MathWebSearch at NTCIR-11: Keywords, Frontend, & Scalability

Radu Hambasan & Michael Kohlhase & Corneliu Prodescu

http://kwarc.info/kohlhase
Computer Science
Jacobs University Bremen, Germany

NTCIR-11, December 11. 2014
Math Markup e.g. in MathML and \( \text{\LaTeX} \)

- MathML3 is a W3C Recommendation for representing Formulae \([ABC^{+10}]\)
- Idea: Combine the presentation and content markup and cross-reference

\[
\frac{3}{(x+2)}
\]

- use e.g. for semantic copy and paste.

- But: Formulae are mostly written in \( \text{\LaTeX} \), e.g. \( \frac{3}{(x+2)} \)

- Solution: Write \( \text{\LaTeX} \), convert to HTML5 \( \equiv \) HTML+MathML+SVG
Substitution Tree Indexing in MathWebSearch

- Represent Mathematical Formulae in Content MathML extended with query variables
- Insert them into an in-memory “index”: a formula structure tree that shares common substructures
- unification by “dropping queries through tree”
- leaves correspond to unifiable formulae
- leaves are mapped to result occurrence URIs $u_i$ (in database)
Index statistics

▶ **Experiment**: Indexing the arXiv (1M documents, $\sim 10^8$ non-trivial formulae)

▶ **Results**: indexing up to 15 M formulae on a standard laptop

Query Times

- query time is constant ($\sim 15$ ms)
- memory footprint seems linear ($\sim 500 \frac{B}{\text{formula}}$) (as expected; goes by depth $\times$ symbols) (expected more duplicates)

▶ So we need ca. 100 GB RAM for indexing the whole arXiv.

▶ Can index all published Math ($\hat{=} 5 \times$ arXiv) on a large server (.5 TB RAM). (ZBL $\hat{=} 3.5$M art.)
MathWebSearch System Architecture

- crawlers for MathML, *OpenMath*, and OAI repositories. (convert your’s?)
- multiple search servers based substitution tree indexing (formula search)
- a RESTful server that acts as a front-end for multiple search servers.
- various front ends tailored to specific applications (search appliances)
  - a Google-like web front end for human users (search.mathweb.org)
  - a TeX-based front-end for the arXiv (http://arxivdemo.mathweb.org)
  - special integrations for theorem prover libraries (MizarWiki, TPTP)
A Front-End for Zentralblatt Math

 zbMATH
the first resource for mathematics

fractional derivative

\int_{a}^{b} |f(x)g(x)| \, dx \leq r

\int_{a}^{b} |f(x)g(x)| \, dx \leq r
A Front-End for Zentralblatt Math

\[ \int_{a}^{b} |f(x)g(x)| \, dx \leq r \]


http://zbmath.org/?q=an:1195.26008

Title: Caputo fractional multivariate Opial type inequalities on spherical shells.
Author(s): Anastassiou, George A.
Published: 2010
Class: 26A33 26D10 26D15
Doctype: serial article
Keywords: Opial inequality; fractional inequality; fractional derivative; radial derivative
Language: EN

\[ \int_{a}^{b} |f(x)g(x)| \, dx \leq r \]

The classical Opial inequality was proven in 1960 and establishes that if \( a \) is a positive number and \( y:[0, a] \rightarrow \mathbb{R} \) is continuously differentiable and \( y(0) = y(a) = 0 \), then

\[ \int_{0}^{a} |y(x)| \, dx \leq \frac{a}{4} \int_{0}^{a} (y'(x))^2 \, dx. \]

Moreover, equality holds if and only if \( y(x) = x \) on \([0, a/2]\) and \( y(x) = a - x \) on \([a/2, a]\). Several multivariate Opial-type inequalities are established. The proofs strongly rely on the notion of Caputo fractional radial derivative defined on a spherical shell. An application to the uniqueness of the resolution of a class of radial differential equations on the shell is also provided.
Formula/Text Search Combination?

- **Observation**: MathWebSearch is similar to a one-word IR algorithm, except unification directly matches one search term against lots of search terms.

