Searching the Space of Mathematical Knowledge

Michael Kohlhase & Mihnea Iancu

http://kwarc.info/kohlhase
Center for Advanced Systems Engineering
Jacobs University Bremen, Germany

Classical Math Search Engines
Instead of a Demo: Searching for Signal Power

Math WebSearch
A SEMANTIC SEARCH ENGINE

Search for:
\[ \int e^{nx} \, dx \]

XML Query | String

Examples | Help | API | About | Contact

Kohlhase & Iancu: Searching the Math Knowledge Space 3 MIR 2012
Instead of a Demo: Search Results

Other integrals (5 formulas)

Matched term:

\[
\int \frac{e^{3z/4}}{\left( -2 + e^{3z/4} \right) \sqrt{-2 + e^{3z/4} + e^{3z/2}}} \, dz = \frac{2}{3} \left( \log(-2 + e^{3z/4}) - \log\left(4\sqrt{-2 + e^{3z/4} + e^{3z/2}} + 5e^{3z/4} - 2\right) \right)
\]

Rank: 100%

Used substitution:

\[
\begin{align*}
n & \rightarrow 3z4^{-1} \\
r & \rightarrow \left( \left( -2 + e^{3z4^{-1}} \right) \left( -2 + e^{3z4^{-1}} + e^{3z2^{-1}} \right)^{1/2} \right)^{-1} \\
x & \rightarrow z
\end{align*}
\]

XML Source
Instead of a Demo: \texttt{\LaTeX\textregistered}-based Search on the arXiv

\[
\lim_{x \to 0} \frac{m[h(t)]}{h_0}.
\]

Examples - \LaTeX\texttt{\textregistered} queries
- Generic subscript search
- Specific subscript search
- Specific integral search
- Physical constant search
- All limits approaching zero
- Text in math search

Generalized off-equilibrium fluctuation–dissipation relations in random Ising systems
Author: Federico Ricci-Tersenghi <ricci@chimera.roma1.infn.it>

Behavior of the reflection function of a plane-parallel medium for directions of incidence and reflection tending to horizontal directions
Author: Daphne Stam <d.m.stam@srn.nl>

Behavior of the reflection function of a plane-parallel medium for directions of incidence and reflection tending to horizontal directions
Instead of a Demo: Applicable Theorem Search in Mizar

definition
let k, n be Ordinal;
pred k divides n means :Def3: :: MTEST1: def 3
ex a being Ordinal st n = k ^ a;
reflexivity
proof
let n be Ordinal; :: thesis:
thus ex a being Ordinal st n = n ^ a ;

ATP Proof not found

status: Timeout
Suggest hints, Unification query,

Suggested hints
t73_card_2, t39_ordinal2,

Try SPASS, Export problem to SystemOnTPTP

:: thesis:
end;
• **Classical Setup:** they all work more or less the same:
  - crawl the resources (the Web or a corpus)
  - index the search-relevant information (formulae, words, structures, . . .)
  - process user queries (via tf/idf, unification, . . .)
  - rank/process the hits (needs work!)

• **Question:** Is this enough for the working Mathematician?

• **Answer:** depends on what you want.
  - Yes, if we restrict ourselves to what is explicitly written in books, papers, etc.
  - No, if we are looking for “Mathematical Knowledge”! (and I claim we should be)

• **Observation 1** *Mathematical knowledge is induced by combinations of explicitly represented facts.* (that’s why we usually ask humans)

• **Example 2** Combine mathematical facts (no, we don’t need theorem proving!)
  - **Theorem 3.1:** *Idempotent monoids are Abelian.* (from course Algebra I)
  - **Lemma 2:** *(S, #) is an associative, unital, idempotent magma.* (you just found out)
  - Search for \( x \# y = y \# x \) (Find it as an instance of Theorem 3.1)
Modular Representation of Mathematics
Modular Representation of Math (Theory Graph)

• **Idea**: Follow mathematical practice of generalizing and framing
  • framing: If we can view an object $a$ as an instance of concept $B$, we can inherit all of $B$ properties (almost for free.)
  • state all assertions about properties as general as possible (to maximize inheritance)
  • examples and applications are just special framings.

