意味と構造を考慮した数式検索手法の提案

横井 啓介 [1]　Minh-Quoc NGHIEM [2]　松林 優一郎 [3]　相澤 彰子 [1,3]

1 東京大学 情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1
2 総合研究大学院大学 複合科学研究科 〒101-8430 東京都千代田区一ツ橋 2-1-2
3 国立情報学研究所 コンテンツ科学研究系 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: {kei-yoko, nqminh, y-matsu, aizawa}@nii.ac.jp

あらまし　本稿では，類似数式を検索する手法について述べ，その有効性を検証する．具体的には，科学論文中に存在する数式に焦点をあて，XML で表現された数式固有の構造，および数式周辺のテキストから抽出した変数や関数記号の定義・説明表現を用いて類似した数式を検索する手法を提案する．また，論文中の数式に対して類似性を提示する環境を実際に構築した上で，検索結果の適合性判定を行い，提案手法の性能を調べる．実験の結果，比較手法に対する優位性，および提案手法中に用いられている類似性の尺度の有効性が確かめられた．

キーワード　数式検索，MathML，Content Markup，情報抽出

An Approach to Mathematical Search Considering Structures and Semantics

Keisuke YOKOI[1] Minh-Quoc NGHIEM[2] Yuichiroh MATSUBAYASHI[3] and Akiko AIZAWA[1, 3]

1 Department of Computer Science, University of Tokyo　7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan
2 Department of Informatics, The Graduate University for Advanced Studies　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
3 Digital Content and Media Sciences Research Division, National Institute of Informatics　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
E-mail:　{kei-yoko, nqminh, y-matsu, aizawa}@nii.ac.jp

Abstract　This paper proposes a method to search for similar mathematical expressions and examines the efficiency of the method. Concretely speaking, we focus on mathematical expressions in scientific papers and propose a new method to search for similar mathematical expressions. In our method, we consider both the structural similarity based on the XML format representations and also the semantic similarity based on the definitions and explanations of variables/functions used in the expressions which we automatically extract from surrounding texts. In order to investigate the effectiveness of our method, we prepare 100 manually annotated papers and implement a mathematical search system. In addition, we propose a new way of evaluation about relevancy of search results, although many past works cannot evaluate in clear way.

Keyword　mathematical search，MathML，Content Markup，information extraction

1. はじめに

数式は，科学論文等において，主張や議論を明確に伝えるための重要な媒体である．数式は自然言語とは大きく異なる独特な構造を持つため，従来の自然言語を対象とした検索システムでは扱うことができない．このため，数式に特化したシステムが必要であるが，現時点では数式を対象とした有効な検索システムは存在していない．数式を検索することの難しさは，主に以下の 2 点に起因すると考えられる．一つは，数式独特の複雑な構造である．数式構造を正確に，そして一意に表すのは非常に難しく，類似した数式を直接検索する手段が未だ確立していない．もう一つは数式の曖昧性である．数式は抽象化された表現で，その正確な理解には数式単独では不可能である．数式は，その意味や使い方に対する説明があって初めて意味をなす．

そこで我々は，これら 2 つの問題を解決するための新たな類似数式検索手法を提案する．まずはじめに，複雑な数式構造を検索可能にするために，数式を木構造で表現する MathML (Mathematical Markup Language) を用いて，木構造の類似度から数式の類似度を計算する．また，数式により深い意味を持たせ，曖昧性を解決するため，数式の周辺テキストに着目し，そこから数式に含まれる変数や関数といった記号の名前や定

義，説明表現を抽出する．最後にこれらの情報を用いて数式を検索する手順について触れ，さらにこれまで非常に難しいとされていた数式の類似性の評価の問題に対し，科学論文中の数式に目を向け実際の数式の利用状況に即した評価を定めることで解決をはかる．

