

# Faceted Search in Mathematics

Hambasan Radu

*r.hambasan@jacobs-university.de*



JACOBS  
UNIVERSITY

Bachelor Thesis in Computer Science

May 12, 2015

# Overview

## Introduction

## Preliminaries

MathWebSearch

Elasticsearch

## The Formula Schematizer

Purpose

Working Principle

## Evaluation of Results

SchemaSearch

TeMa v2

## Future Work

## 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.


[About](#)
[Contact](#)
[General Help](#)
[Reviewer Service](#)
[Subscription](#)
[Preferences ▾](#)
[Log](#)
**zbMATH**

[Documents](#)
[Authors](#)
[Journals](#)
[Classification](#)
[Software](#)
[Formulæ](#)
[Structured Sea](#)


[Fields ▾](#)
[Opera](#)

## 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](#)

[BibTeX](#)
[Full Text DOI](#)
[WorldCat\\*](#)
[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](#)

[BibTeX](#)
[Full Text DOI](#)
[WorldCat\\*](#)
[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*.

$$\int_M \Phi(d_p f) d\text{vol}$$

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

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

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.



- ▶ Idea: use the index to generate similar schemata

- Idea: use the index to generate similar schemata

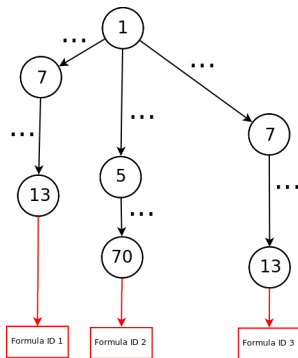


Figure : Simplified Index at depth 1



# 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

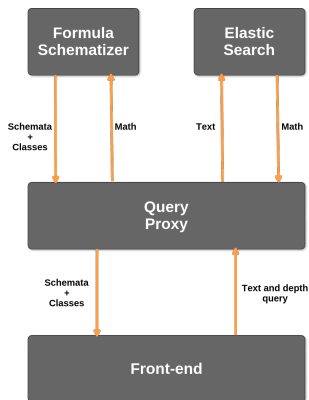


Figure : FS Engine Architecture



# Working Principle

# 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.

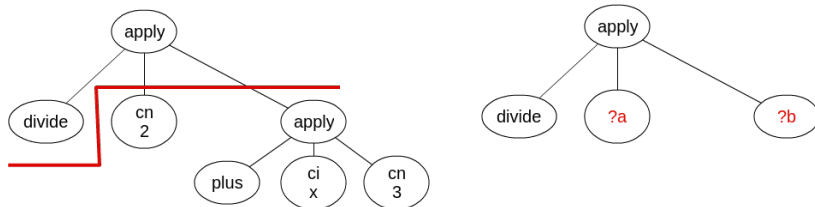


Figure : Dynamic Cutoff





### 3. Process the table.

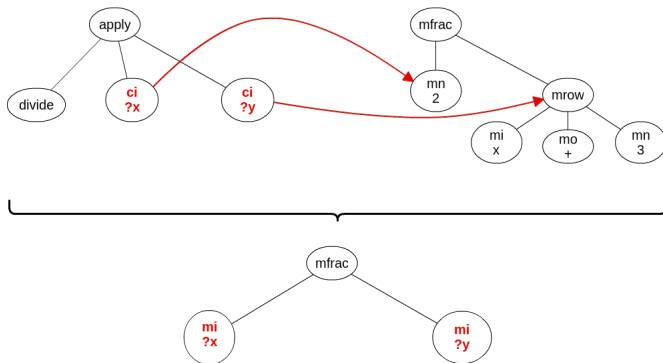




3. Process the table.
4. Generate Content MathML schemata.

## Working Principle

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

# SchemaSearch

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 augments 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.

Search Text	3	<input type="checkbox"/> R	Search
-------------	---	----------------------------	--------

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



Search

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

$$S = \langle a_1, \dots, a_\nu \rangle$$

12

$$\mathfrak{su}(2) + u^?a^3$$

10

$$\epsilon_{IJK} \left( ?a^?b + ?c^?d \right),$$

9

$$\mathfrak{su}(2) \oplus u^?a^3$$

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

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

$$R_{\lambda}^{(1)}(s, x) = \frac{e^{ix^2/4s}}{(4\pi is)^{3/2} |x - 2\lambda TG(1)|} \times \int dy e^{-iy \cdot x/2s} \left( e^{iy^2/4s} - 1 \right) \left( e^{iZ \int_0^s \frac{dr}{|2rp - 2\lambda TG(1)|}} e^{-ix \cdot A(T)} \psi_T \right) (y) .$$

$$R_{\lambda}^{(2)}(s, x) = \frac{e^{ix^2/4s}}{(4\pi is)^{3/2}} \int dy e^{-iy \cdot x/2s} \left( e^{iy^2/4s} - 1 \right) \times \left( \frac{1}{|2sp - 2\lambda TG(1)|} e^{iZ \int_0^s \frac{dr}{|2rp - 2\lambda TG(1)|}} e^{-ix \cdot A(T)} \psi_T \right) (y) .$$

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

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



# 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

## 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.

Examples▼

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

Examples▼

Search

 $?a^{?n} + ?b^{?n} = ?c^{?n}$ 

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

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

« 1 »

arxiv.org: : *Oration for Andrew Wiles*

arxiv.org: :

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



### Math Facets

$$?a = ?b$$

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

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

$$?a + ?b = ?c^{?d} = ?e + ?f + ?g$$

$$?a^{?b} + ?c^{?d} = ?e_{?f}^2$$

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

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

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

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

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

# Future Work

- ▶ SimSearch
- ▶ NNexus

## Introduction

## Preliminaries

MathWebSearch

Elasticsearch

## The Formula Schematizer

Purpose

Working Principle

## Evaluation of Results

SchemaSearch

TeMa v2

## Future Work

## Discussion