Mathematical Formula Retrieval

Megha Chakraborty
Oshita Saxena
Yash Jain
Yash Raj

December 2021

Commonly used search engines are quite inadequate in providing formula retrieval function by traditional retrieval techniques used in full-text information retrieval system. The main reason is that it is difficult to extract the keywords of the mathematical formulas due to their structural complexities. Mathematical formula retrieval systems exist to specifically cater to math formula information needs. Existing formula retrieval models may be categorized as text-based, tree-based or embedding-based.

Text-based models use string representations for formula indexing and retrieval. For example, MIaS, a Math-aware full-text search engine based on Apache Lucene linearizes MathML representations of formula appearance, and then incorporates text tokens generated from formulas in a TF-IDF retrieval model. It takes queries in the form of LaTeX code.

Tree-based models use hierarchical representation(s) of formula appearance and operation syntax, with similarity computed using paths and/or subtrees, or entire trees (e.g., using tree edit distances). For example in Tangent-3 Formula Search Engine, the formulas are first converted into trees. The trees are then transformed into sets of tuples. The candidate formulae are ranked by tuple matching to obtain top-k candidate formulae. The trees for these top-k formulae are then compared with the query tree using maximum subtree similarity. The candidate tree with the highest maximum subtree similarity is then retrieved.

In embedding models, formulae are represented as trees and then converted to vectors, with retrieval performed by identifying neighboring formulas from their vector representations.The Tangent-CFT model uses an n-gram embedding model to generate vectors from OPT and SLT tree representations. Representing math formulas as embedding vectors produces current state-of-art results for partial matching on formula retrieval tasks.

The effectiveness of retrieval models and formula representations can be studied to identify their relative strengths and weaknesses. Applying a learning to rank model on different models and representations enables users to implement the best model for the intended task.