DOCUMENT AND PATTERN RECOGNITION LAB

R·I·T

Creating User-Friendly Math Search Engines

Richard Zanibbi

Document and Pattern Recognition Laboratory

Department of Computer Science

Rochester Institute of Technology, NY, USA

Imaging/Document Processing Seminar

IUT de Nantes, June 30, 2014

IUT NANTES
Université de Nantes

What is Math Notation?

Mathematical Notation

Mathematical notation may represent:

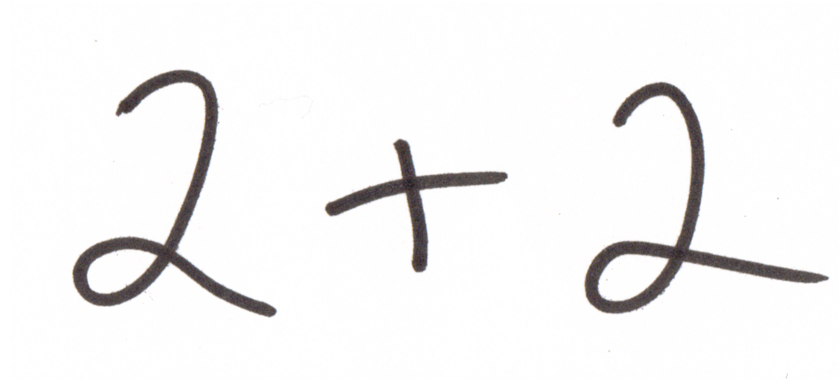
- quantities or values (e.g. real numbers, boolean vars.)
- structures (e.g. matrices, graphs, sets)
- operations on quantities and structures (e.g. +, ∪, ¬)
- relationships (e.g. $x = 2$, $y > 1$)

History of Math Notation: see Cajori, "History of Mathematical Notations" (2 Vols.), 1929

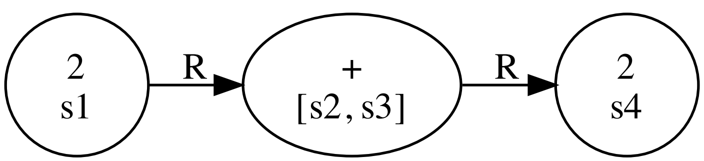
A *natural* visual language: adapted by authors for their own purposes.

e.g. Consider definitions for 'f' or 'x' - *dialects*

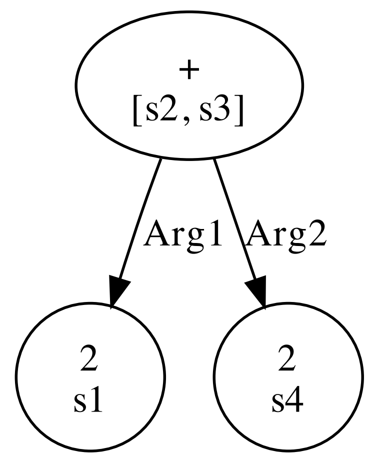
Structure in Math Expressions



Symbol Layout Tree
(Appearance)



Operator Tree
(Math Syntax)



Primitives: 4 pen strokes (2, →, ↓, 2)

Symbols: 3 (2, +, 2)

Why Do This?

(i.e. math recognition and retrieval research)

Survey:

R. Zanibbi and D. Blostein (2012) Recognition and Retrieval of Mathematical Expressions, Int'l. Journal on Document Analysis and Recognition 15(4): 331-357.

