Tangent-3 at the NTCIR-12 MathIR Task

Kenny Davila, Richard Zanibbi
Rochester Institute of Technology, USA

Andrew Kane, Frank Wm. Tompa
University of Waterloo, Canada

NTCIR-12 MathIR Task Session
June 10, 2016

Tangent: Evolution

Tangent-1 - MSc thesis by D. Stalker (2013) extending T. Schellenberg's MSc thesis (2011). **Bag of symbol pairs** with inverted index for formula retrieval.

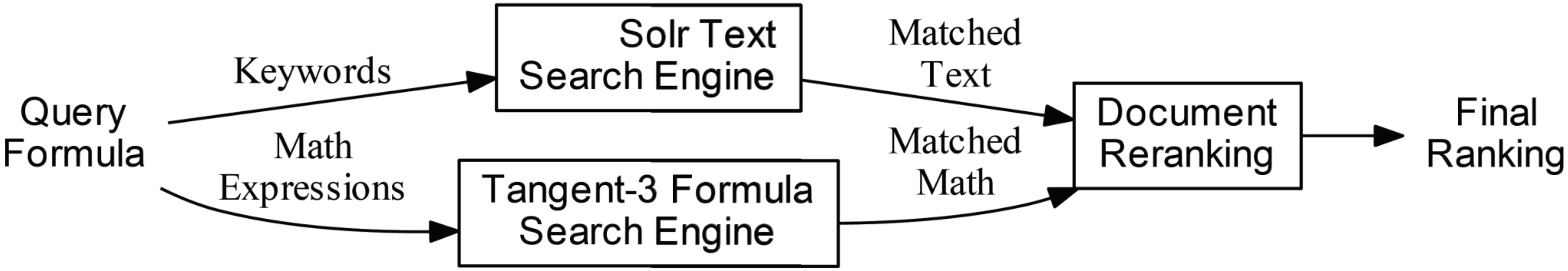
Tangent-2 added matrix support + text search (Lucene); strong results for NTCIR-11 Math-2 subtask at NTCIR-11 (N. Pattaniyil, MSc project 2014). *Large indices; slow retrieval.*

Tangent-3* improved formula representation, faster retrieval, improved wildcard support, unification of arguments (*numbers, ids*), and re-ranking by query recall in subexpressions (*Maximum Subtree Similarity*)
e.g., ‘x^2’ in ‘x^2’ and ’$x^2 + 1$’ treated as equally strong matches.

> **Text/keyword retrieval** via same independent Lucene index from Tangent-2. Linearly combine text and formula match scores.

**R. Zanibbi, K. Davila, A. Kane, and F. Tompa. Multi-stage math formula search: Using appearance-based similarity metrics at scale. SIGIR, 2016.*

Tangent-3 Text + Math Retrieval



Formulae and Keywords retrieved independently.

