SEARCH IN MATHEMATICAL DATABASES

Abdou Youssef

The George Washington University

And

The National Institute of Standards and Technology

(DLMF)

Outline

- Context of the DLMF Math Search Project
- The Project’s Short-Term Goals
- Where we are: A Demo
- Technical Issues and Techniques
- Goals and Issues for the Longer Term

Context of the 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

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

Sample Queries: understanding math, eq. search & highlighting

Form	Entry
$\int_0^\infty \sin(\frac{1}{3}t^3 + xt)$	n int_0^infinity sin((1/3)t^3+xt) n int sin((1/3)t^3+xt)
$\sqrt{\mathrm{Ai}^2 + \mathrm{Bi}^2}$	sqrt(Ai^2+Bi^2)
$\Gamma(\lambda\text{-}\$+\$)$	Gamma(lambda-$+$)
J_ν or J_0	J_nu or J_0
Ai and J	n Ai and J n 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
- 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, *TexSN*ize:
  1. Textualization of math symbols (illust.)
  2. Scoping of the various parts of terms/exprs
  3. Normalization of the orders of parts (illust.)
- *TexSN*ize the Contents Offline before Indexing
- *TexSN*ize 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

Architecture of the System

- Surrogate Files
- Indexing System Architecture
- Search System Architecture

Goals for the Longer Term

- Development of a 2nd Generation Math Search System
  - Based on Content MathML+XPath/XQuery
  - More Precise/Expressive Query Language
    - Higher resolution search
    - Keyword 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
  - $2\pi x$ inside an argument of a function
- Predicate queries
  - z^k, where k is an integer that ranges from –4 to 4
- Term-substitution queries
  - $g(\omega)=z^2 + z +1$, where $z = e^{i\omega}$
- Abstraction support and similarity search
  - x^2 + y^2 =1 whatever x and y

Candidate Syntax (Based on 1st Order Logic)

- /(… sin x …)
- @(… 2πx …)
- z^$k where integer($k) & abs($k)<5
- x-y in matrix[2,3]
- x^2 in matrix [$k,$j] where abs($k-$j) < 2
- $A where matrix($A) &
  (forall $k) (forall $j < $k): $A[$k,$j]>0
- $S where set($S) & (∃ $x in $S): integer($x) & $x>0
- x<1 in condition(set)

Issues that Need to Be Faced

- 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/computing of matched functions?
- Use of Knowledge of Users’ Needs/Preferences
  - More relevant search features & capabilities
  - Better design of the search user interface
  - Better relevance-ranking of search results

We Are Barely Scratching the Surface

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

The end