The Tangent Math Search Engine at NTCIR 2014

Nidhin Pattaniyil (npatta01@gmail.com) and Richard Zanibbi (rlaz@cs.rit.edu)
Document and Pattern Recognition Lab, Rochester Institute of Technology, USA

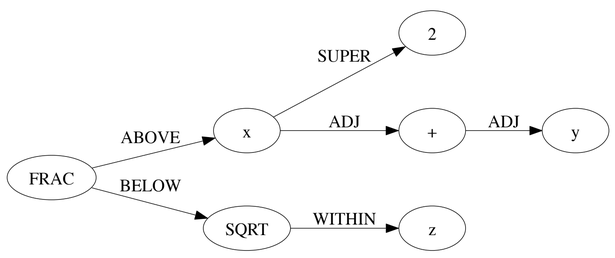
Math Search and Tangent

It may be useful to search technical documents using formulas as well as keywords [5]. In a previous study math experts could not identify uses for formula queries [7], but non-experts have identified a number of them, e.g. interpreting unfamiliar notation [4]. *Text-based* and *tree-based* techniques for formula search have been developed [3,6].

Design: We have extended the Tangent formula search engine [3] to include support for matrices/tabular layouts, prefix sub/superscripts, wildcard variables, and text search integration (Lucene/Solr). **Formula Inverted Index** [8]: defined over name and *relative* position of symbol pairs, and additional tuples for matrix structure. Maps tuples to expressions/documents containing them. **Text Index:** Modified Lucene index with formulae 'text' replaced by identifiers to represent formulae locations only (TF-IDF based). **Final Ranking:** The most similar formula is used for the document formula score. Formula and text search engines scores are combined using: $\alpha \cdot textScore(d) + (1-\alpha)\ formulaScore(d)$. (Note: formula lists are supported)

Formula Structure Representation

$$\frac{x^2+y}{\sqrt{z}}$$



(a) Formula and Symbol Layout Tree

Parent	Child	Dist.	Vert.
FRAC	x	1	1
FRAC	2	2	2
FRAC	+	3	1
FRAC	y	3	1
FRAC	SQRT	1	-1
FRAC	z	2	-1
x	2	1	1
2	**None**	**0**	**0**
x	+	1	0
x	y	2	0
+	y	1	0
y	**None**	**0**	**0**
SQRT	z	1	0
z	**None**	**0**	**0**

(b) Symbol Pair Tuples

Fig. 1. Quartuples are defined for every descendant of a symbol in a symbol layout tree. Symbols without children have child 'None.' In (b), *Dist.* is the path length from the parent to child symbol in the layout tree, and *Vert.* is a sum of vertical displacements along this path: +1 for each superscript/above edge, -1 for each subscript/below edge, and 0 for each horizontally adjacent or within edge

NTCIR-11 Math-2 Retrieval Tasks [1]

Main Task: 50 formula and keyword queries for 100,000 technical articles (from www.arxiv.org) broken into fragments ranging from a couple words to multiple paragraphs. The 8,301,578 document fragments contain 39,008,971 unique formulae.

Wikipedia Subtask: 100 formula queries for approximately 35,000 articles from English Wikipedia containing 387,947 unique LaTeX expressions.

Formula Query: $\mathbb{P}(\boxed{X} \geq \boxed{t}) \leq \frac{\mathbf{E}(\boxed{X})}{\boxed{t}}$

Keyword: Markov inequality

a) Math-2 Task Query #39

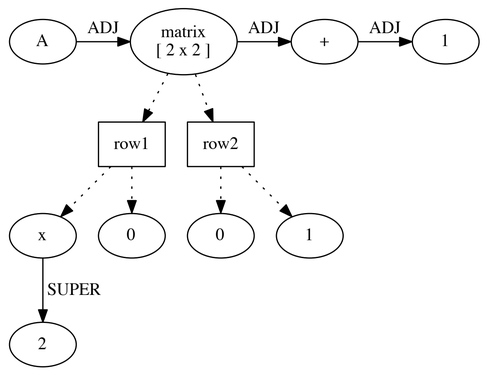
$$\mu(A) = \begin{cases} 1 & \text{if } 0 \in A \\ 0 & \text{if } 0 \notin A. \end{cases}$$

b) Wikipedia Subtask Query #49

Fig. 2. Sample queries. Wildcard variables are shown as red symbols in boxes. We converted queries from a Presentation MathML (XML) representation to symbol pair tuple sets (see below)

