

NTCIR-11 Math-2 Task Overview

Akiko Aizawa & Michael Kohlhase & Iadh Ounis & Moritz Schubotz

<http://kwarc.info/kohlhase>
Computer Science
Jacobs University Bremen, Germany

NTCIR-11, Decemter 10. 2014

1 Introduction & Motivation for Math-2 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** [Lvl10]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** [Lvl10]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!** (↪ NTCIR-11 Math-2 Task)

- **Example 1.1 (The Wolfram Functions Site)** contains $\geq 307k$ Formulae

WOLFRAM RESEARCH **functions.wolfram.com** OTHER WOLFRAM SITES ►

Search Site Formula Search Search Tips

FUNCTION CATEGORIES VISUALIZATIONS NOTATIONS GENERAL IDENTITIES ABOUT THIS SITE



Exp
Exponential function



Mathematica Notation: $\text{Exp}[z]$

Traditional Notation: $\text{exp}(z) = e^z$


VIEW RELATED INFORMATION IN

- [The Mathematica Book](#)
- [MathWorld](#)

DOWNLOAD FORMULAS FOR THIS FUNCTION

-  [Mathematica Notebook](#)
-  [PDF File](#)

DOWNLOAD SOURCE FOR VISUALIZATIONS

-  [Mathematica Notebook](#)

Elementary Functions ► [Exp\[z\]](#) ► [Theorems](#) ▼

Fourier transformation and Parseval relation (1 formula)

$$\hat{f}(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{i y x} dx \Leftrightarrow f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(y) e^{-i y x} dy,$$
$$\int_{-\infty}^{\infty} f_1(t) f_2(x-t) dt = \int_{-\infty}^{\infty} \hat{f}_1(y) \hat{f}_2(y) e^{-i y x} dy.$$

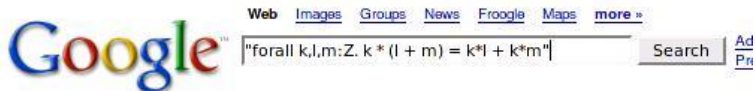
Applications of Math Information Retrieval

- ▶ Potential Applications (some in prototype state)
 - ▶ Literature search/Related Work (where have I seen this before?)
 - ▶ Applicable Theorem Search (I am stuck in a derivation)
 - ▶ Plagiarism detection (not just for the humanities)
 - ▶ Formulae in Excel spreadsheets (are just formulae as well)
 - ▶ Computations/Documentation in mathematical/symbolic software
 - ▶ time series search (e.g. via polynomial interpolations)
- ▶ Production systems in math information systems
 - ▶ MIAS in EU-DML, MathWebSearch in Zentralblatt Math

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)
- ▶ **Sicenario:** Try finding the distributivity property for \mathbb{Z}
($\forall k, l, m \in \mathbb{Z}. k \cdot (l + m) = (k \cdot l) + (k \cdot m)$)

Searching for Distributivity



Web

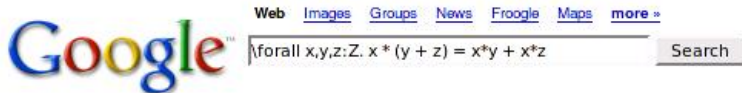
Tip: Try removing quotes from your search to get more results.

Your search - **"forall k,l,m:Z. k * (l + m) = k*l + k*m"** - did not match any documents.

Suggestions:

- ◆ Make sure all words are spelled correctly.
- ◆ Try different keywords.
- ◆ Try more general keywords.

Searching for Distributivity



Searching for Distributivity



Web Images Groups News Froogle Maps more »

`\forall\text{forall } a,b,c:\mathbb{Z}. a * (b + c) = a*b + a*c`

Search

Web

[Mathematica - Setting up equations](#)

Try `*Reduce*` rather than `*Solve*` and use `*ForAll*` to put a condition on x , y , and z . In[1]:=

`Reduce[ForAll[{x, y, z}, 5*x + 6*y + 7*z == a*x + b*y + c*z], ...`

www.codecomments.com/archive382-2006-4-904844.html - 18k - Supplemental Result -

[Cached](#) - [Similar pages](#)

[\[PDF\] arXiv:nlin.SI/0309017 v1 4 Sep 2003](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)

7.2 Appendix B. Elliptic constants related to $g(\mathbb{N}, \mathbb{C})$ 1 for all $s \leq j$. (4.14). The first condition means that the traces (4.13) of the Lax operator ...

www.citebase.org/cgi-bin/fulltext?format=application/pdf&identifier=oai:arXiv.org:nlin/0309017 -

Supplemental Result - [Similar pages](#)

[\documentclass{article} \usepackage{axiom} \usepackage{amssymb ...](#)

`i+1) bz := (bz - 2**i)::NNI else bz := bz + 2**i z.bz := z.bz + c z x * y == z ... b,i-1]] be := reduce("**, m)`

