A Survey on Recognition and Retrieval of Mathematical Expressions

Bijoy Barman[1] and Swapan Biswas[2]

[1]Department of Physics, Abhayapuri College, Abhayapuri, Assam. India
[2]Department of Computer Science, Abhayapuri College, Abhayapuri, Assam. India

Abstract: Document recognition and retrieval technologies complement one another, providing improved access to increasingly large document collections. While recognition and retrieval of textual information is fairly mature, with wide-spread availability of Optical Character Recognition (OCR) and text-based search engines, recognition and retrieval of graphics such as images, figures, tables, diagrams, and mathematical expressions are in comparatively early stages of research. This paper surveys the state of the art in recognition and retrieval of mathematical expressions, organized around four key problems in math retrieval (query construction, normalization, indexing, and relevance feedback), and four key problems in math recognition (detecting expressions, detecting and classifying symbols, analyzing symbol layout, and constructing a representation of meaning).

Keywords: Math Recognition, Graphics Recognition, Mathematical Information Retrieval (MIR), Content Based Image Retrieval (CBIR), Human-Computer Interaction (HCI)

1. INRRODUCTION

1.1 What is Information Retrieval (IR)

Information Retrieval (IR), one of the most well-known applications of Natural Language Processing (NLP), has gotten a lot of attention over time. The term "information retrieval" refers to a set of techniques for storing and retrieving text, image, and video data. Various strategies for efficiently retrieving text information from documents have been popular in the past [1]. These text-based techniques, on the other hand, are insufficient for retrieving mathematical data, which is complex to encode and involves two-dimensional symbol alignment rather than strings of characters. Scientific texts and mathematical terminology abound in scientific documents. Furthermore, because mathematical material is frequently supplemented by language, a math search engine is necessary to search mathematical topics using either text or formula. The mathematical expressions, in most situations, cannot be explained in detail and searched in a few

words; rather, these formulas augment the meaning of the texts. As a result, it is necessary to make the mathematical information retrieval and search system more user-friendly.

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)

Mathematical expressions are a type of semi-formal visual language [2]. Math symbols, music signs, and chemical notations are all examples of graphical languages that are used to communicate meaning. However, there are a few stopping hurdles to figuring out the math equation [3].

1.2 Difference between information retrieval and data retrieval

SL. No	Data Retrieval	Information Retrieval
1	Retrieves data based on the keywords in the query entered by the user.	Retrieves information based on the similarity between the query and the documents.
2.	There is no room for errors since it results in complete system failure.	Small errors are tolerated and will likely go unnoticed.
3.	It has a defined structure with respect to semantics.	It is ambiguous and doesn't have a defined structure.
4.	Provides solutions to the user of the database system.	Does not provide a solution to the user of the database system.
5.	Data Retrieval system produces exact results.	Information Retrieval system produces approximate results
6.	Displayed result are not sorted by relevance	Displayed result are sorted by relevance
7.	Data Retrieval model is deterministic by nature.	Information Retrieval model is probabilistic by nature.

1.3 Components of an Information Retrieval System

A general information retrieval functions in the following steps. It is shown in Figure 1.1

1. The system browses the document collection and retrieves documents is Crawling

2. The system creates a document index is Indexing

3. User gives the query

4. The system displays to the user documents from the index that are relevant to the user's query is Ranking

5. User may give relevance feedback to the search engine - Relevance Feedback.

CHAPTER 1. INTRODUCTION

Figure 1.1: Important Processes in Web information Retrieval

1.4 How to encode Math formulae?

There are many different ways to encode math formulae, including LATEX, MathML, and its extensions.

- Formulae in scientific texts contain a large number of symbols that can be difficult to distinguish from one another.
- A huge number of characters with varying font scripts and typefaces complicates symbol recognition.
- Notations can contain notable ambiguity, with the same symbols having distinct meanings. A dot symbol, for example, can be used to represent a fractional value or a multiplication operator.
- Depending on the domain, the same symbol might have multiple meanings. The Lambda constant, for example, might be a variable, a constant, or a binding function.

