We (partially) have a WDML
What we really need is Semantics!
Then we can do Search and Much More®

Michael Kohlhase

http://kwarc.info/kohlhase
Center for Advanced Systems Engineering
Jacobs University Bremen, Germany

June 1. 2012, WDML 2012

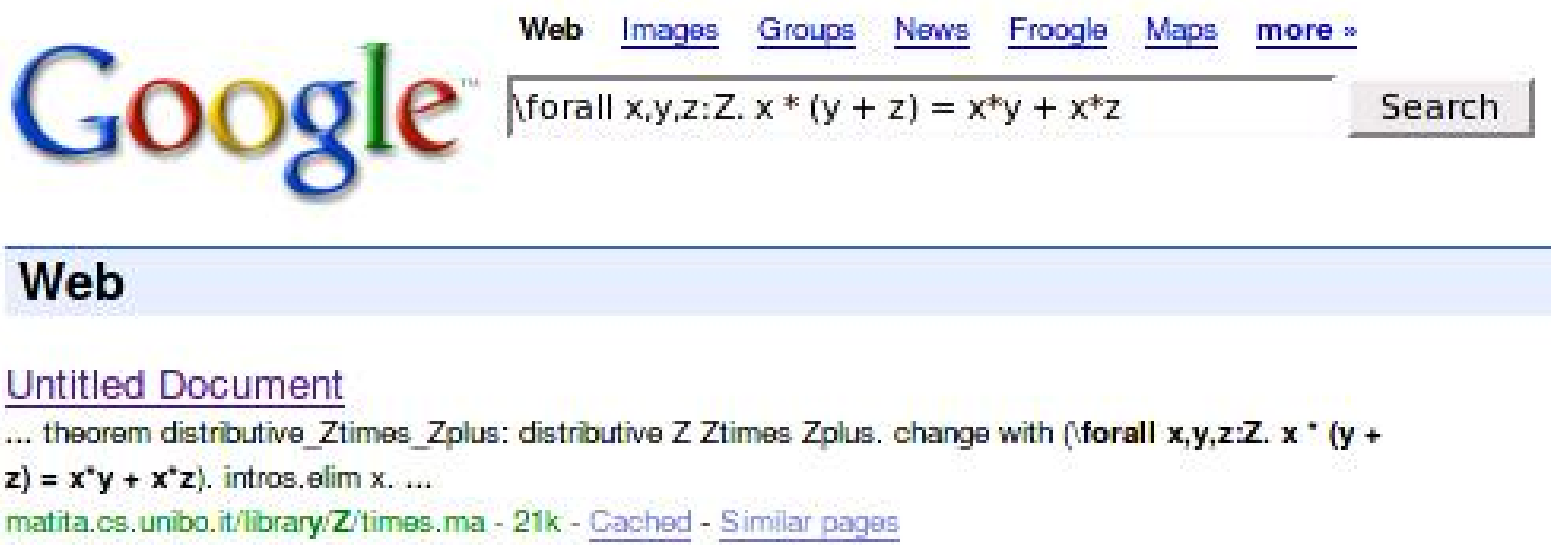
More Mathematics on the Web

- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- ...
- Question: How will we find content that is relevant to our needs
- Idea: try Google (like we always do)
- Scenario: Try finding the distributivity property for $\mathbb{Z}$

$$(\forall k, l, m \in \mathbb{Z}.k \cdot (l + m) = (k \cdot l) + (k + m))$$

Searching for Distributivity

Searching for Distributivity



Searching for Distributivity

Of course Google cannot work out of the box

- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables.
    (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, MG11])
  . . . and works surprisingly well (using Lucene as an indexing engine)
- Idea: Use database techniques (extract metadata and index it)
  Indeed this is done for the `Coq`/`HELM` corpus ([AGC+06])
- Our Idea: Use Automated Reasoning Techniques
  (free term indexing from theorem prover jails)

