



ADVANCED MATH SEARCH: ISSUES & TECHNIQUES

Abdou Youssef

The George Washington University

And

The National Institute of Standards and
Technology
(DLMF)



Outline

- Context of our Math Search Project
- The Project's Short-Term Goals
- Where we are: A Demo
- Issues faced
- Goals and Issues for the Longer Term



Context of our Math Search Project

- The Digital Library of Mathematical Functions (DLMF) at NIST
 - Web+Book Replacement of the Abramowitz and Stegun Handbook
 - Special functions, Analysis, Functions of Number Theory, Combinatorial Analysis, Numerical Methods, Statistical Methods, ...
 - DLMF: Mostly Equations – Need Math Search
- Support from NIST and NSF



Short-Term Goals

- Build a math search system that
 1. Understands math symbols & structures
 2. Returns equations directly, not just hit-titles
 3. Highlights matched equations in documents
 4. Understands dialects (Latex, Mathematica, Maple)
 5. Provides different search modes (TOC, Index, Free-style search, and Menu-driven search)

Where we Are



Demo of the Search System
(Not available online)

Sample Queries: understanding math, eq. search & highlighting

Form	Entry
$\int_0^{\infty} \sin\left(\frac{1}{3}t^3 + xt\right)$	<ul style="list-style-type: none"> ■int_0^infinity sin((1/3)t^3+xt) ■int sin((1/3)t^3+xt)
$\sqrt{A_i^2 + B_i^2}$	sqrt(Ai^2+Bi^2)
$\Gamma(\lambda-\$+\$)$	Gamma(lambda-\\$+\\$)
J_ν or J_0	J_nu or J_0
A_i and J	<ul style="list-style-type: none"> ■Ai and J ■Ai and BesselJ



Sample Queries: different dialects

- BesselJ(nu,z)
- BesselJ(nu,)
- BesselJ(,zeta)
- JacobiP(n,alpha,beta,x)
- JacobiP(, alpha, ,)
- LaguerreL(, , x)
- LegendreP[, mu ,]
- LegendreP[, ,sqrt(z)]



Search Modes

- One extreme: (static and limited)
 - Table of contents (thematic, coarse-grained)
 - Index (alphabetical, fine-grained)
- Opposite extreme: (dynamic and unlimited)
 - Free-style search
- Another mode: (a middle ground)
 - Menu-driven search
 - based on an ontology
 - constrained/standard vocabulary
- Hybrids of the above



Issues Faced

- Recognizing and Indexing Math Symbols and Structures Pre- MathML & XPath/XQuery
- Highlighting Matched Equations (GIF Images) inside HTML Documents
- Development of a Query Language that is Intuitive, Natural, Rich, and Consistent
- Obtaining/Deriving Metadata for Equations
- Development of a Math Taxonomy/Ontology Suitable for Menu-Driven Search



Techniques: for Handling Math Symbols and Structures

- For Recognizing and Indexing Math Symbols and Structures, *TexSNize*:
 1. Textualization of math symbols
 2. Scoping of the various parts of terms/exprs
 3. Normalization of the orders of parts
- TexSNize the Contents Offline before Indexing
- TexSNize each Query before the search



Techniques: for Equation Search and Highlighting

- Create a data model that logically decouples equations from their native documents.
- Assign a unique ID to each equation
- For Returning Equations directly:
 - algorithm that uses a hit list of equation IDs to generate online a document containing the equations
- For Highlighting Equations:
 - Use the IDs of matched equations to locate the latter in a to-be-displayed document
 - add coloring HTML markup to doc before display



Goals for the Longer Term

- We Embarked on Building a 2nd Generation Math Search System
 - Based on Content MathML+XPath/XQuery
 - More Precise/Expressive Query Language
 - Higher resolution search
 - Term-occurrence search
 - Predicate search
 - Search with term substitution
 - Similarity Search (for Sci. Data Mining)



Examples of Future Query Types

- Queries specifying subparts
 - $\sin x$ in a denominator
 - $x-y$ in a 3rd row of a matrix
- Predicate queries
 - z^k , where k is an integer that ranges from -4 to 4
- Term-substitution queries
 - $g(?) = z^2 + z + 1$, where $z = e^i$?
- Abstraction support and similarity search
 - $x^2 + y^2 = 1$ whatever x and y



Candidate Syntax

- $f(\dots \sin x \dots)$
- $@(\dots 2px \dots)$
- $z^{\$k}$ where $\text{integer}(\$k) \ \& \ \text{abs}(\$k) < 5$
- $x-y$ in $\text{matrix}[2,3]$
- x^2 in $\text{matrix} [\$k, \$j]$ where $\text{abs}(\$k - \$j) < 2$
- $\$A$ where $\text{matrix}(\$A) \ \&$
 $(\text{forall } \$k) (\text{forall } \$j < \$k): \$A[\$k, \$j] > 0$
- $\$S$ where $\text{set}(\$S) \ \& \ (? \ \$x \text{ in } \$S): \text{integer}(\$x) \ \& \ \$x > 0$
- $x < 1$ in $\text{condition}(\text{set})$



Issues that Need to Be Faced under C-MathML & XPath/XQuery

- Canonical Normal Forms of Contents
 - Math equivalences: ab/c : $(a*b)/c$ or $a*(b/c)$
 - Notational equivalences: \int_a^b or $\int_{[a, b]}$
 - Distributed definitions
- Uniform Symbolic Notation
- Standard Ontologies
- Development of Metadata
 - Automated extrapolation of metadata
 - Manual (by authors and communities)

Issues to Be Faced (Contd.)



- What Users Need/Want/Prefer
 - What modes of search?
 - What kinds of information?
 - definitions, equations, theorems, proofs, proof techniques, step-by-step evaluations, themes, theories, expositions, etc.?
 - What granularity of retrieval unit?
 - What interactive features?
 - Definition of terms, plotting of matched functions, computing of matched functions?
- Use of Knowledge of Users' Needs/Preferences
 - Better design of the search UI
 - Better relevance-ranking of search results



We Are Barely Scratching the Surface

- The Possibilities Are Endless
 - Search + Automated Reasoning
 - Search + Computing + Visualization