- Some math formulae may be handwritten, making symbol segmentation more difficult.
- It is laborious to recognize spatial association such as $P\left(\frac{A}{B}\right)$ represents the conditional probability or constant A divide by constant B and the result multiplied with constant P.
- Alternative representations are available for a number of math formulae. For instance, square root of x can be represented in three different ways: $\sqrt{x}, x^{\frac{1}{2}}$ or $\sqrt[2]{x}$.

These findings indicate a lack of understanding of precedence and the link between operations, which makes retrieving math information more challenging. Mathematical information retrieval (MIR) is different from textual information retrieval; however, in textual information retrieval, the user enters a few keywords related to the needed information rather than the entire needed information, and the retrieval system searches for the related document based on those keywords. In MIR, however, the user enters the formula, which contains all of the necessary data. As a result, the MIR system should be able to prioritize articles that contain a complete version of the user question. In the recent years, conventional IR systems have been improved significantly in regard to the indexing and searching mechanisms to facilitate MIR. Moreover, the conventional indexing mechanisms are inefficient in dealing with scientific symbols, foreign terms, and composite notations. Such notations are either ignored or misinterpreted, which eventually affects the retrieval performance. It is observed that, Indexing of mathematical information performs different stipulations like- canonicalization, tokenization, structural unification and different representation of math expression. This leads to retrieval of relevant search results. Canonicalization deals with the representation of math notations and it are almost similar with minor differences in their syntax. Therefore, it reduces the redundancy by indexing them on the same position. Similarly, tokenization and structural unification assist in finding semantically similar formulae and sub-formulae [4, 5].

The input to a math recognition system can be realized in three forms: vector graphics (such as PDF), strokes (such as pen strokes on a data tablet), or a document image. The processing that which is needed to extract expressions and recognize characters depends mostly on the form of input.

1.2 Key Challenges

Math recognition serves a variety of functions. For example, a user may create an expression by hand and then insert the recognition result (such as a LATEX string or a picture) into a document. A computer algebra system like Maple or Mathematica can also be used to evaluate a recognised expression. Another possibility is to use the recognised phrase as a query to find documents with similar math notation. The development and implementation of pen-based math entry systems is further motivated by recent work in human-computer interaction. Bunt et al. [15] study mathematicians in a research setting and find that in order to be useful, CAS systems must support annotation, provide multiple levels of formality, and provide more transparency for

the operations that they apply; they suggest that pen-based math systems could be used to meet these needs.

Tutoring systems make use of math recognition as well. When middle and high school students tried a math tutoring prototype (based on FFES/DRACULAE), students who used pen entry completed their math tutoring sessions in half the time as those who typed, with no significant difference in pre-to-post test score gains.

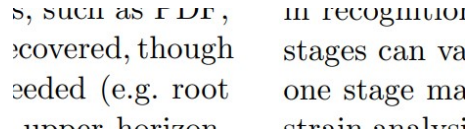
In the recognition of math notation, the following issues arise.

1. Expression Detection: Expression detection requires first identifying and segmenting expressions. Offset expression recognition methods are fairly reliable, however recognising expressions embedded in text lines remains a challenge.

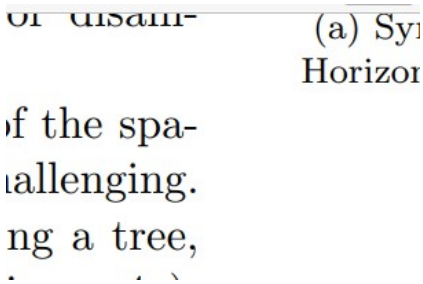
2. Symbol Extraction or Symbol Recognition: In cases like vector-based representations such as PDF, symbol locations and labels can be recovered, but some handling of special cases is needed (e.g. root symbols are often typeset with the upper horizontal bar represented separately from the radical sign, √). In raster image data and pen strokes, detecting symbol location and identity is challenging. There are hundreds of alphanumeric and mathematical symbols used, but many of these symbols are similar in appearance. Hence, some use of context is necessary for disambiguation (e.g. O, o, 0).