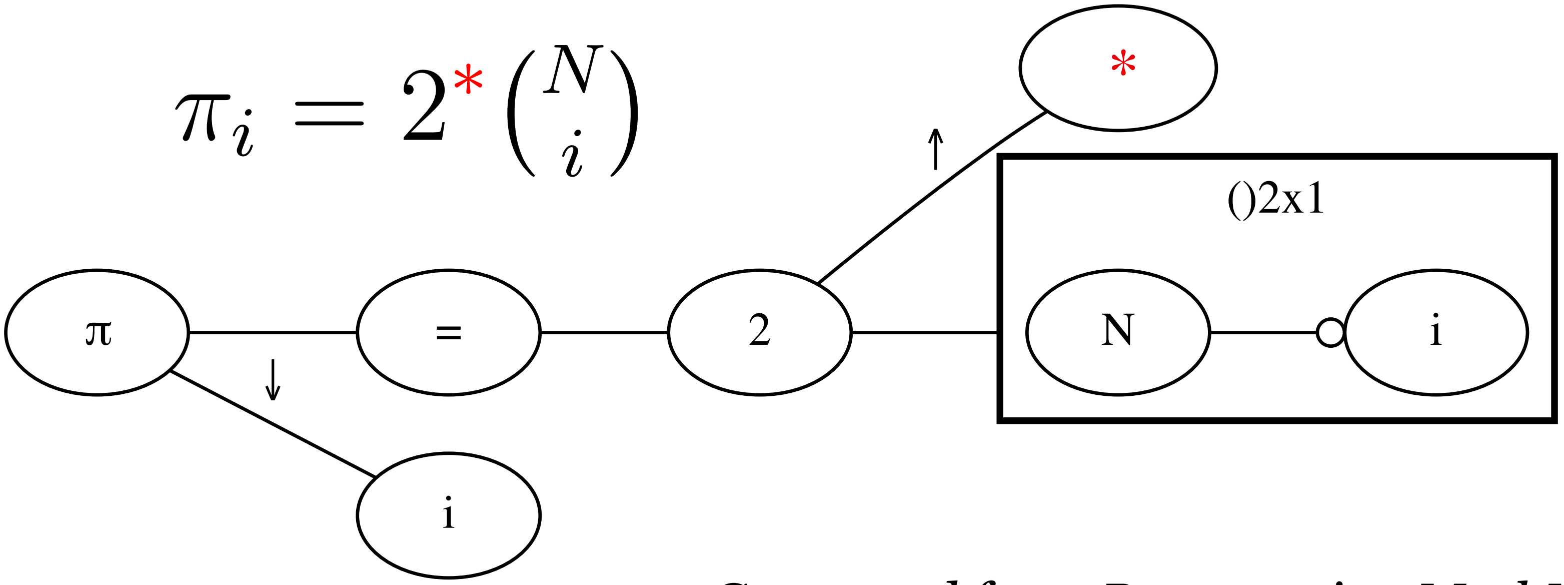
Documents ranked using linear combination of best formula match scores in each document, and Lucene document scores.

Formula Retrieval in Tangent-3

Structure Representation
Indexing
Wildcards and Unification
Re-ranking

Formula Representation
Symbol Layout Tree (SLT, Appearance-Based)

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



Generated from Presentation MathML

All groupings (matrices, vectors, parens, etc.) represented identically. Unlike Tangent-2, distinguishes *above* from *superscript*, and *below* from *subscript*.

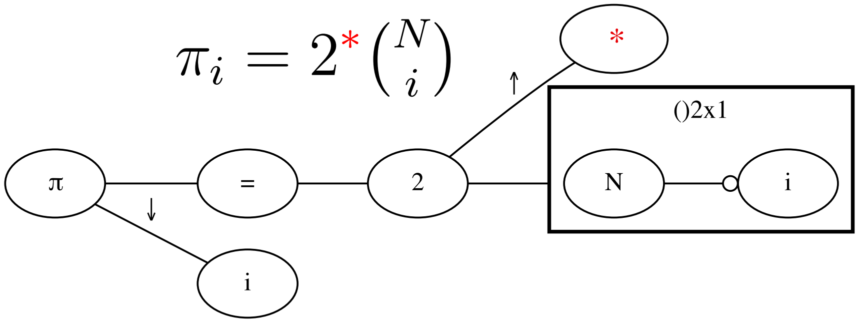
Formula Indexing

1. Inverted Index for Symbol Pairs
   Keys: Pairs of symbols/groupings
   Values: Posting lists of unique formula ids

2. Formula uid → Formula/Doc Index

   e.g., ‘x^2’ uid to *all (doc:offset)* entries

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



Symbol Pairs with Relationships

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$
V!π	**V!**i	$\rightarrow\rightarrow\rightarrow\boxed{\cdot}\multimap$	1

**For SLTs with tree height less than 3, symbols at end of writing lines also indexed.*

Wildcard Matching and Unification

Case	Query	Match
Unrestricted	$x + *$	$x+1$ $x+y+z+sin(x)$ $y+x+z=\frac{\pi}{4}$
	e^{*}	$f(x)=e^{x+1}+2$
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	$*+1$	$x+y+z+1$ $\alpha=f(x+y+1,x^{2})$ $f(x,y)=\frac{1}{x+y+1}$

Greedy Wildcard Expansion: use exact symbol matches first, then ‘flood fill’ with constraints.

