Faceted Search in Mathematics

Hambasan Radu
Supervisor: Michael Kohlhase

r.hambasan@jacobs-university.de

JACOBS UNIVERSITY

Bachelor Thesis in Computer Science

May 13, 2015

Overview

1. Introduction
2. Preliminaries
   - MathWebSearch
   - Elasticsearch
3. The Formula Schematizer
   - Purpose
   - Working Principle
4. Evaluation of Results
   - SchemaSearch
   - TeMa v2
5. Future Work
6. Discussion

Why math search?

Why math search?

- Textual search engines cannot index math.

Why math search?

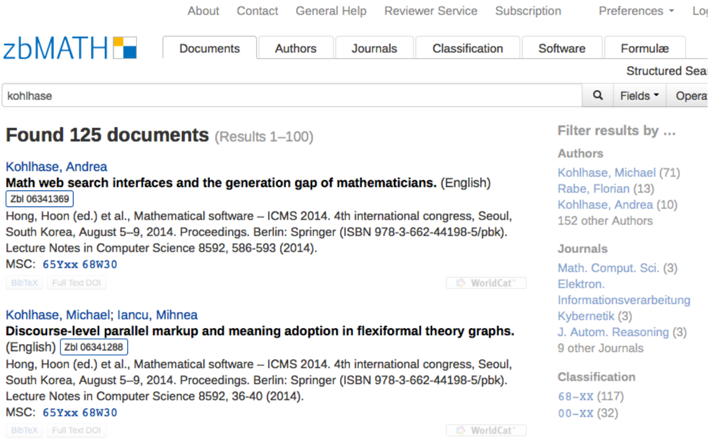
- Textual search engines cannot index math.

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Figure : Typical Formula

Why faceted search?

- It’s important to allow the user to refine the query.



Found 125 documents (Results 1–100)

Kohlhase, Andrea
Math web search interfaces and the generation gap of mathematicians. (English) Zbl 06341369
Hong, Hoon (ed.) et al., Mathematical software – ICMS 2014. 4th international congress, Seoul, South Korea, August 5–9, 2014. Proceedings. Berlin: Springer (ISBN 978-3-662-44198-5/pbk). Lecture Notes in Computer Science 8592, 586-593 (2014).
MSC: 65Yxx 68W30

Kohlhase, Michael; Iancu, Mihnea
Discourse-level parallel markup and meaning adoption in flexiformal theory graphs. (English) Zbl 06341288
Hong, Hoon (ed.) et al., Mathematical software – ICMS 2014. 4th international congress, Seoul, South Korea, August 5–9, 2014. Proceedings. Berlin: Springer (ISBN 978-3-662-44198-5/pbk). Lecture Notes in Computer Science 8592, 36-40 (2014).
MSC: 65Yxx 68W30

Filter results by ...

Authors
Kohlhase, Michael (71)
Rabe, Florian (13)
Kohlhase, Andrea (10)
152 other Authors

Journals
Math. Comput. Sci. (3)
Elektron. Informationsverarbeitung Kybernetik (3)
J. Autom. Reasoning (3)
9 other Journals

Classification
68-XX (117)
00-XX (32)

Figure : Faceted Search Example

Formula Facets

Formula Facets

- Another dimension for refining a query.

Formula Facets

- Another dimension for refining a query.
- Based on the **meaning** of formulae.

Formula Facets

- Another dimension for refining a query.
- Based on the **meaning** of formulae.
- Can be described by *formula schemata*.

Formula Facets

- Another dimension for refining a query.
- Based on the **meaning** of formulae.
- Can be described by *formula schemata*.

$$\int_M \Phi(d_p f)dvol$$

$$\lambda X.h(H^1 X)\cdots H^n X$$

$$\frac{\Gamma \vdash A \gg \alpha}{D}$$

Figure : Formula Schemata as Formula Facets

Preliminaries

- MathWebSearch (MWS)
- Elasticsearch (ES)

MathWebSearch

MathWebSearch

- content-based search engine for math

MathWebSearch

- content-based search engine for math
- indexes MathML using Substitution Tree Indexing

MathWebSearch

- content-based search engine for math
- indexes MathML using Substitution Tree Indexing
- formulae are inserted in the index according to their DFS traversal

MathWebSearch

- content-based search engine for math
- indexes MathML using Substitution Tree Indexing
- formulae are inserted in the index according to their DFS traversal
- the index nodes are unique integers corresponding to MathML elements.

