

Investigating the Characteristics of Mathematical Notations

Clare So

clare@scl.csd.uwo.ca

Stephen Watt

watt@scl.csd.uwo.ca



ORCCA Joint Lab Meeting. UWO. April 8, 2005.

Why Studying Math Notations?

- Guide the mathematical expressions handwriting recognizer to eliminate results of recognition that does not make sense
- Preserve semantical information of expressions when translating between $\text{T}_\text{E}\text{X}$ and MathML



Math Notation is a Language

- A natural language have certain *patterns* of usages
 - For example, “x” and “z” are not letters commonly used in English words
- Usages of math notation mainly based on conventions
 - f and g for functions
 - i and j for integers
 - A and B for sets
 - i for $\sqrt{-1}$
 - ...

Previous Knowledge of Language Important

- Why we recognize the words “cat” and “hat” although “H” and “A” are written similarly?
- In natural language handwriting recognition, a built-in dictionary is used to eliminate results that does not make sense



CAT HAAT

The image shows two words written in a simple, hand-drawn style. The first word is 'CAT', with a capital 'C', a capital 'A', and a capital 'T'. The second word is 'HAAT', with a capital 'H', two capital 'A's, and a capital 'T'. The 'A's in 'HAAT' are drawn with a horizontal bar that is slightly curved, making them look very similar to the 'A' in 'CAT'. This illustrates how a handwritten word like 'HAAT' could be misinterpreted as 'CAT' if not for context or a dictionary.

How About Math?

- We need a dictionary for math in a handwriting recognizer
- Not many studies related to math “linguistics”
- This project provides a foundation of building a math dictionary



A handwritten mathematical expression $a + b + c$ is shown. The expression is enclosed in large, hand-drawn square brackets. The characters are drawn with simple, slightly irregular lines, characteristic of handwriting. The plus signs are also hand-drawn and slightly irregular.

Data Collection

- All mathematical articles in ArXiv database from 2000 to 2005 are collected
- Articles are put into their respective top-level *Mathematical Subject Classification* (2000). There are 63 classification in total representing a field in advanced mathematics.
- Math expressions are extracted from articles TeX source and translated to MathML

Analyzing Each Classification

- We keep track of the following information in each classification
 - Common expressions
 - Frequency of single character symbols
 - * Frequency of identifiers
 - * Frequency of different operators
 - Patterns of expressions
- Histograms are to be built for keeping track of symbols' frequencies

Generating Patterns by Antiunification

- We want to recognize common parts in expressions
 - Let's say we have expressions $F(x, G(x, y), z)$ and $F(a, G(a, b), z)$.
It would be useful to obtain the pattern $F(\alpha_1, G(\alpha_1, \alpha_2), z)$
- Concept of antiunification was first discussed by Robinson in 1965
- Algorithm for generating such patterns (aka antiunifiers) was discovered by Plotkin in 1970
- We are performing antiunification to MathML expression trees
 - For example, x^1 , x^2 and x^3 all have x^α as the antiunifier

Results

- The most common identifiers in classifications:
 - 08 (General algebraic systems):
 $n \quad A \quad , \quad a \quad i$
 - 12 (Field theory and polynomials):
 $, \quad n \quad x \quad i \quad K$
 - 31 (Potential theory):
 $x \quad , \quad n \quad z \quad d$
 - 42 (Fourier analysis):
 $, \quad n \quad x \quad k \quad j$
 - 83 (Relativity and gravitational theory):
 $, \quad i \quad x \quad M \quad t$

Results

- The most popular expressions in classifications:

- 08 (General algebraic systems):

`<mrow><mi>A</mi></mrow>`

`<mrow><mo>-</mo><mn>1</mn></mrow>`

`<mrow><mi>G</mi></mrow>`

`<td/>`

`<mrow><mo>(</mo><mi>X</mi><mo>)</mo></mrow>`

- 12 (Field theory and polynomials):

`<mrow><mo>(</mo><mi>x</mi><mo>)</mo></mrow>`

`<mrow><mo>-</mo><mn>1</mn></mrow>`

`<mrow><mi>K</mi></mrow>`

`<mrow><mi>p</mi></mrow>`

`<mrow><mi>n</mi><mo>-</mo><mn>1</mn></mrow>`

Results

- The most popular expressions in classifications:

- 31 (Potential theory):

(x)

(z)

$\langle \text{mtd}/\rangle$

-1

(r)

- 42 (Fourier analysis):

(x)

-1

$\langle \text{mtd}/\rangle$

(z)

L^2

Results

- The most popular expressions in classifications:
 - 83 (Relativity and gravitational theory):

$\langle \text{mtd}/\rangle$

$\langle \text{mrow}\rangle\langle \text{mo}\rangle-\langle \text{mo}\rangle\langle \text{mn}\rangle 1\langle \text{mn}\rangle\langle \text{mrow}\rangle$

$\langle \text{mrow}\rangle\langle \text{mo}\rangle(\langle \text{mo}\rangle\langle \text{mi}\rangle t\langle \text{mi}\rangle\langle \text{mo}\rangle)\langle \text{mo}\rangle\langle \text{mrow}\rangle$

$\langle \text{mrow}\rangle\langle \text{mo}\rangle(\langle \text{mo}\rangle\langle \text{mi}\rangle x\langle \text{mi}\rangle\langle \text{mo}\rangle)\langle \text{mo}\rangle\langle \text{mrow}\rangle$

$\langle \text{mrow}\rangle\langle \text{mo}\rangle(\langle \text{mo}\rangle\langle \text{mn}\rangle 1\langle \text{mn}\rangle\langle \text{mo}\rangle)\langle \text{mo}\rangle\langle \text{mrow}\rangle$

Results

- Some antiunifiers:

- 00 (General):

```
<mfrac><mi></mi><orccaPattern index="0"/></mfrac>
```

```
<mfrac><mi>k</mi><orccaPattern index="0"/></mfrac>
```

```
<mfrac><mn>1</mn><mrow><mn>1</mn><mo>-</mo><orccaPattern index="0"/>
  </mrow></mfrac>
```

```
<mfrac><mn>1</mn><mrow><mn>5</mn><mo></mo><mn>1</mn><msup><mn>0</mn>
  <orccaPattern index="0"/></msup></mrow></mfrac>
```

```
<mfrac><mrow><mi>d</mi><orccaPattern index="0"/></mrow><mrow><mi>d</mi>
  <orccaPattern index="1"/></mrow></mfrac>
```

```
<mfrac><mn>25</mn><orccaPattern index="0"/></mfrac>
```