

Mathematical Formula Retrieval

Presentation by

Megha Chakraborty
Oshita Saxena
Yash Jain
Yash Raj

Contents

① Why Formula Retrieval?

② Formula Representation

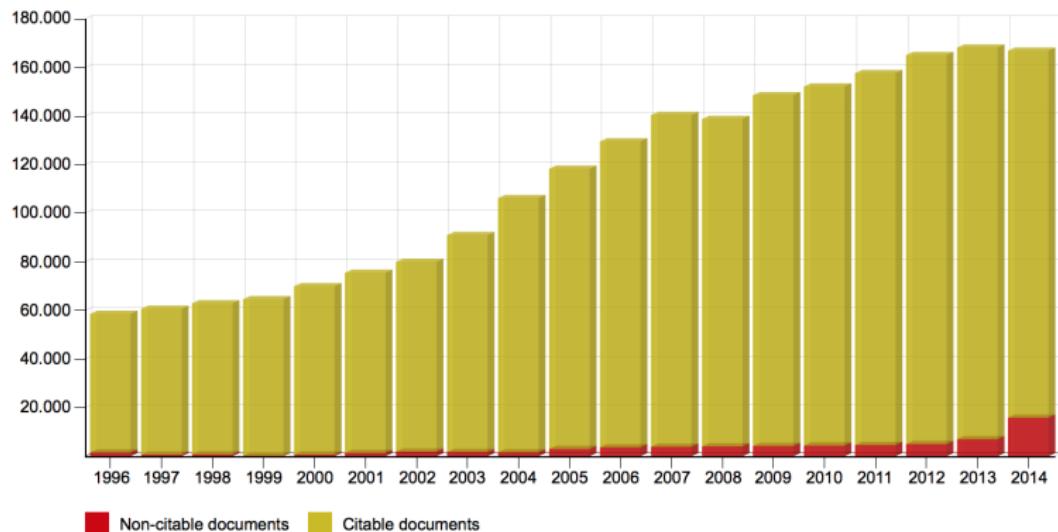
③ Retrieval Models

Text Based Model

Tree Based Model

Embedding Models

Why Formula Retrieval?



Number of Math publications every year

Why Formula Retrieval?

How to solve your problem

$(a + b)^2$

Simplify

1 Expand the square ^

$(a + b)^2$

$(a + b)(a + b)$

2 Distribute v

3 Distribute v

Solution

$a^2 + 2ab + b^2$

2 more steps

Search result for $(a + b)^2$

c^mi | CHENNAI
MATHEMATICAL
INSTITUTE

Why Formula Retrieval?



lim n tends to infinity $1/n^3 \cos^{21}/n^3$



All

Images

Shopping

Videos

News

More

Tools

About 4,02,00,000 results (0.70 seconds)

<https://www3.nd.edu> › Math10560 › Work PDF



SOLUTIONS TO EXAM III, MATH 10560 1. Determine which ...

$n=1 \cdot 1 \cos^2 n+1 \cdot (5)$. Notice $\lim_{n \rightarrow \infty} 3(\pi/2)n = \infty$, so by the divergence test $\sum_{n=1}^{\infty} 3(\pi/2)n$ diverges. $\sum_{n=1}^{\infty} \cos^2 n \cdot 2^n$ converges.

8 pages

<https://math.stackexchange.com> › questions › evaluate-s...



Evaluate $\sum_{n=1}^{\infty} \frac{n}{n^4+n^2+1-n^3}$

And I go, since $n^4+n^2+1-n^3$ and we know that $\sum_{n=1}^{\infty} n^3$ converges, so $\sum_{n=1}^{\infty} n$...

6 answers · Top answer: HINT: As $n^4+n^2+1=(n^2+1)^2-n^2=(n^2+1-n)(n^2+1+n)$ and $(n^2+1+n)-(n^2+1-n)$...

<https://math.stackexchange.com> › questions › evaluate-l...



Evaluate $\lim_{n \rightarrow \infty} \left(\frac{\cos(1/n) - \sin(1/n)}{1/t} \right)^n$

As the larger fraction tends to 1, it can be ignored. Now, with $t=\sin 1/n$, we have.