QUERY

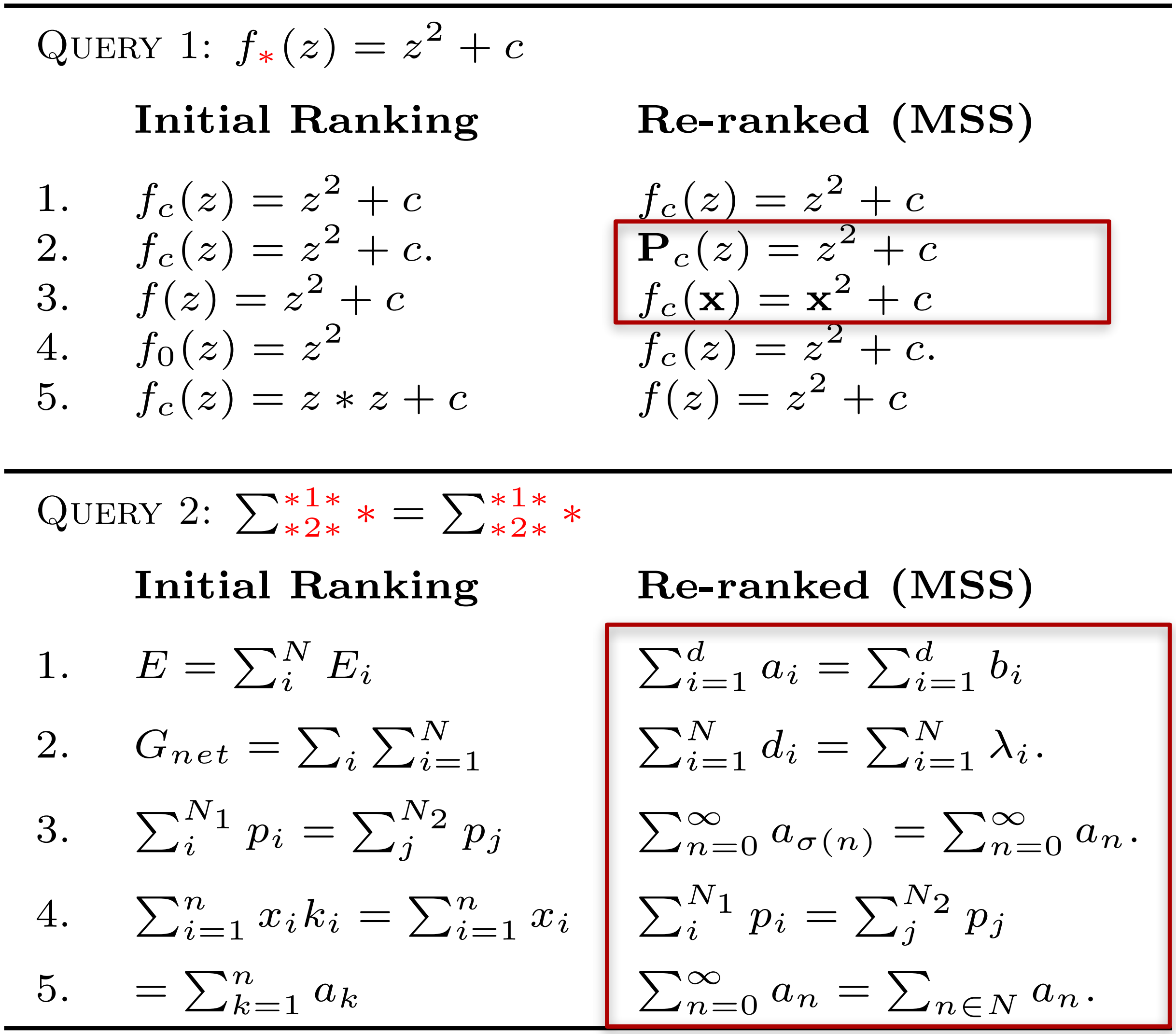
$$x^2+y^2=*$$

MATCH

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

blue: exact match, red: wildcard match; orange: unification.

