TANGENT-3 AT THE NTCIR-12 MATHIR TASK

KENNY DAVILA,[1] RICHARD ZANIBBI,[1] ANDREW KANE,[2] AND FRANK WM. TOMPA[2]

[1]ROCHESTER INSTITUTE OF TECHNOLOGY, USA {KXD7282,RLAZ}@CS.RIT.EDU [2]UNIVERSITY OF WATERLOO, CANADA {ARKANE,FWTOMPA}@UWATERLOO.CA

MATH. INFORMATION RETRIEVAL WITH TANGENT-3

Mathematical Information Retrieval (MIR [1,4,5]) is concerned with finding information on mathematical topics, using a combination of keywords and formulae. Information needs for MIR differ with users' mathematical expertise [1,4,5], e.g., queries to define unfamiliar notation, vs. queries for properties of mathematical objects.

The **Tangent-3 math-aware search engine** [2,3,6] processes queries as in the following.

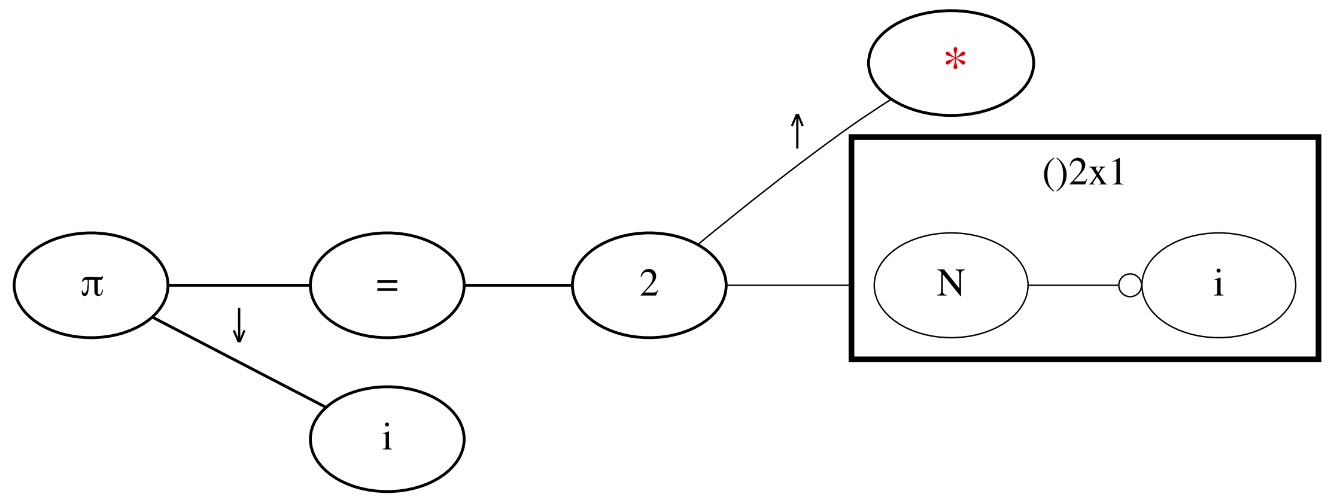
1. **Text** (T) retrieved using Solr
2. **Formulae** (F) retrieved via symbol pairs and their spatial relationships. Matching formulae ranked by approx. Dice coefficient of symbol pair matches: $2RP/(R+P)$
   - Best formula match used to score each document for a formula query; for multiple query formulae use a linear combination of best match scores
   - Optional **re-ranking** of top-k (for NTCIR-12, $k = 1000$)
3. **Final score ($\alpha T + (1-\alpha)F$):** linear combination of Text and Formulae scores

Parameters Explored

1. Text vs. Formula score weighting (α, uniform vs. proportional to query tokens)
2. Multiple query formula weighting (uniform vs. size-proportional)
3. Formula hit re-ranking
4. Wildcard matching (symbol vs. subexpression), Unification (none vs. num + id)

FORMULA STRUCTURE REPRESENTATION

$$\pi_i = 2^{*}\binom{N}{i}$$



(a) Formula and Symbol Layout Tree

SYM-1	SYM-2	PATH	COUNT
V!π	**V!**i	$\downarrow$	1
V!π	=	$\rightarrow$	1
=	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\boxed{\cdot}$	1
V!N	**V!**i	$\multimap$	1
V!π	**N!**2	$\rightarrow\rightarrow$	1
=	*	$\rightarrow\uparrow$	1
⋮	⋮	⋮	⋮
V!π	**V!**i	$\rightarrow\rightarrow\rightarrow\boxed{\cdot}\multimap$	1

(b) Symbol Pair Tuples

Fig. 1. Indexing a Symbol Layout Tree (SLT) obtained from Presentation MathML. (b) shows SLT symbol pairs at different depths with corresponding counts. For SLTs with tree height $\leq$ 2 symbols at the end of writing lines are also indexed (e.g., 2, N, and i). **Formula index sizes: Wiki 580.5 MB, arXiv 8.3 GB on disk.**

NTCIR-12 MATHIR TASKS [4]

Tangent was used in three of the four MathIR Tasks.

arXiv Main Task. 29 formula and keyword queries for 100,000 technical articles (from www.arxiv.org) broken into fragments ranging from two words to multiple paragraphs. The 8,301,578 document fragments contain 39,008,971 unique formulae.

