1F1-3

MathMLを用いた数式検索

Search of Mathematical Formulas using MathML

小田切 健一[*1] 村田 剛志[*1]
Kenichi Otagiri Tsuyoshi Murata

[*1]東京工業大学 大学院 情報理工学研究科 計算工学専攻
Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology

It is not easy to search mathematical formulas properly by existing search engines. This is because existing search engines are based on text search. Mathematical formulas have their structures, and the structures can be used for improving search. This paper proposes a mathematical formulas search engine that has abilities of searching foumulas exactly with complex conditions by using MathML DOM structure.

1. 序論

数式は理工学・社会・経済などのあらゆる分野で使われ、各分野の文書内で重要な要素の記述に用いられている。そのため、数式を対象とした検索が実現できれば非常に有益だと考えられる。しかし、現在のテキスト検索を利用した検索エンジンでは数式を適切に検索することが困難である。これは数式が構造を持っており、テキスト検索で利用される「単語の有無」という条件では特性を捉えきれないためである。例えば「sinを含む積分」という数式を検索したい場合、テキスト検索では「sin AND ∫」というクエリを投げるが、「$\sin x \times \int_a^b xdx$」のような数式もこの条件に合致してしまう。この様に、一般的なテキスト検索用検索エンジンで数式を適切に検索するのは難しく、数式検索に適した検索システムを開発する必要がある。

近年、数式をXMLを用いて表現するMathML (Mathematical Markup Language)[MathML] という規格が普及し始めている。MathMLは数式をXMLの階層構造を使用して前置式で表現しているため、計算機で扱いやすいという特長がある。現在、対応ソフトウェアが増えつつあり、Webや異種ソフトウェア間で数式をやり取りをする際の標準規格として期待されている。このMathMLで表現された数式を検索するシステムとしては、岸本らの類似数式検索 [岸本 2003] がある。しかしこれは、クエリとして「数式」を入力し、クエリと類似した数式を出力するシステムである。今回の目的である「sinを含む積分」といった条件による検索には利用できない。もう一つの関連研究に、橋本らによるMathMLのためのインデックスの調査 [橋本 2007] がある。この研究では、MathMLに含まれるXPathを用いた転置インデックスを作成することでMathMLを高速に検索可能にしている。しかし、入力された数式と完全一致する物を検索するために設計されたインデックスであるので、直接今回のような条件による検索に利用することはできない。この様に、MathMLを用いた数式検索の研究では、「数式」を入力しそれに類似あるいは完全一致する物を検索するという手法が用いられてきた。本論文では問い合わせのために「数式」そのものではなく、数式の部分の形や関数の引数といった条件を利用する。これにより、従来不可能であった「sinを含む積分」「引数が〜（である／を含む）cos」といった条件による検索が可能になる。さらに、問い合わせに数式そのものを用いるよりも記述量が少なくなるという利点もある。

連絡先: 小田切 健一，東京工業大学 大学院 情報理工学研究科 計算工学専攻, 〒 152-8552 東京都目黒区大岡山 2-12-1 W8-59，otagiri@de.cs.titech.ac.jp

2. 実装した検索システムの概要

2.1 実装した問い合わせ言語

今回のシステムでは「sinを含む積分」といった問い合わせに対する検索を正確に実現にするのが目的である。しかし、テキスト検索のような単語の有無を指定する問い合わせでは実現できない。そこで、構造を持った条件を記述することができる問い合わせ言語を設計した。基本的には問い合わせたい数式を「+,-,*,/, 関数」等を用いて記述するが、任意の数値や任意の合成式を表現する「N」「X」などの特殊記号を導入したため、完全一致ではなく柔軟性を持った検索が可能である。また、関数の引数に含むものを指定するといった機能もある。(表1参照) このため、「$\&sin(a*b)$」(変数*変数が引数であるsin) といったような数式に近いレベルから、「$\&int\{\&sin\{\}, !\&cos\{\}\}$」(sinを含むがcosを含まない積分) のような抽象的なレベルまで、広い範囲の問い合わせが可能である。(問い合わせ例は表2参照)

