathML, and LaTeX (23 GiB)
- NTCIR-10 Topics:
  Three Math Retrieval Subtasks: (write LaTeX+?, transform to MathML Query)
  - Formula Search (FS; automated): Formulae with named wildcards e.g. $\int_{?l}^{?h} ?f(x)^2 dx$
  - Full Text Search (FT; automated): (Formulae and keywords)
    e.g. Bell curve in the form of $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Open Math IR (OMIR; semi-automated): human-readable questions
    e.g. For which n and k is $PSL(n, k)$ not commutative?

  Math Understanding Task: Manually created content MathML as gold standard

NTCIR-Math Pilot Task: Task Design & State

- NTCIR-10 Math Dataset: 100.000 Documents transformed to HTML5 from `http://arxiv.org` (10.000 for dry run)
  - 63 GiB overall size, 35 MFormulae, 297 MSubformulae (size challenge for systems)
  - every formula given in content MathML, presentation MathML, and LaTeX (23 GiB)
- NTCIR-10 Topics:
  Three Math Retrieval Subtasks: (write LaTeX+?, transform to MathML Query)
  - Formula Search (FS; automated): Formulae with named wildcards e.g. $\int_{?l}^{?h} ?f(x)^2 dx$
  - Full Text Search (FT; automated): (Formulae and keywords)
    e.g. Bell curve in the form of $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - Open Math IR (OMIR; semi-automated): human-readable questions
    e.g. For which n and k is $PSL(n, k)$ not commutative?

  Math Understanding Task: Manually created content MathML as gold standard
- State of Play: Establishing community (16/6 Teams), ran successful Task (made mistakes, learnt a lot)

Participation, Evaluation & Results

NTCIR-10 Math Pilot Task Participants

Group ID	Organization
BRKLY	University of California, USA
FSE	Technische Universität Berlin, Germany
KWARC	Jacobs University, Germany
MCAT	National Institute of Informatics, Japan
MIRMU	Masaryk University, Czech Republic
NAK	Keio University, Japan

Number of runs for each subtask category.

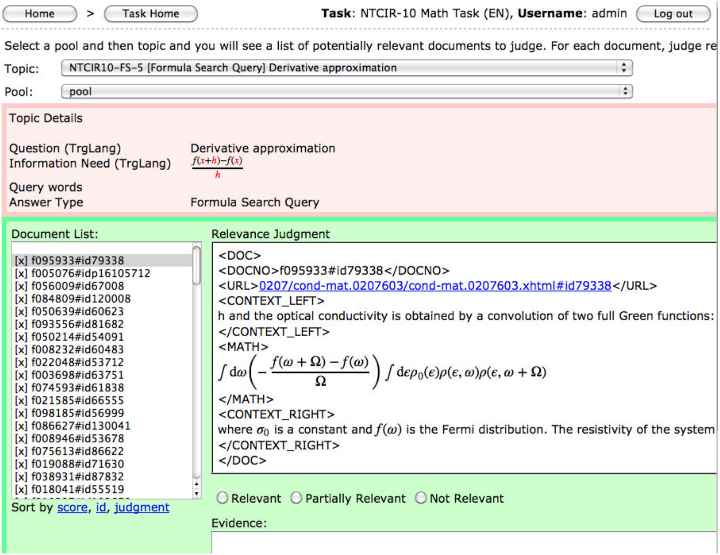
Group ID	Subtasks			
	MIR/FS	MIR/FT	MIR/OIR	MU
BRKLY	4	1*	–	–
FSE	1	1	–	–
KWARC	1	–	–	–
MCAT	1	2	–	4
MIRMU	4	1*	1*	–
NAK	1	–	–	–
Total	12	3(2*)	0 (1*)	4

∗ Reported only document URIs without formula IDs and were not included in the relevance judgment pool.

Total number of topics.

Query type	Distributed	Evaluated
Formula Search	22	21
Full Text Search	15	15
Open Search	19	0

Assessment: Math Extension for SEPIA



Math Retrieval Subtask: Pooling

- Select formulae as evenly as possible from all the runs
- The current top ranked formulae were taken from all the ranking lists, and added to the pool if they were not found.
- This process was repeated until the total size of the pool becomes equal or greater than 100.

Pooling Results

Query ID	Relevance 4	3	2					
FS-1	0	1	1					
FS-2	0	0	1					
FS-3	10	3	12					
FS-4	8	6						
FS-5	38	0						
FS-6	0							
FS-7	10		27					
FS-8	45	0	6	0	50	101	993	0.77
FS-9	0	0	40					
FS-10	0	0	13					
FS-11	0							
FS-12	0	0						
FS-13	2	0						
FS-14	1	0	34					
FS-15	3	0	0					
FS-16	19	0	2					
FS-18	44	0	32					
FS-19	0	0	24					
FS-20	32	0	27					
FS-21	27	0	12					
FS-22	0	0	72					
Total	239	10	438	60	1,381	2,128	6,496	0.45

