An Approach to Similarity Search for Mathematical Expressions using MathML

University of Tokyo

Keisuke Yokoi

University of Tokyo
National Institute of Informatics

Akiko Aizawa

The Purpose of the Research

- Building a search system for mathematical expressions which returns *similar* ones with a query.

Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Background

- Mathematical expressions have their own unique structures.
  - It is not an easy task for traditional search systems targeting natural languages to deal with them.
  - A new search scheme is required that takes their structures into consideration.

Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

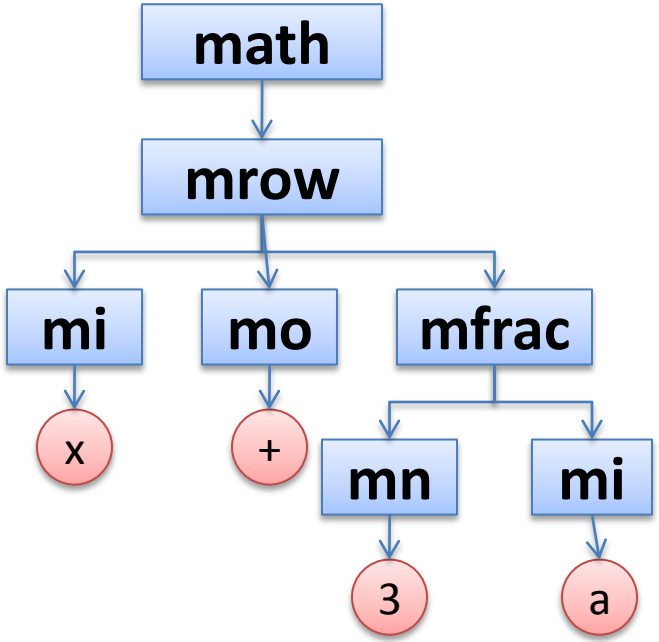
Mathematical Markup Language (MathML)

- Worldwide standard to describe mathematical contents
- A kind of XML
- Two representations
  - Presentation Markup
    - Formatting and displaying
  - Content Markup
    - Semantic construction

Example expression

- Presentation Markup

```
<math>
  <mrow>
    <mi> x </mi>
    <mo> + </mo>
    <mfrac>
      <mn> 3 </mn>
      <mi> a </mi>
    </mfrac>
  </mrow>
</math>
```

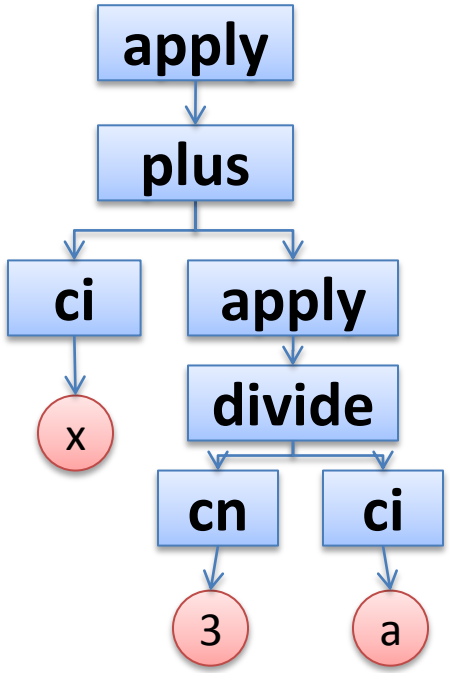


$$x + \frac{3}{a}$$

Example expression

- Content Markup

```
<apply>
  <plus/>
  <ci>x</ci>
  <apply>
    <divide/>
    <cn>3</cn>
    <ci>a</ci>
  </apply>
</apply>
```



