数式概念検索のための情報抽出手法に関する検討

横井 啓介[†1]　Minh NGHIEM[†2]　相澤 彰子[†3]
東京大学情報理工学系研究科コンピュータ科学専攻[†1†3]
総合研究大学院大学情報学専攻[†2†3]
国立情報学研究所[†3]

1　はじめに

本稿では，電子媒体としての科学論文，さらにその中にある数式に焦点をあて，類似式検索手法とそれを強化するための変数や式の定義の抽出手法について述べ，その有効性について検証する．

現在，研究者は論文をどのように探しているだろうか．多くの研究者は，論文の参考文献を調べて広げていく，あるいは既存研究や特定分野において調べたいキーワードをクエリとして入力することで論文を検索している．しかし，そのキーワードは現時点では基本的には自然言語にしか対応していない．論文には自然言語の他にも図・表・数式など，多くの情報を持っており，これらの情報を何らかの方法で検索に役立てることができれば，より利便性の高い検索が期待できる．

本稿の構成は以下の通りである．まず 2 章では，既存の数式検索手法について触れる．そして 3 章でその問題点・改善点を考慮した情報抽出手法を提案する．その有効性を調べる実験を 4 章で説明，5 章で結果を述べ，最後にまとめを述べる．

2　既存手法

数式検索技術は，自然言語以外の情報を検索に役立てる試みの一つとして，様々な研究が行われている．中でも数式を Web 上に表記するための標準でもある MathML(Mathematical Markup Language) [1] を利用した研究は多い．MathML には，Web 上に数式を視覚的に表現することを重視した Presentation Markup と，数式の構造を表現することを重視した表記である Content Markup の 2 種類の記法が存在する．橋本らは MathML の Presentation Markup 構造の Xpath を用いて類似式の検索を行っている [2]．それに対し，我々は数式の構造を重視して Content Markup 表記による木構造の類似度から類似式の算出を行う研究を行ってきた [3].

これらの技術は，確かに，見かけ上類似した式を取り出すことができている．しかし，数式単独では変数や関数を単なる記号としか捉えることができず，そのため理解に充分な情報を得ることはできない．たとえば F = mg と y = ax という二つの式は，数式単独で見た場合はどちらも同じ構造を持つため，従来手法では高い類似度を持つ．しかし一般的には，前者は重力 F は質量 m と重力加速度 g の積で与えられることを表し，後者は一次関数の式を表している．このように，式の構造が同じであってもそれぞれの変数や関数の用途が異なる場合に，それを同じものとみなすことは好ましくない．逆に違う変数名であっても同じ意味として使われている場合もあり，その対応を考慮する必要がある．

3　提案手法

一方、情報獲得分野において自然言語から情報を抽出する研究も多く存在する．小林ら [4] は Web 上の文書から共起パターンを用いて評価情報を抽出している．杉木ら [5] は同様にパターン変換を用いて口コミサイトから意見情報を収集している．科学論文中において，数式中の変数や関数の定義は，明確かつ典型的な表現で記述される場合がほとんどであり，パターンを用いた情報抽出手法は我々の目的に対しても有効であると考えられる．

そこで我々は，数式とその周辺テキストを同時に考えることで，数式中の変数の意味，式の定義などそれぞれの数式の理解や利用に必要な情報を文書中から抽出し，数式が表す深い意味構造を踏まえた検索の実現を「数式概念検索」として，単なる「数式検索」と区別し，よりユーザーの思考に深くマッチした類似式の検索の実現を目指した．検索全体としての処理の流れを図 1 提案手法を含めた検索の流れに示す．

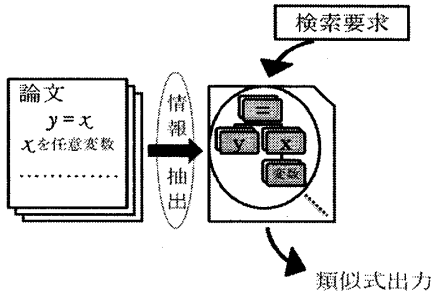

図 1 提案手法を含めた検索の流れ

4 実験方法

パターンを用いた定義抽出の有効性を測るため，情報処理学会関係の論文のうち，比較的数式が多い論文 58 本に対し，数式付き文書認識 OCR である InftyReader [6]を利用して数式に MathML 情報を付与したものから利用した．まず，パターンを導出するための学習用サンプル論文を 5 本，全体から無作為に選び，変数や関数の説明文，定義文を抜き出し，形態素解析を行って定義パターンを抜き出した．そして新たに評価用論文をサンプル論文とは別に 5 本無作為に選び，各文に対して形態素解析を行い，先のパターンから定義を抽出し，人手で評価を行った．それぞれの形態素解析には形態素解析エンジン Mecab [7]を用いた．作業の流れを図 2 作業プロセスにまとめる．