Instantiation Queries

- Application: Find partially remembered formulae
- **Example 1** An engineer might face the problem remembering the energy of a given signal $f(x)$
  - Problem: hmmmm, have to square it and integrate
  - Query Term: $\int_{\boxed{min}}^{\boxed{max}} \boxed{f}(x)^2 dx$ ($\boxed{i}$ are search variables)
  - One Hit: Parseval's Theorem $\frac{1}{T}\int^{T_0} s^2(t)dt = \sum_{k=-\infty}^{\infty} \|c_k\|^2$ (nice, I can compute it)
- This works out of the box (has ween working in `MathWebSearch` for some time)
- Another Application: Underspecified Conjectures/Theorem Proving
  - during theory exploration we often have some freedom
  - express that using metavariables in conjectures
  - instantiate the conjecture metavariables as the proof as the proof dictates

  applied e.g. in Alan Bundy's "middle-out reasoning" in proof planing

Generalization Queries

- Application: Find (possibly) applicable theorems
- **Example 2** A researcher wants to estimate $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ from above
  - Problem: Find inequation such that $\int_{\mathbb{R}^2} |\sin(t)\cos(t)|dt$ matches left hand side.
  - e.g. Hölder's Inequality: ($\boxed{i}$ are universal variables)

$$\int_{\boxed{D}} \left|\boxed{f}(x)\boxed{g}(x)\right| dx \leq \left(\int_{\boxed{D}} \left|\boxed{f}(x)\right|^p dx\right)^{\frac{1}{p}} \left(\int_{\boxed{D}} \left|\boxed{g}(x)\right|^q dx\right)^{\frac{1}{q}}$$

  - Solution: Take the instance

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

Problem: Where do the index formulae come from in particular the universal variables (we'll come back to that later)

Where do the universal variables come from

- Problem: we need to have e.g. Hölder's Inequality in the index:

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

  - How do we know what symbols are "universal" (to be instantiated?)
  - what is their scope (when are different occurrences of f different?)
  - we have no sources with explicit quantifiers, but ([Wikipedia])

    Let (D, Σ, μ) be a measure space and let $1 \leq p, q \leq \infty$ with $1/p + 1/q = 1$. Then, for all measurable real- or complex-valued functions f and g on D, ...

- Solution: Use techniques from computational linguistics and integrate them into the indexing pipeline. (we have started a bit on the arXiv)

What is the Point (I want to make)

- Search/Navigation is the prerequisite and killer application for a WDML
- Search/Navigation/Access has to be semantic (that's how humans think)
- Generally:The WDML needs a semantic Layer (Then we can access it)

  - The Elephant in the Room: Semantization
  - Possible Answers:
    - for new documents: write with semantic annotations (e.g. in sTEX)
    - for legacy documents: extract semantics with linguistic methods (or retype)
  - If you are interested, contact me

- BTW: NTCIR-10 (Information Retrieval Challenge like TREC) has a Math Pilot Task in 2013! (see http://ntcir-math.nii.ac.jp/)


Andrea Asperti, Ferruccio Guidi, Claudio Sacerdoti Coen, Enrico Tassi, and Stefano Zacchiroli.
A content based mathematical search engine: Whelp.
In Jean-Christophe Filliâtre, Christine Paulin-Mohring, and Benjamin Werner, editors, *Types for Proofs and Programs, International Workshop, TYPES 2004, revised selected papers*, number 3839 in Lecture Notes in Computer Science, pages 17–32. Springer Verlag, 2006.


Paul Libbrecht and Erica Melis.
Methods for Access and Retrieval of Mathematical Content in ActiveMath.
In N. Takayama and A. Iglesias, editors, *Proceedings of ICMS-2006*, number 4151 in LNAI, pages 331–342. Springer Verlag, 2006.
`http://www.activemath.org/publications/Libbrecht-Melis-Access-and-Retrieval-ActiveMath-ICMS-2006.pdf`.


Jozef Misutka and Leo Galambos.
System description: Egomath2 as a tool for mathematical searching on wikipedia.org.
In James Davenport, William Farmer, Florian Rabe, and Josef Urban, editors, *Calculemus/MKM*, number 6824 in LNAI, pages 307–309. Springer Verlag, 2011.


Rajesh Munavalli and Robert Miner.

Mathfind: a math-aware search engine.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–735, New York, NY, USA, 2006. ACM Press.

Bruce R. Miller and Abdou Youssef.
Technical aspects of the digital library of mathematical functions.
Annals of Mathematics and Artificial Intelligence, 38(1-3):121–136, 2003.