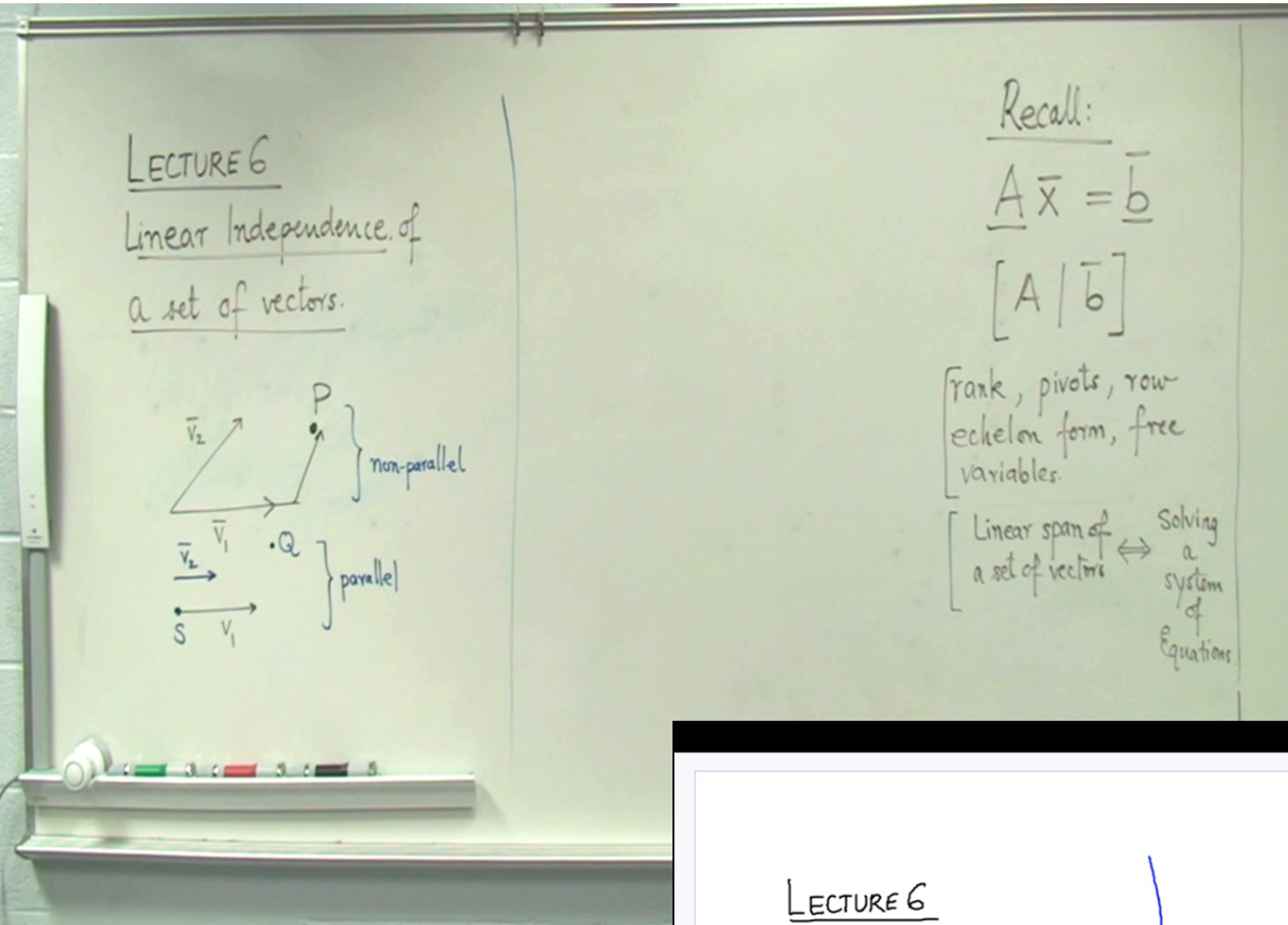
I. Social Motivation

Mathematical Literacy

Make it easier for persons of all ages and walks of life to create and find mathematical material.

Initial Emphasis: Non-experts and children

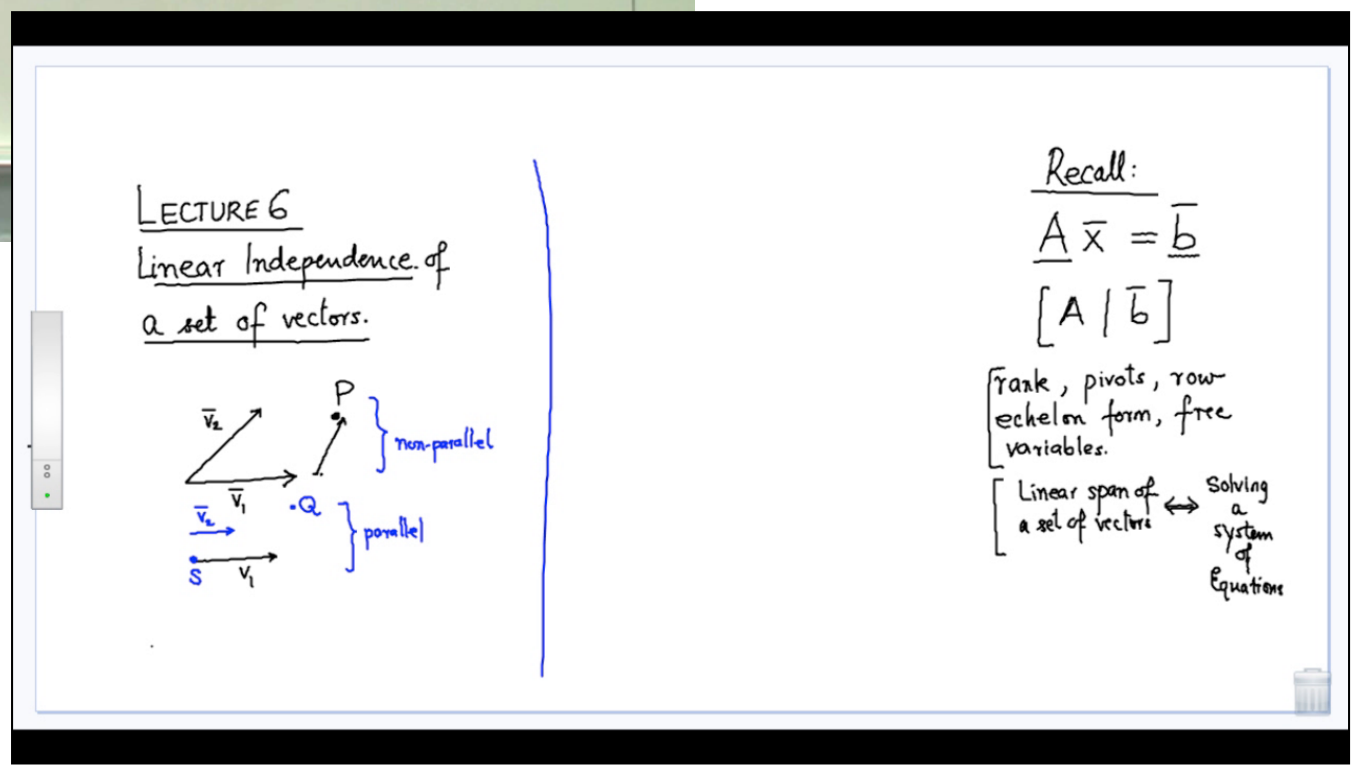
(Zhao et al., 2008): Mathematicians/mathematical experts often know names for common formulas/metrics/theories, use these in web searches - current tools adequate?



AccessMath

Project Set-up: Video + Mimio

(w. Stephanie Ludi, Roger Gaborksi, Anurag Agarwal)



Goal: iPad app for low-vision students using image and audio queries to search math lectures

II. Retrieval Motivation

Structured and Image-Based Retrieval

- Given hierarchical structure, formulae a good domain for graph-based retrieval research
- Many online expressions are images - opportunity to study image-based retrieval in a constrained setting (vs. ‘natural scenes’)
- If we improve math search, can we improve retrieval for other notations (e.g. chemical diagrams)?

Studied since early 2000’s (Miller and Youssef - DLMF)

III. Recognition Motivation

Math as Structural Pattern Recognition Problem

Recognition involves central PR problems:

- Classification (What), Segmentation (Where), Parsing (How objects are structured)
- Optimizing the interaction: Machine Learning

Inputs relatively small

Output language(s) well-constrained

But *non-trivial* - this is *visual* Natural Language Processing

Studied since the late 1960's (Anderson's PhD (MIT))

[DOCUMENT AND PATTERN RECOGNITION LAB]

$\mathrm{m_{in}}$: A Multimodal Math Search Interface

C. Sasarak, K. Hart, R. Pospesel, D. Stalnaker, L. Hu, R. LiVolsi, S. Zhu, and R. Zanibbi. (2012) min: A Multimodal Web Interface for Math Search. *Symp. Human-Computer Interaction and Information Retrieval*, Cambridge, MA (online, 4pp).