本稿の構成は以下の通りである．まず次節では，これまで行われてきた数式検索に関する先行研究を紹介する．そして 3 節で我々の提案する数式検索に関する一連の流れを，MathML を用いた数式構造の類似度計算，数式に含まれる変数や関数表現の定義・説明表現の抽出手法，数式検索システムの評価方法，以上 3 つに分けて説明する．そして最後に提案手法に関して実際に評価を行い，結論として考察と今後の展望について述べる．

2. 関連研究

これまでの数式検索に関する研究は，大きく 2 つに分けることができる．1 つは数式の独特な構造をどのように扱い，いかにして数式を探すかを目的としたものである．そしてもう 1 つは数式を含む文書に目を向け，数学的な知識獲得を行うことを目的としたものである．以下，それぞれについて例を挙げ，簡単に説明する．

数式を探すことに関する研究は，いずれも数式をどのように表現し，どのように類似性を定義するかに主題を置いている．Adeel らは MathML 式から，正規表現を用いて root，matrix といった数式の特徴要素を抽出し，それら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検索を行う手法を提案した[1]．Mistuka らは前処理として変数や数式の形を標準形に変換を行うことで，柔軟な前文テキスト検索を提案した[2]．橋本らは MathML の XPath でインデックス付けを行うことで，高速な類似数式検索を可能にしている[3]．小田切らは MathML をクエリに適した形に拡張を行うことで，数式の部分的な同定検索を可能にしている[4]．そして我々もまた，MathML 式の部分構造を利用した類似度算出法を提案した[5]．これらの手法は，確かに見た目上似ている数式や，同じような表記を持つ数式を探すことはできているが，数式の変数や関数の意味など，その曖昧性を消すことはできておらず，その結果実際に「似ている」式であるか，状況に適した数式が得られているかに対する評価は明確でない．

一方，知識獲得を目的とした研究は，数式を含む文書から一般的な数学的知識を獲得しようとするものである．Jaschke らは数式を含む文書中から自動的に数学的なオントロジーを抽出し，データベースを構築する枠組を提唱している[6]．これは LaTeX で書かれた数式を MathML に変換したのち，構文解析等，自然言語処理による数学概念の関係抽出，その後にそれらの情報をまとめてグラフ化を行っている．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，その関係をデータベースに蓄積させて知識として利用する手法を提案している[7]．これらの研究は数学概念の一般的な知識を得ようとしているが，数式の持つ複雑な構造に対して十分であるとは言えない．

このように，従来研究における，数式のみを考慮した類似検索や周辺文書と表面上の数式表記のみを考えた知識獲得では，それぞれに曖昧性，そして構造といった解決が難しい問題を含んでいる．そこで我々はその両面を同時に考えていくことで，実際の状況に有効である数式検索手法を提案すると共に，これまでこの分野で難しかった性能評価の可能性を見出すことを目指す．

3. 提案手法

本節でははじめに構造を利用した類似性の計算手法，および数式の周辺テキストを利用した，変数や関数などの数学記号に対する定義・説明表現の抽出手法についてそれぞれ述べ，その後それらを同時に考えることによる数式検索の可能性と，実用的な評価方法について説明する．

3.1. 数式構造を用いた類似度計算

MathML は W3C が提唱している数式を Web 上に表現するための標準的な表記である．MathML には，数式の表示に特化した Presentation Markup と，より構造

```
<apply>
    <plus />
    <ci> x </ci>
    <apply>
        <divide />
        <mn> 3 </mn>
        <mi> a </mi>
    </apply>
</apply>
```

図 1 Content Markup コード例

的な意味の表現を意識した Content Markup の 2 種類の記法が存在する．関連研究の多くは，その入手の容易さから Presentation Markup を用いた研究が多いが，今回我々は意味構造をより深く考慮している Content Markup を利用して数式間の類似度を計算する．Content Markup の例を，数式$a+b$を用いてそのコードを図 2 に，木構造を図 2 に示す．