• Modern expositions of Mathematics follow this rule (radically e.g. in Bourbaki)
• formalized in the **theory graph paradigm** (little/tiny theory doctrine)
  • theories as collections of symbol declarations and axioms (model assumptions)
  • theory morphisms as mappings that translate axioms into theorems

• **Example 3 (MMT: Modular Mathematical Theories)** MMT is a foundation-indepent theory graph formalism with advanced theory morphisms.

**Problem**: With a proliferation of abstract (tiny) theories readability and accessibility suffers (one reason why the Bourbaki books fell out of favor)
Modular Representation of Math (MMT Example)

\[ \text{NatNums} \]
\begin{align*}
\text{NatNums} &::= \mathbb{N}, s, 0 \\
P1, \ldots P5
\end{align*}

\[ \vartheta = \begin{cases} 
m \mapsto e \\
a \mapsto c \end{cases} \]

\[ \psi = \begin{cases} 
G \mapsto \mathbb{Z} \\
o \mapsto + \\
e \mapsto 0 \end{cases} \]

\[ \psi' = \begin{cases} 
i \mapsto - \\
g \mapsto f \end{cases} \]

\[ \phi = \begin{cases} 
G \mapsto \mathbb{N} \\
o \mapsto \cdot \\
e \mapsto 1 \end{cases} \]

\[ \psi = \begin{cases} 
G \mapsto \mathbb{Z} \\
o \mapsto + \\
e \mapsto 0 \end{cases} \]

\[ \psi' = \begin{cases} 
i \mapsto - \\
g \mapsto f \end{cases} \]
The MMT Module System

• **Central notion:** theory graph with theory nodes and theory morphisms as edges

• **Definition 4** In MMT, a **theory** is a sequence of constant declarations – optionally with type declarations and definitions

• MMT employs the Curry/Howard isomorphism and treats
  • axioms/conjectures as typed symbol declarations (propositions-as-types)
  • inference rules as function types (proof transformers)
  • theorems as definitions (proof terms for conjectures)

• **Definition 5** MMT had two kinds of theory morphisms
  • **structures** instantiate theories in a new context (also called: definitional link, import)
    they import of theory \( S \) into theory \( T \) induces theory morphism \( S \to T \)
  • **views** translate between existing theories (also called: postulated link, theorem link)
    views transport theorems from source to target (framing)

• together, imports and views allow a very high degree of re-use

• **Definition 6** We call a statement \( t \) **induced** in a theory \( T \), iff there is
  • a path of theory morphisms from a theory \( S \) to \( T \) with (joint) assignment \( \sigma \),
  • such that \( t = \sigma(s) \) for some statement \( s \) in \( S \).

• In MMT, all induced statements have a canonical name, the **MMT URI.**
Searching for Induced statements
Simple Idea: We have all the necessary components: MMT and MathWebSearch

Definition 7 The bsearch system is an integration of MathWebSearch and MMT that
- computes the induced formulae of a modular mathematical library via MMT (aka. flattening)
- indexes induced formulae by their MMT URIs in MathWebSearch (hits are MMT URIs)
- uses MathWebSearch for unification-based querying (compute the actual formula)
- uses the MMT to present MMT URI
- generates explanations from the MMT URI of hits.

Implemented by Mihnea Iancu in ca. 10 days (MMT harvester pre-existed)
- almost all work was spent on improvements of MMT flattening
- MathWebSearch just worked (web service helpful)
Recall: \texttt{bsearch (MathWebSearch really)} returns a MMT URI as a hit.

Question: How to present that to the user? \hspace{1cm} (for his/her greatest benefit)

Fortunately: MMT system can compute induced statements (the hits)

Problem: Hit statement may look considerably different from the induced statement

Solution: Template-based generation of NL explanations from MMT URIs. MMT knows the necessary information from the components of the MMT URI.
Modular Representation of Math (MMT Example)

\[
\begin{align*}
\text{IntArith} & : \mathbb{Z} := \mathbb{N} \cup -\mathbb{N} \\
& -: n+0=n, n+s(m) = s(n+m) \\
& \times: n\cdot 1 = n, n\cdot s(m) = n\cdot m + n \\
\text{NatArith} & : +, \cdot \\
& n+0=n, n+s(m) = s(n+m) \\
& n\cdot 1 = n, n\cdot s(m) = n\cdot m + n \\
\text{NatNums} & : \mathbb{N}, s, 0 \\
& P_1, \ldots, P_5 \\
\end{align*}
\]

\[
\begin{align*}
\varphi &= \{ G \mapsto \mathbb{N}, o \mapsto \cdot, e \mapsto 1 \} \\
\psi &= \{ G \mapsto \mathbb{Z}, o \mapsto +, e \mapsto 0 \} \\
\psi' &= \{ i \mapsto -, g \mapsto f \} \\
\end{align*}
\]
Example 8. Search search result u?IntArith?c/g/assoc for query 
\[(\overline{\overline{x}} + \overline{\overline{y}}) + \overline{\overline{z}} = \overline{\overline{R}}.\]