表 1: 実装した問い合わせ言語の主な書式

形	意味
&xxx(yyy)	yyy を引数に持つ関数 xxx
&xxx{yyy}	yyy を引数に含む関数 xxx
&xxx{}	関数 xxx（引数は問わない）
0 - 9	数値
N	任意の数値
X	合成式（単一の変数以外）
+ - * / ^	演算子
その他の英字	変数

2.2 問い合わせと数式のマッチング

今回設計した問い合わせ言語は単純な単語の羅列でないため、構文解析をする必要がある。構文解析された問い合わせは木構造の形でメモリ上に置かれる。この木構造は内部で用いられるだけなので、処理が正しく行えればどのような形でも問題ない。今回のシステムでは、マッチング部分を簡潔にし、マッチングの説明を分かりやすくするために、構文解析時に問い合わせをMathMLを一部拡張した形の木構造に変換する。なお一部拡張とは、MathMLには存在せず問い合わせ言語にのみ存在する「〜を含む／含まない（$\langle contains \rangle, \langle in\_unit \rangle, \langle notin\_unit \rangle$）」

表 2: 問い合わせ例

問い合わせ例	意味
$a * b$	任意の変数同士の掛け算
$a\hat{}N$	変数を任意の数値で累乗
$\&sin(a)$	引数が単一の変数である sin
$\&sin(X)$	引数が単一の変数でない sin
$\&int(\&sin(a))$	sin(単一の変数) のみの積分 （積分内に sin 以外を含まない）
$\&int\{\&sin(a)\}$	sin(変数一つ) を含む積分 （積分内に sin 以外を含んでも良い）
$\&int\{\&sin\{\}\}$	sin（引数の条件なし）を含む積分
$\&int\{\&sin\{\},\&cos\{\}\}$	sin と cos を含む積分
$\&int\{\&sin\{\},!\&cos\{\}\}$	sin を含むが cos を含まない積分

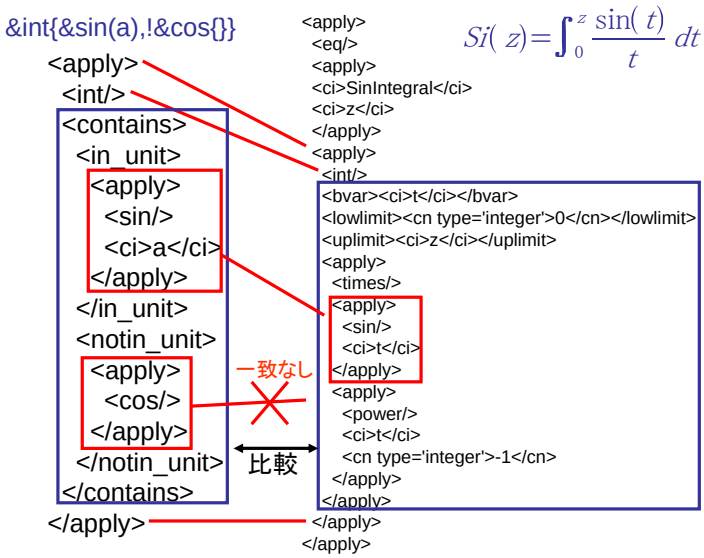

図 1: マッチング例

「合成式（$\langle applyX \rangle$）」といった表現である。このような特殊な動作も含むが、基本的にマッチング時は二つの同じ形の木構造をたどっていくだけであるので比較的簡単に実装できる。図 1 に変換後の問い合わせと数式のマッチング例を示す。なお、マッチング時には変数同士は変数名が違っていてもマッチングがなされるため、表記の揺れに対応することができる。

