Search Engine Crawler and Indexer for Mathematical Formulas

Project Report

submitted in partial fulfillment of the

requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

Name	Roll No.
Siddharth Joshi	R2142220664
Bhavya Agrawal	R2142221157
Sumit Verma	R2142220184

under the guidance of

Dr. Pankaj Kundan Dadure

School of Computer Science

University of Petroleum & Energy Studies

Bidholi, Via Prem Nagar, Dehradun, Uttarakhand

November – 2024

CANDIDATE'S DECLARATION

We hereby certify that the project work entitled **"Search Engine Crawler and Indexer for Mathematical formulas"** in partial fulfilment of the requirements for the award of the Degree of BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING with specialization in **Artificial Intelligence & Machine Learning** and submitted to the Department of Systemics, School of Computer Science, University of Petroleum & Energy Studies, Dehradun, is an authentic record of my/ our work carried out during a period from **August**, **2024** to **December, 2024** under the supervision of **Dr. Pankaj Kundan Dadure , Assistant Professor SoCS**.

The matter presented in this project has not been submitted by us for the award of any other degree of this or any other University.

Siddharth Joshi	**R2142220664**
Bhavya Agrawal	**R2142221157**
Sumit Verma	**R2142220184**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: 27th November 2024

Dr.Pankaj Kundan Dadure

Project Guide

ACKNOWLEDGEMENT

We wish to express our deep gratitude to our guide **Dr.Pankaj Kundan Dadure**, for all advice, encouragement and constant support he has given us throughout our project work. This work would not have been possible without his support and valuable suggestions.

We sincerely thanks to our respected **Prof. Vijaysekra Chellaboina, Head Department of SoCS,** for his great support in doing our project in **Search Engine Crawler and Indexer for Mathematical Formulas (Minor Project - I).**

We are also grateful to Dean SoCS UPES for giving us the necessary facilities to carry out our project work successfully. We also thanks to our Course Coordinator, **Mr Sandeep Chand Kumain** and our Activity Coordinator **Mr Suneet Gupta** for providing timely support and information during the completion of this project.

We would like to thank all our **friends** for their help and constructive criticism during our project work. Finally, we have no words to express our sincere gratitude to our **parents** who have shown us this world and for every support they have given us.

Name	**Siddharth Joshi**	**Bhavya Agrawal**	**Sumit Verma**
Roll No.	**R2142220664**	**R2142221157**	**R2142220184**

Contents

Abstract

Mathematical Information Retrieval (MIR) is a specialized domain within information retrieval focused on indexing, searching, and retrieving mathematical expressions from large datasets. The complexity of MIR lies in the unique syntactical and semantical structure of mathematical formulas, which differ significantly from natural language text. Traditional text-based retrieval methods fall short when applied to mathematical content due to the need to understand the precise mathematical meaning and relationships within expressions. In this study, we addressed these challenges by proposed a vector-based embedding technique that transformed mathematical formulas into a syntactically and semantically balanced format. Our method aims to capture the inherent structural and contextual meaning of mathematical expressions, making them accountable for effective retrieval. To validate our approach, we utilize the MathTagArticles of Wikipedia corpus of NTCIR-12 Dataset, a comprehensive collection of mathematical formulas from online forums. This dataset provides a rich ground for testing the retrieval capabilities of our embedding technique. We evaluate the performance of our method using standard precision metrics, specifically Precision at 5 (P@5), Precision at 10 (P@10).

Our approach leverages the MathTagArticles from the Wikipedia corpus of the NTCIR-12. Dataset, which comprises mathematical formulas in multiple formats, including Presentation MathML, Content MathML, and LaTeX. For this research, we specifically focused on Presentation MathML format, preprocessing the formulas and transforming them into binary vectors of 202 bits using a Bit Position Table. To enhance efficiency and retrieval performance, we applied a clustering technique to organize the data into 13 clusters. The system's effectiveness was evaluated using standard metrics, such as Precision at 5 (P@5) and Precision at 10 (P@10), demonstrating its capability to improve the retrieval of mathematical expressions. This study contributes to advancing MIR by optimizing formula representation and retrieval accuracy.

1 Introduction

Information Retrieval (IR) is a process to retrieve the relevant or needed information with respect to the user's entered query in a rank set of order from the collection of documents [12].An effective Information Retrieval system must organize and represent data in a way that ensures easy and effective access to meet user needs. IR focuses on the representation, storage, organization, and retrieval of information. The organization and representation of information should facilitate easy access for users to find what they need. Over the past 20 years, the field of IR has expanded beyond its initial goals of text indexing and document searching. Modern IR research now encompasses areas such as modeling, content classification and categorization, device architecture, user interfaces, data visualization, filtering, and language processing. Furthermore, the web has evolved into a global repository of text, images, and videos, enabling unparalleled exchange of ideas and knowledge on a massive scale. Its effectiveness is bolstered by a consistent user interface, regardless of the underlying computing framework. Additionally, developers can create their own website content and link to other pages without restrictions, making the web a modern, open publishing medium accessible to everyone. This accessibility allows any web user to easily share their ideas at minimal cost. Since its inception, this border-less nature has attracted billions of users worldwide, transforming how individuals interact with machines and perform daily tasks. Despite significant advancements, the web presents new challenges, particularly in finding valuable information, which remains time-consuming and difficult.

While searching for text, images, and videos using text queries is well-established and achieves state-of-the-art results, searching for mathematical or textual information using math-based queries is still in its early research stages. This area requires further development and the incorporation of innovative technologies. The continuous growth in Science, Technology, Engineering, and Mathematics (STEM) research generates

a vast number of scientific documents that include text, images, and mathematical formulas, necessitating advanced IR techniques to manage this complex information.[26].
To understand the concept behind scientific documents, mathematical formulas play a crucial role. The meaning of scientific documents emerges through the interaction of two contexts: textual and mathematical. In scientific texts, the surrounding text provides meaning to the formulas. For example, the Pythagoras Theorem [43] states that "In a right-angled triangle, the square of the hypotenuse side is equal to the sum of the squares of the other two sides." The sides of this triangle are referred to as Perpendicular, Base, and Hypotenuse. Based on this definition, the formula for the Pythagoras theorem is given as:

$$c^2 = a^2 + b^2 \tag{1}$$

In equation 1, based on the surrounding text, the variable a, b and, c have assigned a meaning Perpendicular, Base, and Hypotenuse. This behaviour of the textual and mathematical context witnessed their collaborative work to delivered the clear idea of scientific documents. Mathematical Information Retrieval enables the users to search for the mathematical formula and/or concept based on the text and/or math based query. Moreover, the retrieval of textual information is qualitatively and quantitatively different from the retrieval of mathematical information. Over the period of time, the MIR research witnessed the several preprocessing operations [39] to refine and makes the formula compatible for retrieval system. For instance: The documents on the web can emerge from different sources, which contains similar mathematical expressions with trivial differences. Canonicalization accesses the single canonical form for such syntactically similar formulas or sub-formulas and indexes them in same position. However, canonicalization operation is unable to perform the syntactic manipulation in the mathematical notation. For example, although the notations "x+y" and "y+x" are semantically same, canonicalization fails to consider equivalence of such notations. Hence, ordering of operand is recorded and such notations are converted into single canonical form. Moreover, mathematical information retrieval system does not restrict for exact match formula, for that the tokenization is performed to obtain the partial similar formula. Sometimes same mathematical equations can be represented with the different variable name, for example, $x^2 + y^2 = z^2$ which can be represented as $a^2 + b^2 = c^2$. The meaning of both equations is the same but different in their representation. The unification operation replaced the variable name with an identifier and numeric value with constant to find the unified formula. The unified form for both the mathematical equation $x^2 + y^2 = z^2$ and $a^2 + b^2 = c^2$ is $id_1^{const} + id_2^{const} = id_3^{const}$ which makes the bridge for the retrieval of semantically similar formula.