MathWebSearch

- content-based search engine for math
- indexes MathML using Substitution Tree Indexing
- formulae are inserted in the index according to their DFS traversal
- the index nodes are unique integers corresponding to MathML elements.
- a FormulaID is assigned to each formula.

Example

- Formula: $\frac{2}{x+3}$

Example

- Formula: $\frac{2}{x+3}$

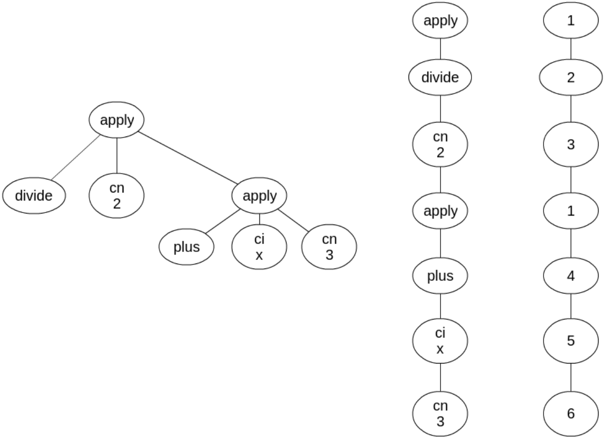


Figure : MWS Index

Elasticsearch

- powerful & efficient text search and analytics engine

Elasticsearch

- powerful & efficient text search and analytics engine
- built on top of Lucene

Elasticsearch

- powerful & efficient text search and analytics engine
- built on top of Lucene
- massively scalable & fault tolerant

Elasticsearch

- powerful & efficient text search and analytics engine
- built on top of Lucene
- massively scalable & fault tolerant
- provides faceted search features (aggregations)

Elasticsearch

- powerful & efficient text search and analytics engine
- built on top of Lucene
- massively scalable & fault tolerant
- provides faceted search features (aggregations)
- We can use it to run aggregations on formulae.

