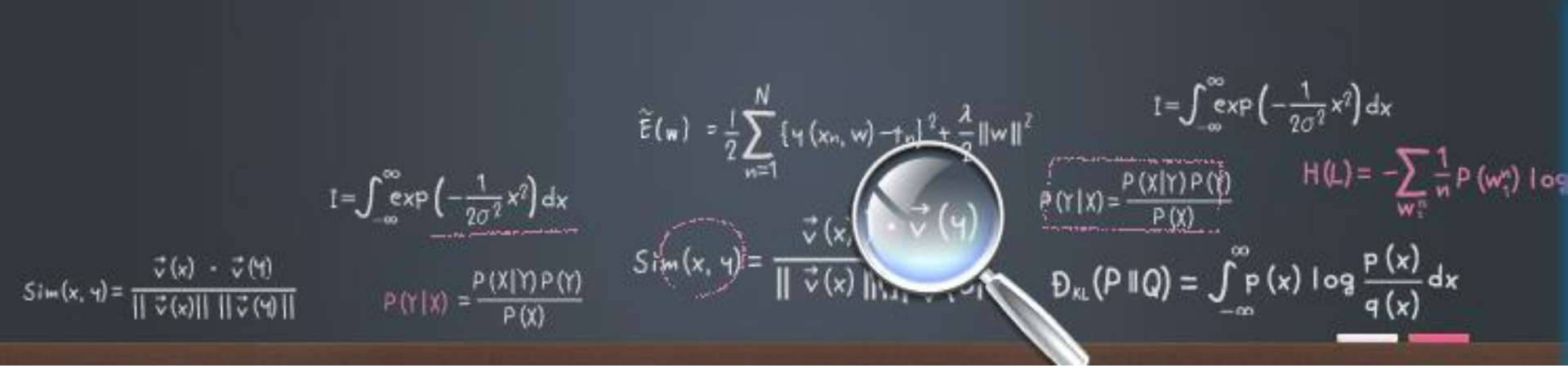


Which One Is Better: Presentation-Based or Content-Based Math Search?

Minh-Quoc NGHIEM, Giovanni Yoko KRISTIANTO,
Goran TOPÍC, Akiko AIZAWA

The Graduate University for Advanced Studies
Sokendai

TRUONG DAI HOC KHOA HOC TU NHIEN
KHTN
TP. HO CHI MINH

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

Outline

- Introduction
- Math Search Systems
- Method
- Evaluation
- Conclusion

Introduction

- Math Search
  - Presentation-based
    - LaTeX
    - Presentation MathML
  - Content-based
    - Content MathML
    - OpenMath
- NTCIR Math Track
  - http://ntcir-math.nii.ac.jp/

Introduction

- Content-based systems use SnuggleTeX or LaTeXML for semantic enrichment
- No evaluation of how semantic enrichment module contribute to search system
- Which one is better: content-based search or presentation-based search

Mathematical Search Systems

- Presentation-based systems
  - Springer LaTeX Search
  - MathFind
  - The Digital Library of Mathematical Functions
  - EgoMath
  - Math Indexer and Searcher
  - ActiveMath
  - …