2.3 実装したランキング方法

今回の検索システムは数式の条件を入力し、その条件を満たす数式を出力する。特定の数式を検索したい場合（例えば、参考書に書かれている数式と全く同じ数式を用いた文書を検索したい場合）は、条件を順次追加して絞り込んでいけばよい。しかし、「sin を含む積分の例を探したい」といった漠然とした探し方の場合、検索結果は膨大な数となり、なんらかのランキングが必要となる。特定の数式を探しているのでなければ、「簡単な数式」「指定した条件が分かりやすく現れている数式」といった条件を満たす事が求められると考え、この二つの特性に着目した 3 つのランキングと比較用の 1 つのランキングの実装・評価を行った。

2.3.1 式の大きさ

「簡単な数式である」という特性を「短い数式である」と読み替えて実装をしたランキングである。MathML 内のノード数が小さい順に順位付けをする。

2.3.2 一致部分の大きさ

「簡単な数式である」という特性を「条件に一致した部分が単純（短い）である」と読み替えて実装を行ったランキングである。条件に一致した MathML の部分木のノード数が小さい順に順位付けをする。なお、「$\&\sin(a)$」のように大きさが固定の条件では一致する部分の大きさが毎回同じため、このランキングの意味はない。

2.3.3 一致部分の大きさの割合

「指定した条件が分かりやすく現れている」という特性を「指定した条件を満たす部分が数式の大きな割合を占めている」と読み替えて実装したランキングである。条件と一致した部分木のノード数と数式全体のノード数の比率で順位付けを行う。

2.3.4 TF-IDF

これは、上記の二つの特性のどちらも用いていないランキングである。条件中に含まれる単語の TF-IDF 値と数式中に含まれる単語の TF-IDF 値を計算し、値が近い順に順位付けを行う。今回想定した特性を考慮せずに、単純にテキスト検索のランキング方法を利用している。今回実装した三つのランキングとの比較のために実装した。

なお、単語ごとの TF-IDF の計算式は以下の通りである。

n_i は単語 i の出現回数
$|D|$ は数式の総数
$|d : d \ni t_i|$ は単語 i が出現する数式の総数

$$tf\text{-}idf_i = tf_i \times idf_i \tag{1}$$

$$tf_i = \frac{n_i}{\Sigma_k n_k} \tag{2}$$

$$idf_i = \log \frac{|D|}{|\{d : d \ni t_i\}|} \tag{3}$$

3. 実験

3.1 概要

上記の問い合わせ言語およびランキングを実装し、実際に数式検索システムで実験を行った。システム概観を図 2 に示す。左のテキストボックスに問い合わせを入力し、右のセレクトボックスよりランキング方法を選択する。実験で使用した環境は表 3 の通りである。なお、数式データは「The Wolfram Functions Site」[Wolfram] から収集した約 8000 個の MathML を用いた。

表 3: 実験環境

ブラウザ	Firefox 2.0.0.11
言語	Java SE 1.6.0_01
XML パーサ	Java 標準ライブラリ JAXP
サーブレット	Tomcat 6.0.10