$\lim_{t \rightarrow 0} (\sqrt{1-t^2}-1)/t$. Using Taylor, $\sqrt{1-t^2}-1=t+o(t)$.

Search results for

$$\lim_{n \rightarrow +\infty} \frac{1}{n^3} \cos^2 \frac{1}{n^3}$$

Why Formula Retrieval?

Query

How can I evaluate $\sum_{n=0}^{\infty} (n+1)x^n$?

Asked 8 years, 5 months ago Active 4 months ago Viewed 34k times

How can I evaluate

384

$$\sum_{n=1}^{\infty} \frac{2n}{3^{n+1}}$$

I know the answer thanks to [Wolfram Alpha](#), but I'm more concerned with how I can derive that answer. It cites tests to prove that it is convergent, but my class has never learned these before so I feel that there must be a simpler method.

146

In general, how can I evaluate

$$\sum_{n=0}^{\infty} (n+1)x^n$$

[sequences-and-series](#) [convergence](#) [power-series](#) [faq](#)

edited Sep 24 '17 at 12:09 by  Pardly Taxil 51.7k 13 80 120

asked Apr 3 '11 at 21:41 by  Backus 2,072 5 12 8

Search Results

- 1** No need to use Taylor series, this can be derived in a similar way to the formula for geometric series. Let's find a general formula for the following sum:
$$S_m = \sum_{n=1}^m nr^n.$$

...
- 2** It is equivalent to $x(x+1)(x+5)(x+6) + 96 = 0$

Now

...

$$(x^2 + 6x)(x^2 + 6x + 5) + 96 = 0$$
- 3** If you want a solution that doesn't require derivatives or integrals, notice that
$$1 + 2x + 3x^2 + 4x^3 + \dots = 1 + x + x^2 + x^3 + \dots$$
$$+ x + x^2 + x^3 + \dots$$
$$+ x^2 + x^3 + \dots$$

...

Answer Retrieval

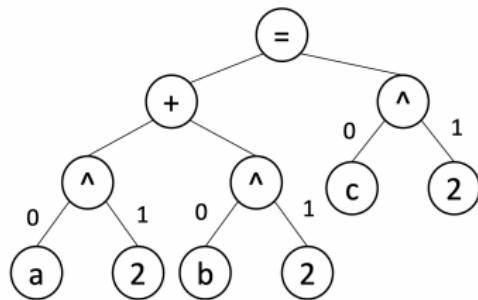
Why Formula Retrieval?

Query	Search Results
<p>How can I evaluate $\sum_{n=0}^{\infty} (n+1)x^n$?</p> <p>Asked 8 years, 5 months ago Active 4 months ago Viewed 34k times</p> <p>How can I evaluate</p> <p>384</p> <p>I know the answer thanks to Wolfram Alpha, but I'm more concerned with how I can derive that answer. It cites tests to prove that it is convergent, but my class has never learned these before so I feel that there must be a simpler method.</p> <p>In general, how can I evaluate</p> $\sum_{n=0}^{\infty} (n+1)x^n?$ <p>sequences-and-series convergence power-series faq</p> <p>edited Sep 24 '17 at 12:09 Pardy Taxel 51.7k 13 80 120</p> <p>asked Apr 3 '11 at 21:41 Backus 2,072 3 12 8</p>	<p>1 $\sum_{n=0}^{\infty} (n+1)x^n$</p> <p>2 $\sum_{n=0}^{\infty} (n+1)x^n$</p> <p>3 $\int_0^1 \frac{\ln(x+1)}{x^2 + 1} dx$</p>

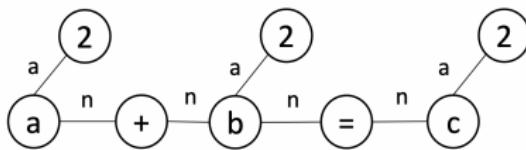
Formula Retrieval

Formula Representation

$$a^2 + b^2 = c^2$$



(a)



(b)