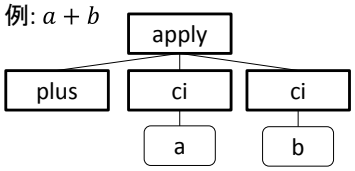

図 2 Content Markup 構造例

このように，Content Markup は関数とその引数の対応関係が明確で，数式の構造関係を表すのに適している．この構造を生かしつつ計算量を抑えた類似判定を行うため，市川らがテキスト構文構造の類似度の尺度として提案した *Subpath Set* を数式構造に応用した[8]．これは，部分経路(subpath)を，「根から葉までの経路とその一部」として定義し，「各々の構文木の部分経路の集合」を Subpath Set と定義したものである(図 3)．この定義に従って，各数式から得られた Subpath Set に対し，いくつかの集合の類似度の尺度の中で試行錯誤した結果，Jaccard 係数を数式間の類似度として利用した．

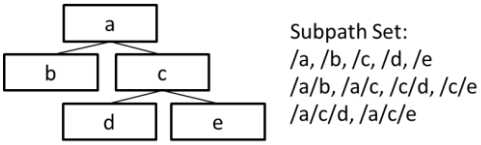

図 3 Subpath Set

しかし，Subpath Set は親子関係に重点を置いた構造表現であるが，Content Markup の Subpath では関数とその引数の関係を表現できない．そこで我々と引数を兄弟関係にしている“apply”タグに注目し，「“apply”タグの長男ノードを親ノードである“apply”タグに置き換え，“apply”タグを取り除く」という操作を行うことで，関数と引数の関係を Subpath 中に含めることを可能にした．この新たな数式構造を *Apply-free* Content Markup と呼ぶ．図 4 に *Apply-free* Content Markup への変換例を載せる．この例は$a+b$という数式を扱っているが，変換前の Subpath Set には，「何かを足す」「変数を扱う」という意味を表す Subpath は含まれるものの，それらをまとめた「変数を足す」という Subpath は得ることができない．そこで，親ノードである“apply”タグをその長男の“plus”タグに置き換えることで，右のような *Apply-free* Content Markup 構造ができる．この構造は，返還前よりはシンプルな構造となっているものの，意味のある Subpath は全て含まれていることがわかる．

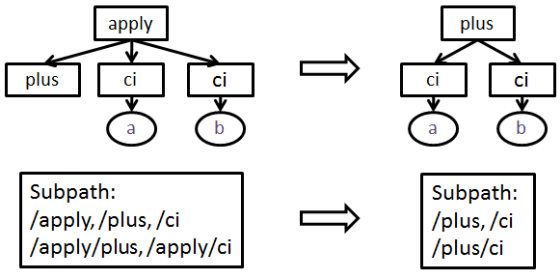

図 4 *Apply-free* Content Markup

数式間の類似度の計算法をまとめると，各数式の Content Markup 表現を *Apply-free* に変換を行い，その後に Subpath Set を構築，そして両式の Subpath Set の Jaccard 係数の値を算出することでその値を数式間の類似度と定義した．

実際に本手法を用いて検索した結果(上位 5 式)を表 1 に載せる．今回は検索クエリとして，三角関数(sine)の加法定理を用いた．The Wolfram Functions Site[9]よりクロールした 155,607 の数式を検索対象としている．The Wolfram Functions Site には，数学やその他科学において用いられる公式を中心とした数式が多く存在し，それぞれの数式について MathML の記述が，Presentation Markup，Content Markup 共に付与されていることが特徴である．検索結果を見ると，sine の加法定理に加え，cosine の加法定理や複素数の加法定理など，柔軟に構造の似ている数式が得られていることがわかる．部分的な構造のみから類似数式を算出する従来研究に比べ，Subpath Set により木構造全体を眺めた類似性が判断できていることがわかる．

表 1 検索結果例

No.	
	$\sin(a+\mathrm{b})$ $=\sin a\cos b+\cos a\sin b$
1	$\sin(a+\mathrm{b})$ $=\sin a\cos b+\cos a\sin b$
2	$\sin(a-\mathrm{b})$ $=\sin a\cos b-\cos a\sin b$
3	$\sin(a+i\mathrm{b})$ $=\sin a\cos b+i\cos a\sin b$
4	$\cos(a-\mathrm{b})$ $=\cos a\cos b+\sin a\sin b$
5	$\cos(a+\mathrm{b})$ $=\cos a\cos b-\sin a\sin b$