Purpose

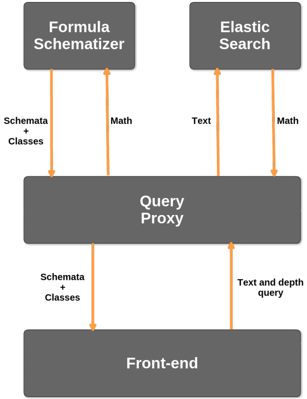


Figure : FS Engine Architecture

Working Principle

- Idea: use the index to generate similar schemata

Working Principle

- Idea: use the index to generate similar schemata

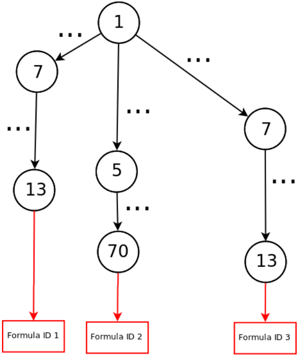


Figure : Simplified Index at depth 1

Working Principle

Working Principle

1. Obtain MathML representation of formulae set.

Working Principle

1. Obtain MathML representation of formulae set.
2. Create the table of signatures using a cutoff heuristic.

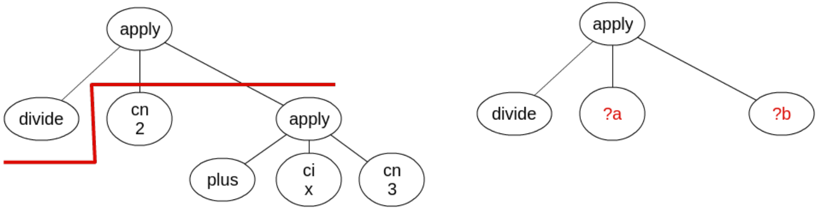
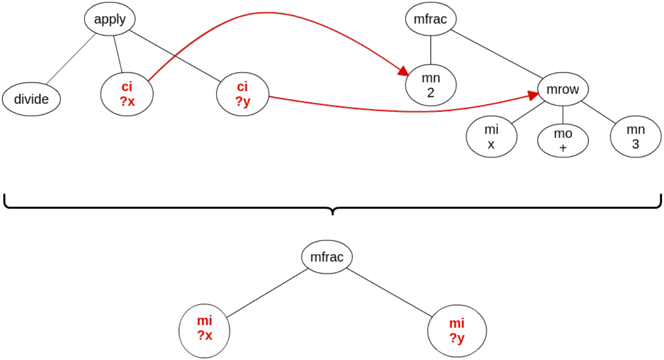


Figure : Dynamic Cutoff

3. Process the table.

3. Process the table.
4. Generate Content MathML schemata.

3. Process the table.
4. Generate Content MathML schemata.
5. Create Presentation MathML schemata (**presentation by replacement**).



SchemaSearch

SchemaSearch

- Text-only search

SchemaSearch

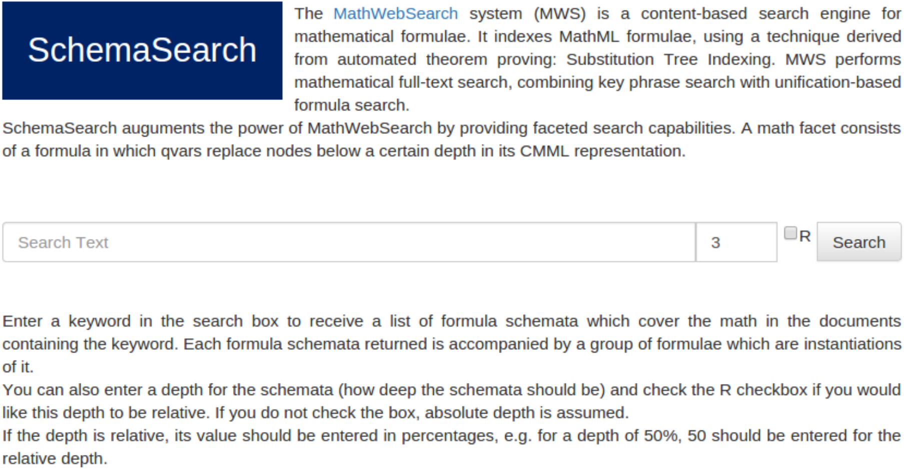
- Text-only search
- Ideal for exploring a corpus

SchemaSearch

- Text-only search
- Ideal for exploring a corpus
- It returns schemata & formula classes

SchemaSearch

- Text-only search
- Ideal for exploring a corpus
- It returns schemata & formula classes
- Used mainly to showcase to Schematizer



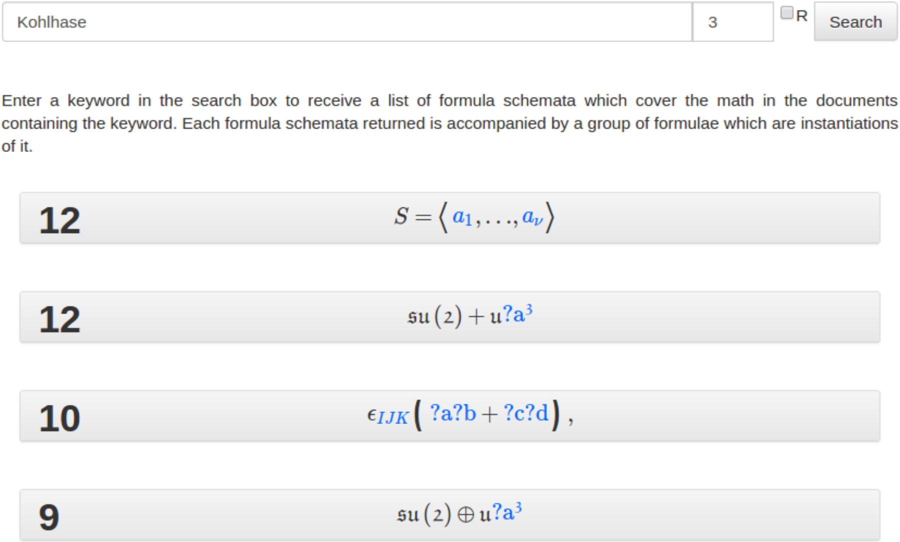
The MathWebSearch system (MWS) is a content-based search engine for mathematical formulae. It indexes MathML formulae, using a technique derived from automated theorem proving: Substitution Tree Indexing. MWS performs mathematical full-text search, combining key phrase search with unification-based formula search.