(Easy) R: 45 PR:6 N: 50

NTCIR10-FS-8:

$$?a?x^2+?b?x+?c$$

(Hard) R: 0 PR:13 N: 87

NTCIR10-FS-10:

$$?f^{?}n(?z)?f^{(?k)(?}a?z)\,6=?c$$

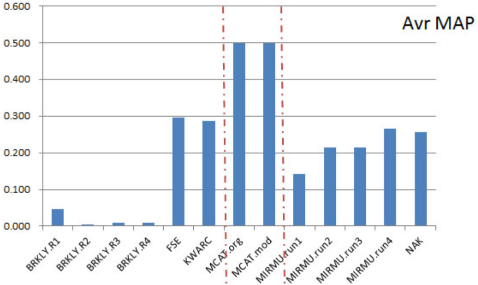
12% relevant formulae, 23% partially-relevant formulae

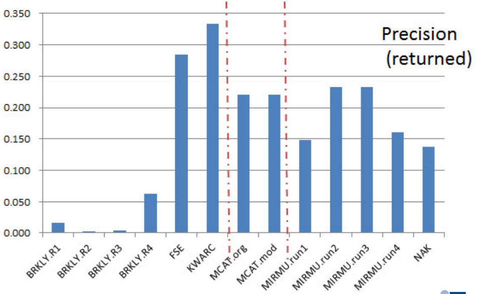
Relevance judgment statistics (Formula Search).

Query ID	Relevance score 4	3	2	1	0	Total judged	Total hit	Uniq ratio
FS-1	0	1	1	30	69	101	155	0.30
FS-2	0	0	1	1	102	104	453	0.25
FS-3	10	3	12	10	66	101	284	0.33
FS-4	8	6	17	19	52	102	278	0.56
FS-5	38	0	25	0	38	101	274	0.34
FS-6	0	0	25	0	77	102	261	0.53
FS-7	10	0	27	0	68	105	382	0.46
FS-8	45	0	6	0	50	101	993	0.77
FS-9	0	0	40	0	63	103	361	0.58
FS-10	0	0	13	0	87	100	281	0.49
FS-11	0	0	42	0	58	100	161	0.29
FS-12	0	0	26	0	74	100	135	0.26
FS-13	2	0	0	0	98	100	245	0.49
FS-14	1	0	34	0	65	100	231	0.40
FS-15	3	0	0	0	98	101	304	0.23
FS-16	19	0	2	0	81	102	357	0.38
FS-18	44	0	32	0	28	104	610	0.58
FS-19	0	0	24	0	76	100	195	0.29
FS-20	32	0	27	0	41	100	100	0.00
FS-21	27	0	12	0	61	100	178	0.31
FS-22	0	0	72	0	29	101	128	0.22
Total	239	10	438	60	1,381	2,128	6,496	0.45

Evaluation Measure & Results

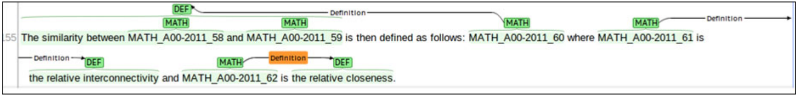
- Formula-based evaluation: It turned out document-based evaluation cost too much for human assessors!
- Evaluation measures:
  - Trec-eval (MAP, P-5, P-10) (similarity-based systems)
  - P-hit: The ratio of the relevant and the submitted hits for all the queries. (matching-based systems)





Math Understanding Task: Dataset & Participants

- Development Data: 35 papers selected from ArXiv.org dataset which were also used in Math Retrieval Task.
- Data for Formal Run: 10 papers selected from ArXiv.org dataset which were also used in Math Retrieval Task.



- There was only one participant
- The achieved performance was
  - 0.45-0.55 (F1-measure) for strict matching
  - 0.55-0.65 (F1-measure) for soft matching
- The best precision for soft matching was 0.87

Lessons Learnt & Conclusions

Four Types of Approaches

- Math-agnostic IR systems: (BRKLY)
  - Keyword-based search
  - Did not perform well
- Batch Math processors: (FSE)
  - Distributed system, but does not use a search index
  - Not suitable for interactive IR
- Matching/Unification-based Math IR systems: (KWARC, NAK)
  - Return exact instances of the query
- Similarity-search Math IR systems: (MIRMU, MCAT)
  - Return large sets of similar formulae scored by closeness.
  - Query variables are similar to any sub-formula

Analysis of Retrieval Performances

- Without explicitly dealing with mathematical formulae, it is very difficult (impossible?) to achieve high effectiveness
  - Math-agnostic systems did not perform well
- Similarity-search Math IR systems did better overall
  - Especially in terms of MAP
  - Investing into partial matches can be rewarding
- Matching /Unification-based Math IR systems perform better if unanswered topics are discarded

Relating the Subtasks

- Math Search has normally two components:
  - Math Understanding (semantic extraction)
  - Semantic Search
- Math Understanding can therefore be seen as a step towards more intelligent Math Search
- In the Formulae Search Subtask, all groups used a possible baseline (LaTeXML) to approximate Math understanding
- In the future, the efforts in creating more sophisticated and effective Math understanding systems can directly feed into Formulae Search enriching the indexing of documents