Figure: (a) Operator Tree and, (b) Symbol Layout Tree

Nodes and Edges

Node labels: **Type!**value

- Numbers (**N!**2)
- Variable names (**V!**x)
- Text fragments (**T!**'lim')
- Fractions (**F!** 1/2)
- Matrices (**M!** () 2×3)
- Wildcard (*)
- Commutative operator (**U!**)
- Non-commutative operator (**O!**)

Edge labels

- **SLTs:** edge labels give spatial arrangement of symbols
 - Next (**n** or \rightarrow)
 - Within (**w** or \square)
 - Above (**a** or \uparrow)
 - Below (**b** or \downarrow)
 - Under (**U**)
- **OPTs:** edge labels indicate argument position
 - 0/1

Approaches for Formula Retrieval

- **Text-based** (formulae are converted to a string such as \LaTeX)
- **Tree-based** (formulae are represented as trees like SLTs/OPTs)
- **Embedding-based** (formulae are represented as trees and then converted to vectors)

MlaS: A Text Based Model

MIaS: A Text Based Model

MIaS-Math Indexer and Searcher

- Math-aware full-text search engine based on Apache Lucene
- Joins textual and mathematical querying
- Takes MathML or L^AT_EX as input

How to write query

x^2+y^2 exponential distribution

d

Search in: MREC 2011.4.439 ▾ Search

Total hits: 15973, showing 1-30. Searching time: 584 ms

[Andreev bound states in normal and ferromagnet/high-T_C superconducting tun ...](#)

... close from the [110] surface when the symmetry is $d_{x^2+y^2}$.

score = 1.1615998

arxiv.org/abs/cond-mat/0305446 - cached XHTML

[Particle trajectories and acceleration during 3D fan reconnection](#)

... at $\sqrt{x^2 + y^2} = 1$ and ...

score = 1.0577431

arxiv.org/abs/0811.1144 - cached XHTML

[Pairing symmetry and long range pair potential in a weak coupling theory of ...](#)

... does not mix with usual $s_{x^2-y^2}$ symmetry gap in an anisotropic band structure.

score = 1.0254444

arxiv.org/abs/cond-mat/9906142 - cached XHTML

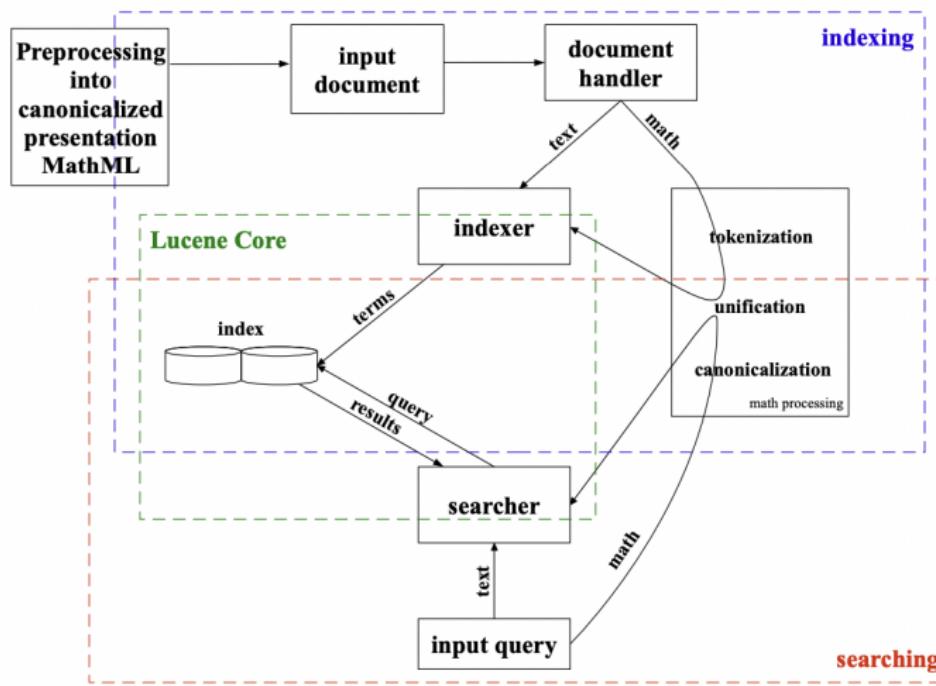
MlaS: A Text Based Model

- In text retrieval: index word stems only instead of word forms
- From text to math: have multiple representations of a formula (with different 'near synonyms') put in the index
- \LaTeX notation for querying
- MathML for indexing
- Uses Lucene's standard practical scoring function:

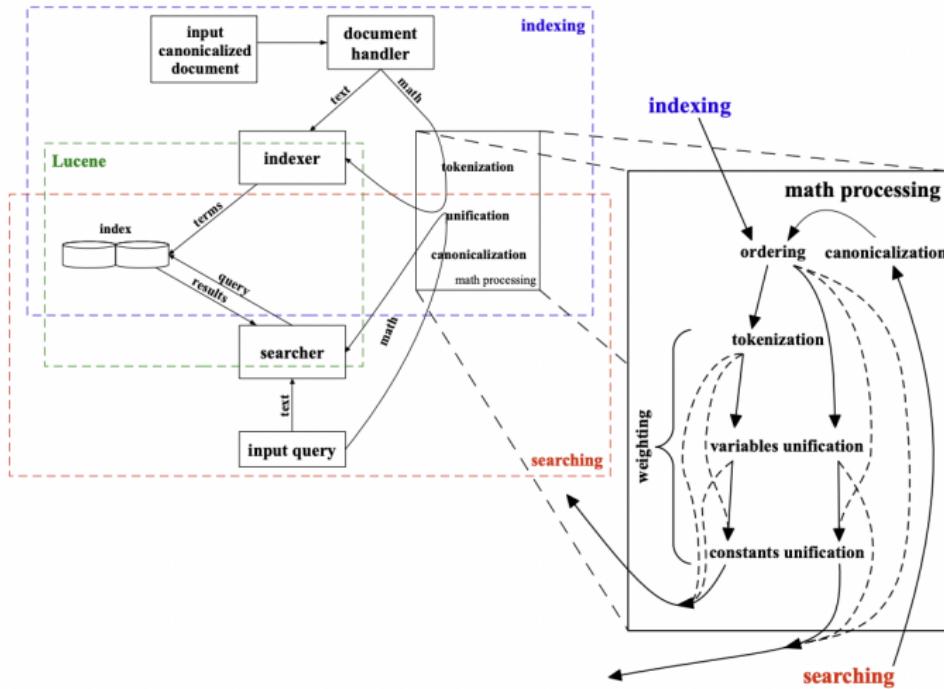
$$\text{score}(q,d) = \text{coord}(q,d) \cdot \text{queryNorm}(q) \cdot \sum_{t \text{ in } q} (\text{tf}(t \text{ in } d) \cdot \text{idf}(t)^2 \cdot \text{t.getBoost()} \cdot \text{norm}(t,d))$$