Examples: Formula Re-ranking



Query 1: $f_{*}(z) = z^2 + c$

	Initial Ranking	Re-ranked (MSS)
1.	$f_c(z) = z^2 + c$	$f_c(z) = z^2 + c$
2.	$f_c(z) = z^2 + c.$	$\mathbf{P}_c(z) = z^2 + c$
3.	$f(z) = z^2 + c$	$f_c(\mathbf{x}) = \mathbf{x}^2 + c$
4.	$f_0(z) = z^2$	$f_c(z) = z^2 + c.$
5.	$f_c(z) = z * z + c$	$f(z) = z^2 + c$

Query 2: $\sum_{*2*}^{*1*} * = \sum_{*2*}^{*1*} *$

	Initial Ranking	Re-ranked (MSS)
1.	$E = \sum_i^N E_i$	$\sum_{i=1}^d a_i = \sum_{i=1}^d b_i$
2.	$G_{net} = \sum_i \sum_{i=1}^N$	$\sum_{i=1}^N d_i = \sum_{i=1}^N \lambda_i.$
3.	$\sum_i^{N_1} p_i = \sum_j^{N_2} p_j$	$\sum_{n=0}^\infty a_{\sigma(n)} = \sum_{n=0}^\infty a_n.$
4.	$\sum_{i=1}^n x_i k_i = \sum_{i=1}^n x_i$	$\sum_i^{N_1} p_i = \sum_j^{N_2} p_j$
5.	$= \sum_{k=1}^n a_k$	$\sum_{n=0}^\infty a_n = \sum_{n\in N} a_n.$

Unifiable Types

- identifier
- number
- groupings (e.g., matrix)

Identifiers

- Variables
- Function Names
- etc.

Re-Rank Scoring

From best subexpression

Results

Participated in three tasks:

1. Wikipedia Formula Browsing Task
2. arXiv Main Task
3. Wikipedia Main Task

Index Sizes and Retrieval Times

Wiki formula index: 580.5 MB

arXiv formula index: 8.3 GB

	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

***Note:** Core formula engine implemented in C++; re-ranking functions in Python (4-10 times slower)*

Single Threaded Execution

Ubuntu Linux 14.04

24 Intel Xeon 2.93 GHz Processors

96 GB RAM

Tangent-2: >3 mins for each arXiv query at NTCIR-11, parallel retrieval over 9 shards.

NTCIR

Wiki Formula Browsing Task

Formula Similarity Metrics

1. Core Engine: Dice Coefficient of Symbol Pairs, *2RP/(R+P)*
2. Core + Dice Coefficient for best subexpression
3. Core + Dice Coefficient for best subexpression w. unification
4. Core + Maximum Subtree Similarity (MSS) Vector w. unification

MSS: Dice Coeff. for SLT symbol and edge *recall:* $2R_sR_e/(R_s+R_e)$

$S(k)$	>	$P(k)$	>	$\omega^2(k)$	>	(k)	>	$V^{(k)}$
(1, 0, 3)		(1, 0, 2)		(1, -1, 2)		(0.6, 0, 2)		(0.6, -1, 2)

MSS Scoring for Query S(k). Ranking triples contain MSS (1), and the number of candidate symbols that are unmatched (2) and exactly matched (3). Parentheses count as one symbol.

Formula Browsing Task Results (40 queries)

	Relevant				*Partially-Relevant*			
	P@5	*P@10*	*P@15*	*P@20*	*P@5*	*P@10*	*P@15*	*P@20*
MCAT	**0.5150**	**0.4050**	**0.3450**	**0.3000**	**0.9300**	**0.8650**	**0.8300**	**0.8012**
Core (Dice Coeff.)	0.4300	0.3400	0.2933	0.2450	0.8400	0.7800	0.7533	0.7225
Core + Subexp. Dice	0.4450	0.3675	0.3100	0.2687	0.8550	0.8125	0.7833	0.7638
Core +SDice+Unif.	**0.4900**	**0.3750**	**0.3283**	0.2812	0.8750	0.8175	0.7833	0.7563
Core + MSS	**0.4900**	**0.3750**	0.3217	**0.2937**	**0.9000**	**0.8250**	**0.8033**	**0.7762**
Upper Bound (Top-1k)	0.7450	0.5625	0.4433	0.3700	1.0000	0.9925	0.9683	0.9375
Ideal Pool (all sys.)	0.7900	0.6400	0.5383	0.4725	1.0000	1.0000	0.9933	0.9800

Overall, subexpression-based ranking, subexpression wildcards and unification help.

