### **Exploring the One-brain Barrier:** a Manual Contribution to the NTCIR-12 MathIR Task WIKIPEDIA The Free Encyclopedia Moritz Schubotz, Norman Meuschke, Marcus Leich, and Bela Gipp



### Math Information Retrieval

- Formulae: integral part of language in STEM
- Information Retrieval
- Today: Textual content
- Recently: Image and Video
- "If  $\boldsymbol{v}$  is a vector which is not zero, then it is an eigenvector of a square matrix A if Av is a scalar multiple of v."
- Exponential growth in #publications Fig. 2 Math as integral part of text
- Applications
- Applicable Theorem Search
- Plagiarism Detection
- Related work search
- Math IR needs
- Datasets with structured mathematical formulae
- Topics
- Evaluation methods

berlin

Technische

Berlin

Universität

# Example result

Query

Difference between  $Log x_1$ and  $\log x_1$ 

Physikerwelt 0 F 0 Talk Sandbox Preferences Beta



Contents

Featured content

#### Common logarithm

In mathematics, the common logarithm is the logarithm with base 10. It is also known as the decadic logarithm and also as the decimal logarithm, named after its base, or Briggsian logarithm, after Henry Briggs, an English mathematician who pioneered its use, as well as "standard logarithm". It is indicated by  $\log_{10}(x)$ , or sometimes Log(x) with a capital L (however, this



D pproac 5

Current events	notation is ambiguous since it can also mean the complex	
Random article	natural logarithmic multi-valued function). On calculators it is usually "log", but mathematicians usually mean natural logarithm (logarithm with base $e \approx 2.71828$ ) rather than common logarithm	
Donate to Wikipedia Wikipedia store		
Interaction Help About Wikipedia Community portal Recent changes Contact page	when they write "log". To mitigate this ambiguity the ISO 80000 specification recommends that $\log_{10}(x)$ should be written lg ( <i>x</i> ) and $\log_{e}(x)$ should be ln ( <i>x</i> ).	
	Contents [hide]	
Tools What links here Related changes Upload file Special pages	<ul> <li>2 History</li> <li>3 Numeric value</li> <li>4 See also</li> <li>5 Notes</li> <li>6 Deferences</li> </ul>	
<		

## Conclusion

- Gold standard dataset as big step forward to develop a math aware search engine for Wikipedia
- System strengths:
  - -Definition lookup queries
  - -Applications lookup

## Future work

- Improve description of information need
- taking into account our focussed mir task categories
  - 1. Definition look-up
  - 2. Explanation look-up
  - 3. Proof look-up
  - 4. Application look-up
  - 5. Computation assistance

0 N utlook <u>
</u> <u>S</u>. 0

 $\bigotimes$ 

• System weaknesses

-Low precision

-No standard interface to specify query type

6. Gernal formula search

- Improve query syntax for similarity search Develop a math search engine for wikipedia with the help of the gold standard dataset and new Mathematical Language Processing Technology





