# Combining TF-IDF Text Retrieval with an Inverted Index over Symbol Pairs in Math Expressions:

## The Tangent Math Search Engine at NTCIR 2014

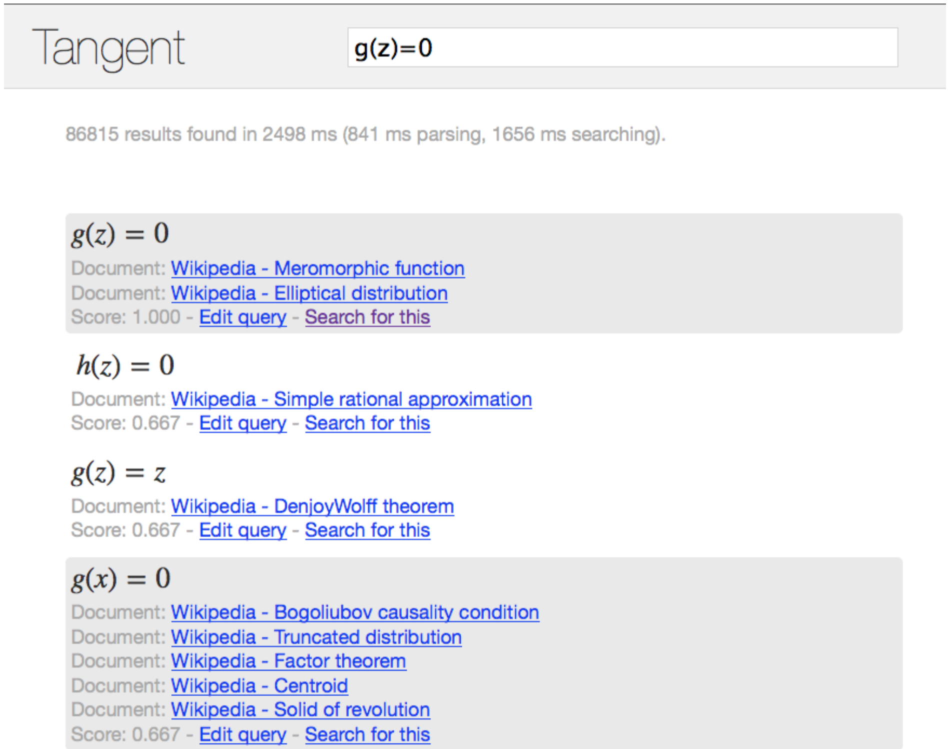
Nidhin Pattaniyil and Richard Zanibbi
**Document and Pattern Recognition Laboratory**
Department of Computer Science
Rochester Institute of Technology, NY, USA

**NTCIR-11 (2014) Math-2 Task Presentation**
National Institute of Informatics (NII), Tokyo, Japan

Dec. 11, 2014

NTCIR

# Tangent




saskatoon.cs.rit.edu/tangent

www.cs.rit.edu/~dprl/Software.html

## A Formula Search Engine

Previously used for expressions in Wikipedia (Stalnaker, 2013). Appearance-based retrieval model using relative positions of symbols in LaTeX or Presentation MathML

## NTCIR-11 Modifications

- Represent matrices, prefix scripts
- Support wildcard query variables
- Support multiple query expressions
- Support keywords (Lucene)
- Reduced storage requirements

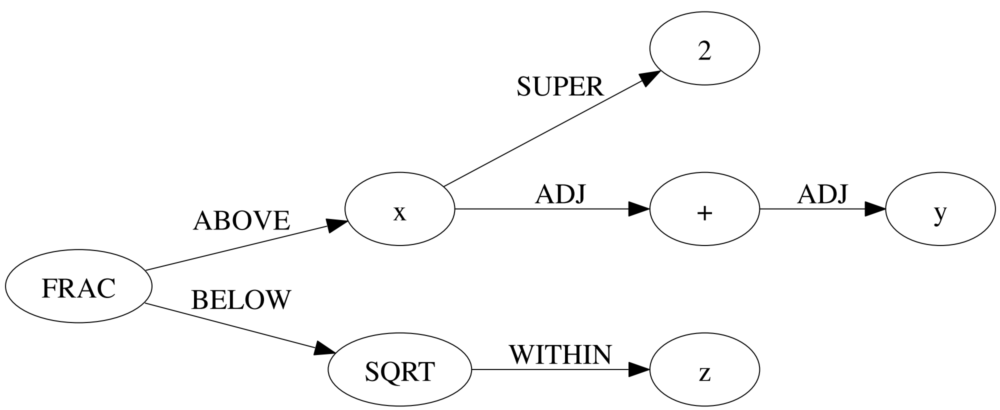
*Stalnaker, D. and Zanibbi, R. (2015) Math expression retrieval using an inverted index over symbol pairs. Proc. Document Recognition and Retrieval, San Francisco (to appear Feb. 2015).*

