Discovering Namespaces in Mathematical Notation

Moritz, Alexey, Sergey, Marcus, Howard, Volker

ABSTRACT

While modern programming languages use name-spaces for means of modularity and expandability, mathematical notation has no such concept. However, in most scientific communities a standard notation for mathematics has been established. We claim that the sharing of notation corresponds to the taxonomic distance of the research fields. Nowadays, where digital communication plays a significant role in the transportation of concepts and ideas expressed using mathematical notion, we see advantages in using name-spaces for mathematical notation to reduce ambiguity and increase the widespread of ideas across community boarders. In this paper, we extract identifier-definition-tuples from Wikipedia, and map them to classification sachems for mathematics and physics. Thereby, we get a hierarchy of identifier definition tuples for pairs. In addition, we investigate scientific articles from arXiv to test our method on a more specialized corpus.

1. MOTIVATION

1.1 The vision of namespaced Mathematics

Motivation and introduction goes here

Namespaces turned out to be useful for Programming Languages so they'll probably be useful for Mathematics as well.

Use cases:

- Enhance Math Search:

A certain level of semantics is required to search for mathematical expression. Two prototypical problems are identifier disambiguation (does E stand for *Energy* or *Expectation Value*) and canonicalization of synonymously used identifiers such as σ or SD, which both denote standard deviation.

- Enhance Math Paper writing:

Fixed naming conventions simplify the writing process. Basis for visual editors. Possibility to create semantic LaTeXsources

- Enhance readers experience:

Annotate meanings of identifiers in equations via tool-tips.

2. BACKGROUND

Math meats information retrieval

Keep this section brief. Find additional references from other research fields like linguistics?

3. OUR METHOD

3.1 The Machinery for Namespace Discovery

Brief description of the tools
our schema: documents(id,content)
formulae(id, content, documentId)
identifier(id)
formulaeIdentifer(identiferId, formulaeId)
documentIdentifier(identifierId, documentId)
categorization(id, parentId)
documentCategory(documentId, categorizationId)

1. Identifier identification in formulae
2. Identifier extraction from text
3. Identifier definition tuple candidate extraction
4. formulae, d

3.2 The Wikipedia Case Study

Result of the Master Thesis Fuse bilingual result.

3.3 Name spacing the ArXiv

Needs to be done. In general this follows the same pattern.

Discuss with Deyan Ginev to get the full HTML5 corpus.

See if clustering methods can cope with the data volume? 100M formula probably 1 billion identifiers.

4. RESULTS

5. DISCUSSION

5.1 Learning outcome: Namespaced Identifiers

Write evaluation

6. CONCLUSION

A long road ahead

Write future work

7. ACKNOWLEDGEMENT

APPENDIX

	E	m	c	λ	σ	μ
Linear algebra	matrix	matrix	scalar	eigenvalue	related permutation	algebraic multiplicity
General relativity	energy	mass	speed of light	length	shear	reduced mass
Coding theory	encoding function	message	transmitted codeword		natural isomorphisms	
Optics		order fringe	speed of light in vacuum	wavelength	conductivity	permeability
Probability	expectation	sample size		affine parameter	variance	mean vector

TableTable 1: Definitions for selected identifiers and namespaces extracted from the English Wikipedia.