D-011

数式とその周辺情報を利用した数式概念検索の実現
The Search for Concept of Mathematical Equations by Appending Information around Them

横井 啓介† Minh Nghiem Quoc†† 松林 優一郎††† 相澤 彰子†,†††
Keisuke Yokoi† Minh Nghiem Quoc†† Yuichiroh Matsubayashi††† Akiko Aizawa†,†††

1. はじめに

科学的な情報伝達において、数式はしばしば非常に重要な役割を果たしている．単に計算を行うためだけではなく，科学的な知識を正確に相手に伝えるために数式は必要な媒体である．しかしながら現在のところ，情報検索や情報提示に関する研究では自然言語のみを扱うものが主流であり，数式を扱えるものはほとんど行われておらず，多くの場合数式に関する情報は無視されている．

そこで我々は，特に科学文書中の数式に焦点を当て，数式が持つ独特の構造情報と，その周辺にある数式に関する自然言語による説明記述の両方を考慮に入れた数式意味検索の実現を目的とする．これまでの数式に関する研究の多くは数式そのものの構造のみを扱っているが，数式のみでは書き手や分野，あるいは文書の形式によっても曖昧性が生じてしまう．我々は周辺記述も取り入れることで数式に深い意味情報を取り入れ，より柔軟かつ意味を深く考えた情報検索・情報提示を目指す．そのための第一歩として，今回我々は数式中に現れる変数や関数などに対し，その定義記述や説明記述を周辺文章から探し，それらを関係づける手法について提案する．

今回の研究には実験のためのデータセットとして，情報処理学会論文誌より発行されている論文のうち 100 本を選択，使用した．まず初めにそれらの論文に含まれる数式を，数式に対応した OCR を用いて Mathematical Markup Language (MathML) [1] へと変換し，OCR による細かい認識ミスを手作業で修正した．MathML は数式の構造情報を保持した表記法であり，Web 上で数式を表すための標準とも言えるものである．そうしてできたデータから，我々は MathML で書かれた数式とそれに対応した定義記述や説明記述を抽出し，整形してコーパスを作成した．これを用いることで，機械学習において，数式と対応した定義・説明記述を自動的に抽出し，またそれを数値的に評価することが可能になった．

本稿の構成は以下の通りである．まず 2 章で数式を対象とした関連研究を紹介する．3 章で本研究に用いたデータセットの作成手順について述べ，実際にそれを用いて行う数式対応情報の抽出手法について 4 章で提案する．5 章では実際に実験を行い，それに対しての結果，考察を共に述べる．最後に 6 章で我々の貢献や結論，および今後の展望についてまとめる．

2. 関連研究

我々の研究において，数式は MathML を用いて表されている．MathML は World Wide Web Consortium (W3C) によって定められた Web 上に数式を表記するための標準とも言える記法である．MathML に対応したソフトウェアも増えてきており，数式を含む文章の編集，TeX や OpenMath など他の数式の表記法との相互変換，そして紙面に印刷された数式を含む文書が認識可能である数式 OCR など様々である．

これまでに MathML やその他の数式記述言語を用いて行われてきた試みはいくつかあるが，それらは大きく二つに大別できる．一つは数式検索を目的としたもの，もう一つは数学的な知識獲得と辞書構築を目的としたものである．以下にそれぞれについて述べる．

数式検索を目的とした研究では，電子媒体および Web 上の数式を含む科学的文書の情報を検索に利用している．それらの数式は表現が多様でかつ曖昧性を含むため，多くの研究では多様性を許す表記を用いたり，数式の構造から類似度を計算したりするなどの手法をとっている．Adeel らは MathML 式から正規表現を用いて数式の特徴をキーワードとして抽出し，それを従来の自然言語を用いた検索システムのクエリとして用いている [2]．Mišutka は数式をより一般的な表現に変換することで，多様な表現に対応した拡張全文検索を実現している [3]．橋本らは MathML 式の XPath を利用し，インデックス付けを行うことで高速に類似式の検索することを可能にしている [4]．そして我々もまた，MathML 式の Content Markup を利用し，数式を木構造として捉えてその類似度の算出式を独自に定義することで，関数や変数に柔軟性を許した類似数式検索手法を提案した [5]．

