Towards Structure-Aware Information Retrieval

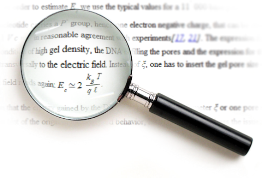
Petr Sojka et al

Masaryk University, Faculty of Informatics, Brno, Czech Republic
<https://mir.fi.muni.cz/>

Informatics Colloquium, Faculty of Informatics, MU, Brno, Czech Republic
October 25th, 2014

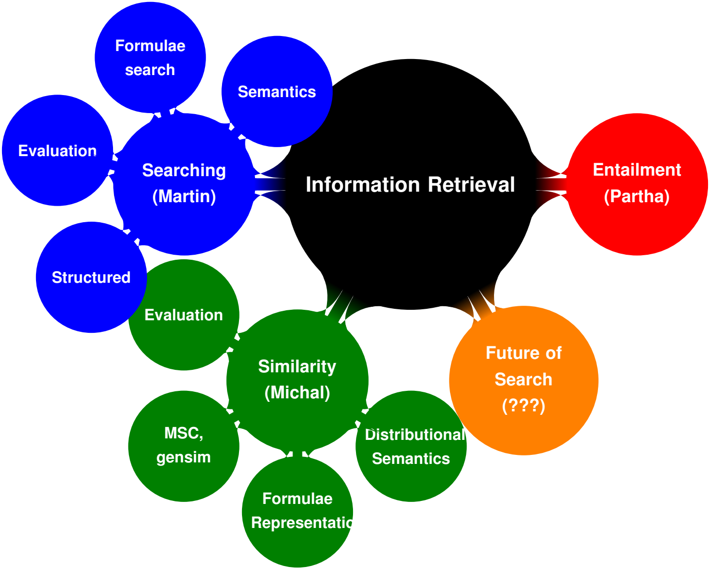
EuDML
The EUROPEAN DIGITAL MATHEMATICS LIBRARY

FACULTAS ARTIS INFORMATICAE • UNIVERSITAS MASARYKIANA •



Illustrations by Jiří Franek.

Talk topics and take-home message

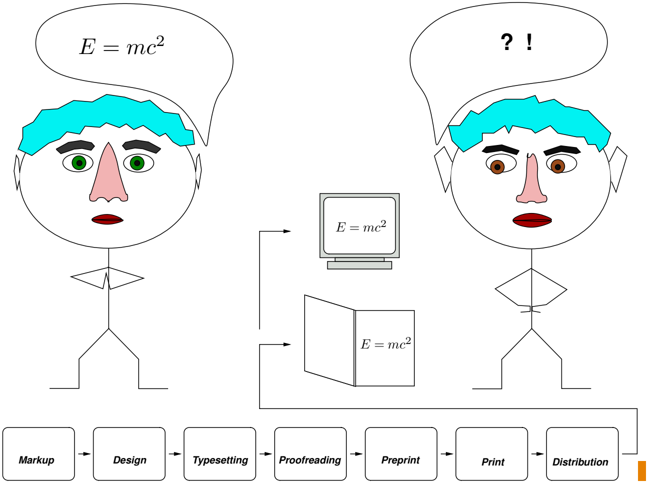


Outline

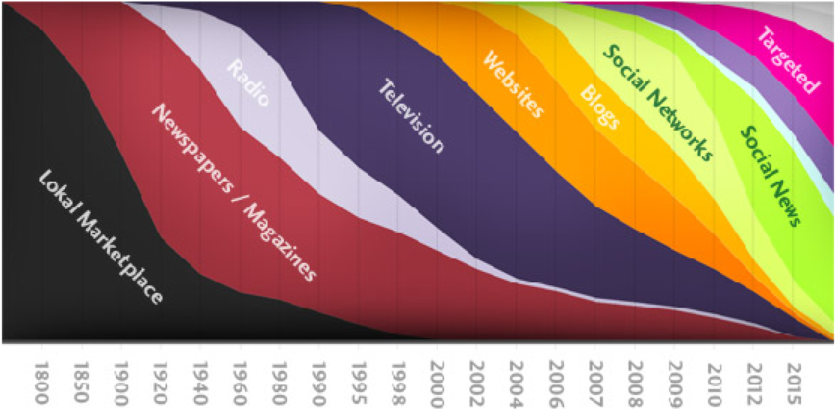
1. Motivation
2. Searching: MIaS
3. MIaS at NTCIR
4. Similarity
5. Entailment
6. Summary and future work

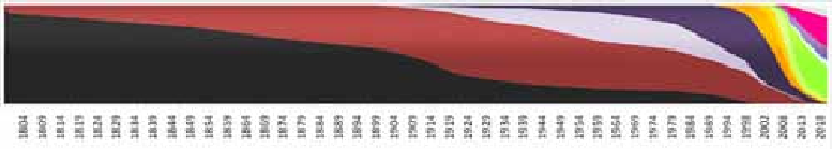
Dependency on Information Retrieval: Information Society Now!

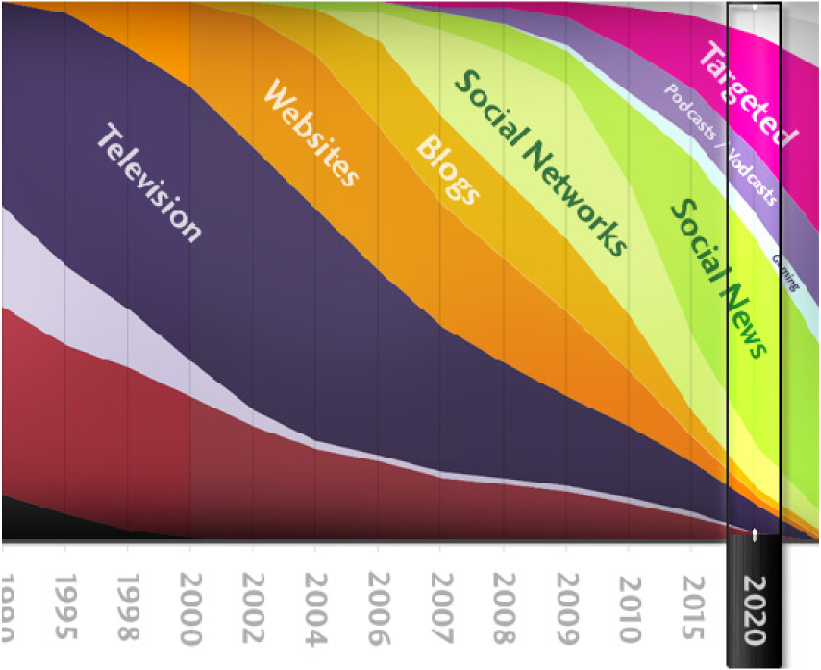
Scholarly STEM Communication *via Digital Math Libraries*



History of *information retrieval*: gradual speedup of changes







Television
Websites
Blogs
Social Networks
Social News
Podcasts / Vodcasts
Gaming
Targeted
1990
1995
1998
2000
2002
2004
2006
2007
2008
2009
2010
2015
2020

Search: A gate to knowledge

Querying and *search*ing *similar* structures more and more important.

