社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DBS－142 (10)
2007－ＦＩ－ 87 (10)
2007／5／31

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

橋本 英樹† 土方 嘉徳†† 西田 正吾††

大阪大学大学院基礎工学研究科システム創成専攻
〒560-8531 大阪府豊中市待兼山 1-3
E-mail: †hasimoto@nishilab.sys.es.osaka-u.ac.jp, ††{hijikata,nishida}@sys.es.osaka-u.ac.jp

概要 現在一般的に用いられているインターネット検索システムでは，数式をクエリとして検索することは出来ない．それは数式が単に数字や記号などを一列に並べたものではなく，分数や指数を始めとした構造を持って表現されるためである．そこで，MathML オブジェクトの DOM 構造を用いて転置インデックスを作成することにより数式の構造を利用した検索を実現する数式検索システムを提案する．本研究ではインターネット上で公開されている数式コンテンツを対象として調査と実験を行い，転置インデックスの構成方法を検討する．

A Survey of index formats for the search of MathML objects

Hideki HASHIMOTO†, Yoshinori HIJIKATA††, and Shogo NISHIDA††

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, JAPAN
E-mail: †hasimoto@nishilab.sys.es.osaka-u.ac.jp, ††{hijikata,nishida}@sys.es.osaka-u.ac.jp

Abstract Users cannot search information by mathematical formulas as queries in existing search engines. This is because mathematical formulas are not expressed in text in a row. Some formulas are expressed in a complex structure like fractional numbers and index numbers. We proposed a search engine for MathML objects using the structure of mathematical formulas. The system makes the inverted indices by using the DOM structure of the MathML object. We also proposed some indices for this system. We conducted an experiment to see the effectiveness of those indices by using the mathematical contents which are open to the public on the Internet.

1. はじめに

数式は，理工学・社会・経済などあらゆる分野において自然現象・社会現象・工学技術等の知識を表現する最良の方法の１つである．しかし，現在一般的に用いられているインターネット検索システムでは，数式をクエリとして検索することは出来ない．それは数式が単に数字，記号，アルファベットを一列に並べたものではなく，分数や指数を始めとした構造を持って表現されるためである．例えば xy と x^y とでは x と y というアルファベットを含むという点では同様であるが，意味としては全く異なる．従って x と y というアルファベットの出現情報だけを用いて検索を行うシステムでは意図した情報を検索することはできない．数式の検索を実現するためには，上記のような数式の構造に関する情報を利用した検索システムを開発する必要がある．

一方，インターネットなどで数式を表現するための規格として，MathML (Mathematical Markup Language) [1]が普及し始めている．MathML は XML の階層構造を使って数式の持つ構造を記述することができる．中西らは MathML を対象としてベクトル空間モデルによる類似数式検索手法を提案している[2]．しかし，特定の記号などの有無の情報を用いてベクトル空間を生成しており，数式の構造が十分に表現されていない．そのため，同じような記号が含まれていれば異なる構造の数式であっても検索結果として出力されてしまう問題があった．また，ベクトル空間モデルは，検索対象が多くなると検索速度が劇的に遅くなるという問題がある．商用レベルの大規模な検索システムへの応用にはベクトル空間モデルより転置ファイルを用いた手法が適している．

そこで，本研究では，MathML オブジェクトの DOM 構造を用いて転置インデックスを作成することにより数式の構造を利用した検索を実現する数式検索システムを提案する．また，インターネット上で公開されている数式コンテンツを対象として転置インデックスの構成方法の調査と実験を行った．

2. 数式検索システムの概要

本研究で提案する数式検索システムの構成を図１に示す．検索システムは，事前に検索対象の数式コンテ

ンツの情報を収集して転置インデックスに格納するサブシステムと，ユーザーからの検索要求があったときに転置インデックスの問い合わせを行い検索結果を出力するサブシステムの２つのサブシステムから構成されている．

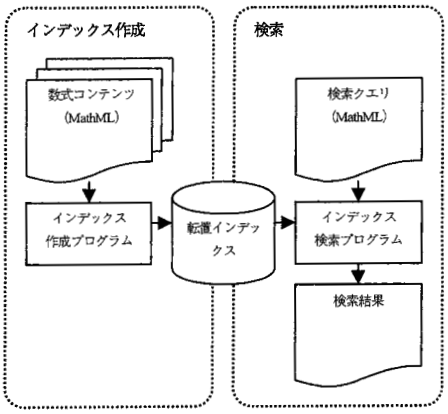

図 1　数式検索システムの構成

検索対象となる Web 文書は，図２のような MathML で記述された数式を含む HTML 文書または XML 文書である．msqrt タグは平方根，msup タグは上付き文字，mi タグは変数などの識別子，mn タグは数字，mo タグは演算子を表している．

