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.

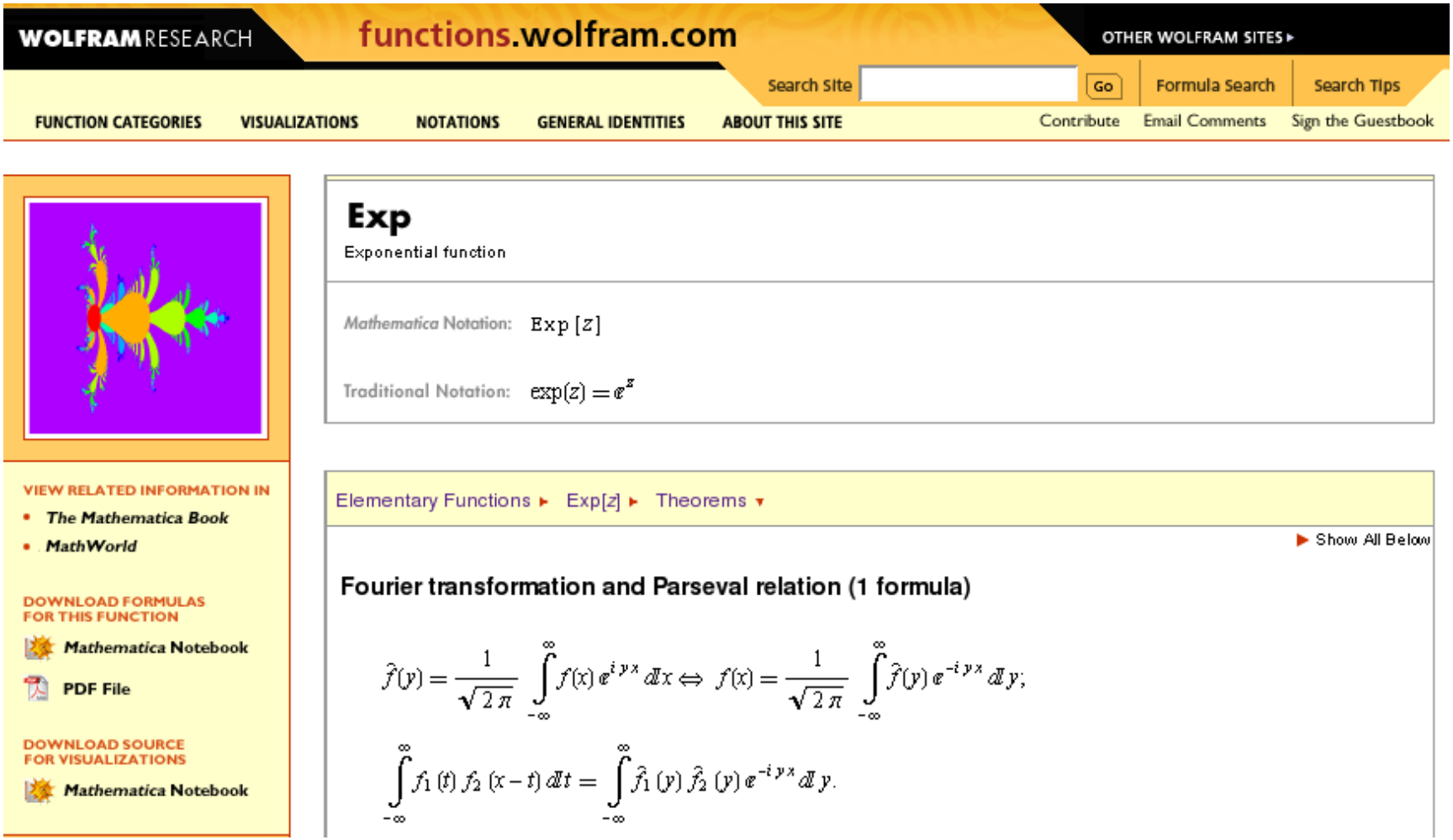
Introduction/Background

- 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! ($\leadsto$ NTCIR-11 Math-2 Task)