`c = 1 => be c::Ex * be coerce(x): Ex == tl ...`

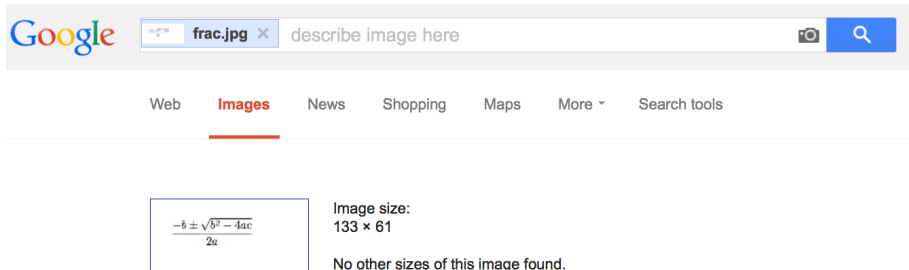
wiki.axiom-developer.org/axiom-test-1/src/algebra/CliffordSpad/src - 20k - Supplemental Result -

[Cached](#) - [Similar pages](#)

Does Image Search help?

- ▶ Math formulae are visual objects, after all

(let's try it)



The screenshot shows a Google search interface. The search bar contains the text "describe image here" and a camera icon. Below the search bar, the "Images" tab is selected. The search results display a thumbnail of a math formula:
$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$
. To the right of the thumbnail, the text reads "Image size: 133 x 61" and "No other sizes of this image found." Below the search results, there is a horizontal line.

Tip: Try entering a descriptive word in the search box.

Your search did not match any documents.

Suggestions:

- Try different keywords.

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

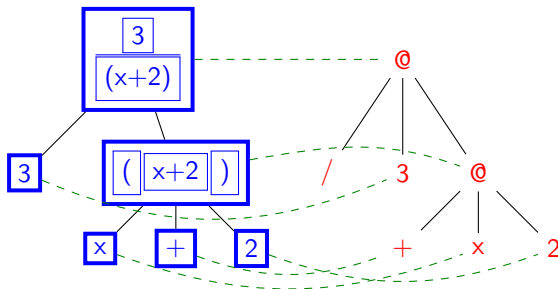
- ▶ **Formulae are not words:**
 - ▶ $a, b, c, k, l, m, x, y,$ and z are (bound) variables. (do not behave like words/symbols)
 - ▶ where are the word boundaries for “bag-of-words” methods?
- ▶ **Idea:** Need a special treatment for formulae (translate into “special words”)
Indeed this is done ([MY03, 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)

Math Markup e.g. in MathML and \LaTeX

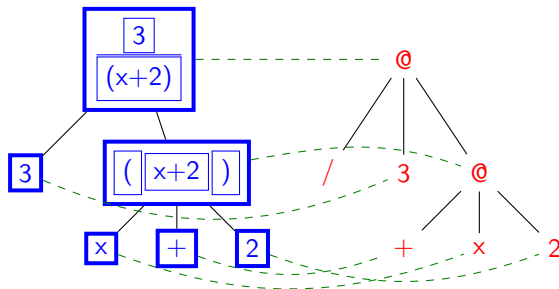
- ▶ MathML3 is a W3C Recommendation for representing Formulae [ABC⁺10]
- ▶ **Idea:** Combine the **presentation** and **content** markup and cross-reference



- ▶ use e.g. for semantic copy and paste.
(click on **presentation**, follow link and copy **content**)

Math Markup e.g. in MathML and \LaTeX

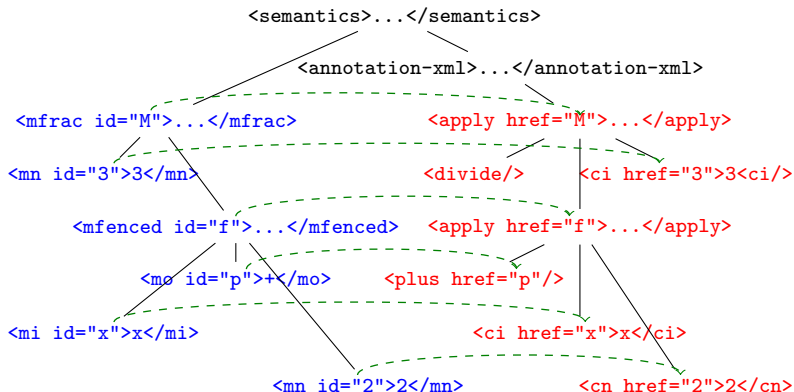
- ▶ MathML3 is a W3C Recommendation for representing Formulae [ABC⁺10]
- ▶ **Idea:** Combine the **presentation** and **content** markup and cross-reference



- ▶ 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. $\text{\frac{3}{(x+2)}}$
- ▶ **Solution:** Write \LaTeX , convert to $\text{HTML5} \hat{=} \text{HTML} + \text{MathML} + \text{SVG}$

Parallel Markup Markup in MathML

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





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 LNCS, 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, *Calcuemus/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.