数式検索のための数式周辺テキストの言語解析手法

横井 啓介[1]　Minh-Quoc NGHIEM[2]　松林 優一郎[3]　相澤 彰子[1,3]
東京大学[1]　総合研究大学院大学[2]　国立情報学研究所[3]
{kei-yoko, nqminh, y-matsu, aizawa}@nii.ac.jp

1. はじめに

数式は，単なる計算の記法ではない．科学的な議論の場において，明確かつ厳密に物事を記述する必要不可欠な媒体である．数式が検索できるということは，その数式が何を意味するのか，なぜその場面で使われるのか，そしてどの数式から変換されたか等を知る手がかりを得ることに等しい．しかしながら，数式検索に関する研究は未だ少なく，現存する実用的な検索システムのほとんどにおいて数式媒体は対象とされていない．

我々は数式の持つ意味を解析することによって数式検索の新たな枠組みを作り出すことを目的としている．数式は抽象化された表現であるが故に，それ単独では十分に情報を引き出すことが難しい．我々は，数式の持つ独特な構造と数式周辺の自然言語による説明記述の両方を考慮することで，状況を踏まえた柔軟な数式検索が可能になると考えている．その第一段階として，数式に対応する説明記述をその周辺テキストから抽出する手法について考える．

まず初めに，抽出のタスクを厳密に定義すると共に，学習および精度測定のため，100編の日本語論文からのデータセットの作成について述べる．そして説明記述の抽出法について，パターンマッチングと機械学習の2種類のアプローチを利用した手法をそれぞれ提案し，実験を通して有効性を検証する．

2. 関連研究

数式に関する情報を扱っている研究は，数式検索に関する研究と数学的知識獲得に関する研究の大きく2つに分類できる．

数式検索に関する研究は，数式の持つ独特な構造をどのように表現し，その構造を用いて数式間の類似度をどのように計算するかという問題を主に扱っている．Adeel らは Mathematical Markup Language (MathML) で表された数式に対して，正規表現を用いて数式中の特徴表現(根号, 行列など)を抽出し，それらをクエリとして従来の自然言語を用いた検索システムを利用することで数式情報を検索している[1]．橋本らは MathML 式の XPath を用いてインデックス付けを行い, 共通の XPath を持つ数式を高速に検索する仕組みを構築している[2]．我々もまた, MathML によって木構造で表現された数式に対し，その部分構造を用いた類似度の計算手法を提案している[3]．これらの研究は，数式の構造を考えることによって表面的に類似した数式を獲得できているが，数式中の変数や関数について意味的な側面を考えていないため，それらの持つ曖昧性を解消できていない．

一方，数学的な知識獲得を目的とした研究は，一般的な数学の知識，公式，意味のデータベースを自動的に構築することを重視している．Jeschke らは LaTeX により書かれた科学文書を, 数式を MathML に変換したのちに構文解析等を行うことで数学概念を抽出し，データベースを構築する枠組を提案している[4]．このアプローチは文書を解析して意味情報を取得しているが，数式独特の構造情報を生かしていない．

我々は数式の持つ，意味と構造の両面を考えた数式検索の実現を目指す．今回はそのうちの意味情報に焦点をあて，文書ごとに定義された数式の説明記述の抽出を目的とした．

3. データセット構築

数式の説明記述に関する情報抽出に有効なデータセットは知る限りでは存在しない．そこで我々は学習や評価のため，データセットを人手で作ることから始めた．一連の流れを図1に示す．

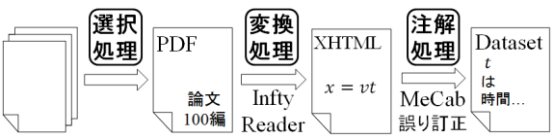

図1 データセット構築の流れ

選択処理は，データセットに用いる論文を選出することが目的である．我々は少ないデータから数式間の関係や類似性を発見できるように，比較的近いトピックの論文を選出することを心掛けた．具体的には，情報処理学会論文誌 Vol.41～49 のうち，表 1 中の語を含む 214 の論文を候補とし，その中から数式が少ないもの，数式に関する説明記述が少ないと思われるものを除き，さらに同じ文献を参照している，またはお互いが参照/非参照関係にある論文を重視し，最終的に 100 編の論文を選出した．