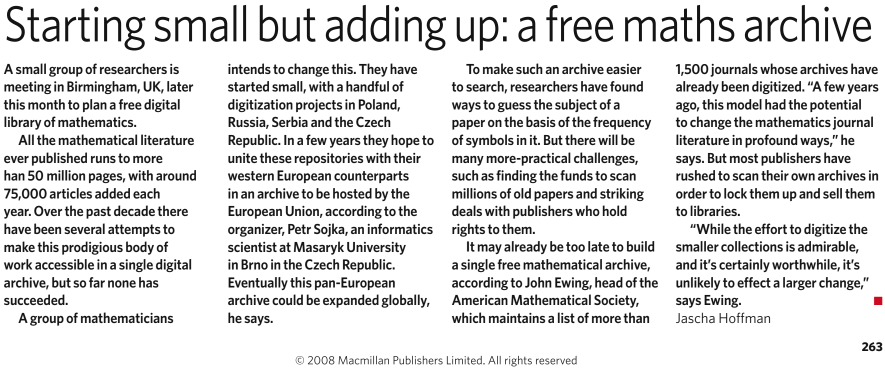
Structures: math formulae, syntactic or sentence dependency trees, compositional named entity terms, knowledge base terms.

<http://google.cz/search?q=Kovacik+Rakosnik>

```
$L^{p(x)}$
https://www.google.cz/search?q="L^{p(x)}"
```

+ without quotes or figures :-).

Nature 454, 263 (2008) | doi:10.1038/454263b



Starting small but adding up: a free maths archive

A small group of researchers is meeting in Birmingham, UK, later this month to plan a free digital library of mathematics.

All the mathematical literature ever published runs to more han 50 million pages, with around 75,000 articles added each year. Over the past decade there have been several attempts to make this prodigious body of work accessible in a single digital archive, but so far none has succeeded.

A group of mathematicians intends to change this. They have started small, with a handful of digitization projects in Poland, Russia, Serbia and the Czech Republic. In a few years they hope to unite these repositories with their western European counterparts in an archive to be hosted by the European Union, according to the organizer, Petr Sojka, an informatics scientist at Masaryk University in Brno in the Czech Republic. Eventually this pan-European archive could be expanded globally, he says.

To make such an archive easier to search, researchers have found ways to guess the subject of a paper on the basis of the frequency of symbols in it. But there will be many more-practical challenges, such as finding the funds to scan millions of old papers and striking deals with publishers who hold rights to them.

It may already be too late to build a single free mathematical archive, according to John Ewing, head of the American Mathematical Society, which maintains a list of more than 1,500 journals whose archives have already been digitized. "A few years ago, this model had the potential to change the mathematics journal literature in profound ways," he says. But most publishers have rushed to scan their own archives in order to lock them up and sell them to libraries.

"While the effort to digitize the smaller collections is admirable, and it's certainly worthwhile, it's unlikely to effect a larger change," says Ewing. ■

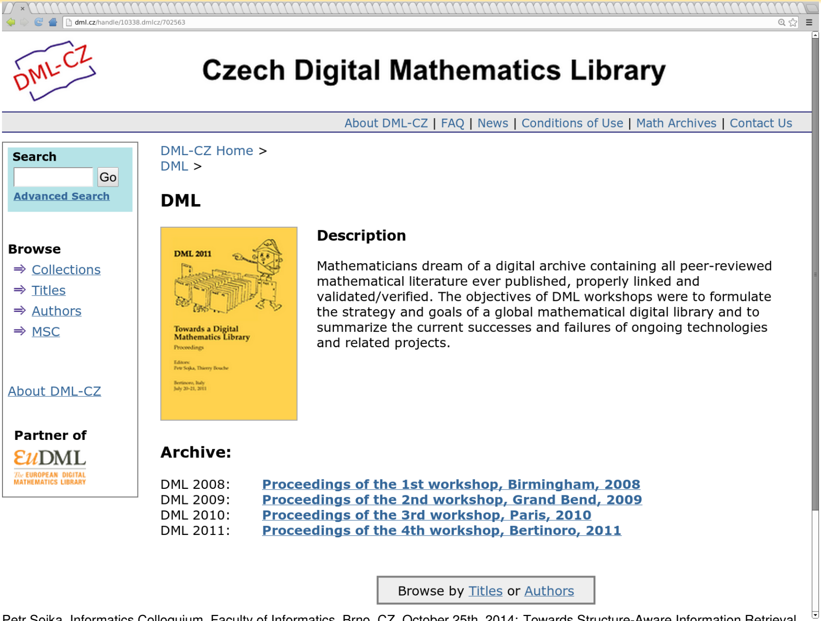
Jascha Hoffman

263

© 2008 Macmillan Publishers Limited. All rights reserved

Workshop series *Towards a Digital Mathematics Library* founded to tackle numerous challenges identified during DML-CZ project.

DML workshop series archived in DML-CZ


dml.cz/handle/10338.dmlcz/702563
DML-CZ

Czech Digital Mathematics Library

About DML-CZ | FAQ | News | Conditions of Use | Math Archives | Contact Us

Search

Go

Advanced Search

Browse

- ⇒ Collections
- ⇒ Titles
- ⇒ Authors
- ⇒ MSC

About DML-CZ

Partner of

EuDML The European Digital Mathematics Library

DML-CZ Home >
DML >

DML

DML 2011
Towards a Digital Mathematics Library
Proceedings
Editors:
Petr Sojka, Thierry Bouche
Bertinoro, Italy
July 20–21, 2011

Description

Mathematicians dream of a digital archive containing all peer-reviewed mathematical literature ever published, properly linked and validated/verified. The objectives of DML workshops were to formulate the strategy and goals of a global mathematical digital library and to summarize the current successes and failures of ongoing technologies and related projects.

Archive:

DML 2008:	Proceedings of the 1st workshop, Birmingham, 2008
DML 2009:	Proceedings of the 2nd workshop, Grand Bend, 2009
DML 2010:	Proceedings of the 3rd workshop, Paris, 2010
DML 2011:	Proceedings of the 4th workshop, Bertinoro, 2011

Browse by Titles or Authors

Aggregation of data from building bricks of regional repositories: EuDML

14 data and technology providers plus associated partners as ZMath, Göttingen library,…

DML content providers serve mostly publisher's or regional more or less established DML repositories: The Czech Digital Mathematics Library DML-CZ, NUMDAM, DML-PL, DML-PT, DML-GR, DML-BG, DML-ES,…

Aggregation via standard OAI-PMH protocol (OAI servers run by data providers).

<http://eudml.org>

Math aware Search and Indexing

- Conventional searching approaches are not applicable for math
- Usage of existing mathematical search engines (MathDex, EgoMath, LaTeXSearch, LeActiveMath, MathWebSearch) problematic
- new Math Indexer and Searcher (MIaS) developed at MU

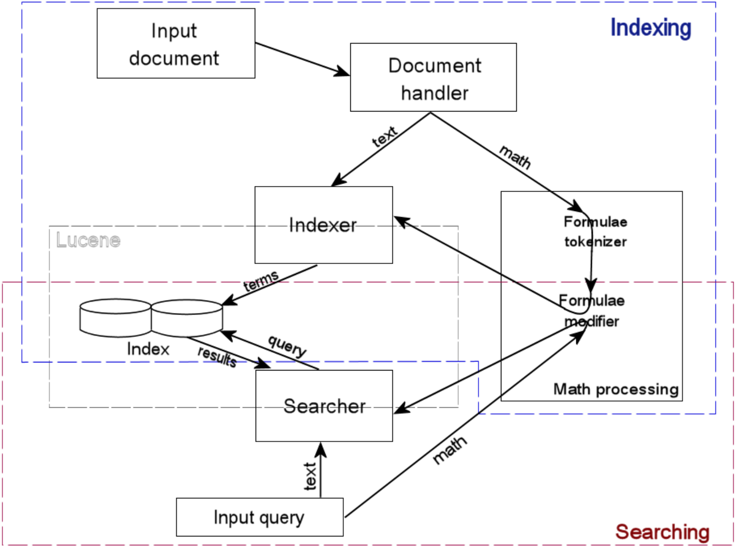
MIR systems comparison

	Input documents	Internal representation	Used converters	Approach	α-eq.	Query language	Queries	Indexing core
MathDex	HTML, TeX/LaTeX, Word, PDF	Presentation MathML (text)	jtidy, blahtex, LaTeXML, Hermes, Word+MathType, pdf2tiff->Infty	syntactic	×	?	text, math, mixed	Apache Lucene
LeActiveMath	OMDoc, OpenMath	OpenMath (text)	-	syntactic	×	OpenMath (palette editor)	text, math, mixed	Apache Lucene
LaTeXSearch	LaTeX	LaTeX(text)	-	syntactic	×	LaTeX	titles, math, DOI	?
MathWeb Search	Presentation MathML, Content MathML, OpenMath	Content MathML, OpenMath (substitution trees)	-	semantic	✔	QMath, LaTeX, Mathematica, Maxima, Maple, Yacas styles (palette editor)	text, math, mixed	Apache Lucene (for text only)
EgoMath	Presentation MathML, Content MathML, PDF	Presentation MathML (text)	Infty	mixed	×	LaTeX	text, math, mixed	EgoThor

