Retrieval and Disambiguation of Mathematical Expressions for Mathematical Information Access

論 文 の 内 容 の 要 旨

論文題目　Retrieval and Disambiguation of Mathematical Expressions for Mathematical Information Access
(数学情報アクセスのための数式表現の検索と曖昧性解消)

氏　名　クリスティアント　ギオヴァニ　ヨコ
(Giovanni Yoko Kristianto)

Mathematical expressions are important for communication of scientific information, for instance, to give formal definitions of concepts written in natural language. In this dissertation, we propose a mathematical information access (MIA) system which can help people access and understand math expressions in scientific documents. MIA has been studied in the digital mathematics library and information retrieval communities for searching for math expressions based on their token elements (e.g. identifiers, numbers, and operators) and structures (e.g. fractions, scripting, and matrices). The major focus of the existing work on MIA has been the development of algorithms to store the tree representation of math expressions into a database. On the other hand, the descriptive text of the math expression (hereinafter, we call it textual information) has not yet fully exploited for MIA. The use of textual information has potential to improve the retrieval performance of an MIA system and helps the MIA user understand the definitions of math expressions.

This dissertation presents a framework of MIA that supports math search and math understanding of the users by utilizing math structure and text similarities. We introduce three core ideas which are essential for realizing MIA.

The first core idea in this dissertation is the development of a math information retrieval (MIR) system that exploits the structures and the textual information of math expressions to allow effective search for mathematical knowledge. Following the convention of the current digital math library research community,

we assume that a query is given as the combination of math expressions and textual keywords. Our proposed math search system takes advantage of multiple types of textual information to enable high-recall and high-precision retrieval. An evaluation in NTICR-12 MathIR, a mathematics information retrieval shared task, shows that our search system achieved the best performance over other existing math search systems in retrieving highly relevant paragraphs containing the users' requested math expressions.

The second core idea in this dissertation is the enrichment of the textual information of a math expression considering the relationships with other math expressions within the same document. The motivation behind is that textual descriptions of the component subexpressions or identifiers are useful to explain a complex math expression, yet these descriptions may not be captured within the context of the target expression. Therefore, to enrich the textual information of each math expression, while keeping its capability to enable high precision search, we utilize the dependency relationships (e.g. formulae-variables relationships) between math expressions. An evaluation shows that this approach has a significant impact in improving search precision.

In addition to the development of math search system, this dissertation formulates a task of determining the identity of math expressions in documents by linking these expressions to their corresponding entities in knowledge base, such as Wikipedia. This task is denoted as math entity linking (MEL). We propose a supervised learning based approach using math related features, such as math and text similarities, as well as the location and importance of the math expression within the document. Our evaluation shows that the proposed approach can determine correct links for math expressions in a higher precision than a straightforward application of an MIR system.

To conclude, this dissertation proposes the use of math structures and multiple types of textual information incorporated with the dependency relationships between math expressions to capture the semantics inherent in math expressions. The proposed math search system shows the highest search performance among other four state-of-the-art search systems. In addition, we propose a MEL module reliable enough to link math expressions to their best non-null corresponding entities in knowledge base. Since math expressions are essential part of

scientific information, our proposed approach has an important implication for the applications of information access in a wide range of scientific fields. Finally, this dissertation is a step towards enabling effective formula search and formula browsing in digital library practices.