Mathematics Resources on the Web

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



Applications of Math Information Retrieval

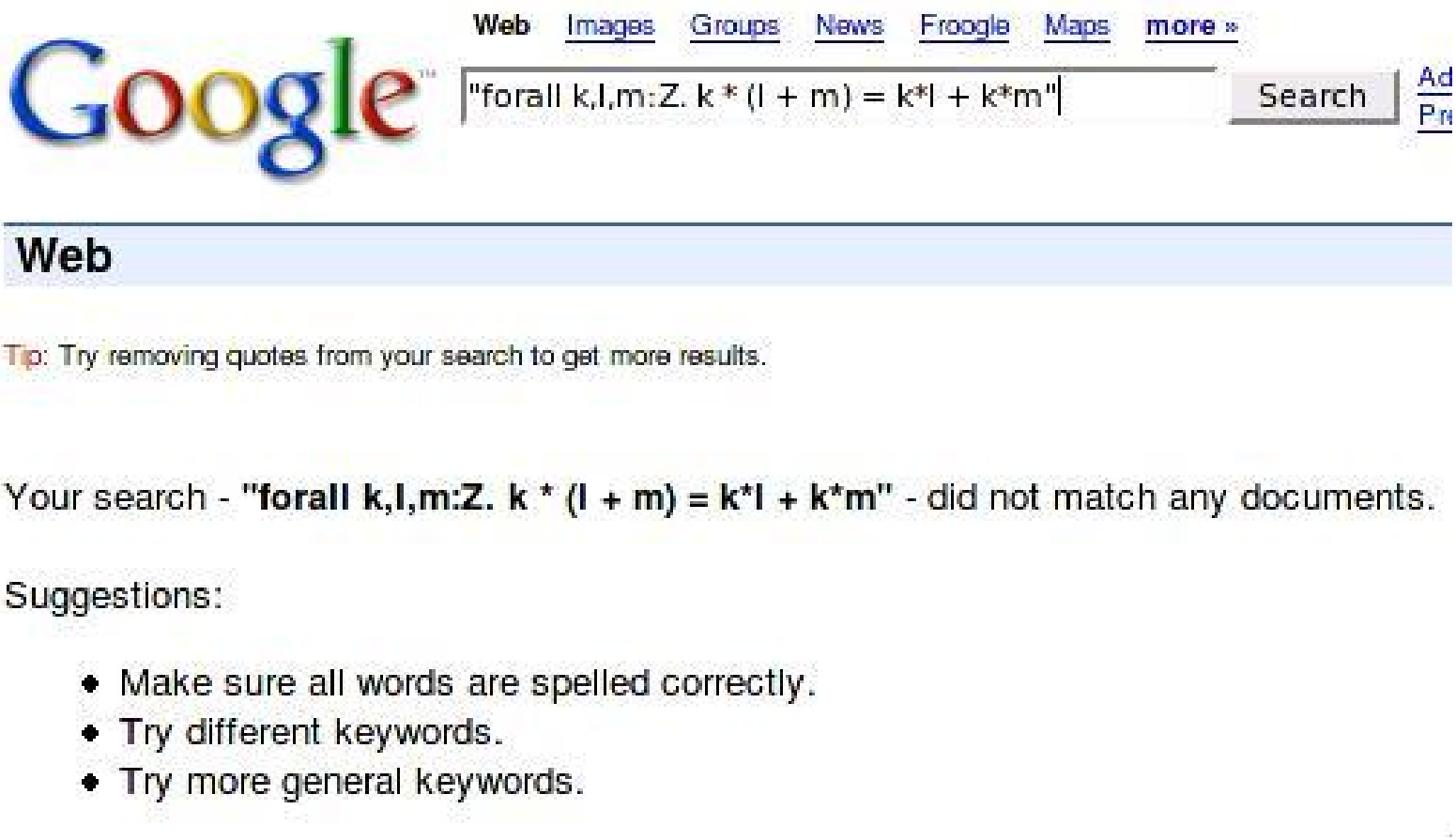
- Potential Applications (some in prototype state)
  - Literature search/Related Work (where have I seen this before?)
  - Applicable Theorem Search (I am stuck in a derivation)
  - Plagiarism detection (not just for the humanities)
  - Formulae in Excel spreadsheets (are just formulae as well)
  - Computations/Documentation 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