1.1 Digital Representation of Formula

Mathematical expressions are defined as a semi-formal visual language [24]. They serve as a graphical medium for depicting complex interactions among primitives, objects, and symbols [2]. Understanding mathematical knowledge involves dealing with a variety of challenges including a vast array of symbols, ambiguities, context-dependencies, layout issues, and semantics. Mathematical expressions use numerous terms to represent constants (e.g., π, e, 0), variables (e.g., α, β), operators (e.g., +, −), functions, and relations (e.g., $\int$, cos, <). The effectiveness of information retrieval systems largely hinges on the appropriate representation of information, which impacts the performance of algorithms, storage efficiency, and system accuracy. Typically, textual information is presented in a linear sequence (a sequence of words), while mathematical information can appear in a variety of formats [4].
LaTeX [25] is one of the most widely used digital formats for representing mathematical formulas. Compared to other formats, LaTeX presents mathematical formulas in a more compact manner. However, despite its simplicity in representation, processing LaTeX can be complex due to its package dependencies.

Additionally, LaTeX faces issues with ambiguity, as the same symbols can be represented either through Unicode or embedded commands.

Additionally, MathML [4] is another standard format for representing mathematical formulas, first proposed by the World Wide Web Consortium in 1998 as the recommended language for representing mathematical expressions in web documents. MathML offers several advantages over LaTeX, including its XML-based framework that simplifies algorithmic processing. MathML provides two methods for representing mathematical formulas: Presentation MathML and Content MathML.
Presentation MathML, also known as the symbol layout tree (SLT), focuses on the visual layout of the formula, while Content MathML, known as the operator tree (OPT), is concerned with the semantic meaning of the formulas, with nodes representing symbols and explicit aggregates [23]. OpenMath [3] is another widely used markup language for formula representation that can be combined with Presentation MathML to include additional semantic information. It is mainly used to define mathematical concepts within standard content dictionaries. The above-mentioned formats requires compiler to translate and process the mathematical formulae. Among these formats, MathML provides the finer representation length and easy to process.

2 Related Work

In recent years, there has been a growing interest in mathematical information retrieval, with many researchers worldwide contributing to this field. At NTCIR-10 math task [1], team MIRMU [20] conducted various preprocessing operations, converting mathematical information into a compact string. This preprocessed formula was then handled by a conventional full-text search mechanism, demonstrating that mathematical information retrieval is qualitatively and quantitatively different from textual information retrieval. A semantic enhancement technique [19] extracted valuable semantic information from the layout format, improving operand order optimization and generalization of substructures. Additionally, the hierarchical generalization of substructures generated index terms to support substructure and fuzzy matching. The approach of combining textual information with mathematical expressions [21] in documents and queries has shown superior performance in MIR systems. To retrieve exact matches, similar matches, and sub-formulas, a query expansion technique was used, resulting in the best search results with a MAP of 0.215.
The WikiMir system of ICST [11] utilizes a keyword-based approach that takes into account the structural information of formulas and their significance within a document. This system introduces an innovative hybrid indexing and matching model to support both exact and fuzzy matching. To enhance the relevance of the results, the system re-ranks the top-k formulas by matching the query formula with regular expressions.
At NTCIR-12, team FSE [37] implemented a straightforward method to manually perform the Wiki-main task. Physicists and computer scientists examined the queries and entered the titles of related Wikipedia pages into the search interface at *en.wikipedia.org*. For certain queries, FSE utilized the German version of Wikipedia and used inter-language links to find the corresponding English Wikipedia page. Also at NTCIR-12, team MCAT introduced an indexing technique based on the Apache Solr database [15]. The MCAT search system processed textual information in three levels of granularity: math, paragraph, and document. For mathematical information, it leveraged the dependency relationships between math expressions, score normalization, cold-start weights, and unification operations. The dependency relationships and unification significantly enhanced search accuracy. However, the cold-start weights did not positively impact search performance due to negative weights in several fields within their database.
The MIaS system developed by team MIRMU [35] leverages the relevance judgments from the NTCIR-11 Math-2 Task. Additionally, an evaluation platform was created to rigorously assess combinations of new features and identify the most promising ones for the NTCIR-12 evaluation. The main objectives of these new features include further canonization of MathML input, structural unification of formulas for

syntactically similar searches, and query expansion to improve results for combined text and math queries. The Tangent-3 system from team RITUW [8] employed two indices: a Solr-based index for textual information and a custom inverted index for mathematical information. For efficient searching, the custom inverted index utilized token pairs of symbols and their spatial relationships. Text and math indices were queried separately, and documents were ranked based on a linear combination of math expression and keyword similarity scores, with equal weights assigned to both keyword matches and formula matches. Constraining unification positively affected formula retrieval, along with advanced similarity metrics that better utilized the high recall for formulas returned from the index. The Math Search System of SMSG5 [42] used Elasticsearch as the primary ranking mechanism. To refine this, an innovative ranking technique was introduced to re-rank documents and formulas, namely the Borda countbased hybrid ranking technique, which is based on the doc2vec model, latent Dirichlet allocation algorithm, and a pattern-based approach.
For applications such as information retrieval, document clustering, or classification, capturing semantics is often done probabilistically. The initial step typically involves representing words or documents in a vector space. Recently, neural network approaches like word embedding [17] have become important for representing textual content, as they can identify distributed word representations and capture semantic similarities between words. Concurrently, various neural retrieval methods have been developed, with some showing significant performance improvements [28]. For example, the EqEmb model [16], based on the concept of word embeddings (Bernoulli embeddings [34]), takes a dataset of words and equations as input to derive semantic representations. The EqEmb model posits that a good semantic representation of equations can be achieved by expanding the original word context to include any equations appearing within a larger window. Additionally, the EqEmb-U model [16] (Equation unit embedding) treats mathematical formulas as sentences, where the words are the equation variables, symbols, and operators, referred to as units.

The advanced version of the MIaS system [40] handles text and math data separately. The text data is tokenized, and words are stemmed to their base forms. MIaS requires math data in MathML format and utilizes open tools such as Tralics[1] and LATEXML[2] to convert LaTeX formulas to MathML. The math data is then preprocessed through canonicalization, ordering, tokenization, and unification. MIaS also features a graphical interface that allows users to input queries combining text and math, with native LaTeX support provided by Tralics and MathJax
[5].

In the retrieval of mathematical information, both formulas and textual content are crucial for achieving state-of-the-art results and meeting users' needs, whether the query is formula-based, text-based, or a combination of both (formula + text). The variable typing approach [41] assigns meaning to variables using the surrounding text within the same sentence. Types are multi-word phrases typically used to denote mathematical terminologies such as objects (e.g., "set"), algebraic structures (e.g., "monoid"), and instantiable notions (e.g., "cardinality of a set"). To assess the effectiveness of the variable typing approach, two baseline systems—nearest type and the SVM proposed by [13][14]—along with three newly proposed methods—an extended version of the SVM baseline, a convolutional neural network, and a bidirectional LSTM [18][33]—were employed. Among these approaches, the bidirectional LSTM achieved the most remarkable results.
To develop better models in the MIR domain, a signature-based hash indexing approach [9] offers a more suitable alternative to text-based models. This method involves obtaining mathematical formulas from scientific documents and converting them into structured encoded strings (SES). These strings are then used

[1] https://www-sop.inria.fr/marelle/tralics/
[2] https://dlmf.nist.gov/LaTeXML/

as input for a hash-based indexing technique, which transforms them into bit vectors or signatures. These bit vectors are compared with the user-entered query to retrieve relevant search results.

In formula-based search engines, vector-based approaches have demonstrated significant performance. For example, the Binary Vector Transformation of Math Formula (BVTMF) approach [31] extracts MathML formulas from documents, preprocesses them, and creates fixed-sized binary vectors where each mathematical symbol's presence is represented by '1' and its absence by '0'. It also processes textual information using Apache Lucene[3]. Both textual and mathematical information are retrieved independently and documents are ranked based on assigned priorities.

Inspired by the strong performance of LSTM for sequence-to-sequence tasks, the LSTM-based Formula Entailment (LFE) approach [29] has effectively identified entailment between formulabased user queries and formulas in scientific documents. The LFE approach was trained and validated using a symbol-level Math Formula Entailment (MENTAIL) dataset.

