NTCIR-11 Math-2 Task Overview

Akiko Aizawa & Michael Kohlhase & Iadh Ounis & Moritz Schubotz

http://kwarc.info/kohlhase
Computer Science
Jacobs University Bremen, Germany

NTCIR-11, Decemter 10. 2014

1 Introduction & Motivation for Math-2 Task

Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!

Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- There is a lot of documents with maths
  - there are 120.000 journal articles per year in pure/applied math, 3.5 Million overall
  - 50 million science articles in 2010 [Jin10] with a doubling time of 8-15 years [Lvl10]

  And this excludes gray literature, engineering, and school textbooks.
  - Even in the Renaissance, polymaths like Leonardo de Vinci were a rare exception.

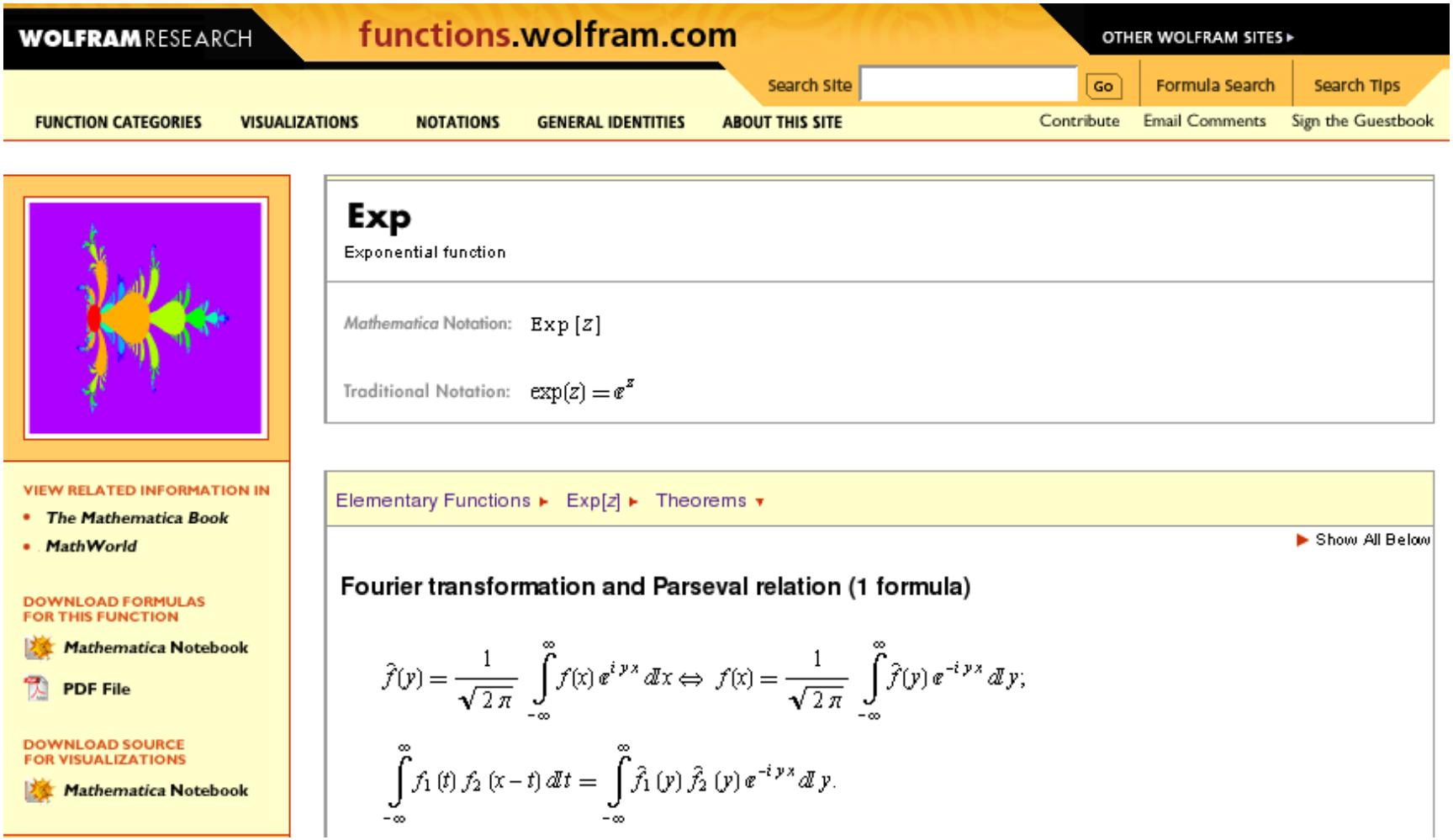
Introduction/Background

- Mathematics plays a fundamental role in Science, Technology, and Engineering
  (learn from Math, apply for STEM)
- Mathematical knowledge is rich in content, sophisticated in structure, and technical in presentation!
- There is a lot of documents with maths
  - there are 120.000 journal articles per year in pure/applied math, 3.5 Million overall
  - 50 million science articles in 2010 [Jin10] with a doubling time of 8-15 years [Lvl10]

  And this excludes gray literature, engineering, and school textbooks.
  - Even in the Renaissance, polymaths like Leonardo de Vinci were a rare exception.
- We need IR support to deal with this! ($\leadsto$ NTCIR-11 Math-2 Task)

Mathematics Resources on the Web

- **Example 1.1 (The Wolfram Functions Site)** contains $\geq$ 307k Formuae



Applications of Math Information Retrieval

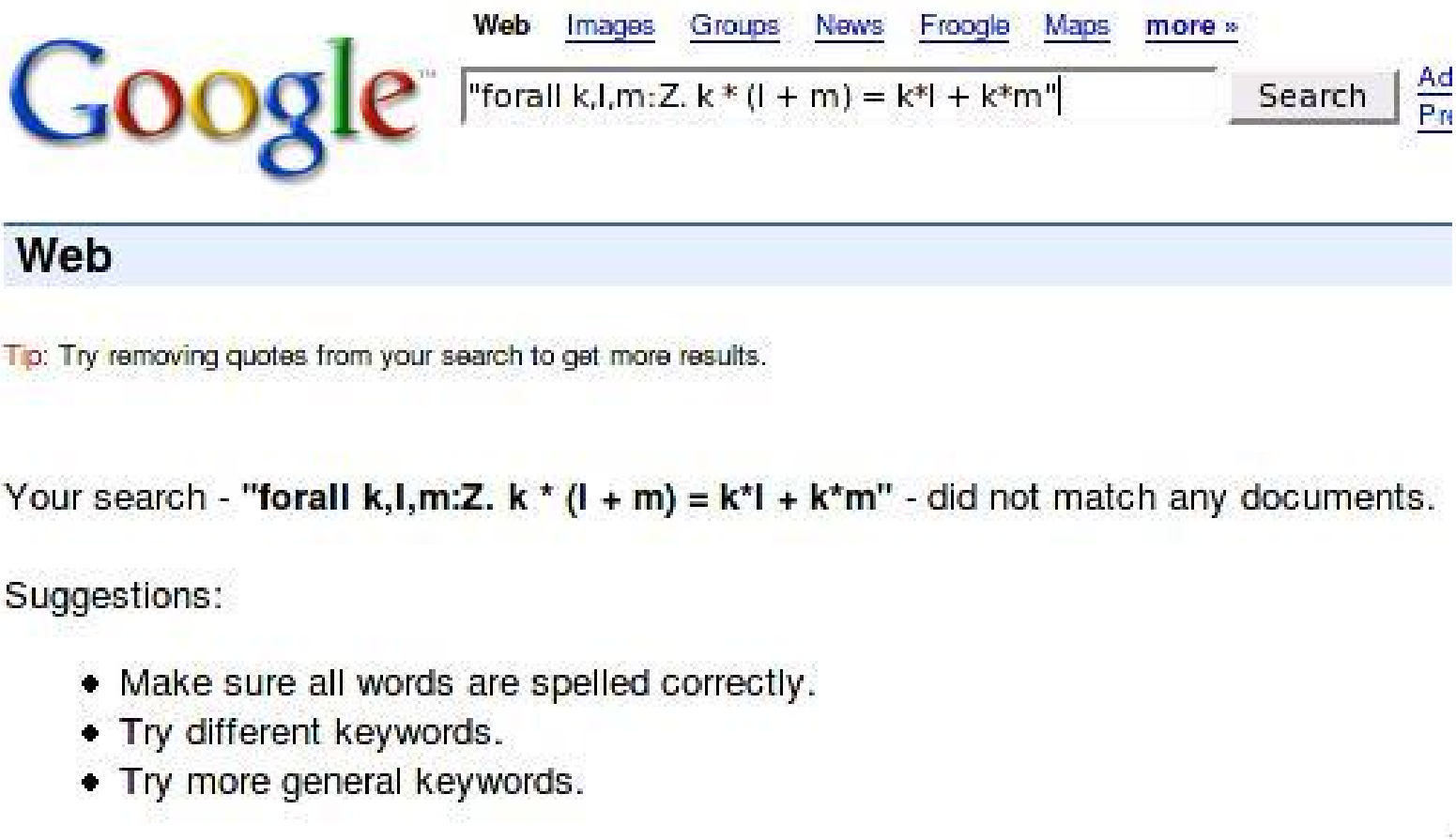
- Potential Applications (some in prototype state)
  - Literature search/Related Work (where have I seen this before?)
  - Applicable Theorem Search (I am stuck in a derivation)
  - Plagiarism detection (not just for the humanities)
  - Formulae in Excel spreadsheets (are just formulae as well)
  - Computations/Documentation in mathematical/symbolic software
  - time series search (e.g. via polynomial interpolations)
- Production systems in math information systems
  - MIAS in EU-DML, MathWebSearch in Zentralblatt Math

More Mathematics on the Web

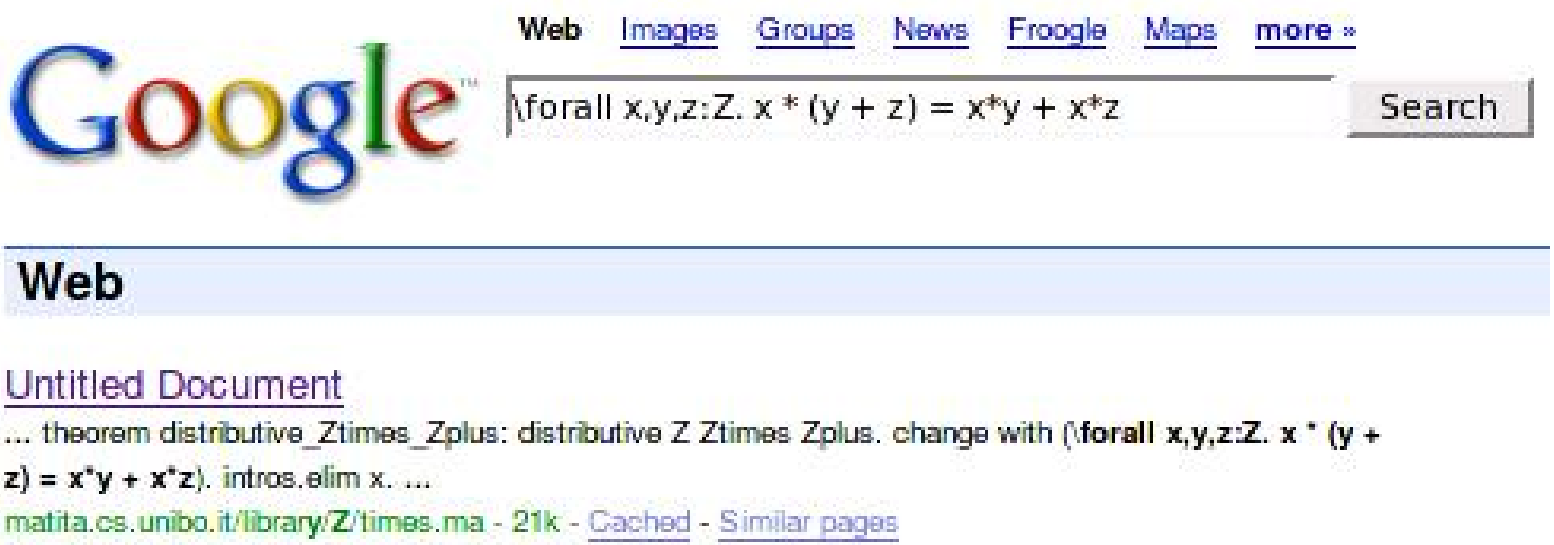
- The Connexions project (http://cnx.org)
- Wolfram Inc. (http://functions.wolfram.com)
- Eric Weisstein's MathWorld (http://mathworld.wolfram.com)
- Digital Library of Mathematical Functions (http://dlmf.nist.gov)
- Cornell ePrint arXiv (http://www.arxiv.org)
- Zentralblatt Math (http://www.zentralblatt-math.org)
- . . . Engineering Company Intranets, . . .
- Question: How will we find content that is relevant to our needs
- Idea: try Google (like we always do)
- Sicenario: 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