Matrix Representation

$$\mathrm{A}\begin{bmatrix} x^2 & 0 \\ 0 & 1 \end{bmatrix} + 1$$



(a) Formula and Symbol Layout Tree

Matrix Structure			
Parent	**Child**	**Row**	**Column**
matrix	dimensions	2	2
matrix	'x^2'	1	1
matrix	'0'	1	2
matrix	'0'	2	1
matrix	'1'	2	2
Subexpressions			
Parent	**Child**	**Dist.**	**Vert.**
A	**matrix2x2**	1	0
A	+	2	0
A	1	3	0
matrix2x2	+	1	0
matrix2x2	1	2	0
+	1	1	0
1	None	0	0
x	2	1	1
2	None	0	0
0	None	0	0
0	None	0	0
1	None	0	0

(b) Tuples

Fig. 3. At the topmost level of the expression, matrices are treated as a single symbol (e.g. **'matrix2x2'**). This topmost expression along with all subexpressions in matrix cells are represented as at left. Additional tuples are used to represent matrix dimensions, and the contents of matrix cells (represented as 'Child' symbols)

Wildcards and Formula Retrieval

Wildcard Tuples: Two additional indices group tuples with common parent or child symbols. For example, the tuple $(?i, 2, 1, 1)$ refers to symbols with a superscript 2 (e.g. x^2, n^2, $)^2$), and tuple $(x, ?i, 1, 1)$ refers to any superscript of an x (e.g. x^2, x^3, $x^($). **Wildcard-wildcard relationships are not indexed.**

Formula Retrieval: 1) Look up query formula tuples in regular and wildcard indices to retrieve expressions. 2) Sort by match count, keep top $k = 1000$. 3) Greedy wildcard matching: iteratively select wildcard/symbol unification matching the most unmatched query tuples. 4) Score by F-measure, $F = 2RP/(R + P)$, where R and P are the number of matched query and candidate pairs, respectively.

Retrieval Results

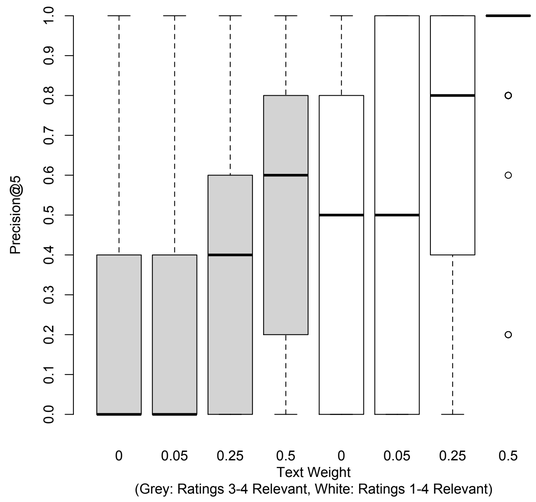


Fig. 4. Tangent Precision@5 (Main Task) for 50 queries combining one or more formulas with keywords, for different text vs. formula score weightings

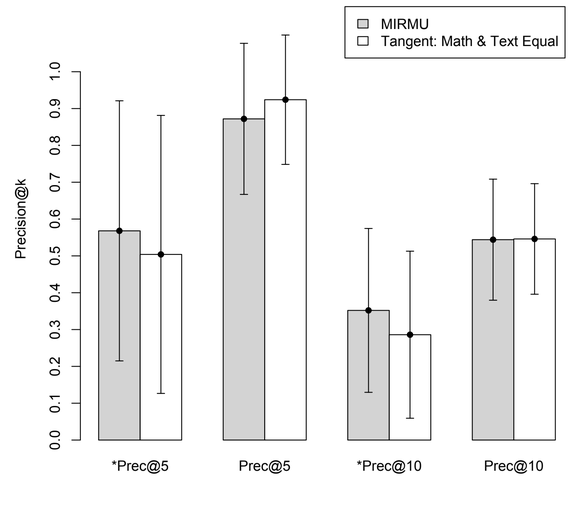


Fig. 5 MIRMU [2] System vs. Tangent (Main Task). *Prec@ indicates precision for high-relevance hits (rated 3-4); Prec@ for hits rated higher than 0

