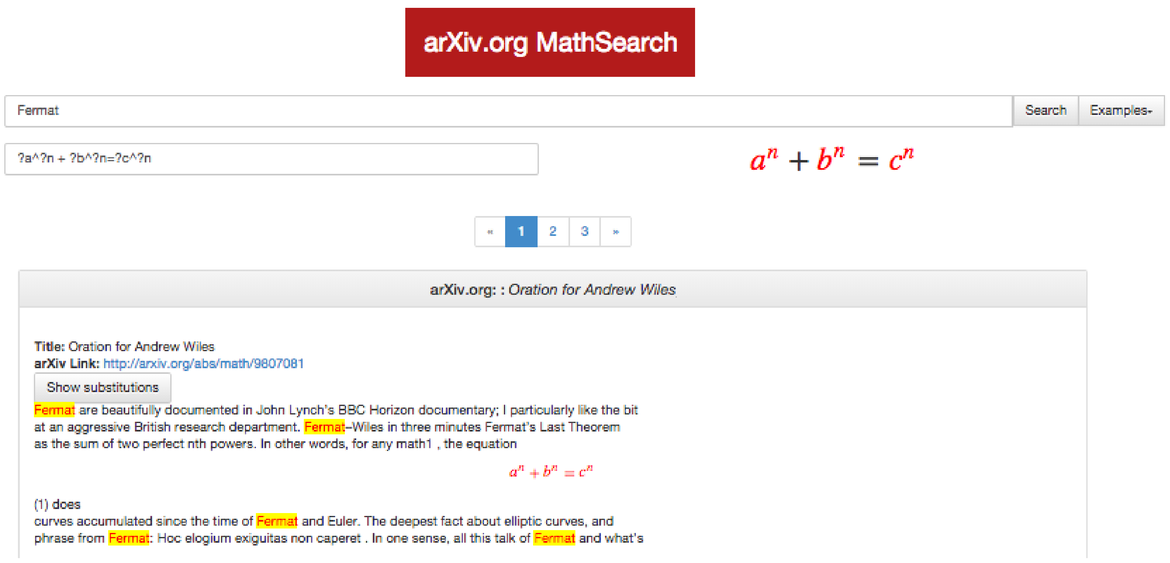
MathWebSearch: Low-Latency Unification-based Full-Text Search

Radu Hambasan, Michael Kohlhase, Corneliu Prodescu
http://kwarc.info

JACOBS UNIVERSITY

MathWebSearch is a content-based search engine that focuses on fast query answering for interactive applications. It is currently restricted to exact formula search via unification queries combined with keyword search.
NTCIR-11 System: http://arxivsearch.mathweb.org/ Zentralblatt Math: https://zbmath.org/formulae/ Excel Search: http://search.mathweb.org/xl/

Formula Search with Named Wildcards/Keywords



Unification Queries: Applicable Theorem Search

Approximate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ from above? $\rightsquigarrow$ Ask MathWebSearch!
It finds Hölder's inequality with universal variables in the index

$$\int_D |f(x)g(x)|\,dx \leq \left(\int_D |f(x)|^p\,dx\right)^{\frac{1}{p}} \left(\int_D |g(x)|^q\,dx\right)^{\frac{1}{q}}$$

with substitution $x \mapsto t, f \mapsto sin, g \mapsto cos, D \mapsto \mathbb{R}^2 \rightsquigarrow$ Solution:

$$\int_{\mathbb{R}^2} |\sin(t)\cos(t)|\,dt \leq \left(\int_{\mathbb{R}^2} |\sin(t)|^p\,dt\right)^{\frac{1}{p}} \left(\int_{\mathbb{R}^2} |\cos(t)|^q\,dt\right)^{\frac{1}{q}}$$

