MathWebSearch: Low-Latency Unification-based Search

Michael Kohlhase & Corneliu Prodescu

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

NTCIR-10, June 21. 2013

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]

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 [Lyl10]

More Math: Gray literature, engineering, and school textbooks.

- 1 M Techreports at `http://ntrs.nasa.gov/` (including the Apollo reports)
- Boeing Engineer told me they have similar collection (but in Word 3,4,5,...)

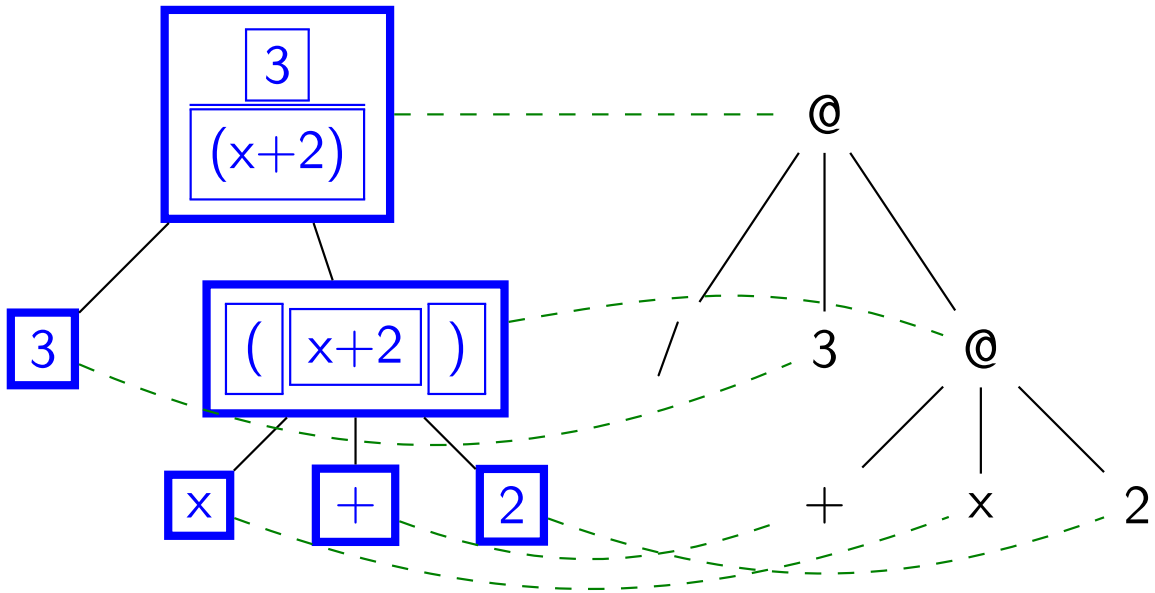
Introduction/Background

- 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 [Lyl10]

  More Math: Gray literature, engineering, and school textbooks.
  - 1 M Techreports at `http://ntrs.nasa.gov/` (including the Apollo reports)
  - Boeing Engineer told me they have similar collection (but in Word 3,4,5,...)
- We need IR support to deal with this! ($\rightsquigarrow$ NTCIR-10 Math Pilot Task)