3. Layout Analysis: It's difficult to analyse the spatial relationships between symbols. A tree, known as a symbol layout tree, is frequently used to show spatial structure. Symbol layout trees are comparable to LATEX math expressions in that they represent information. They show subscript, superscript, above, below, and containment relationships, as well as which sets of horizontally adjacent symbols share a baseline (writing line). To ease later processing, symbols can be merged into tokens (e.g. function names and numeric constants).

4. Mathematical Content Interpretation: The variables, constants, operands, and relations represented in an expression are recovered by interpreting the symbol layout. This analysis generates an operator tree, which is a syntax tree for an expression. An operator tree can be used to evaluate an expression if it has definitions for symbols and operations, such as after mapping the tree to an expression in a CAS (Computer Algebra System) language like Matlab, Maple, or Mathematica. However, establishing the correct symbol and structural mapping can be challenging, especially when context is limited.



a. Symbol Layout tree. The tree is rooted at left (‘(‘) Horizontally adjacency relationship edges are unlabeled



b. Operator tree. The tree represents the addition of a and b, squared.

Figure 1.2. Symbol Layout tree & Operator tree for $(a + b)^2$.

5. Normalization: Both the query and the searchable documents are normalised to reduce variance. Expressions must be reduced to canonical forms in text-based retrieval to avoid mismatches between equivalent expressions with various representations. Normalization of symbol layout trees, for example, imposes a distinct ordering on spatial relationships. Enumeration of variables in operator trees, for example, allows variables to be matched regardless of their exact symbol identities.

6. Query Languages and Query Formulation: LATEX, MathML [16], and OpenMath [17] have inspired modern query languages for retrieving mathematical information. Determining what types of inquiries are valuable and practical, as well as providing an efficient user interface for query formation, are both challenges in query formulation.

7. Indexing and Matching: The document representation chosen and the similarity criteria used to match queries to the index have a significant impact on retrieval performance. Different indexing and retrieval methods are required for vector, picture, and stroke data. Currently, only a small amount of effort has been done on indexing and retrieving handwritten mathematical documents.

8. Relevance Feedback: The user can provide relevance comments while seeing a retrieval result, allowing the system to automatically generate an improved query. This is an essential but untapped research area for math retrieval systems at the moment. In text and image-based retrieval systems, relevance feedback has been extensively investigated.

Besides these four key problems, the evaluation of a math retrieval system is also difficult.

Mathematical Information Retrieval (MIR) is a relatively new research area, lying at the intersection of text-based information retrieval, content-based image retrieval and Mathematical Knowledge Management (MKM). Mathematical knowledge management deals with the representation, archiving, extraction, and use of mathematical information.

MIaS (Math Indexer and Searcher)

The Math Indexer and Searcher [23] is a maths-aware **full-text** based search engine. It is based on the state-of-the-art system Apache Lucene, however, its maths processing capabilities are standalone and can be easily integrated into any Lucene/Solr based system, as in **EuDML** [24] search service. MIaS processes documents containing mathematics encoded in MathML [19] format and in several steps allowing formulae similarity search transforms problem of matching XML structures to regular full-text searching. For calculating the relevance to the user's query, MIaS uses a **heuristic weighting of indexed mathematical terms**, which accordingly affects scores of matched documents and thus the order of results.

Math Web Search

Math Web Search System [18], a search engine that can index mathematics found in content representation on the web. Math Web Search is a system which employs substitution tree indexing and provides specific math querying capabilities. It retrieves documents based on user generated XML queries or math formula queries generated by the interactive JavaScript-based Sentido interface. It is a very useful tool in the information extraction enterprise, because it can understand and index mathematics in the Content MathML format.