- localize the result in the theory u?IntArithf with

  \textit{Induced statement} \( \forall x, y, z : \mathbb{Z}. (x + y) + z = x + (y + z) \) found in http://cds.OMDoc.org/cds/elal?IntArith \textit{(subst, justification)}.

- Justification: from MMT info about morphism c (source, target, assignment)

  \textit{IntArith is a CGroup if we interpret} \( \circ \) \textit{as} + \textit{and} \( G \) \textit{as} \( \mathbb{Z} \).

- skip over g, since its assignment is trivial and generate

  \textit{CGroups are SemiGrps by construction}

- ground the explanation by

  \textit{In SemiGrps we have the axiom} assoc : \( \forall x, y, z : G. (x \circ y) \circ z = x \circ (y \circ z) \)
The LATIN Logic Atlas

- **Definition 9** The LATIN project (Logic Atlas and Integrator) develops a logic atlas, its home page is at http://latin.omdoc.org.

- **Idea**: Provide a standardized, well-documented set of theories for logical languages, logic morphisms as theory morphisms.

```
Isabelle --→ IMPS --→ PVS

skl1      partial1   sthol

pl1       skl0       stlc

pl0       lambda-calc

ind       truthval   undef   subst   simple-types

dep-types
```

- **Technically**: Use MMT as a representation language logics-as-theories
- **Integrate** logic-based software systems via views.
LATIN: Representing Logics and Foundations as Theories

- Logics and Foundations as Theories:
  - Logics and foundations represented as theories
  - Meta-relation between theories
  - Models represented as theory morphisms
  - e.g. $v_1$ interprets monoid in integers using meta-morphism $v_3$

- The LATIN atlas in numbers: it currently contains (tiny theories approach)
  - 449 theories with 2310 symbol declarations (avg. = 5.14 declarations/theory)
  - and 1072 direct imports (including metas) (avg = 2.39 imports/theory)
  - 382 views between theories.
  - Size: 123.9 MB in native OMDoc format
search on the LATIN Logic Atlas

- Flattening the LATIN Atlas (once):

<table>
<thead>
<tr>
<th>type</th>
<th>modular</th>
<th>flat</th>
<th>factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>declarations</td>
<td>2310</td>
<td>58847</td>
<td>25.4</td>
</tr>
<tr>
<td>library size</td>
<td>23.9 MB</td>
<td>1.8 GB</td>
<td>14.8</td>
</tr>
<tr>
<td>math sub-library</td>
<td>2.3 MB</td>
<td>79 MB</td>
<td>34.3</td>
</tr>
<tr>
<td>MathWebSearch harvests</td>
<td>25.2 MB</td>
<td>539.0 MB</td>
<td>21.3</td>
</tr>
</tbody>
</table>

- simple bsearch frontend at http://cds.omdoc.org:8181

---

FlatSearch DEMO

\[ X + Y \]


assoc:== (+ (+ X Y) Z) (+ X (+ Y Z))

Justification
Induced statement found in http://latin.omdoc.org/math?IntArith
IntArith is a AbelianGroup if we interpret over view &
AbelianGroup contains the statement assoc

Conclusions and Recap

- From searching documents to searching knowledge spaces!
- bSearch implemented from existing components
  - MMT for modular representations of mathematical knowledge
  - MMT URIs name induced statements
  - flattening to compute all induced statements
  - generate human-oriented explanations of induction paths
- Prototypical implementation for the LATIN logic atlas
- Future work: we have only just begun \((\text{most work in MMT though})\)
  - Flattening away other language features, e.g. patterns \((\leadsto F. \text{ Horozal})\)
  - Avoiding duplication from structures.
  - Integrating graph structure constraints into MathWebSearch
  - Extending MMT (and flattening) to informal Math! \((\text{redo Bourbaki})\)