SchemaSearch auguments the power of MathWebSearch by providing faceted search capabilities. A math facet consists of a formula in which qvars replace nodes below a certain depth in its CMML representation.

Enter a keyword in the search box to receive a list of formula schemata which cover the math in the documents containing the keyword. Each formula schemata returned is accompanied by a group of formulae which are instantiations of it.

You can also enter a depth for the schemata (how deep the schemata should be) and check the R checkbox if you would like this depth to be relative. If you do not check the box, absolute depth is assumed.

If the depth is relative, its value should be entered in percentages, e.g. for a depth of 50%, 50 should be entered for the relative depth.



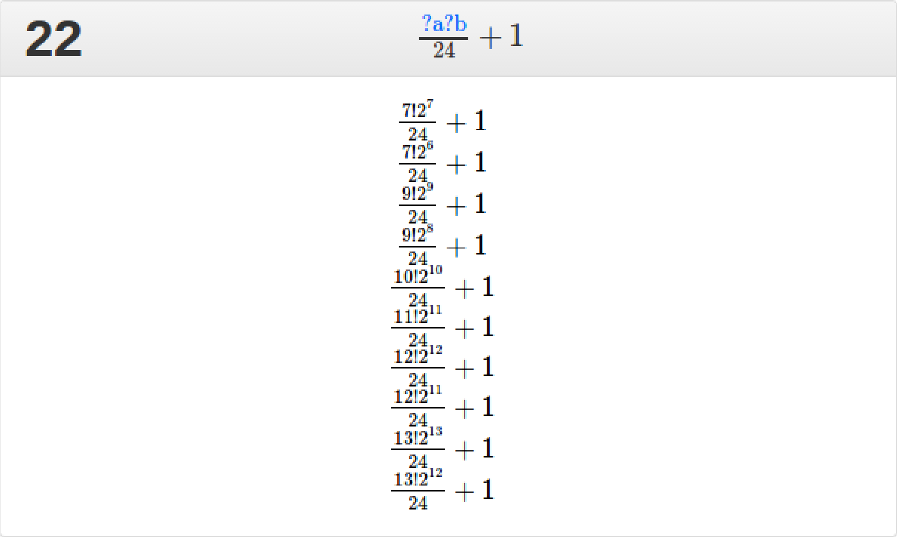
Kohlhase

3

□R

Enter a keyword in the search box to receive a list of formula schemata which cover the math in the documents containing the keyword. Each formula schemata returned is accompanied by a group of formulae which are instantiations of it.

12	$S = \langle a_1, \ldots, a_\nu \rangle$
12	$\mathfrak{su}(2) + \mathfrak{u}?\mathrm{a}^3$
10	$\epsilon_{IJK}\left(?\mathrm{a}?\mathrm{b} + ?\mathrm{c}?\mathrm{d} \right),$
9	$\mathfrak{su}(2) \oplus \mathfrak{u}?\mathrm{a}^3$