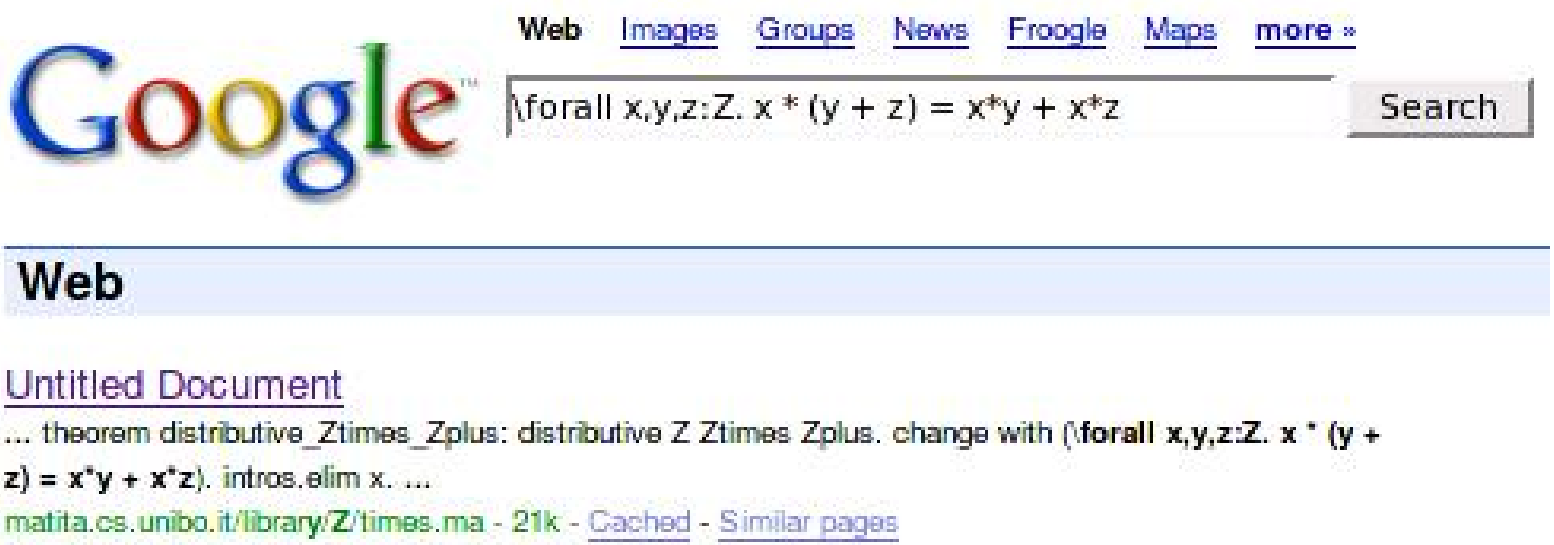
- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- . . . 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 $\mathbb{Z}$