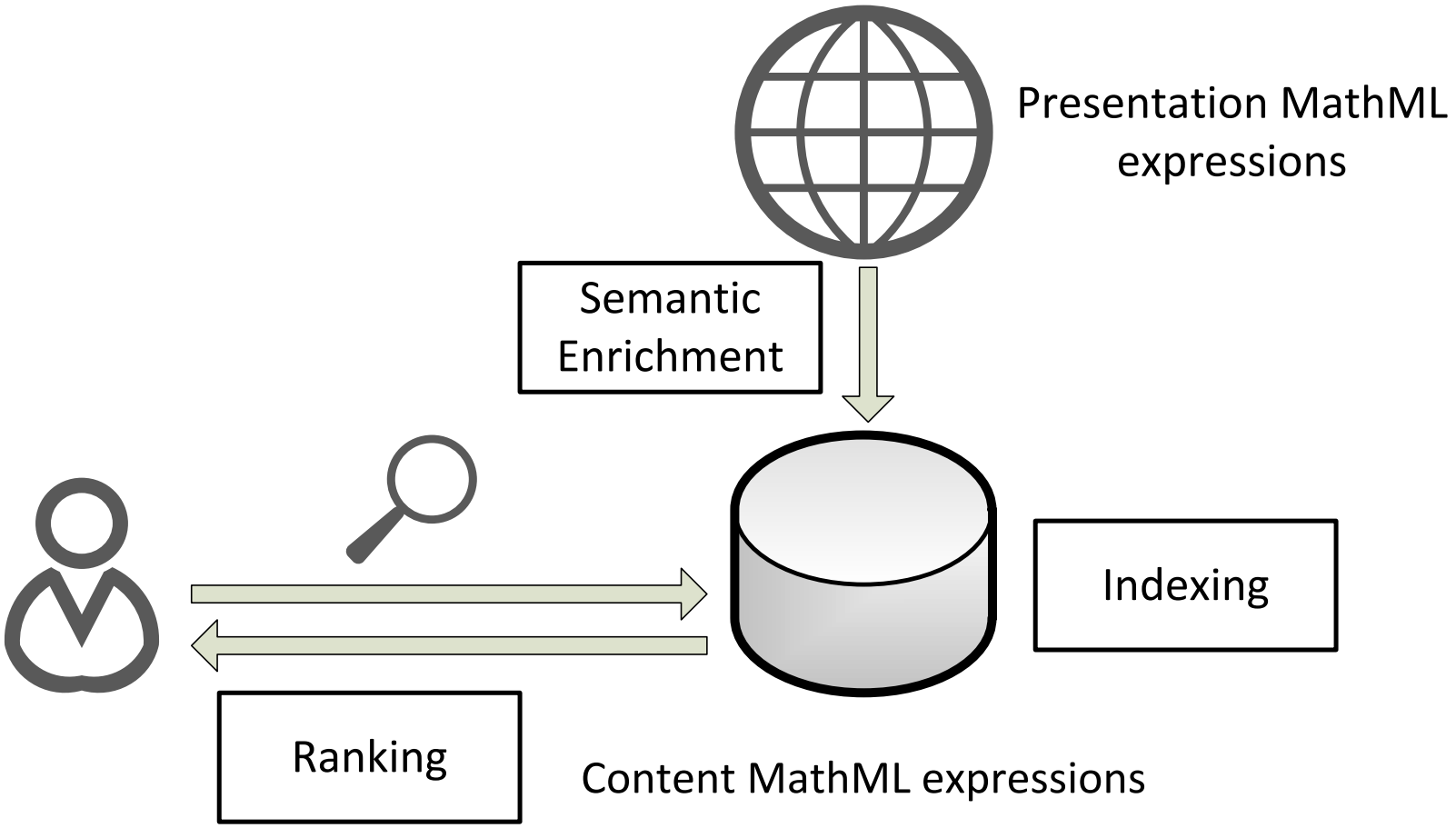
Mathematical Search Systems

- Content-based systems
  - Wolfram Function
  - MathWebSearch
  - MathGO!
  - MathDA
  - The system of Nguyen et. al
  - …

Method

- Use Semantic Enrichment module to convert Presentation to Content MathML
- Use Content MathML for Indexing
- Allow user to input query in Presentation MathML

System framework



Semantic Enrichment

- Semantic Enrichment method of Nghiem et. al (CICM 2013)
  - Segmentation rules: segment Presentation MathML trees into smaller trees
  - Translation rules: translate Presentation MathML trees to Content MathML trees
  - Each rule is associated with a probability

Indexing

- Indexing method of Topic et. al (NTCIR 2013)
  - Opaths: path in XML tree with order
  - Upaths: no order
  - Sisters: sister nodes in subtree

Evaluation

- Data
  - 20k Math expressions in WFS
  - 15 queries (modified from NTCIR)
- Systems
  - Presentation MathML (PMathML)
  - Content MathML (CMathML)
  - Semantic Enrichment (SE)

