

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

Ștefan 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 \Rightarrow Presentation MathML)
- The **LAMAPUN Architecture** - semantic enrichment of LATEXML output (XMATH + context + user \Rightarrow * Content MATHML)

MATHWEBSEARCH - Search for Mathematical Structure

Math WebSearch
A SEMANTIC SEARCH ENGINE

exp(x) ln(y)

$e^x \ln(y)$

Input language: QMathen

Arithmetic

$a + b$	ab	a^b	$n!$
$a - b$	$\frac{a}{b}$	$\sqrt[n]{a}$	\sqrt{a}
$\sum_k = a^k$	$\prod_k = a^k$	$ x $	$-a$
$\text{lcm}(a, b)$	$\text{gcd}(a, b)$	$\text{round}(x)$	
$\text{quot}(a, b)$	$a \bmod b$	$\text{trunc}(x)$	
$\min(a, b)$	$\max(a, b)$	$ x $	$ x $
e	π	i	
$a + bi$	$\Re c$	$\Im c$	
r^a	$\Im c$	argc	

Transcendental functions

e^x	$\ln x$	$\log_a x$
$\sin x$	$\tan x$	$\sec x$
$\cos x$	$\cot x$	$\csc x$
$\arcsin x$	$\arctan x$	$\text{arcsec} x$

Variables

Variable	Generic	Sequence	Function
y	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
x	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Search

MathSearch: Local process:

Examples Queries API XML API About Contact

lnx

Found 61 results, main query time: 0.47s

1-10 11:20 21:30 31:40 -- Last

$$V_{BI} = \left(1.1 - \left(0.025 \ln \left(\frac{10^{38}}{N_d N_a} \right) \right) \right) \text{volts}$$

P-N Junction: Part II

Advanced discussion about P-N junction, especially focus on the affect of the depletion region in energy band diagram.

$$S(P_2) - (S(P_1)) = R \ln \left(\frac{P_2}{P_1} \right) = - \left(R \ln \left(\frac{P_1}{P_2} \right) \right)$$

Equilibrium and the Second Law of Thermodynamics

Equilibrium and the Second Law of Thermodynamics 1.3 2005/01/12
14:34:19 UTCentral 2007/07/03 15:24:32 308 GMT-5 John Steven
mathchris@hutch.ch@rice.edu john.steven@hutch.ch@rice.edu
jshutch@rice.edu

$$qV_{BI} = E_g - (E_c - E_f) - (E_f - E_v) = E_g - \left(kT \ln \left(\frac{N_c}{N_d} \right) \right) - \left(kT \ln \left(\frac{N_v}{N_a} \right) \right) = E_g - \left(kT \ln \left(\frac{N_c N_v}{N_d N_a} \right) \right)$$

<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 (~ 85000 terms)

APPLICABLE THEOREM SEARCH - Search for Theorems

Found 95 results, main query time: 0.105s

1:10 11:20 21:30 31:35

The Functions e^x and e^{-x} : The Euler and De Moivre Identities

The Functions e^x and e^{-x} : The Euler and De Moivre Identities: The Functions e^x and e^{-x} : The Euler and De Moivre Identities 1.4 2009/02/26 23:29:51 US:Central 2009/09/27 13:40:47.453 GMT-5

Idioms: if H1 then C1

Conclusion: we see that Pascal's triangle keeps track of the number of occurrences of $x^n - k y^k$.

Hypotheses: we think of a left-hand path as an occurrence of an x and a right-hand path as an occurrence of a y .

Scope: if we think of a left-hand path as an occurrence of an x and a right-hand path as an occurrence of a y , then we see that Pascal's triangle keeps track of the number of occurrences of $x^n - k y^k$.

DFT and FFT: An Algebraic View

DFT and FFT: An Algebraic View: Polynomial Algebras and the DFT/Polynomial Algebra/Chinese Remainder Theorem (CRT)/Polynomial Transform/DFT as a Polynomial Transform/Algebraic Derivation of the Cooley-Tukey

Idioms: if H1 then C1

Conclusion: \mathbb{F}^n can be expressed as a matrix.

Hypotheses: we choose bases b, c, d in the three polynomial algebras.

Scope: if we choose bases b, c, d in the three polynomial algebras, then \mathbb{F}^n can be expressed as a matrix.

Theorem: If we take $n \geq 0$, then we know that

$$\sum_{i=0}^n i = \frac{n \cdot (n+1)}{2}.$$

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

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

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

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 infer semantics 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**, **pantarrhei**, **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 → **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 L^AT_EX_ML, to reach content-level semantics \Rightarrow MathML, OpenMath
- **Preprocessing:** correct the math-related human encoding mistakes (e.g. “ $\$1^{\{st\}}\$$ ”, “ $\{\bf x\} - \{\bf y\}$ ”, $\$last(x)\$ \rightarrow l \cdot a \cdot s \cdot t(x)$)
- **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

LaMaPU Architecture