一方で，数学的な知識獲得，辞書構築を目的とした研究では，数式を含む文書から数学的な知識を自動的に獲得し，知識ベースの辞書を構築することを目的としている．それゆえこれらの研究では数式周辺の自然言語を用いた解析に重点を置いている．Kohlhase らは数式に関する符号，定義，そして証明などの記述を抽出し，それらの関係をデータベースとして蓄積させて知識として利用する手法を提案した [6]．Jaschke らは数式を含む文書から自動的に数学的なオントロジーを抽出するための枠組みを述べている [7]．これは LaTeX で書かれた数式を MathML に変換，構文解析等の自然言語処理による数学概念の関係の抽出，抽出された情報をまとめてグラフ化，の 3 ステップからなっている．この枠組みは一般性が高く多くのシステムに適用可能であると考えられるが、数式の構造解析についてはまだ手がつけられていない．

このように，数式検索についての研究は主に数式自身の構造的な部分に重点を置いており，一方で知識獲得に関する研究は数式周辺の情報を取り込んだ意味的な部分に重点を置いている．我々は数式検索の改善を目標に挙げているが，数式に関してより深い解析を可能にするためにも，数

†東京大学情報理工学系研究科コンピュータ科学専攻 Department of Computer Science, The University of Tokyo
††総合研究大学院大学複合科学研究科情報学専攻 Department of Informatics, The Graduate University for Advanced Studies
†††国立情報学研究所コンテンツ科学研究系 Digital Content and Media Sciences Research Division, National Institute of Informatics

式構造・意味両面を考慮した手法を目指している．その中でも今回は数式記述とその文書中の自然言語記述の対応関係に着目し，それを効果的に抽出する手法を後の章で述べていく．

3. データセット

我々の提案する新しい数式検索の枠組みを実現するために，数式とその周辺テキスト中の説明記述の対応付けを自動的に抽出することは不可欠であるが，それを検証する適切なデータセットを見つけられなかったため，我々は実際にデータセットを作成することから始めた．この章では提案手法を検証するためのデータセットの作成手順について述べる．手順の概要を以下の図 1 に示す．

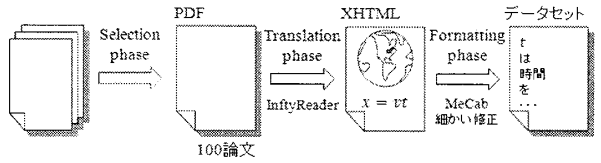

図 1　データセット作成手順

まず初めに，今回の研究対象とする科学論文を集めた．今回は少ないデータで密な学習結果を得るために関連の深い論文を集めることを目的として，「機械学習」をテーマに設定し，関連するキーワードを含む 214 論文を候補とした．用いたキーワードを表 1 に示す．その候補論文のうち，数式が明らかに少ない 52 論文を候補から除き，さらにそれぞれの論文から，内容が類似している，参照/被参照関係にあるなど，関連の深い 104 論文を残した．さらに数式とその説明や定義に関する記述の関係が確認できないものを除き，最終的に 100 論文を選び出した．

表 1　論文選択に用いたキーワード

キーワード
機械学習
教師あり学習
教師なし学習
サポートベクターマシン
SVM
ニューラルネットワーク

次に，これらの 100 論文の PDF 形式のデータを，InftyReader を用いて MathML 式を含む XHTML 形式に変換した [8]．InftyReader は数式に対応した OCR ソフトウェアであり，自然言語の認識に加えて，数式も解析し MathML 形式で出力することができる．また，XHTML は XML の一種である MathML を確認するために必要な形式である．

こうして得られた XHTML 形式のデータに対して，人手で一つずつチェックし，数式とそれに対応する説明・定義記述のペアを抜き出して記録する．ここで数式とは，InftyReader によって MathML 記述に変換されたもの全てのことを指す．すなわち文中に存在する単なる一変数や関数なども数式に含まれる．またそれと同時に，OCR の誤認識の訂正を人手により行った．その後，各文に対し形態素解析器を用いて形態素に分解した．今回の実験では形態素解析器として MeCab を用いた．なお，形態素解析のために数式はそれぞれ "Exp" という文字列に変換を行い，また簡単のため，名詞 (数詞も含む，ただし"Exp"は除く) の連なりは複合名詞であると仮定し，それらはまとめて一つの形態素とみなした．最後に，先に抽出した数式とその説明記述のペア情報および形態素情報を利用し，もし数式が対応した説明記述を持つならばその記述の形態素 ID を数式に正解情報として付与し，最終的なデータセットを構築した．

データセットのフォーマットを，例文を通して表 2 に示す．今回の例文として「ここで，分布 f はパラメータ ϕ の事前確率分布を示す」という文を用いた．この例には 2 つの数式が含まれているが，区別するため形式的に Exp1, Exp2 と分けて記述した．この文において，Exp1 は対応する説明・定義記述が 2 つある(「分布」，「事前確率分布」)ため，2 つの対応する形態素 ID を持っている．また形態素情報は機械学習の素性として様々な要素の可能性を考えるため，MeCab の解析による情報を全て載せている．