MathML

Another way to formally specify mathematics is, of course, the MathML [19] XML format, which can encode both the presentation and the semantic form of mathematical formulas. The presentation form, Presentation MathML, is row-oriented format, focused on visual representation. The semantic form, called Content MathML is, as the name suggests, oriented towards content and the underlying semantics of the formula, thus posing no ambiguity related to the formal mathematical meaning. Difficulties, arising mostly from the different interpretations of mathematical symbols in complex formulas are the main issues tackled by mathematical formula disambiguation, which needs to be done in order to reach a semantic computer parseable format.

SciBorg

The SciBorg project [11] tries to combine several existing tools for a unified analysis, using one basic information format. Focusing several Natural Language Processing (NLP) tools on working with Robust Minimal Recursion Semantics (RMRS), the project attempts the extraction of information from scientific papers in the field of chemistry. The project appeals to classical NLP tools, like tokenizers, part-of-speech (POS) taggers and parsers for various

analyses of the same text. The RASP system is the backbone of the SciBorg processing architecture, performing sentence splitting, tokenizing, parsing and POS tagging.

Various approaches for ME detection have been investigated in recent years [8]. Page segmentation techniques [9] are used initially in traditional detection procedures to acquire text lines and words. The MEs are then determined by extracting features from the text lines and words. To detect MEs, existing methods have incorporated several feature extractions and classifiers. Lee and Wang [10] proposed a set of heuristic criteria for obtaining MEs from a printed page image. The parameters of the height and line spacing of text lines are computed for isolated expression detection. The layout, context, and identification of the characters are all used to detect inline expressions. The error correcting step is included to increase the ME detection accuracy. They validated their method using small data sets.

To detect MEs, Garain [11] uses a system that extracts multiple layouts and linguistic aspects. Several properties of expressions (e.g., symbol size variation, text line size variation, and the presence of special symbols) have been proposed for the detection of an isolated expression. The linguistic properties of textual words are examined to find inline expressions. The features are derived from the results of character recognition. This approach is also used to develop context information for inline expressions. To detect the expressions, some heuristic thresholds are established after the feature extraction. The approach was evaluated on a tiny data set, making estimating the thresholds problematic. The method in [12] integrated ME detection into the OCRopus system. First, text lines in documents are segmented using the one-column projection algorithm supported by OCRopus. The method then extracts five layout features of text lines to detect isolated expressions. After feature extraction, a support vector machine classifier is applied for detection.

ME detection is used in the approach described in [13], which is divided into two steps. For heterogeneous document images, a low-cost text line segmentation algorithm is applied in the first stage. To detect isolated expressions, a layout feature extraction of the resultant text lines is designed. Words are divided into normal text lines. To determine inline expressions, features such as variations in the information position or character size) of segmented words are explored. For ME identification, two SVM classifiers were trained and optimised. For isolated expressions, the technique achieves competitive accuracy; nevertheless, inline expression identification accuracy needs to be improved.

DNNs have been tuned in recent years to improve the accuracy of ME identification in complicated documents [14]. On a large data set, Ohyama et al. trained a CNN inspired by U-net for ME identification. The technique separates document pictures into sections. The CNN is then trained using the picture blocks. The detection of mathematical symbols is 95.2% accurate in this study; nevertheless, the layout analysis of the symbols requires refinement in order to acquire the final MEs. The different sorts of textual word styles (italics and bold) resulted in inline expression recognition issues.

In a recent study by Phong et al [5] , a two-stage technique for ME detection was created. To extract the text lines and words, page segmentation is used first. Then, to boost detection accuracy, a late fusion of handmade and deep learning feature extraction was developed. The qualities of a quick Fourier transform and a projection profile were presented for isolated and inline expressions, respectively, in handmade feature extraction. The random forest was fine-tuned to detect the MEs after that. AlexNet and Resnet-18 transfer learning were employed in a deep learning technique. To obtain correct expression position information, post-processing is required.