転置インデックスへの登録時には，システムは，まず Web 文書から MathML オブジェクトの DOM 構造（図 3）を獲得し，DOM 構造のルートから葉ノードまでのパスと葉ノードの値を XPath 形式で表現する（図 4）．そして，この XPath 表記をキーとして文書 ID とともに転置インデックスに格納する．このとき，数式の内容に関係のない mrow，mstyle，semantics，annotation の各タグは，これらのタグの有無による検索漏れを防ぐため，あらかじめ XPath 表記から取り除いておく．ここで mrow は描画時における子要素の水平的な表示位置の調整を表すタグである．mstyle は文字の色など描画属性の指定に用いる．semantics と annotation は注釈のためのタグである．図 5 に転置インデックスの例を示す．図 5 の最初の例は文書 ID が 1,52,70 などの文書に分数の中に記号∂がある数式が含まれていることを表している．

検索を行うときには，システムはまず検索クエリとして与えられた MathML オブジェクトの DOM 構造のルートから葉ノードまでのパスと葉ノードの値を XPath 形式で表現する．そして，この XPath 表記をキーとして転置インデックスから該当する文書 ID のリストを取り出す．次に，リストの各文書に対しその文書に含まれる MathML オブジェクトを抽出し，この MathML オブジェクトに含まれる全てのパスが検索クエリの MathML オブジェクトに適合するかどうかを確認する．検索クエリに適合した文書のリストを検索結果として出力する．

元の数式　$\sqrt{x^2+y^2}$

MathML による記述

```
<math>
  <msqrt>
    <msup>
      <mi>x</mi>
      <mn>2</mn>
    </msup>
    <mo>+</mo>
    <msup>
      <mi>y</mi>
      <mn>2</mn>
    </msup>
  </msqrt>
</math>
```

図 2　MathML で記述された数式コンテンツの例

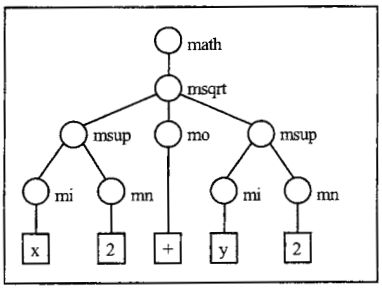

図 3　MathML オブジェクトの DOM 構造の例

```
math/msqrt/msup/mi[text()="x"]
math/msqrt/msup/mn[text()="2"]
math/msqrt/mo[text()="+"]
math/msqrt/msup/mi[text()="y"]
```

図 4　MathML オブジェクトから抽出した XPath 表記の例

XPath 表記	文書 ID リスト
/math/mfrac/mo[text()="∂"]	1,52,70,271,,,
/math/msqrt/mfrac/mn[text()="1"]	2,16,55,102,,,
/math/mfrac/mfrac/mi[text()="π"]	22,93,181,,,

図 5　転置インデックスの例

3. 予備調査

本研究で提案する XPath 表記をキーとした転置インデックスの有効性を検討するために，Wolfram 社の数学公式集サイト [3] で公開されている約 87000 件の

MathML オブジェクトをテスト対象の文書集合として調査を行った．この文書集合は大学レベル以上の専門的な数式で構成されている．

図 6 と図 7 に文書集合のそれぞれの MathML オブジェクトの DOM ツリーに含まれるパスの数とツリー構造の深さの分布を示す．パスの数はその数式に含まれている記号や数字の数を反映している．パスの深さは分数，指数，行列など数式の構造の複雑さを反映している．パスの数と深さが大きい MathML オブジェクトほどより複雑な数式を表しているといえる．図から，パス数が 30 から 60 の文書の割合が大きく．一方，100 以上のパスを持つ文書数も少なくないことがわかる．また，深い階層を持つパスの比率は比較的少ないことがわかる．１つの MathML オブジェクトに含まれるパスの数が多いことから，その数式を特徴付けるためには１つのパスの情報だけでなく複数のパスの情報を組み合わせる必要があると予測される．