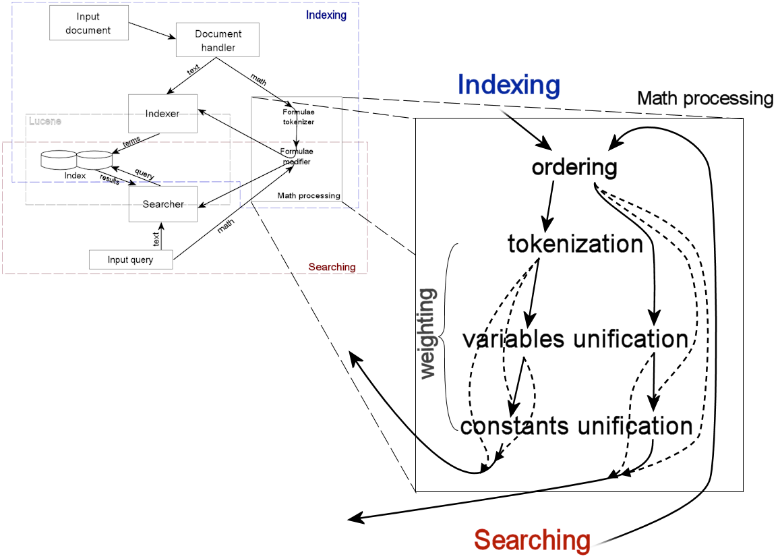
Math Indexer and Searcher MIaS — features

- Inspired mostly by MathDex and EgoMath
- Presentation and now also Content MathML
- Allows *similarity* (not only exact match) between query and matched term, *distributional representation* of formulae
  - Commutativity
  - Unification of variables and constants
  - Subformulae matching
- Level of similarity calculation for expressions
- Mixed mathematical-textual queries
- Based on full text state of the art Apache Lucene core

Math Indexer and Searcher — Design



Math Indexer and Searcher — Design II



Input document

Document handler

Indexing

text

math

Indexer

Lucene

Formulae tokenizer

terms

Formulae modifier

Index

query

results

Searcher

Math processing

text

math

Input query

Searching

Indexing

Math processing

ordering

tokenization

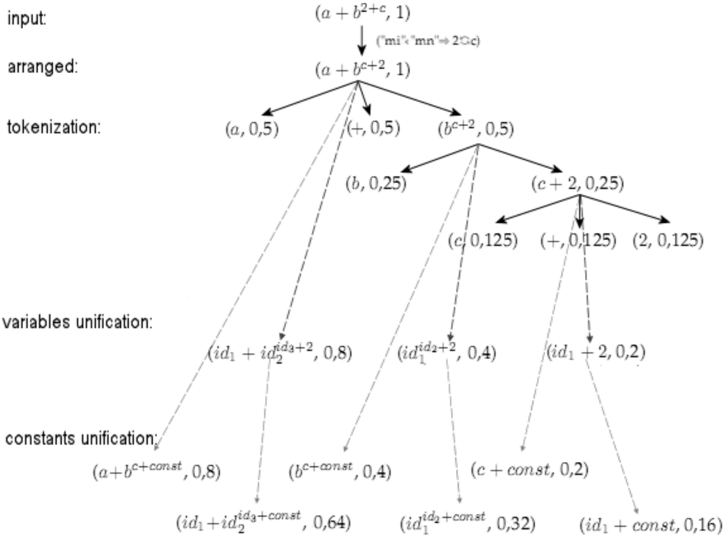
weighting

variables unification

constants unification

Searching

Formula processing weighting example



input: $(a+b^{2+c}, 1)$

("mi"<"mn"⇒ 2↔c)

arranged: $(a+b^{c+2}, 1)$

tokenization: $(a, 0{,}5)$ $(+, 0{,}5)$ $(b^{c+2}, 0{,}5)$

$(b, 0{,}25)$ $(c+2, 0{,}25)$

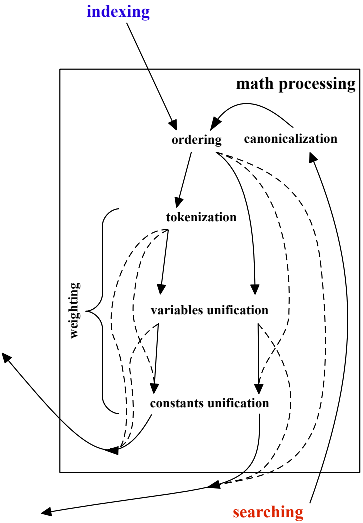
$(c, 0{,}125)$ $(+, 0{,}125)$ $(2, 0{,}125)$

variables unification: $(id_1+id_2^{id_3+2}, 0{,}8)$ $(id_1^{id_2+2}, 0{,}4)$ $(id_1+2, 0{,}2)$

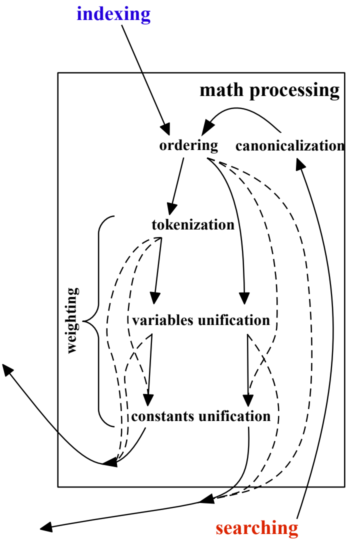
constants unification: $(a+b^{c+const}, 0{,}8)$ $(b^{c+const}, 0{,}4)$ $(c+const, 0{,}2)$

$(id_1+id_2^{id_3+const}, 0{,}64)$ $(id_1^{id_2+const}, 0{,}32)$ $(id_1+const, 0{,}16)$