# Formula Representation and Indexing

**Formula Index**: inverted index from tuples → formulae

**Presentation MathML to SLT Conversion**: Depth-First Traversal

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




(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** |

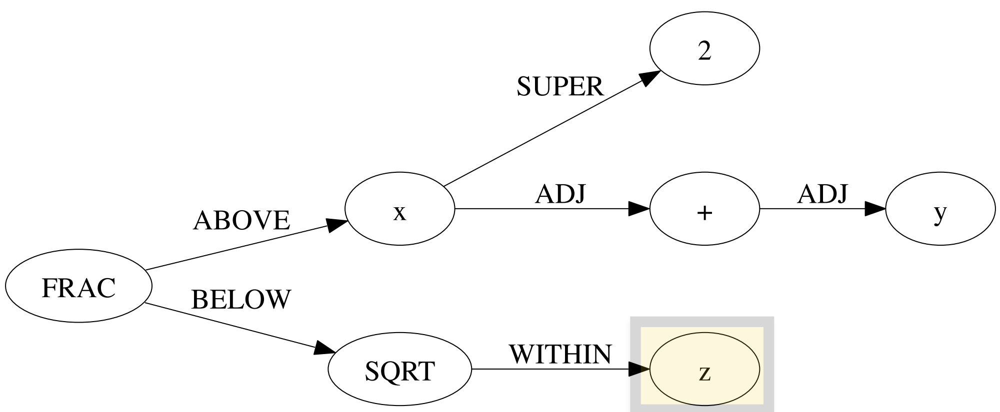
(b) Symbol Pair Tuples

# Formula Representation and Indexing

**Formula Index**: inverted index from tuples → formulae

**Presentation MathML to SLT Conversion**: Depth-First Traversal

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




(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** |

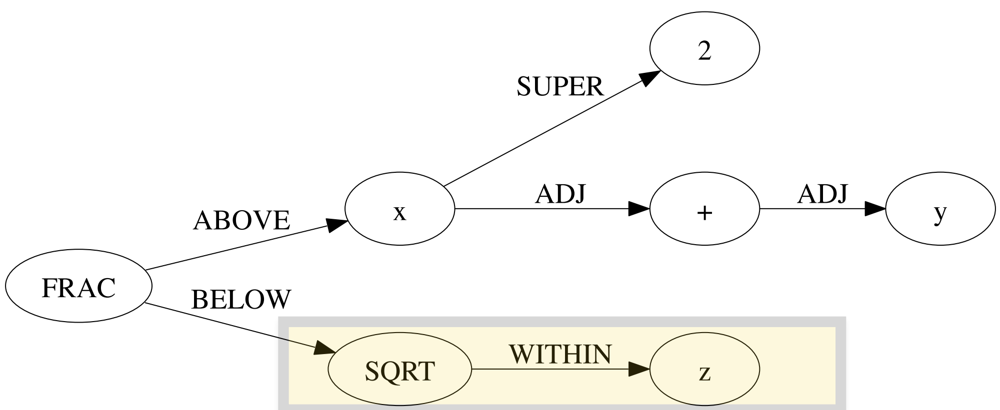
(b) Symbol Pair Tuples

# Formula Representation and Indexing

**Formula Index**: inverted index from tuples → formulae

**Presentation MathML to SLT Conversion**: Depth-First Traversal

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




(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** |

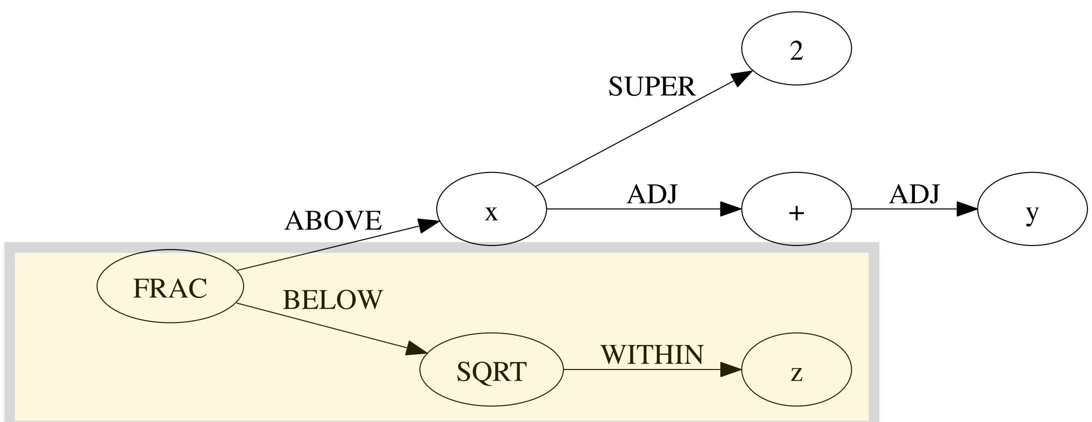
(b) Symbol Pair Tuples