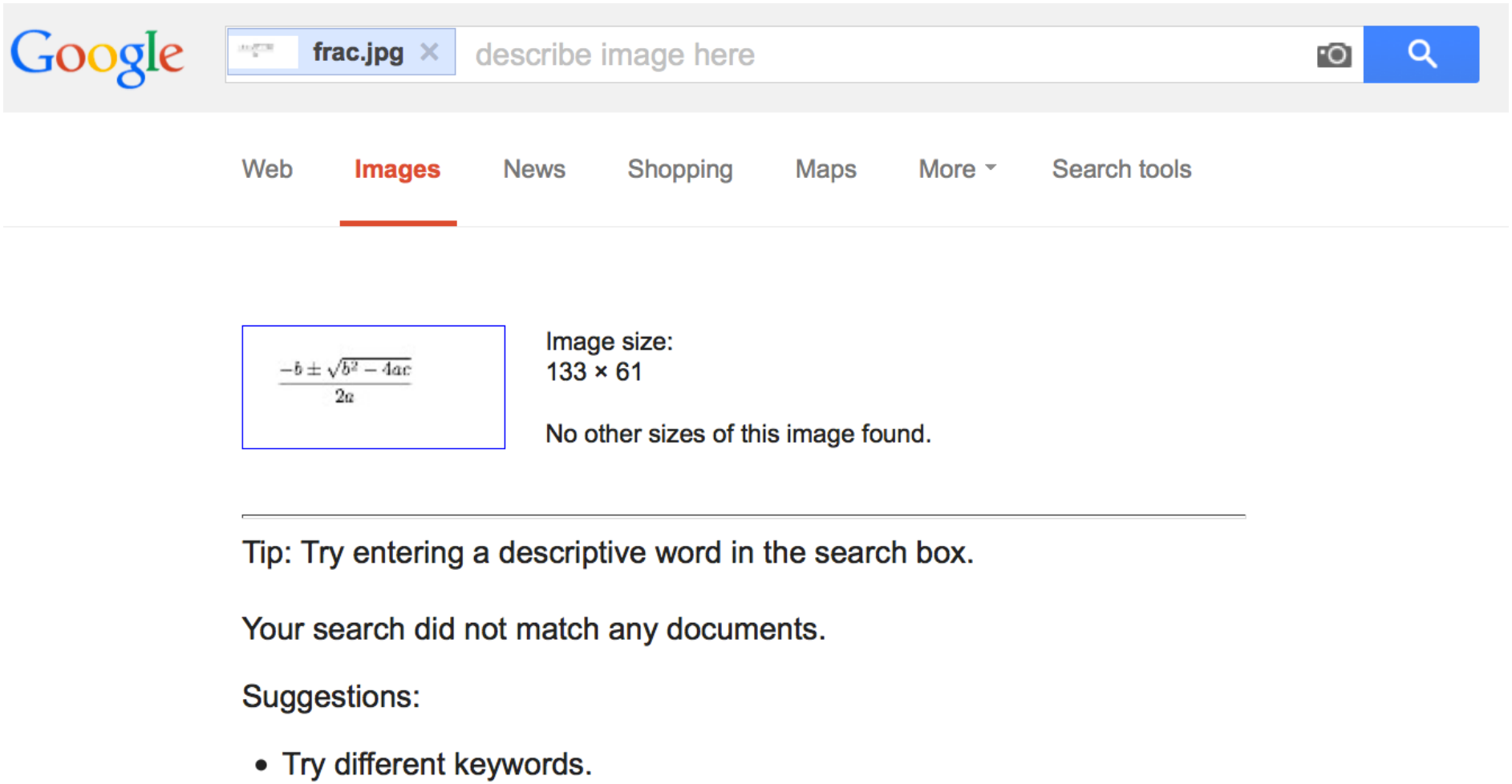
Searching for Distributivity



Searching for Distributivity

Does Image Search help?

- Math formulae are visual objects, after all (let's try it)



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?

Of course Google cannot work out of the box

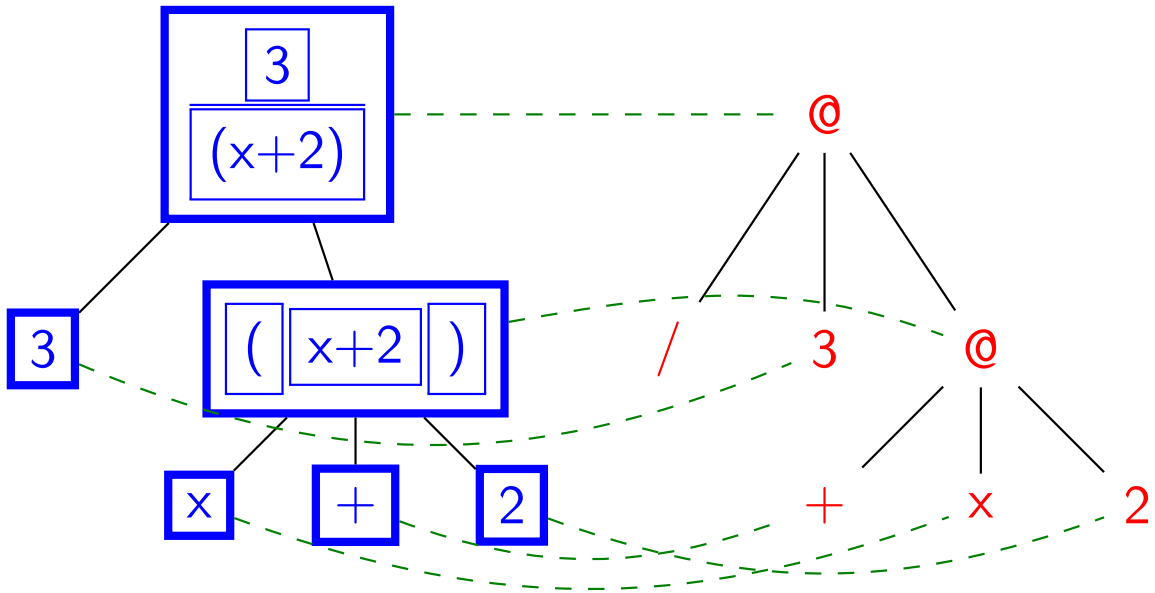
- Formulae are not words:
  - a, b, c, k, l, m, x, y, and z are (bound) variables.
    (do not behave like words/symbols)
  - where are the word boundaries for "bag-of-words" methods?
- Idea: Need a special treatment for formulae (translate into "special words")
  Indeed this is done ([MY03, MM06, LM06, 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])
- Idea: Use Automated Reasoning Techniques
  (Term Indexing [Nor06, KŞ06, KMP12])
- Idea: Use standard IR techniques (Learn from the NTCIR crowd?)

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])
- Idea: Use Automated Reasoning Techniques
  (Term Indexing [Nor06, KŞ06, KMP12])
- Idea: Use standard IR techniques (Learn from the NTCIR crowd?)
- Which one is best?: We do not really know, evaluation is very difficult
- Future: maybe even mix/integrate the respective best features (once we know)

Math Markup e.g. in MathML and LaTeX

- MathML3 is a W3C Recommendation for representing Formulae [ABC+10]
- Idea: Combine the presentation and content markup and cross-reference

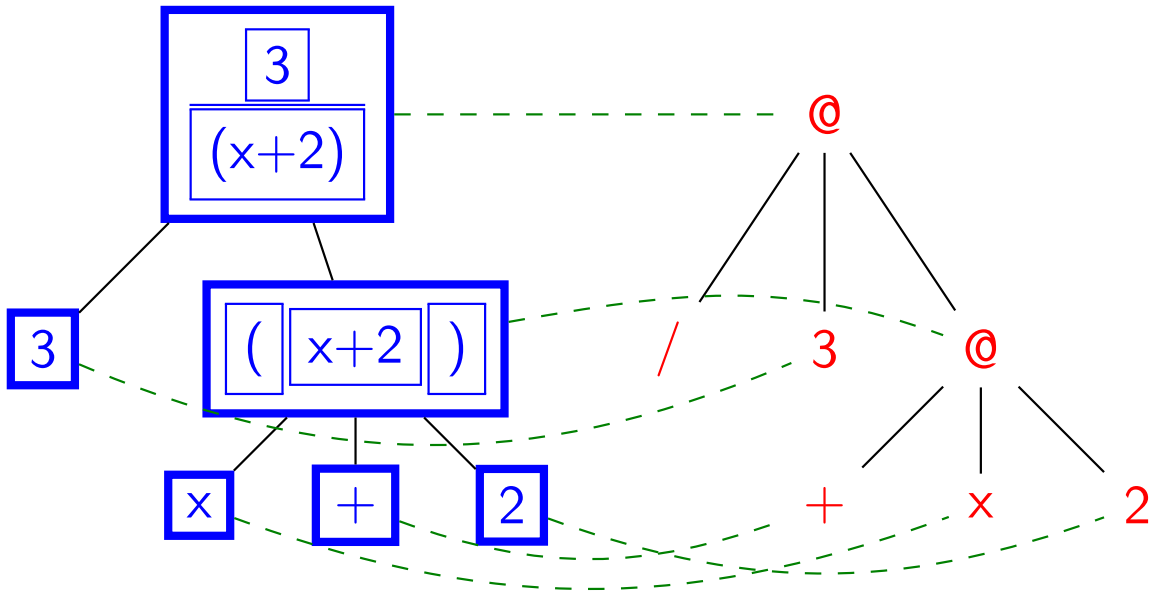


- use e.g. for semantic copy and paste.

(click on presentation, follow link and copy content)

Math Markup e.g. in MathML and LaTeX

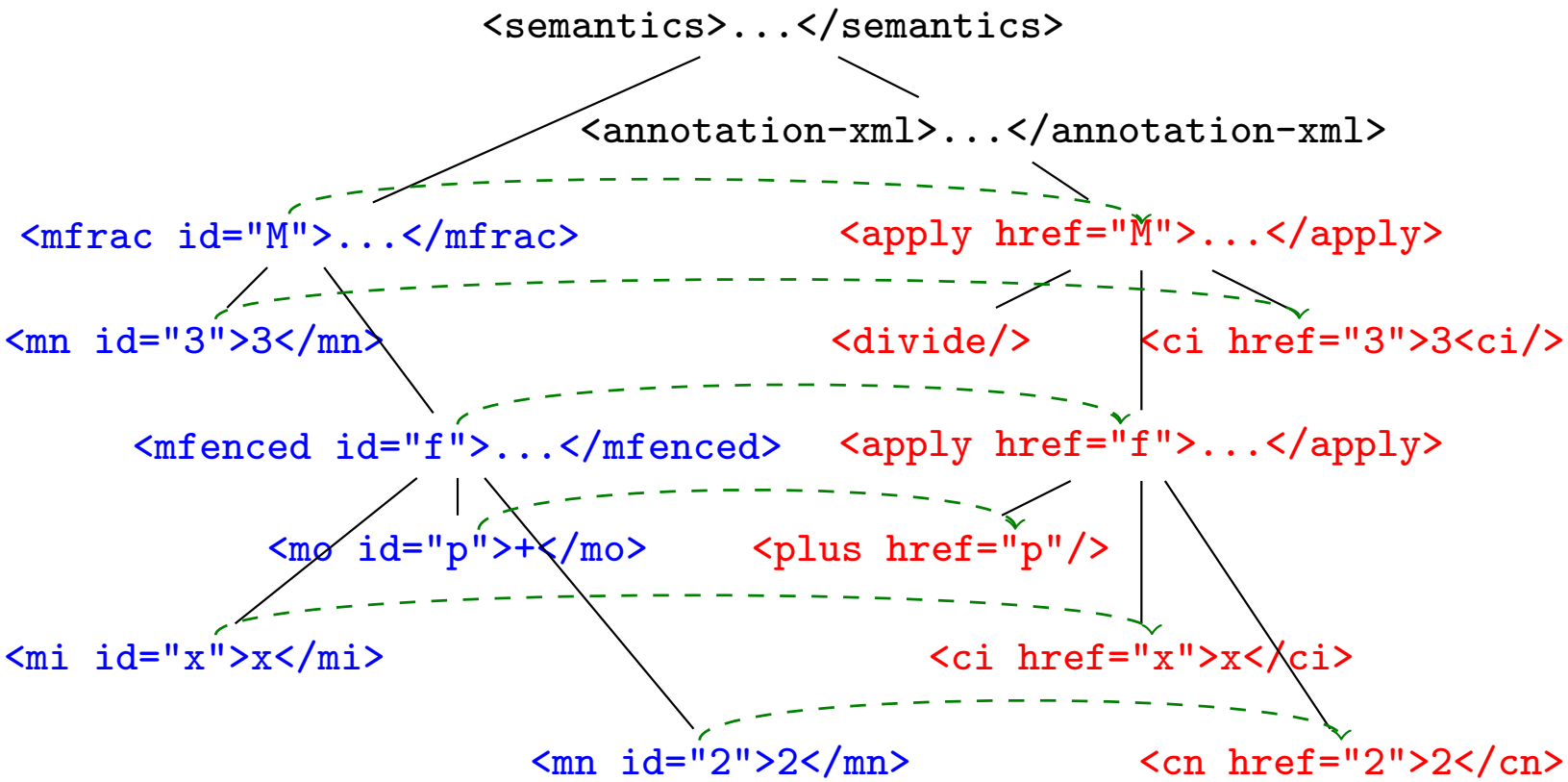
- MathML3 is a W3C Recommendation for representing Formulae [ABC+10]
- Idea: Combine the presentation and content markup and cross-reference