A Mondol et al [20], proposed a mathematical equation description (MED) model, a novel end-to-end trainable deep neural network-based approach that learns to generate a textual description for reading mathematical equation images. This model consists of a convolution neural network as an encoder that extracts features of input mathematical equation images and a recurrent neural network with attention mechanism which generates description related to the input mathematical equation images.

In [23], the authors focus on the resolving ambiguities in mathematical symbols and propose a novel recognition technique that has been tested over large number of ambiguous mathematical symbols obtained from different categories including factoring formula, algebra identity, geometric progression, integral expressions, quadratic equations, area function, and geometric progression formulas for mathematical expressions. In this paper, authors propose a novel feature extraction technique which helps in reducing the ambiguity within mathematical symbols. Once, the identification of ambiguous isolated mathematical symbols from different categories of mathematical expressions is completed. Then the features are extracted using feature extraction technique which includes Fourier descriptor and chain code. Two different classifiers SVM and K-NN are used for classification purpose. Fivefold cross validation technique are used for computing recognition results.

In [24] the authors have proposed a new method for detection of displayed expression. The process is that the image of document layout is analyzed, and ordinary text lines and displayed expressions are extracted using OCRopus, popular open source software. After that, displayed expression is evaluated using the combination of Fast Fourier Transformation (FFT) and Mean Square Error (MSE) to extract features. Finally, threshold and Support Vector Machine (SVM) are used to classify. In this, two benchmark datasets: Harvard Mathematical Textbooks Dataset [25] and InftyCDB-2 [26].

Conclusion

Recognition and retrieval of mathematical notation are challenging, interrelated research areas of great practical importance. In math retrieval, the key problem with ground truth, and increased availability of datasets for math recognition and retrieval. There will be advances in performance metrics for computing errors in layout, segmentation, parsing, classification, and representation of meaning. Performance evaluation needs to be carried out in reference to tasks a user is trying to accomplish. Research is needed to obtain a better understanding of different models of

relevance for mathematical information retrieval. Relevance depends on a number of factors, including the expertise of the user, the task underlying the user's information need, and the type of resource(s) sought. In math recognition, future directions and open problems include the detection of inline expressions, the automatic detection of mathematics in vector graphics documents, and the processing of matrix and tabular structures. We predict refinements of layout analysis, including development of new techniques and combination of existing methods via parser combination. More sophisticated language models will be developed to incorporate statistical information about mathematical notation; this information can be used during recognition or post-processing. Stochastic language models will be become increasingly sophisticated; stochastic gram mars, as initially proposed by Chou [34] can be ex tended using different segmentation and/or parsing approaches. A challenge is to identify usable notation sets with invariants that can be easily adapted to dialects; the goal is to scale this up to the index set used by the Mathematical Subject Classification (MSC) [121]. In conclusion, the combination of math retrieval and math recognition technologies provides rich possibilities for math-aware computer interfaces, and for intelligent search and retrieval tools for math in documents.

This section presents a short review of the of the most relevant current research projects focused on mathematics processing

References:

1. *Carpineto, C., Romano, G., 2012. A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR) 44, 1–50.*
2. *Zanibbi, R., Blostein, D., 2012. Recognition and retrieval of mathematical expressions. International Journal on Document Analysis and Recognition (IJDAR) 15, 331–357.*
3. *Schubotz, M., Youssef, A., Markl, V., Cohl, H.S., 2015. Challenges of mathematical information retrievalin the ntcir-11 math wikipedia task. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 951–954.*
4. *Ruzicka, M., Sojka, P., Liska, M., et al., 2016. Math indexer and searcher under the hood: Fine-tuning query expansion and unification strategies. In: Proc. of the 12th NTCIR Conference on Evaluation of Information Access Technologies. Noriko Kando, Tetsuya Sakai, and Mark Sanderson, (Eds.) NII Tokyo, pp. 331–337.*
5. *Phong B., Hoang T.M., & Le T. L. (2020). A hybrid method for mathematical expression detection in scientific document images. IEEE Access, 8, 83663-83684.*
6. *Lin X., Gao L., Tang Z., Baker J., &Sorge V (2014). Mathematical formula identification and performance evaluation in pdf documents. International Journal on Document Analysis and Recognition, 17, 239-255.*
7. *Katherine L., Abdollahian G., Boutin M., &Delp E.J. (2011). A low complexity sign detection and text localization method for mobile applications. IEEE Transactions on Multimedia, 13, 922-934.*