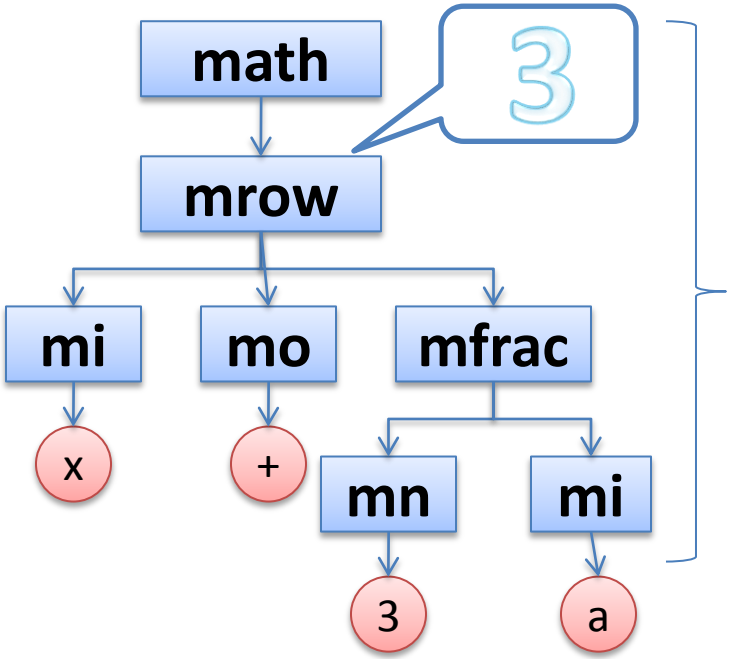
$$x + \frac{3}{a}$$

Characteristics of tree constructions

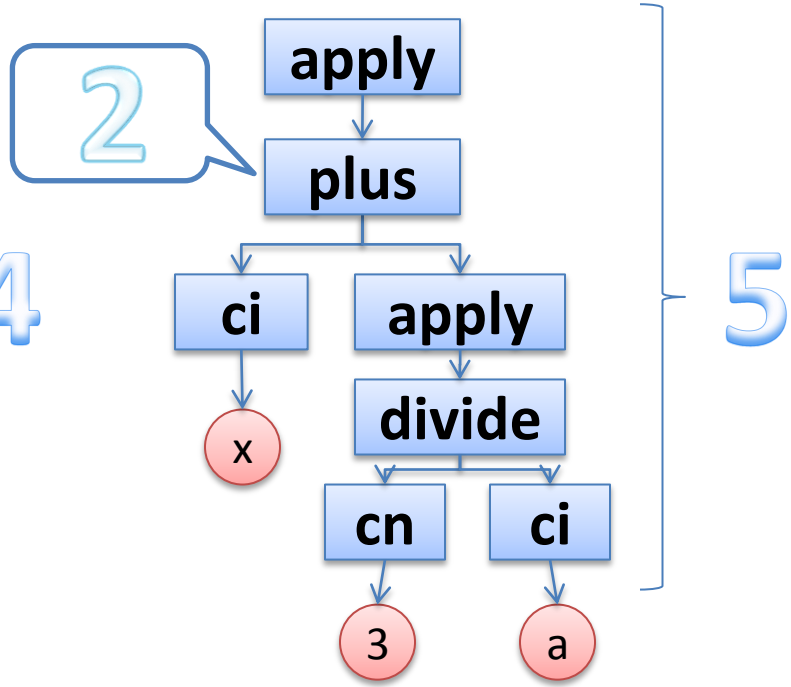
- Presentation Markup
  - broaden width
- Content Markup
  - broaden height

Example expression

- Presentation Markup



- Content Markup



Related works

- Adeel et al. [2008]
  - Math GO!
  - Generating keywords
    - By using regular expressions
  - Throwing them to conventional search systems