- use e.g. for semantic copy and paste.
  (click on presentation, follow link and copy content)
- But: Formulae are mostly written in LaTeX, e.g. `\frac{3}{(x+2)}`
- Solution: Write LaTeX, convert to HTML5≙ HTML+MathML+SVG

Parallel Markup Markup in MathML

- Concrete Realization in MathML: `semantics` element with presentation as first child and content in `annotation-xml` child




Ron Ausbrooks, Stephen Buswell, David Carlisle, Giorgi Chavchanidze, Stéphane Dalmas, Stan Devitt, Angel Diaz, Sam Dooley, Roger Hunter, Patrick Ion, Michael Kohlhase, Azzeddine Lazrek, Paul Libbrecht, Bruce Miller, Robert Miner, Murray Sargent, Bruce Smith, Neil Soiffer, Robert Sutor, and Stephen Watt.
Mathematical Markup Language (MathML) version 3.0.
W3C Recommendation, World Wide Web Consortium (W3C), 2010.


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 LNCS, pages 17–32. Springer Verlag, 2006.


Arif Jinha.
Article 50 million: an estimate of the number of scholarly articles in existence.
Learned Publishing, 23(3):258–263, 2010.


Michael Kohlhase, Bogdan A. Matican, and Corneliu C. Prodescu.
MathWebSearch 0.5 – Scaling an Open Formula Search Engine.

In Johan Jeuring, John A. Campbell, Jacques Carette, Gabriel Dos Reis, Petr Sojka, Makarius Wenzel, and Volker Sorge, editors, *Intelligent Computer Mathematics*, number 7362 in LNAI, pages 342–357. Springer Verlag, 2012.

Michael Kohlhase and Ioan Şucan.
A search engine for mathematical formulae.
In Tetsuo Ida, Jacques Calmet, and Dongming Wang, editors, *Proceedings of Artificial Intelligence and Symbolic Computation, AISC'2006*, number 4120 in LNAI, pages 241–253. 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.

Peder Olesen Larsen and Markus von Ins.
The rate of growth in scientific publication and the decline in coverage provided by science citation index.
Scientometrics, 84(3):575–603, 2010.

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.

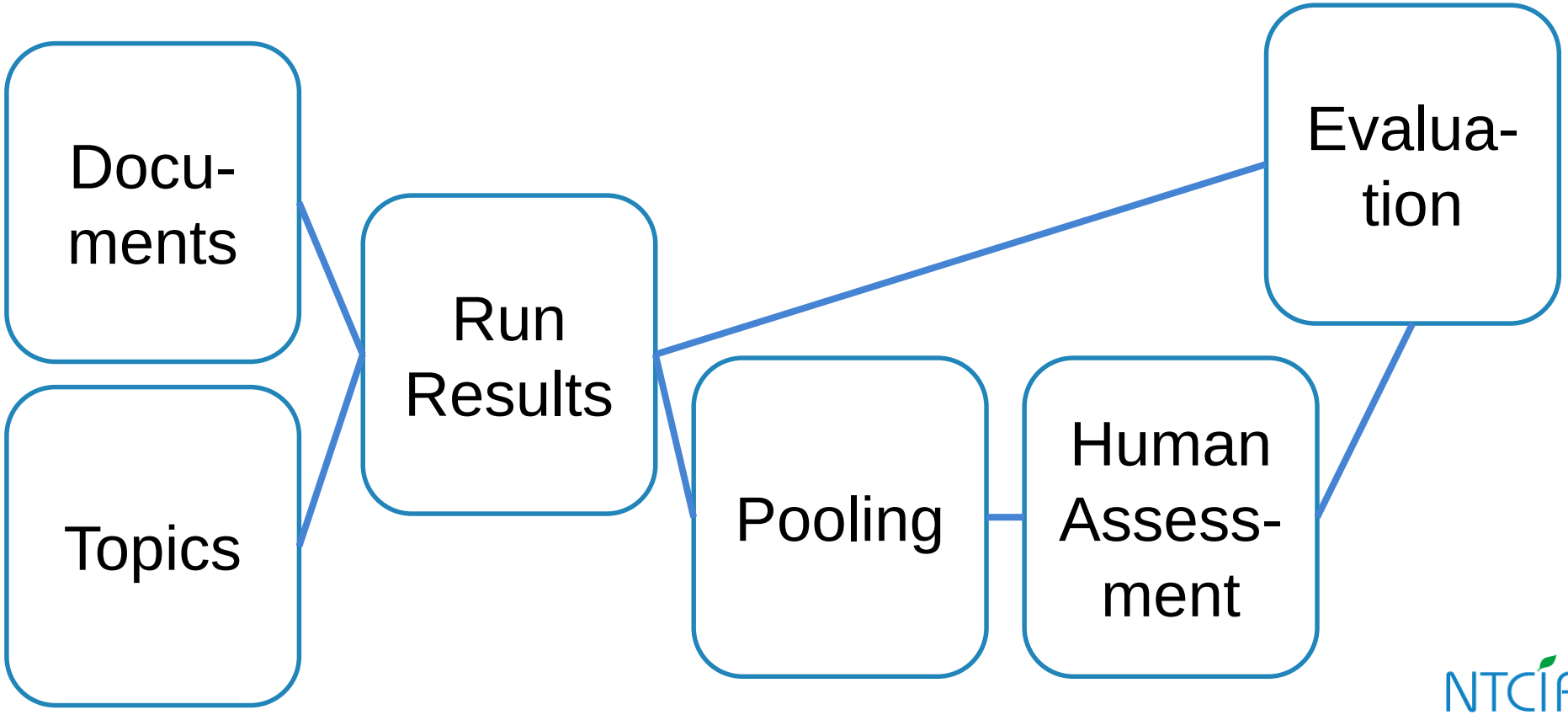

Rajesh Munavalli and Robert Miner.
Mathfind: a math-aware search engine.
In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 735–735, New York, NY, USA, 2006. ACM Press.


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.


Immanuel Normann.
Extended normalization for *e*-retrieval of formulae.
2006.

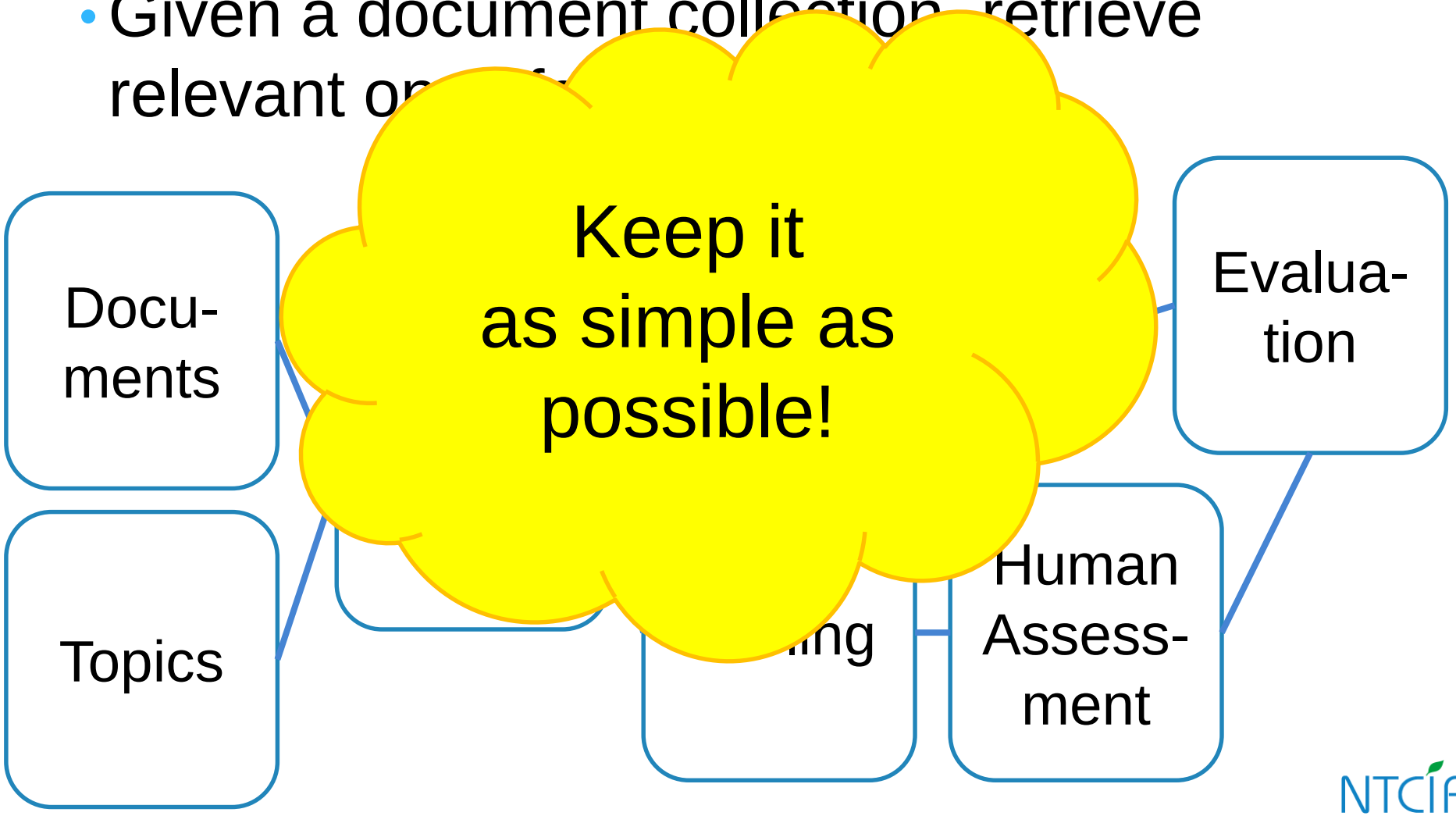
Math-2 Task Outline

- Given a document collection, retrieve relevant ones for a given query.