# Formula Representation and Indexing

**Formula Index**: inverted index from tuples → formulae

**Presentation MathML to SLT Conversion**: Depth-First Traversal

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



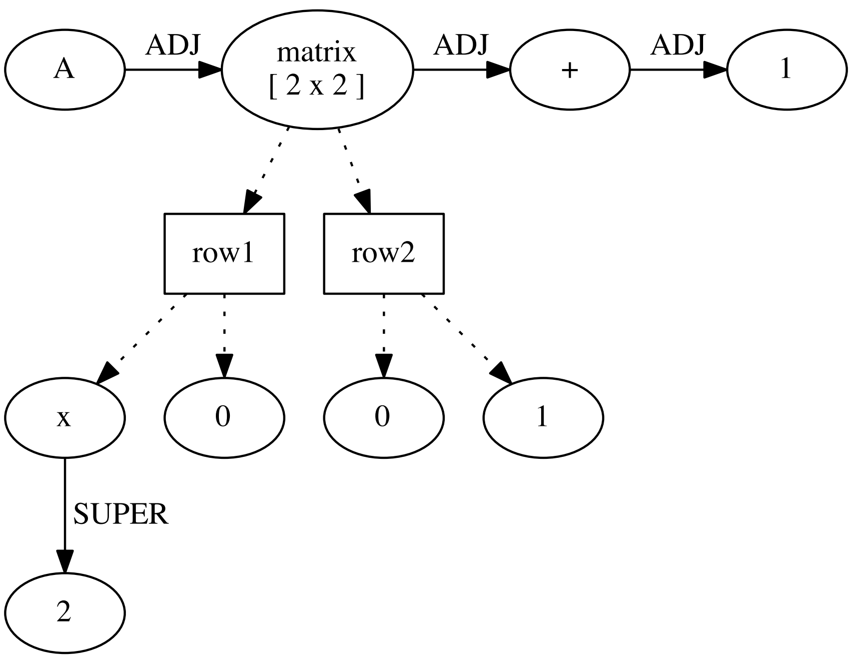

(a) Formula and Symbol Layout Tree

| 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

# Matrix Representation

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



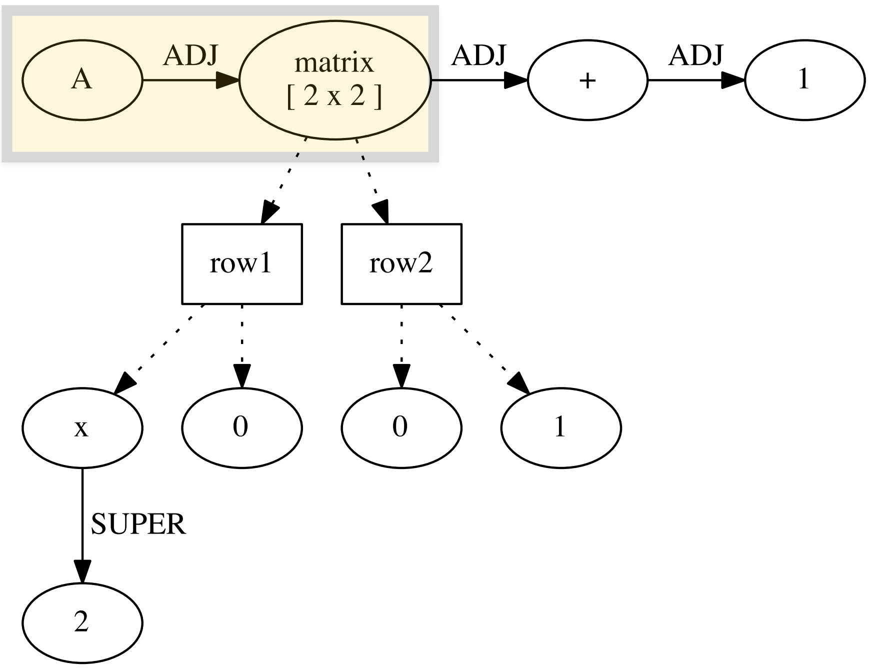

(a) Formula and Symbol Layout Tree

| **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

# Matrix Representation

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




(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 |

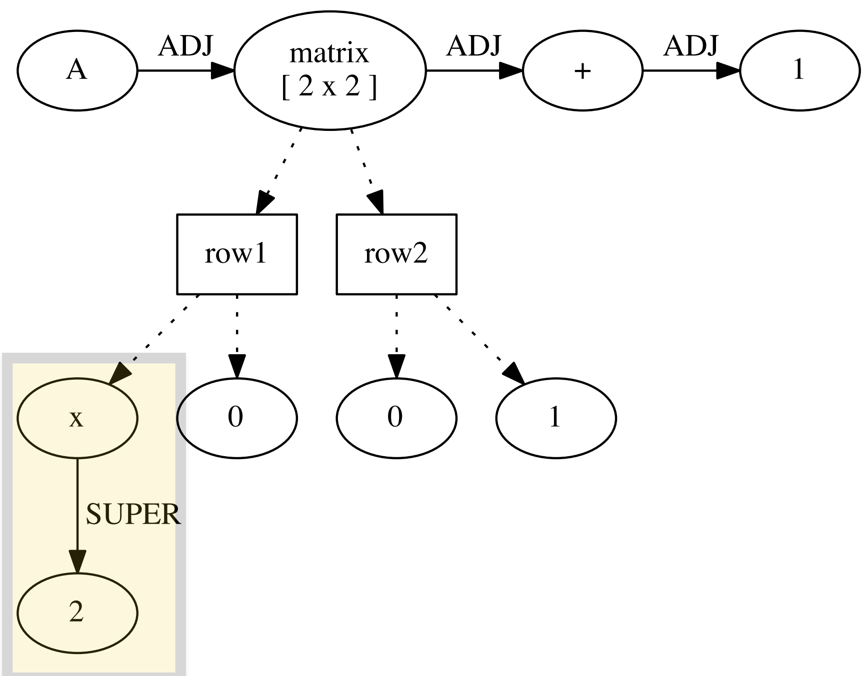
**Subexpressions**

| **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

# Matrix Representation

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




(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

# Wildcards

**Formula Query**: $\mathbb{P}[\boxed{X} \geq \boxed{t}] \leq \frac{\mathbf{E}[\boxed{X}]}{\boxed{t}}$

**Keyword**: Markov inequality

To handle query wildcards, two inverted indices group formula index entries with common parents/children ('Left' and 'Right' wildcard inverted indices)

Examples

(?i, 2, 1, 1): any symbol with superscript 2, e.g. $x^2$, $n^2$, $)^2$

(x, ?i, 1, 1): x with any superscripted symbol, e.g. $x^2$, $x^n$, $x^($

Wildcard-wildcard relationships are not retrieved

# Retrieval Model

## Text Score

Filter: ‘text’ for formulae replaced by formula identifiers

Lucene used for TF-IDF-based keyword retrieval; Lucene score used as *textScore*

## Formula Score

1) Look up query tuples in formula tuple and L/R wildcard indices to retrieve expressions

2) Sort by match count, keep top k = 1000 formulae

3) Wildcards: iteratively select unifications that match max. no. unmatched query tuples

4) For each document *d,* select formula with max. F = 2RP / (R + P ) ( *formulaScore )*

R: # matched query tuples P: # matched candidate tuples

4*) Multiple formulae: sum of top-1 score for each query expression in document, weighted by relative sizes of query expressions

$$m(d, e_1, ..., e_n) = \frac{|e_1|}{\sum\limits_{i=1...n} |e_i|} t_1(d, e_1) + \ldots + \frac{|e_n|}{\sum\limits_{i=1...n} |e_i|} t_1(d, e_n)$$