22

$$\frac{?a?b}{24}+1$$

$$\frac{7!2^{7}}{24}+1$$

$$\frac{7!2^{6}}{24}+1$$

$$\frac{9!2^{9}}{24}+1$$

$$\frac{9!2^{8}}{24}+1$$

$$\frac{10!2^{10}}{24}+1$$

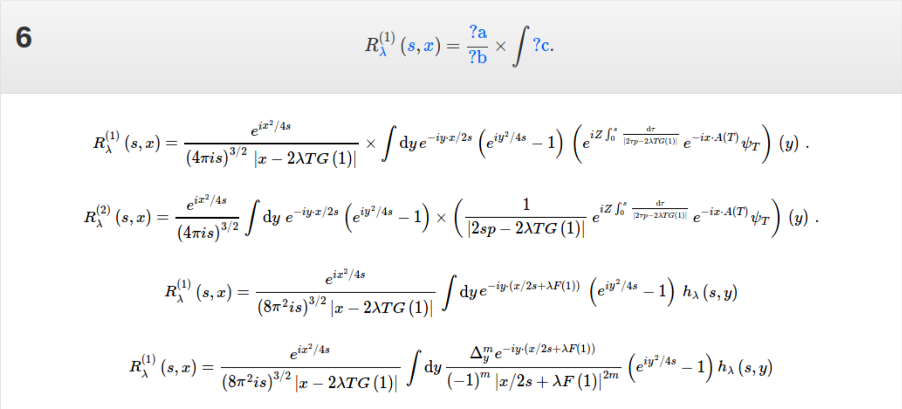
$$\frac{11!2^{11}}{24}+1$$

$$\frac{12!2^{12}}{24}+1$$

$$\frac{12!2^{11}}{24}+1$$

$$\frac{13!2^{13}}{24}+1$$

$$\frac{13!2^{12}}{24}+1$$



6

$$R_{\lambda}^{(1)}\left(s,x\right)=\frac{?\mathrm{a}}{?\mathrm{b}}\times\int ?\mathrm{c}.$$

$$R_{\lambda}^{(1)}\left(s,x\right)=\frac{e^{ix^{2}/4s}}{\left(4\pi is\right)^{3/2}\left|x-2\lambda TG\left(1\right)\right|}\times\int\mathrm{d}y\,e^{-iy\cdot x/2s}\left(e^{iy^{2}/4s}-1\right)\left(e^{iZ\int_{0}^{s}\frac{\mathrm{d}\tau}{\left|2\tau p-2\lambda TG(1)\right|}}e^{-ix\cdot A(T)}\psi_{T}\right)(y)\;.$$

$$R_{\lambda}^{(2)}\left(s,x\right)=\frac{e^{ix^{2}/4s}}{\left(4\pi is\right)^{3/2}}\int\mathrm{d}y\;e^{-iy\cdot x/2s}\left(e^{iy^{2}/4s}-1\right)\times\left(\frac{1}{\left|2sp-2\lambda TG\left(1\right)\right|}e^{iZ\int_{0}^{s}\frac{\mathrm{d}\tau}{\left|2\tau p-2\lambda TG(1)\right|}}e^{-ix\cdot A(T)}\psi_{T}\right)(y)\;.$$

$$R_{\lambda}^{(1)}\left(s,x\right)=\frac{e^{ix^{2}/4s}}{\left(8\pi^{2}is\right)^{3/2}\left|x-2\lambda TG\left(1\right)\right|}\int\mathrm{d}y\,e^{-iy\cdot\left(x/2s+\lambda F(1)\right)}\left(e^{iy^{2}/4s}-1\right)h_{\lambda}\left(s,y\right)$$

$$R_{\lambda}^{(1)}\left(s,x\right)=\frac{e^{ix^{2}/4s}}{\left(8\pi^{2}is\right)^{3/2}\left|x-2\lambda TG\left(1\right)\right|}\int\mathrm{d}y\,\frac{\Delta_{y}^{m}e^{-iy\cdot\left(x/2s+\lambda F(1)\right)}}{\left(-1\right)^{m}\left|x/2s+\lambda F\left(1\right)\right|^{2m}}\left(e^{iy^{2}/4s}-1\right)h_{\lambda}\left(s,y\right)$$