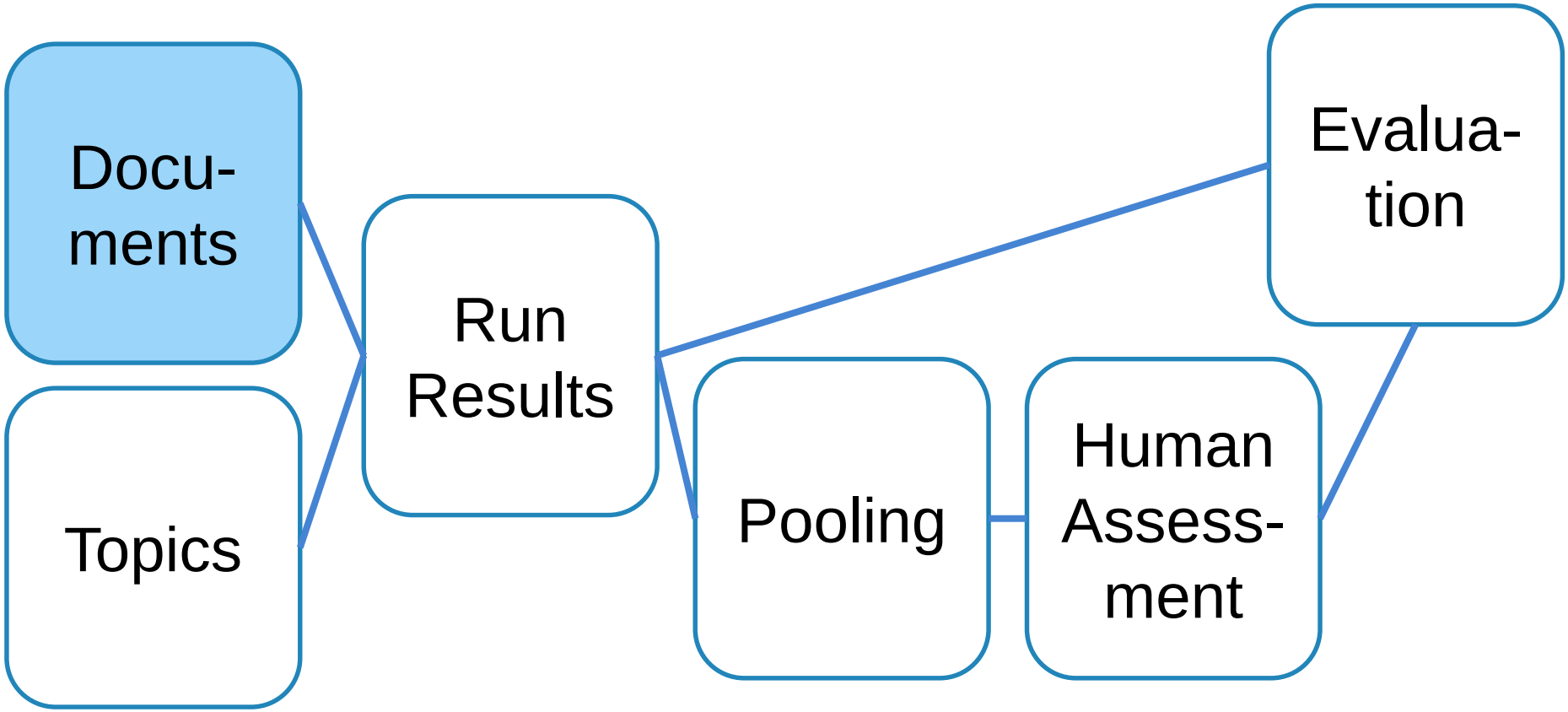


Math-2 Task Outline

- Given a document collection, retrieve relevant op



Math-2 Task Outline



Document Collection

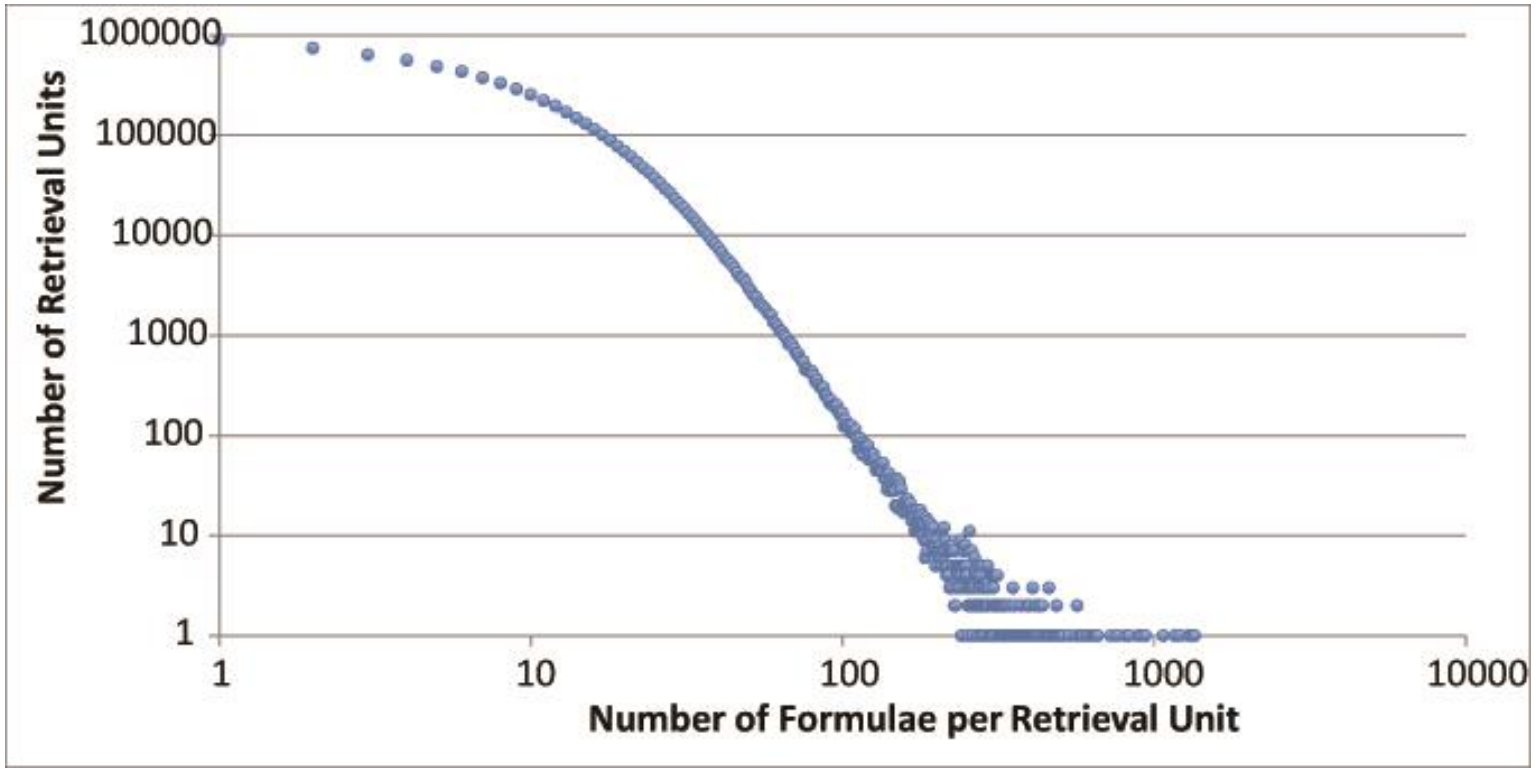
- About 105,120 scientific articles (in English) from arXiv
  - arXiv categories: math, cs, physics:math-ph, stat, physics:hep-th, physics:nlin
- Converted from LaTeX to an HTML+MathML-based format by the KWARC project (http://kwarc.info/).
- 8,301,578 search units with about 60 millions math formulae

Document Collection

- About 105,120 scientific articles (in English) from arXiv
  - arXiv categories: math, physics:hep-th, physics:nlin
- Converte… …ased format b… …).
- 8,3… …h about 60 … formulae