Math formulae indexing processing



Example



indexing searching

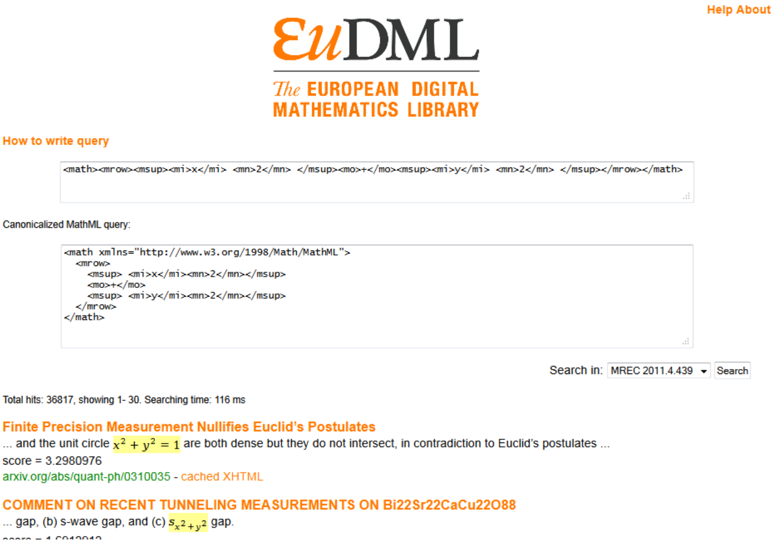
$$\begin{tikzcd}
x^y+y^3 \arrow[d] & & x^y+y^2 \arrow[d] \\
x^y+y^3 \arrow[d] & & x^y+y^2 \arrow[dd] \\
x^y+y^3, x^y, y^3, x, y, 3, + \arrow[d] & & \\
x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, id_1^{id_2}, id_1^3 \arrow[d] & & x^y+y^2, id_1^{id_2}+id_2^2 \arrow[d] \\
\begin{array}{c} x^y+y^3, x^y, y^3, x, y, 3, +, id_1^{id_2}+id_2^3, \\ id_1^{id_2}, id_1^3, x^y+y^{const}, y^{const}, id_1^{id_2}+id_2^{const}, id_1^{const} \end{array} \arrow[dr] & & \begin{array}{c} x^y+y^2, id_1^{id_2}+id_2^2, \\ x^y+y^{const}, id_1^{id_2}+id_2^{const} \end{array} \arrow[dl] \\
& x^y+y^{const}, id_1^{id_2}+id_2^{const} &
\end{tikzcd}$$

Match!

Implementation

- Java
- Solr + Lucene
- scalable (indexing $10^{10}+$ formulae without problems
- Mathematical part implements Lucene's interface Tokenizer — able to integrate to any Solr/Lucene based system as DSpace, Elasticsearch…

Search demonstration



```
<math><mrow><msup><mi>x</mi> <mn>2</mn> </msup><mo>+</mo><msup><mi>y</mi> <mn>2</mn> </msup></mrow></math>
```

Canonicalized MathML query:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup> <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup> <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

Search in: MREC 2011.4.439 ▾ Search

Total hits: 36817, showing 1- 30. Searching time: 116 ms

Finite Precision Measurement Nullifies Euclid's Postulates

... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...

score = 3.2980976

arxiv.org/abs/quant-ph/0310035 - cached XHTML

COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88

... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.

Formulae search demonstration comments

Demo web interface: https://mir.fi.muni.cz/webmias-ntcir/

- MathML/TEX input (LaTeXML for conversion to MathML)
- Canonicalization of the query – our own MathCanEval canonicalizer (developed as part of Dean's program at FI MU)
- Matched document snippet generation
- MathJax for nicer math rendering and better portability
- Snuggle TeX for on-the-fly as-you-type rendering

All up and ready on the EuDML system: <http://eudml.org/search/>

How to evaluate math-aware systems like MIaS

IR tradition of evaluation competitions: TREC, CLEF, NTCIR, FIRE,…

Since 2013 there is a new *Math task* at NTCIR for evaluation of math-aware systems.

NTCIR-11 is going to be held in Tokyo, Dec 9–12th: Math task 2, Wikipedia math task.

100,000 arXiv documents to index, splitted on paragraphs. 50 queries, containing *several* textual keywords and *math* formulae.

Up to four runs, and up to thousands ranked answers for every query.

Pooling technique, experts mark pool of most frequent relevant documents in the range from 0 to 4.

Metrics evaluated: P@5, P@10, AVG.

NTCIR-11 Math Task 2

<http://research.nii.ac.jp/ntcir/ntcir-11/program-poster.html#math>

1. (Tokyo) G. Y. Kristianto, G. Topic, F. Ho, and Akiko Aizawa: The MCAT Math Retrieval System for NTCIR-11 Math Track
2. (Braunschweig) G. Pinto, J. Maria, S. Barthel, and W-T. Balke: QUALIBETA at the NTCIR-11 Math 2 Task: An Attempt to Query Math Collections
3. (Bremen) R. Hambasan, M. Kohlhase, and C-C. Prodescu: MathWebSearch at NTCIR-11
4. (Berlin, Washington) M. Schubotz, A. Youssef, V. Markl, H. Cohl and J. Li: Evaluation of Similarity-Measure Factors for Formulae based on the NTCIR-11 Math Task
5. (Rochester) N. Pattaniyil, and R. Zanibbi: Combining TF-IDF Text Retrieval with an Inverted Index over Symbol Pairs in Math Expressions: The Tangent Math Search Engine at NTCIR 2014
6. (Brno) M. Růžička, P. Sojka, and M. Líška: Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy
7. (Vienna) A. Lipani, L. Andersson, F. Piroi, M. Lupu, and A. Hanbury: TUW-IMP at the NTCIR-11 Math-2
8. (Beijing) L. Gao, Y. Wang, L. Hao, and Z. TangThe: The ICST system at NTCIR-11 Math-2

MIaS4NTCIR: data indexing statistics

Table: Index statistics

Indexing times [min]		Index
Wall Clock	CPU	size [GiB]
1,940.0	3,413.55	68

Table: Formulae count statistics

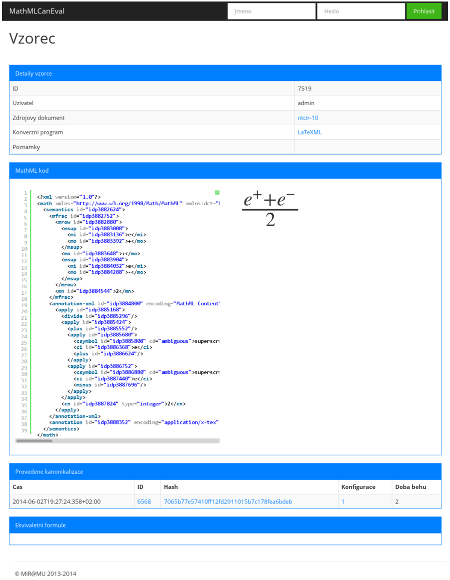
	Formulae	
Documents	Original	Indexed
8,301,545	59,647,566	3,021,865,236

MIaS4NTCIR: canonicalization

We have designed, implemented and continually improve a converter<https://mir.fi.muni.cz/mathml-normalization/> for *both* Presentation and Content MathML for this task.

MathCanEval application developed by Michal Růžička (lead), David Formánek, Dominik Szalai, Robert Šiška, Jakub Adler is designed and developed for evaluation of the canonicalizer.