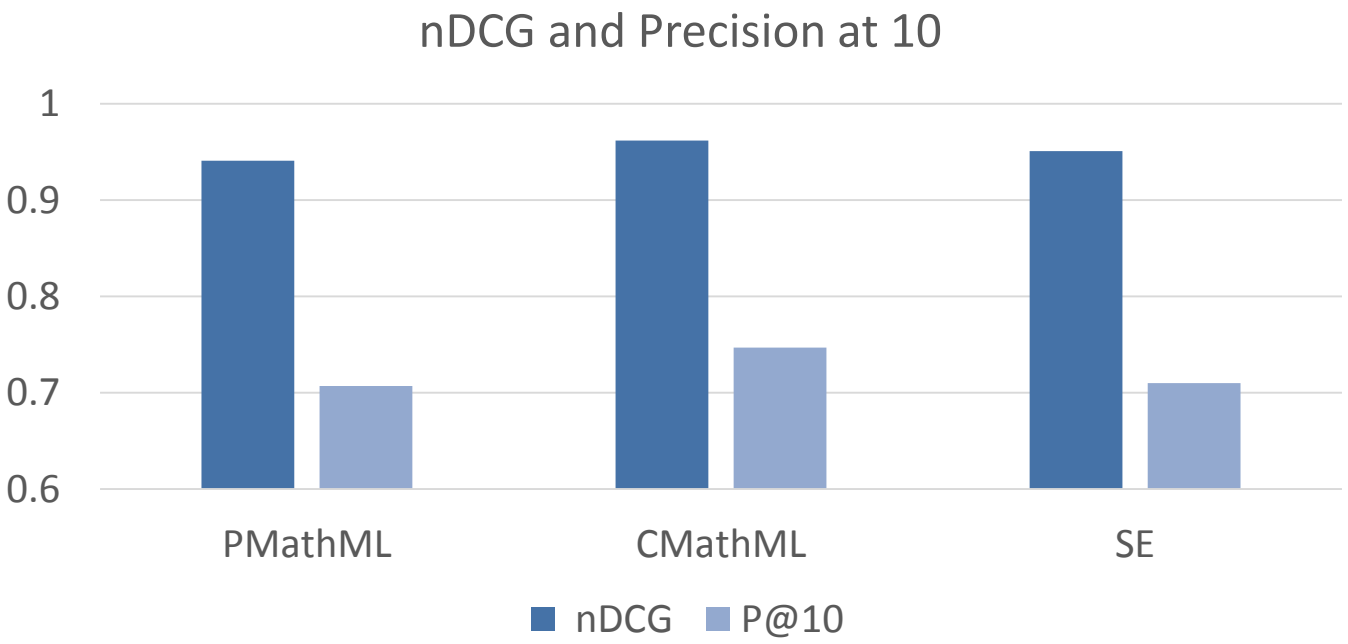
Evaluation

- Metrics
  - Precision at 10 (P@10)
    - Precision in top k results
  - Normalize Discounted Cumulative Gain (nDCG)
    - Ranking quality

Queries

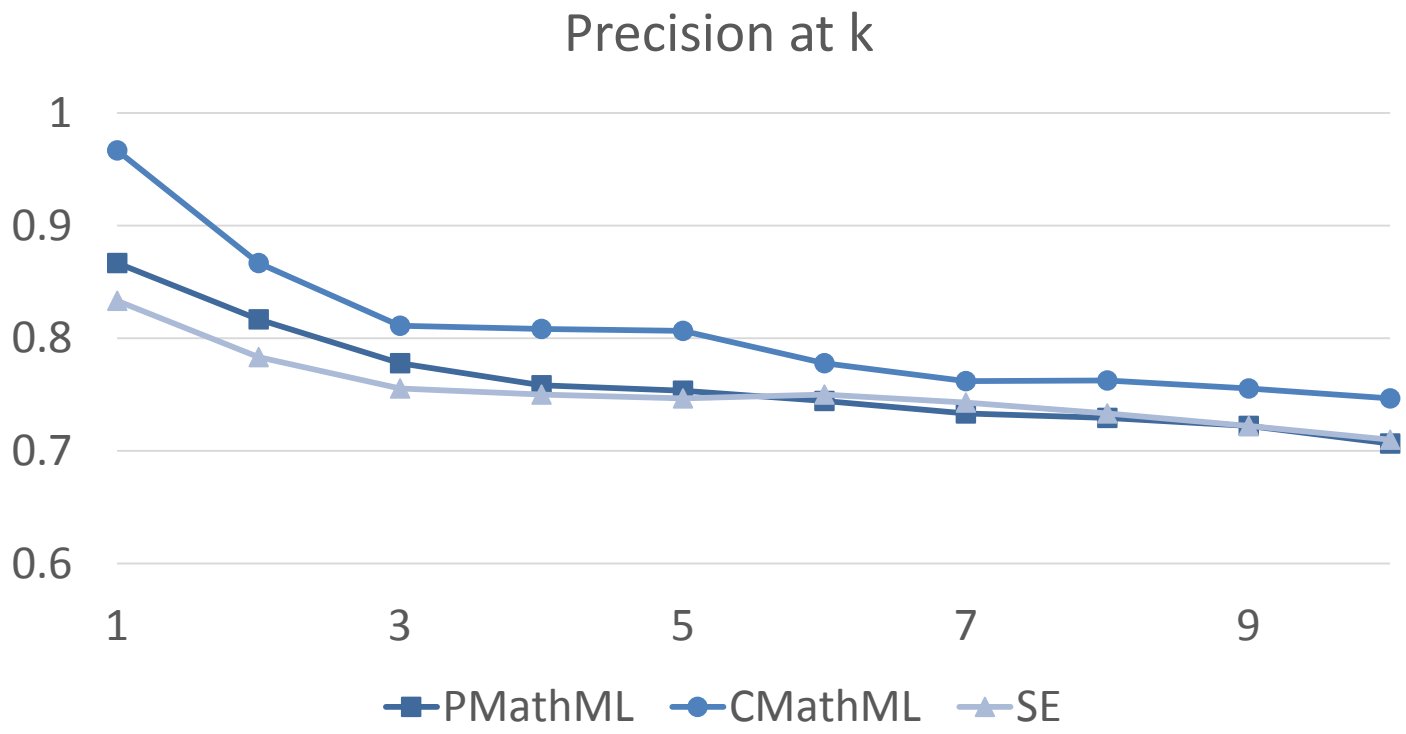
$\int_0^\infty x dx$	$x^2 + y^2$	$\int_0^\infty e^{-x^2}\, dx$
$arcsin(x)$	k^2	$\frac{\cosh ez + \sinh ez}{e}$
$\mathcal{R}_z \psi^\nu(z), \widetilde{\infty}$	$\int \frac{a^{d+bz}}{z} dz$	$\lim_{\nu \to \infty} \frac{L_{\alpha+\nu}}{L_\nu}$
$\mathcal{BP}_z \mathfrak{P}_\nu^\mu(z)$	$\nu \in \mathbb{N}$	$\psi^\nu(z)$
$\log(z+1)$	$H_n(z)$	$\frac{1}{\pi}\int_0^\pi \cos tn - z\sin t$