**Combined Score:** score(d) = $\alpha$ textScore(d) + (1- $\alpha$) formulaScore(d)

# Effect of Text Weight (Main Task)

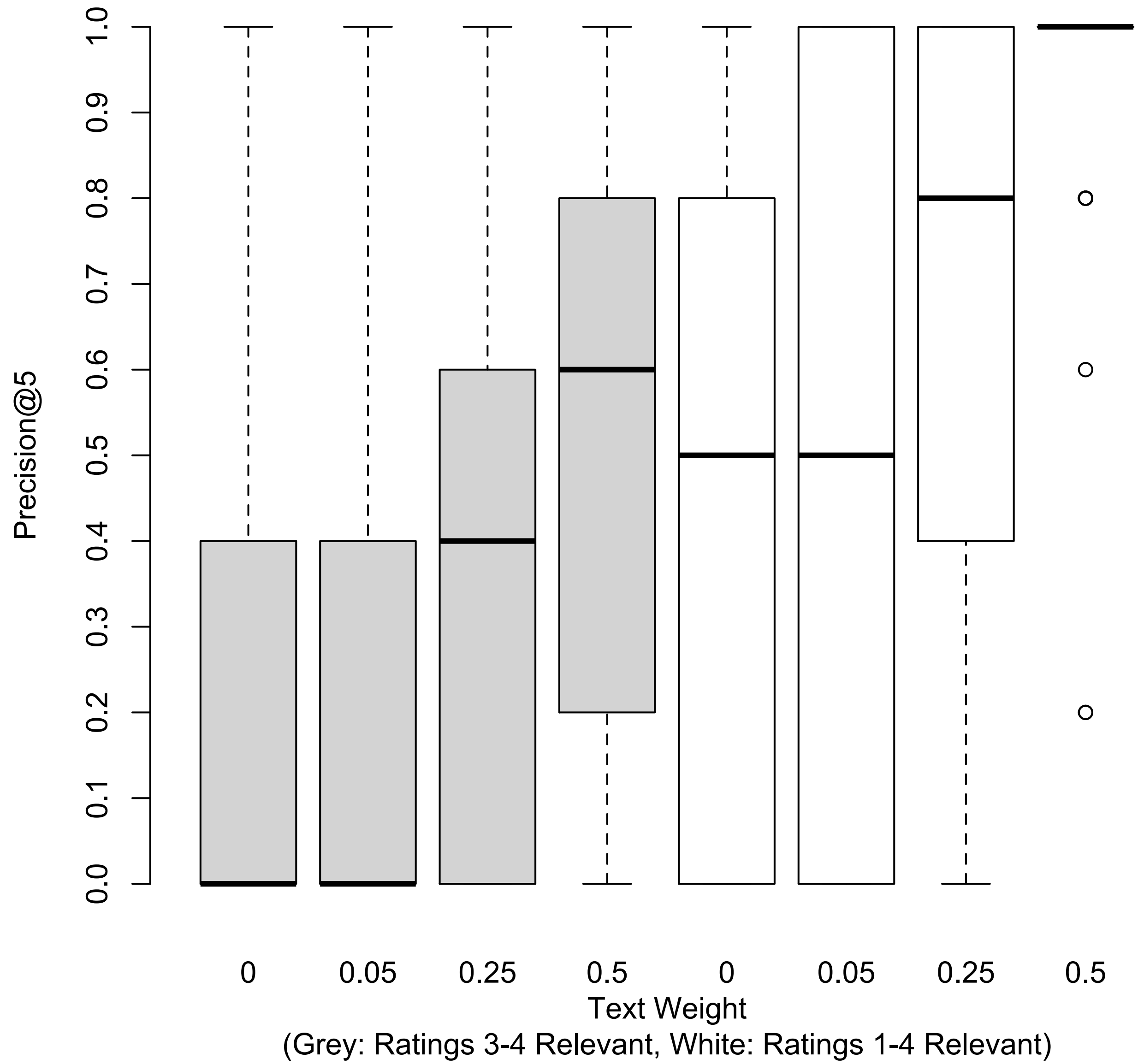



**Figure 6: Tangent Precision@5 (Main Task).**

# Retrieval Results

**Formula Query**: $\mathbb{P}[\boxed{X} \geq \boxed{t}] \leq \frac{\mathbf{E}[\boxed{X}]}{\boxed{t}}$

**Keyword**: Markov inequality

a) Math-2 #39

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

b) Wikipedia #49

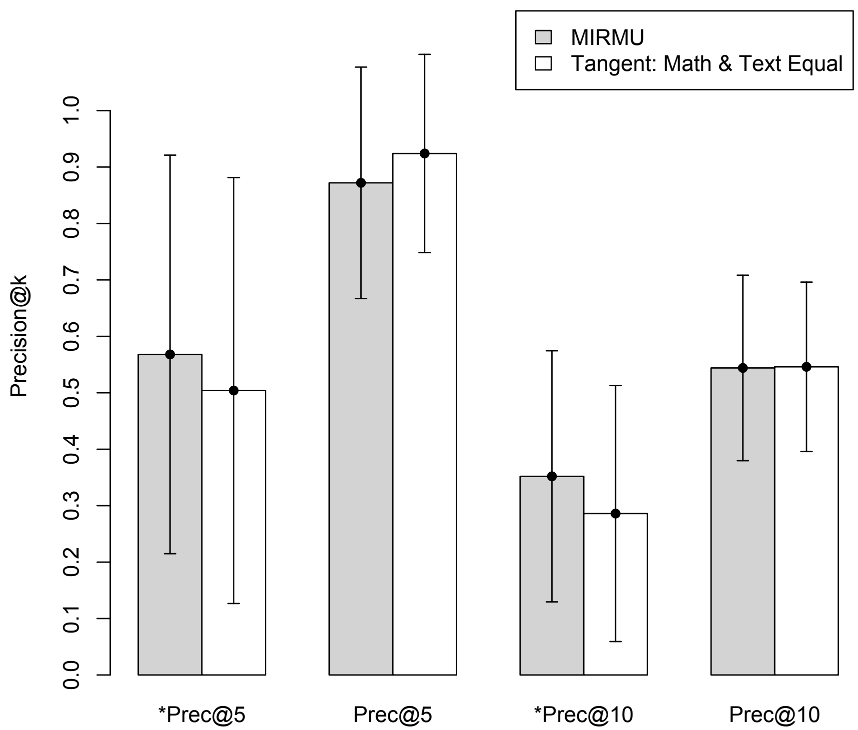



Figure 7: MIRMU System vs. Tangent (Main Task).

