Semantic Extraction and Enrichment of Natural Language and Mathematical Discourse for Mathematical Search

Ştefan Anca

KWARC - Knowledge Adaptation and Reasoning for Content
Jacobs University, Bremen, Germany

September 9, 2009

Context: Scientific Documents with Math Formulas

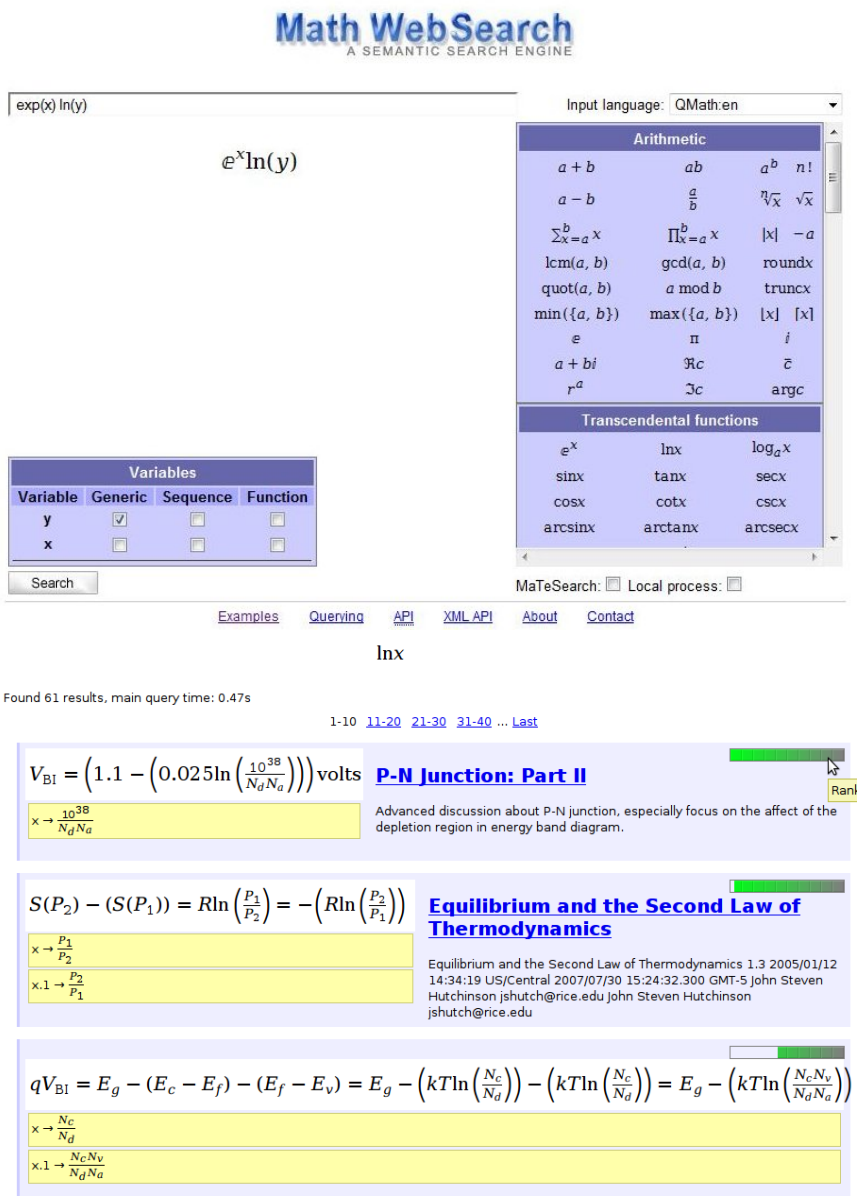
Corpora of interest:

- arXMLiv translated corpus from the Cornell University arXiv Library
  - Presentation MathML or intermediary XML math (XMath) → **non-semantic** formats (not useful for search)
  - Over 400000 converted documents
  - **Very high incidence** of mathematical formulas and standard formulations (theorems, definitions, proofs, etc.)
- Connexions online platform for publishing user content
  - Content MathML→ **semantic** format (useful for search)
  - Over 12000 documents, 3400 with semantic mathematical formulas
  - **Low incidence** of standard mathematical formulations

Tools employed:

- LaTeXML - converts arXiv (arxiv.org) into arXMLiv (arxmliv.kwarc.info) (LaTeX math ⇒ Presentation MathML)
- The LaMaPUn Architecture - semantic enrichment of LaTeXML output (XMath + context + user ⇒* Content MathML)

MathWebSearch - Search for Mathematical Structure



http://search.mathweb.org

- Can only index formulas in semantic representation (Content MathML or OpenMath)
- Stores all the mathematical terms by their structure in a substitution-tree
- Provides instantiation, generalization and unification search
- Sentido editor for math query input
- Indexes the Connexions repository ($\sim$ 85000 terms)

kwarc JACOBS UNIVERSITY

Applicable Theorem Search - Search for Theorems

Theorem: **If** we take $n \geq 0$, **then** we know that $\sum_{i=0}^{n} i = \frac{n\cdot(n+1)}{2}$.