$$(\forall k, l, m \in \mathbb{Z}.k \cdot (l + m) = (k \cdot l) + (k + m))$$

Searching for Distributivity



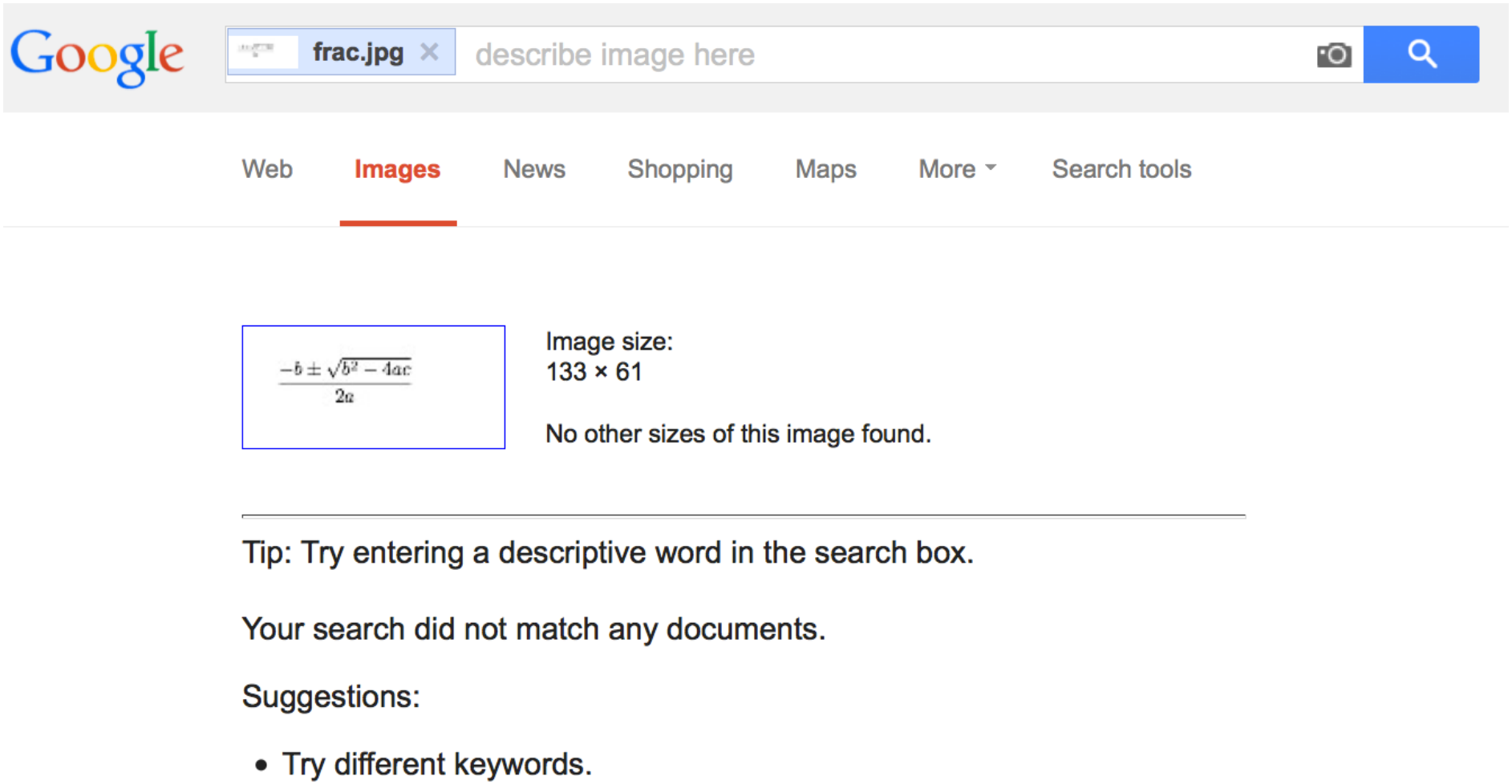
Searching for Distributivity



Searching for Distributivity

Does Image Search help?

- Math formulae are visual objects, after all (let's try it)