MIaS4NTCIR: canonicalization II



MathMLCanEval Jmeno Heslo Prihlasit

Vzorec

Detaily vzorce

ID	7519
Uzivatel	admin
Zdrojovy dokument	ntcir-10
Konverzni program	LaTeXML
Poznamky	

MathML kod

```
<?xml version="1.0"?>
<math xmlns="http://www.w3.org/1998/Math/MathML" xmlns:dct="
  <semantics id="idp3882624">
    <mfrac id="idp3882752">
      <mrow id="idp3882880">
        <msup id="idp3883008">
          <mi id="idp3883136">e</mi>
          <mo id="idp3883392">+</mo>
        </msup>
        <mo id="idp3883648">+</mo>
        <msup id="idp3883904">
          <mi id="idp3884032">e</mi>
          <mo id="idp3884288">-</mo>
        </msup>
      </mrow>
      <mn id="idp3884544">2</mn>
    </mfrac>
    <annotation-xml id="idp3884800" encoding="MathML-Content
      <apply id="idp3885168">
        <divide id="idp3885296"/>
        <apply id="idp3885424">
          <plus id="idp3885552"/>
          <apply id="idp3885680">
            <csymbol id="idp3885808" cd="ambiguous">superscr
            <ci id="idp3886368">e</ci>
            <plus id="idp3886624"/>
          </apply>
          <apply id="idp3886752">
            <csymbol id="idp3886880" cd="ambiguous">superscr
            <ci id="idp3887440">e</ci>
            <minus id="idp3887696"/>
          </apply>
        </apply>
        <cn id="idp3887824" type="integer">2</cn>
      </apply>
    </annotation-xml>
    <annotation id="idp3888352" encoding="application/x-tex"
  </semantics>
</math>
```

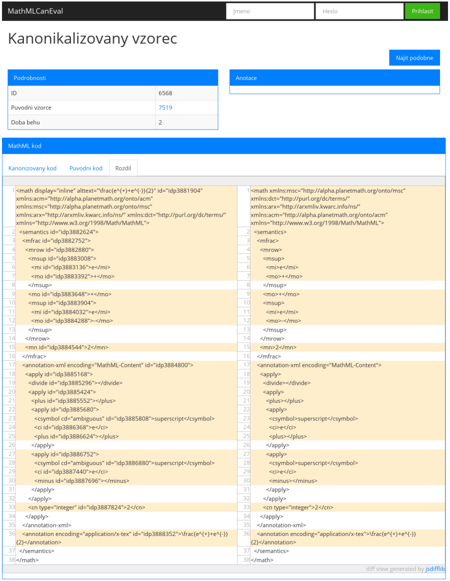
$$\frac{e^{+}+e^{-}}{2}$$

Provedene kanonikalizace

Cas	ID	Hash	Konfigurace	Doba behu
2014-06-02T19:27:24.358+02:00	6568	7065b77e57410ff12fd2911015b7c178fea6bdeb	1	2

Ekvivaletni formule

© MIR@MU 2013-2014



MathMLCanEval Jmeno Heslo Prihlasit

Kanonikalizovany vzorec

Najit podobne

Podrobnosti

ID	6568
Puvodni vzorce	7519
Doba behu	2

Anotace

MathML kod

Kanonizovany kod Puvodni kod Rozdil

```
<math display="inline" alttext="\frac{e^{+}+e^{-}}{2}" id="idp3881904"
xmlns:acm="http://alpha.planetmath.org/onto/acm"
xmlns:msc="http://alpha.planetmath.org/onto/msc"
xmlns:arx="http://arxmliv.kwarc.info/ns/" xmlns:dct="http://purl.org/dc/terms/"
xmlns="http://www.w3.org/1998/Math/MathML">
 <semantics id="idp3882624">
  <mfrac id="idp3882752">
   <mrow id="idp3882880">
    <msup id="idp3883008">
     <mi id="idp3883136">e</mi>
     <mo id="idp3883392">+</mo>
    </msup>
    <mo id="idp3883648">+</mo>
    <msup id="idp3883904">
     <mi id="idp3884032">e</mi>
     <mo id="idp3884288">-</mo>
    </msup>
   </mrow>
   <mn id="idp3884544">2</mn>
  </mfrac>
  <annotation-xml encoding="MathML-Content" id="idp3884800">
   <apply id="idp3885168">
    <divide id="idp3885296"></divide>
    <apply id="idp3885424">
     <plus id="idp3885552"></plus>
     <apply id="idp3885680">
      <csymbol cd="ambiguous" id="idp3885808">superscript</csymbol>
      <ci id="idp3886368">e</ci>
      <plus id="idp3886624"></plus>
     </apply>
     <apply id="idp3886752">
      <csymbol cd="ambiguous" id="idp3886880">superscript</csymbol>
      <ci id="idp3887440">e</ci>
      <minus id="idp3887696"></minus>
     </apply>
    </apply>
    <cn type="integer" id="idp3887824">2</cn>
   </apply>
  </annotation-xml>
  <annotation encoding="application/x-tex" id="idp3888352">\frac{e^{+}+e^{-}}
{2}</annotation>
 </semantics>
</math>
```

```
<math xmlns:msc="http://alpha.planetmath.org/onto/msc"
xmlns:dct="http://purl.org/dc/terms/"
xmlns:arx="http://arxmliv.kwarc.info/ns/"
xmlns:acm="http://alpha.planetmath.org/onto/acm"
xmlns="http://www.w3.org/1998/Math/MathML">
 <semantics>
  <mfrac>
   <mrow>
    <msup>
     <mi>e</mi>
     <mo>+</mo>
    </msup>
    <mo>+</mo>
    <msup>
     <mi>e</mi>
     <mo>-</mo>
    </msup>
   </mrow>
   <mn>2</mn>
  </mfrac>
  <annotation-xml encoding="MathML-Content">
   <apply>
    <divide></divide>
    <apply>
     <plus></plus>
     <apply>
      <csymbol>superscript</csymbol>
      <ci>e</ci>
      <plus></plus>
     </apply>
     <apply>
      <csymbol>superscript</csymbol>
      <ci>e</ci>
      <minus></minus>
     </apply>
    </apply>
    <cn type="integer">2</cn>
   </apply>
  </annotation-xml>
  <annotation encoding="application/x-tex">\frac{e^{+}+e^{-}}
{2}</annotation>
 </semantics>
</math>
```

diff view generated by jsdifflib

MIaS4NTCIR: representation of math for indexing

Concepts of *similarity* and *distributional representations* are central in the design of MIaS. Every formulae is represented in the index as a *set of weighted tokens (subformulae, features)* that grab both structure and content of indexed mathematical formulae. The weighting is computed via small set of rules reflecting similarity distance of indexed tokens to the original formulae: the more similar is token to the original (in size, variable naming, constants used, …), the higher weighting score is stored in the index for a token. On average, currently the formulae representation is distributed over about 30 indexed weighted tokens.

MIaS4NTCIR: query expansion

subquery 1 (the original query):	f_1	f_2	k_1	k_2	k_3
subquery 2:	f_1	f_2	k_1	k_2	
subquery 3:	f_1	f_2	k_1		
subquery 4:	f_1	f_2			
subquery 5:	f_1		k_1	k_2	k_3
subquery 6:			k_1	k_2	k_3

Figure: Complete sequence of subqueries derived from the original user's query

Results merging, finally.

Query expansion results' insight

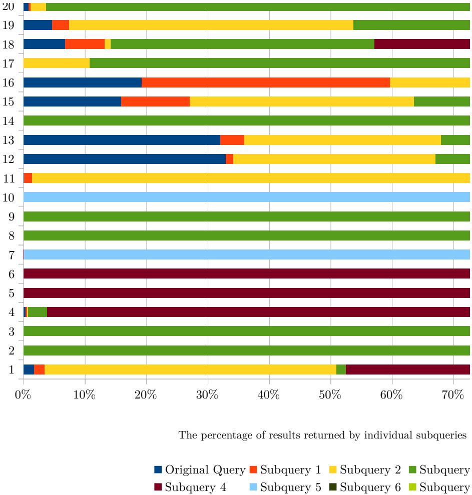


Figure: Relative number of results found using different subqueries for every query in CMath run

MIaS Results: 4 runs PMath, CMath, PCMath, $\TeX$

Table: Results of submitted runs with Relevance Level ≥ 3 (Relevant). Main task team rank is in [ ] for our best runs (in bold).