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

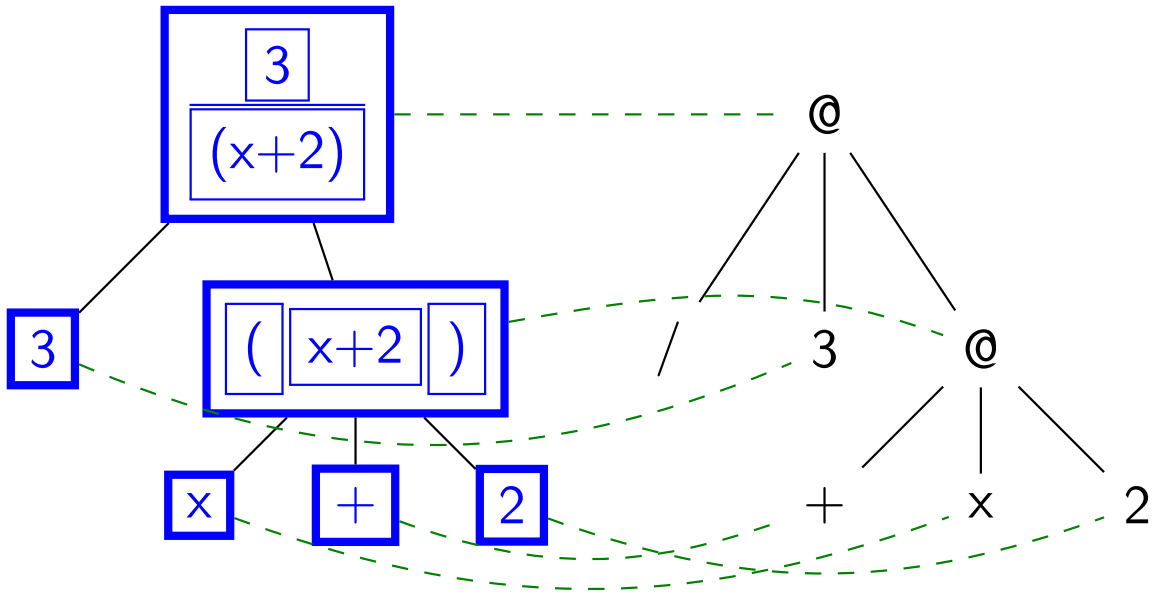


- use e.g. for semantic copy and paste.

(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

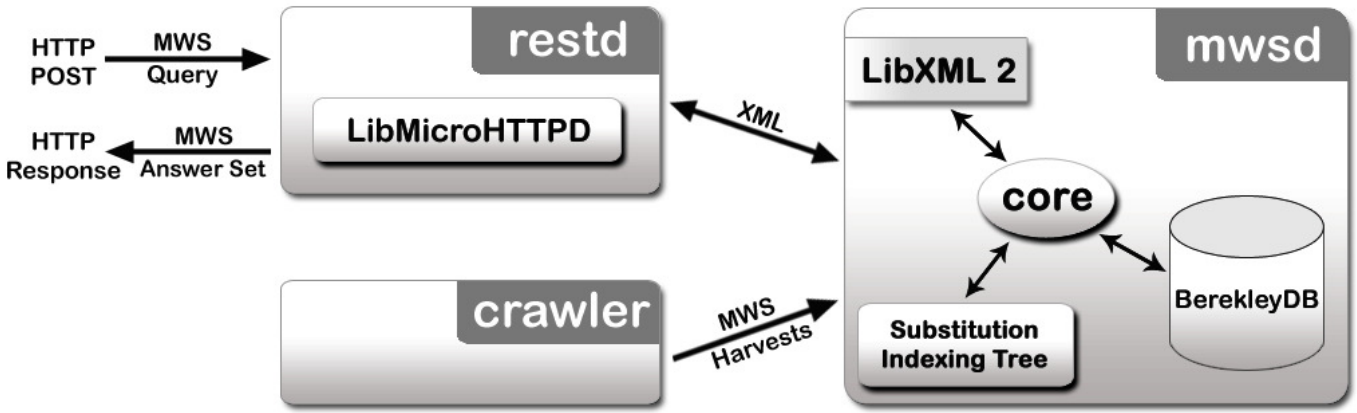
MathWebSearch only does Formula Search

- MathWebSearch is a content-based formula search engine
  - focuses on fast query answering for interactive applications.
  - restricted to exact formula search via unification queries,
  - i.e. no similarity search and no full-text search.