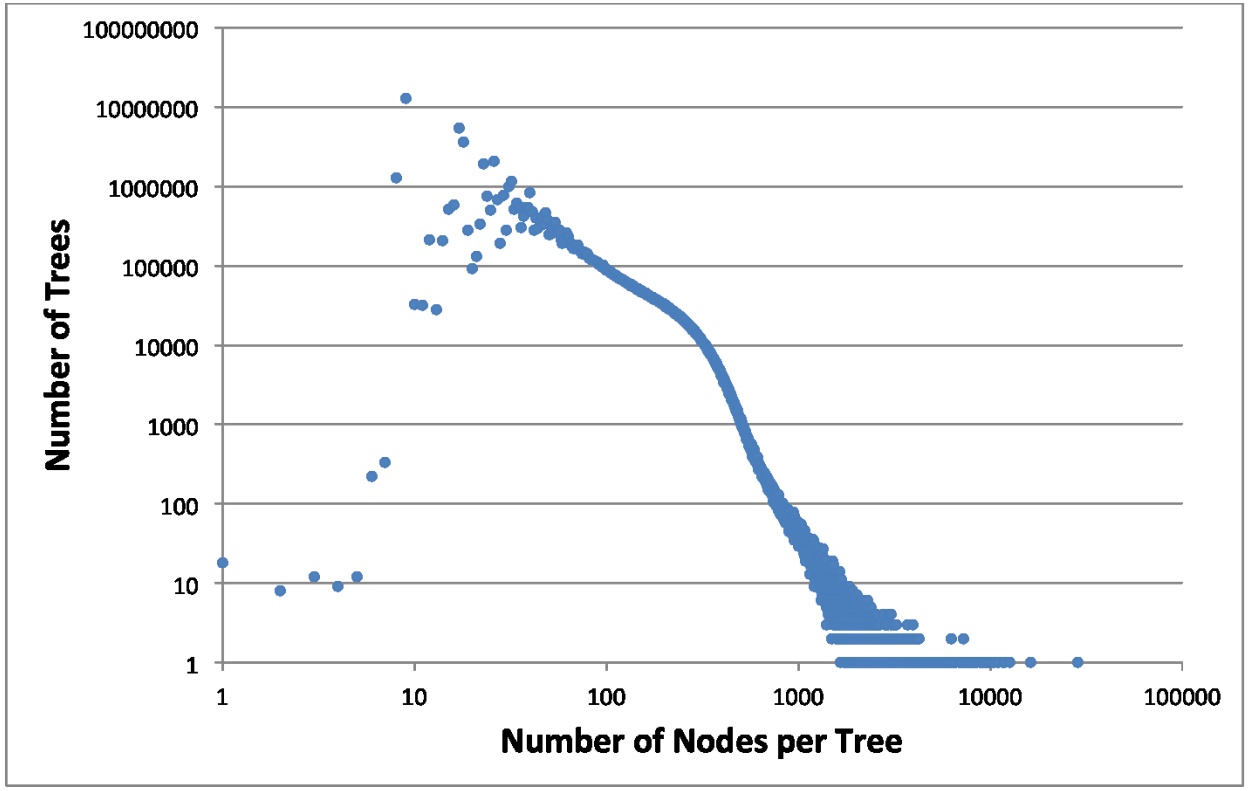
Real
&
Large-scale
XML-data

Search Units: Statistics



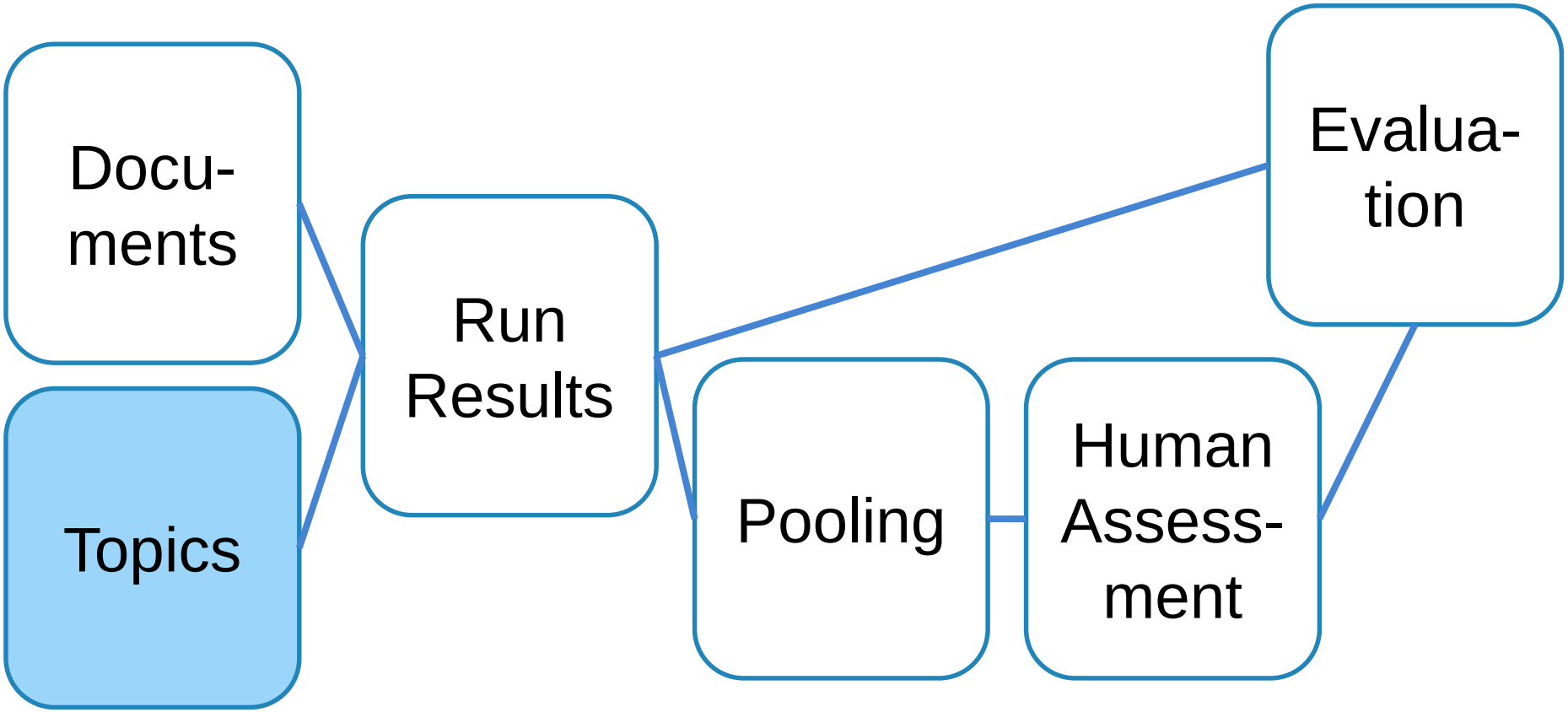
- Number of formulae per retrieval unit: 95% of the retrieval units have 23 math formulae or less

Formula Tree: Statistics



- Number of Nodes per formula tree: much heterogeneity in the target collection

Math-2 Task Outline



Topics

- 50 topics
  - For participants:
    - Query (formulae + keywords)
    - Forumlae are expressed using Latex, MathML-Presentation Markup, and MathML-Content Markup
    - May contain named query variables that act as wildcards
  - For judges:
    - Narrative (precise description of the user situation and information need and relevancy criteria)
    - Example of relevant documents

Topic example

NTCIR11-Math2–16

Formula Query: $f(x) = x$
Keyword: fixed point

NTCIR11-Math2–17

Formula Query: $f(z) = z^{d} + c$
Keyword: Mandelbrot
Keyword: dynamical plane

NTCIR11-Math2–18

Formula Query: $\frac{az+b}{cz+d}$
Keyword: Mobius
Keyword: Möbius
Keyword: automorphism

Topic example

NTCIR11 Math2

Query variables

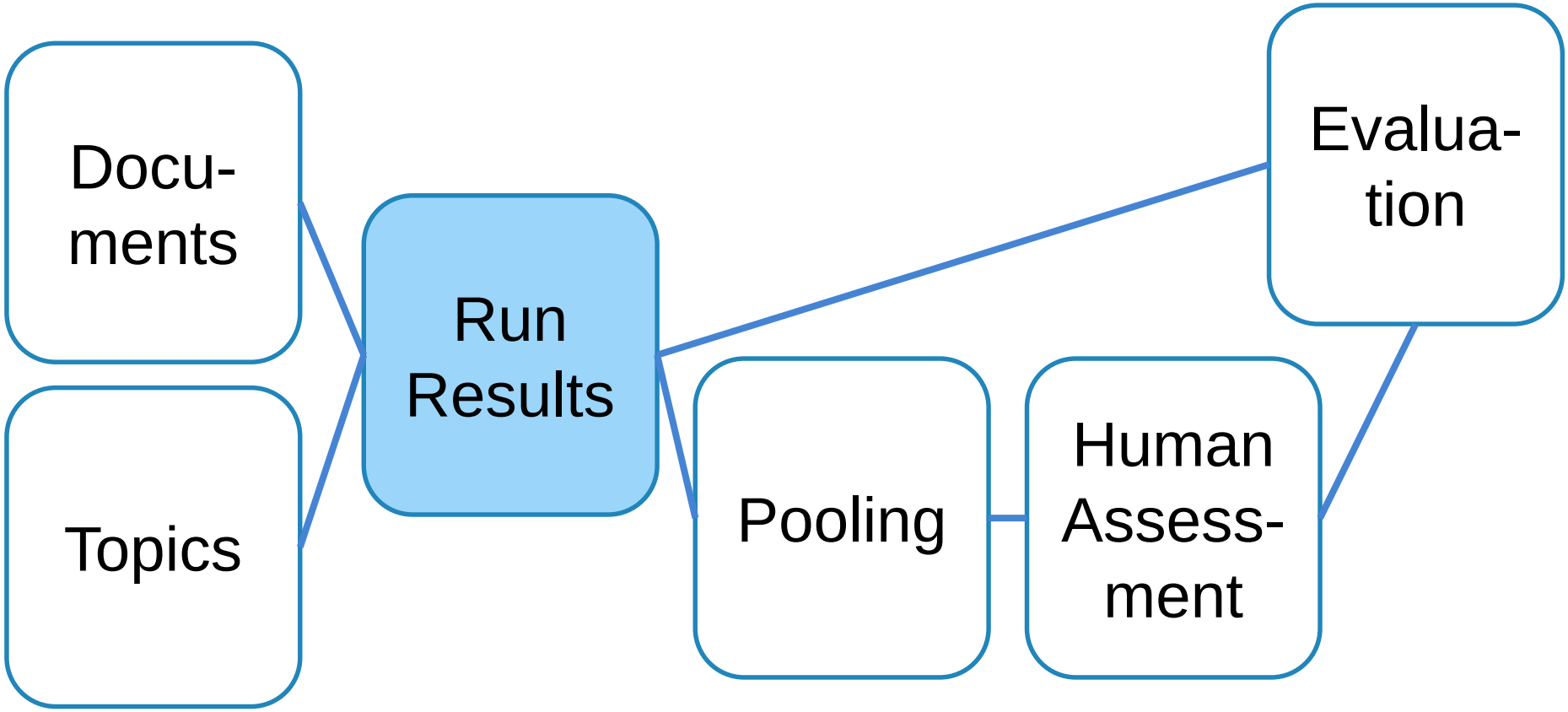
$f(x) = x$ $f(*) = *$

unification

f(x)=y
Is not a “fixed point” !

Keyword: Mo
Keyword: Möbius
Keyword: automorphism

Math-2 Task Outline



Submission

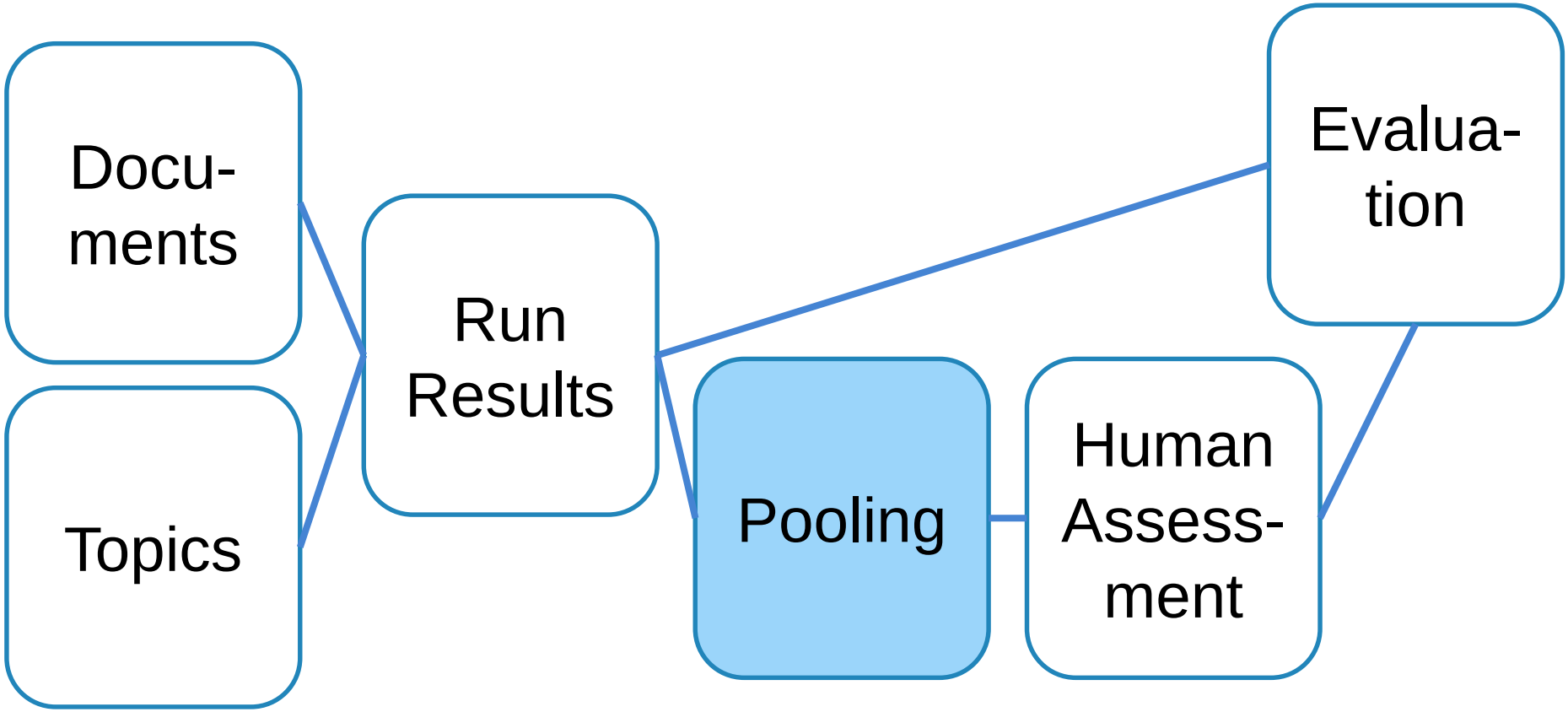
#	Group ID	Organization in English	from
1	ICST	Peking University	CN
2	IFISB	TU Braunschweig	DE
3	FSE	TU Berlin	DE
4	KWARC	Jacobs University Bremen	DE
5	MCAT	National Inst. of Informatics	JP
6	MIRMU	Masaryk University	CZ
7	RIT	Rochester Inst. of Technology	US
8	TUW-IMP	Vienna Univ. of Technology	AT

