表記が異なる同義の数式の高速な検索法

大橋 駿介† 高須 淳宏‡* 相澤 彰子‡†

†東京大学大学院 情報理工学系研究科 〒113-8656 東京都文京区本郷 7-3-1
‡国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2
*総合研究大学院大学 〒240-0115 神奈川県三浦郡葉山町上山口 1560-35

E-mail: {sohashi, takasu, aizawa}@nii.ac.jp

あらまし 従来情報検索の分野では、文献中のテキストに着目した検索手法が中心に研究されてきた。しかし学術論文などではテキストとともに「数式」が重要であり、数式を対象とした検索手法が有用である。数式の検索には、数式を木構造で表現し、pq-gram などの一般的な木構造の類似度計算アルゴリズムを用いることができるが、表記や木の構造が異なるが同義の数式が検索できないため、十分ではない。

そこで本研究では、あらかじめ用意した値を引数として数式に代入し、計算した値を特徴量として検索インデックスに利用することで、表記が異なっていても外延的に同値な関数を高速に検索する手法を提案する。また、実際の論文に含まれる数式データを用いて提案手法を評価する。

キーワード 数式検索，木構造索引付け，pq-gram，MathML

1. はじめに

数式は科学技術のコミュニケーションにおいて言語の壁を超えて用いられる、普遍性と高い表現力をもつテキストである。常時増え続ける学術論文にも多くの数式が含まれており、これらの文献から目的の情報を取り出す数式検索技術の重要性は高い。また、数式検索は数学教育の分野での需要もあり、よい数式検索技術は科学者・技術者に限らない多くの人々に求められている。

ところが、現在普及している Web 検索エンジンは文字や単語の並びである単純テキストを想定して実装されており、構文木によって表される複雑な構造を持つ数式を検索するには不十分である。同じく木構造を持つ XML データに対する研究も行われているが、これらは一般の木構造を対象とするものであり、数式の意味論を考慮した検索を行うことは難しい。

これらの背景に基づき、本研究では数式の意味を考慮したインデックスを作成して数式の検索を行う手法を提案し評価する。数式がいくつかの変数を含む場合、これらに何らかの値を代入して計算した値は、その数式の意味に関する手がかりを含むものと考えられる。本研究ではこの点に着目し、あらかじめ決定した乱数をデータベース中の数式の変数に代入して、得られる計算値を数式の特徴量として検索インデックスに用いる。また、既存の XML データ検索手法との比較実験を行い、提案手法を評価する。

以下、本稿では 2 節で関連研究を紹介し、3 節で本研究において使用した数式検索に関連する技術を説明する。4 節で数式の意味を考慮したインデックスを提案し、5 節で既存の XML 検索手法との比較実験を通して提案した手法を評価する。6 節で提案した手法に考察を加える。7 節で結論を述べる。

2. 関連研究

2.1. XML データ検索

XML データ検索に用いられる柔軟なマッチングを行うアルゴリズムを紹介する。Tree Edit Distance (TED) [3] は文字列に対する編集距離の自然な拡張であり、XML 木構造の類似度として用いることができる。ところが、TED の計算には現在知られている最も時間計算量が小さいアルゴリズム [5] でも木の頂点数 n に対して $O(n^3)$ 時間が必要であり、計算を繰り返し行う必要がある検索タスクには適切でない。pq-gram [4] はラベル付き順序木に対する距離の尺度である。pq-gram は TED へのよい近似でありながら、O(n log n) 時間で計算できる。これを用いて XML 文書を検索する手法 [6] も提案されている。pq-gram の詳細については 5 節にて改めて説明する。

これらの手法を用いると、XML の構造を反映した類似度を計算することができるが、XML 自体は汎用的なデータ記述言語であり、構造に対して特定の意味を仮定していない。したがって、TED や pq-gram などの既存手法も、特定のドメインで定義される構造が持つ意味については取り扱うことができない。

2.2. 数式検索

数式を対象とした検索手法を紹介する。

Kamali ら [1] は、数式検索の手法を、クエリと完全一致する数式を取り出す厳密マッチング、データベース中の数式の部分構造をインデックスとする部分構造マッチング、そしてクエリにワイルドカードを許した