表 1 論文選択のためのキーワード

No.	キーワード
1	機械学習
2	教師あり学習
3	教師なし学習
4	サポートベクターマシン(SVM)
5	ニューラルネットワーク

次に変換処理として，これら 100 編の PDF 論文を数式対応 OCR である InftyReader[5]を用いて，XHTML 形式に変換した．この際に数式は MathML として保存される．この成果物には OCR による誤りも存在し，文章構成や数式構造に影響を与えると思われる誤認識に関しては人手で修正を行った．

最後に注解処理として，数式と，それに対応した説明記述の組を手作業で抽出し，それらを記入したデータセットを作成した．まず各文章に対して，HTML タグを取り除き，数式部分に関しては一時的に“Exp”と変換した．その文章に MeCab[6] による形態素解析を施し，数式ごとに各説明記述部分の対応に応じて BIO タグを付与した．説明記述は名詞節や修飾語など様々な形が考えられるが，今回は簡単のため，名詞もしくは複合名詞のみとした．

データセットの一例を表 2 に載せる．この文は数式が 2 式あるため，データ部分は 2 列で表現される．Exp1 式に対しては「分布」と「事前確率分布（先に述べた説明記述の定義により，『パラメータ Exp2 の事前確率分布』とはならない)」が，Exp2 式に対しては「パラメータ」が説明記述にあたる．

4. 説明記述の抽出法

本節では，数式に対応した説明記述を抽出するために，パターンマッチング・機械学習をそれぞれ用いた 2 つのアプローチの手法を提案する．

4.1 基本アプローチ

我々の目的は，数式が与えられたときにその数式の説明記述を見つけることである．ここで説明記述とは，その数式の意味や定義，名前に関して述べている記述のことである．今回は問題の単純化のため，数式の説明記述は (1) すべて名詞もしくは複合名詞である，(2) 数式と同じ文中に存在する，という 2 つの制限を設けた．この制限により，それぞれの数式に対して文中の各名詞が説明記述であるかを判定する二値分類の問題に帰着する．

表 2 データセット例

ID	形態素	タグ	
0	ここ	O	O
1	で	O	O
2	,	O	O
3	分布	B	O
4	Exp1	Pred	O
5	は	O	O
6	パラメータ	O	B
7	Exp2	O	Pred
8	の	O	O
9	事前	B	O
10	確率	I	O
11	分布	I	O
12	を	O	O
13	示す	O	O

我々の手法の基本的なアルゴリズムは以下の通りである．まず文章を文ごとに分け，それぞれの文から数式と，形態素解析により名詞と判断された単語を取り出す．その際に，連続した名詞は一つの複合名詞として以降扱うこととする．続いて，後に述べるそれぞれの手法によって，各数式に対し，各名詞が対象数式の説明記述であるかを判断する．先の例(表 2)を考えると，2 つの数式に対して 4 つの名詞が存在するため，合計 8 組の二値分類問題を解くことになる(図 2)．

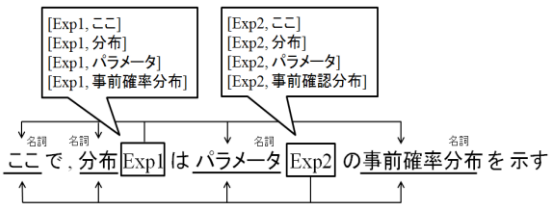

図 2 分類リスト

4.2 パターンマッチング

ここではパターンマッチングを用いた抽出手法について述べる．本手法は，数式の説明記述は形式的な表現が多く，限られた数のパターンである程度網羅できるという仮説の下で発案している．今回我々は情報処理学会論文誌よりデータセットに用いたものとは異なる適当な 5 論文を選び，手作業で論文中に頻出する 8 つの説明記述の表記パターンを得た．それらのパターンを表 3 に示す．

[名詞]は判定対象の名詞を指し，[Exp]は同じく判

表 3 獲得した頻出パターン

No.	パターン
1	[名詞] (+[他 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” (+…) +[Exp]+“で”+“表す”
6	[名詞]+“を/は”+[Exp]+“,”
7	[Exp]+“を/は”+[名詞]+“,”
8	[Exp]+“は” (+…) +[名詞]+“を”+<示す>