表 2　データセットの形式例

ID	形態素	形態素情報	対応 ID
0	ここ	名詞,代名詞,一般,...	
1	で	助詞,格助詞,一般,...	
2	,	名詞,サ変接続,...	
3	分布	名詞,サ変接続,...	
4	Exp1	名詞,一般,...	3,9
5	は	助詞,係助詞,...	
6	パラメータ	名詞,一般,...	
7	Exp2	名詞,一般,...	6
8	の	助詞,連体化,...	
9	事前確率分布	名詞,サ変接続,...	
10	を	助詞,格助詞,一般,...	
11	示す	動詞,自立,..,五段・サ行, 基本形, 示す,...	

4. 数式対応情報抽出手法

この章では，先の章で述べたデータセットを用いて，自動的に MathML 式に対応する説明・定義記述を抽出するし 2 つの手法について述べる．まずデータセットを使って行う問題と評価法について明確に定義し，その後，パターンマッチングを主体とした手法と機械学習を主体とした手法についてそれぞれ説明する．

4.1 問題定義

本研究での目的は，MathML 式が与えられたときに，その数式の意味，名前，説明，あるいは定義を表している箇所 (以下，まとめて説明記述とする) を発見する，というものである．今回の実験では，問題を単純化するために以下のいくつかの制限を設ける．

- 数式の説明記述は，すべて名詞句/節とする．
  今回はその句/節の中で，最後の形態素を抽出することを目的としている．
- 数式の説明記述は，すべてその対応する数式と同一文中に存在する．
  実際には文をまたぐ説明記述も考えられるが，今回はその例は考えない．また，数式が改行されて文を区切っている場合は，区切られたそれぞれを一つの文とみなして考える．

以上の制限から，問題は各 MathML 式に対して，それぞれその説明として最も適する説明記述の名詞句/節を選択する (説明記述が存在しない数式もある) という問題に帰着する．具体的には，既に形態素に分けられているデータセットを用いて，各数式と各名詞形態素の組それぞれに対し，その名詞が数式の説明記述となっているかを決定する二値問題で表される．3 章で用いた例 (表 2) についてもう一度考えると，この文には 2 つの数式 (Exp1 と Exp2) と 4 つの名詞 (ここ，分布，パラメータ，事前確率分布) が存在するため，全部で 8 つの二値分類問題を考えることになる ([Exp1, ここ], [Exp1, 分布], [Exp1, パラメータ], [Exp1, 事前確率分布], [Exp2, ここ], [Exp2, 分布], [Exp2, パラメータ], [Exp2, 事前確率分布]) (図 2).

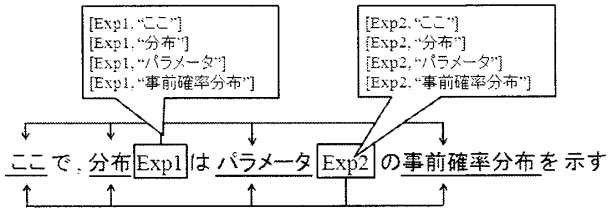

図 2　問題設定の例

今回の実験では，先に 100 論文から得たデータセットを訓練データ，検証データ，テストデータの 3 つに分類する．今回は訓練データ 60，検証データ 20，テストデータ 20 とした．学習を用いない手法に関しては，20 本のテストデータのみ用いることとなる．それぞれのデータセットの正例，非例の数を表 3 に示した．

表 3　データセットごとのデータ数

データセット名	論文数	正例数	非例数
訓練データ	60	3,857	50,289
検証データ	20	1,172	12,519
テストデータ	20	1,488	17,223

4.2 パターンマッチング手法

我々の先行研究により，数式の説明記述は著者により大きく変化するというものではなく，その記述のパターンに関しても大部分はわずかな数のパターンにあてはまるということがわかっている．その時の実験で得られた 8 パターンを用いて，今回の実験データに対しても適応できるかどうかを検討する．用いるパターンを表 4 に示す．これらは，情報処理学会学会誌の論文の中から比較的数式が多く含まれていると思われる論文を 5 本選び，その中の数式と説明記述が共に含まれるパターンを人手で抜き出したものである．

表において，[名詞] は対象となる説明記述の候補を表し，[Exp] は数式を表す．スラッシュ (/) で並んだ語はどちらかにマッチすれば良いことを示し，山括弧 (<>) は，動詞の原型が山括弧内の語であることを示している．また，「...」は空文字列を含む任意の文字列を指している．なお，パターン 1 は[名詞] の後に対象の [Exp] が続いているか，または間に他の [Exp] とカンマ (,) のみを含むということを表している．4.1 節の例(図 2) をもう一度考えると，Exp1 に関しては [Exp1, 分布] のペアがパターン 1 を，