Université de Nantes

Existing Tools for Math Search

Existing Search Engines

Designed for text; Term Frequency-Inverse Document Frequency (TF-IDF) of words provides basis for many retrieval systems + statistics (e.g. n-grams), word proximity, etc.

Structure represented in string languages, e.g. `1/2 as \frac{1}{2}` in LaTeX

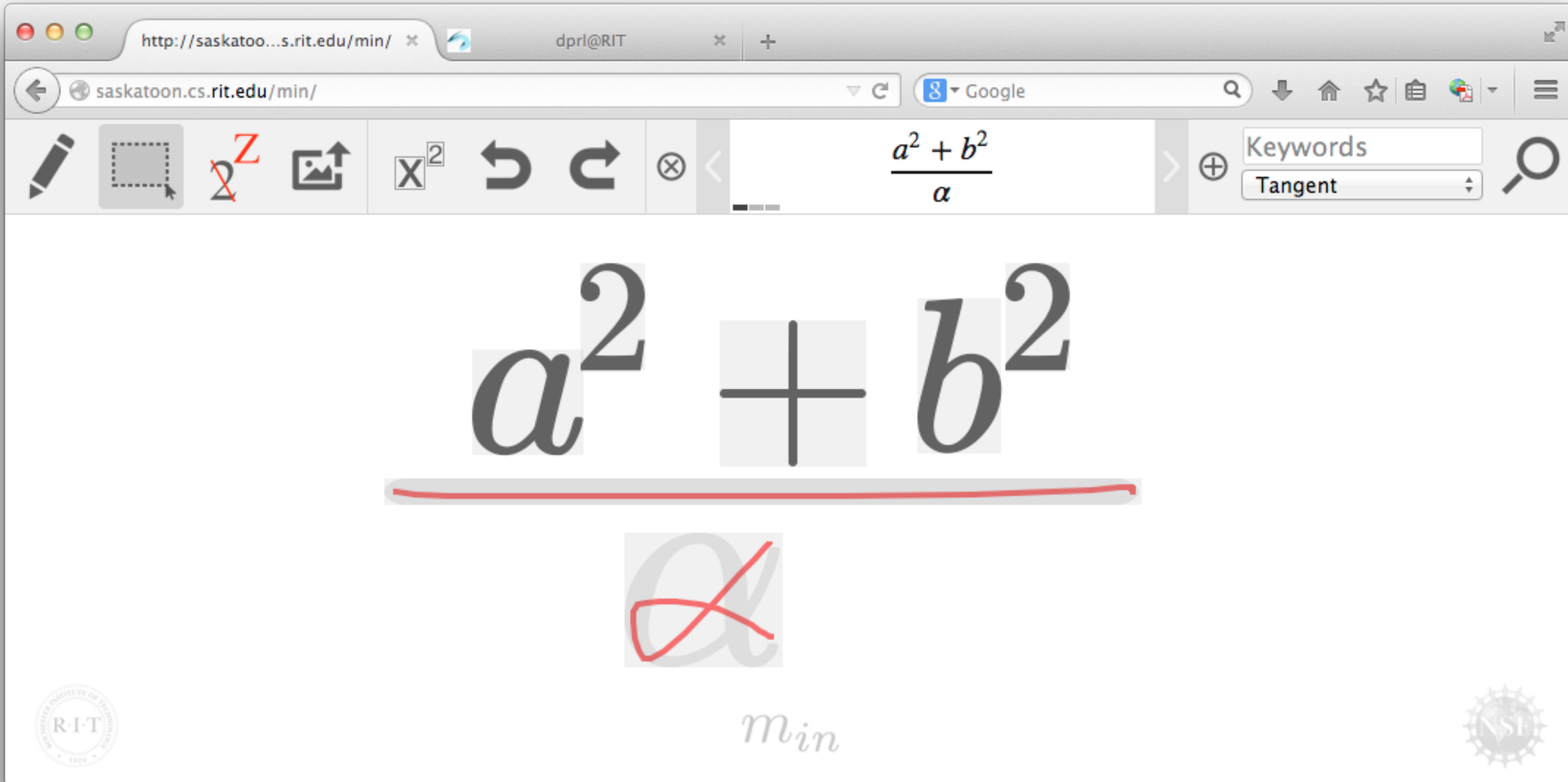
Limitations for Math Search with Current Engines

Many are unfamiliar with string languages used to represent symbols (e.g. greek letters) and structures in math

Making structural comparisons directly on "flattened" representations introduces problems:

- String-based difference measurements for what is a tree-based (i.e. hierarchical) structure leads to very coarse structural matching (e.g. missing sub-expressions between a query and candidate expression)

Tree-based distances expensive (e.g. EMERS (Sain et. al) is $O(n^4)$ - in general, edit distance on unordered trees is NP-complete)



m_{in} search interface

- Mouse/touch, keyboard, and image input
- Keywords + LaTeX sent to chosen search engine
- http://saskatoon.cs.rit.edu/min code: https://github.com/DPRL

[DOCUMENT AND PATTERN RECOGNITION LAB]

R·I·T

Preliminary User Study for m_{in}

Del Valle Wangari, K., Zanibbi, R. and Agarwal, A. (2014) [Discovering real-world use cases for a multimodal math search interface.] Proc. ACM SIGIR, Gold Coast, Australia (to appear, July 2014).

Université de Nantes

Study Design

Questions:

1. Does using m_{in} change search behavior for mathematical non-experts?

2. Can users identify real-world scenarios for using a multimodal math search interface?

Search Tasks

Designed four search tasks with Prof. Agarwal who teaches Math at RIT, in "peer-assist" style.