	PMath	**CMath**	**PCMath**	**$\TeX$**
MAP avg	0.3073	**0.3630 [1]**	0.3594	0.3357
P@10 avg	0.3040	**0.3520 [1]**	0.3480	0.3380
P@5 avg	0.5120	**0.5680 [1]**	0.5560	0.5400

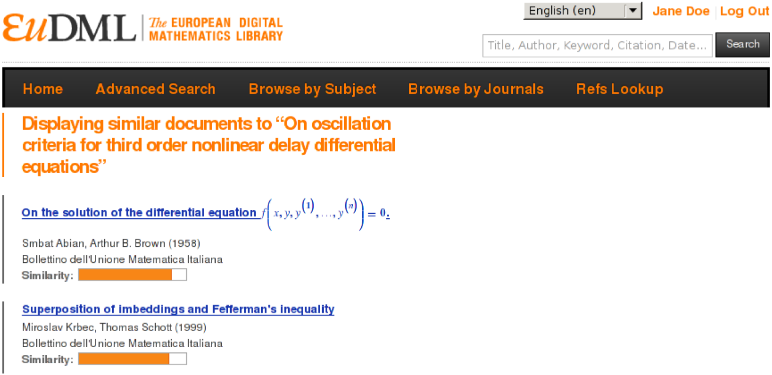
Table: Results of submitted runs with Relevance Level ≥ 1 (Partially Relevant). Number in [ ] is team rank of all runs.

	PMath	**CMath**	**PCMath**	**$\TeX$**
MAP avg	0.2557	**0.2807 [2]**	0.2799	0.2747
P@10 avg	0.5020	0.5440	**0.5520 [1]**	0.5400
P@5 avg	0.8440	**0.8720 [2]**	0.8640	0.8480

Martin wins poster session at FI MU with NTCIR-11 poster :-)

Content Similarity in EuDML: <http://eudml.org>

We have developed and delivered technology DocSim for document *similarity* with Gensim by Radim Řehůřek—„the most robust, efficient and hassle-free piece of software to realize unsupervised semantic modelling from plain text“: <http://radimrehurek.com/gensim/>



Displaying similar documents to “On oscillation criteria for third order nonlinear delay differential equations”

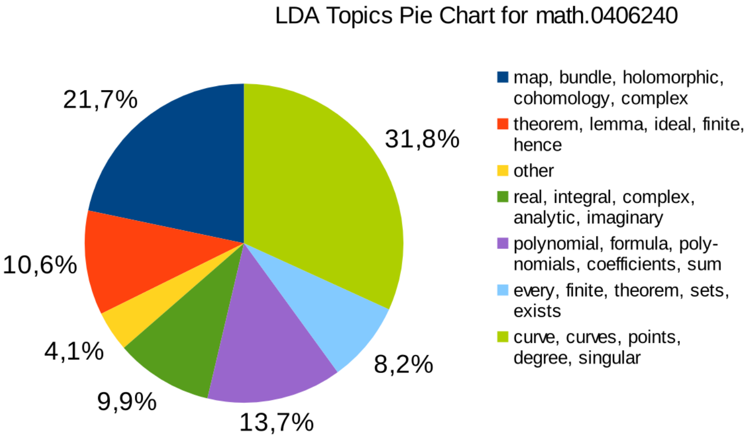
On the solution of the differential equation $f\left(x, y, y^{(1)}, \ldots, y^{(n)}\right) = 0.$

Smbat Abian, Arthur B. Brown (1958)
Bollettino dell'Unione Matematica Italiana
Similarity:

Superposition of imbeddings and Fefferman's inequality

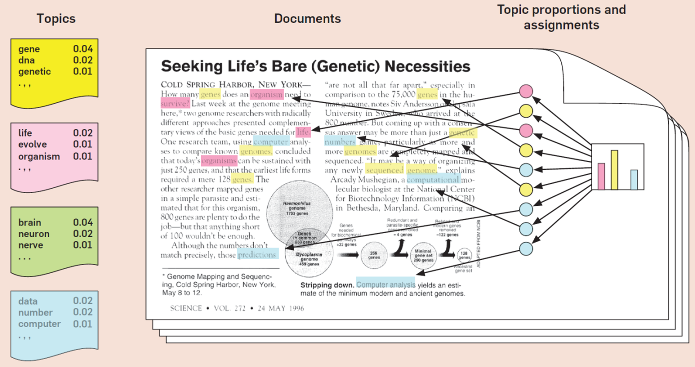
Miroslav Krbec, Thomas Schott (1999)
Bollettino dell'Unione Matematica Italiana
Similarity:

Example I: Automated Meaning Picking from Texts



Probabilistic Topical Modeling: Latent Dirichlet Allocation

- topic: weighted list of words
- document: weighted list of topics


Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
, , ,
life 0.02
evolve 0.01
organism 0.01
, , ,
brain 0.04
neuron 0.02
nerve 0.01
. . .
data 0.02
number 0.02
computer 0.01
, , ,
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

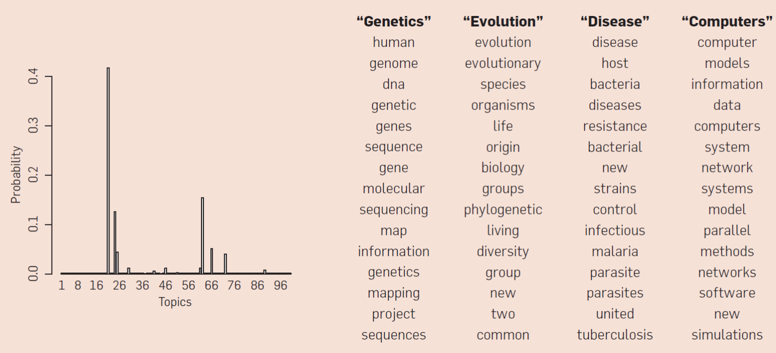
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topical Modeling: Latent Dirichlet Allocation II

- all topics computed automatically from document corpora

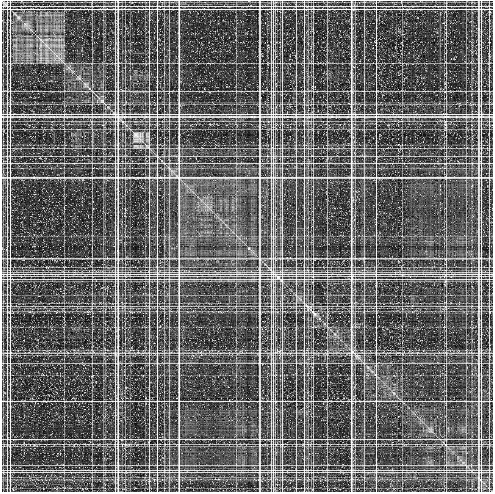


“Genetics”	“Evolution”	“Disease”	“Computers”
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

How math formulae affect document similarities?