20 submitted runs

1 run X 4 groups, 4 runs x 4 groups

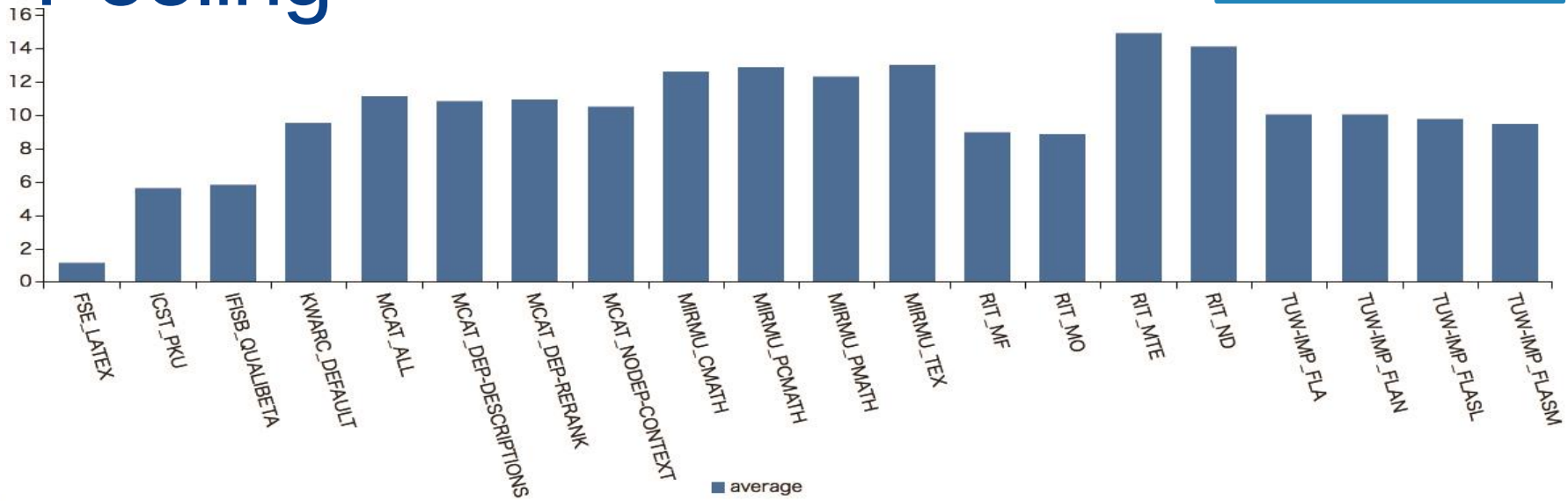
NTCIR

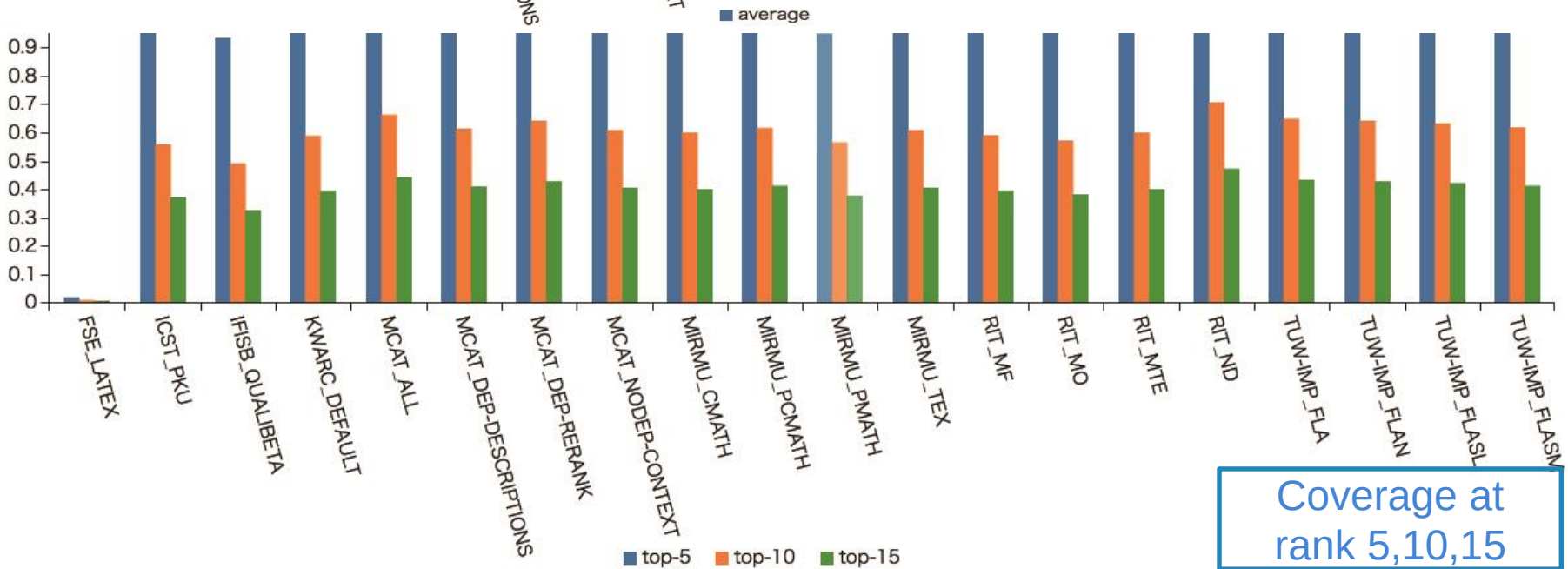
Math-2 Task Outline



Pooling

Assessment Allocation

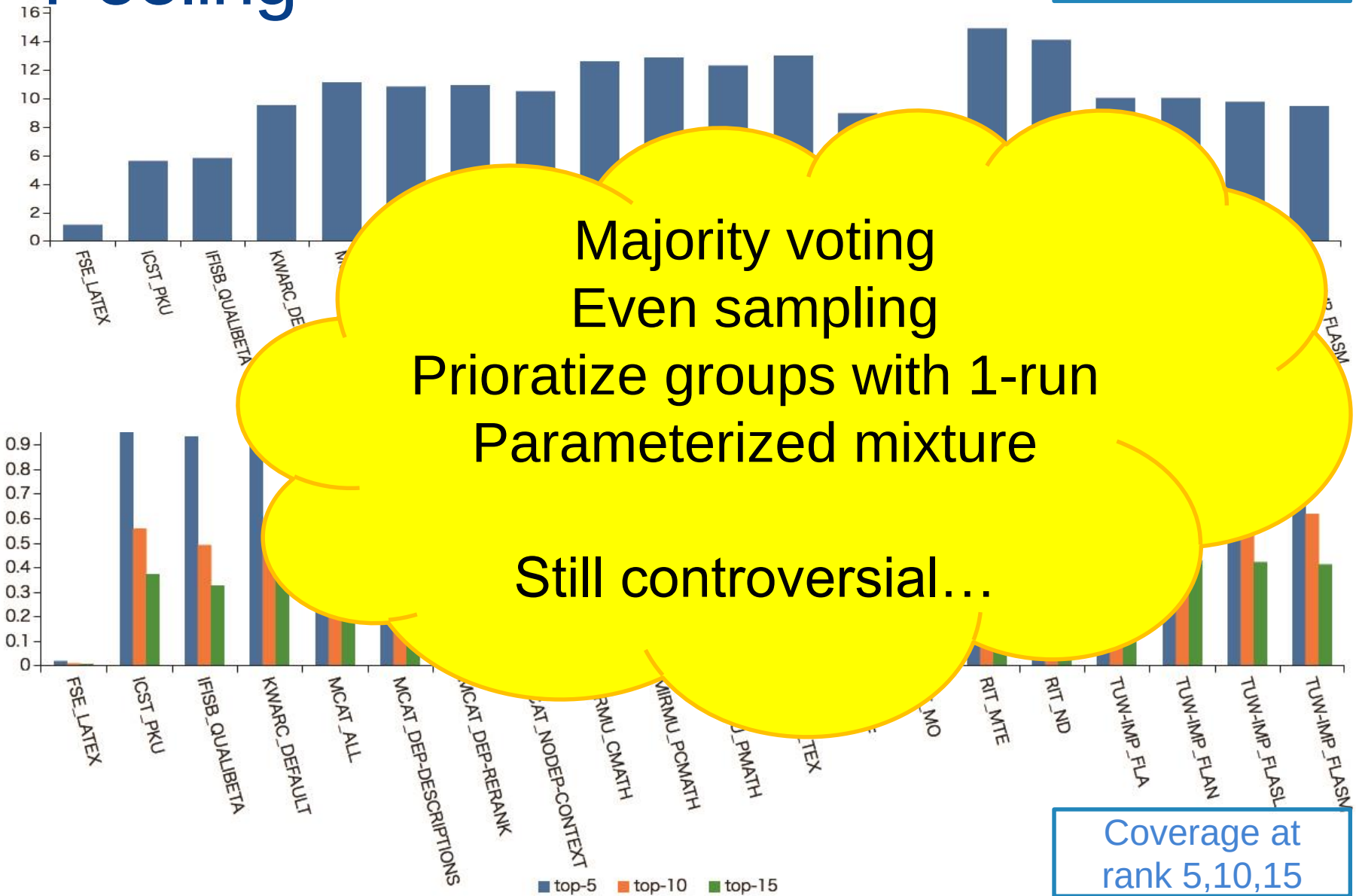




Coverage at rank 5,10,15

Pooling

Assessment Allocation

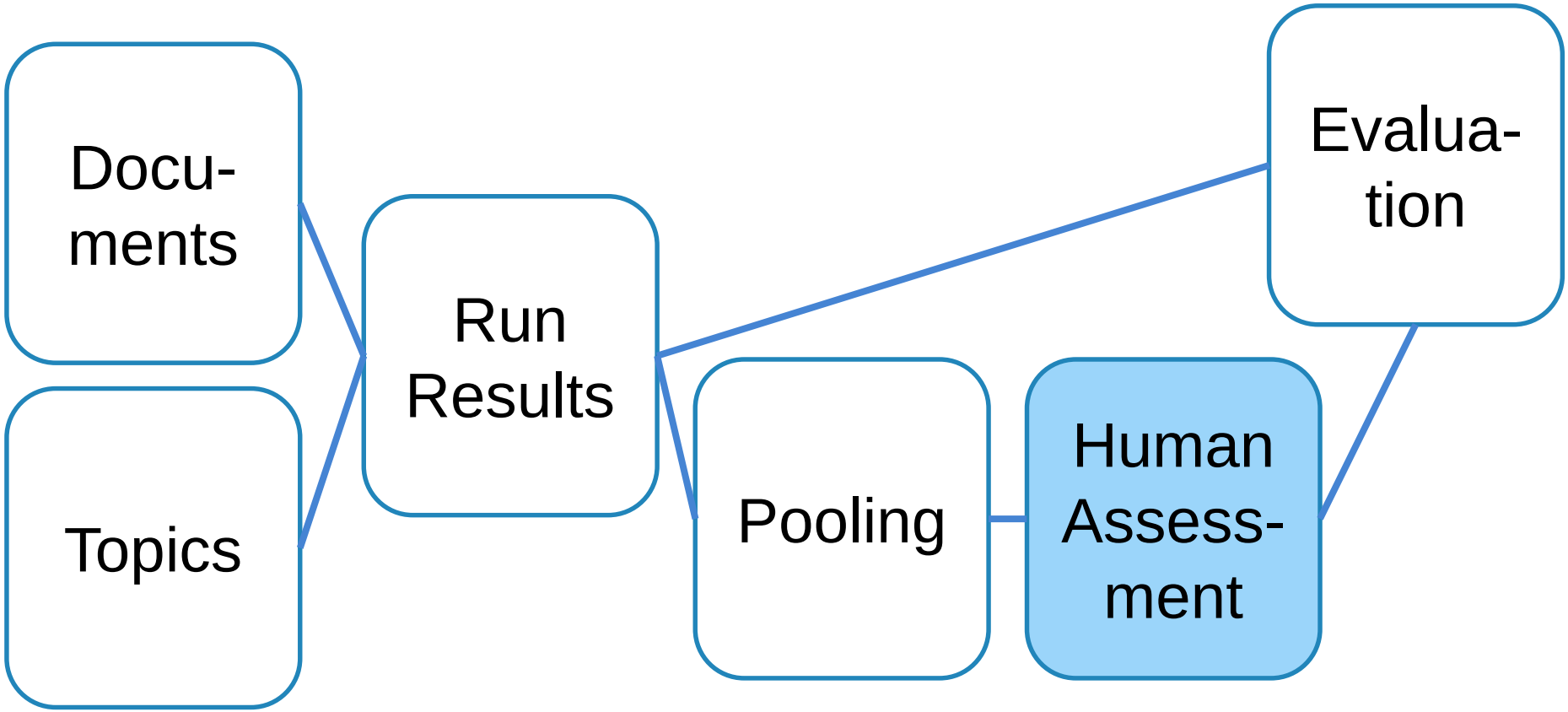


Majority voting
Even sampling
Prioratize groups with 1-run
Parameterized mixture

Still controversial…

Coverage at rank 5,10,15

Math-2 Task Outline



Human Assessment

- 50 retrieval units per topic, 2 assessments per retrieval unit
- Relevant, Partially relevant, Non relevant
- Use SEPIA with Math Extension

Agreement score for the two judges	Relevant	Partially relevant
Fleiss Kappa Agreement	0.5442	0.5783
Pearson Correlation value	0.5481	0.5914

SEPIA Snapshot

Home > Task Home **Task**: NTCIR11 Math Task (EN), **Username**: admin Log out

Select a pool and then topic and you will see a list of potentially relevant documents to judge. For each document, judge relevance to the topic.