Template Rules	Mapped Keyword
<mo>[\(\[]</mo>\s*(<mrow>)?\s*(<mtable>\s*(<mtr>(\s*<mtd>\s*\p{Graph}+\s*</mtd>){2,}\s*</mtr>){2,}\s*</mtable>)\s*(</mrow>)?\s*<mo>[\)\]]+</mo>	Matrix
<m(?:sqrt|root)>\s*(?:(<mrow>\s*)?<mn[^>]*>\d+</mn>\s*(</mrow>\s*)?)+</m(?:sqrt|root)>	Root

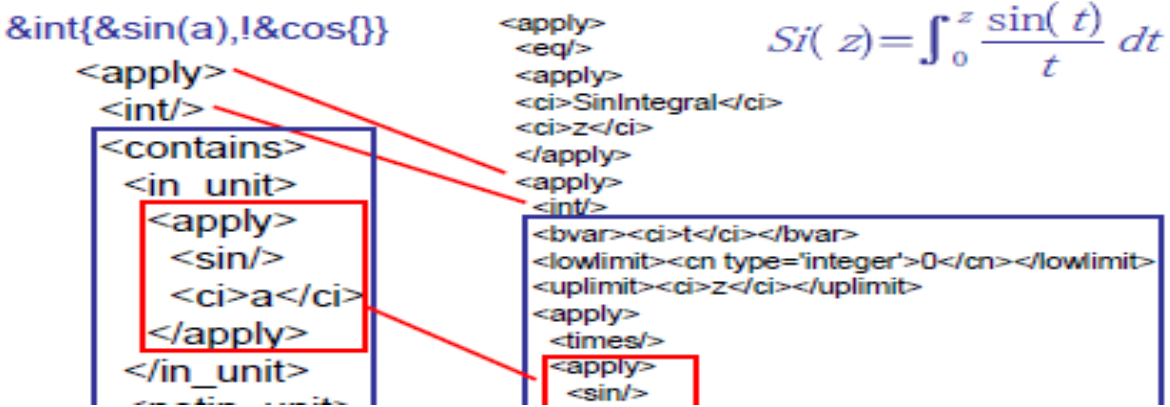
Related works

- Adeel et al. [2008]
  - Using conventional search systems
    - Scalability and Compatibility
    - Difficulty
      - narrow down the answer expressions

Template Rules	Mapped Keyword
<mo>[\(\[]</mo>\s*(<mrow>)?\s*(<mtable>\s* (<mtr>(\s*<mtd>\s*\p{Graph}+\s*</mtd>){2,}\s*</mtr>){2,} \s*</mtable>)\s*(</mrow>)?\s*<mo>[\)\]]+</mo>	Matrix
<m(?:sqrt|root)>\s*(?:(<mrow>\s*)?<mn[^>]*> \d+</mn>\s*(</mrow>\s*)?)+</m(?:sqrt|root)>	Root

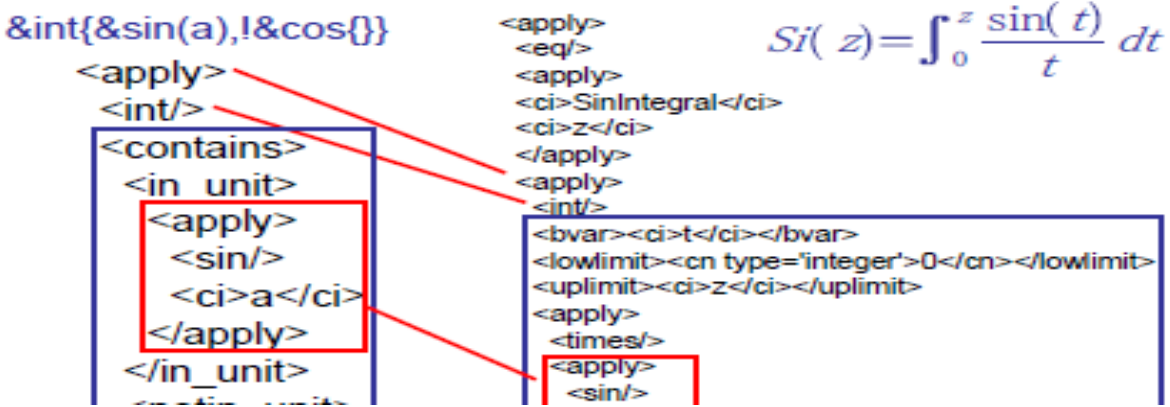
Related works

- Otagiri et al. [2008]
  - Using their own query language
  - Searching by matching tree constructions