定対象の数式を指す．<動詞>はその動詞の活用形にも対応することを表し，スラッシュ(/)は論理和，(+…)は任意の 0 以上の単語列を指す．(+[他 Exp]+“,”+…)は，0 以上の対象以外の数式とカンマの組があることを示す．すなわちパターン 1 は，対象の名詞が対象数式の直前に存在する，あるいは間に数式とカンマ以外の要素が存在しない時にあてはまる．表 2 の例に適用すると，[Exp1，分布]，[Exp2，パラメータ]はパターン 1 に，[Exp1，事前確率分布]はパターン 8 に該当する．

これらのパターンを用い，いずれかのパターンにあてはまる数式と名詞の組に対しては真，いずれのパターンにもあてはまらない場合は偽を返す分類器を作成した．

4.3 機械学習

機械学習アプローチとして，先の節でも述べたパターン情報に加え，その他自然言語に関する情報を特徴素として教師あり学習を行う手法を提案する．4.1 節で述べた通り，今回は全て二値分類を行うタスクであるため，サポートベクターマシンの二値分類モデルを用いた．学習に用いた特徴素を表 4 に列挙する．これらの特徴素は大きく 4 つに分類できる．1 つは 4.2 節で述べたパターンに関する特徴素，2 つめは対象の名詞と数式の間に存在する記号や助詞などに関する特徴素，3 つめは対象の名詞と数式それぞれの近傍の単語に関する特徴素，最後に Cabocha[7] による係り受け解析に基づく特徴素である．

これらの特徴素を，3 節で述べたデータセットの正解データに対して用い，機械学習を行った．アルゴリズムとして Primal Estimated sub-GrAdient SOlver (Pegasos)[8] に基づく，L2 正則化 L1 損失関数サポートベクターマシンを，実際に機械学習ソフトウェアとして Classias[9] を用いた．

表 4 機械学習の特徴素性

大分類	特徴素
パターン	パターン(1～8)にマッチする
対象ペアの関係	単語数(1,2,…, -1, -2,…)，順序
	数式/カンマ/開括弧/閉括弧/は/を の有無
対象名詞/数式周辺情報	対象名詞の名称/複合名詞か否か
	直前/直後の単語の名称/品詞
	両方から後ろにある直近の動詞の原型
係り受け情報	対象名詞を含む節と対象数式を含む節との係り関係の有無/同じ節かどうか

5. 実験と考察

ここでは，4 節で述べた手法の有効性を確かめるため，3 節で説明したデータセットを用いて実験を行った．まず 100 の論文からなるデータセットを訓練データ 60，検証データ 20，テストデータ 20 に分けた．それぞれ 3,867, 1,267, 1,193 の正例, 53,153, 17,440，16,219 の負例が含まれる．それぞれの手法の精度を，Precision，Recall，および F 値から評価を行った．またベースラインとして，対象の名詞が対象の数式の直前にある時のみ真を返す単純な手法を用いている．評価結果を表 5 に示す．機械学習手法に関しては，全ての特徴素(all features)，パターンを除いたもの(w/o patterns)，係り受けを除いたもの(w/o depend)，パターンと係り受けを除いたもの(w/o pat&dep)の 4 つのモデルを評価している．

表 5 各手法の評価結果

手法	Precision	Recall	F 値
ベースライン	0.8990	0.5817	0.7064
パターン	0.8709	0.7125	0.7838
w/o pat&dep	0.8735	0.8106	0.8409
w/o depend	0.8743	0.8106	0.8412
w/o pattern	0.8732	0.8139	0.8425
all features	0.8732	0.8139	0.8425

実験結果に関して，最も良い結果を得たのは機械学習手法，その中でも係り受け情報を用いたモデルであった．全体的に機械学習手法はパターンマッチングによる手法より良い結果を得ている．また，機械学習手法ではパターンの特徴素の有無は結果に大きく影響しないことがわかる．これらの結果より，手動で抽出したパターンは特徴素の組み合わせにより補えると言える．また，係り受け情報は文章の構造を知る手がかりとして，有効に働いていると言える．また 4.1 節で述べた連続した名詞を複合名詞として扱うという制約により，誤って名詞を複合している場合が存在しており，それにより F 値で約 6%の精度減少が確認されている．そのような学習以前の問題も考えると，機械学習手法によって得られた 84%の精度は結果として良好であると言える．