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

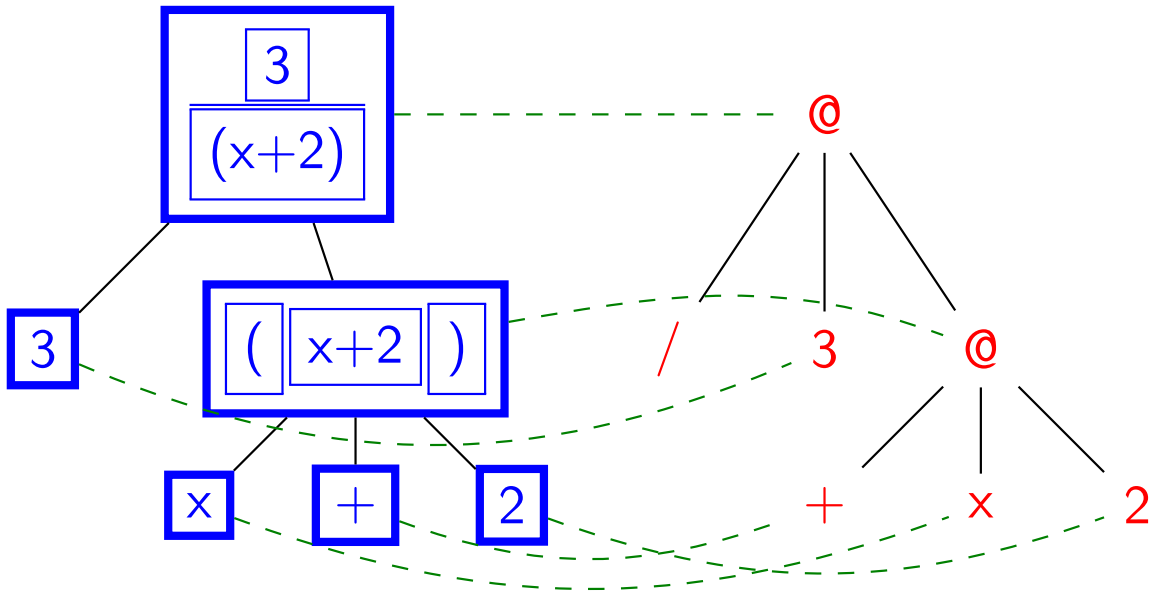
- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables.
    (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, 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

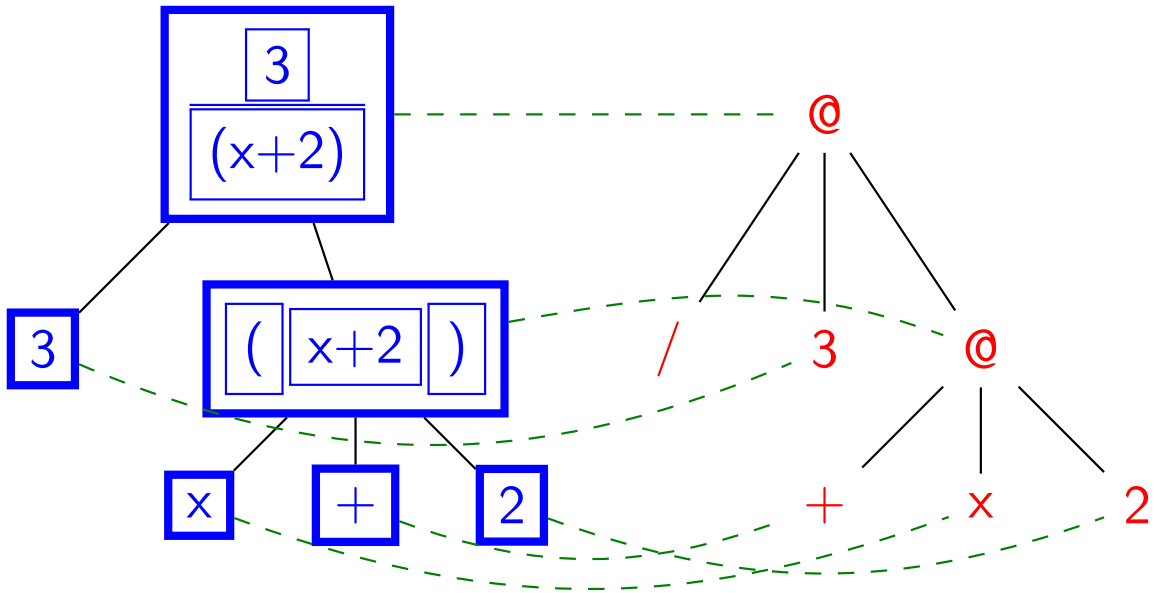


- use e.g. for semantic copy and paste.