Query: $\sum_{k=0}^{25} k$

Match: $\sum_{i=0}^{n} i = \frac{n\cdot(n+1)}{2}$, $i \to k$, $n \to 25$.

Conclusion: we know that $\sum_{i=0}^{n} i = \frac{n\cdot(n+1)}{2}$.

Hypothesis: we take $n \geq 0$

http://betasearch.mathweb.org

- Search for *fixed-structure natural language patterns* (idioms) which express theorem relations: **if** H **then** C, **let** H **then** C, etc.
- Use idioms as natural language patterns for semantic information extraction
- Find and extract (index) theorems with mathematical universals
- Use MathWebSearch generalization search on queries with constants to retrieve applicable theorems
- Use natural language context to infersemantics about mathematical content:
  For all x, **there exists** a y, **such that** $4^x = 2^y$

Knowledge Adaptation and Reasoning for Content

Group Details

- Homepage: http://kwarc.info
- Based at **Jacobs University**, Bremen, Germany
- Led by **Prof. Dr. Michael Kohlhase**
- Main Research Focus: knowledge representation with a view towards applications in knowledge management, especially for documents with mathematical content

Projects involving:

- Representing documents with mathematical content through semantic mark-up (**OMDoc** (http://omdoc.org), **sTeX**), **browsing and annotating** (**SWiM, panta rhei, CPoint**)
- Management of change for structured documents (**locutor, TNTBase, CCWord**)
- Semantic extraction from XML documents (**Krextor, Idiom Spotter**)
- Semantic enrichment of mathematical terms in XML documents (**LaMaPUn**)
- Processing, validating and rendering OMDoc documents (**JOMDoc, MMT**)
- Integrating web services into interactive mathematical documents (**JOBAD**)
- Converting the arXiv database of LaTeX documents to an XML format (**arXMLiv**)
- Semantic search on XML documents with mathematical content (**MathWebSearch, MaTeSearch, Applicable Theorem Search**)

Idiom Spotter - Semantic Extraction from Informal Math

Corpus	Connexions	Saarbrücken
Total files	11712	10239
Files with idioms	451	9947
Idioms found	1794	215044
Avg idioms per file	0.15	21

Idiom	Freq. Cnx	Freq. Saarb
assume H1 then C1	29	1755
conclude D1 is D2	22	3176
define D1 to be D2	58	4911
given H1 then C1	43	1809
H1 if and only if C1	56	25979
H1 implies C1	170	30593
C1 only if H1	102	27964
C1 only when H1	35	1553
if H1 then C1	1195	108633
let H1 then C1	61	6915
suppose H1 then C1	23	1756
Theorem patterns	**1714**	**206957**

- **Idiom**: natural language formulation which follows a certain fixed word and syntax pattern
- Extract semantic relations from scientific texts $\longrightarrow$ **structured knowledge**
- **Connexions corpus** - user-authored online content
- **Saarbrücken corpus** - selection of math publications from arXMLiv
- All idioms extracted contain at least a mathematical term
- **Saarbrücken corpus** obviously better for Theorem extraction
- No correct content representation of math in Saarbrücken corpus!

Language and Mathematics Processing and Understanding

The LaMaPUn Architecture: A project pursuing semantic enrichment, ambiguity resolution of mathematics in the arXMLiv corpus.

- Semantically enrich the XML math outputted by an initial stage of LaTeXML, to reach content-level semantics ⇒ MathML, OpenMath
- **Preprocessing**: correct the math-related human encoding mistakes (e.g. "1^{st}", "{\bf x} - {\bf y}", $last(x)$ → $l \cdot a \cdot s \cdot t(x)$)
- **Semantic Blackboard**: represent the XML documents in an RDF Database
- **Semantic Analysis Modules**: plug into the Blackboard and perform semantic processing, results stored as stand-off annotations
  - Mathematical Formula Disambiguation
  - Content-Based Formula Disambiguation
- **Semantic Result** and **Output Generation**: merge annotations with original documents to obtain the semantically enriched result, outputted as XHTML or OMDoc with Content/Presentation MathML.
- **Visualization and Feedback**: allow users/authors to review/correct inferred semantics

LaMaPUn Architecture