表 4　抽出パターン

No.	パターン
1	[名詞] [Exp] (, [Exp] , [Exp] , ...)
2	[名詞] を ... [Exp] と <する/表す>
3	[Exp] を ... [名詞] と <する>
4	[Exp] は ... [名詞] で <ある>
5	[名詞] と 呼び ... [名詞] で <表す>
6	[名詞] を/は [Exp] ,
7	[Exp] を/は [名詞] ,
8	[Exp] は ... [名詞] を <示す>

[Exp1, 事前確率分布]がパターン 8 を満たし，Exp2 に関しては[Exp2, パラメータ] のペアがパターン 1 を満たしているため，上記の 3 つが数式と対応する説明記述であるとみなされ，残りの 5 つは非例であると判断される．

これらのパターンを用いて，各文中のすべての数式とすべての名詞に対しパターンにマッチするかを見てラベル付けを行った．

4.3 機械学習手法

我々は先に述べたパターン記述やその他データセットに含めた形態素情報などを手掛かりとして，教師あり学習を行う手法を提案する．4.1 節で定めた問題定義の通り，今回の実験で挑戦する問題は各名詞が各数式に対して対応した説明記述となっているかどうかの二値分類であり，これに対しサポートベクターマシン(SVM) の二値分類モデルを適用する．今回の分類実験に用いた素性を表 5 に列挙する．

表 5　用いた基本素性

大分類	基本素性
パターン	パターン (1～8) にマッチする
[名詞]と[Exp]の関係	距離(1,2,...,10, 11 以上, -1,-2,...-10, -11 以下)
	間に 数式/カンマ/開括弧/閉括弧 の有無
	(直前直後を除く)間に は/を/カンマ の有無
	間に カンマ/数式 以外の要素の有無
	[名詞]が[Exp]より先にある/後にある
[名詞]	自身の 名称/複合名詞か否か
	直前/直後 の単語の 名称/品詞
[Exp]	直前/直後 の単語の 名称/品詞
[名詞][Exp]	両方の後ろにある動詞の原型

また，これらのうち，有効と思われる組み合わせ (2～6 個)を人手で組み合わせたものも新たな素性として加えた．

4.1 節で述べた訓練データを用いて，L2 正則化付き L1 損失関数 (ヒンジ損失関数) を Primal Estimated sub-GrAdient SOlver (Pegasos) アルゴリズムにより最小化した．パラメータの推定等，実際の機械学習アルゴリズムとしては Classias [9]を使用した．

5. 実験結果，考察

この章では，3 章で述べたデータセットを用い，4 章で定めた問題と提案手法を用いて実験を行った結果を示す．またその結果に対する信憑性を含めて考察を行う．

- ベースライン

提案手法との比較のためにベースラインを定めた．ベースラインでは，説明記述候補がそれに対応する数式の

直前にある場合に限り正しいと判断する手法である．表 2 の例についてこの手法を用いて考えると，[Exp1, 分布]，[Exp2, パラメータ] の 2 つのペアは正しいと判断され，残りのペアは非例と判断されることになる．

5.1 実験結果

それぞれの手法を用いた場合の結果を表 6 に示す．ここで評価尺度として，正解ペアに関しての適合率 (P)，再現率 (R)，F 値 (F) を用いた．

表 6　実験結果

手法	適合率	再現率	F 値
ベースライン	0.4538	0.5058	0.4784
パターンマッチング	0.7241	0.6156	0.6655
機械学習	0.8193	0.6062	0.6968

結果より，パターンマッチング，機械学習共に比較的効果的に抽出できていると言える．また，機械学習手法はパターンマッチングより適合率の数値が大幅に高く，これに伴い F 値も高くなっている．さらに有効な素性を加えることができればより精度を高めることができると思われる．

5.2 データセットの信憑性に関する検証

今回用いたデータセットは我々自身が用意したものであり，かつその作成に関しては，数式とその対応する説明記述のペアの抽出，抽出されたペアと形態素解析結果の結び付けなど，手作業に依存する部分が多く存在する．それゆえ今回の実験結果がどこまで信用できる値であるかを検証する必要があると判断し，データセットの信憑性を確認するための追加実験を行った．まずテストデータの中からランダムに 5 本の論文を選び，それぞれの数式と名詞のペアに付与されたラベルが正しいかどうかを人手でカウントした．その結果を表 7 に示す．この実験から，データセットの信憑性はおおよそ 85%であることがわかり，それに伴って今回の実験の値で見込める上限も 85%前後であると思われる．