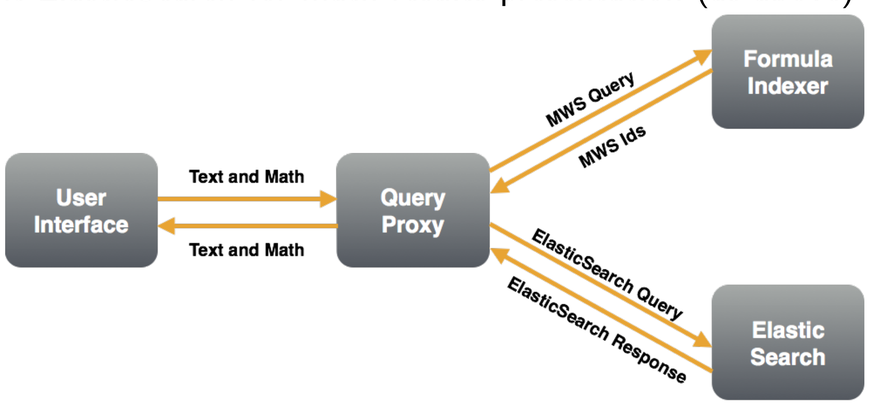
Variant query $\int_{\mathbb{R}^2} |\sin(t)\cos(2t)|dt$ will not find Hölder's inequality since that would introduce inconsistent substitutions $x \mapsto t$ and $x \mapsto 2t$.

The MathWebSearch backend is realized as a RESTful web service that keeps a formula index in memory and hit URIs in database. MathWebSearch front-ends post MathML queries via HTTP and receive XML results.

System Architecture: Formula Search as Query Expansion

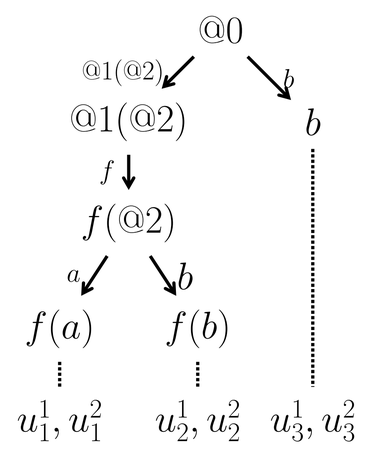
Idea: Formula IDs as "words" in ElasticSearch, MathWebSearch for Query Expansion

i) Replace text formulae by their index id $\rightsquigarrow$ index in ElasticSearch
ii) Unify query formulae via MathWebSearch $\rightsquigarrow$ replace by ids in query
iii) Augment ElasticSearch for math results presentation (as above)



MathWebSearch aims at high-quality hits only (randomly extend to 1000)
- **high precision** (matching formulae + text) $\rightsquigarrow$ 26/50 hits only (32.1/query)
- **low precision** (matching only text) $\rightsquigarrow$ 23/50 hits
- **no hits** $\rightsquigarrow$ 1/50, common keyword (two spellings) no formula matches.

Substitution Tree Indexing



- 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^j (in database)

Results Evaluation: NTCIR-11 dataset (~ 8.3 million paragraphs from 105,539 XHTML+MathML documents)
$\rightsquigarrow$ 224 GB harvest data, 584 M SubFormulae (63M unique) [16h harvesting]
$\rightsquigarrow$ 15.9GB Formula Index (in RAM) + 63 GBs ElasticSearch Index (on disk) [45h indexing, 90s restore from disk]
$\rightsquigarrow$ query answer times $3 - 70ms$ (avg $= 11ms$) for MathWebSearch, longer for ElasticSearch, even longer for result presentation

MathWebSearch aims at high-quality hits only (randomly extend to 1000)
- **high precision** (matching formulae + text) $\rightsquigarrow$ 26/50 hits only (32.1/query)
- **low precision** (matching only text) $\rightsquigarrow$ 23/50 hits
- **no hits** $\rightsquigarrow$ 1/50, common keyword (two spellings) no formula matches.

a) 50% of top5 hits judged "relevant", 79% "partially relevant"
b) excellent precision for formula queries with ≥ 2 query variables
c) side-effect of MathWebSearch query expansion: keywords used for ranking formula hits.

Current Work: Faceted Search, Ranking, Extensions, Embedding, Unit Search

- Generating formula schemata for faceted search — Anti-Unification
- Special treatment of literal data types (e.g. numbers) — search for ranges
- Physics: Search for quantities modulo unit conversion — via flatsearch
- Semantic search via query expansion — $e = \sum_{n=0}^{\infty} \frac{1}{n!}$ = Euler's Number

Searching the Mathematical Knowledge Space