Note: *If we break apart 20 queries with from 20 queries without wildcards, Core + MSS is not always the best ranking procedure.*

ArXiv/Wikipedia Main Tasks

Fixed formula similarity metric

Core + Dice Coefficient for best subexpression w. Unification

Keyword vs. Math score weighting

Uniform (50-50)

Dynamic (proportional to number of query terms)

Multiple query formulae weighting

Uniform

Proportional to query formula sizes (# symbols)

**Total of 4 weighting combinations per task (4 runs)*

Total of 4 weighting combinations per task (4 runs)

ArXiv Main Task Results (29 Queries)

	Relevant				*Partially-Relevant*			
	P@5	*P@10*	*P@15*	*P@20*	*P@5*	*P@10*	*P@15*	*P@20*
MCAT	**0.2897**	**0.2448**	**0.2276**	**0.2000**	**0.5793**	**0.5552**	**0.5402**	**0.5121**
Uniform, Pr.Size	0.2552	**0.2000**	0.1586	0.1345	**0.5517**	0.4517	0.3908	0.3483
Uniform, Uniform	**0.2621**	**0.2000**	**0.1632**	**0.1362**	0.5448	0.4552	0.3908	0.3517
Pr.Terms, Pr.Size	0.1862	0.1552	0.1425	0.1259	0.5448	0.4931	0.4575	0.4414
Pr.Terms, Uniform	0.1862	0.1586	0.1425	0.1276	0.5310	**0.5034**	**0.4644**	**0.4448**
Ideal Pool	0.6966	0.5586	0.4644	0.4086	0.9655	0.9552	0.9172	0.8828

For Relevant hits, uniform weighting of query terms and combined text and math scores works best.

For Partially Relevant hits, proportional weighting for text/math or query formula sizes obtain best results at different ranks.

Wikipedia Main Task Results (30 Queries)

	Relevant				*Partially-Relevant*			
	P@5	*P@10*	*P@15*	*P@20*	*P@5*	*P@10*	*P@15*	*P@20*
ICST	**0.4733**	**0.3767**	**0.2978**	**0.2617**	**0.8533**	**0.79**	**0.7133**	**0.66**
Uniform, Pr.Size	0.2467	0.2333	0.2156	**0.2050**	**0.4933**	0.4900	**0.5000**	**0.4850**
Uniform, Uniform	**0.2533**	**0.2500**	**0.2200**	**0.2050**	**0.4933**	**0.4933**	0.4867	0.4767
Pr.Terms, Pr.Size	0.1600	0.1267	0.1222	0.1250	0.3867	0.3667	0.3689	0.3567
Pr.Terms, Uniform	0.1533	0.1400	0.1289	0.1250	0.3800	0.3667	0.3600	0.3550
Ideal Pool	0.8400	0.6967	0.5956	0.5133	0.9467	0.9400	0.9289	0.9217

Similar result; uniform weightings do best at higher ranks.

Systems that use text features/context performed much better on this task, due to the text available in the full articles.

Conclusion

What worked…and what did not.

Summary: Our Observations

Don’t use independent indices for text and math. Consider interactions between text and formulas in context.

Query formula relevance appears to be **independent of size**.

Core formula retrieval results produce an initial Top-1000 with **high recall**. Good for ranking exact matches and partial matches with many missing terms; **room to improve re-ranking.**

Scoring formulae using subexpression matching helps, but **good partial matches missed by our subexpression matching method** (connected component-based).

Unified formula matches 'good' when candidates are very similar to query; constraints needed (e.g., *sin* unifies with *x*).

Overall MSS reranking produced best Wiki Formula Browsing results, but Core results best for P.Rel concrete, local Dice re-ranking best for Rel. wildcard queries. Differences may be due to **constrained matching and unification**.

Thank you.

Source code: www.cs.rit.edu/~dprl/Software.html

This material is based upon work supported by the National Science Foundation (USA) under Grants 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.