表 7　データセットの信憑性検証

サンプル	適合率	再現率
1	0.8846 (46 / 52)	0.9020 (46 / 51)
2	0.8673 (98 / 113)	0.8991 (98 / 109)
3	0.8980 (44 / 49)	0.7097 (44 / 62)
4	0.8000 (28 / 35)	0.7778 (28 / 36)
5	0.8514 (63 / 74)	0.7875 (63 / 80)
計	0.8638 (279 / 323)	0.8254 (279 / 338)
平均	0.8597	0.8152

6. 実験結果，考察

本論文では，科学論文中に含まれる数式に対し，その周辺テキストの中でそれらの数式に対応した名前，定義，説明などの情報を与える部分を抽出する手法を提案した．科学論文における数式の説明記述は，ある程度特徴的で形式が整っており，著者による違いはあまり多くないと考えられる．よってパターンマッチングを用いた方法が効果的に作用することは予想される結果であった．一方で，本稿での実験結果では，機械学習を用いて説明記述を抽出する手法もまた有効に働くことが示された．現在機械学習で用いている素性はまだ一部の情報しか用いておらず，有効な素性は他にも存在し得る．また他にも，各文に対して係り受け解析を行い，明らかに説明記述ではありえない関係を判断する．数学的概念に関する用語などを辞書として利用するなど精度向上のための関連手法は考えられる．柔軟性や応用性，手間の低さを考慮すると機械学習の適用はパターンマッチング以上に有望であると考えられる．

今後の展開として，まず挙げられるのはデータセットの改善である．データセットの検証を通して，手作業によるラベル付けの信頼性の問題が明らかになった．より信頼できる判断基準のために，今後は，様々な自動抽出手法と組み合わせてラベル付けの精度を高める必要がある．そのためにも，まずは現在のデータセットの誤りを具体的に確かめ，対策を練る必要がある．

二つ目は，問題設定の幅を広げることである．今回は簡単のために，説明記述が同一文中に存在し，かつ名詞の集合であるとする制限をかけ，かつ名詞のみの抽出を目的としていた．しかし現実的に抽出した情報を情報提示やより深い意味をもった数式検索に応用するにあたり，名詞だけでは非常に情報量が少ない．名詞句/節ごと取り出すべきであり，また制限も取り払うことが今後の発展のために重要だと思われる．データセットのフォーマットを再考察し，情報を付け加え，タスクを再設定して取り組めば良い．

三つ目に，言語の拡張である．本稿で提案した情報検索・情報提示のための枠組は言語に依存するものではない．数式は自然言語とは異なり，世界共通で通用する言語であるため，これらの手法を日本語だけでなく他の言語に対しても適用できるように取り組んでいく．

参考文献

[1] Mathematical Markup Language (MathML) Version 2.0 (Second Edition), World Wide Web Consortium. http://www.w3.org/TR/MathML2.

[2] Muhammad Adeel, Hui Siu Cheung, and Sikandar Hayat Khiyal: Math GO! Prototype of a Content Based Mathematical Formula Search Engine. Journal of Theoretical and Applied Information Technology, Vol 4, No 10, pp. 1002-1012, 2008.

[3] J. Mišutka: Indexing Mathematical Content Using Full Text Search Engine. WDS '08 Proceedings of Contributed Papers, Part I, 240-244, 2008.

[4] 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

[5] Keisuke Yokoi and Akiko Aizawa: An Approach to Similarity Search for Mathematical Expressions using MathML. DML 2009, 2nd workshop, Towards a Digital Mathematics Library, Ontario, Canada, 2009.

[6] Michael Kohlhase, Andreas Franke: MBase: Representing Knowledge and Context for the Integration of Mathematical Software Systems, Journal of Symbolic Computation; Special Issue on the Integration of Computer Algebra and Deduction Systems, pp. 365-402, 2001

[7] Jeschke, S., Wilke, M., Blanke, M., Natho, N. M., and Pfeiffer, O. F.: Information Extraction from Mathematical Texts by Means of Natural Language Processing Techniques. In Proceedings of the international Workshop on Educational Multimedia and Multimedia Education, Augsburg, Bavaria, Germany, 2007.

[8] Masakazu Suzuki, Fumikazu Tamari, Ryoji Fukuda, Seiichi Uchida, Toshihiro Kanahori: Infty: an integrated OCR system for mathematical documents, Proceedings of the 2003 ACM symposium on Document engineering, Grenoble, France, 2003.

[9] Naoaki Okazaki: Classias: a collection of machine-learning algorithms for classification. http://www.chokkan.org/software/classias, 2009.