8. *Lee H., & Wang J. (1997). Design of Mathematical expression understanding system. Pattern Recognition Letters, 18, 289-298.*
9. *Garain U. (2009). Identification of mathematical expressions in document images. International Conference on Document Analysis and Recognition.*
10. *Yamazaki S., Furukori F., Zhao Q., Shirai K., & Okamoto M. (2011). Embedding a mathematical OCR module into OCRopus. In International Conference on Document Analysis and Recognition, IEEE.*
11. *Chu W. and Liu F. (2013). Mathematical formula detection in heterogeneous document images. Proceeding of the International Conference in Technologies and Applications of Artificial Intelligence.*
12. *Ohyama W., Suzuki M. & Uchida S. (2019). Detecting mathematical expressions in scientific document images using a u-net trained on a diverse dataset. IEEE Access, 7, 144030-144042.*
13. *A. Bunt, M. Terry, and E. Lank. Friend or foe? ExaminingCAS use in mathematics research. In Proc. Int'l Conf. Human Factors in Computing Systems, pages 229–238, NewYork, 2009.*
14. *R. Ausbrooks, S. Buswell, D. Carlisle, G. Chavchanidze,S. Dalmas, S. Devitt, A. Diaz, S. Dooley, , R. Hunter, P. Ion,M. Kohlhase, A. Lazrek, P. Libbrecht, B. Miller, R. Miner,C. Rowley, M. Saregent, B. Smith, N. Soiffer, R. Sutor, andS.Watt. Mathematical markup language (MathML) version3.0, W3C recommendation (http://www.w3.org/math/),2010.*
15. *The OpenMath Society. http://www.openmath.org/.*
16. *S. Anca. Natural Language and Mathematics Processing for Applicable Theorem Search. Master's thesis, Jacobs University, Bremen, Aug. 2009.*
17. *Mathematical Markup Language (MathML) https://www.w3.org/Math/.*
18. *arXiv Corpus -http://www.arxiv.org.*
19. *H. Phong, T. M. Hoang and T. -L. Le, "A Hybrid Method for Mathematical Expression Detection in Scientific Document Images," in IEEE Access, vol. 8, pp.83663-83684, 2020.*
20. *A. Mondal and C. V. Jawahar, "Textual Description for Mathematical Equations," 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1300-1307.*
21. SOJKA, Petr and Martin LÍŠKA. The Art of Mathematics Retrieval. In Matthew R. B. Hardy, Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60. ISBN 978-1-4503-0863-2.

22. W. Sylwestrzak, J. Borbinha, T. Bouche, A. Nowi´nski, and P. Sojka. EuDML—Towards the European Digital Mathematics Library. In P. Sojka, editor, Proceedings of DML 2010, pages 11–24, Paris, France, July 2010. Masaryk University. EuDML https://eudml.org/.
23. S. A. Naik, P. S. Metkewar and S. A. Mapari, "Recognition of ambiguous mathematical characters within mathematical expressions," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2017, pp. 1-4.
24. B. H. Phong, T. M. Hoang and T. Le, "A new method for displayed mathematical expression detection based on FFT and SVM," 2017 4th NAFOSTED Conference on Information and Computer Science, 2017, pp. 90-95.
25. “Harvard dataset,” http://www.math.harvard.edu/ shlomo/, accessed: 2017-07-14.
26. M. S.Uchida, A.Nomura, “Quantitative analysis of mathematical documents,”International Journal on Document Analysis and Recognition, vol. 7, no. 4, pp. 211–218, 2005.