Task 1: *Your classmate is having difficulty recognizing polynomials. Find one or more resources to help explain to your classmate why* $x^2 - 7x + 2$ *is a polynomial and why* $\frac{x^2-7x+2}{x+2}$ *is not a polynomial.*

Task 2: *Your classmate has heard of Pascal's triangle but doesn't understand how it relates to math. Find one or more resources to help explain to your classmate how the equation* $(x+y)^2 = x^2 + 2xy + y^2$ *relates to Pascal's Triangle.*

Task 3: *Your classmate is struggling with binomial coefficients. Find one or more resources to help explain to your classmate how to find the value of* $\binom{4}{2}$.

Task 4: *Your classmate is having trouble understanding the prime counting function. Find resources that help explain why* $\pi(2) = 1$.

Search Tool Conditions

All participants did the following, in order:

1. 'Free' - choice of textbooks, notes, websites and online search.
2. Online search *without* m_{in}

   (demonstration; brief set of questions about min)

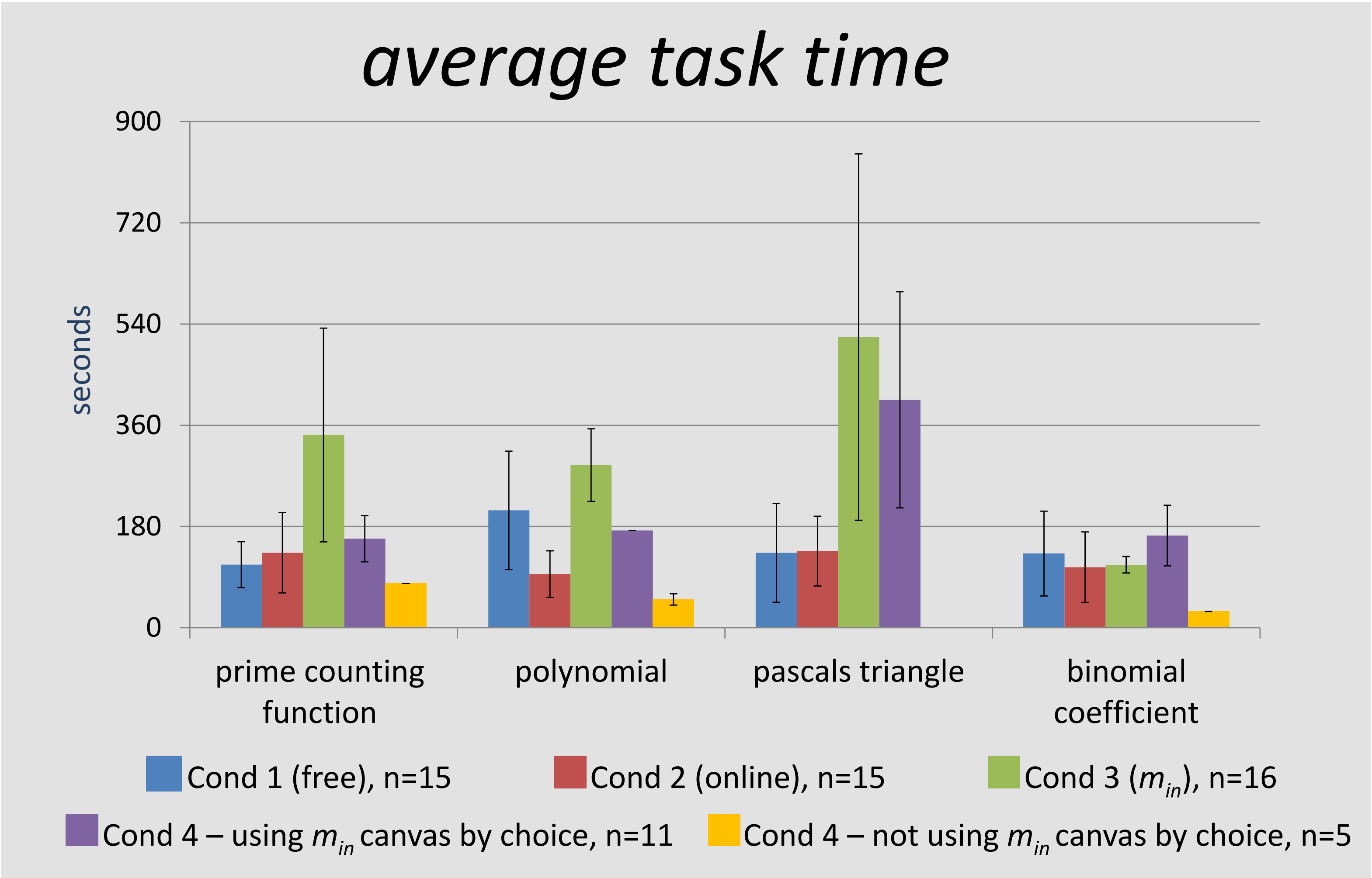
3. Online search with m_{in}
4. Online search with m_{in} optional

**Search tasks counter-balanced to avoid order effects*

Results

- The 16 participants were 18 or older, currently enrolled in a first- or second-year college math course, self-rated as Beginner or Intermediate level in math knowledge, and self- rated as Comfortable or Very Comfortable using the internet. All were students in College of Science or College of Computing at RIT.
- Sessions were videotaped in a quiet room.
- **No participant used LaTeX or a structure editor, though some knew of these.**
- **12/16 (75%) of participants could identify scenarios where they could use min**; studying for math tests (in particular, working with Calculus, integrals, complex math problems and expressions with lots of Greek letters), taking notes, collaborating with remote students on assignments, and exporting expressions as image files or LaTeX for use in reports.

Search Task Times and Success



Self-reported success rates were nearly identical for m_{in} vs. non-m_{in} conditions.

Results, Continued

Despite the longer entry/search times, 11/16 participants (69%) reported that m_{in} made it easy to enter expressions.

"Like 4 choose 2 – that's really hard to 'write' but it knew what I meant and it accurately translated what I was trying to say to it."

Search behavior: condition 2 (online search) - no expressions entered; condition 3 (m_{in}) expressions used by all participants, and 10/11 in condition 4 using m_{in}.