Topic: NTCIR11-Math-1

Pool: pool

Topic Details

Question (TrgLang)
Information Need (TrgLang)

$square(phi) = id$ (id=f1.0)
$phi \neq id$ (id=f1.1)
hyperelliptic surface (id=w1.0)
Riemann (id=w1.1)

Memo

title	Forgotten relevant concept name in a definition I'm looking for
relevance	I am looking for a definition of hyperelliptic Riemann surfaces that references an involution (defined by formulae of the form $\phi \neq id$ $\phi^2 = id$
examplehit	http://arxiv.org/pdf/1008.2233.pdf

Answer Type

Document List:

[Not-Relevant] [x] math991119
[Partially-Relevant] [x] hep-th0
[Not-Relevant] [x] math060113
[x] math0512368_1_53
[x] 1207.0190_1_174
[x] nlin0701038_1_13
[x] 1304.6518_1_8
[x] 1304.6518_1_7
[x] 1311.1931_1_65
[x] hep-th0103181_1_1
[x] math0109124_1_453
[x] math0110242_1_9
[x] 0803.0289_1_99
[x] math9912016_1_182
[x] 1211.2089_1_270
[x] 1307.8227_1_96
[x] math0410244_1_18

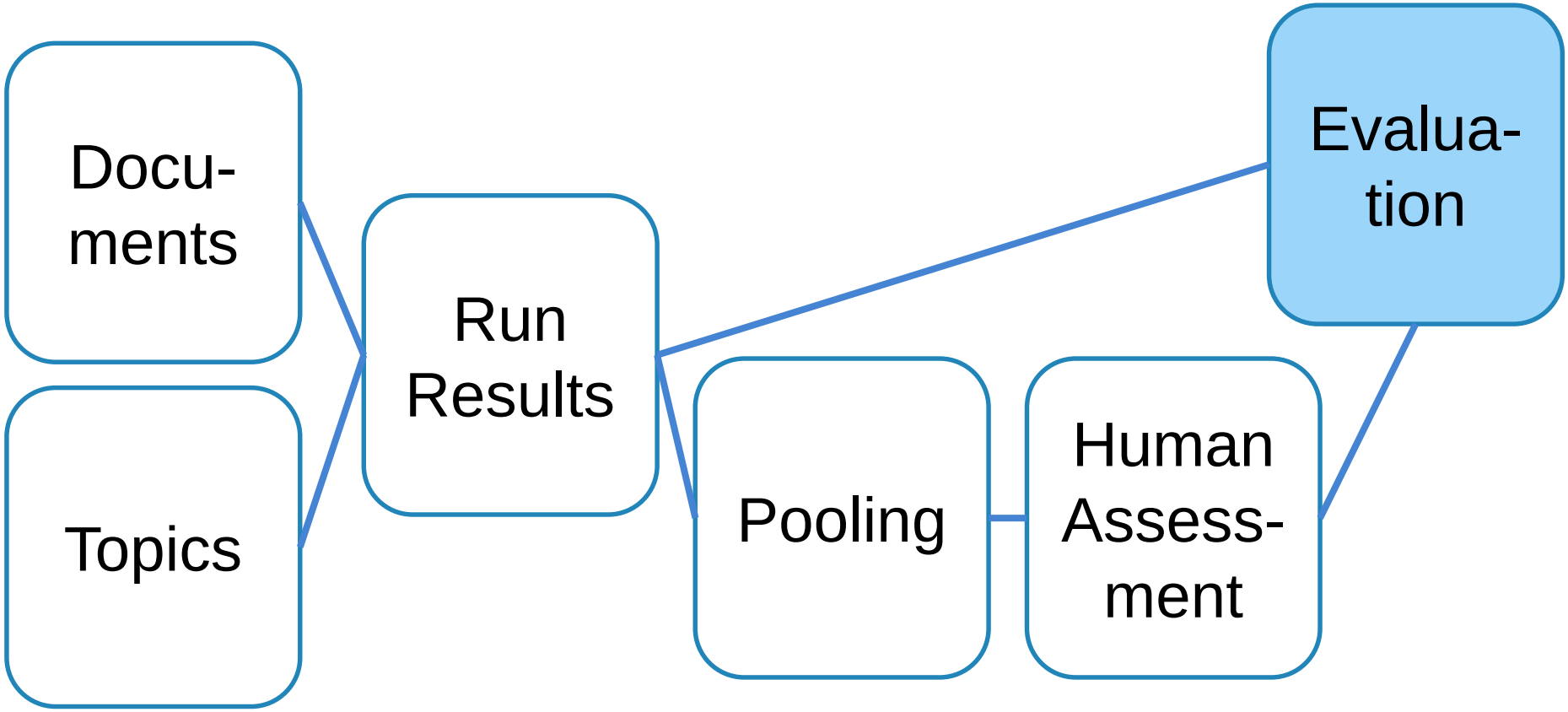
Relevance Judgment toggle

<DOC>
<DOCNO>hep-th0508030_1_43</DOCNO>
<URL>https://archive.org/details/arxiv-hep-th0508030</URL>
<CONTEXT>

- Proof: Let us first define several (quasi-)elliptic differential operators. Recall that $\mathcal{G} = -\mathcal{J}\mathcal{J}$ defines a positive-definite metric on $T \oplus T^*$, and it induces a natural Hermitian inner product on $\Omega^\bullet(M, \mathbb{C})$ (therefore also on $C^\infty(\wedge^\bullet E^*)$) that generalizes the familiar Hodge inner product in ordinary Riemannian geometry [9]. Let us denote this Hermitian inner product by (α, β) for $\alpha, \beta \in \Omega^\bullet(M, \mathbb{C})$. Let $\bar{\partial}_H, \partial_H$ be the generalized Dolbeault operators associated

NTCIR

Math-2 Task Outline



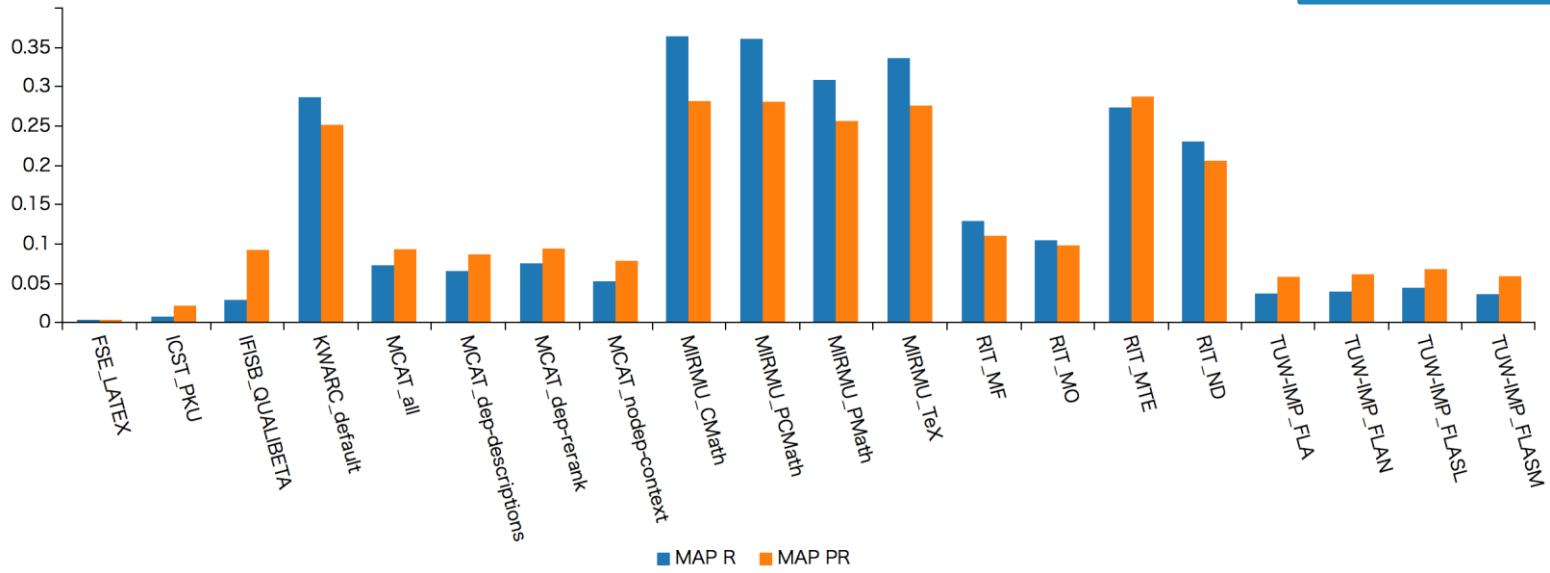
Performance Measures

trec-eval tool was used for the evaluation

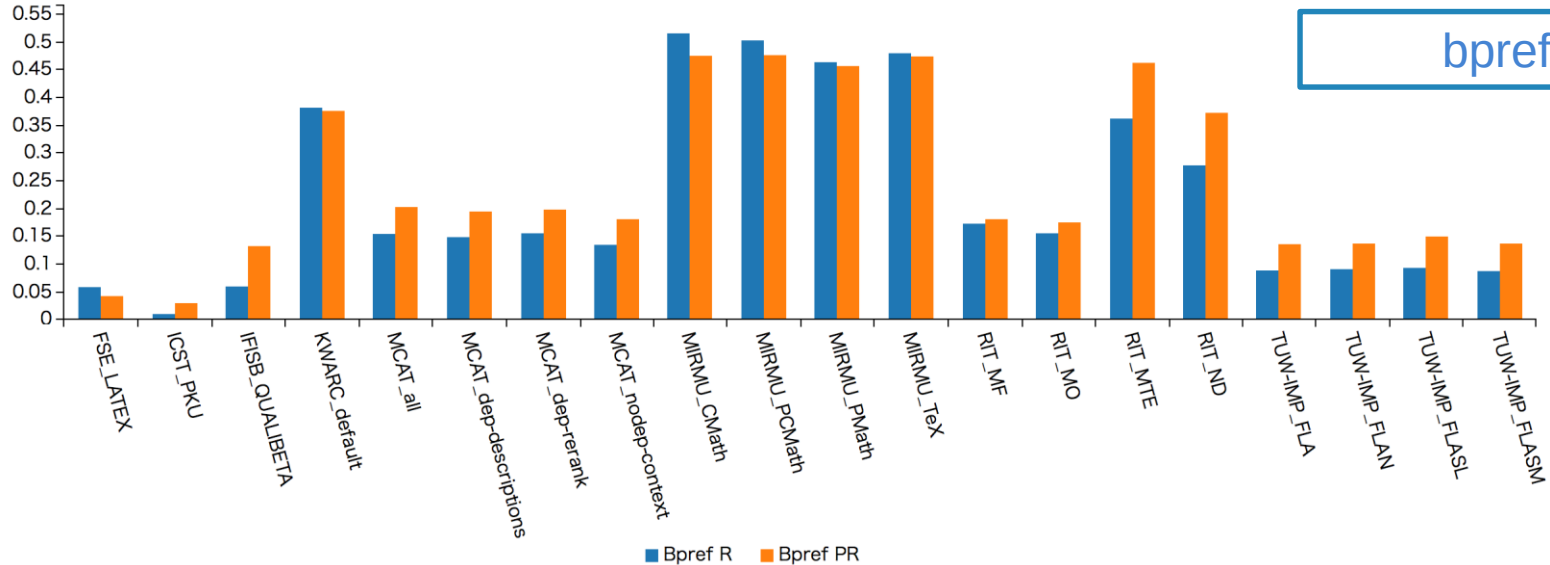
- MAP: Mean average precision over judgment groups
- Bpref: Preference-based information retrieval measure for incomplete relevance judgment
- P-5: Precision at rank 5
- P-10: Precision at rank 10