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

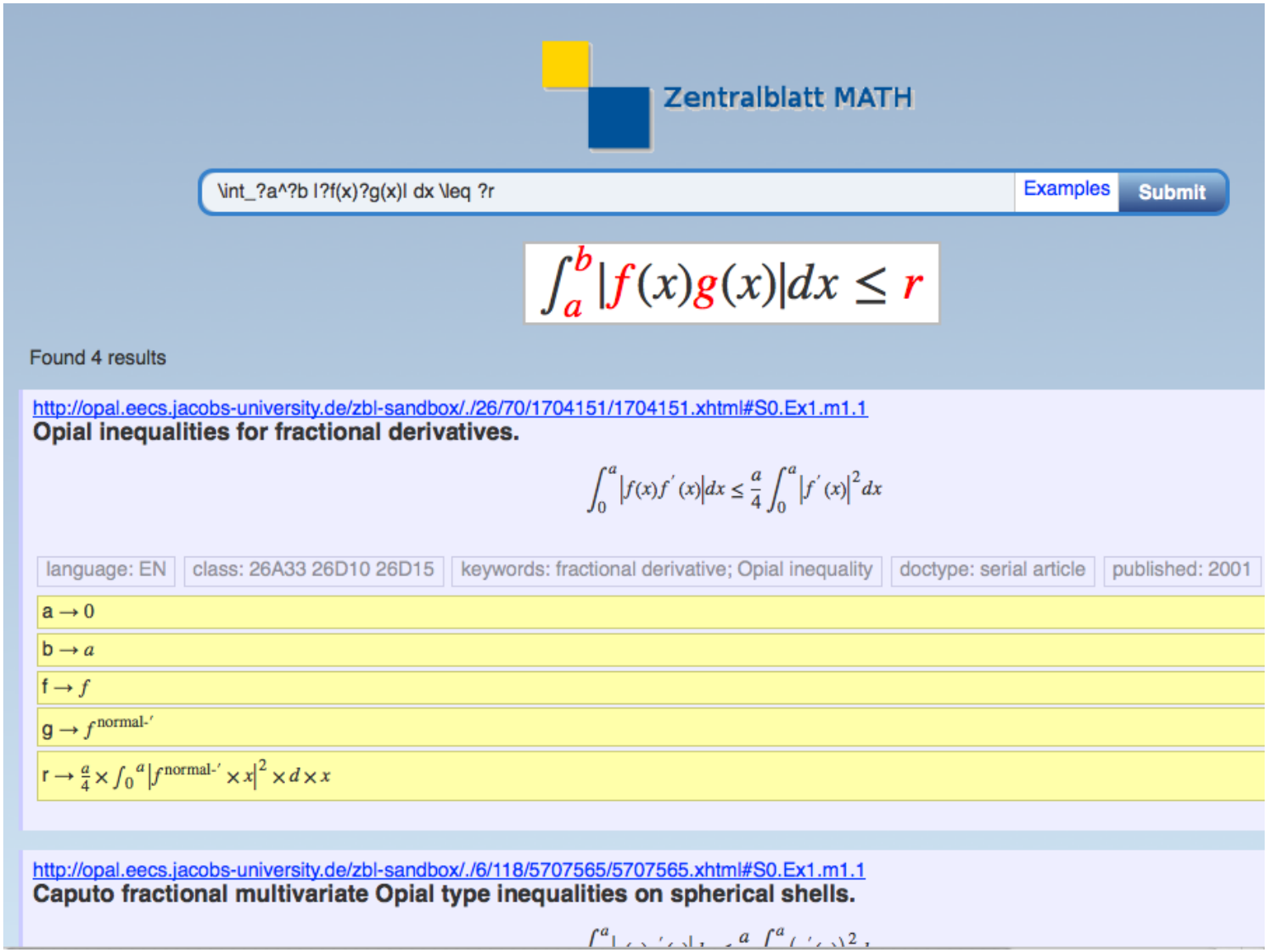
∗ Reported only document URIs without formula IDs and were not included in the relevance judgment pool.

System Architecture

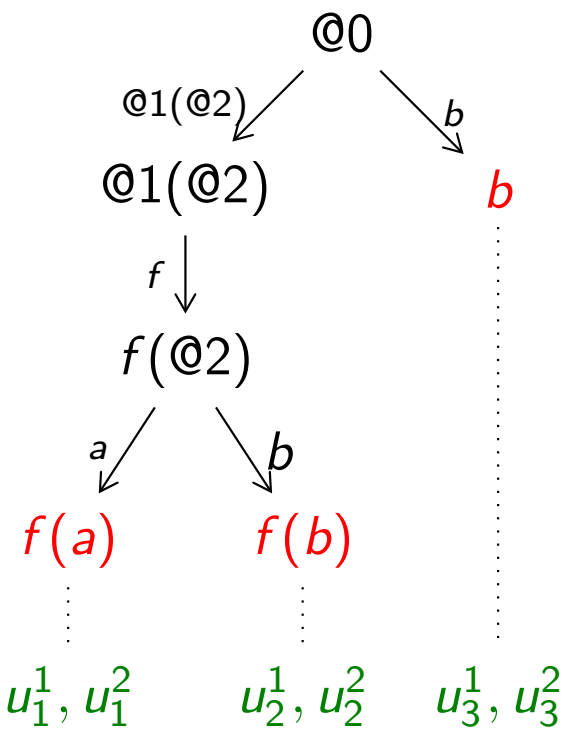


- crawlers for MathML, `OpenMath`, and OAI repositories. (convert your's?)
- multiple search servers based substitution tree indexing (formula search)
- a RESTful server that acts as a front-end for multiple search servers.
- various front ends tailored to specific applications (search appliances)
  - a Google-like web front end for human users (`search.mathweb.org`)
  - a LaTeX-based front-end for the arXiv (`http://arxivdemo.mathweb.org`)
  - special integrations for theorem prover libraries (MizarWiki, TPTP)