Related works

- Otagiri et al. [2008]
  - Flexible queries
  - Only equations that exactly matching the query



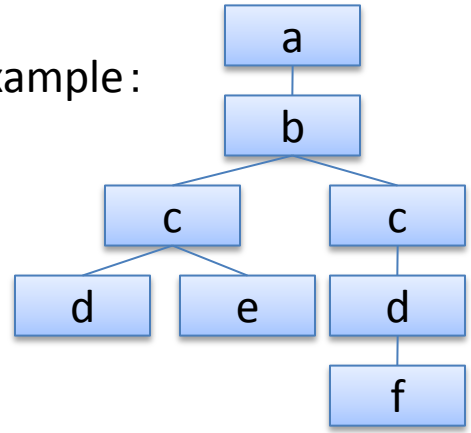
Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Subpath Set

- A similarity measure originally proposed by [Ichikawa ‘05]
- Subpath: the path from the root to the leaves and all the sub-paths of that

Example:



/a, /b, /c, /d, /e, /f
/a/b, /b/c, /c/d, /c/e, /d/f
/a/b/c, /b/c/d, /b/c/e, /c/d/f
/a/b/c/d, /a/b/c/e, /b/c/d/f
/a/b/c/d/f

Subpath Set

- In this experiments, Jaccard coefficient is used for scoring the overlap of the Subpath Sets.

Jaccard coefficient: $\dfrac{\left\| S(t_1) \cap S(t_2) \right\|}{\left\| S(t_1) \cup S(t_2) \right\|}$

t_i : a tree
$S(t_i)$: Subpath Set of t_i

Contents

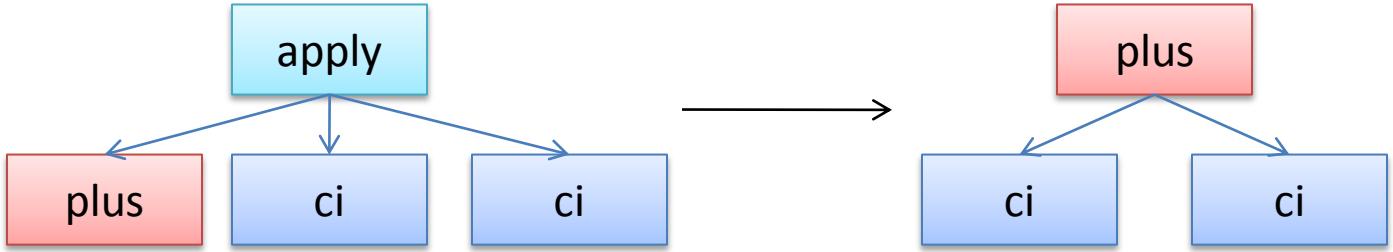
- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Transformation of Content-Based MathML

- The “apply” symbol
  - To apply an operator which is their first child to arguments
  - To be used whenever any functions and operators are used
  - Merit
    - Useful to know the range a function or operation applies
  - But in search, they cause
    - Memory consumption
    - Disguise meaningful sequences of function operators on the sub-paths
- Original Content Markup -> *apply-free* Content Markup

Transformation of Content-Based MathML

- The first children of the “apply” symbols replace their parents while other children remain the same position.



Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Experiments

- 155,607 mathematical expressions were used as targeted formulas.
  - Crawled from The Wolfram Functions Site (http://functions.wolfram.com/)

Experiments

1. Search results of example queries
   - Efficiency of Subpath Set
2. Evaluation by the rank of expected expressions
   - Superiority of *Apply-free* Content Markup

Experiments (1)

- Search results of example queries
  - Select a few sample queries
  - Check top-5 rankings
  - Use the *apply-free* Content Markup

Experiments (1)

rank	Results	
Query	$\sin(a+b) = \sin a \cos b + \cos a \sin b$	$\tan(z) = \frac{\sinh(iz)}{\sinh(i\pi/2 + iz)}$
1	$\sin(a+b) = \sin a \cos b + \cos a \sin b$	$\tan(z) = \frac{\sinh(iz)}{\sinh(i\pi/2 + iz)}$
2	$\sin(a-b) = \sin a \cos b - \cos a \sin b$	$\tan(z) = \frac{\sinh(iz)}{\sinh(i\pi/2 - iz)}$
3	$\sin(a+ib) = \sin a \cosh b + i\cos a \sin b$	$\sec(z) = \frac{i}{\sinh(i\pi/2 + iz)}$
4	$\cos(a-b) = \cos a \cos b + \sin a \sin b$	$\sec(z) = \frac{i}{\sinh(i\pi/2 - iz)}$
5	$\cos(a+b) = \cos a \cos b - \sin a \sin b$	$\cot(z) = \frac{\sinh(i\pi/2 + iz)}{\sinh(iz)}$

Experiments (2)

- Evaluation by the rank of expected expressions
  - Compare different forms of tree constructions
    - Presentation Markup
    - Content Markup
    - *Apply-free* Content Markup
  - Select an ‘expected’ answer manually for each query
  - Examine their ranks of them

Experiments (2)

Query	Expected Answer	Present	Content	Apply-free
$\sin(a+b) = \sin a\cos b + \cos a\sin b$	$\sin(a-b) = \sin a\cos b - \cos a\sin b$	x	6	2
$\int \sin z dz = -\cos z$	$\int \sin(az) dz = -\frac{\cos(az)}{a}$	x	39	23
$\int ze^{az} dz = \frac{e^{az}(-1+az)}{a^2}$	$\int z^3 e^{az} dz = \frac{e^{az}(-6+6az+3a^2z^2+a^3z^3)}{a^4}$	x	17	5
$\int (e^{cz})^v dz = \frac{(e^{cz})^v}{cv}$	$\int \sqrt{e^{cz}} dz = \frac{2\sqrt{e^{cz}}}{c}$	x	5	2
$\sin^{-1} z = \frac{3\pi}{4} - \frac{1}{2}\tan^{-1}\left(\frac{1-2z^2}{2z\sqrt{1-z^2}}\right)$	$\cos^{-1} z = -\frac{\pi}{4} + \frac{1}{2}\tan^{-1}\left(\frac{1-2z^2}{2z\sqrt{1-z^2}}\right)$	33	79	16

Results (1)

- Search result of example queries
  - Those proposed methods are capable of evaluating structural similarity of the trees.

Results (2)

- Evaluation by the rank of expected expressions
  - Presentation Markup is not suitable.
    - “mo”
    - short Subpath
  - Apply-free Content Markup showed slightly-better performance than other forms.
    - The system using Content Markups answered some unexpected expressions near the top.

Contents

- Background
- Methods
  - Mathematical Markup Language
  - Subpath Set
  - Transformation of MathML
- Experiments
- Conclusion

Conclusion

- Proposed a similarity search scheme for mathematical expressions
  - Similarity measure based on Subpath Set
  - A MathML conversion which is suitable for math search
- Demonstrated these techniques' effectiveness

Future Works

- Scalability
  - The similarity calculation may become the bottleneck.
- Consideration of symbol values
  - My search systems does not perceive the actual values.
- Difficulty in evaluation
  - The type of searching is different from previous works.

Thank you for listening!

mailto: kei-yoko@nii.ac.jp