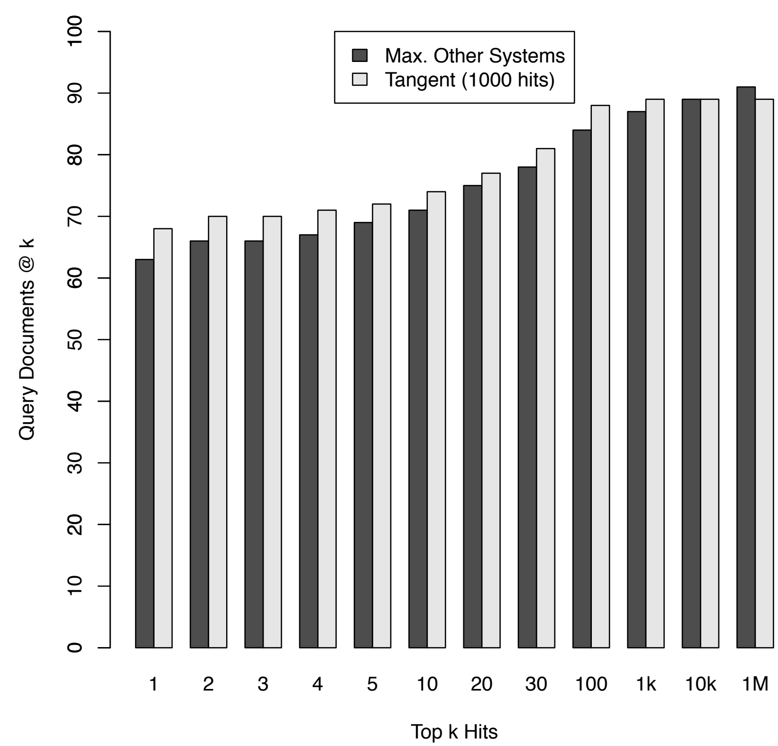



Figure 8: Wikipedia Math Search Subtask Results.

# System Performance

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

**Main task:** Nine EC2 instances used to index formulas, one for Lucene, and one instance to process queries and access text and formula engines (Python-based)

**Wikipedia subtask**: One instance 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.

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

# Thank You.

## Acknowledgements

David Stalnaker

Frank Wm. Tompa (Univ. Waterloo, Canada)

Math-2 Task Organizers:

Akiko Aizawa, Michael Kohlhase, Iadh Ounis

Moritz Schubotz

This material is based upon work supported by the National Science Foundation (USA) under Grant No. IIS-101681. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

dprl [DOCUMENT AND PATTERN RECOGNITION LAB]

NTCIR

# Sample Results

# Example Text + Math Query

## NTCIR11-Math2–47

**Formula Query**: $P_n = 2P_{n-1} + P_{n-2}$
**Keyword**: recurrence relation
**Keyword**: Pell number

**Result 1**

Example 3.3 An obvious example of Remark 3.2 is the Mersenne number $M_n = 2^n - 1$ $(n \geq 0)$, which satisfies the linear recurrence relation of order 2: $M_n = 3M_{n-1} - 2M_{n-2}$ ( with $M_0 = 0$ and $M_1 = 1$) and the non-homogeneous recurrence relation of order 1: $M_n = 2M_{n-1} + 1$ (with $M_0 = 0$). It is easy to check that sequence $M_n = \left(k^n - 1\right) / \left(k - 1\right)$ satisfies both the homogeneous recurrence relation of order 2, $M_n = \left(k + 1\right) M_{n-1} - kM_{n-2}$, and the non-homogeneous recurrence relation of order 1, $M_n = kM_{n-1} + 1$, where $M_0 = 0$ and $M_1 = 1$. Here, $M_n$ is the IRS with respect to $E_2 = \{3, -2\}$. Another example is Pell number sequence that satisfies both homogeneous recurrence relation $P_n = 2P_{n-1} + P_{n-2}$ and the non-homogeneous relation $\overline{P}_n = 2\overline{P}_{n-1} + \overline{P}_{n-2} + 1$, where $P_n = \overline{P}_n + 1/2$.

# Example Text + Math Query

**NTCIR11-Math2–47**

**Formula Query**: $P_n = 2P_{n-1} + P_{n-2}$
**Keyword**: recurrence relation
**Keyword**: Pell number

**Result 2**

The Fibonacci numbers $F_n$ satisfy the recurrence $F_{n+1} = F_n + F_{n-1}$ with $F_0 = F_1 = 1$ and $F_2 = 2$. The Lucas numbers $L_n$ satisfy the recurrence $L_{n+1} = L_n + L_{n-1}$ with $L_0 = 1, L_1 = 3$ and $L_2 = 4$. And the Pell numbers $P_n$ satisfy the recurrence $P_{n+1} = 2P_n + P_{n-1}$ with $P_0 = 1, P_1 = 2$ and $P_3 = 5$. Thus we can conclude the following result from Corollary 3.17.