- how weight metadata, full texts, formulae?
- how represent formulae representations for similarity computation?
- which learning methods?
- how to evaluate performance?
- MSC – mathematical subject classification *mandatory* for math publications (ZMath, MathSciNet)
- MSC induces equivalence: similarity of papers of the same primary top-level MSC should have lower variance than with other
- picked papers with just one primary MSC for evaluation of math representation and methods
- winner is the method with lowest mean of variances within same MSC document blocks

Matrix 33 Variance Mean: 3390.8107



Variance Matrix Mean

3390.8107

Method

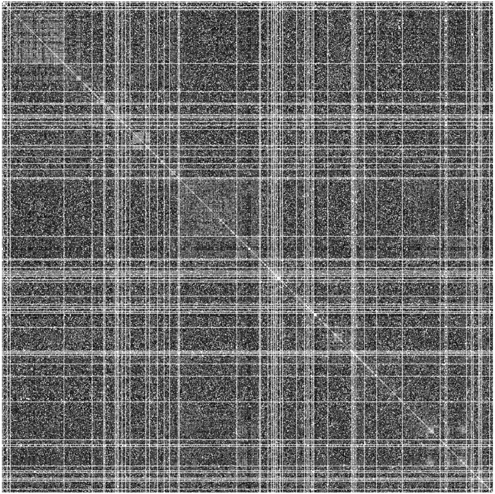
TfIdf-LSI (200 topics)

MTerm Weight Conversion

min(trunc(10 * mtermWeight), 4)

Description	0
SimilarityText	6
Authors	30
Language	0
Category	0
Id	0
Title	30
Keywords	50
MscCodes	0
MathMathML	0
MathBeginingElements	0
MathElements	0
MathMterms	0
MathWightedMterms	1

Matrix 15 Variance Mean: 4117.3155



Variance Matrix Mean

4117.3155

Method

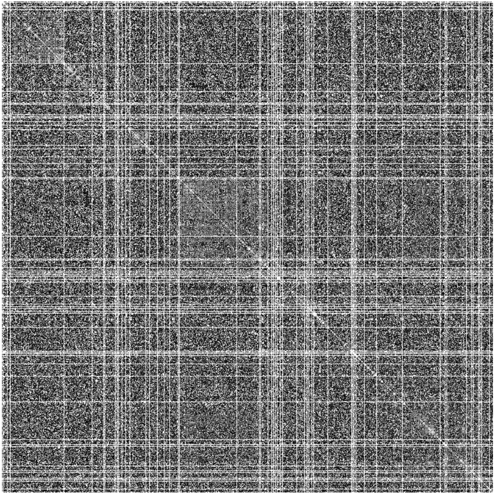
TfIdf-LSI (200 topics)

MTerm Weight Conversion

min(trunc(10 * mtermWeight), 4)

Description	0
SimilarityText	0
Authors	30
Language	0
Category	0
Id	0
Title	30
Keywords	50
MscCodes	0
MathMathML	0
MathBeginingElements	0
MathElements	0
MathMterms	0
MathWightedMterms	1

Matrix 15 Variance Mean: 6971.8214



Variance Matrix Mean

6971.8214

Method

TfIdf-LSI (200 topics)

MTerm Weight Conversion

min(trunc(10 * mtermWeight), 4)

Description	0
SimilarityText	0
Authors	30
Language	0
Category	0
Id	0
Title	30
Keywords	50
MscCodes	0
MathMathML	0
MathBeginingElements	0
MathElements	0
MathMterms	0
MathWightedMterms	0

Evaluation framework for math, knowledge representation and machine learning methods

Yesterday's first results:

```
Matrix 30 Variance Mean: 3517.1352
Matrix 27 Variance Mean: 3562.7631
Matrix 21 Variance Mean: 3591.9553
Matrix 24 Variance Mean: 3631.0433
Matrix 18 Variance Mean: 3657.6139
Matrix 15 Variance Mean: 4117.3155
Matrix 9  Variance Mean: 4290.0905
Matrix 12 Variance Mean: 5365.2903
Matrix 3  Variance Mean: 6888.0026
Matrix 6  Variance Mean: 6914.4168
Matrix 36 Variance Mean: 6971.8214
```

confirms hyphotesis that *math matters* and that our math (distributional) representation gives best results.

Semantic gap between lexical surface of the text and its meaning in [M]IR

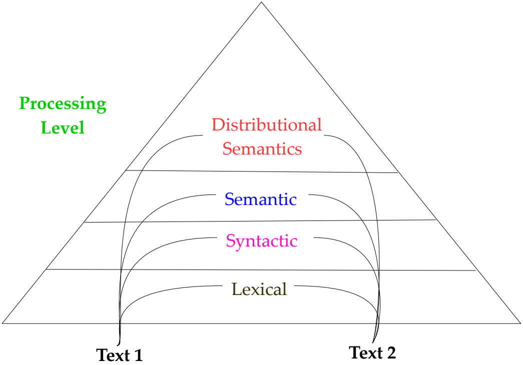
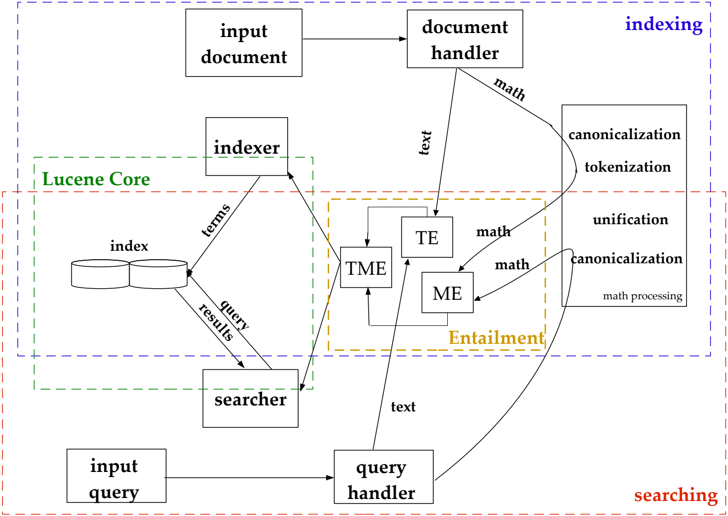
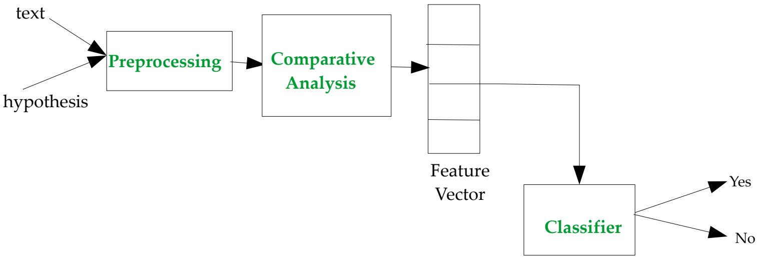


Figure: Natural language processing levels

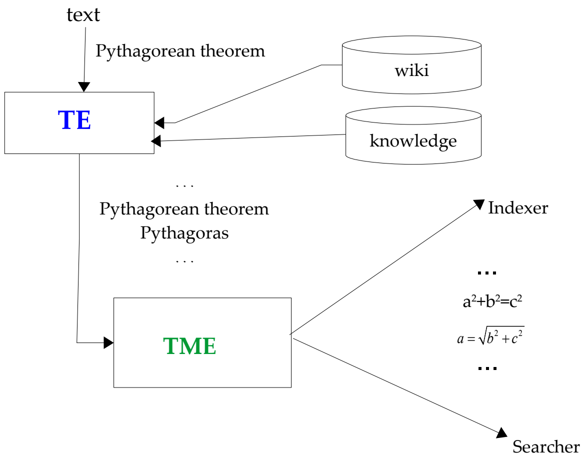
New MIaS architecture with textual and math entailment modules



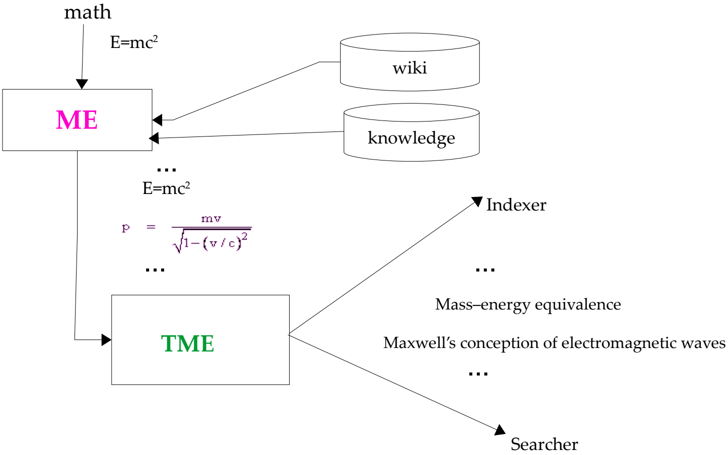
General Textual Entailment architecture



Data flow in TE and TME modules



Data flow in ME and TME modules



Future work?