A Front-End for Zentralblatt Math



Substitution Tree Indexing

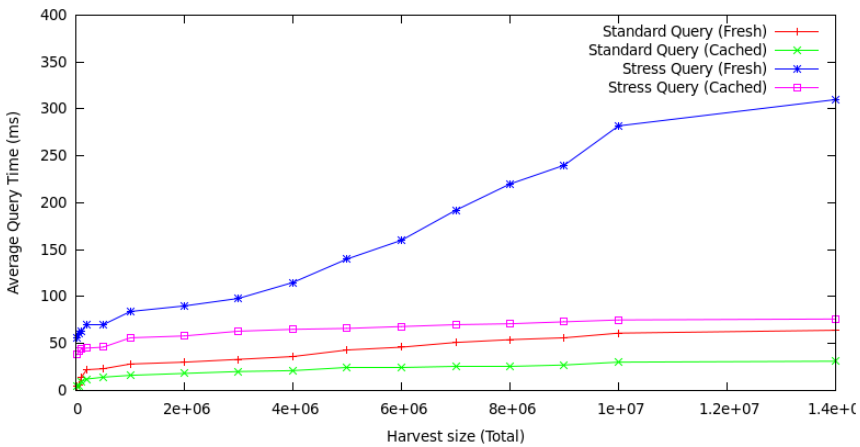


- Represent Mathematical Formulae in Content MathML extended with query variables
- Insert them into an in-memory "index": a formula structure tree that shares common substructures
- unification by "dropping queries through tree"
- leaves correspond to unifiable formulae
- leaves are mapped to result occurrence URIs u_i^j (in database)

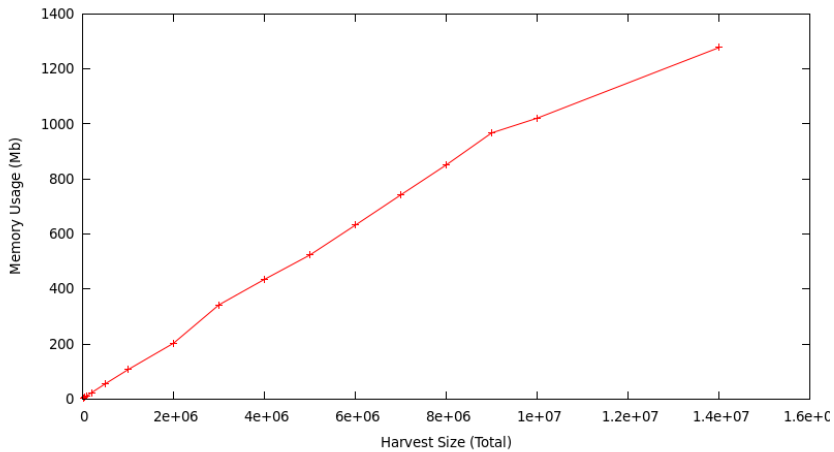
Index statistics

- Experiment: Indexing the arXiv (700k documents, $\sim 10^8$ non-trivial formulae)
- Results: indexing up to 15 M formulae on a standard laptop

Query Times



Memory Footprint



  - query time is constant ($\sim Q50$ms) (as expected; goes by depth $\times$ symbols)
  - memory footprint seems linear ($\sim Q100\frac{\text{B}}{\text{formula}}$) (expected more duplicates)
- So we need ca. $Q200$GB RAM for indexing the whole arXiv.
- Can index all published Math ($\hat{=}$ 5 $\times$ arXiv) on a large server ($Q1$TB RAM).
  (ZBL $\hat{=}$ 3M art.)