Wikipedia Main Task. 30 queries containing keywords and math expressions for 30,000 English Wikipedia articles containing more than 500,000 formulas.

Wikipedia Formula Browsing Task. 40 queries containing isolated formulae. The first 20 are concrete (without wildcards), while the remaining 20 are constructed by deleting or replacing subexpressions with wildcards in the concrete queries.

FORMULA MATCHING USED IN RE-RANKING

QUERY

$$x^2 + y^2 = *$$

MATCH

$$\alpha^2+\beta^2=\gamma^2$$

(a) Query match with identical symbols (blue), wildcard match (red), and unification (orange). Numbers and identifiers are unified.

Case	Query	Match
Unrestricted	$x + *$ e^{*}	$x+1$ $x+y+z+sin(x)$ $y+x+z=\frac{\pi}{4}$ $f(x)=e^{x+1}+2$
Children	$*^2+1$	x^2+y^2+1 x^2+y+1 $x^2+(y+z)^2+1$
Binding	$*1*^2+*1*+1$	x^2+x+1 $(x+1)^2+(x+1)+1$ x^2+y+1
Fill right	$x+*+1$	$x+y+1$ $x+y+z+1$ $x+y-z+1$ $x+\frac{1}{2+y}-3z+1$
Fill left	$*+1$	$x+y+z+1$ $\alpha=f(x+y+1,x^2)$ $f(x,y)=\frac{1}{x+y+1}$

(b) Wildcard expansion. Wildcards are matched after identical symbols and relationships are found, using the cases above.

Fig. 2. Formula Matching with Wildcard Expansion and Unification. For re-ranking, a greedy algorithm locates the best matching subexpression (i.e., connected component) on a candidate formula.

SIMILARITY METRICS

D - Approximated Dice Coefficient. Global Dice coefficient for matching symbol pairs between expressions; wildcards match individual symbols. ***Produces Top-1000 hits for re-ranking.**
Wildcards: single symbols, *Unification:* none

D + DS - Dice Coefficient for Best Matching Subexpression. Rerank by *local* Dice coefficient for best matching subexpression (connected component-based), wildcards match subexpressions.
Wildcards: subexpressions, *Unification:* none

D + DSU - Dice Coefficient with Unification. Rerank per *DS*, but with symbol unification, scoring unified matches lower than exact matches.
Wildcards: subexpressions, *Unification:* num + id

D + MSU - Maximum Subtree Similarity (MSS) [6]. Rerank by harmonic mean of query symbol and relationship matches; penalize unmatched symbols, then prefer identical symbols.
Wildcards: subexpressions, *Unification:* num + id

RESULTS

Table 1. Wikipedia Formula Browsing Task Results. Avg. Precision@K shown for Top-20 hits provided. Each formula hit rated by two students (MSc + ugrad). *Re-Rank Upper Bound:* P@k results from sorting initial Top-1000 hits (ranked by *D*) in decreasing order of rating.

		RELEVANT				PARTIALLY RELEVANT			
QUERIES (40)	SUBMISSION	P@5	P@10	P@15	P@20	P@5	P@10	P@15	P@20
CONCRETE (20)	Run-1, *D*	0.4800	0.3550	0.2900	0.2375	**0.9400**	**0.8850**	**0.8267**	**0.7950**
	Run-2, $D + DS$	0.4200	0.3300	0.2667	0.2300	0.9200	0.8550	0.8000	0.7700
	Run-3, $D + DSU$	0.5200	0.3500	0.2933	0.2500	0.9100	0.8600	0.8133	0.7750
	Run-4, $D + MSU$	**0.5300**	**0.3700**	**0.3167**	**0.2775**	0.9100	0.8250	0.8067	0.7700
	Re-rank Upper Bound	0.7200	0.5400	0.4167	0.3375	1.0000	1.0000	0.9800	0.9325
WILDCARD (20)	Run-1, *D*	0.3800	0.3250	0.2967	0.2525	0.7400	0.6750	0.6800	0.6500
	Run-2, $D + DS$	**0.4700**	**0.4050**	0.3533	0.3075	0.7900	0.7700	0.7667	0.7575
	Run-3, $D + DSU$	0.4600	0.4000	**0.3633**	**0.3125**	0.8400	0.7750	0.7533	0.7375
	Run-4, $D + MSU$	0.4500	0.3800	0.3267	0.3100	**0.8900**	**0.8250**	**0.8000**	**0.7825**
	Re-rank Upper Bound	0.7700	0.5850	0.4700	0.4025	1.0000	0.9850	0.9567	0.9425

Table 2. Retrieval Times for Single Threaded Execution. *System:* Ubuntu Linux 14.04, 24 Intel Xeon 2.93 GHz Processors, 96 GB RAM.