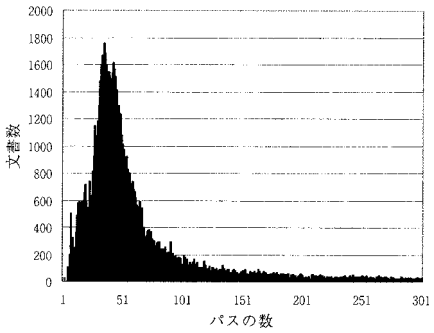

図 6　MathML の DOM ツリーに含まれるパスの数

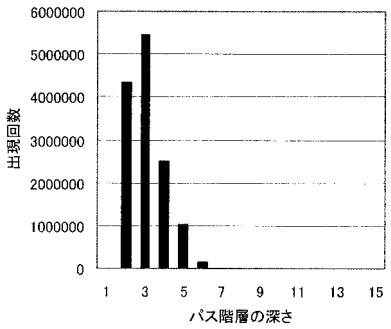

図 7　MathML の DOM ツリーに含まれるパスの深さ

文書集合に含まれる全ての MathML オブジェクトに対して，DOM 構造のルートから全ての葉ノードまでのパスと葉ノードの値の XPath 形式による表現を抽出し，それぞれの XPath 表記の出現頻度を調査した．その結果 8813 通りの XPath 表記が抽出された．この中から，出現頻度上位 10 位までの XPath 表記と出現回数を表 1 に，出現頻度 1001 位から 1010 位までの XPath 表記と出現回数を表 2 に示す．また，XPath 表記の出現頻度順位を横軸に，出現回数を縦軸にとったグラフを図 8 にしめす．

表 1　出現頻度上位 10 位までの XPath 表記と出現回数

順位	Xpath 表記	出現回数 (頻度%)
1	math/mo[text()=" ⁢"]	857574(6.369%)
2	math/mo[text()=" ("]	540280(4.012%)
3	math/mo[text()=")"]	540279(4.012%)
4	math/mfrac/mo[text()=" ⁢"]	480413(3.568%)
5	math/msup/mo[text()=" -"]	381608(2.834%)
6	math/msup/mo[text()=" ⁢"]	341416(2.535%)
7	math/mo[text()=" -"]	272379(2.023%)
8	math/mo[text()=" +"]	222359(1.651%)
9	math/mfrac/mn[text()=" 2"]	221889(1.648%)
10	math/msup/mo[text()=")"]	202749(1.506%)

表 2　出現頻度 1001 位から 1010 位までの XPath 表記と出現回数

順位	Xpath 表記	出現回数 (頻度%)
1001	math/msup/msup/mi[text()=" sin"]	320(0.00238%)
1002	math/msqrt/msup/mo[text()=" -"]	320(0.00238%)
1003	math/msub/mi[text()=" ℬ𝒫"]	319(0.00237%)
1004	math/mo[text()=" ∈"]	318(0.00236%)
1005	math/mfrac/msup/mi[text()=" sn"]	315(0.00234%)
1006	math/mfrac/msub/msub/mi[text()=" b"]	312(0.00232%)
1007	math/munderover/mo[text()=" Σ"]	311(0.00231%)
1008	math/msup/mfrac/msqrt/mn[text()=" 1"]	311(0.00231%)
1009	math/mfrac/msup/mi[text()=" β"]	310(0.00230%)
1010	math/mtable/mtr/mtd/mi[text()=" μ"]	309(0.00229%)

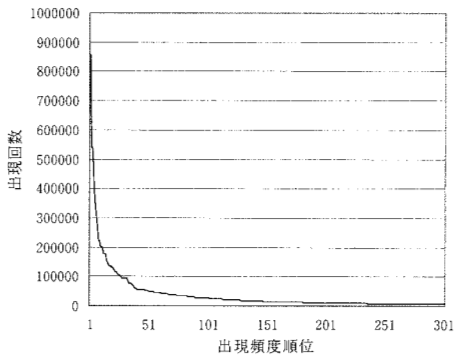

図 8　パスの出現頻度順位と出現回数との関係

例えば，出現頻度順位 1001 位の XPath 表記の出現回数は 320 回である．従ってこの 1001 位の XPath 表記を使った絞り込みでは該当する文書が 320 件出力されることになる．表 1 と表 2 を比較すると，大まかな傾向として，出現頻度 1001 位以降では，深い階層構造を持ち，元の数式を特徴付けるような数学記号を含む XPath 表記の比率が高い．一方，出現頻度 10 位以内は，浅い階層構造を持ち，InvisibleTimes（積 ab のように印字を省略した積記号），括弧，マイナス記号など，どのような数式でもよく用いられる記号を示す XPath 表記がほとんどを占めていることがわかる．図 8 からは上位 40 位までの出現回数が特に高いことがわかる．

