文書ファイルからの数式抽出と式変形の試み

Extraction of Mathematical Equations from Textbook and Equation Conversion

物材機構[1]　○吉武 道子[1]

NIMS[1], ○Michiko Yoshitake[1]

E-mail: yoshitake.michiko@nims.go.jp

マテリアルインフォマティクスの領域では、材料の密度や屈折率などの数値データを機械学習する例がほとんどを占めるが、ビジネスの世界ではチャットボットや PRA のように言語（自然言語）を対象としたインフォマティクスも盛んである。マテリアルインフォマティクスの分野で自然言語を扱う例としては、特許文献の分類・抽出や、科学技術文献などから材料名と数値データをペアにして抽出して、数値データの機械学習用の数値データベースを作成する試みがある。一方、マテリアルインフォマティクスのような自然科学の分野では、教科書などに数式（テキストデータの一種、数値データと密接に関係する）が掲載されている場合が多い。数式中に現れる変数と数式をペアにしてコンピュータで処理できれば、個々の数値データが無くても物性値の間の関係性がわかり、実験値や計算値が無くても別の物性値の実験値・計算値から数式によって値を計算することも可能になると期待される。

発表者は、教科書などのような PDF データから XHML 変換して得たデータ、またはウェブから直接得られるデータ、から presentation MathML 部分を抜き出し、その部分を数式としての意味を付与された content MathML へ変換し、content MathML から数式として扱える形式への変換、数式の変形、変数とそれが表す物性量の名称の紐づけ、などを行うことで、物性間の関係性を定量的にネットワーク構造に作り込み、目的の物性変数を、直接つながっていない変数で表現することが可能なシステムの構築を目指している。今回、式変形までを試みた結果を報告する。

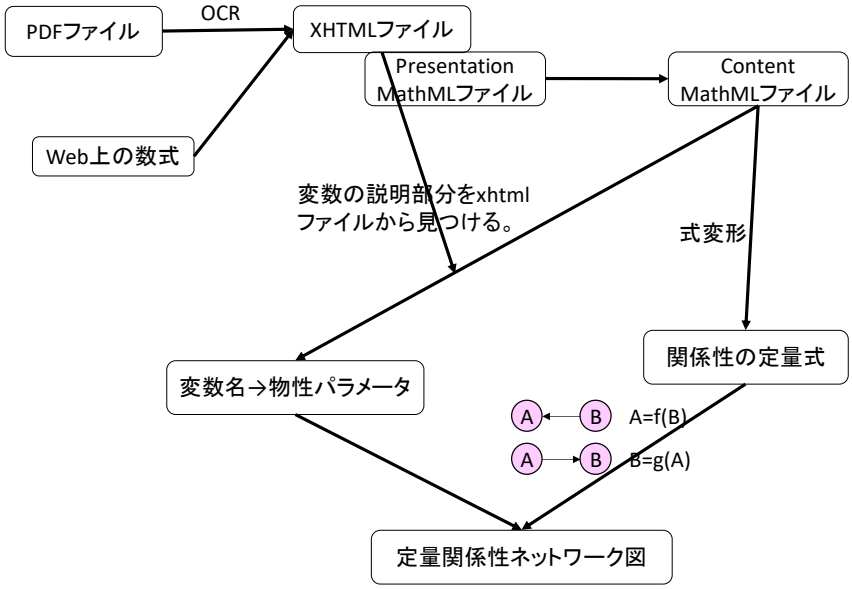

図　本報告で試みているシステムの概要

© 2020年 応用物理学会