

Math Information Retrieval: User Requirements and Prototype Implementation

Jin Zhao, Min-Yen Kan and Yin Leng Theng

Why Math Information Retrieval?

Examples:

- Looking for formulas
- Collect teaching resources
- Keeping updated on research development

Generic search engines ineffective in such situations

- Unaware of user needs and math expressions

[Fourier transform - Wikipedia, the free encyclopedia](#)

In mathematics, the continuous **Fourier transform** is one of the specific forms of Fourier analysis. As such, it transforms one function into another, ...
[en.wikipedia.org/wiki/Fourier_transform-111k-Cached-Similar pages](#)

[Fast Fourier transform - Wikipedia, the free encyclopedia](#)

Survey and history of FFT algorithms from Wikipedia.
[en.wikipedia.org/wiki/Fast_Fourier_transform-62k-Cached-Similar pages](#)

[Fourier Transform – from Wolfram MathWorld](#)

are sometimes also used to denote the **Fourier transform** and inverse **Fourier transform**, respectively (Krantz 1999, p. 202). ...
[mathworld.wolfram.com/FourierTransform.html-89k-Cached-Similar pages](#)

[Image Transforms - Fourier Transform](#)

The **Fourier Transform** is an important image processing tool which is used to decompose an image into its sine and cosine components. ...
[homepages.inf.ed.ac.uk/rbf/HIPR2/fourier.htm-22k-Cached-Similar pages](#)

[Fourier Transform](#)

The **Fourier Transform** is an important tool in Image Processing, and is directly related to filter theory, since a filter, which is a convolution in the ...
[student.kuleuven.be/~m0216922/CG/fourier.html-85k-Cached-Similar pages](#)

[Fast Fourier Transform](#)

This document describes the Discrete Fourier Transform (DFT), that is, a **Fourier Transform** as applied to a discrete complex valued series. ...
[local.wasp.uwa.edu.au/~pbourke/other/df/ -21k-Cached-Similar pages](#)

[MATHEMATICS OF THE DISCRETE FOURIER TRANSFORM \(DFT\) WITH AUDIO.](#)

MATHEMATICS OF THE DISCRETE FOURIER TRANSFORM (DFT) WITH AUDIO APPLICATIONS SECOND EDITION.
[ccrma.stanford.edu/~jos/mdft/ -33k-Cached-Similar pages](#)

[The Fourier Transform](#)

The **Fourier Transform**. by Bartosz Milewski How do we split sound into frequencies? Our ears do it by mechanical means, mathematicians do it using **Fourier** ...
[www.relisoft.com/science/physics/sound.html-11k-Cached-Similar pages](#)

[Fourier Transforms](#)

The **Fourier transform** defines a relationship between a signal in the time domain and its representation in the frequency domain. Being a **transform** ...
[www.see.ed.ac.uk/~mjj/dspDemos/EE4/tutFT.html-11k-Cached-Similar pages](#)

[PDF 3: Fourier Transforms](#)

File Format: PDF/Adobe Acrobat - [View as HTML](#)
and further generalized to derive the **Fourier Transform**. Forward **Fourier Transform**: ... Inverse **Fourier Transform** maps the series of frequencies ...
[www.cs.columbia.edu/~hns/teachinn/ais/slides/03-fourier.pdf-Similar pages](#)

[Advanced Search](#)
[Preferences](#)

Linked Expression:
 $a^2 + b^2 = c^2$

Search: the web pages from Singapore

Filter by > **resource category**, experience, specificity

Tutorials, Slides, Problem Solution Set, Applets, Tools, Data, Algorithms

[Pythagorean theorem - Wikipedia, the free encyclopedia](#)

Definition...

If we let c be the length of the hypotenuse, and a and b be the lengths of the other two sides, then the Pythagorean theorem states that

$a^2 + b^2 = c^2$
[en.wikipedi](#)

[Pythagor](#)

[Pythagoras](#)
[www.cut-the](#)

Goal: a *user-centric* and *math-aware* DL

[The Pythagorean Theorem Lesson](#)

A lesson on the **pythagorean theorem** with the objective that the student discovers or figures out by him or herself the actual **theorem**.