At ARQMath-2020 [46], the DPRL research team from Rochester Institute of Technology's Pattern Recognition Lab introduced the Tangent-CFT system [22]. This system uses both SLT and OPT representations to account for both the appearance and syntax of formulas. As an extension of Tangent-CFT, Tangent+CFT employs two vector representations for each formula: Formula Vector and Text Vector. The Formula Vector is derived using Tangent-CFT, while the Text Vector represents formulas by treating them as words within their surrounding context, utilizing the fastText model.

Additionally, team MIRMU [27] presented two approaches for formula retrieval: Soft Cosine Measure (SCM) and Formula2Vec. The SCM approach integrates TF-IDF with unsupervised word embeddings to create interpretable representations of math documents and formulas. On the other hand, the Formula2Vec approach uses Doc2Vec to generate embeddings for documents and formulas, employing the Doc2Vec DBOW model for this purpose.

This work proposes a novel approach for mathematical formula retrieval by converting formulas into binary vectors using a Bit Position Information Table. The binary vectors enable efficient similarity-based retrieval. Despite its innovative approach, the study highlights the absence of several key mathematical entities such as % and &, which can affect the robustness of the retrieval system.[30]

Building on the earlier work, this study refines the transformation of mathematical formulas into binary vectors. It assigns equal importance to operands and operators while factoring in formula context. However, the lack of differentiation in priority between these elements can lead to reduced precision in capturing the relevance of mathematical entities.[31]

This paper enhances formula retrieval by representing formulas as Symbol Layout Trees (SLTs) and Operator Trees (OPTs). It employs embedding models for vector representation, using n-gram embedding to capture formula structure. However, the n-gram approach, while effective for local sequences, overlooks broader semantic relationships, potentially limiting comprehensive formula understanding.[23]

The integration of the CA-YOLOv5 model with HFS facilitates efficient recognition and retrieval of mathematical expressions from layout images. This method effectively handles indexed expression similarity evaluation. Nevertheless, YOLO-based models struggle with documents having complex layouts, such as overlapping symbols, which limits detection accuracy. [44]

MathBERT is a pre-trained model developed to understand mathematical formulas and their structural and semantic relationships. The model uses OPT for semantic representation, enabling robust comprehension. However, its reliance solely on OPT neglects visual layout information, which might be essential for tasks requiring appearance-based analysis.[32]

This study introduces a retrieval method combining hesitant fuzzy sets (HFS) and local semantic distillation (LSD). The approach leverages spatial and semantic features of symbols and refines pre-trained language

[3] http://lucene.apache.org/

models for semantic matching. However, its dependency on predefined attributes within HFS restricts its ability to capture nuanced relationships [10].
MathUSE employs the Universal Sentence Encoder to generate fixed-dimensional embeddings for mathematical formulas using their LaTeX representations. While effective for formulabased retrieval, it does not adequately explore the combined potential of text and formula integration, nor does it compare performance with advanced transformer-based models [7].
This study evaluates the mathematical proficiency of six LLMs, including GPT-4 and others, in generating answers and retrieving relevant results. While the embedding-based approach simplifies mathematical expressions for processing, it risks losing critical syntactic and structural nuances essential for accurate retrieval and reasoning [36].

3 Problem Statement

Followings are the challenges that shows the varied range of problems faced by traditional information retrieval system to retrieve the mathematical information [38].

- Mathematical formulas are inherently recursive, while textual information follows a linear structure.
- The conventional search engine have unavailability of math editor to enter the formula based query.
- In mathematical information, some mathematical notations posses the alternative representation like a permutation of k objects chosen from n distinct objects have several representations: P_k^n, ${}_nP_k$, $P_{n,k}$, ${}^nP_k$ and P(n, k). The conventional search engine unable to treat all such differently expressed formulae as equivalent.
- Formula are ambiguous in nature for instance f(x) which means that either multiplication of variables ‘f’ and ‘x’ or a function of ‘x’. Sometimes, the formula have different representation structure but holds same meaning for instance $a^2 + b^2$ and $x^2 + y^2$.
- Mathematics employs numerous symbols to express scientific concepts and ideas, with these symbols varying in scripting styles and typefaces based on different formula representation strategies, leading to symbol identification challenges.
- L^AT_EX and MathML are very common representation formats used by the researcher to illustrate the mathematical formulae. The existing text-based search engines are efficient to handle the strings of character, but insufficient to handle the scripting style of L^AT_EX and MathML.
- Users may sometimes be unaware of the exact form of the information they need. Therefore, a Mathematical Information Retrieval system should be capable of searching for subformulas, similar formulas, and parent formulas to fully meet user needs.
- LaTeX, MathML, OpenMath, SGML, and Unicode are common encoding techniques for representing mathematical formulas. These techniques often include redundant elements and attributes, which contribute to the creation of syntactically and semantically rich formulas. Consequently, existing Information Retrieval and MIR systems must rigorously reprocess these formulas to handle redundancy and make the data suitable for IR and MIR systems.

4 Motivation

In recent years, the emergence of full-text math sources has opened up exciting opportunities for advancing the processing of mathematical knowledge. It is widely believed that directly searching for mathematical information, such as formulas, can provide both effective and enhanced access to mathematical knowledge. Therefore, advancements in formula representation, formula-context mapping, and ranking have generated enthusiasm for retrieving mathematical knowledge beyond traditional library and information science

communities. This indicates that the time is ripe for innovative methods in mathematical retrieval. These methods are particularly relevant for the use of scientific documents, which are often not retrieved by traditional systems, and highlight that AI and NLP offer distinct strategies for classifying and searching mathematical information. In the future, successful search and retrieval of mathematical information will enhance areas like math recommender systems, mathematical plagiarism detection, and mathematical answer finding.

5 Objectives

- To perform formula preprocessing operations.
- To design a formula representation and indexing approach.
- To design a crawling and indexing approach for mathematical information retrieval.
- To design a mathematical information retrieval approach for formula-based query.

6 Data Description

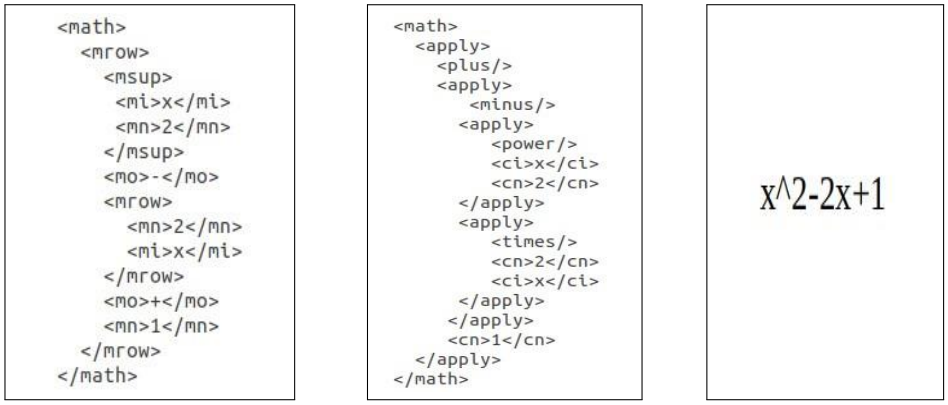
The MathTagArticles of Wikipedia corpus of NTCIR-12 [4] [45] contains 31,839 math articles which constitutes the 579,608 formulas. The documents contained in MathTagArticles, constitutes a textual as well as mathematical information. In math articles of Wikipedia, the mathematical information has been represented in three distinct formats namely Presentation MathML, Content MathML and $LaTeX$ format. The metadata about the corpus is shown in Table **??**. The math formulae inside the articles are written in three distinct forms: Presentation MathML, Content MathML and $LaTeX$. The Figure 1 shows the three distinct forms for mathematical expression $x^2 - 2x + 1$ where Figure 1(a) represents the formula in Presentation MathML format, Figure

Table 1: MathTagArticles of Wikipedia of NTCIR-12 Corpus Description

Source	MathTagArticles of Wikipedia of NTCIR-12
Size	2.1 GB
No. of Articles	31,839
Article Format	HTML
Formula Format	$LaTeX$, Presentation MathML, and Content MathML
No. of Formulas	5,79,608
No. of Test Queries	40 (Wikipedia Formula Browsing Task)

Figure 1(b) represents the formula in Content MathML format and Figure 1(c) represents the formula in $LaTeX$ format.