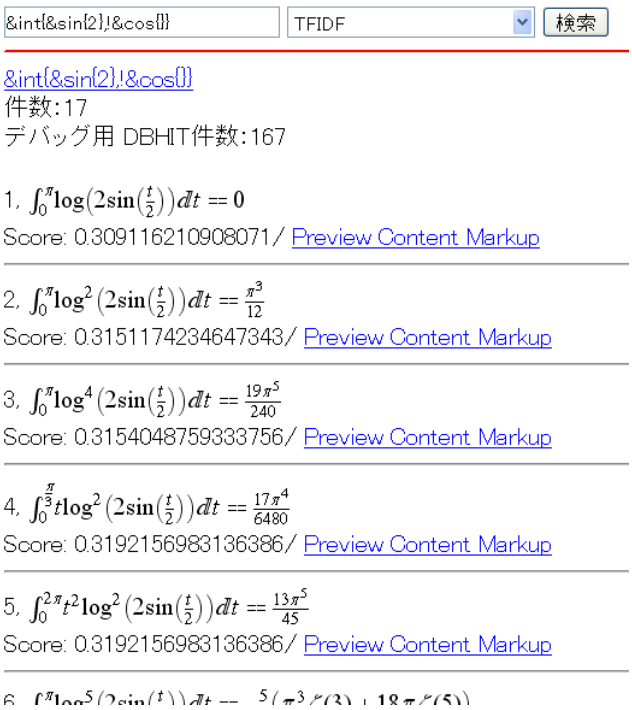

図 2: 検索画面

3.2 検索結果

いくつかの検索を行い、その検索結果が条件を満たしていることを確認した。

表 4: 「a*b」の検索例

1	$\mu(mn) == \mu(n)\mu(m)/;gcd(m,n) == 1$
2	$\int Ai'(az)Bi'(az)\boldsymbol{d}z == \frac{1}{3a}(Ai(az)(Bi'(az) - a^2z^2...$
3	$\int z^2Ai(az)Bi(az)\boldsymbol{d}z == \frac{1}{5a^3}(Ai(az)((a^3z^3 - 1)...$

表 4 は、「$a*b$」という問い合わせの検索結果である。一致部分は上から「mn」「az」「az」である。変数名が違っても正常にマッチングが行われている。

表 5: 「&cos(a)」の検索例

1	$\cos(-z) == \cos(z)$
2	$\cos(\bar{z}) == \bar{\cos(z)}$
3	$\cos^3(z) == \frac{1}{4}(\cos(3z) + 3\cos(z))$

表 5 は、「$\&cos(a)$」という問い合わせの検索結果である。全て「$cos(z)$」という部分に一致している。

表 6 は、「$\&cos\{\}\hat{}N$」という問い合わせの検索結果である。条件なし cos の任意の累乗という検索である。1 番目では「$cos(z)$ の三乗」、2 番目では「$cos(a)$ の二乗」、3 番目では「$cos(z)$ の二乗」に一致している。

表 6: 「&cos{} ^ N」の検索例

1	$\cos^3(z) == \frac{1}{4}(\cos(3z) + 3\cos(z))$
2	$\cos^2(a) - \sin^2(b) == \cos(a-b)\cos(a+b)$
3	$\cos(2z) == \cos^2(z) - \sin^2(z)$

表 7: 「&cos(a+b)」の検索例

1	$\cos^2(a) - \sin^2(b) == \cos(a-b)\cos(a+b)$
2	$\cos(a+b) == \cos(a)\cos(b) - \sin(a)\sin(b)$
3	$\cos(a+z)Y_\nu(z) == \frac{1}{\sqrt{2}}G^{2,2}_{3,5}...$

表 7 は、「$\&cos(a+b)$」という問い合わせの検索結果である。引数の条件が複雑になっているが、正常に「$cos(a+b)$」「$cos(a+b)$」「$cos(a+z)$」といったものと一致している。ここでも変数名が違っていても正常に処理されていることが確認できる。

表 8: 「&int(&sin(X))」の検索例

1	$\cos(z) == 1 - z\int_0^1 \sin(zt)\boldsymbol{d}t$

表 8 は、「$\&int(\&sin(X))$」という問い合わせの検索結果である。「合成式を引数にもつ sin の積分」という検索内容であり、その通りの「$\int_0^1 \sin(zt)dt$」という数式がヒットしている。(これは非常に複雑な条件設定であるが、条件を満たさない誤った数式は現れていない。このように、どのような条件に対しても誤判定を出さないのがテキストではなく MathML を利用した本システムの利点である)

表 9: 「&int{&sin(X)}」の検索例

1	$\int_0^{\pi}\log(2\sin(\frac{t}{2}))\boldsymbol{d}t == 0$
2	$\cos(z) == 1 - z\int_0^1 \sin(zt)\boldsymbol{d}t$
3	$\gamma == \log(2) - \pi\int_0^{\frac{1}{2}}\int_0^1 \tan(\frac{\pi t}{2})(\frac{\sin(\pi tu)}{\sin(\pi u)} - t)\boldsymbol{d}t\boldsymbol{d}u$

表 9 は、「$\&int\{\&sin(X)\}$」という問い合わせの検索結果である。「合成式を引数にもつ sin を"含む"積分」という検索で、先ほどより条件がゆるくなっている。結果を見れば、「$\frac{t}{2}$」「zt」「πtu」といった合成式を引数にもつ sin を含んだ積分が含まれている。これも問い合わせどおりに動作していることが確認できる。