# Example Text + Math Query

## NTCIR11-Math2–47

**Formula Query**: $P_n = 2P_{n-1} + P_{n-2}$
**Keyword**: recurrence relation
**Keyword**: Pell number

**Result 3**

(excerpt)

The Pell numbers $P_n$ are given by

$$P_0 = 0\ , \quad P_1 = 1 \text{ and } P_n = 2P_{n-1} + P_{n-2} \text{ for } n \geq 2.$$

It is easy to check that

$$P_n = \frac{\left(1+\sqrt{2}\right)^n - \left(1-\sqrt{2}\right)^n}{2\sqrt{2}}\ .$$

Hence for odd prime $p$, we have

$$P_p = \frac{\left(1+\sqrt{2}\right)^p - \left(1-\sqrt{2}\right)^p}{2\sqrt{2}} \equiv \frac{2\left(\sqrt{2}\right)^p}{2\sqrt{2}} = 2^{(p-1)/2} \equiv \left(\frac{2}{p}\right) \pmod{p} ..8$$

Define the $q$-Pell numbers $\mathscr{P}_n\left(q\right)$ and $\widehat{\mathscr{P}}_n\left(q\right)$ by

# Example Text + Math Query

## NTCIR11-Math2–47

**Formula Query**: $P_n = 2P_{n-1} + P_{n-2}$
**Keyword**: recurrence relation
**Keyword**: Pell number

**Result 4**

Let $m \geq 3$. The determinant of $THK\left(m, 2\right)$ is the $m$th Pell number $P_m$ where $P_1 = 1, P_2 = 2$, and $P_m = 2P_{m-1} + P_{m-2}$ for $m \geq 3$.

$$G_{k,\sigma}(y) = 1 - (1 + ky/\sigma)^{-1/k}$$ (NTCIR11-Math-92)

---

1. 0.99 $G_{k,\sigma}(y) = 1 - (1 + ky/\sigma)^{-1/k}$
2. 0.46 $G_{k,\sigma}(y) = 1 - e^{-y/\sigma}$
3. 0.34 $0.187859\ldots = \sum_{k=1}^{\infty}(-1)^k(k^{1/k} - 1) = \sum_{k=1}^{\infty}\left((2k)^{1/(2k)} - (2k-1)^{1/(2k-1)}\right).$
4. 0.33 $a_{\mathrm{dual}}(Z) = 2Z^d \left(\frac{1+Z}{2}\right)^A q_{\mathrm{dual}}(1 - (Z + Z^{-1})/2)$
5. 0.33 $a_{\mathrm{prim}}(Z) = 2Z^d \left(\frac{1+Z}{2}\right)^A q_{\mathrm{prim}}(1 - (Z + Z^{-1})/2)$

---

$$K_{\boxed{x0}}^{\boxed{x1}}(k) := T^*(k^\times)/(a \otimes (1-a)) \qquad \text{(NTCIR11-Math-72)}$$

---

1. 0.95 $K_*^M(k) := T^*(k^\times)/(a \otimes (1-a))$

2. 0.95 $K_*^M(k) := T^*(k^\times)/(a \otimes (1-a))$

3. 0.50 $K_*^M(F) := T^* F^\times/(a \otimes (1-a)),$

4. 0.41 $K_2(k) = k^\times \otimes_{\mathbf{Z}} k^\times / \langle a \otimes (1-a) \mid a \neq 0, 1 \rangle.$

5. 0.33 $T(n) = T(1)\left(B + \frac{1}{n}(1-B)\right)$

---

$$\frac{\partial L}{\partial q_i} = \boxed{x0}\,\frac{\partial\,\boxed{x1}}{\partial\,\boxed{x2}}.$$ (NTCIR11-Math-86)

---

1. 0.72 $M_i = \frac{v_i}{a} = \frac{1}{a}\frac{\partial \Phi}{\partial x_i}.$

2. 0.66 $\frac{\partial L}{\partial q_i} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial L}{\partial \dot{q}_i}.$