[4] http://ntcir-math.nii.ac.jp/data/



(a)PresentationMathML (b)ContentMathML (c) $L^A T_E X$ Format

Figure 1: Different Forms to Represent Math Formula

7 System Architecture

The proposed system architecture primarily focuses on vector generation using formula embedding and optimizing these vectors through clustering techniques. The architecture is divided into two components: the client side and the server side. Each module in the proposed system architecture operates sequentially and cooperatively. A pictorial representation of the proposed system architecture is provided in Figure 2. The functionality of each module in the proposed system is explained in detail in the subsections below.

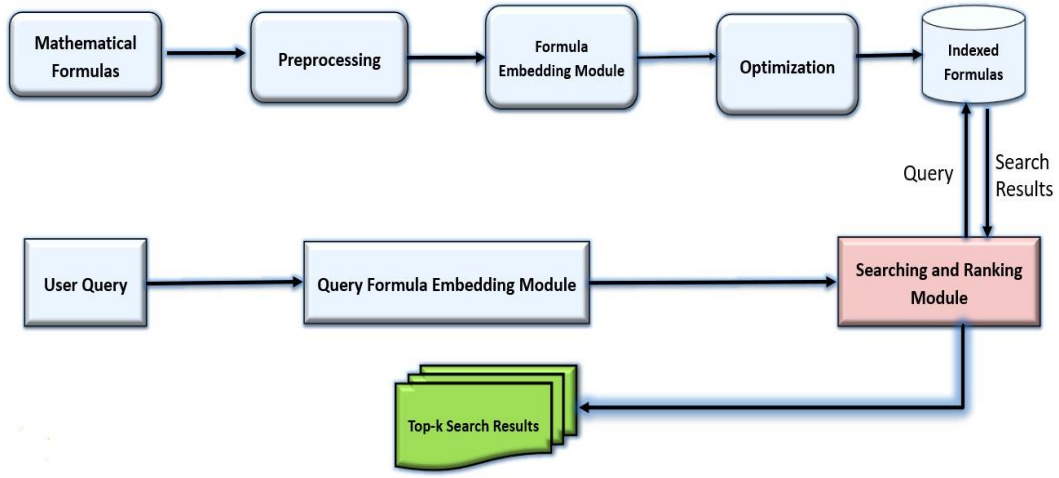


Figure 2: System Architecture

7.1 Formula

In the MathTagArticles of Wikipedia from NTCIR-12, mathematical formulas are represented in three distinct formats: Presentation MathML, Content MathML, and LaTeX. Each format has its own unique syntax tailored to its representation style. Researchers in the field of Mathematical Information Retrieval have developed various retrieval techniques to handle these formats effectively. They found that rule-based techniques are best suited for Presentation MathML, tree-based techniques are ideal for Content MathML, and deep learning approaches work well for LaTeX.

1. **Rule-Based Technique:** A rule-based technique refers to the approach used to represent and display mathematical notations based on predefined rules or patterns that define how symbols, operators, and structures are arranged visually. Rule Based Technique is better for presentation MathML.
2. **TreeBasedTechnique:** A tree-based technique is used to represent the semantic structure of mathematical expressions. Content MathML emphasizes the meaning and logical structure of the math, often represented as a tree.
3. **Deep Learning Technique:** In LaTeX, a deep learning technique is often employed for tasks like automatic equation recognition, document parsing, and natural language processing-based text generation. While LaTeX itself is a typesetting system and doesn't inherently involve deep learning, various tools and extensions integrate deep learning to process LaTeX documents, especially for automating and enhancing tasks related to math and text handling.

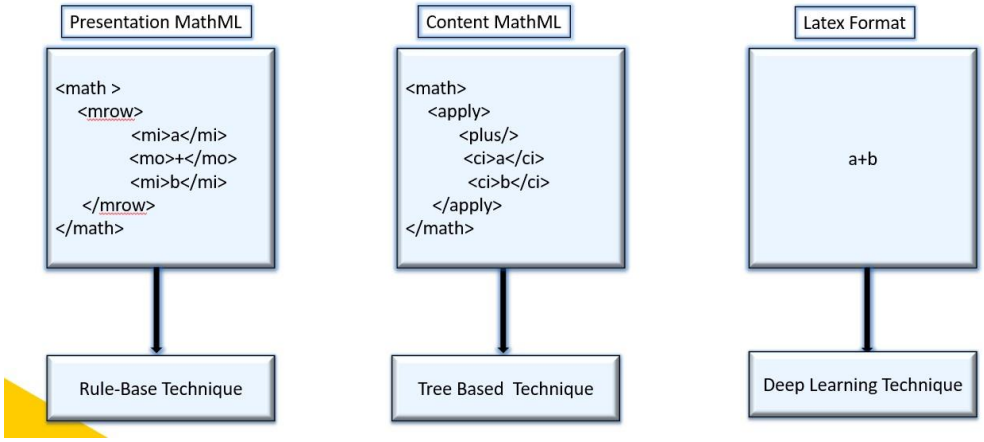


Figure 3: Representations of mathematical formulas

7.2 Document Preprocessing

Document preprocessing is a crucial step in the development of the Mathematical Information Retrieval system, as it ensures that only relevant mathematical data is extracted and processed, optimizing the system's performance. The given data consists of both textual and mathematical information, with the mathematical content available in multiple formats such as presentation MathML, content MathML, and LaTeX. For the purpose of our project, we focus on processing mathematical formulas in the presentation MathML format, as it aligns best with our rule-based technique.

Extraction of Mathematical Formulas The preprocessing begins with isolating the mathematical formulas from the textual data. This is achieved by extracting formulas in their respective formats—presentation MathML, content MathML, and LaTeX—from the input documents. These formulas serve as the foundation for further analysis and processing. Among the extracted formats, presentation MathML was selected for subsequent steps because it represents the visual structure of the formulas, which is well-suited for rule-based matching and categorization.

ProcessingPresentationMathML Once the mathematical formulas in the presentation MathML format are identified, they undergo a standardization process to prepare them for efficient comparison and retrieval. The standardization involves simplifying the structure of the MathML representation by removing redundant attributes and tags that do not contribute to the core semantics of the formula. This ensures uniformity in representation, reduces noise, and improves computational efficiency.