From videos, long tasks times with m_{in} largely from recognizer errors, and participant errors interpreting recognition results. *(recognition feedback modified)*

Study Conclusions

Questions:

1. Does using m_{in} change search behavior for mathematical non-experts?

Use of expressions in queries was increased.

2. Can users identify real-world scenarios for using a multimodal math search interface?

Yes (studying; writing; course work; collaboration)

How Important is it to Render Math in Search Hits?

Reichenbach, M., Agarwal, A. and Zanibbi, R. (2014) Rendering expressions to improve accuracy of relevance assessment for math search. Proc. ACM SIGIR, Gold Coast, Australia (to appear, July 2014).

An Example

Top: Google search hit Bottom: with rendered expression

Linear Algebra WebNotes. Part 3.

The first indexes form a permutation of the set {**1**,2,3,4}. .... A-**1** = (**1/det(A)**) **adj(A)**. **Proof**. Indeed, if A is invertible then by the third theorem about determinants ...

Linear Algebra WebNotes. Part 3.

The first indexes form a permutation of the set {**1**,2,3,4}. ....

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$$

Proof. Indeed, if A is invertible then by the third theorem about determinants ...

Questions

1. Does properly formatting expressions increase accuracy in *relevance assessment* for search hits?

2. Does properly formatting expressions decrease time needed to assess relevance?

Search Tasks

‘Hits’ were taken from Google search results (control), using LaTeX for math in the queries.

‘Relevant’ hits contained both a portion of the query expression and the accompanying keyword or semantically equivalent term. Five ‘relevant’ and five ‘irrelevant’ hits were selected for each task.

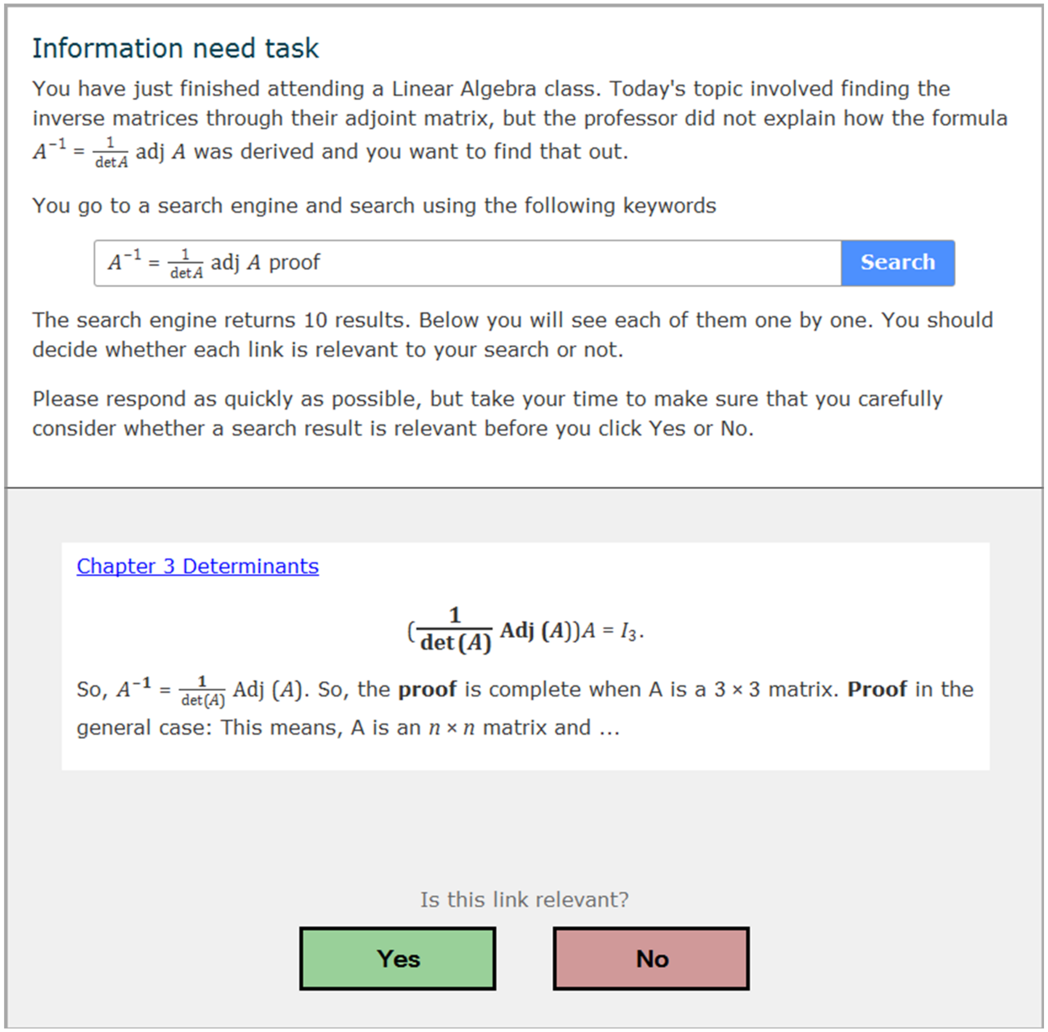
Informational Need

You have just finished attending a Linear Algebra class. Today’s topic involved finding the inverse matrices through their adjoint matrix, but the professor did not explain how the formula $A^{-1} = \frac{1}{\det A} \cdot \operatorname{adj} A$ was derived and you want to find that out. You go to a math search engine and search for ’$A^{-1} = \frac{1}{\det A} \cdot \operatorname{adj} A$ proof.’

Resource Need

Your friend is having trouble understanding derivatives of polynomials and you have agreed to help him. You need to be prepared to explain that to him so you want to find tutorials showing $\frac{d}{dx} ax^b = abx^{b-1}$. You go to a math search engine and search for ’$\frac{d}{dx} ax^b = abx^{b-1}$ tutorial.’

Study Design



Study: Human evaluation for different presentation styles of search hits containing mathematical expressions.

Participants: 38 college students having taken at least 2 college-level math courses.

Protocol: Familiarization task, two experimental tasks, exit questionnaire.