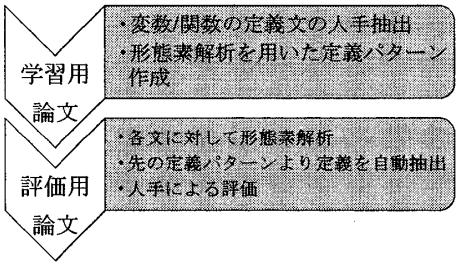

図 2 作業プロセス

5 実験結果

今回の実験で 5 本の学習サンプルから抽出した定義パターンを以下に示す．

表 1 抽出パターン一覧

	パターン
1	[N] + [Exp] (+ “, ” + [Exp] + ...)
2	[N] + “を/は” + [Exp] + “, ”
3	[Exp] + “を/は” + [N] + “, ”
4	[N] + “を” (+...) + [Exp] + “と” + <する/表す>
5	[Exp] + “を” (+ ...) + [N] + “と” + <する>
6	[Exp] + “は” (+ ...) + [N] + “で” + <ある>
7	[Exp] + “は” (+ ...) + [N] + “を” + <示す>
8	[N] + “と呼び” (+ ...) + [Exp] + “で” + <表す>

表 1 において，[N]は名詞句，[Exp]は MathML 情報の付加された数式を表す．また<>で括った動詞は原型に直した場合の表記であることを示す．評価サンプルの定義抽出情報を人手で判定した結果は以下の通りである．

表 2 抽出結果

	R	N	F	Precision	Recall
1	19	2	5	90.5%	79.2%
2	43	4	6	91.5%	87.8%
3	67	4	6	94.4%	91.2%
4	9	4	0	69.2%	100.0%
5	72	14	23	83.8%	75.6%
計	210	28	40	88.2%	84.0%

表 2 において，R は正しい定義であると判断された数，N は抽出されるべき定義が抽出されなかった数，F は誤った定義を抽出した数を示す．Precision，Recall はそれぞれ R/(R + N)，R/(R + F) と定義する．

今回の実験は，少ない学習データの量にも関わらず，Precision, Recall ともに 80% を超える高い数字を得ることができた．以上から，数式における変数や関数の定義は形式的な表現が使われることが多く，パターンを用いた情報抽出が有効であることがわかる．

6 まとめ

本稿では，パターンを用いて類似式検索に有効な変数，関数，およびそれらの定義を抽出する手法を述べた．そして実験を通じて，変数や関数の定義文は著者によってあまり大きな差異は生じず，比較的少数のサンプル論文からでも有効な定義パターンの多くを発見することができるということがわかった．

今後の展望として，変数，関数定義の自動抽出が挙げられる．形式的な表現が多く使われるということは，機械学習等，統計的な手法も有効であることが期待される．また，現在はまだ定義を抽出しただけであるため，数式構造の検索への結びつけを行うことや，数式周辺テキストから定義を抽出するだけでなく，論文中における数式変形の記述を利用して，数式自体から情報を抽出するなど，数式概念検索のさらなる応用が考えられる．

引用文献

1. Mathematical Markup Language Version 2.0. http://www.w3.org/TR/MathML2/.
2. 橋本 英樹，土方 嘉徳，西田 正吾：MathML を対象とした数式検索のためのインデックスに関する調査．情報処理学会研究報告，2007-DBS-142，pp.55-59，2007.
3. Keisuke Yokoi and Akiko Aizawa: An Approach to Similarity Search for Mathematical Expressions using MathML, 2nd workshop Towards a Digital Mathematical Library (DML 2009), 2009.
4. 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一：テキストマイニングによる評価表現の収集．情報処理学会研究報告，NL-154，pp.77-84，2003.
5. 杉木健二，松原茂樹：消費者の意見に基づく商品検索．情報処理学会論文誌，Vol.49，No.7，pp.2598-2603，2008.
6. Masakazu Suzuki, Toshihiro Kanahori, Nobuyuki Ohtake and Katsuhito Yamaguchi: An Integrated OCR Software for Mathematical Documents and Its Output with Accessibility, ICCHP 2004, LNCS 3118, pp.648-655, 2004.
7. MeCab: Yet Another Part-of-Speech and Morphological Analyzer http://mecab.sourceforge.net/.