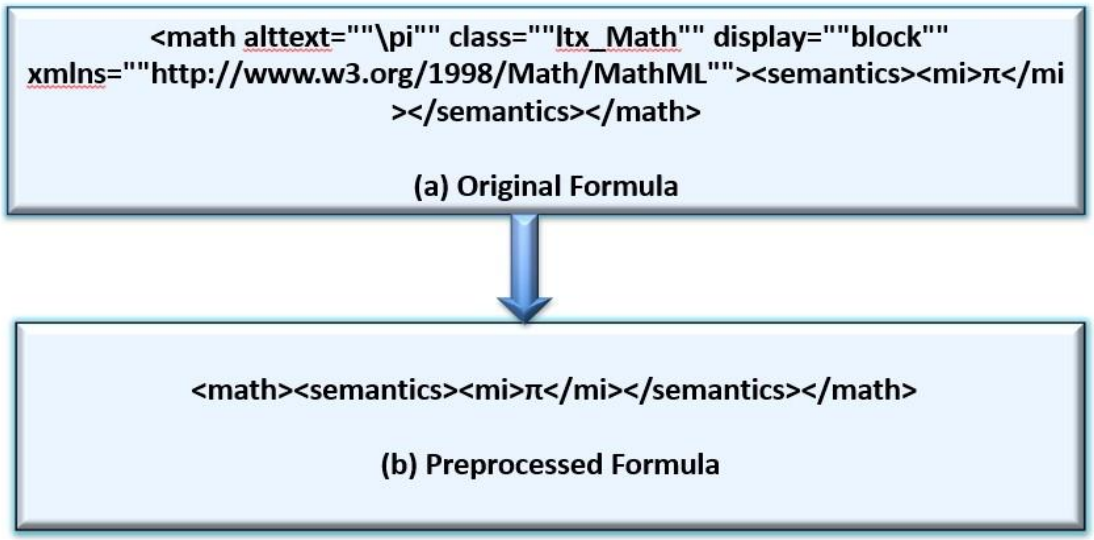


Figure 4: Workflow of the preprocessing module

7.3 Formula Embedding

The formula embedding module takes processed formula as input and outputs corresponding binary vector.Careful observation of the Presentation MathML formulae guides us to the fact that <mi> and <mo> elements form the key constituents of nearly all formulae. Thus, the module parses processed MathML formula to fetch contents from all occurrences of <mi> and <mo> elements. Besides, the module also accounts for all other MathML elements, such as <msqrt>, <msub>, <msup> and so on, present in the formula. Eventually, the module uses fetched information contents and the bit position information table to set respective bits of the corresponding formula vector, which has a fixed length of 202 bits.
A bit position information table (figure 5) depicts bit positions assigned to different mathematical entities. A vector of length 202 suffices to encode mathematical contents of all the documents in our corpus. However, the module offers flexibility to vary (increase or decrease) length of formula vector, if required. Figure 5 shows typeset representation and formula vector a processed MathML formula.
The following points about the bit position information table and formula vector:-

1. Bit position 0-25, 57-65 and 71-100 represents to <mi> (mathematical identifier) tag. Bit position 26-45, 66-70 and 101-149 refer to <mo> (mathematical operator) tag and bit position 46-56 refer to essential MathML tags which contribute to the semantic formula.
2. In future version of the proposed system, the length of formula vector will be extended to account for more math symbols and their semantics.
3. In order to retrieve specific results ahead of non-specific ones, distinction is maintained between single alphabet variables, by assigning them different bit positions ranging from 0-25. However, the table makes no distinction between the cases of variables, and the different cases of same variable are assigned same bit position. For example, ‘a’ and ‘A’ are assigned same bit position to 0. Since the entities ‘exp’ and ‘e’ are interchangeably used, they have been assigned same bit position equal to 4.
4. For the same reason ‘log’ and ‘ln’ are assigned same bit position. For the query system term involving trigonometric ratio such as “sin”, “cos”, “tan”, “cot” and so on. Thus, all such ratios are assigned same bit position equal to 90.

5. An entity which can be a part of <mi> as well as <mo> tags, i.e. the one which can act as variable as well as operator, is assigned two distinct bit position. One such entity is Σ which is assigned bit position 76 and 112, designating variable and operator respectively.
6. Bit position 65 designates a multi character variable, whose name is not a standard variable name (such as lim, log, gcd and so on).
7. The formula vector does not account for multiple occurrences of the entities. Even though an entity occurs more than once in the formula, only once corresponding bit of the formula vector is set. This limitation, however, causes inability to retrieve relevant search results if the user query contains repetition of an entity.

7.4 Optimization and Index

Optimization is a critical component in developing the Mathematical Information Retrieval system to ensure efficiency and scalability, particularly when handling large-scale formula comparisons. Initially, the system relied on a brute-force comparison approach that required a computationally expensive process of evaluating each user query against a database containing 579,608 formulas, each represented by a 202-bit vector. This resulted in an overwhelming computational complexity, calculated as: $1 query \times 579{,}608 formulas \times 202 - bit vector$ This naive approach posed significant challenges in terms of response time and computational resource usage, especially as the size of the formula database expanded. A clustering-based optimization technique was introduced to address these issues and improve performance.

Clustering-Based Optimization Technique:

Clustering is a powerful method for reducing the computational load by grouping similar formulas into predefined categories, thereby limiting the scope of comparison for a given query. For this system, the formula database was organized into 13 distinct clusters based on mathematical categories. These clusters are: Arithmetic, Calculus, Statistics, Measurement, Letter-like Symbols, Set/Logic, Geometric, Equivalence, Latin/Greek Symbols, Digits/Numerals, Structural Symbols, Arrows, and Trigonometric. By categorizing formulas, the system reduces the overall number of comparisons. A user query is first classified into one of these clusters using rule-based techniques, and comparisons are limited to the formulas within the corresponding cluster rather than the entire database.

Table 2: Bit Position Information Table

Entity	Position	Entity	Position	Entity	Position	Entity	Position
a/A to z/Z	0-25	<msqrt>	56	*log, ln*	88	‡	120
exp	4	Z	57	!	89		121
=	26	N	58	Trigo. Ratios	90	°	122
Product	27	Q	59	R	91	⋀	123
-	28	¬	60	ϑ	92		124
,	29	α	61	gcd	93	∃	125
+	30	γ	62	xor	94	lim	126
∇	31	ω	63	τ	95		127
∂	32	ϑ	64	η	96	<, ⟨	128
→	33	Var. Name	65	σ	97	>, ⟩	129
.	34	Null	66	Ω	98	⊗	130
(	35	†	67	#	99	▸	131
)	36	:	68	⌜	100	H	132
≡	37	dist	69	⁄=	101	H	133
	38	∓	70	{	102	⋃	134
∝	39	$\phi\,\varphi, \Phi$	71	}	103	∩	135
≈	40	$\hbar$	72	Ⓢ	104		136
/	41	π	73	≤	105	⃗	137
⊆	42	Δ	74	∈	106	⊂	138
⊕	43	μ	75	[	107	det	139
~	44	Σ	76	]	108	∏	140
|	45	ϵ	77	*, x	109	mod	141
<mfrac>	46	. . .	78	∉	110	sup	142
<mn>	47	δ	79	^	111	≥, ⩾, ≳	143
<msub>	48	ψ, Ψ	80		112	dim	144
<msup>	49	Γ	81	∑	113	:=	145
<msubsup>	50	∞	82	;	114	≅	146
<mover>	51	ρ	83	⇐⇒, ⇔	115	max	147
<munderover>	52	β	84	⇒	116	inf	148
<munder>	53	λ	85	⌈	117	min	149
<mtable>	54	ξ	86	|	118	a/A to z/Z present in superscript	150-175
<mmultiscripts>	55	□	87	∀	119	a/A to z/Z present in subscript	176-201

Reduction in Time Complexity:

Before clustering, the time complexity of handling a single query was proportional to the product of the total number of formulas and the vector size. After clustering, the complexity is significantly reduced as follows:

1. The query is first assigned to a cluster based on its characteristics, a process with negligible overhead due to efficient rule-based techniques.
2. Comparisons are limited to the formulas within the selected cluster, drastically reducing the number of computations.