- **Idea**: combine unification indexing with the vector space model for a "bag-of-formulae" (instead of standard IR’s "bag-of-words") method . . .

- **at Indexing time**: when we index a math document $D$,
  - insert the formulae into the MathWebSearch index (remember dbid)
  - replace all formulae in $D$ with their dbid to get $D'$
  - index $D'$ in a bag-of-words index (e.g. Elastic Search or Terrier)

- **At query time**: (essentially query expansion)
  - query $Q$ consists of a set $Q_f$ of formulae and a set $Q_w$ of words.
  - run $Q_f$ through MathWebSearch to get set $I_f$ of matching dbids.
  - run $Q' = Q_w + I_f$ through nutch to get a set $R$ of document fragments URIs.

- we return $R$ together with the fragments of $D$ they point to.

- we can even inherit the ranking mechanisms from nutch. (see if they help)
TeMaSearch Realization

- interleave harvesting with MathWebSearch formula indexing (dbid replacement)
- use MathWebSearch as query expansion in ElasticSearch.
Scalability/Stability Issues in MWS 1.0

- Reduced Memory footprint of formula index to $\sim 35\% \,(16\text{GB in RAM for NTCIR})$
- Formula Index Persistence: write/read index to/from disk in 90s
  
  (cf. 20h index creation)
- profiling index creation
  
  (20-40% speedup now)
- Full release on GitHub
  
  (https://github.com/KWARC/mws)
- Watchdog processes for MathWebSearch web services
- Production System at http://zbmath.org
  
  (structured/faceted search)
- NTCIR demo at http://arxivsearch.mathweb.org
  
  (try it!)
Conclusion & Future Work

- **MathWebSearch at NTCIR-11**
  - full text search
  - Scalability/Stability work
  - much improved web front-end
  - MathWebSearch 1.0 did well at NTCIR-11

- **Android App for MathWebSearch on Google Play**

  (formula search as query expansion)
  (production ready)
  (cross-browser, result highlighting)
  (without any tuning – no time)
Conclusion & Future Work

- **MathWebSearch at NTCIR-11**
  - full text search
  - Scalability/Stability work
  - much improved web front-end
  - MathWebSearch 1.0 did well at NTCIR-11

- Android App for MathWebSearch on Google Play

- Future Directions
  - Classifying formula schemata for full faceted search
  - number ranges & search modulo unit conversion
  - semantics extraction for more semantic search

(formula search as query expansion)  (production ready)  (cross-browser, result highlighting)  (without any tuning – no time)  (for physics)  (see Pre-NTCIR talk)
Submitted one run

- Results for 49/50 queries: avg. 112.5 hits/query.
- High precision results: matching formula and text
- Low precision results: matching only text
- Had to fill up with randomly sampled items.

(no time for tuning/variants)

Results for 49/50 queries: avg. 112.5 hits/query.

- High precision results: matching formula and text (26 queries 32.15 hits/query)
- Low precision results: matching only text (23 queries)
Analysis MathWebSearch Results at NTCIR-11

▶ Submitted one run (no time for tuning/variants)
  ▶ Results for 49/50 queries: avg. 112.5 hits/query.
  ▶ high precision results: matching formula and text (26 queries 32.15 hits/query)
  ▶ low precision results: matching only text (23 queries)
  ▶ had to fill up with randomly sampled items.

▶ Result evaluation
  ▶ 50% of top5 hits judged relevant, 79% partially relevant.
  ▶ ergo: MathWebSearch is precise for first-page results (top five hits)
  ▶ excellent precision for queries with ≥ 3 query variables (constraining query)
  ▶ MathWebSearch is better at ranking relevant (MAP: 29%) than partially relevant results (MAP: 25%)
  ▶ General Observation: MAP for relevant hits better with formula match
  ▶ Intuition: high precision via exact formula search + recall via keyword search.

Mathematical Markup Language (MathML) version 3.0.

Arif Jinha.
Article 50 million: an estimate of the number of scholarly articles in existence.

Peder Olesen Larsen and Markus von Ins.
The rate of growth in scientific publication and the decline in coverage provided by science citation index.