TeMa v2

TeMa v2

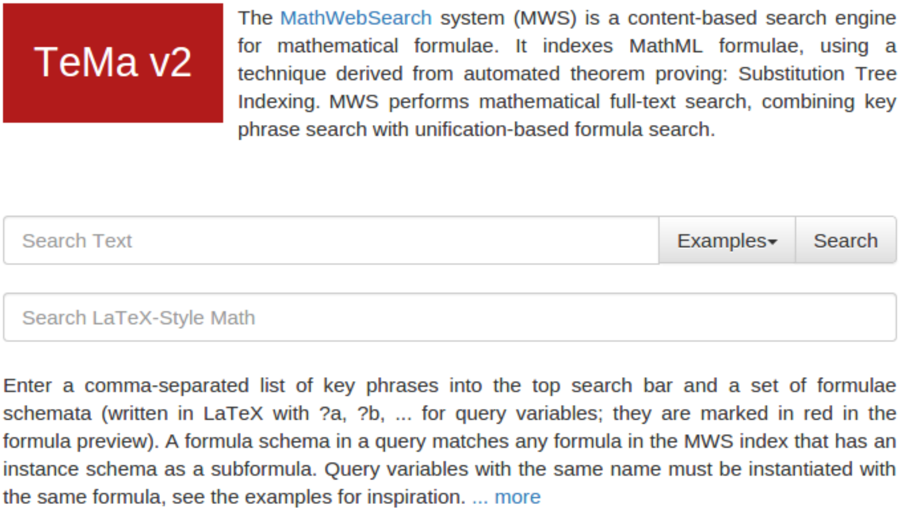
- Text and formula search

TeMa v2

- Text and formula search
- Schemata filter query results

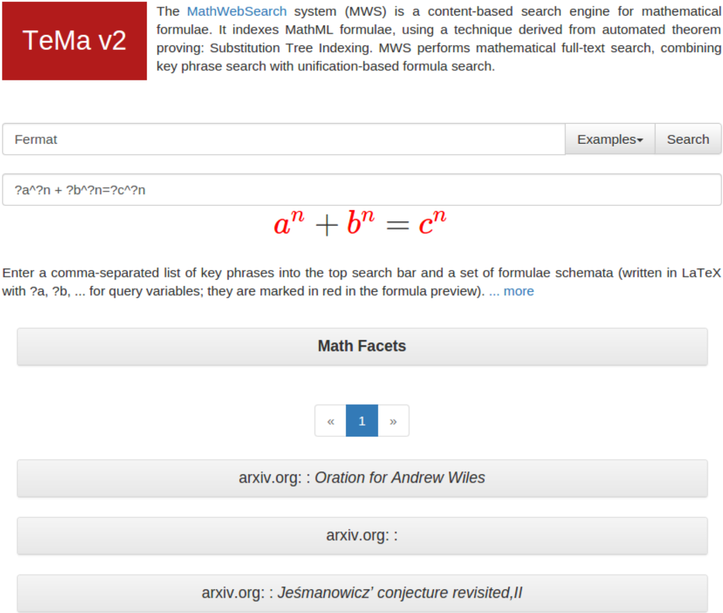
TeMa v2

- Text and formula search
- Schemata filter query results
- Main demo for faceted search



The MathWebSearch system (MWS) is a content-based search engine for mathematical formulae. It indexes MathML formulae, using a technique derived from automated theorem proving: Substitution Tree Indexing. MWS performs mathematical full-text search, combining key phrase search with unification-based formula search.

Enter a comma-separated list of key phrases into the top search bar and a set of formulae schemata (written in LaTeX with ?a, ?b, ... for query variables; they are marked in red in the formula preview). A formula schema in a query matches any formula in the MWS index that has an instance schema as a subformula. Query variables with the same name must be instantiated with the same formula, see the examples for inspiration. ... more



TeMa v2

The MathWebSearch system (MWS) is a content-based search engine for mathematical formulae. It indexes MathML formulae, using a technique derived from automated theorem proving: Substitution Tree Indexing. MWS performs mathematical full-text search, combining key phrase search with unification-based formula search.