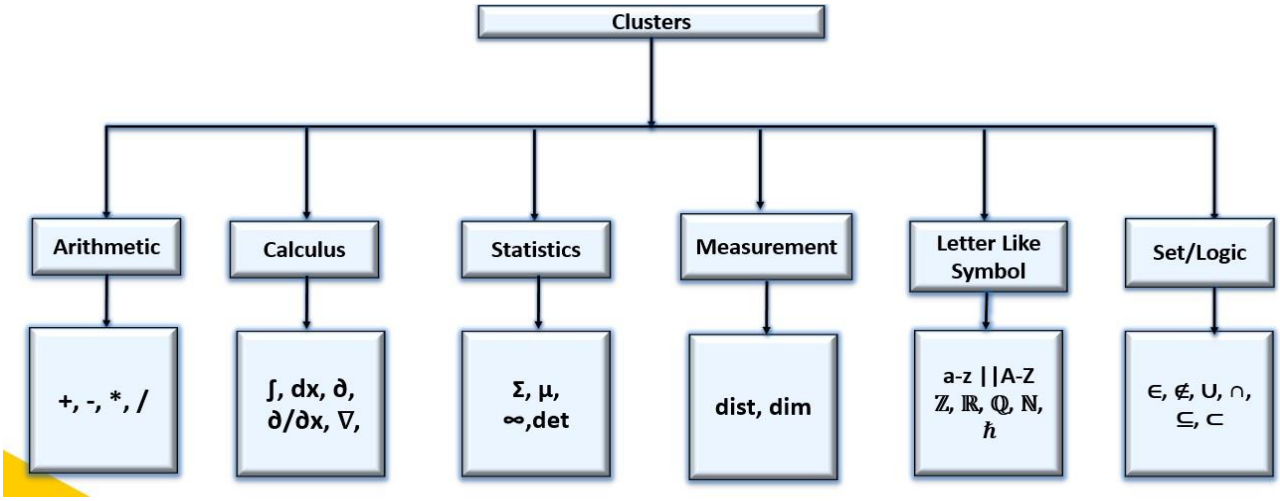


Figure 5: Clusters

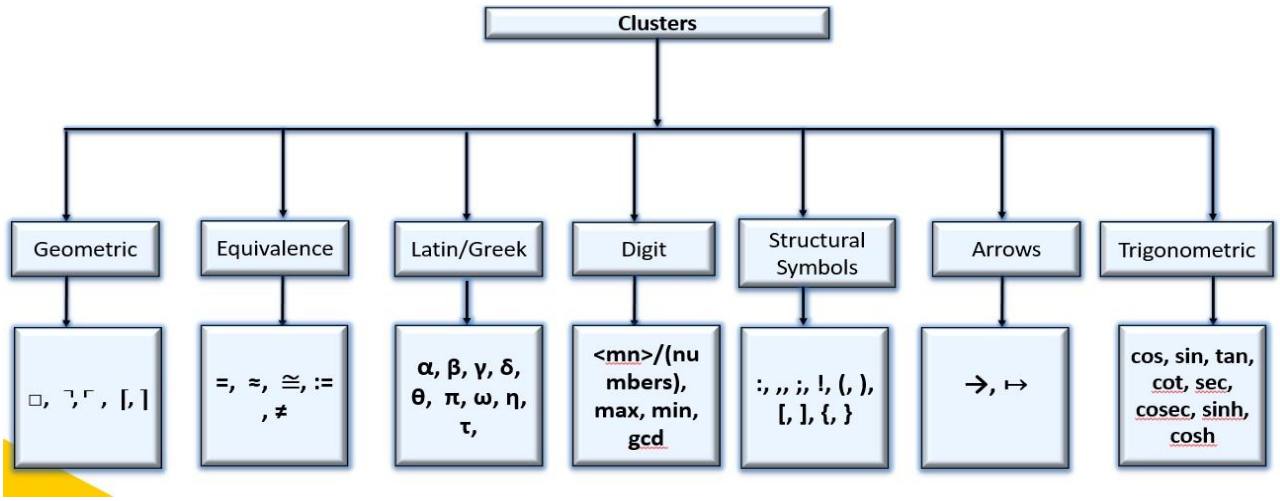


Figure 6: Clusters

1. **Arithmetic** Arithmetic operators are fundamental mathematical symbols or functions used to perform calculations like addition, subtraction, multiplication, and division on numerical values.

   Arithmetic Operators:

   - Addition: +
   - Subtraction: −

- Multiplication: ∗ or ×
- Division: /
- Product: ∏
- Modulus: mod

2. **Calculus** Calculus is a branch of mathematics that focuses on studying change and motion. It is divided into two main branches: Differential Calculus: Concerned with the rate of change of quantities (e.g., slopes, derivatives). Integral Calculus: Concerned with the accumulation of quantities (e.g., areas under curves, integrals).
   - Nabla (Gradient): ∇
   - Partial Derivative: ∂
   - Limit: limx→a
   - Integral: ∫
   - Differential: dx
   - Determinant: det
   - Partial Derivative with Respect to x: $\frac{\partial}{\partial x}$
3. **Statistics** Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, presenting, and organizing data. It provides tools and methods for making informed decisions based on data.
   - Summation: Σ
   - Mean (mu): μ
   - Infinity: ∞
   - Determinant: det
4. **Measurement:** Measurement is the process of quantifying an object's physical properties or attributes using a standard unit. It provides a numerical value to represent length, mass, time, temperature, or other quantities.
   - Distance (dist): dist
   - Dimension (dim): dim
   - Top-left corner symbol (): ⌜
   - Top-right corner symbol (): ⌝
5. **Letter Like Symbol:** Letter-like symbols are a special category of symbols frequently used in mathematics andscientific notation to represent distinct sets, constants, and variables. These symbols oftenappear in a stylized format to differentiate them from regular letters, providing clarity andavoiding confusion in complex expressions.

$\mathbb{Z}$, $\mathbb{R}$, $\mathbb{Q}$,$\mathbb{N}$, $\bar{h}$, $a \parallel A$, $b \parallel B$, $c \parallel C$, $d \parallel D$, $e \parallel E \parallel \exp$, $f \parallel F$, $g \parallel G$, $h \parallel H$, $i \parallel I$, $j \parallel J$, $k \parallel K$, $l \parallel L$, $m \parallel M$,$n \parallel N$, $o \parallel O$, $p \parallel P$, $q \parallel Q$, $r \parallel R$, $s \parallel S$, $t \parallel T$,

$u \parallel U$, $v \parallel V$, $w \parallel W$, $x \parallel X$, $y \parallel Y$, $z \parallel Z$

6. **Set/Logic:** Set theory introduces operations like union, intersection, difference, and complement to combine or compare sets. For example, the union of two sets includes all their elements, while their intersection consists of elements common to both. Subsets, power sets, and Cartesian products are other key concepts that extend the utility of sets. Sets play a cru- cial role in various fields, such as defining events in probability, structuring data in computer science, and forming the basis of abstract algebra. By organizing elements systematically, sets provide a versatile framework for reasoning and problem-solving in mathematics and beyond.

   - $\in$ \in (Element of)
   - $\notin$ \notin (Not an element of)
   - $\cup$ \cup (Union)
   - $\cap$ \cap (Intersection)
   - $\subseteq$ \subseteq (Subset or equal)
   - $\subset$ \subset (Proper subset)
   - $\parallel$ \parallel (Parallel)
   - $\land$ \land (Logical AND)
   - $\lor$ \lor (Logical OR)
   - $\neg$ \neg (Logical NOT)
   - $\forall$ \forall (For all)
   - $\exists$ \exists (There exists)
   - $\otimes$ \otimes (Tensor product)
   - $\oplus$ \oplus (Direct sum)

- ⊙. \odot (Circle dot)
- ⊢ \vdash (Proves)
- ⊓ \sqcap (Square intersection)
- ⊔ \sqcup (Square union)
- ∼ \sim (Tilde or equivalence)
- xor xor (Exclusive OR)
- ∅ \emptyset (Empty set)
- inf \inf (Infimum)
- sup \sup (Supremum)

7. **Geometric:** Geometry is a branch of mathematics that deals with the study of shapes, sizes, and the properties of space. It provides the foundation for understanding the physical world by describing the relationships between points, lines, surfaces, angles, and solids. Geometry is divided into various subfields, such as plane geometry, which focuses on two-dimensional shapes like triangles, circles, and polygons, and solid geometry, which deals with three-dimensional objects like spheres, cubes, and cylinders.
   - Square symbol (□): □
   - Top-right corner symbol (): ⌝
   - Top-left corner symbol (): ⌜
   - Tensor product symbol (): ⊗
   - Maps-to symbol (): ↦
   - Composition symbol (): ∘
   - Left ceiling symbol (): ⌈
   - Right ceiling symbol (): ⌉
8. **Equivalence** Equivalence in mathematics refers to a fundamental concept where two ob- jects, expressions, or structures are considered to have the same value, properties, or function under a specific context or set of rules. The idea of equivalence is crucial across various mathematical domains, enabling simplification, categorization, and abstraction. Equiva- lence also arises in algebra, calculus, and logic. In algebra, two expressions are equivalent if they yield the same value for all permissible variable assignments. In calculus, equiv- alence of functions may involve comparing their limits, derivatives, or integrals. Logical equivalence occurs when two statements are true under the same conditions.
   - Equal symbol (=): =
   - Approximately equal symbol (≈): ≈
   - Congruent symbol (): ≅
   - Defined as symbol (:=):

- Not equal symbol (≠): /=
- Identically equal symbol (≡): ≡
- Precedes or equal symbol (): ≤
- Less than or equal to symbol (≤): ≤
- Precedes symbol (): ≺
- If and only if symbol (): ⇔
- Greater than or equal to symbol (≥): ≥
- Succeeds symbol (): >

9. **Latin/Greek:** The Greek alphabet is widely used in mathematics today, with symbols such as (pi), (al- pha), and (sigma) representing constants, angles, and summations. Latin, the language of scholarly communication during the Renaissance, played a key role in disseminating Greek mathematical works throughout Europe, often translating and preserving Greek texts. Terms like ”radius,” ”quadratic,” and ”integer” are derived from Latin, reflecting its influence on mathematical vocabulary.