パターンマッチングに分類し、これらの特徴を分析した。これによると、厳密マッチングは部分構造マッチングやパターンマッチングに比べて正確に候補を見つけることができるが、発見できる候補の数が少ない。部分構造マッチングとパターンマッチングではパターンマッチングの方がややよい結果が得られたが、よい結果を得るためにユーザーは検索結果を見ながらクエリの修正をする必要があると報告されている。

また、数学検索システムの性能比較のために NTCIR-10 Math pilot task [2] において、評価用データセットが構築されている。この評価用データセットを用いた参加システムどうしの比較によると、部分構造マッチングに基づく検索システムは MAP 性能などの統合的な検索性能に優れているが、ランキング結果に不適合の数式が含まれるケースが避けられない。一方、パターンマッチングに基づく検索システムでは、検索結果として得られた数式については正解率が高いが、クエリによっては 1 件も数式がヒットしない場合がある。なお、厳密マッチングでは検索可能な数式に対する制約が強すぎるため、NTCIR-10 Math では厳密マッチングに基づく方式を用いた参加チームはいなかった。

以上より本稿では、XML データ検索でも従来から検討されている、部分構造マッチングに基づく手法に焦点をあてて、その数式検索への適用方法について検討する。

3. 関連技術

3.1. MathML

MathML [7] は W3C によって勧告された数式表現用のマークアップ言語である。XML をベースとしており、Web ページ上での数式表現を中心に、コンピュータ上で数式を扱う際に標準的に用いられている規格の 1 つである。数式にはその外観と意味の 2 つの側面があるが、MathML ではそれぞれ Presentation Markup と Content Markup によってそれらを表現することができる。表 1 は式 xy に対する Presentation または Content Markup である。Content Markup においては乗算を表す<times>タグと関数適用を表す<apply>タグによって表現されているが、実際の数式において乗算記号は省略されているので Presentation Markup においては変数を並べたものとして式が表現されている。

表 1: 2 種類の MathML Markup

Presentation Markup	Content Markup
<math> <mrow> <mi>x<mi/> <mi>y</mi> <mrow/> </math>	<math> <apply> <times/> <ci>x</ci> <ci>y</ci> </apply> </math>

3.2. Mathematica

Mathematica [8] は、ウルフラム・リサーチ社が開発している数式処理システムである。

Mathematica は図 1 に示すように外部ファイルに MathML などの形式で保存された数式をインポートして使用することができる。Mathematica は記号操作に基づいた高度な数式処理を実現しており、本研究においては数式に含まれる変数を抽出する、代入操作を実行するといった操作を利用して数式の特徴量計算に利用した。

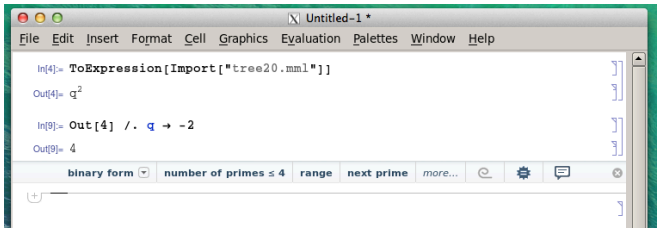


図 1:Mathematica で数式を取り込んだ例

4. 手法

4.1. 数式の意味を踏まえた特徴量

数式は、その表記としては記号列であるがその表記に従って自然言語やプログラミング言語と同様に構造を持ち、この構造に基づく意味を持っている。例えば自然言語において否定語の影響する部分は構文木の構造によって決定されこれは文全体の意味を左右する。また、プログラミング言語においては抽象構文木の評価順序や構造から値への対応付け、すなわち意味論が定まっており、これは構造が持つ意味にほかならない。Mathematica に代表される数式処理系は、数式に対して定めた評価規則を実装したシステムであるといえる。

したがって、数式処理系を通じて数式を評価すると評価規則に基づく意味を含む出力を得られるとみなすことができる。出力を特徴量として扱うためにはそれが実数値であることが望ましいが、この際にほとんどの数式には変数が含まれていることが問題となる。

そこで、本研究ではこれらの変数に対して事前に生成した乱数を代入して評価することで得る実数値を特徴量とする手法を提案する。これにより、評価値が実数値となるような数式に対して意味論に基づく特徴量を得ることができる。