[www.arcytech.org/java/pythagoras/](#) - 5k - [Cached](#) - [Similar pages](#)

Tutorial

[Pythagorean Theorem -- from Wolfram MathWorld](#)

The various proofs of the **Pythagorean theorem** all seem to require application of some version or consequence of the parallel postulate: proofs by dissection ...

[mathworld.wolfram.com/PythagoreanTheorem.html](#) - 52k - [Cached](#) - [Similar pages](#)

Proof

Outline

- Introduction
- **Literature Review**
 - Domain-specific Information Seeking Studies
 - Current Math Resources
 - Math Information Retrieval
- **User Study**
- **Prototype Implementation**
- **Conclusion**



Domain-Specific Information Seeking Studies

- **Brown 1999**

- Monograph being the major source of information for math
- Predominant source of information for math

Key requirements:

- **Wiberley**

- Technological resources that contain relevant content

**Usefulness,
Usability, and**

- **Tibbo 2001**

- Growing importance of online resources acknowledged but coupled with usability and accessibility problems

Accessibility

Current Math Resources Online

From Math Web Search

```
<mq:query xmlns:mq="http://mathweb.org/MathQuery">
  <math xmlns="http://www.w3.org/1998/Math/MathML">
    <apply><int/>
      <domainofapplication mq:generic="domain"/>
      <bvar> <ci>t</ci> </bvar>
      <apply><power/>
        <apply><ci mq:generic="fun"/>
```

From Wolfram Function Site

• **Functions**

From the category , the function

• **Constants**

The constant

1. Hamper Accessibility

2. Limited search capability and hard to judge usefulness

1. Lack of cross-reference and subscription required

2. Different degree of math-awareness

- Math-unaware
- Syntactically Math-aware
- Semantically Math-aware



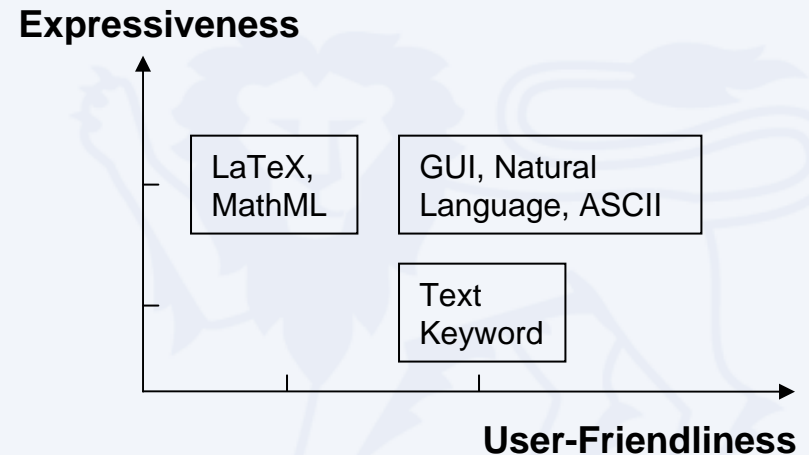
Current Math Information Retrieval

Expression Matching

- **Text-based approaches**
 - Match expressions on the surface
 - Notational Variation Problem: " $a^2+b^2=c^2$ " \neq " $x^2+y^2=z^2$ "
- **Non-text-based approach**
 - Tree matching

Query language

- Text keywords
- Math authoring language
- Expression-input friendly language



Unanswered Issues

Whether the information needs of the users are satisfied by such resources

- What do the user really need?
- How do they perform information seeking?
- What are the difficulties encountered?

Whether the current research focus is appropriate

- Do they really need/prefer expression search?