    - Alpha (α): α
    - Beta (β): β
    - Gamma (γ): γ
    - Delta (δ): δ
    - Theta (θ): θ
    - Pi (π): π
    - Omega (ω): ω
    - Eta (η): η
    - Tau (τ): τ
    - Theta variant (): ϑ
    - Capital Gamma (Γ): Γ
    - Capital Psi (Ψ): Ψ
    - Rho (ρ): ρ
    - Sigma (σ): σ
    - Capital Delta (Δ): Δ
    - Exponential function (exp): exp
    - Null set (Null or): ∅ or ∅
    - Epsilon (): ϵ or ε
    - Xi (ξ): ξ

- Lambda (λ): λ
- Mu (μ): μ
- Phi (): ϕ or φ
- Empty set or null set (): ∅
- Capital Omega (Ω): Ω

10. **Digit/Numeral:** In mathematics, digits and numerals are fundamental concepts, yet they serve distinct pur-poses. A digit refers to any of the ten basic symbols in the decimal system: 0, 1, 2, 3, 4, 5,6, 7, 8, and 9. Digits are the building blocks of numbers and play a crucial role in forming numerals and representing quantities. A numeral, on the other hand, is a representation of a number using one or more digits. For instance, ”5” is both a digit and a numeral, while ”25” is a numeral made up of two digits.
    - Division (e.g., m/n): $\frac{m}{}$ or m/n
    - Maximum (max): max
    - Minimum (min): min
    - Greatest common divisor (gcd): gcd
11. **Structural Symbols:** Structural symbols in mathematics are essential notations that define relationships, organize expressions, and establish the framework for mathematical reasoning. Unlike numbers or variables, structural symbols do not carry numerical value but provide the syntax necessary for mathematical operations. Common examples include parentheses (), brackets [], and braces , which are used to group terms, dictate the order of operations, and simplify complex expressions.
    - Colon (:): :
    - Comma (,): ,
    - Semicolon (;): ;
    - Exclamation mark (!): !
    - Left parenthesis ((): (
    - Right parenthesis ())):)
    - Left bracket ([): [
    - Right bracket (]):]
    - Left brace (): {
    - Right brace (): }
    - Left angle bracket (): ⟨

- Right angle bracket (): ⟩
- Less than (’<’): <
- Greater than (’>’): >
- Vertical bar (|): | or |
- Overline (¯): $\overline{x}$
- Ceiling left (): [
- Ceiling right (): |

Mathematical Structures latex Copy code

- Fraction (< mfrac >): $\frac{a}{}$
- Subscript (< msub >): x_a
- Superscript (< msup >): x^a
- Subscript and superscript (< msubsup >): x^b
- Over (< mover >): $\overset{\text{label}}{}$
- Under and over (< munderover >): $\underset{b}{\overset{a}{}}$
- Under (< munder >): $\underset{\text{label}}{x}$
- Table (< mtable >): Use an array:

$$\begin{array}{c|c} a & b \\ c & d \end{array}$$

- Multiscripts (< mmultiscripts >): Use scripts with _ and^: x^b
- Square root (< msqrt >): $\sqrt{x}$

12. **Arrows:** In mathematics, the arrow is a versatile symbol that carries a range of meanings dependingon the context, serving as a visual representation of direction, transformation, or mapping.
    - Right arrow (→): →
    - Maps-to symbol (): ›→

13. **Trigonometry:** Trigonometric concepts are foundational to mathematics, dealing with the relationships be-tween the angles and sides of triangles, particularly right-angled triangles. Trigonometry isa branch of mathematics that explores the relationships between the angles and sides of tri-angles, with its applications extending into virtually every field of science and engineering.Rooted in geometry, trigonometry was initially developed to solve problems in astronomy,such as calculating the positions of stars and planets, and later expanded to include broadermathematical applications.

- cos: Cosine
- sin: Sine
- tan: Tangent
- cot: Cotangent
- sec: Secant
- csc: Cosecant
- sinh: Hyperbolic Sine
- cosh: Hyperbolic Cosine

8 Experimental Results and Analysis

8.1 Trec Eval

It is the standard tool used by the TREC community for evaluating an ad hoc retrieval run, given the results file and a standard set of judged results. A retrieval system takes the input of some information need represented by a query and generates a list of documents that are relevant to that query.

8.2 Precision

Precision are used to calculate the performance of information retrieval systems. Precision is defined as the fraction of relevant items among all retrieved items. The mathematical definition of the Precision are shown in below equations:

The formula for precision is:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Positives (FP)}}$$

Where, TP is the true positive rate, that is the number of items which are relevant and which the model correctly considered as relevant. fp is the false positive rate, that is the number of items which are not relevant but which the model incorrectly considered as relevant.

8.3 Experimental Results

Impact of Clustering on System Performance: The introduction of clustering has had a profound impact on the overall performance of the MIR system. Key benefits include:

- **Improved Query Response Time**: By narrowing down the search space, the system processes user queries much faster, resulting in a more responsive user experience.

Table 3: Comparison table of proposed approach with the existing results

Method	**P@5**	**P@10**
Method 1 [31]	0.64	0.54
Method 2 [20]	0.34	0.26
Method 3 [6]	0.37	0.31
Proposed Approach	0.86	0.60

- Reduced Computational Overhead: The optimization minimizes the resources required for formula comparison, enabling the system to handle larger query volumes or scale the formula database without degradation in performance.
- Scalability: With the clustering approach, the system is better equipped to accommodate future expansions in the formula database while maintaining efficient operations.
- Enhanced Categorization: The cluster-based approach aligns with the logical organization of mathematical formulas, improving the accuracy and relevance of retrieval results.

9 Conclusions and Future Scope

In the project, we have proposed the vector-based formula embedding approach which transform the formula into a vector of 202 bits and optimizes them using clustering. Primarily we have performed the preprocessing of scientific documents and mathematical formulas containing MathTagArticles of Wikipedia of the NTCIR-12 Corpus dataset. Moreover, by deploying the BPIT we have generated the formula vector of size 202 bit where 1 represents the presence of entity and 0 represents the absence of mathematical entity in the formula. Afterward, to optimize the search process, we designed the formula clustering technique that clustered the vector into 13 clusters. Afterward, the performance of the proposed approach was analyzed with 9 formula-based queries provided by the organizer of NTCIR-12, and the results in terms of precision were shown at 5 and 10.

Overall, clustering has proven to be a transformative optimization strategy, enabling the MIR system to meet the demands of modern applications while delivering accurate and timely results. It has effectively reduced time complexity, resource usage, and query latency, providing a robust foundation for future enhancements.

References

[1] Aizawa, A., Kohlhase, M., Ounis, I.: Ntcir-10 math pilot task overview. In: Proceedings of the 10th NTCIR Conference, June 18-21, 2013, Tokyo, Japan, pp. 654–661 (2013)

[2] Blostein, D., Lank, E., Zanibbi, R.: Treatment of diagrams in document image analysis. In: International Conference on Theory and Application of Diagrams, pp. 330–344. Springer (2000)

[3] Caprotti, O., Carlisle, D.: Openmath and mathml: semantic markup for mathematics.