MlaS: A Text Based Model

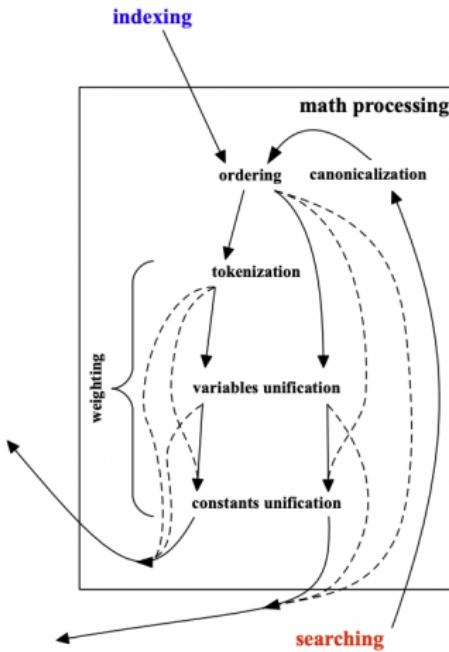
MlaS Framework



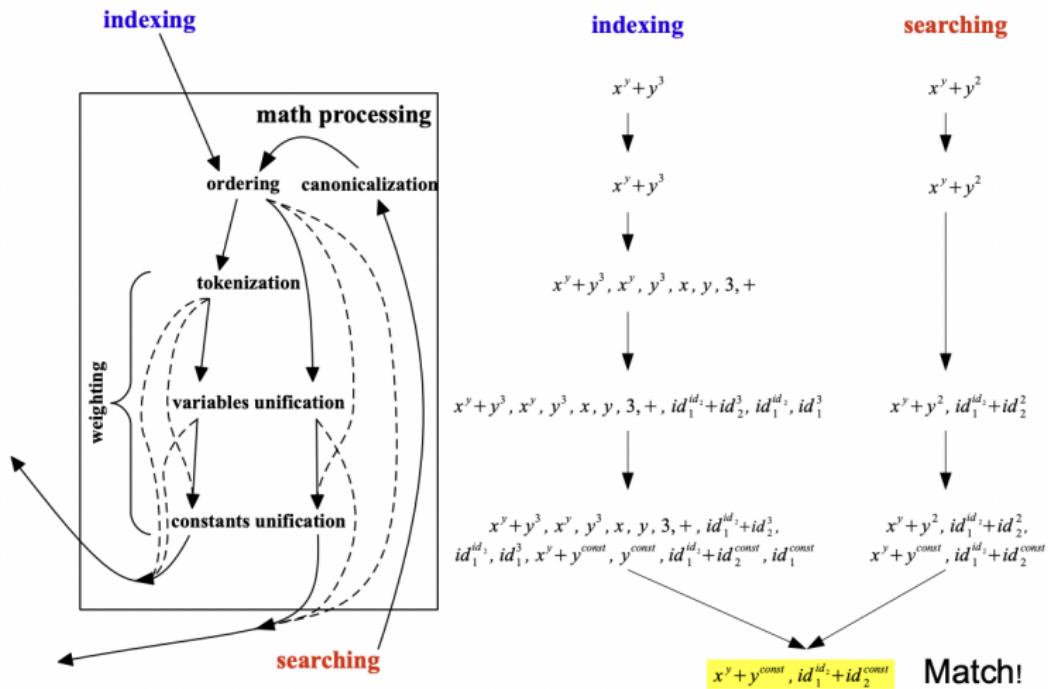
Indexing Framework



Index Processing



Example



Formula Processing Example – Subformula Weighting

input:

$$(a + b^{2+c}, 0.125)$$

ordering:

$$(a + b^{c+2}, 0.125)$$

tokenization:

$$(a, 0.0875) \quad (+, 0.0875) \quad (b^{c+2}, 0.0875)$$

variables
unification:

$$(id_1 + id_2^{id_3+2}, 0.1)$$

constants
unification:

$$(a + b^{c+const}, 0.0625)$$

$$(b^{c+const}, 0.04375)$$

$$(c+2, 0.06125)$$

$$(b, 0.06125)$$

$$(c, 0.042875)$$

$$(+, 0.042875)$$

$$(2, 0.042875)$$

$$(id_1^{id_3+2}, 0.07)$$

$$(id_1+2, 0.0343)$$

$$(c+const, 0.030625)$$

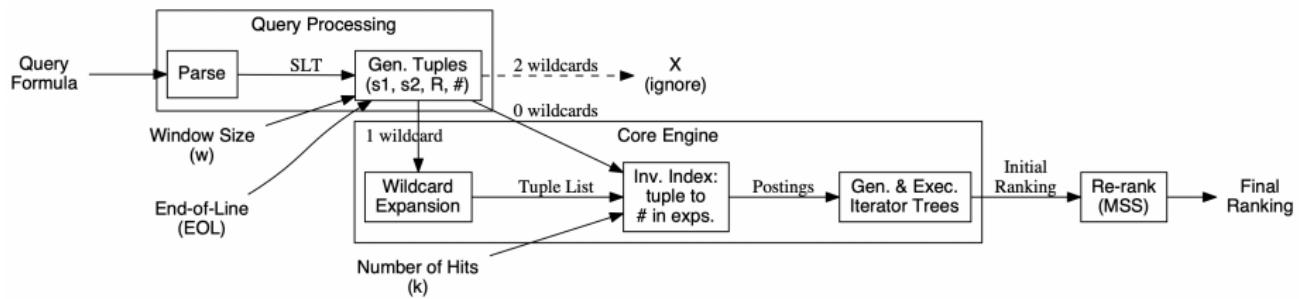
$$(id_1+const, 0.01715)$$

$$(id_1^{id_3+const}, 0.035)$$

Tangent-3: A Tree Based Model

Tangent-3: A Tree Based Model

Tangent-3 Formula Search Engine



Tree to Tuples

SLT TUPLES				OPT TUPLES			
(PARENT,	CHILD,	PATH,	PFR)	(PARENT,	CHILD,	PATH,	PFR)
(V! a,	N! 2,	a,	-)	(U! eq,	U! plus,	0,	-)
(N!2,	eob,	n,	a)	(U! plus,	O! SUP,	0,	0)
(V! a,	+,	n,	-)	(O! SUP,	V!a,	0,	00)
:				:			

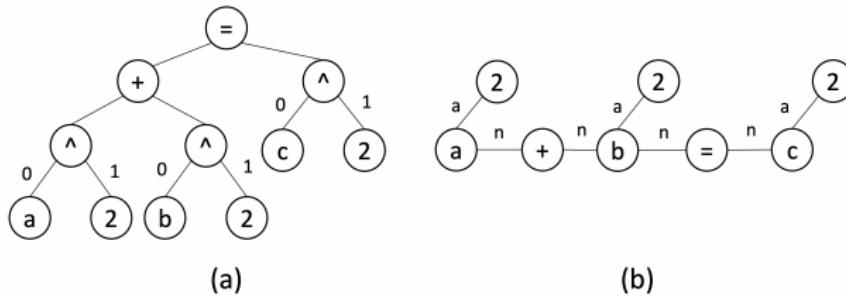


Figure: (a) Operator Tree and, (b) Symbol Layout Tree

Indexing

Inverted index built over document-formula-tuple relationships

- D_f : formula \rightarrow formID
- D_t : tuple \rightarrow tupleID
- D_d : document \rightarrow docID
- D_w : wildcardtuple \rightarrow wildcardtupleID
- P_t : tupleID \rightarrow (formID, count) $^+$
- P_f : formID \rightarrow (docID, position) $^+$
- P_w : wildcardtupleID \rightarrow (tupleID) $^+$

Core engine

Greedy wildcard expansion

Behavior	Query	Match
Unrestricted	$x + *$	$x+1$ $x+y+z+\sin(x)$
	e^*	$y+x+z = \frac{\pi}{4}$ $f(x) = e^{x+1} + 2$
Restricted by children	$*^2+1$	$x^2 + y^2 + 1$ $x^2 + y + 1$ $x^2 + (y + z)^2 + 1$
Restricted by binding	$*1*^2+*1*+1$	x^2+x+1 $(x+1)^2+(x+1)+1$ x^2+y+1
Horizontal expansion (right)	$x + *+1$	$x+y+1$ $x+y+z+1$ $x+y-z+1$ $x+\frac{1}{2+y}-3z+1$
Horizontal expansion (left)	$*+1$	$x+y+z+1$ $\alpha = f(x+y+1, x^2)$ $f(x, y) = \frac{1}{x+y+1}$

Candidate Selection

The engine uses Dice's coefficient over tuples as a simple ranking algorithm, counting number of tuples that overlap between query and candidate formula

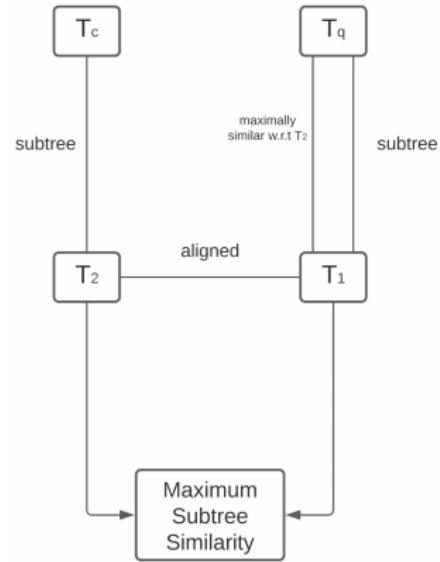
Dice's Coefficient

Given query tree T_q and candidate tree T_c with tuple sets F_q and F_c ,
Dice's coefficient of similarity is given by $\frac{2|F_q \cap F_c|}{|F_q| + |F_c|}$