Participants timed as they evaluated search hits for tasks one-at-a-time in a web interface (at left) in Control or Rendered condition *(Guan & Cutrell SIGCHI 2007)*

Presentation of queries counter-balanced to avoid order effects

Results

Table 1: Relevance assessment accuracies and response times. Task 1 required locating a proof; Task 2 required locating a tutorial. Groups: Control $n = 19$; Rendered $n = 19$; Total $n = 76$

		Accuracy (%)		Response Time (s)	
Task	Summary	μ	σ	μ	σ
1	Control	69.47	13.11	12.58	4.55
	Rendered	83.10	12.01	14.06	5.11
2	Control	69.71	20.78	12.39	4.79
	Rendered	80.00	15.63	12.70	4.35
1 & 2	(Total)	75.57	16.60	12.93	4.66

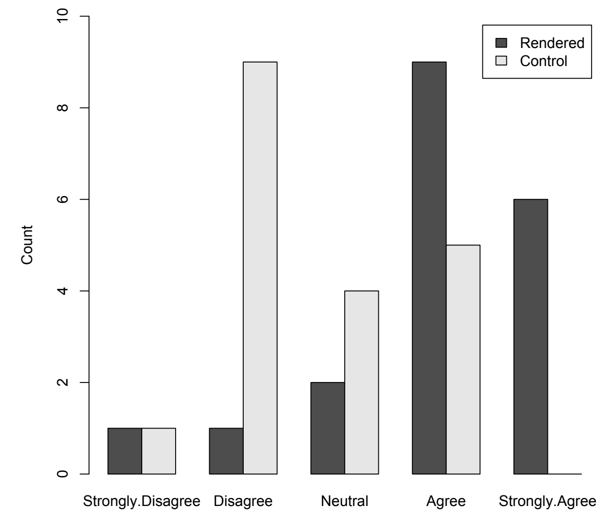


Figure 3: Participant responses from the Rendered and Control summary style conditions for the statement "I had no problems reading the results presented."

Assessment accuracy changed by summary style ($F(1,36) = 8.73$, $p < 0.01$) - rendered condition mean 17.18% higher.

Rendered style reported easier to read ($p < 0.05$ Mann-Whitney Ind. Samples Test)

Small negative correlation between time and accuracy in control condition ($r = -0.114$, $p < 0.05$ (Pearson Corr.)).

No effect for rendering condition on response time ($p > 0.05$)

Study Conclusions

1. Does properly formatting expressions increase accuracy in relevance assessment?

Confirmed by results; 17.18% increase in study

2. Does properly formatting expressions decrease the time needed to assess relevance?

Surprisingly, not observed. Possible that normal speed-accuracy trade-off violated due to low discriminability (negative correlation for control).

[DOCUMENT AND PATTERN RECOGNITION LAB]

Tangent: Query-by-Expression via Matching Symbol Pairs

D. Stalnaker (2013) Math Expression Retrieval Using Symbol Pairs in Layout Trees. Master's Thesis, Rochester Institute of Technology (Computer Science), NY, USA (August 2013).

Université de Nantes

Query-by-Expression

Definition: Retrieving mathematical expressions using a math expression as a query

Existing Approaches

- Text-Based: linearize expression (e.g. as LaTeX) and use existing TF-IDF methods (e.g. Lucene) (Miller & Youssef, 2003)
- Tree-Based: Tree edit-distance (Kamali et al., 2013); Substitution trees (Kohlhase and Sucan, 2006); Local structural techniques (Nguyen et al., 2012; Hiroya and Saito, 2013)

Tangent (Stalnaker, 2013)

A 'Local Tree-Based' Method

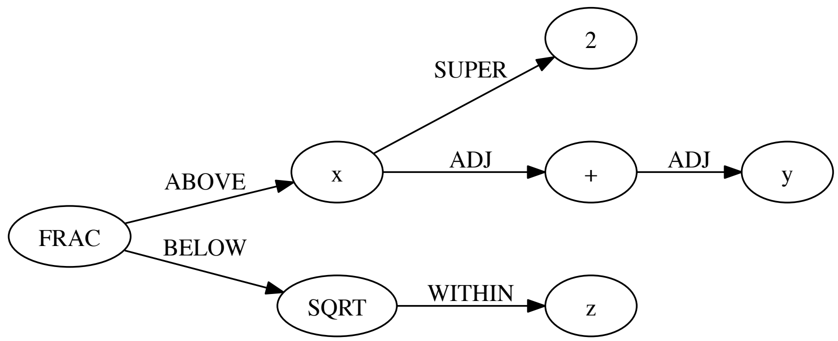
Main Ideas:

- Use symbol pairs to capture local and global expression structure.
  - Using *specific* symbols (no 'wildcards')
- Store pairs in an *inverted index*, commonly used for fast text retrieval to map words to documents containing them.

Indexing Expressions

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

(a) Expression



(b) 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
x	y	2	0
x	+	1	0
+	y	1	0
SQRT	z	1	0

(c) Symbol Pair Tuples

Expressions in LaTeX or MathML format converted to a Symbol Layout Tree, and then a list of quartuples.

Inverted index from quartuples to list of matching expressions is created.

Retrieval

1. Convert query expression to a list of tuples *(symparent, symchild, dist, ver. offset)*
2. Lookup each quartuple in the inverted index. Add entries to a hash table using expression identifiers as keys.
3. Rank matched expressions using recall (% query tuple matches) and precision (% candidate tuple matches), e.g. by F-measure, $2RP/(R+P)$

Questions

1. Can we obtain more relevant results using Tangent than a conventional TF-IDF system used to index math?
2. Is Tangent fast enough for use in real-time?

Study Design

20 students and professors participated in the experiment. English Wikipedia expression set (476,238 expressions).