4.2. 手法の対象

4.1 において提案した特徴量には、評価の結果が実数値でない（例えば行列やベクトルである）数式には適用することはできない。ゆえに、本手法の対象は学術的文献等に含まれる、実数値を値として持つ数式の検索である。

4.3. 手法の詳細

4.3.1. 特徴量の計算

数式に対して次の操作を行い、特徴量を計算する。

1. 等号・不等号を含む数式は、それをいくつかの部分に分割する。すなわち、等号・不等号を N (≧0) 個含む数式は N+1 個の部分に分かれる。このとき、各部分はなんらかの実数を表現する式となっている必要がある。

2. 各部分の変数に対して、予め生成した共通の乱数を代入する。この乱数は、システム全体で共通する 1 つのセットを用いる。これによって 1 つの部分に対して 1 つの特徴量を得ることとなる。

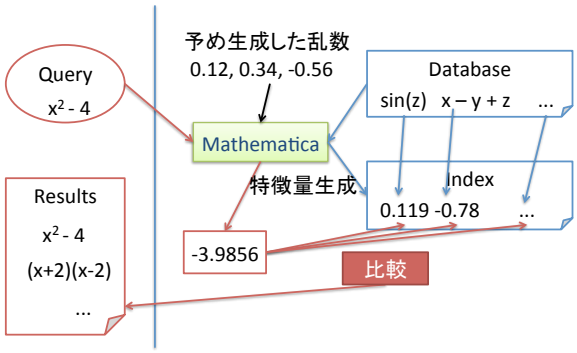
これによって、数式に対してその部分の個数分の特徴量を持つ集合を得る。

4.3.2. 索引付け

4.3.1 節で述べた手法で全ての数式に対して特徴量を計算し、これを転置インデックスによって索引付けする。

4.3.3. 検索

クエリを処理する際には、まずクエリの特徴量の集合を得る。そして、特徴量集合の要素全ての転置インデックスを参照し、その指し示す要素を全て出力とする。



4.4. データフロー

4.3 節で述べた操作を図 2 に示す。図 2 において青線部は前処理の段階での操作、赤線部はクエリを受け取った時の操作を示す。

図 2 提案手法の模式図

5. 実験

5.1. 実験手法

東京書籍株式会社が出版する高等学校数学科の教科書に含まれる数式データ（データ数 30,569）をデータベースとし、データベース中の数式自身をクエリとして、pq-gram によるインデックスと、提案手法によるインデックスそれぞれを用いて検索を行い、その出力を比較した。

5.1.1. pq-gram

本実験においては、一般的な XML 文書類似度である pq-gram [4] を用いた検索と比較した。pq-gram はラベル付き順序木に対する類似度で、pq-gram と呼ばれる元の木の部分木をその特徴として抽出し、それらを集めた pq-gram profile の集合としての類似度によって木の類似度とするものである。本実験においては類似度 0.5 以上のものを、20 個を限度として出力した。

5.2. 評価尺度

再現率の尺度として、平均出力件数を用いた。平均出力件数は、各クエリに対して検索結果として得たデータ数の平均である。出力結果が適合しているか、すなわち出力結果がクエリと数学的に関連しているかについては自動的に判断することが困難であるため、いくつかの実際の出力を確認することで評価した。

5.3. 結果

表 2 に平均出力件数を示す。提案手法に対しては、クエリ 30,569 個中 9,869 個（約 32%）の数式に対して特徴量を求めることができず、それに伴い検索に失敗したが、それらのクエリの検索結果は 0 件であるとして計算した。

特徴量の計算に失敗したケースには、(1) 出力がベクトル・行列である式 (2) 集合の包含関係の式 (3) 文書内の別の箇所で定義した関数を含む式 などが見られた。

表 2 平均出力件数

pq-gram	提案手法
19.40	11.77

5.4. 出力例

表 3 にクエリとそれに対する提案手法の出力例を示す。

これらの出力結果は、表記が異なるため pq-gram では類似していないとみなされるが変数名の書き換えを除いてクエリと数学的に同値な数式である。表記でなく数学的意味に基づいた検索ができていることを示している。

