MathWebSearch: Low-Latency Unification-based Search

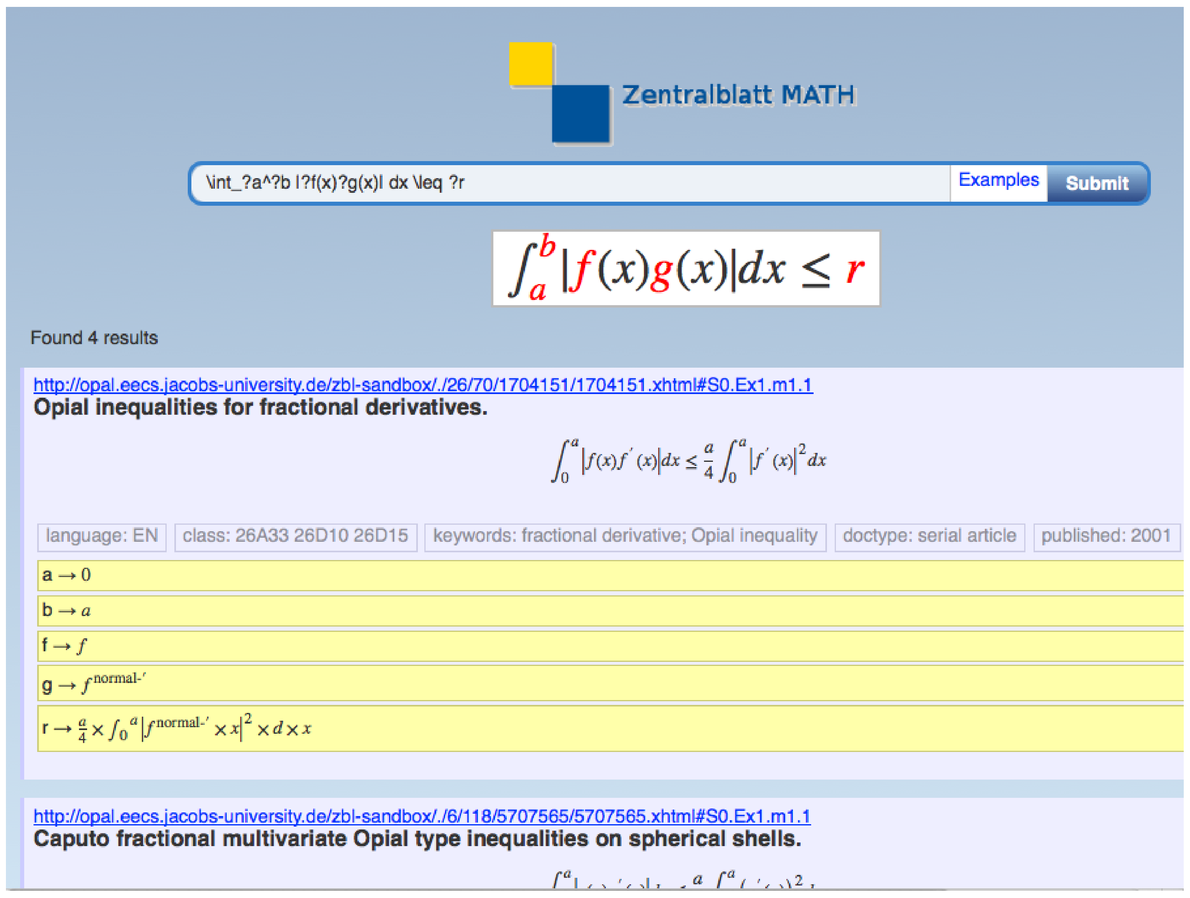
Michael Kohlhase <m.kohlhase@jacobs-university.de>

Corneliu Prodescu <c.prodescu@jacobs-university.de>

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, i.e. no similarity search and no full-text search.

Formula Search with Named Wildcards on ZBMath



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

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

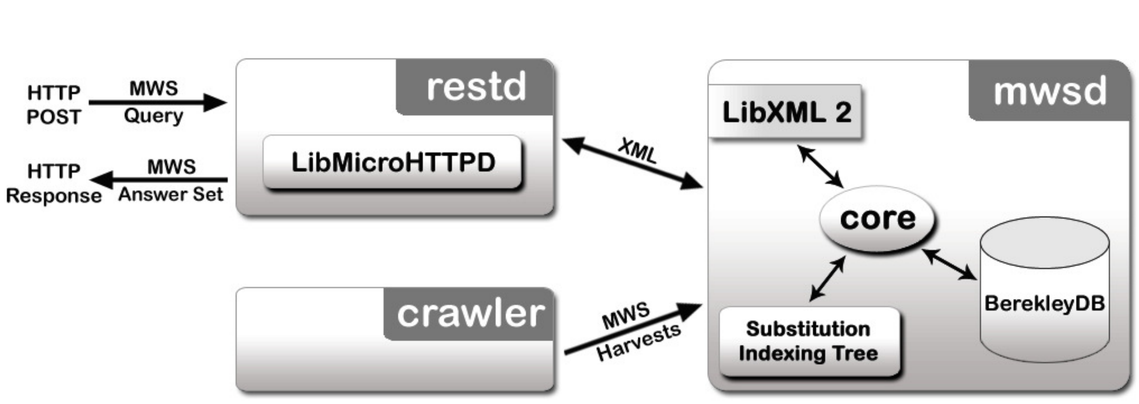
with substitution $x \mapsto t, f \mapsto sin, g \mapsto cos, D \mapsto \mathbb{R}^2 \rightsquigarrow$ Solution:

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

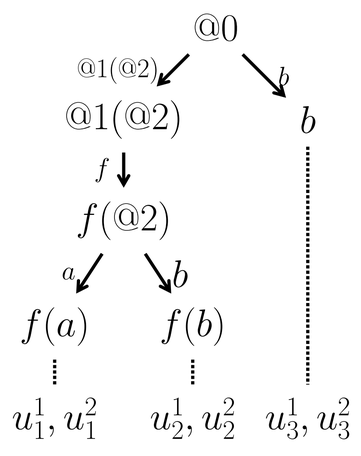
Variant query $\int_{\mathbb{R}^2} |\sin(t)\cos(2t)|dt$ will not find Hölder's inequality since that would introduce inconsistent substitutions $x \mapsto t$ and $x \mapsto 2t$.

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: Web Service with multiple Front-Ends



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 dataset (100,000 XHTML+MathML documents: 63GBs, 297 MFormulae)
$\rightsquigarrow$ 10GBs RAM + 43 GBs URIs (on disk) $\rightsquigarrow$ query answer times $3-70ms$ (avg $= 11ms$)

MathWebSearch aims at high-quality hits only, reported 434 hits:

1	2	3	4	5	6	7	8	9	10	11	2	13*	14	15	16	17*	18	19	20*	21	22
0	9	21	4	49	0	51	100+	100+	0	0	0	0	0	0	0	0	100+	0	0	0	0

a) lots of hits $(100+)$: general queries with multiple query variables
b) few hits: specific queries with precise expressions.
c) no hits: errors* in query or missing exact matches

Current Work: Full-Text Search, Ranking, Extensions, Embedding

Full Text Search (TeMaSearch)

Idea: Formulae as words in Lucene, MathWebSearch for Query Expansion

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

Searching other Corpora

e.g. Spreadsheet formulae $\rightsquigarrow$ just encode in content MathML!

Searching the Mathematical Knowledge Space