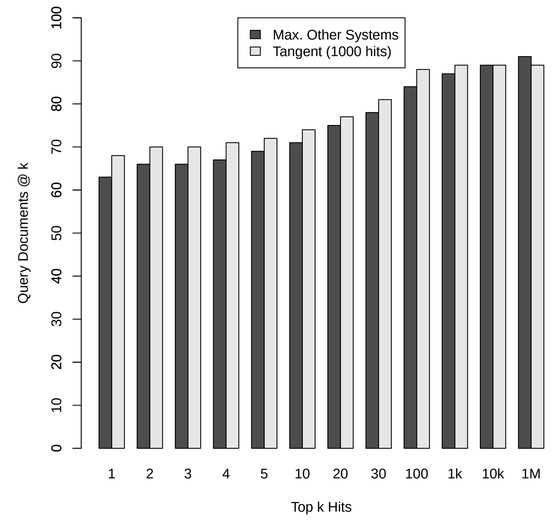


Fig. 6. Wikipedia Subtask Results (100 formula queries). 'Query Documents @k' is a specific-item recall measure, giving the percentage of articles from which queries are taken in the first k hits

Implementation and System Performance

We used the Amazon EC2 web service: a memory-optimized configuration (r3.4xlarge) with 16 vCPUs, 2.5 GHz, Intel Xeon E5-2670v2, 122 GB memory, and 1 x 320 GB Disk.

Main task: Nine EC2 instances were used to index formulas in the collection, one instance for Solr/Lucene, and one instance to parse queries and access the text and formula engines (Python-based). **Wikipedia subtask:** A single machine was sufficient for indexing and retrieval.

Table 1. MySQL database table sizes for formula indices. For the main task 81,774,641 symbol pairs are defined across nine indices (with repetitions)

Table	Rows	Size(MB)	Idx(MB)
arXiv (main)	**Shown: 1 of 9 Indices**		
symbol pairs	14,791,465	2600	692
expression-docs	5,927,284	183	147
expression	5,636,077	313	78
symbol-ids	195,960	6	10
Wikipedia	**Shown: Complete Index**		
symbol pairs	3,002,881	305	141
expression-docs	387,975	12	9
expression	387,947	775	6
symbol	56,437	2	3

Table 2. Indexing & retrieval times for formula retrieval. Search times shown are for 50 main task queries, and 100 Wikipedia subtask queries.

Collection	Time (minutes) Index	Search
NTCIR-main (arXiv)	$420 \times 9 \approx 3380$	150
Wikipedia	33	8

Notes: wildcard support increased retrieval time slightly; missing symbol name synonymns (e.g. TEX vs. unicode for '>'); database (MySQL) organization for symbol pairs can be compressed/reorganized.

References

[1] Aizawa, A., Kohlhase, M., Ounis, I., and Schubotz, M. NTCIR-11 Math-2 Task overview. In *Proc. NTCIR-11* (2014).

[2] Michal Růžička, P. S., and Líška, M. Math indexer and searcher under the hood: History and development of a winning strategy. In *Proc. NTCIR-11* (2014).

[3] Stalnaker, D., and Zanibbi, R. Math expression retrieval using an inverted index over symbol pairs. In *Proc. DRR XXII* (2015).

[4] Wangari, K., and Zanibbi, R. Discovering real-world usage scenarios for a multimodal math search interface. In *ACM SIGIR* (2014), pp. 947–950.

[5] Youssef, A. Roles of math search in mathematics. In *Proc. MKM*, no. 4108 in LNAI. Springer, 2006, pp. 2–16.

[6] Zanibbi, R., and Blostein, D. Recognition and retrieval of mathematical expressions. *IJDAR 15*, 4 (2012), 331–357.

[7] Zhao, J., Kan, M., and Theng, Y. Math information retrieval: user requirements and prototype implementation. *Proc. J. Conf. Digital Libraries* (2008), 187–196.

[8] Zobel, J., and Moffat, A. Inverted files for text search engines. *ACM Computing Surveys 38*, 2 (2006), 56 pp.

Acknowledgements

We thank David Stalnaker, Frank Tompa, Akiko Aizawa and Moritz Schubotz. This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-1016815. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Links: Source & Demo

Code: cs.rit.edu/~dprl/Software.html
Demo: saskatoon.cs.rit.edu/tangent
Lab: cs.rit.edu/~dprl