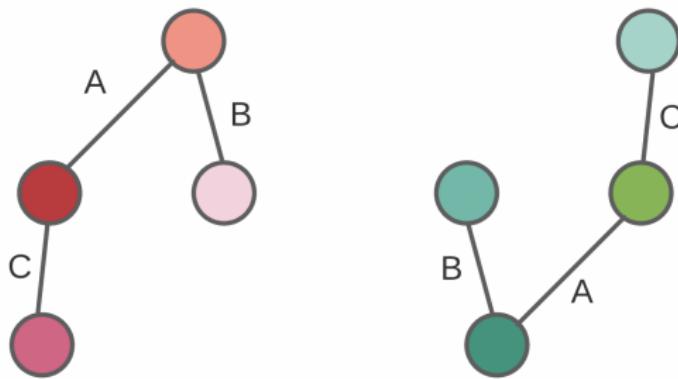
Core Engine

Re-ranking

- T_q : query tree
- T_c : candidate tree
- T_1 : subtree of T_q
- T_2 : subtree of T_c



Aligned Subtrees



- Similar structure
- Nodes may not match

Maximally Similar Subtree

- Let $m = \#$ matching nodes between aligned trees T_1 and T_2
- Dice's coefficient = $\frac{2m}{|T_1| + |T_q|}$
- = measure of similarity of T_1 to T_q with respect to T_2
- T_1 is maximally similar to T_q if root of T_1 can be matched with root of T_2 and has maximum dice's coefficient

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.	$= \sum_{k=1}^n a_k$	$\sum_{n=0}^{\infty} a_n = \sum_{n \in N} a_n.$

Tangent-CFT: An Embedding Based Model

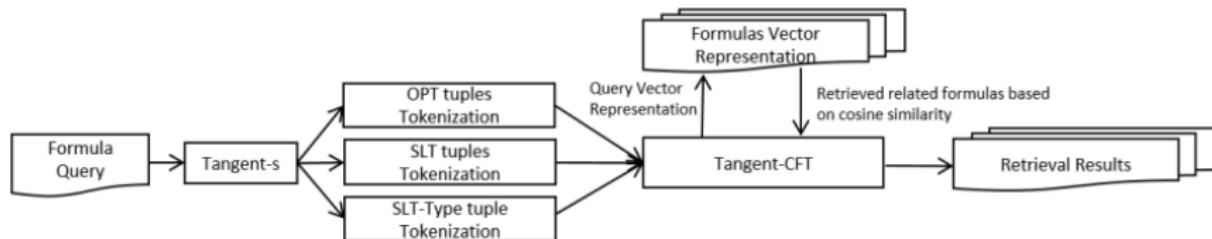
Introduction to Embedding Model: Tangent-CFT

- Tangent Combined with fastText
- n-gram EM for OPT and SLT representations
- Current SOTA for partial matching on NTCIR-12 formula task
- Drawback: loses structural information- less effective for near-exact matches

Similarity Features for Embedding Models

- Provide dense vector representations - provide better matching for partially relevant formulae
- Originally:
 - 3 n-gram embedding models are trained separately on SLT, OPT and SLT Type trees
 - Final representation was sum of 3 embedding vectors
- For Learning to Rank Models:
 - Produce similarity scores separately for each of the three representations
 - Combine these using LtR model

Framework



Generating the Embedding Vectors

- 1 Tuple sequence generation
- 2 Tokenizing formula tuples: each character (token) is encoded using a unique identifier
- 3 Applying fastText to formulae
- 4 Combining SLT and OPT embeddings (we are avoiding this step)

Thank You!