ここで，自然言語を対象とした一般のテキスト検索システムとの比較を考えると，自然言語のキーワードによる検索では最頻出キーワードの出現頻度は全キーワードの 0.01%程度と報告されている[4]．これに対し，本実験の最頻出の XPath 表記の出現頻度は 6.37%と非常に高い頻度である．また，本実験で抽出された XPath 表記のバリエーションの総数は 8813 件となっているが，日本語辞書の単語数が数十万語レベルである[5]のと比較して大幅に小さな値となっている．これは自然言語による文書はトピックが多岐に渡っているが，数式が表す内容はそれほどの多様性を含まないことが要因の 1 つと考えられる．数式の検索効率を向上させるためには，特定の数式を特徴付けるのに適したキーを用いて転置インデックスによる索引付けを行う必要がある．例えば，一般のテキスト検索システムでは出現頻度上位のキーワードを索引付けの対象から除外することにより検索性能が向上することが報告されている[6]．本研究でも出現頻度が高く数式を特徴付けるためのキーとしては適当でない XPath 表記をインデックス作成時に除外することによって検索効率の向上が期待できる．

4. 実験

予備調査により，MathML オブジェクトの DOM 構造のルートから全ての葉ノードまでのパスのうち 1 つのパスを選び，それをキーとすることで元の数式をある程度特徴付けできることが示された．しかし，1 つのパスだけを使ったインデックスでは検索結果の絞り込みが不十分であることも判明した．そこで，MathML オブジェクトに含まれる全てのパスの中から 2 つのパスを選び，その 2 つのパスの XPath 表記の組み合わせをキーとして転置インデックスによる索引付けを行う数式検索システムを試作し，実際に検索を行って検索能力を評価した．

実験には予備調査で使用した約 87000 件の数式を用いた．このそれぞれの数式に対し，MathML オブジェクトの DOM 構造のルートから全ての葉ノードまでのパスを抽出し，その中から 1 番はじめのパスと最も階層構造が深いパスを選択した．これら 2 つのパスと葉ノードの値を組み合わせたキーを作成し転置インデックスに格納した．実験では 7398 通りのキーが作成された．1 番はじめのパスの情報を転置インデックスに含むことにより，対象となる数式がどのような書き出しで始まるかを特徴付けることが出来る．また，階層構造の一番深いパスは，対象となる数式の最も特徴的な部分を指している可能性が高いと考えられる．作成した転置ファイルの例を図 9 にしめす．図 9 では 2 つの XPath 表記をカンマでつないだ表記をインデックスのキーとしている．

2 つの XPath 表記を組み合わせたキー	文書 ID
/math/mi[text()="cot"],/math/mover/mi[text()="∞"]	47206, 47208, 47276,,,
/math/munderover/mo[text()="Σ"],/math/msup/mfrac/mn[text()="1"]	850, 1211, 1379,,,
/math/mo[text()="∫"],/math/mfrac/msqrt/msup/mi[text()="sin"]	18679, 39824, 39839,,,

図 9　実験で作成した転置インデックスの例

文書集合の中から 1 つの数式を選び，この数式をクエリとして転置インデックスから該当する数式を抜き出した結果を図 10 および図 11 に示す．図 10 の例では，クエリとした数式から一番初めのパスとして数式が tan で始まることを表す XPath 表記が抽出されている．また，最も階層の深いパスとして分数の中にさらに分数がありその中に虚数を表す i が含まれていることを表す XPath 表記が抽出されている．検索結果にはこれら 2 つのパスを含む数式が 4 件出力されている．この 4 件の中から厳密にクエリと同一の数式だけを検索結果として出力する場合は，この 4 件に対してそれぞれの数式の MathML オブジェクトに含まれる全てのパスが検索クエリの MathML オブジェクトに含まれるパスに適合するかどうかを確認する．全てのパスが適合するものだけを検索結果として出力する．