Formula Browsing Task Results (20 concrete queries)

	Relevant				*Partially-Relevant*			
	P@5	*P@10*	*P@15*	*P@20*	*P@5*	*P@10*	*P@15*	*P@20*
Core (Dice Coeff.)	0.4800	0.3550	0.2900	0.2375	**0.9400**	**0.8850**	**0.8267**	**0.7950**
Core + Subexp. Dice	0.4200	0.3300	0.2667	0.2300	0.9200	0.8550	0.8000	0.7700
Core + Sub Dice + Unif.	0.5200	0.3500	0.2933	0.2500	0.9100	0.8600	0.8133	0.7750
Core + MSS	**0.5300**	**0.3700**	**0.3167**	**0.2775**	0.9100	0.8250	0.8067	0.7700
Upper Bound (Top-1k)	0.7200	0.5400	0.4167	0.3375	1.0000	1.0000	0.9800	0.9325
Ideal Pool (all sys.)	0.7300	0.5800	0.4733	0.4000	1.0000	1.0000	0.9967	0.9800

Formula Browsing Task Results (20 wildcard queries)

	Relevant				*Partially-Relevant*			
	P@5	*P@10*	*P@15*	*P@20*	*P@5*	*P@10*	*P@15*	*P@20*
Core (Dice Coeff.)	0.3800	0.3250	0.2967	0.2525	0.7400	0.6750	0.6800	0.6500
Core + Subexp. Dice	**0.4700**	**0.4050**	0.3533	0.3075	0.7900	0.7700	0.7667	0.7575
Core + Sub Dice + Unif.	0.4600	0.4000	**0.3633**	**0.3125**	0.8400	0.7750	0.7533	0.7375
Core + MSS	0.4500	0.3800	0.3267	0.3100	**0.8900**	**0.8250**	**0.8000**	**0.7825**
Upper Bound (Top-1k)	0.7700	0.5850	0.4700	0.4025	1.0000	0.9850	0.9567	0.9425
Ideal Pool (all sys.)	0.8500	0.7000	0.6033	0.5450	1.0000	1.0000	0.9900	0.9800

Tangent-3 Formula Retrieval

Step 1. Core Engine (Candidate Hit Generation)

- Retrieves based on *all* symbol pairs; ranked via symbol pair Dice coefficient (*harmonic mean:* 2RP/(R+P))
- Top-k unique formulae returned as candidates.
- Wildcards treated as single symbols.

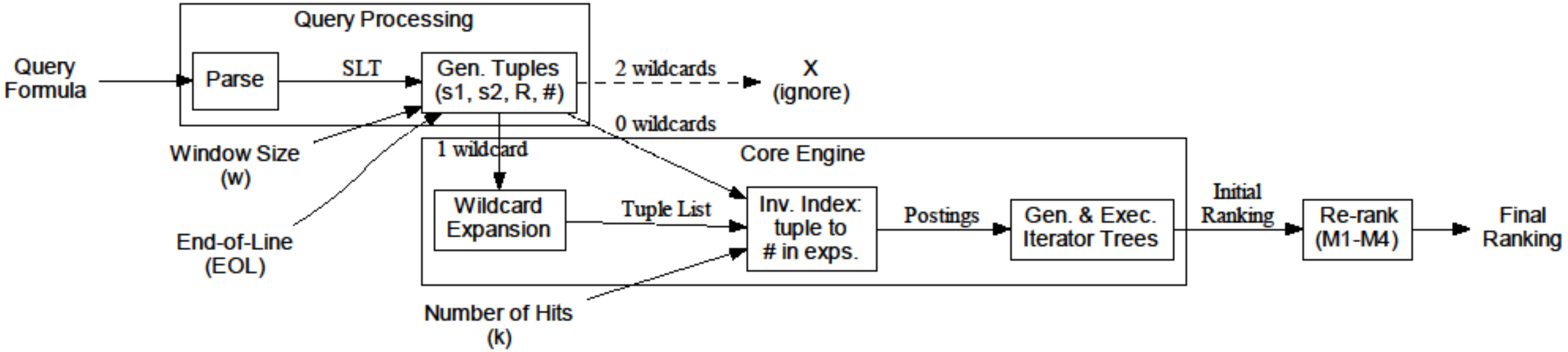
Step 2 (Optional). Re-rank Formula Hits

- Detailed Matching
  - Wildcards may match subexpressions
  - Support unification of numbers, identifiers
  - Find best *subexpression* matching the query
- Scoring vector (variety of similarity metrics considered)

Step 3. Produce 'Math Score' for Documents

- Lookup documents corresponding to matched formulae.
- Use best match for each query formulae on a document for scoring.
- Match scores linearly combined to produce 'math score.'

Tangent-3 Formula Retrieval Model



After final rankings for all query formula hits, complete **Step 3**
(score docs via **linear comb.** of best match scores for query formulae)