XRDS: Crossroads, The ACM Magazine for Students **6**(2), 11–14 (1999)

[4] Carlisle, D.: Openmath, mathml, and xsl. ACM SIGSAM Bulletin **34**(2), 6–11 (2000)

[5] Cervone, D.: Mathjax: a platform for mathematics on the web. Notices of the AMS **59**(2), 312–316 (2012)

[6] Dadure, P., Pakray, P., Bandyopadhyay, S.: Embedding and generalization of formula with context in the retrieval of mathematical information. Journal of King Saud UniversityComputer and Information Sciences **34**(9), 6624–6634 (2022)

[7] Dadure, P., Pakray, P., Bandyopadhyay, S.: Mathuse: Mathematical information retrieval system using universal sentence encoder model. Journal of Information Science **50**(1), 66– 84 (2024)

[8] Davila, K., Zanibbi, R., Kane, A., Tompa, F.W.: Tangent-3 at the ntcir-12 mathir task. In: 12th NTCIR Conference on Evaluation of Information Access Technologies (2016)

[9] Dhar, S., Roy, S.: Mathematical document retrieval system based on signature hashing. Aptikom Journal on Computer Science and Information Technologies, **vol. 4**(1), pp. 45–56 (2019)

[10] Feng, Z., Tian, X.: A scientific document retrieval and reordering method by incorporating hfs and lsd. Applied Sciences **13**(20), 11207 (2023)

[11] Gao, L., Yuan, K., Wang, Y., Jiang, Z., Tang, Z.: The math retrieval system of icst for ntcir-12 mathir task. In: 12th NTCIR Conference on Evaluation of Information Access Technologies (2016)

[12] Kowalski, G.J.: Information retrieval systems: theory and implementation, vol. 1. Springer (2007)

[13] Kristianto, G.Y., Nghiem, M.Q., Matsubayashi, Y., Aizawa, A.: Extracting definitions of mathematical expressions in scientific papers. In: Proc. of the 26th Annual Conference of JSAI, pp. 1–7 (2012)

[14] Kristianto, G.Y., Topić, G., Aizawa, A.: Exploiting textual descriptions and dependency graph for searching mathematical expressions in scientific papers. In: Ninth International Conference on Digital Information Management (ICDIM 2014), pp. 110–117. IEEE (2014)

[15] Kristianto, G.Y., Topic, G., Aizawa, A.: Mcat math retrieval system for ntcir-12 mathir task.

In: 12th NTCIR Conference on Evaluation of Information Access Technologies, pp. 120— 126 (2016)

[16] Krstovski, K., Blei, D.M.: Equation embeddings. arXiv preprint arXiv:1803.09123 (2018)

[17] Lai, S., Liu, K., He, S., Zhao, J.: How to generate a good word embedding. IEEE Intelligent

Systems **31**(6), 5–14 (2016)

[18] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360 (2016)

[19] Lin, X., Gao, L., Hu, X., Tang, Z., Xiao, Y., Liu, X.: A mathematics retrieval system for formulae in layout presentations. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, pp. 697–706 (2014)

[20] Liska, M., Sojka, P., Ruzicka, M.: Similarity search for mathematics: Masaryk university team at the ntcir-10 math task. In: Proceedings of the 10th NTCIR Conference on Evaluation of Information Access Technologies, June 18-21, Tokyo, Japan, pp. 686–691 (2013)

[21] Liska, M., Sojka, P., Ruzicka, M.: Combining text and formula queries in math information retrieval: Evaluation of query results merging strategies. In: Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems, October 23, ACM, Melbourne, Australia, pp. 7–9 (2015)

[22] Mansouri, B., Oard, D.W., Zanibbi, R.: Dprl systems in the clef 2020 arqmath lab. In: Working Notes of CLEF 2020-Conference and Labs of the Evaluation Forum (2020)

[23] Mansouri, B., Zanibbi, R., Oard, D.W.: Learning to rank for mathematical formula retrieval. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 952–961 (2021)

[24] Marriott, K., Meyer, B., Wittenburg, K.B.: A survey of visual language specification and recognition. In: Visual language theory, pp. 5–85. Springer (1998)

[25] Matthews, D.: Craft beautiful equations in word with latex. Nature **570**(7760), 263–265 (2019)

[26] Miller, B.R., Youssef, A.: Technical aspects of the digital library of mathematical functions. Annals of Mathematics and Artificial Intelligence, **vol. 38**(3), pp. 121–136 (2003)

[27] Novotnỳ, V., Sojka, P., Štefánik, M., Lupták, D.: Three is better than one: Ensembling math information retrieval systems. Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum Thessaloniki, Greece (2020)

[28] Onal, K.D., Zhang, Y., Altingovde, I.S., Rahman, M.M., Karagoz, P., Braylan, A., Dang, B., Chang, H.L., Kim, H., McNamara, Q., et al.: Neural information retrieval: At the end of the early years. Information Retrieval Journal **21**(2), 111–182 (2018)

[29] Pathak, A., Pakray, P., Das, R.: Lstm neural network based math information retrieval. In: 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), pp. 1–6. IEEE (2019)

[30] Pathak, A., Pakray, P., Gelbukh, A.: A formula embedding approach to math information retrieval. Computación y Sistemas, **vol. 22**(3), pp. 819–833 (2018)

[31] Pathak, A., Pakray, P., Gelbukh, A.: Binary vector transformation of math formula for mathematical information retrieval. Journal of Intelligent & Fuzzy Systems, **vol. 36**(5), pp. 4685–4695 (2019)

[32] Peng, S., Yuan, K., Gao, L., Tang, Z.: Mathbert: A pre-trained model for mathematical formula understanding. arXiv preprint arXiv:2105.00377 (2021)

[33] Rei, M., Crichton, G.K., Pyysalo, S.: Attending to characters in neural sequence labeling models. arXiv preprint arXiv:1611.04361 (2016)

[34] Rudolph, M.R., Ruiz, F.J., Mandt, S., Blei, D.M.: Exponential family embeddings. arXiv preprint arXiv:1608.00778 (2016)

[35] Ruzicka, M., Sojka, P., Liska, M.: 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 (2016)

[36] Satpute, A., Gießing, N., Greiner-Petter, A., Schubotz, M., Teschke, O., Aizawa, A., Gipp, B.: Can llms master math? investigating large language models on math stack exchange. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2316–2320 (2024)

[37] Schubotz, M., Meuschke, N., Leich, M., Gipp, B.: Exploring the one-brain barrier: A manual contribution to the ntcir-12 mathir task. In: 12th NTCIR Conference on Evaluation of Information Access Technologies, pp. 309–317 (2016)

[38] Schubotz, M., Youssef, A., Markl, V., Cohl, H.S.: 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 (2015)

[39] Sojka, P., Liska, M.: The art of mathematics retrieval. In: Proceedings of the 11th ACM symposium on Document engineering, pp. 57–60 (2011)

[40] Sojka, P., Růžička, M., Novotnỳ, V.: Mias: math-aware retrieval in digital mathematical libraries. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1923–1926 (2018)

[41] Stathopoulos, Y., Baker, S., Rei, M., Teufel, S.: Variable typing: Assigning meaning to variables in mathematical text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 303–312 (2018)

[42] Thanda, A., Agarwal, A., Singla, K., Prakash, A., Gupta, A.: A document retrieval system for math queries. In: In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, June 7-10, Tokyo, Japan, pp. 346–353 (2016)

[43] Veljan, D.: The 2500-year-old pythagorean theorem. Mathematics Magazine, **vol. 73**(4), pp. 259–272 (2000)

[44] Xu, X., Tian, X., Yang, F.: A retrieval and ranking method of mathematical documents based on ca-yolov5 and hfs. Mathematical Biosciences and Engineering: MBE **19**(5), 4976–4990 (2022)

[45] Zanibbi, R., Aizawa, A., Kohlhase, M., Ounis, I., Topic, G., Davila, K.: Ntcir-12 mathir task overview. In: In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, June 7-10, Tokyo, Japan, pp. 299–308 (2016)

[46] Zanibbi, R., Oard, D.W., Agarwal, A., Mansouri, B.: Overview of arqmath 2020: Clef lab on answer retrieval for questions on math. In: International Conference of the Cross-Language
Evaluation Forum for European Languages, pp. 169–193. Springer (2020)