Fermat

?a^?n + ?b^?n=?c^?n

$$a^n + b^n = c^n$$

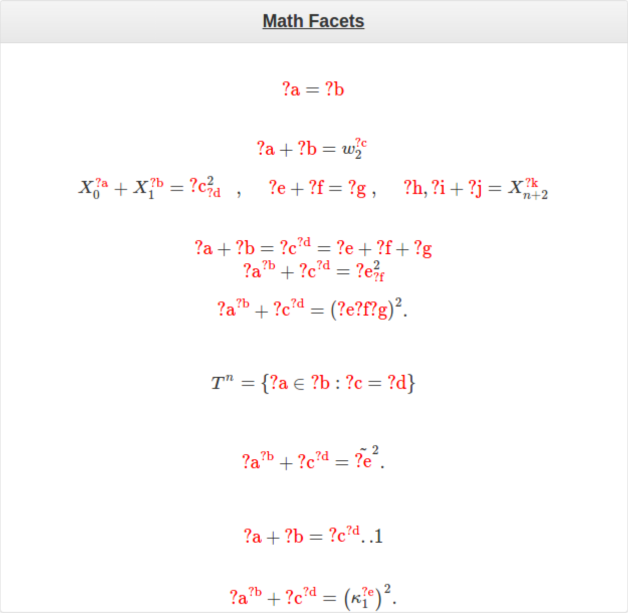
Enter a comma-separated list of key phrases into the top search bar and a set of formulae schemata (written in LaTeX with ?a, ?b, ... for query variables; they are marked in red in the formula preview). ... more

Math Facets

arxiv.org: : *Oration for Andrew Wiles*

arxiv.org: :

arxiv.org: : *Jeśmanowicz' conjecture revisited,II*



Math Facets

$$\mathrm{?a} = \mathrm{?b}$$

$$\mathrm{?a} + \mathrm{?b} = w_2^{\mathrm{?c}}$$

$$X_0^{\mathrm{?a}} + X_1^{\mathrm{?b}} = \mathrm{?c}_{\mathrm{?d}}^2 \quad , \quad \mathrm{?e} + \mathrm{?f} = \mathrm{?g}\ , \quad \mathrm{?h}, \mathrm{?i} + \mathrm{?j} = X_{n+2}^{\mathrm{?k}}$$

$$\begin{gathered} \mathrm{?a} + \mathrm{?b} = \mathrm{?c}^{\mathrm{?d}} = \mathrm{?e} + \mathrm{?f} + \mathrm{?g} \\ \mathrm{?a}^{\mathrm{?b}} + \mathrm{?c}^{\mathrm{?d}} = \mathrm{?e}_{\mathrm{?f}}^2 \end{gathered}$$

$$\mathrm{?a}^{\mathrm{?b}} + \mathrm{?c}^{\mathrm{?d}} = (\mathrm{?e?f?g})^2.$$

$$T^n = \{\mathrm{?a} \in \mathrm{?b} : \mathrm{?c} = \mathrm{?d}\}$$

$$\mathrm{?a}^{\mathrm{?b}} + \mathrm{?c}^{\mathrm{?d}} = \tilde{\mathrm{?e}}^2.$$

$$\mathrm{?a} + \mathrm{?b} = \mathrm{?c}^{\mathrm{?d}}..1$$

$$\mathrm{?a}^{\mathrm{?b}} + \mathrm{?c}^{\mathrm{?d}} = \left(\kappa_1^{\mathrm{?e}}\right)^2.$$

Future Work

- Similarity Search
- Improving NNexus

Demos

- **SchemaSearch**: http://jupiter.eecs.jacobs-university.de/schema
- **TeMa v2**: http://jupiter.eecs.jacobs-university.de/temaV2

1. Introduction
2. Preliminaries
   - MathWebSearch
   - Elasticsearch
3. The Formula Schematizer
   - Purpose
   - Working Principle
4. Evaluation of Results
   - SchemaSearch
   - TeMa v2
5. Future Work
6. Discussion