図 11 は π ではじまり sin の累乗が分数の中にある数式を検索した例である．検索結果には sin の 2 乗，3 乗などを用いた公式とともに，これらを sin の n 乗に一般化した公式も含まれている．

上記の例では，検索クエリと同一の数式だけでなく，検索クエリと類似した構造と意味を持つ数式も出力されている．このような類似数式検索は，検索クエリとした元の数式に関連する数式の知識を得ること目的とする検索に適している．

5. まとめと今後の課題

本研究では，テキスト検索で広く用いられている転置ファイルによる検索手法を，MathML を検索対象とした数式検索システムに適用するために，MathML の DOM 構造のルートから葉ノードまでのパスを用いた

転置インデックスの検討と，簡単な検索システムの実装による検索能力の評価を行った．

今後の課題としては，基本的な検索機能の強化として，XPath 表記の出現頻度に応じた重み付け，検索結果のランキング表示，2 つ以上の MathML オブジェクトをクエリとした AND 検索などが挙げられる．また，本研究は商用レベルの数式検索の実用化を目標としており，これに向けて数式検索モジュールを含む SDK とデモンストレーションの公開を予定している．

謝辞

本研究は，NEDO 産業技術研究助成事業（プロジェクト番号 06A14501d）の助成を頂きました．

参考文献

1) W3C : W3C Math Home, http://www.w3.org/math/ .
2) 中西崇文,岸本貞弥,村方衛,大塚透,櫻井鉄也,北川高嗣: 数式データを対象とした複合連想検索システムの実現，日本データベース学会 Letters, Vol.4, No.1, 2005.
3) Wolfram Research Inc. : The Wolfram Functions Site, http://functions.wolfram.com .
4) 情報通信研究機構: EDR Home Page, http://www2.nict.go.jp/r/r312/EDR/J_index.html .
5) 河野浩之,北村泰彦,山田誠二,高橋 克巳: 情報検索とエージェント，東京電機大学出版局 .
6) 北研二,津田和彦,獅々堀正幹: 情報検索アルゴリズム，共立出版.

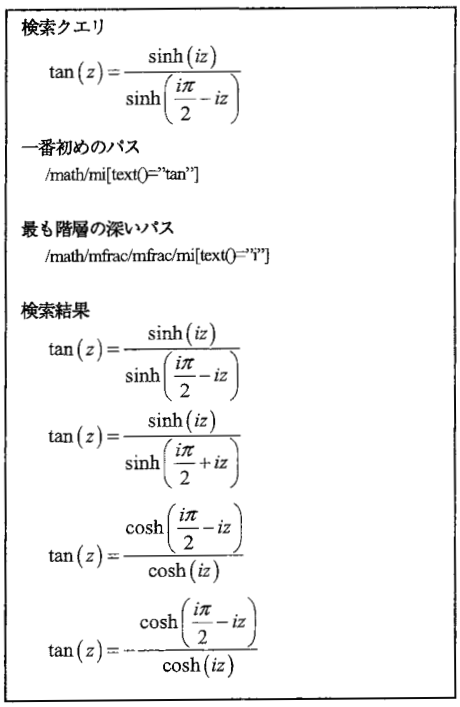

図 10 検索例

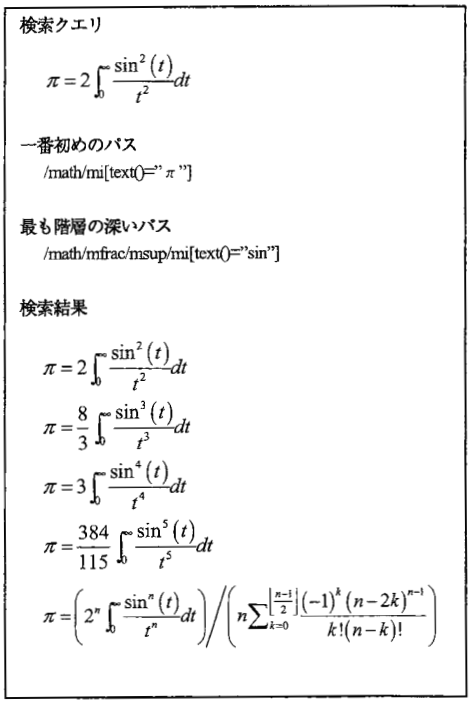

図 11 検索例