Search results were obtained for:

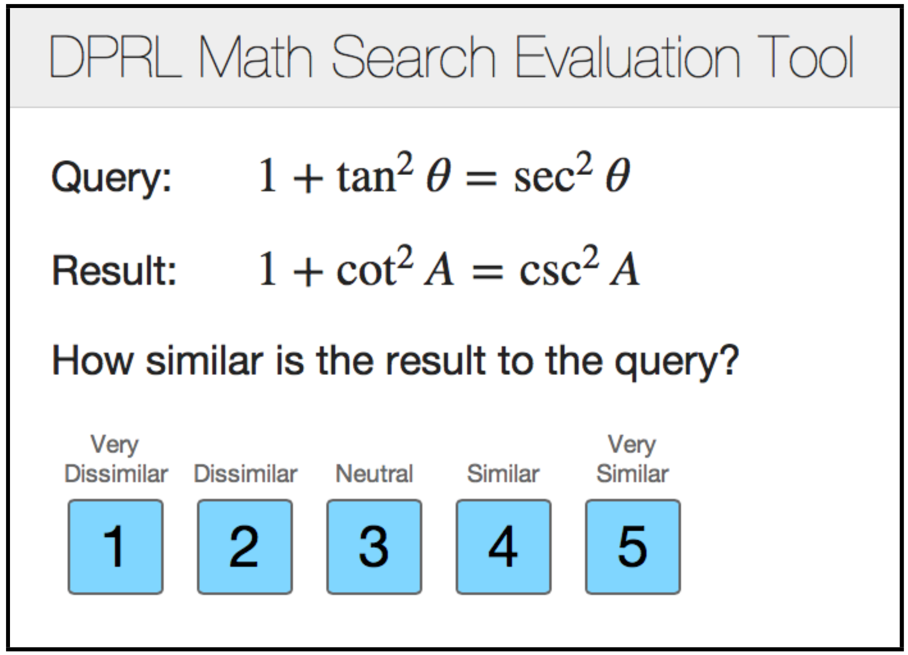
1) Lucene-based system (Zanibbi&Yuan, 2011)

2.) Three Tangent variations (ranking fns)

Search hits were pooled. Queries and their hits were presented in a random order, one-at-a-time.

Study Design

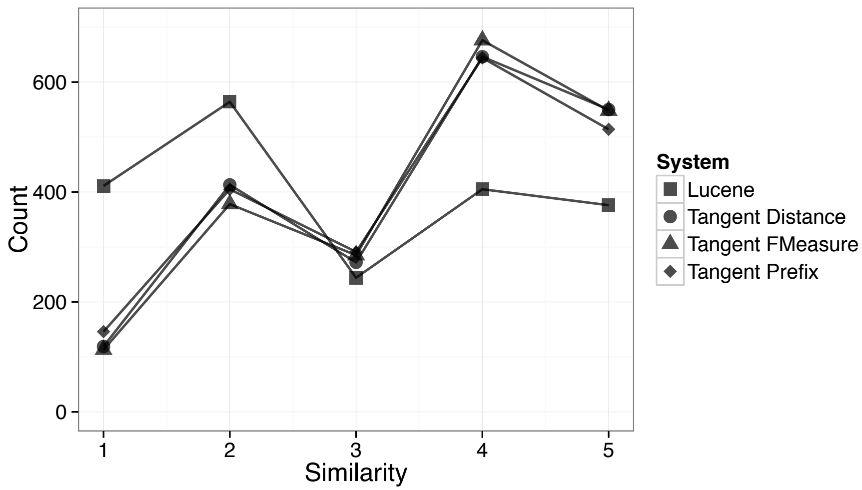
Evaluation Interface



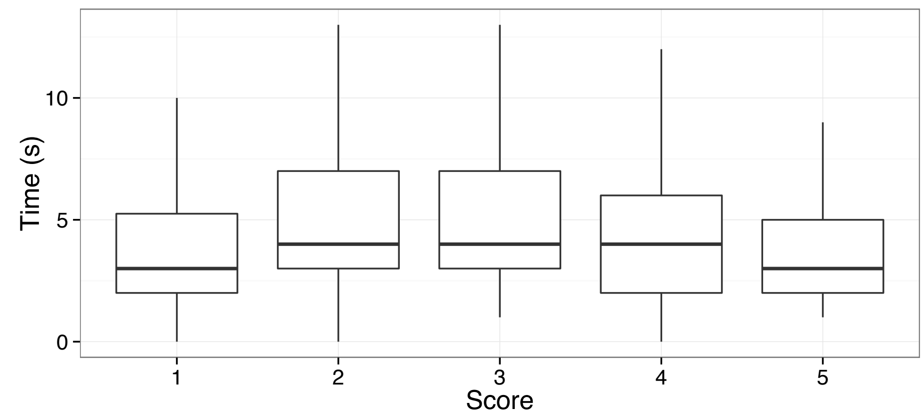
10 Queries

No.	Query	No.	Query
1.	$\widetilde{\rho}$	6.	$\int_a^b f(x)\,dx = F(b) - F(a).$
2.	$\bar{u} = (x, y, z)$	7.	$\left(1/6,\ \sqrt{1/28},\ -\sqrt{12/7},\ 0,\ 0,\ 0,\ 0,\ 0\right)$
3.	$1 + \tan^2\theta = \sec^2\theta$	8.	$\sum_{i=m}^{n} a_i = a_m + a_{m+1} + a_{m+2} + \cdots + a_{n-1} + a_n.$
4.	$cos(\theta_E) = e^{-TR/T_1}$	9.	$f(x;\mu,c) = \sqrt{\frac{c}{2\pi}}\,\frac{e^{-\frac{c}{2(x-\mu)}}}{(x-\mu)^{3/2}}$
5.	$a = g\frac{m_1-m_2}{m_1+m_2}$	10.	$D4\sigma = 4\sigma = 4\sqrt{\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)(x-\bar{x})^2\,dx\,dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\,dx\,dy}}$

Results



(a) Ratings by system



(b) Response times by rating score

Discretized Likert similarity scores into 'similar' (4,5) and 'dissimilar' (1-3). Significant difference between similarity scores (two-way ANOVA system vs. query for Precision@10; $p < 2.2 * 10^{-16}$)

- Prec@1,10: Lucene: (60%, 39%) vs. Tangent: (99%, 60%)

Response time for Lucene-based results slightly slower (mean of 5.84 vs. 5.29 seconds)