Performance Summary

MAP



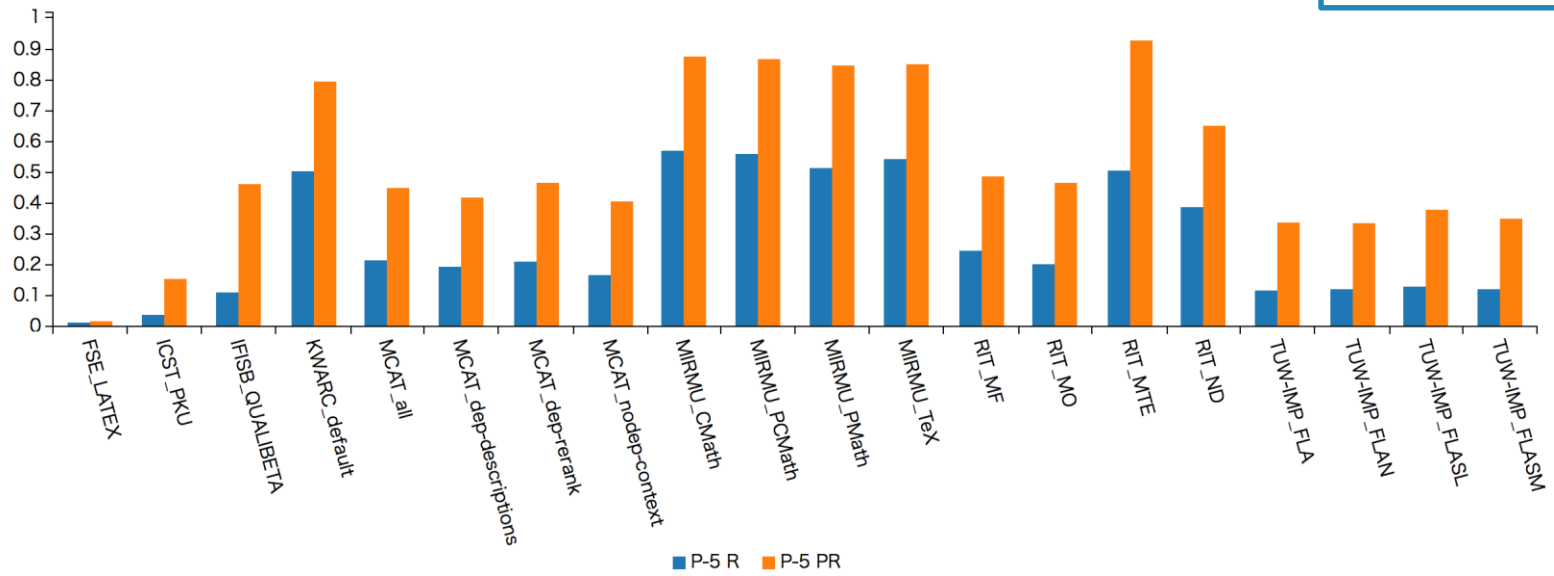
bpref



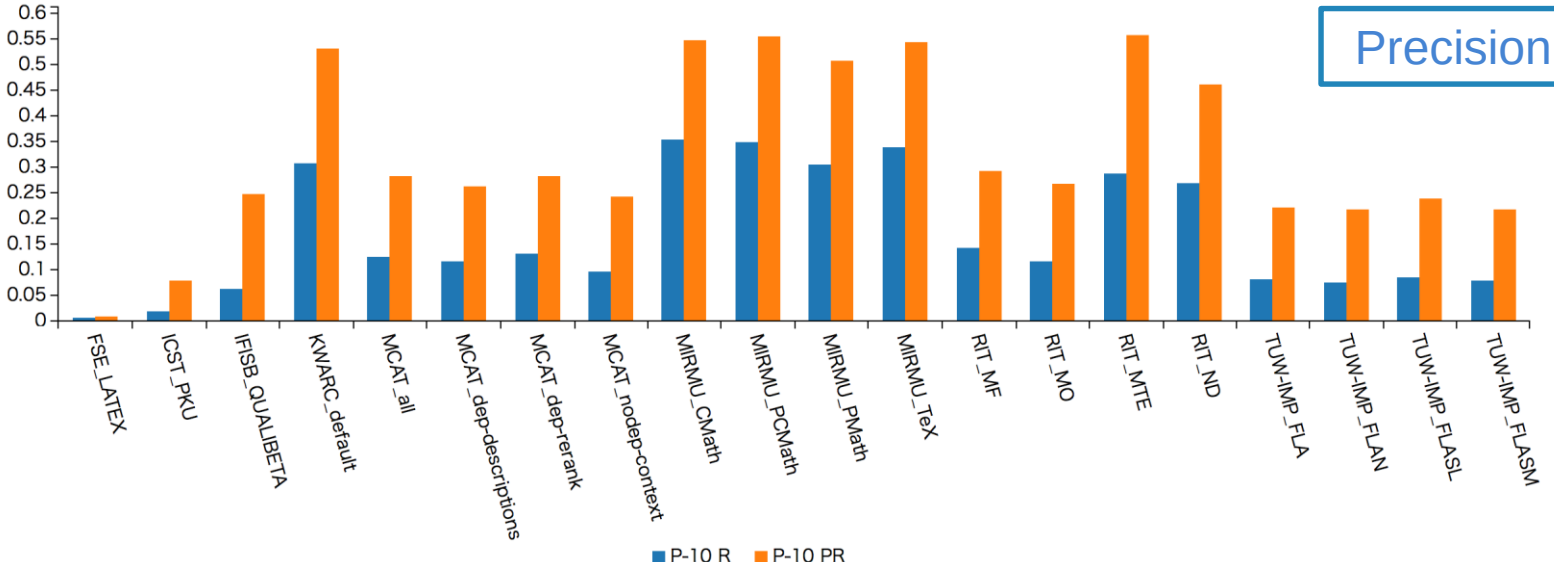
NTCIR

Performance Summary

Precision @ 5



Precision @ 10



NTCIR

Summary of the Submitted Runs

Table 7: Summary

RunID	keywords	math formulae	LaTeX	query variables	search engine
FSE_{latex}	yes	yes	yes	yes	no
$ICST_{pku}$	no	yes	no	no	yes
$IFIS_{QUALIBETA}$	yes	yes	n	yes	yes
$KWARC_{default}$	yes	yes		yes	yes
MCAT	yes	yes		no	yes
$MIRMU_{cmath}$	yes	yes		no[1]	no[2]
$MIRMU_{pcmath}$			no		
$MIRMU_{pmath}$			no		
$MIRMU_{tex}$			yes		
RIT_{mf}	yes	yes	no	yes	yes[3]
RIT_{mo}	no				no
RIT_{mte}	yes				yes[3]
RIT_{nd}	yes				yes[3]
TUW-IMP	yes	yes	no	yes	yes

*1 Not explicitly; but qvar names were shortened t
*2 MIaS extension was used to general search engi
*3 Solr was used just for the text index; a custom n expressions.

All runs used math formulae in the topics

Some runs did not use keywords

NTCIR

Summary of the Submitted Runs

Table 7: Summary of configuration of partici

RunID	keywords	math formulae	format		
			LaTeX	Presentation	Content
FSE_{latex}	yes	yes	yes	yes	yes
$ICST_{pku}$	no	yes	no	yes	no
$IFIS_{QUALIBETA}$	yes	yes	no	no	yes
$KWARC_{default}$	yes	yes	no	no	yes
MCAT	yes	yes	no	yes	no
$MIRMU_{cmath}$	yes	yes	no	no	yes
$MIRMU_{pcmath}$			no	yes	yes
$MIRMU_{pmath}$			no	yes	no
$MIRMU_{tex}$			yes	no	no
RIT_{mf}	yes	yes	no	yes	no
RIT_{mo}	no				
RIT_{mte}	yes				
RIT_{nd}	yes				
TUW-IMP	yes	yes	no	no	yes

*1 Not explicitly; but qvar names were shortened to the unique one letter var
*2 MIaS extension was used to general search engine Lucene.
*3 Solr was used just for the text index; a custom mysql based index was use

The three math representations (LaTeX, MathML Presentation, MathML Content) were all used in the task

NTCIR

Summary of the Submitted Runs

on of participating systems.

RunID	k	Content	tree structure	query variables	search engine
FSE_{latex}		yes	yes	yes	no
$ICST_{pku}$		no	yes	no	yes
$IFIS_{QUALIBETA}$		yes	no	yes	yes
$KWARC_{default}$			yes	yes	yes
MCAT			yes	no	yes
$MIRMU_{cmath}$		yes	yes	no[1]	no[2]
$MIRMU_{pcmath}$		yes			
$MIRMU_{pmath}$		no			
$MIRMU_{tex}$		no			
RIT_{mf}		no	yes	yes	yes[3]
RIT_{mo}					no
RIT_{mte}					yes[3]
RIT_{nd}					yes[3]
TUW-IMP		yes	yes	yes	yes

*1 Not explicitly; but one letter variables.
*2 MIaS extension wa
*3 Solr was used just ndex was used solely for math expressions.

The configurations of the search systems varied

- tree-structure / token-based search?
-consideration of query variables?
-general / specific purpose search engine?

Math-2 Summary

- Large-scale tree structure search
  - Experienced teams implemented online search systems for 60 millions formulae.
- Diverse approaches: all types of Math representations were used in the task
  - (1) LaTeX
  - (2) MathML Presentation Markup (automatically converted from (1))
  - (3) MathML Content Markup (automatically converted from (2))
- Importance of query variables
  - How to handle variables in math formulae was a key issue
  - Deciding which symbols should be considered as variables is left for future investigation

Other Developments in the Task

- docs2harvest: Tool for parsing html / xhtml documents and generate harvest files with the Content Math data only.
  https://github.com/KWARC/mws
- LaTeXML: A LaTeX to MathML converter.
  http://dlmf.nist.gov/LaTeXML/
- PlaneText: converting XML-tagged text into plain text sequences which can be directly input to NLP tools
  http://kmcs.nii.ac.jp/planetext/
- MathML extension to SEPIA
- Math-2 topic and submission format specifications
- Math-2 submission validation script
- Reference to other useful tools: Mathjax, mathml-converter, etc.

WIKIPEDIA OPEN SUBTASK

Wikipedia Open Subtask

- Size and complexity of the main task pose challenges for first-time participants

- Create a tool to soften the learning curve!

Wikipedia Open Subtask

- Size and complexity of
  challenges f
- Create a

Task Compatibility

All the participants to the main task can make a submission with minimum efforts

Wikipedia Open Subtask

smaller corpus + automated evaluation

- All math pages from Wikipedia
  - uniform language, K-18 math
- 100 formulae queries (no keywords) in NTCIR-11 syntax
  - generated from Wikipedia pages
- Evaluation
  - HIT iff found original, NoHIT if not
  - (≠ relevance)

Wikipedia Open Subtask

Experience gained

- 7 teams from 5 countries uploaded 115 runs with 2M hits
  - indeed used for debugging
- Performance example
  - run 111: 94% (top 1000) 85% (top 10), etc.
- Consecutive runs improved performance

- “Development” subtask

SUMMARY

Achievements and Todo's

- HTML5 data set with 8M paragraphs from 105 K papers
- 22 runs submitted by 8 teams, 5000 hits pooled and evaluated
- Solid methodology for dataset/topic creation, good inter-reviewer agreement

- Reusable test collection
- Enhanced task environment
- Participation from both Math and IR sides

- Pooling/evaluation measures needs work
- Probably the largest (non-generated) collection of (formula) trees in existence --> come and try your methods.

- Next task (Math-3) ?

Conclusion

- Math is the queen and language of the sciences and technology
- Regular IR techniques are of limited use for Math IR
- Small but vibrant community who studies Math IR

multidisciplinary endeavor

Acknowledgement

- The work reported here here has been partially supported by the Leibniz association under
- grant SAW-2012-FIZ¥_KA-2 and JSPS KAKENHI Grant Numbers 2430062, 25245084, and 14J09896.

- We are indebted Deyan Ginev for generating the dataset for the Math Retrieval Subtask and
- to Deyan Ginev, Radu Hambasan, Fulya Horozal, Mihnea Iancu, Constantin Jucovschi, Andrea Kohlhase, Sentao Miao, Mihai Munteanu, Corneliu Prodescu, Florian Rabe, Jurgen Schefter, Wolfram Sperber, and Adelin Travers for providing formula search queries.

- We gratefully We gratefully acknowledge the contribution of Goran Topic in administrating a backend system for the entire task and many useful suggestions to the task, and Giovanni Yoko Kristianto, Shunsuke Ohashi, Kazuhiro Hattori and TriAx co. for assisting the task organization.

Thank you for the great supports from NTCIR-Committees, Chairs, and Secretaries!

NTCIR