Evaluation: search performance



Using content markup improve search performance

Evaluation: search performance



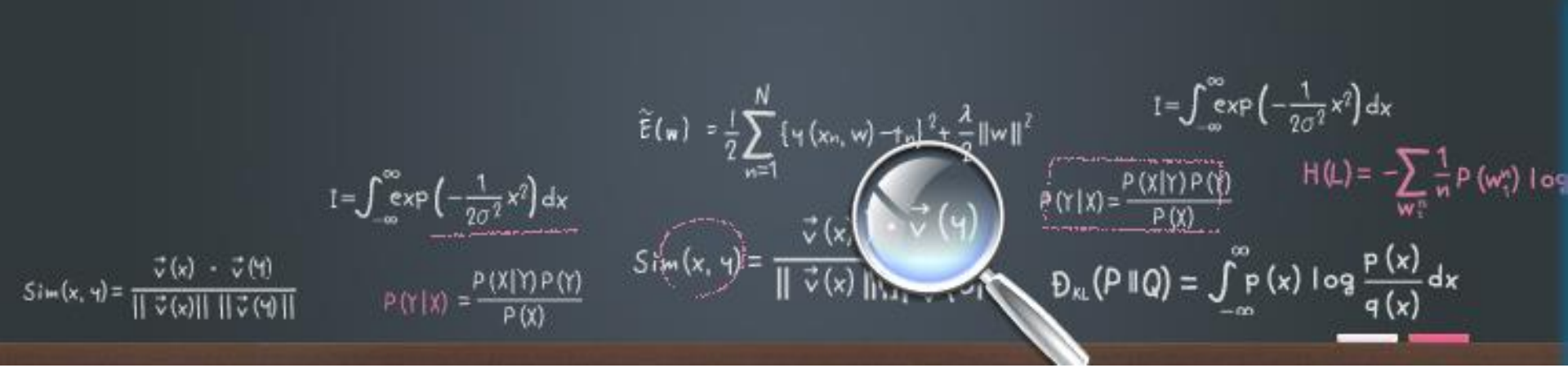
Using content markup improve search performance
Relevant results are ranked higher

PMathML and SE systems

- SE system is better
  - Functions have specific meanings
    - Poly-Gamma, Hermite-H
  - More than one way to represent math expression
    - Sin^{-1} and Arcsin
- PMathML system is better
  - Elementary functions
    - Power, Logarithm, Trigonometric functions

Summary

- Content-based math search is better than presentation-based math search
- Performance of semantic enrichment module affect the math search performance
- Both presentation-based and content-based systems have their strong points



ThΔnk you

∫or your Δtteηtioη!