3.3 ランキングの比較

今回のランキング比較に用いた問い合わせは「$\&sin(X)$(合成式を引数にもつ sin)」である。各ランキングの上位三つを示す。

表 10: 式の大きさによるランキング

1	$\sin(\widetilde{\infty}) ==$ ¿（未定義）
2	$\sin(2\pi) == 0$
3	$\sin(am(z\|m)) == sn(z\|m)$

表 11: 一致部分の大きさによるランキング

1	$Im(z) == \frac{z\sin(Arg(z))}{sgn(z)}/; z \neq 0$
2	$\|z\| == z(\cos(Arg(z) - i\sin(Arg(z)))$
3	$\|z\| == \frac{Im(z)}{\sin(Arg(z))}$

表 12: 一致部分の大きさの割合によるランキング

1	$\cosh(z) == \sin(\frac{\pi}{2} - zi)$
2	$\sin(\frac{3\pi}{2}) == -1$
3	$\cosh(z) == \sin(iz + \frac{\pi}{2})$

表 13: TF-IDF によるランキング

1	$\sin(\bar{z}) == \bar{\sin}(z)$
2	$\sin(-z) == -\sin(z)$
3	$\sin(2\pi) == 0$

3.4 ランキングの考察

一般的に良い測度と思われるのは「式の大きさ」である。単純に小さい数式が出力されるが、一見して見やすい数式になっている。次によいと思われるのが「一致部分の大きさの割合」である。右式や左式全体が条件を満たしているものが上位に出やすい。「一致部分の大きさ」では、条件に一致する部分の大きさが小さいものが上位に順位付けされるため、数式全体ではなくその部分に注目するときは有用だと考えられる。「TF-IDF」はテキスト検索でよく用いられるランキング方法をそのまま数式検索に適用したものである。問い合わせ中の単語の頻度と数式中の単語の頻度が近いものが上位に順位付けされるため、問い合わせの長さと数式の長さの比によって順位付けの傾向が変わってしまう。また、そもそも「TF-IDF の値が近いから問い合わせと似ている」という想定が数式検索においては必ずしも正しいとは考えられない検索結果が得られた。例として「$\&int\{\&cosh\{\}\}$（cosh を含む積分）」を検索した場合、
$\int cosh(4cz)coth(cz)\boldsymbol{d}z \ = \ \frac{4cosh(2cz)+cosh(4cz)+4log(sinh(cz))}{4c}$
の方が $sinh(z) == z\int_0^1 cosh(zt)\boldsymbol{d}t$ より上位にランキングされる。しかし、実際は後者の方が問い合わせにより近いと考えられる。テキスト検索のためのランキングをそのまま数式検索に適用した場合、数式とテキストの性質の相違のため、ランキングの性質が発揮されない場合があると考えられる。

4. 結論

本研究では、テキスト検索・類似数式検索といった手法では不可能だった複雑な条件を使用した数式検索を可能にした。今回設計したシステムにより、変数名が異なった数式も同様に検索し、関数の引数や積分対象を指定した検索が可能になった。また、数式検索においては従来のテキスト検索で用いられた TF-IDF のような尺度をそのままランキングに適用するのは不適という考察が得られた。

今後の課題としては、OR 検索と高速化が挙げられる。現段階では問い合わせ言語には OR 検索の機能がないため「sin か cos を含む積分」という検索ができない。また、現状では逐次処理をしていくため検索速度がとても遅く、インデックスの利用などによる改善が求められるだろう。

参考文献

[MathML] "W3C Math Home", W3C, http://www.w3.org/Math/ .

[岸本 2003] 岸本貞弥, 中西崇文, 櫻井鉄也, 北川高嗣, 栃木敏子 : MathML を用いた類似数式検索方式の実現, 第 14 回データ工学ワークショップ（DEWS2003) 論文集 (2003).

[橋本 2007] 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).

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