表 3 提案手法の出力例

クエリ	検索結果
a^6	$a^{2\times3}$
$(a-c)+c+(b-c)$	$a+b-c$
$(4x^2-3x+10)+(-2x^2+6)$	$(4x^2-3x+10)-2x^2+6$
	$(4-2)x^2-3x+10-6$
	$2x^2-3x+16$
ax^2	PT^2
	CP^2

6. 考察

6.1. 出力の検討

平均出力件数はクエリに対してどれだけ柔軟に候補を見つけ出してきたかを代表する値であり、pq-gramと提案手法では大きな差が付いている。これはpq-gram が部分構造に着目して検索することにより柔軟なマッチングを実現しているのに対し、提案手法は特徴量が等しいかどうかについてのみ、というかなり強い条件のもとで検索をしており柔軟性を欠いていることによると考えられる。pq-gram が 10 未満の候補しか見つけられなかったのは 881 ケースに限られるのに対して提案手法では特徴量計算に失敗したケースを除いても 2,956 ケースあったことも提案手法が柔軟でないことを示唆している。

5.3 節で示した、特徴量の計算に失敗したケースについては、(1) や (2) については提案手法の対象である実数を返す関数でないために生じる、織り込み済みの失敗である。ところが (3) は他の 2 つとは性質を異にしている、というのもこれケースにおいては実数を返す関数という仮定が守られており、実験前には予測できない失敗であるからだ。これは、数式においても自然言語と同様に周辺の文脈を考慮して意味を解釈する必要があることを示していると考えられる。

出力例においては、特に最後の例において顕著であるが、数式の表面的記述では異なるが数学的に同値な多項式が検索結果に出現しており、提案手法の有効性をある程度示しているものと考えられる。

6.2. 数学的解析

数式が変数名の書き換えを除いて同値であるならば、同一の変数に同一の値を代入することで特徴量を一致させることができる。このことから、2 つの数式間で提案した特徴量が一致することはそれら同値であることの必要条件であることがわかる。言い換えると、特徴量が一致することが同値性を調べる上での緩和された条件となっている。このことから、提案した特徴量には同値性を代表する性質があると考えられる。

7. 終わりに

関数の同値性を表現する特徴量を提案し、実験的に評価した。その結果、複雑な数式や実数を扱わない数式などは特徴量を計算することができず失敗する一方で、多項式などの数式に対しては同値な関数を効率的に検索することができ、提案手法の可能性を示すことができた。

今後の課題としては、より多くの数式に対して特徴量を計算できるようにする必要がある。このために、今回想定しなかった実数を扱わない数式などにも実数への対応付けを考えることで検索を可能とする手法を開発する。また、特徴量を完全に同値な数式だけでなく近似的な意味を持つ数式も検索できるように柔軟化する必要がある。このために、単一の値だけでなく複数の値からなる特徴ベクトルへと拡張してその空間での距離などを考える手法を検討する。

8. 謝辞

本研究は、JSPS 科学研究費補助金 2430062, 25245084 の助成を受けたものです。

参 考 文 献

[1] Shahab Kamali and Frank Wm. Tompa, “Retrieving Documents with Mathematical Document”, Proc. of ACM SIGIR, pp. 353-362, 2013.

[2] Akiko Aizawa, Michael Kohlhase and Iadh Ounis, “NTCIR-10 Math Pilot Task Overview”, Proc. of NTCIR-10, pp. 654-661, 2013.

[3] Kuo-Chung Tai, “The Tree-to-Tree Correction Problem”, Journal of the ACM, volume 26, Issue 3, pp 422-433, 1979

[4] Nikolaus Augsten, Michael Böhlen, and Johann Gamper, “The pq-gram distance between ordered labeled trees”, ACM Transactions on Database Systems, Volume 35, Issue 1, Article 4, 2010

[5] Erik D. Demaine, Shay Mozes, Benjamin Rossman, and Oren Weimann, “An Optimal Decomposition Algorithm for Tree Edit Distance”, ACM Transactions on Algorithms, Vol. 6, No. 1, Article 2, 2009

[6] Peisen Yuan, Chaofeng Sha, Xiaoling Wang, Bin Yang, Aoying Zhou, and Su Yang, “XML Structural Similarity Search Using MapReduce”, Proc. of WAIM’10, pp. 169-181, 2010

[7] http://www.w3.org/Math/

[8] http://www.wolfram.com/mathematica/