3.2. 定義・説明表現の抽出

この処理の目的は，数式の周辺テキストを解析し，機械学習を用いて数式中の変数・関数に関する説明記述を抽出することで，数式により多くの意味情報を持たせることである．このような数式を含む文書の情報抽出のタスクに関連する研究は見つからなかったため，まず学習のためにデータセットを手作業で作成した．作業の流れを図 5 に示す．

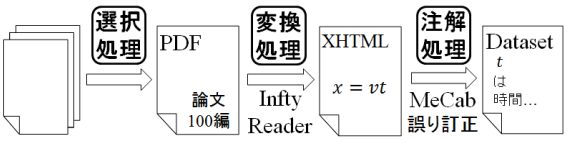

図 5 データセット作成の流れ

まず初めに選択処理として，少ないデータから関連性が認められるように，内容が近い論文 100 編を人手で選択した．今回は情報処理学会論文誌より発行される論文のうち「機械学習」に関係するワードを文書中に持つ論文を選んだ．

次に，変換処理として数式対応 OCR である InftyReader[10]を用いて PDF 形式の論文から XHTML 形式への変換を行った．この際に，数式は MathML に変換される．

最後に，注解処理として，先の処理により得られた XHTML 文書群に対し，まず OCR の誤りのうち，文書構造や数式に大きく影響を与えると思われるものは手作業で訂正を行った．その後に MeCab[11]による形態素解析を行い，数式とその説明記述のペアを人手で探し，抽出・整形を行って，データセットを作成した．例を表 2 に示す．これは「ただしt_xは語xの出現頻度とする」という一文からデータセットを作った場合である．この例においては，数式が二つ使われている(データ中では Exp と変換している)ため，タグは二行与えられる．それぞれの数式の説明表現となっている部分(それぞれ出現頻度,語)に B/I タグを振り分けている．

今回は簡単のために以下の 2 つの制限を設けている．

- 数式の説明記述は，すべて名詞，もしくは複合名詞とする．
- 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．

以上の制限のもとで，各数式に対し，その説明として適する説明記述を(存在すれば)選択する．

学習には，MeCab による形態素情報に加え，Cabocha[12]により得られた係り受け関係から対象となる数式と説明記述候補の名詞の間の関係を特徴素として選び，サポートベクターマシンの二値分類モデルを適用した．特徴素に関する詳細を表 3 に挙げる．その結果として，表 4 のような結果を得た．今回は連続した名詞はすべて複合名詞として扱っており，機械学習の性能とは別に，その区切り方により正解できない例が全体の 6%程度存在することも考えると，高い性能で抽出できていると言える．

表 2 データセット例

ID	単語	タグ	
0	ただし	O	O
1	Exp	Pred	O
2	は	O	O
3	語	O	B
4	Exp	O	Pred
5	の	O	O
6	出現	B	O
7	頻度	I	O
8	と	O	O
9	する	O	O

表 3 機械学習に用いる特徴素

大分類	特徴素
パターン	あらかじめ別論文 5 編より得られた説明記述の頻出出現パターン 8 つのうち，いずれかにマッチする
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
	両方の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/係り先が同じか

表 4 抽出性能

Precision	Recall	F1-measure
0.8732	0.8139	0.8425

3.3. 評価手法

従来研究においては，数式の類似性に関する議論が明示的にされることが少なく，数式検索システムでは性能評価が課題であるとされていた．すなわち，トピックや状況によって求められる数式は異なり，明確な正解は存在しないため，与えられた数式に対して，掲示される数式が適切であるかを，数式の情報のみから判断することはむずかしい．したがって，これまで行われてきたような方法では，類似数式検索に関する評価は曖昧にならざるを得ない．