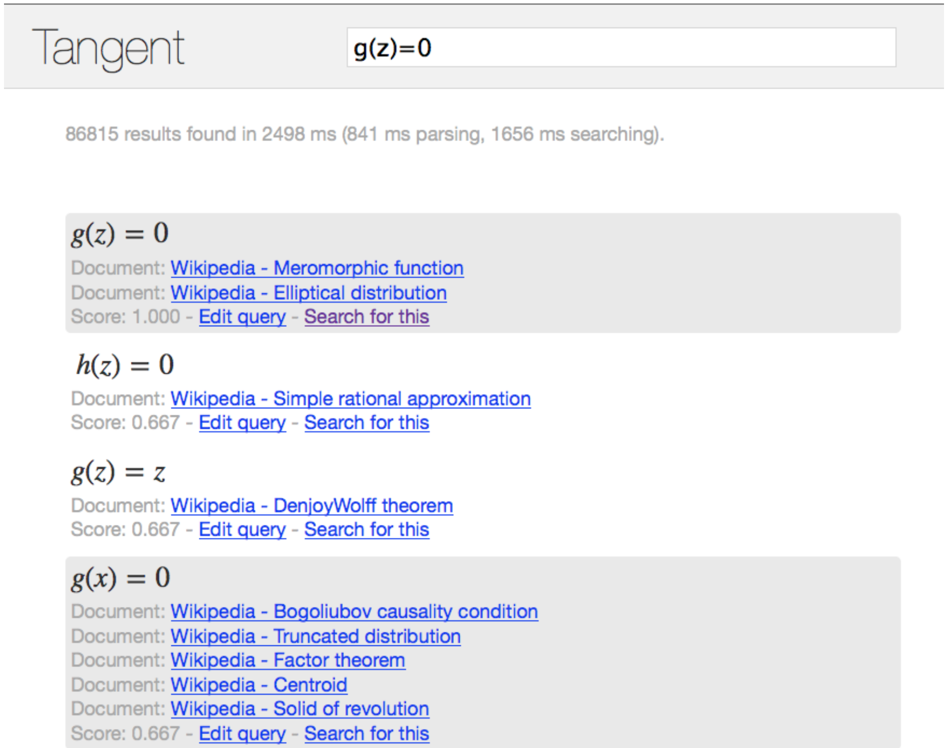
3. 0.61 $\frac{\partial L(t, y(t), \dot{y}(t))}{\partial y} = \frac{d}{dt}\frac{\partial L(t, y(t), \dot{y}(t))}{\partial \dot{y}}.$

4. 0.61 $\mathbf{F}_i = -\nabla V \Rightarrow Q_j = -\sum_{i=1}^{n} \nabla V \cdot \frac{\partial \mathbf{r}_i}{\partial q_j} = -\frac{\partial V}{\partial q_j}.$

5. 0.60 $\frac{dF}{dt} = \sum_i \frac{\partial F(T, V, N)}{\partial N_i}\frac{dN_i}{dt} = \sum_i \mu_i \frac{dN_i}{dt} = -VRT\sum_r (\ln w_r^+ - \ln w_r^-)(w_r^+ - w_r^-) \leq 0$

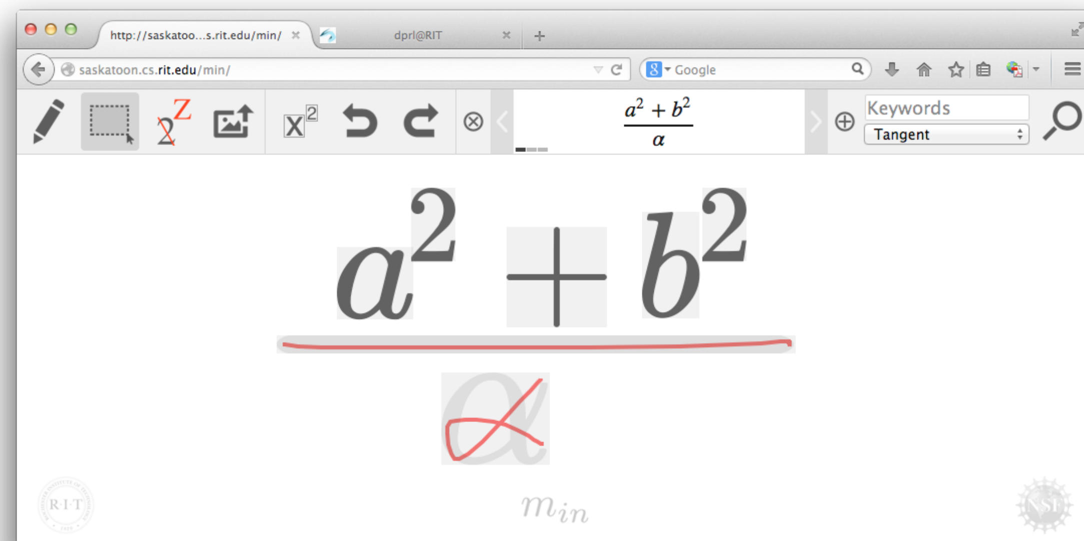
---

# Source Code and Demos




## CODE

www.cs.rit.edu/~dprl/Software.html




## DEMOS

saskatoon.cs.rit.edu/tangent

saskatoon.cs.rit.edu/min