Chapter 5
Mathematical Information Retrieval Trends and Techniques

Pankaj Dadure
National Institute of Technology, Silchar, India

Partha Pakray
National Institute of Technology, Silchar, India

Sivaji Bandyopadhyay
National Institute of Technology, Silchar, India

ABSTRACT

Mathematical formulas are widely used to express ideas and fundamental principles of science, technology, engineering, and mathematics. The rapidly growing research in science and engineering leads to a generation of a huge number of scientific documents which contain both textual as well as mathematical terms. In a scientific document, the sense of mathematical formulae is conveyed through the context and the symbolic structure which follows the strong domain specific conventions. In contrast to textual information, developed mathematical information retrieval systems have demonstrated the unique and elite indexing and matching approaches which are beneficial to the retrieval of formulae and scientific term. This chapter discusses the recent advancement in formula-based search engines, various formula representation styles and indexing techniques, benefits of formula-based search engines in various future applications like plagiarism detection, math recommendation system, etc.

INTRODUCTION

Mathematics is a significant factor in the field of science, technology, engineering, and mathematics (STEM) (Greiner-Petter A. A., 2020)**.** Without a single mathematical expression or symbol, a scientific text is often available. In this digital world, with growing numbers of teaching and learning materials

DOI: 10.4018/978-1-7998-7728-8.ch005

being produced, the explosion of knowledge was indeed inevitable. In the last decade, new techniques, concepts, and tools were created to store, maintain and retrieve this vast array of scientific records. In order to ensure, the users can easily access the information according to their information needs, the information needs to be organized and represented in the most efficient way.

Information retrieval (IR) is a subfield of natural language processing (NLP) that aims to retrieve the needed information from the collection of documents. The general IR system takes the user's query as an input, works on the similarity, and based on that returns the rank of relevant documents (Buttcher, 2016)**.** This is a common methodology used by today's retrieval system like Google search, PubMed, or Apple's Spotlight system. Nowadays, most of the available data on the web is sequential text data. Besides, the demand of the users may change: sometimes users may search for image/video data based on the text data, text based on the image data, based on the cause user's looking for effect related documents, some users interested in the linguistically structured documents. In some cases, users are unsure about what exactly they are looking for. To achieve these, several preprocessing operations have been investigated depend on the domain and the user's requirements (Virmani, 2019)**.** Almost all the retrieval systems are specially written programs, as long as researchers can explain their methodology, can be done for particular types of data. Moreover, the domain of information retrieval is explored since the early 1950s, and as a result, many IR models come into the limelight which mainly lies on the boolean model, vector space model, and probabilistic model. The field of textual information retrieval has been extensively investigated for many years, but mathematical information retrieval (MIR) (Hu, 2013) requires distinctive attention since traditional text recuperation systems cannot retrieve mathematical expressions. The mir systems are formula-based search engines that assist to search for knowledge in mathematical documents. The prime aspect of these MIR systems is to retrieve mathematical formulae which are relevant to a queried formula. In this task, the term 'relevant' encompasses two meaning: first considered the structural similarities to query formula, and the second one considered the conceptual meaning of the formula. Each finds not only formulas that are the exact match of the query formula, but also those which share similarities with it. For example, a retrieved result might contain only part of the query equation or might append terms.

Formulas found in web pages are mainly encoded in latex and/or MathML format. The traditional text-based search engines ignore the structure of these encodings by treating formula as normal text. This creates obstacles for a search engine to retrieve the relevant documents due to bounded structural information about the formula in the search index. In terms of query generation, this is a challenging task for an unfamiliar user with latex or MathML. Also, a recent study confirms that presenting raw math encodings in search results can adversely affect the accuracy of relevance assessment for search hits.

Information in mathematics is conveyed through descriptions, scientific terms, mathematical structures, and symbols which can be predefined in a mathematical expression. In this sense, technical terminology has also a significant role to play in mathematics. Moreover, the use of mathematical notation is dialectic. For example, different communities use different conventions for naming variables and defining operators. Individual authors redefine and adapt notation for their immediate needs. This flexibility is beneficial for authors and readers but makes automatic interpretation very difficult. This hypothesis stemmed from the observation that the mathematical discourse is dense with named mathematical objects, structures, properties, and results. An automatic understanding of equations significantly beneficial for analyzing scientific literature. In addition to this, the useful representations of equations can help to draw connections between articles, improve retrieval of scientific texts, and help to create tools for exploring and navigating scientific literature. Furthermore, the successful searching and retrieval of mathematical