	RETRIEVAL TIMES (SECONDS)			
TASK	μ	min	max	*median*
ARXIV MAIN	27.54	2.77	178.51	16.014
WIKI MAIN	37.83	1.33	176.06	33.84
WIKIPEDIA FORMULA BROWSING				
D (Core, Top-1k)	*2.67*	*0.10*	*64.13*	*1.07*
D + DS	12.75	0.17	109.61	3.61
D + DSU	45.26	0.58	1032.39	8.58
D + MSU	29.80	0.18	718.70	4.67
Concr. (20)	13.05	1.26	66.97	4.50
Wild. (20)	46.55	0.18	718.70	4.82

Wikipedia Formula Browsing Task. Submitted Top-20 D + MSU $P@5$ of Rel 49.0% $P.Rel$ 90.0%, vs. best (MCAT) Rel 51.5% $P.Rel$ 93.0%. Tangent is faster than the MCAT system, and uses only symbol layout. **Re-ranking may be improved (see *Re-rank Upper Bound* in Table 1).**

arXiv Main Task. 2nd-place $P@5$ for submitted Top-20. Rel 26.2% $P.Rel$ 54.5%, vs. best (MCAT) Rel 30.0% $P.Rel$ 57.9%. Run-2, using D + DSU re-ranking, equal text and formula weights, equally weighted query formulae. **Note: arXiv 'documents' contain little text.**

Wikipedia Main Task. 4th-place $P@5$, for submitted Top-20 Rel 25.3% $P.Rel$ 49.3%, vs. best (ICST) Rel 47.3% $P.Rel$ 85.3% (same condition as above). **Integrating text and formula retrieval, and representing referencing within and between articles produces better results.**

CONCLUSIONS

Q1. How should query text vs. formula matches be weighted?
A. Don't use independent indices and weight match scores. Consider interactions between text and formulas in context.

Q2. Should larger query formulae have higher weight?
A. Query formula relevance appears to be independent of size.

Q3. Is the global Dice coefficient over identical symbol pairs effective?
A. Produces an initial Top-1000 with high recall. Good for ranking exact matches and partial matches with many missing terms.

Q4. Does subexpression-based scoring affect Dice coefficient rankings?
A. Good partial matches are lost due to current subexpression matching method (connected component-based).

Q5. Does unification affect the perceived similarity of formula hits?
A. Unified matches perceived as good when result matches query; constraints needed (e.g., prevent $\sin$ unifying with x).

Q6. How do Dice coefficient-based rankings compare with Maximum Subtree Similarity (MSS)?
A. Overall MSS produced best avg. P@k metrics; however global Dice best for P.Rel concrete, local Dice re-ranking best for Rel. wildcard. Differences may be due to constrained matching and unification.

REFERENCES

[1] AIZAWA, A., KOHLHASE, M., OUNIS, I., AND SCHUBOTZ, M. NTCIR-11 Math-2 task overview. In *NTCIR* (2014), pp. 88–98.

[2] PATTANIYIL, N., AND ZANIBBI, R. Combining TF-IDF text retrieval with an inverted index over symbol pairs in math expressions: The Tangent math search engine at NTCIR 2014. In *NTCIR* (2014), pp. 135–142.

[3] STALNAKER, D., AND ZANIBBI, R. Math expression retrieval using an inverted index over symbol pairs. In *DRR* (2015), vol. 9402, pp. 940207–1–12.

[4] ZANIBBI, R., AIZAWA, A., KOHLHASE, M., OUNIS, I., TOPIĆ, G., AND DAVILA, K. NTCIR-12 mathir task overview. In *NTCIR* (2016), National Institute of Informatics (NII).

[5] ZANIBBI, R., AND BLOSTEIN, D. Recognition and retrieval of mathematical expressions. *IJDAR 15*, 4 (2012), 331–357.

[6] ZANIBBI, R., DAVILA, K., KANE, A., AND TOMPA, F. Multi-stage math formula search: Using appearance-based similarity metrics at scale. *SIGIR* (2016).

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-1016815 and HCC-1218801. Financial support from the Natural Sciences and Engineering Research Council of Canada under Grant No. 9292/2010, Mitacs, and the University of Waterloo is gratefully acknowledged.

NSF

NSERC
CRSNG

LINKS

CODE: cs.rit.edu/~dprl/Software.html
DPRL LAB: cs.rit.edu/~dprl

dprl