これに対して我々は，「論文を読む」という状況の下，関連性のある数式が検索できているかという観点から，掲示される数式の適合性をユーザが判定する手法を適用する．数式が似ているかどうかでなく，その数式が論文を読む際に参考になるかどうかで適合性を判定することにより，実際に必要な数式を獲得できるかどうかを判断する．

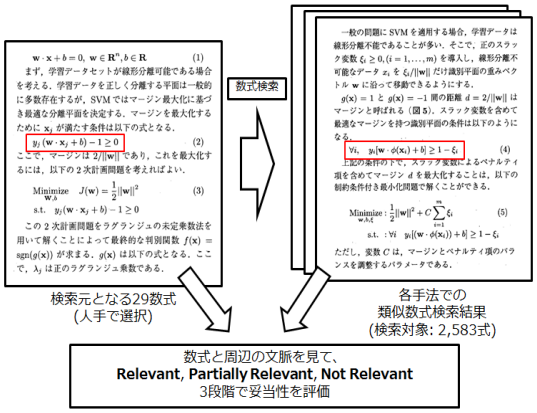

図 6 評価の流れ

評価手法の流れを図 6 に示す．まず初めに，論文中において，著者の主張を理解するために必要と考えられる数式を 1 論文につき 3〜5 式，7 論文から合計 29 の数式を選出した．それぞれの数式に対し，各類似数式検索システムを用いて検索結果を獲得する．検索対象として，今回は 3.2 節でも述べた 100 の論文中に存在する 2,583 の数式を対象としている．この際に，類似度をスコア付けしている提案手法のような手法に対しては，上位 5 件を適合性の判定を行う候補式とした．そして判定者に対して，検索元の数式と検索結果の数式を，それぞれが使われている論文の箇所と共に提示する．判定者は提示された情報に基づき，3 段階の適合性評価をそれぞれの数式ペアに対して行う．それぞれの評価とその基準を以下に述べる．

- Relevant: 検索元の数式に対する類似数式として妥当である．
  対象数式が検索元の数式および論文の内容に対し，明らかに関連している，参考になる，もしくは検索結果として然るべきである
- Partially relevant: 検索元の数式に対する類似数式として，やや妥当である．
  対象数式が検索元の数式および論文の内容に対し，関連性を見つけることができる，参考にできなくはない
- Not relevant: 検索元の数式に対する類似数式として，妥当でない．
  対象数式が検索元の数式および論文の内容に対し，関連性が全くない，全く参考にならない

以上の評価に基づき，判定が割れたものを除いた検索結果の数式の評価の分布から，類似検索の評価を行う．

4. 実験

前節で述べた評価手法に基づき，3.1 節の構造に基づく類似度を用いた類似検索を評価する．比較対象としては橋本らの手法[3]を用いた．橋本らの手法は，Presentation Markup で表現される数式を用いて，クエリとして与えられた数式に対して，いちばんはじめの XPath と最も深い XPath の両方を持つものを類似数式として出力している．この計算の際に，すべての検索対象の数式の XPath をインデックス付けしておくことで，高速な処理が期待できる．一方で，2 つの XPath しか判断していないため，数式構造の全体を見ておらず，類似数式検索の妥当性はやや疑問である．表 5 に適合性に関する性能評価を示す．適合性の計算法として，Relevant の全体に対する割合 (*Fully Relevant*)，Relevant または Partially Relevant の全体に対する割合 (*Partially Relevant*)，の 2 種類を用いた．提案手法はランク付けがされているため，類似度の上位 1〜5 件までを用いた結果をそれぞれ載せている．

表 5 適合性評価

		Fully Relevant	Partially Relevant
提案手法	1	0.483	0.828
	2	0.397	0.672
	3	0.356	0.609
	4	0.310	0.526
	5	0.262	0.469
比較手法		0.250	0.577