パターンマッチング手法における，パターンごとの結果を表 6 に示す．パターン 1 が最も良い結果を示し，数式の説明記述として最も頻繁に使われていることがわかる．一方，パターン 5 とパターン 8 は今回のテストデータには全く出現しなかった．この結果から，数式に関する説明記述パターンは，各論

表 6 パターンごとの抽出精度

No.	Precision	Recall	F 値
1	0.8910	0.6169	0.7291
2	0.9635	0.0344	0.0663
3	0.3000	0.0050	0.0099
4	0.8036	0.0377	0.0721
5	0.0000	0.0000	NaN
6	0.8571	0.0151	0.0297
7	0.6667	0.0050	0.0100
8	NaN	0.0000	NaN
All	0.8709	0.7125	0.7838

文においては種類が多くないものの，論文の分野や著者によって大きく異なることが推定できる．

また，本機械学習手法において，データセットの量の妥当性を考察するため，訓練データの数に応じた精度の変化を確認した．全ての特徴素を用いた際の学習曲線を図 3 に示す．まだ学習曲線は完全に収束しておらず，データセットのサイズを増やすことでパターンマッチングによる手法との差はより開くことが考えられる．

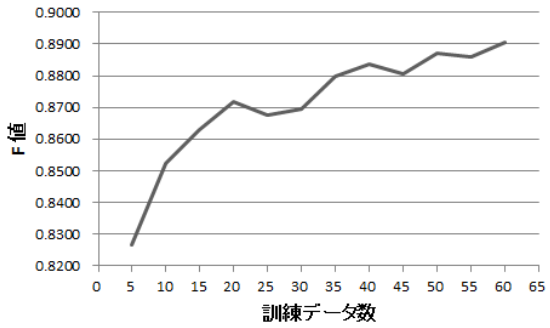

図 3 機械学習手法の学習曲線

6. 結論

我々は科学論文中の数式とその周辺テキストに注目し，数式の説明記述を抽出する手法を提案した．抽出のため，まずデータセットを人手で構築し，実験を通して我々の提案した機械学習手法が効果的に働いていることを示した．今後の課題としては，本手法により得られた数式の説明記述を用いて，数式の状況と意味をより深く考慮した類似数式検索の実装が考えられる．

しかし，今回の説明記述の情報抽出タスクに関しても課題は残っている．まず，今回の手法および実験に課したいくつかの制限による問題である．今回は名詞および複合名詞のみを抽出の対象としたが，それによって「こと」「パラメータ」「変数」など非常に曖昧な言葉しか獲得できない状況も多く存在した．有効な情報を取得するためには対応する名詞節全体が必要となる場合も多く，抽出タスクに関してさらなる考察，必要に応じてデータセットの形式も再考する必要がある．

また，情報抽出のさらなる精度向上も課題である．精度向上には，特徴素の追加，および現在用いている特徴素の有効性の検証が必要である．また，係り受け解析結果によっては学習せずに(主に負例であると)判断できるものも存在すると思われ，精度向上や時間短縮に有効だと考えられる．

最後に，これらの提案手法は日本語に限らず他の自然言語にも対応可能であると考えている．数式の説明記述は，言語による差異はあるが，各々の言語内では一般的な表記が存在するため，機械学習によってそれぞれの特徴を探ることは有効だと考えられる．

参考文献

[1] Adeel, M, Cheung, H., S., Khiyal, S., H.: 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.

[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] Yokoi, K., Aizawa, A.: An Approach to Similarity Search for Mathematical Expressions using MathML. DML, pp.27-35, 2009.

[4] Jeschke, S., Wilke, M., Blanke, M., Natho, N. M., Pfeiffer, O. F.: Information Extraction from Mathematical Texts by Means of Natural Language Processing Techniques. EMME'07, pp.109-114, 2007.

[5] Suzuki, M., Kanahori, T., Ohtake, N., Yamaguchi, K.: An Integrated OCR Software for Mathematical Documents and Its Output with Accessibility. ICCHP '04, LNCS, vol.3118, pp.648-655, 2004.

[6] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

[7] 工藤 拓，松本 裕治: チャンキングの段階適用による日本語係り受け解析．情報処理学会論文誌, Vol.43, No.6, pp.1834-1842, 2002.

[8] Shalev-Shwarz, S., Singer, Y., Srebro, N.: Pegasos: Primal Estimated sub-GrAdient Solver for SVM. ICML '07, vol.227, pp.807-814, ACM, 2007.

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