(click on presentation, follow link and copy content)

Math Markup e.g. in MathML and LaTeX

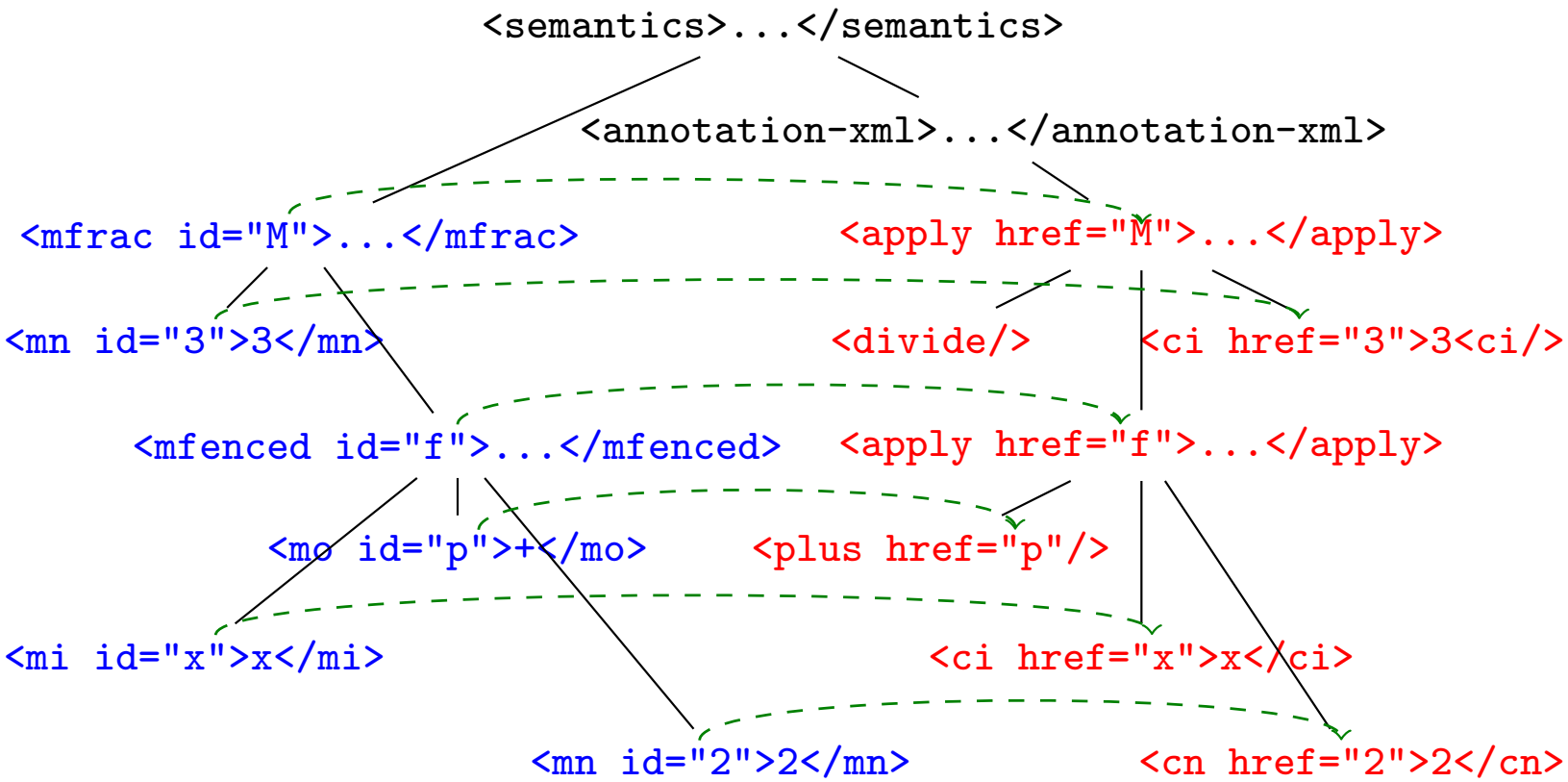
- MathML3 is a W3C Recommendation for representing Formulae [ABC+10]
- Idea: Combine the presentation and content markup and cross-reference



- 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. `\frac{3}{(x+2)}`
- Solution: Write LaTeX, convert to HTML5≙ 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




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, *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.