まず我々の提案手法に対する適合性評価の結果を見ると，検索結果の上位にある数式ほど（n の値が小さいほど）適合性が高いことがわかる．この結果は，提案手法で用いている数式構造の類似度計算がうまく働いていることを示している．また，橋本らの XPath による手法は類似度計算によるランキングを行っていないため，検索元の数式によっては全く結果を返さないこともある．今回の実験に用いた検索式のセットでは 1 数式あたり 1.79 式の類似数式の出力であった．そのため，厳密な数値比較こそできないが，同じような出力件数（たとえば$n = 2$）で比較すると，我々の提案手法の適合性が，橋本らのそれよりも高いことがわかる．

5. 結論

我々は，数式検索が持つ問題である構造の特異性と曖昧性に対応するべく，MathML による数式構造の利用とその周辺テキストからの情報抽出という 2 つのアプローチからなる類似数式検索を提案した．それに加え，これまであやふやにされていた類似数式検索の精度評価に関し，科学論文中の数式に対して実際の数式検索の状況を意識することで，単なる見かけ上似ている数式ではなく，実際にその場面に適した数式が検索できているかを判定する手法を提案した．

今後の課題としては，数式の構造・意味の 2 つの観

点に基づく類似性について，それらをどのように組み合わせれば最良の検索結果を得ることができるかの検討が挙げられる．今回の提案手法では，数式中の変数や関数に「変数」「関数」「こと」など，あまり有意でないと思われる意味が取れている場合も多い．このことからも，意味と構造はそれぞれ別々に類似度を測るのではなく，木構造中に意味情報を埋め込んで類似度を取ることが必須だと思われる．

また，各々の類似度の詳細化・緻密化もさらに検討していきたい．具体的には，構造の類似度に関しては木構造の兄弟関係も視野に入れるべきであり，さらに x^{-1} と $1/x$ のように，構造が異なる場合でも実質的に同じである数式にも対応できるよう，あらかじめ等式変換リストのようなものを作り，それぞれの変形の重みを設定することで，より意味と構造の両方を深く考慮した類似数式検索が可能になると思われる．周辺テキストからの情報抽出に関しては，単なる複合名詞だけでなく名詞節等幅広く抽出するようにタスクを設定することで，より実用的な数式検索システムが可能になると考えている．

謝辞

本研究にあたり，InftyReader による PDF 論文の XHTML 化に協力戴いた九州先端科学技術研究所，サクセスネット(sAccessNet) 代表理事の鈴木昌和先生，および我々の研究成果の実用化，視覚化に多大な協力を戴いたピコラボ相良毅氏に感謝する．

参 考 文 献

[1] M. Adeel, H. S. Cheung, S. H. 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, 2006.

[2] J. Misutka and L. Galambos, “Extending Full Text Search Engine for Mathematical Contente”, Towards a Digital Mathematics Library (DML 2008), pp.55-67, 2008.

[3] 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.

[4] 小田切健一，村田剛志，“MathML を用いた数式検索”，人工知能学会全国大会，2008.

[5] Keisuke Yokoi and Akiko Aizawa, “An Approach to Similarity Search for Mathematical Expressions using MathML”, Towards a Digital Mathematics Library b(DML 2009), pp.27-35, 2009.

[6] S. Jeschke, M. Wilke, M. Blanke, N. Natho, O. F. Pfeiffer, “Information Extraction from Mathematical Texts by Means of Natural Language Processing Techniques”, Proceedings of the International Workshop on Educational Multimedia and Multimedia Education, 2007.

[7] M. Kohlhase, A. 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

[8] 市川宙，橋本泰一，徳永健伸，田中穂積，“テキスト構文構造類似度を用いた類似文検索手法”，情報処理学会研究報告, 2005-DBS-136, pp.39-46, 2005.

[9] Wolfram Research, Inc. The Wolfram Functions Site, http://functions.wolfram.com/.

[10] 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, 2003.

[11] Taku Kudo, “MeCab: Yet Another Part-of-Speech and Morphological Analyzer”, http://mecab.sourceforge.net/.

[12] 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.