MathWebSearchResults & Interpretation

- Recap NTCIR dataset: 100,000 XHTML+MathML documents: 63GBs, 297 MFormulae
- MathWebSearchStatistics: 10GBs RAM + 43 GBs URIs (on disk) $\rightsquigarrow$ query answer times $3 - 70ms$ (avg $= 11ms$)
- Results: MathWebSearchreported 434 hits:

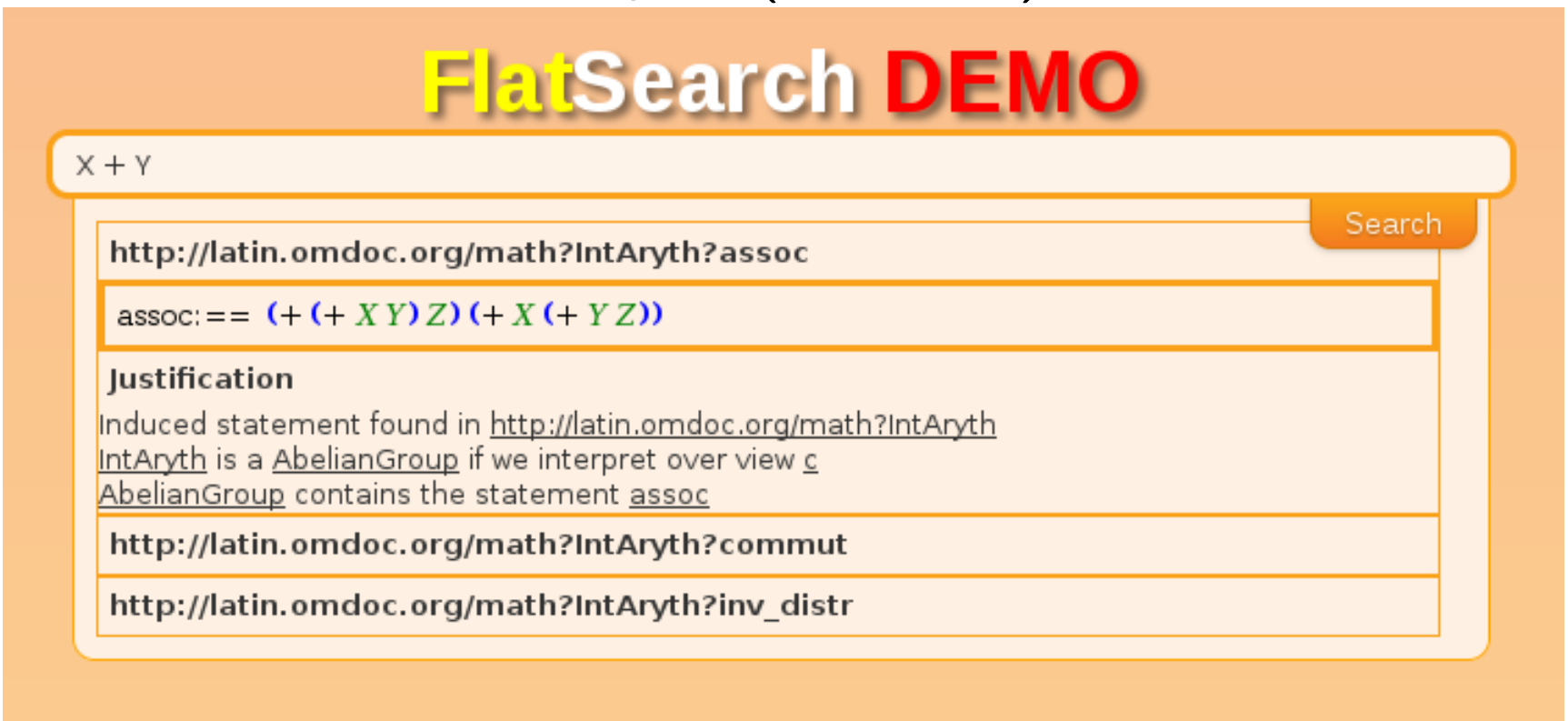
1	2	3	4	5	6	7	8	9	10	11
0	9	21	4	49	0	51	100+	100+	0	0

12	13*	14	15	16	17*	18	19	20*	21	22
0	0	0	0	0	0	100+	0	0	0	0

- Interpretation: MathWebSearchaims at high-quality hits only
  1. lots of hits (100+): general queries with multiple query variables
  2. few hits: specific queries with precise expressions.
  3. no hits: errors* in query or missing exact matches

The Future: Math Understanding, e.g. Theory Graph Structure

- Find out that $\langle \mathbb{Z}, +, * \rangle$ is a Ring, annotate
- Use this information to find induced knowledge
  - **Example 1 (FlatSearch)**



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.

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

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.