Sample Search Results

Table 3: Top 10 results for $1 + \tan^2\theta = \sec^2\theta$ for Lucene and Tangent (F-Measure Ranking)

Rank	Lucene	Tangent	Rank	Lucene	Tangent
1.	$1+\tan^2\theta=\sec^2\theta$	$1+\tan^2\theta=\sec^2\theta$	6.	$\sin^2\theta+\cos^2\theta=1$	$\sqrt{1+\tan^2\theta_o}$
2.	$\tan^2\theta+1=\sec^2\theta$	$1+\tan^2 y=\sec^2 y$	7.	$\cos^2\theta+\sin^2\theta=1$	$\pm\sqrt{1+\tan^2\theta}$
3.	$\sec^2\theta=1+\tan^2\theta$	$\frac{d}{d\theta}\tan\theta=\sec^2\theta$	8.	$1+\cot^2\theta=\csc^2\theta$	$1+\cot^2 A=\csc^2 A$
4.	$1+\tan^2\theta=\sec^2\theta$ and $1+\cot^2\theta=\csc^2\theta.$	$1+\cot^2\theta=\csc^2\theta$	9.	$\cot^2\theta+\ 1=\csc^2\theta$	$1+\cot^2 y=\csc^2 y$
5.	$\cos^2\theta+\sin^2\theta=1$,	$\sec^2\theta=1+\tan^2\theta$	10.	$x=r\cos\theta=2a\sin^2\theta=\frac{2a\tan^2\theta}{\sec^2\theta}=\frac{2at^2}{1+t^2}$	$\tan^2\theta+1=\sec^2\theta$

$$\text{Query 2: } \bar{u}=(x,y,z)$$

Rank	Lucene	Tangent F-Measure	Tangent Distance	Tangent Prefix
1	$f(\bar{u})=f(x,y,z)$	$\bar{u}=(x,y,z)$	$\bar{u}=(x,y,z)$	$\bar{u}=(x,y,z)$
2	$=\mathrm{R}(z,dt)\vert x,y,z\rangle$	$u=(x,y,z)$	$u=(x,y,z)$	$u=(x,y,z)$
3	$(x\vee y)(\bar{x}\vee z)(y\vee z)=(x\vee y)(\bar{x}\vee z)$	$\mathbf{v}=(x,y,z)$	$\mathbf{v}=(x,y,z)$	$\mathbf{v}=(x,y,z)$
4	$\bar{u}=(x,y,z)$	$\mathbf{r}=(x,y,z)$	$\mathbf{r}=(x,y,z)$	$\mathbf{r}=(x,y,z)$
5	$z(x)=\frac{d}{dx}y(x)$	$\mathbf{x}=(x,y,z)$	$\mathbf{x}=(x,y,z)$	$\mathbf{x}=(x,y,z)$
6	$f(t,\bar{u})=f(t,x,y,z)$	$F=(x,y,z)$	$F=(x,y,z)$	$F=(x,y,z)$
7	$\mathrm{P}(X=x\vert Y=y,Z=z)=\mathrm{P}(X=x\vert Z=z)$	$\mathbf{r_0}=(x,y,z)$	$\mathbf{r_0}=(x,y,z)$	$\mathbf{r_0}=(x,y,z)$
8	$=H(p,q)\cdot G(p,q)\vert_{p=\frac{x}{\lambda z},q=\frac{y}{\lambda z}}$	$\vec{x}=(x,y,z)$	$\vec{x}=(x,y,z)$	$\vec{x}=(x,y,z)$
9	$P=\{(x,y,z)\vert 3x+y-2z=10\}$	$\mathbf{x}=(x,y,z)^T$	(x,y,z)	$\mathbf{x}=(x,y,z)^T$
10	$z(x)=Q(y(x),\frac{d}{dx}y(x))$	(x,y,z)	$\mathbf{x}=(x,y,z)^T$	(x,y,z)

Performance

Space

Tangent inverted index (uncompressed, unoptimized) is 6.19 GB in size

Time

Indexing: 53 mins. (Tangent) vs. 8 mins (Lucene) - (25 core Linux server)

Tangent Retrieval: (1.5,1)s (mean,stdev) < 3s max - most time spent on network data transfer

Study Conclusions

1. Can we obtain more relevant results using Tangent than a conventional TF-IDF system used to index math?

   Confirmed; evaluated as significantly more relevant than Lucene-based system results.

2. Is Tangent fast enough for use in real-time?

   Yes; with (significant) room for improvement.

Tangent: Future Work

Optimization of inverted index

Modifications to incorporate matrices and pre-subscripts/superscripts

Integration with text-based search

**N. Pattaniyil made some progress on these problems in early 2014…(NTCIR Competition entry)*

DOCUMENT AND PATTERN RECOGNITION LAB

R·I·T

Handwritten Math Recognition

(work with IRCCyN/IVC)

Mouchere, H., Viard-Gaudin, C., Zanibbi, R. and Garain, U. (2014) **ICFHR 2014 Competition on Recognition of On-line Handwritten Mathematical Expressions (CROHME 2014).** Proc. Int'l Conf. Frontiers in Handwriting Recognition, Crete, Greece (to appear, Sept. 2014).

H. Mouchere, C. Viard-Gaudin, R. Zanibbi, U. Garain, D.H. Kim and J.H. Kim (2013) ICDAR 2013 CROHME: Third International Competition on Recognition of Online Handwritten Mathematical Expressions. *Proc. Int'l Conf. Document Analysis and Recognition,* Washington, DC

R. Zanibbi, H. Mouchere, and C. Viard-Gaudin (2013) Evaluating Structural Pattern Recognition for Handwritten Math via Primitive Label Graphs *Proc. Document Recognition and Retrieval*, Proc. SPIE vol. 8658, pp. 17-1 - 17-11, San Francisco, CA.

R. Zanibbi, A. Pillay, H. Mouchere, C. Viard-Gaudin, and D. Blostein. (2011) Stroke-Based Performance Metrics for Handwritten Mathematical Expressions. *Proc. Int'l Conf. Document Analysis and Recognition*, pp. 334-338, Beijing.

NANTES
Université de Nantes