Conclusions: Achievements

- First time a task dedicated to Math IR was run as part of an evaluation forum
- The NTCIR-9 Pilot Math Task has been successful in creating an experimental platform for conducting Math Retrieval experiments:
  - The development of a new collection of 100K of documents and over 35M of formulae
  - The definition of 2 natural Math Search Subtasks
  - The development of a reusable relevance assessment system for Math Tasks

Conclusions: Assessment of the state

- A great deal of work has been done in the first NTCIR-10 Math Pilot Task:
  - Identification of reasonable baselines
  - Shaping the details of the tackled Math Subtasks
  - Facilitating the formation of a pluri-disciplinary community of researchers

Conclusions: Assessment of the state

- A great deal of work has been done in the first NTCIR-10 Math Pilot Task:
  - Identification of reasonable baselines
  - Shaping the details of the tackled Math Subtasks
  - Facilitating the formation of a pluri-disciplinary community of researchers
- A great more deal of work is still needed:
  - Refinement of the topic development process (e.g. easy vs hard topics)
  - Conducting an inter-assessor agreement study
  - Developing common/standard baselines
  - ... Converging perspectives from the two main types of participating groups: IR scientists and Mathematicians/Logicians

Looking Forward

- It is our intention to run a new iteration of Math IR Task in NTCIR-11
  - Using the same created collection in NTCIR-10
  - Focussing more on the Formulae Search Subtask
  - Achieving a reusable test collection for the Formulae Search Subtask
  - Developing the Math Understanding Subtask
- ...Growing and supporting the Math IR community
  Visit the new community portal: `https://trac.mathweb.org/NTCIR-Math/`

Ron Ausbrooks, Stephen Buswell, David Carlisle, Giorgi Chavchanidze, Stéphane Dalmas, Stan Devitt, Angel Diaz, Sam Dooley, Roger Hunter, Patrick Ion, Michael Kohlhase, Azzeddine Lazrek, Paul Libbrecht, Bruce Miller, Robert Miner, Murray Sargent, Bruce Smith, Neil Soiffer, Robert Sutor, and Stephen Watt.
Mathematical Markup Language (MathML) version 3.0.
W3C Recommendation, World Wide Web Consortium (W3C), 2010.

Andrea Asperti, Ferruccio Guidi, Claudio Sacerdoti Coen, Enrico Tassi, and Stefano Zacchiroli.
A content based mathematical search engine: Whelp.
In Jean-Christophe Filliâtre, Christine Paulin-Mohring, and Benjamin Werner, editors, *Types for Proofs and Programs, International Workshop, TYPES 2004, revised selected papers*, number 3839 in Lecture Notes in Computer Science, pages 17–32. Springer Verlag, 2006.

Arif Jinha.
Article 50 million: an estimate of the number of scholarly articles in existence.
Learned Publishing, 23(3):258–263, 2010.

Michael Kohlhase, Bogdan A. Matican, and Corneliu C. Prodescu.
MathWebSearch 0.5 – Scaling an Open Formula Search Engine.

In Johan Jeuring, John A. Campbell, Jacques Carette, Gabriel Dos Reis, Petr Sojka, Makarius Wenzel, and Volker Sorge, editors, *Intelligent Computer Mathematics*, number 7362 in LNAI, pages 342–357. Springer Verlag, 2012.

Michael Kohlhase and Ioan Şucan.
A search engine for mathematical formulae.
In Tetsuo Ida, Jacques Calmet, and Dongming Wang, editors, *Proceedings of Artificial Intelligence and Symbolic Computation, AISC'2006*, number 4120 in LNAI, pages 241–253. Springer Verlag, 2006.

Paul Libbrecht and Erica Melis.
Methods for Access and Retrieval of Mathematical Content in ActiveMath.
In N. Takayama and A. Iglesias, editors, *Proceedings of ICMS-2006*, number 4151 in LNAI, pages 331–342. Springer Verlag, 2006.
http://www.activemath.org/publications/Libbrecht-Melis-Access-and-Retrieval-ActiveMath-ICMS-2006.pdf.

Peder Olesen Larsen and Markus von Ins.
The rate of growth in scientific publication and the decline in coverage provided by science citation index.
Scientometrics, 84(3):575–603, 2010.

Jozef Misutka and Leo Galambos.

System description: Egomath2 as a tool for mathematical searching on wikipedia.org.
In James Davenport, William Farmer, Florian Rabe, and Josef Urban, editors, *Calculemus/MKM*, number 6824 in LNAI, pages 307–309. Springer Verlag, 2011.

Rajesh Munavalli and Robert Miner.
Mathfind: a math-aware search engine.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–735, New York, NY, USA, 2006. ACM Press.

Bruce R. Miller and Abdou Youssef.
Technical aspects of the digital library of mathematical functions.
Annals of Mathematics and Artificial Intelligence, 38(1-3):121–136, 2003.

Immanuel Normann.
Extended normalization for *e*-retrieval of formulae.
2006.