- full text mining in semantic direction (typesetting^{-1}), higher level NLP
- globalization (Google Scholar), deploying global knowledge bases
- personalization (up to the individual's preferences)
- increase of automation and precision on semantic level

Future challenges

- Math-aware knowledge representation
- Math entailment (Partha Pakray), ‘flexiformat’ processing, ‘canonicalization’ of math formulae
- Math-aware corpora processing
- robust Math OCR is necessary
- robust born-digital PDF2Math conversion is needed as well
- only then challenges as: multilingual math retrieval, MathML indexing and search, math common sense, text and math disambiguation and understanding, mathematical document classification, document similarity could be possible

Challenge of math-aware distributional semantics processing

- Math-aware knowledge representation: handling abstractions, i high-dimensional vector space representations?
- math2vec? ‘smooth’ vector space representation of math formulae learnt by recurrent neural network: math2vec aka word2vec (T. Mikolov from Brno, now Google), GloVe (Stanford’s tool for distributional semantics), COMPOSES Semantic vectors (M. Baroni’s way of distributional semantics)
- Hyperlapsed vector space representation of documents (narrative qualitites, rephrased plagiarism)

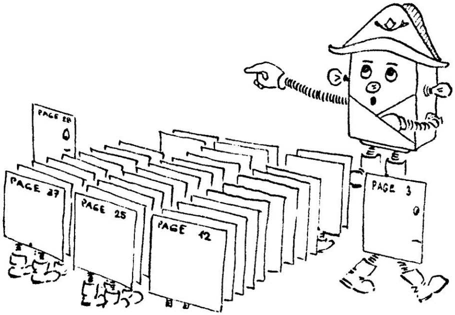
Challenge of math-aware corpora processing and tools

- Canonicalization of math formulae processing (MathCanEval)
- Switching between different levels of structured data
- tools adaptation (handling trees and abstractions), ideally on data acquired and tagged without supervision

Challenge of Evaluation of Math Information Retrieval

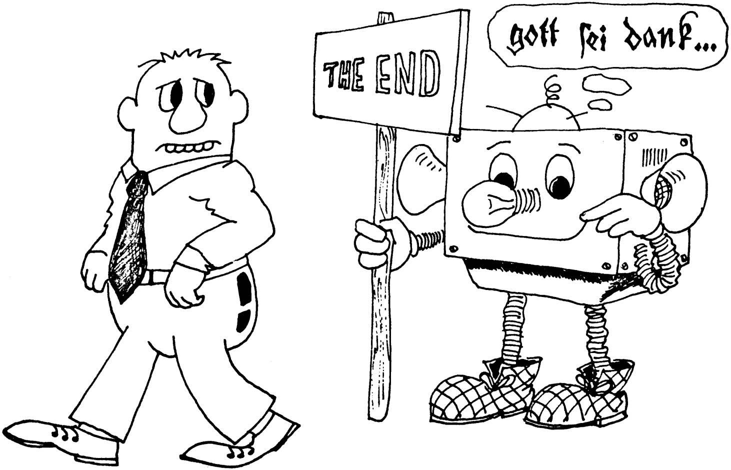
- what works in math-aware IR, UI, pragmatics
- first MIR happening in 2012, now regular Math Tasks at NTCIR-10, NTCIR-11
- deploying MIaS and our tools in the GDML project

Acknowledgments and questions?



Acknowledgements: EuDML and DML-CZ projects (funding), EuDML and DML-CZ colleagues, Martin Líška, *Michal Růžička*, Partha Pakray, Radim Řehůřek, David Formánek, Dominik Szalai, Robert Šiška, Jakub Adler, Radim Hatlapatka, Martin Jarmar, Maroš Kucbel, Zuzana Nevěřilová, Mirek Bartošek, Martin Šárfy, Vlastík Krejčíř, Petr Kovář, Vlastimil Dohnal, and many, many other authors and contributors of tools used.

That's it!



Archambault, D., Moço, V.: Canonical MathML to Simplify Conversion of MathML to Braille Mathematical Notations. In: Miesenberger, K., Klaus, J., Zagler, W., Karshmer, A. (eds.) Computers Helping People with Special Needs, Lecture Notes in Computer Science, vol. 4061, pp. 1191–1198. Springer Berlin / Heidelberg (2006), <http://dx.doi.org/10.1007/11788713_172>

Grimm, J.: Producing MathML with Tralics. In: Sojka [4], pp. 105–117, <http://dml.cz/dmlcz/702579>

MREC – Mathematical REtrieval Collection, <http://nlp.fi.muni.cz/projekty/eudml/MREC/index.html>

Sojka, P. (ed.): Towards a Digital Mathematics Library. Masaryk University, Paris, France (Jul 2010), <http://www.fi.muni.cz/ sojka/dml-2010-program.html>

Sojka, P., Líška, M.: Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In: Davenport, J.H., Farmer, W., Urban, J., Rabe, F., (eds.) Proceedings of CICM Conference 2011 (Calculemus/MKM). Lecture Notes in Artificial Intelligence, LNAI, vol. 6824, pp. 228–243. Springer-Verlag, Berlin, Germany (Jul 2011), <http://dx.doi.org/10.1007/978-3-642-22673-1_16>

Líška, Martin and Petr Sojka and Michal Růžička. Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task. In Proceedings of the 10th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Math Pilot Task. pp. 686-691. NII, Tokyo, 2013. PDF

D. Formánek, M. Líška, M. Růžička, and P. Sojka. Normalization of digital mathematics library content. In J. Davenport, J. Jeuring, C. Lange, and P. Libbrecht, editors, 24th OpenMath Workshop, 7th Workshop on Mathematical User Interfaces (MathUI), and Intelligent Computer Mathematics Work in Progress, number 921 in CEUR Workshop Proceedings, pp. 91–103, Aachen, 2012.

Sojka, Petr and Martin Líška. The Art of Mathematics Retrieval. In Matthew R. B. Hardy , Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2. <http://dx.doi.org/10.1145/2034691.2034703>

Stamerjohanns, H., Ginev, D., David, C., Misev, D., Zamdzhiev, V., Kohlhase, M.: MathML-aware Article Conversion from LaTeX. In: Sojka, P. (ed.) Proceedings of DML 2009. pp. 109–120. Masaryk University, Grand Bend, Ontario, CA (Jul 2009), <http://dml.cz/dmlcz/702561>

Stamerjohanns, H., Kohlhase, M., Ginev, D., David, C., Miller, B.: Transforming Large Collections of Scientific Publications to XML. Mathematics in Computer Science 3, 299–307 (2010), <http://dx.doi.org/10.1007/s11786-010-0024-7>

Sylwestrzak, W., Borbinha, J., Bouche, T., Nowiński, A., Sojka, P.: EuDML—Towards the European Digital Mathematics Library. In: Sojka [4], pp. 11–24, <http://dml.cz/dmlcz/702569>

Martin Líška, Petr Sojka, Michal Růžička, and Petr Mravec.
Web Interface and Collection for Mathematical Retrieval.
In Petr Sojka and Thierry Bouche, editors, *Proceedings of DML 2011*, pages 77–84, Bertinoro, Italy, July 2011. Masaryk University. <http://dml.cz/dmlcz/702604>.

Credits for LDA pictures goes to David M. Blei.

Credits for illustrations goes to Jiří Franek.