Further study needed

Outline

- Introduction
- Literature Review
- **User Study**
 - Study Design and Consideration
 - Findings
 - Desiderata in Math Information Retrieval
- **Prototype Implementation**
- **Conclusion**



User Study

Study Design and Considerations

- Qualitative feedbacks for system design
- Pilot for future user study
- Small scale
- Semi-structured interviews
- Focus on profiling user behavior and analyzing needs
- Findings stabilized towards the end

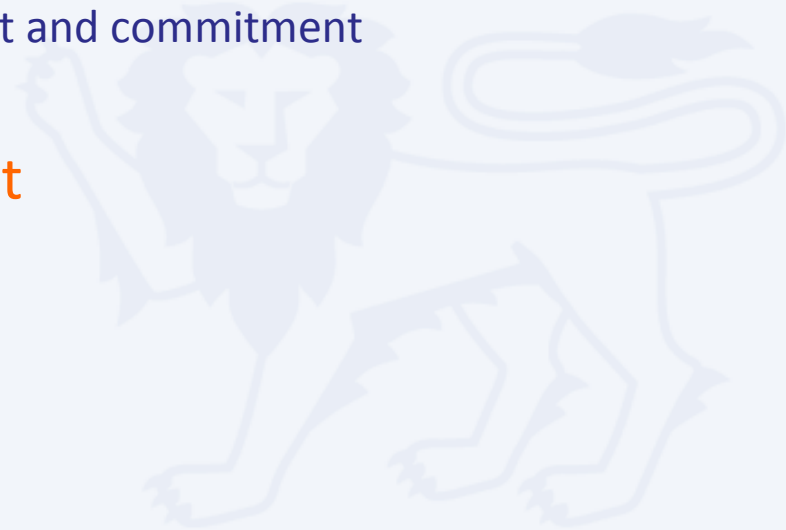


User Study (Findings)

Three Approaches

- Keyword Search
 - Fast, available but disorganized
- Browsing
 - More effective but costly to compile or subscribe to
- Personal Contacts
 - Most effective but requires more effort and commitment

Trade-off between cost and benefit



User Study (Findings)

Expression Search

- Attractive but utility unknown
 - To find homework solutions?
 - Too specific
 - Less prevalent in certain domains
 - More convenient to use keyword

Keyword search still popular and preferred



User Study (Findings)

The multi-faceted user needs

- **Informational / Resource**
 - Definition, example, proof, etc.
 - Slides, tutorial, tools, etc.
- **Two implicit facets for filtering**
 - Specificity
 - Experience
- **The context**
 - Domain
 - Intent

Need to cater for specifically



Desiderata in Math Retrieval

Multi-collection search

- Search through multiple collections on behalf of the user

Enhance the usability and accessibility of collections

Resource Categorization

- Automatically classify the materials according to the different facets of the user needs

Return results that best suit the user needs



Outline

- Introduction
- Literature Review
- User Study
- **Prototype Implementation**
 - Focus on Resource Categorization
- **Future Work**



Prototype Implementation

Multi-collection Search

- Meta-search
- Offline indexing based on open source package
- Easier requirement to meet between the two

Resource Categorization

- Domain-specific text categorization on webpages
- More interesting as a research topic

[Advanced Search](#)
[Preferences](#)
Linked Expression: $a^2+b^2=c^2$

Search: the web pages from Singapore

Filter by > **resource category**, experience, specificity

Tutorials, Slides, Problem Solution Set, Applets, Tools, Data, Algorithms

[Pythagorean theorem - Wikipedia, the free encyclopedia](#) Definition...

If we let c be the length of the hypotenuse and a and b be the lengths of the other two sides, the theorem can be expressed as the equation:

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

en.wikipedia.org/wiki/Pythagorean_theorem - 98k - [Cached](#) - [Similar pages](#)

[Pythagorean Theorem and its many proofs from Interactive ...](#) Proof

Pythagoras' Theorem. 78 proofs of the **Pythagorean theorem.**
www.cut-the-knot.org/pythagoras/index.shtml - 157k - [Cached](#) - [Similar pages](#)

[The Pythagorean Theorem Lesson](#) Tutorial

A lesson on the **pythagorean theorem** with the objective that the student discovers or figures out by him or herself the actual **theorem**.
www.arcytech.org/java/pythagoras/ - 5k - [Cached](#) - [Similar pages](#)

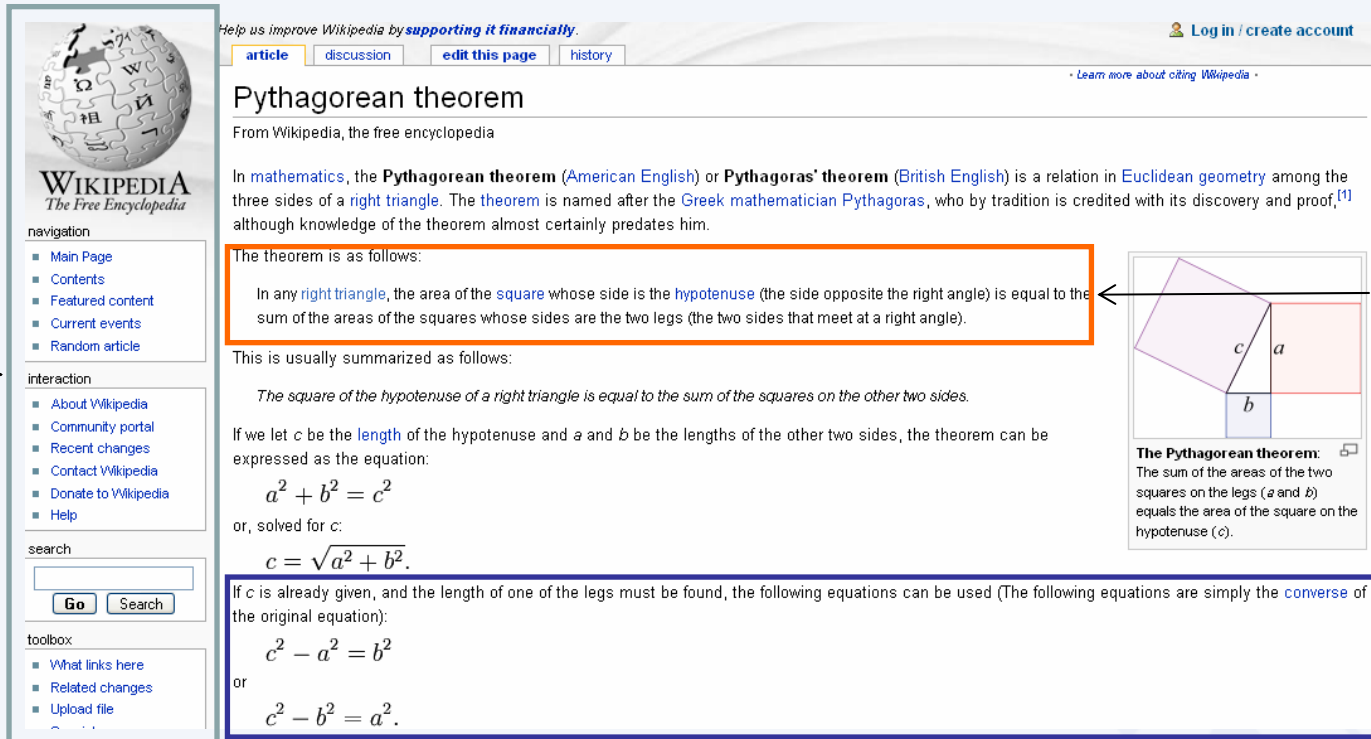
[Pythagorean Theorem -- from Wolfram MathWorld](#) Proof

The various proofs of the **Pythagorean theorem** all seem to require application of some version or consequence of the parallel postulate: proofs by dissection ...
mathworld.wolfram.com/PythagoreanTheorem.html - 52k - [Cached](#) - [Similar pages](#)

Focus of the prototype is on **Resource Categorization**

Webpage Segmentation

Entire page is not a suitable unit for categorization



The screenshot shows a Wikipedia article titled "Pythagorean theorem". The page is annotated with several boxes and arrows to illustrate segmentation:

- Toolbar**: An arrow points to the left sidebar navigation menu.
- Definition**: An orange box highlights the text: "The theorem is as follows: In any right triangle, the area of the square whose side is the hypotenuse (the side opposite the right angle) is equal to the sum of the areas of the squares whose sides are the two legs (the two sides that meet at a right angle)." An arrow points from the word "Definition" to this box.
- Variation**: A blue box highlights the text: "If c is already given, and the length of one of the legs must be found, the following equations can be used (The following equations are simply the converse of the original equation): $c^2 - a^2 = b^2$ or $c^2 - b^2 = a^2$." An arrow points from the word "Variation" to this box.
- Diagram**: A diagram of a right-angled triangle with legs of length a and b , and hypotenuse of length c . Squares are drawn on each side. An arrow points from the "Definition" box to the diagram.

- Vision-based Segmentation (VIPS) used

Resource Categorization

Supervised Machine Learning Pipeline

- Labels
 - Math related / non-math-related
- Features
 - Word, Image, Formatting, Hyperlink, Layout, Context
- Machine Learner
 - SVM
- Training/Testing Data
 - Small corpus of webpages for 5 math topics
 - Manually annotated
 - Kappa-agreement: 0.87



Evaluation

Average accuracy: 0.36 on F_1

- Strength: separating math contents from the rest
- Weakness: identifying their exact type

Feature Utility

- Text → competitive baseline
- Image → filter non-math information
- Formatting → identify section headings etc.
- Hyperlink → separate related concepts and resource from the rest
- Layout → improve precision at the cost of recall
- Context → not effective overall

Potential Sources of Error

- Training Data
 - Insufficient examples
 - Skewed distributions
- Segmentation
 - Over- or under-segmented

Operation on $x(t)$	Resulting signal $y(t)$	$Y(\omega)$
amplitude scale	$bx(t)$	$bX(\omega)$
time shift	$x(t-T)$	$e^{-j\omega T} X(\omega)$
time scale	$x(at), a \neq 0$	$\frac{1}{ a } X(\omega/a)$
time reverse	$x(-t)$	$X(-\omega)$
derivative	$\frac{dx(t)}{dt}$	$j\omega X(\omega)$
running integral	$\int_{-\infty}^t x(\sigma) d\sigma$	$\frac{1}{j\omega} X(\omega) + \pi X(0) \delta(\omega)$

HOME

HISTORY

DIAGRAMS

TEST

LINKS

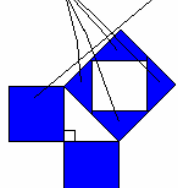
SIGN GUESTBOOK

VIEW GUESTBOOK

E MAIL

4

These pieces make this!



You can view one of the images by clicking **once** on the picture you want.

Pythagoras was a great Mathematician who was the first to create the music scale of today. He also created theorems. One of his most famous theorem was:

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

If You have any information on different proofs e-mail me. I would love to add more proofs to my site. Thank you.

This theorem shown in the diagram was created by Dundeny. You start with a generic right angled triangle. The hypotenuse is labeled c. The bottom of the triangle is b. And the remaining side is labeled a. Now draw squares off each side of the triangle (see diagram 1). Take the square from side b and insert it in the center of the square side c (see diagram 2). The remaining edge of the side c square is the same area as the side a square (see diagram 3 & 4). This proves that $a^2+b^2=c^2$.

Outline

- Introduction
- Literature Review
- User Study
- Prototype Implementation
- **Conclusion**
 - Future Work



Future Work

Iterative Development Process

- Enhance and extend categorization
- Prototype fielding after expanded user testing and requirement analysis

Text-to-Expression Linking

- **Resolve text keywords to expressions**
“Pythagorean Theorem” → : “ $a^2+b^2=c^2$ ” & “ $x^2+y^2=z^2$ ”
 - Reduce the need for expression input
 - Help to solve the notational variation problem
 - Fit well with the rest of the desiderata

Conclusion

To create a *user-centric* and *math-aware* digital library on math materials

Two Desiderata:

- Multi-Collection Search, Resource Categorization

Prototype classification accuracy of 0.36 F_1

Future Text-to-Expression Linking

Thank you for listening
Questions?