数式検索システムの検索機能拡張の試み
Attempts to expand the search functions of Math IR System

櫻井 翼[*1], 宮崎 佳典[*2], 中村 泰之[*3], 田中 省作[*4], 新谷 誠[*2]
Tsubasa SAKURAI [*1], Yoshinori MIYAZAKI[*2], Yasuyuki NAKAMURA [*3], Shosaku TANAKA[*4], Makoto ARAYA[*2]
[*1] 静岡大学 情報学部
[*1]Faculty of Informatics, Shizuoka University
[*2] 静岡大学学術院 情報学領域
[*2]College of Informatics, Shizuoka University
[*3] 名古屋大学 大学院情報学研究科
[*3]Graduate School of Informatics, Nagoya University
[*4] 立命館大学 文学部
[*4]College of Letters, Ritsumeikan University
Email: sakurai.tsubasa.19@shizuoka.ac.jp

あらまし: 我々は STEM 教育分野における自主学習を補助する目的で MathML Presentation Markup ベースの数式検索システムを開発している．その開発過程で，数式の読込処理や正規表現を用いたマッチング処理時に予期しない検索結果を出力する不具合を発見した．本研究ではそれらの原因を解明し改良を重ねることで，同システムのさらなる機能拡張として階層的な構造を含む数式にも対応することを目指す．
キーワード：数式検索システム，MathML Presentation Markup，正規化，正規表現

1. はじめに

近年 STEM 教育として数理分野への取り組みが盛んになっている．この分野を深く理解するための自主学習用ツールとして，当研究室では MathML Presentation Markup を用いた，正規表現を用いた数式用検索システムを開発している．(1)

このシステムを拡張する先行研究では，変形依拠公式提示機能による公式の抽出(2)等によって基本的な数式の分析に着目していた．しかし，これらの研究では考慮されない階層的な構造を含む数式を検索すると検索結果に不具合が生じることがわかった．

本研究では，この不具合の解消によって，システムのより多彩な数式への対応を目指す．この改善により，本システムが応用的な問題を学習する際にも補助が可能になることを目標とする．

2. 数式検索システムの利用方法

本システムは Firefox 上で動作する MathML Presentation Markup（以下 MathML）で記述された数式を対象に検索を行うものである．このシステムは図 1 に示したインタフェースにて操作を行う．

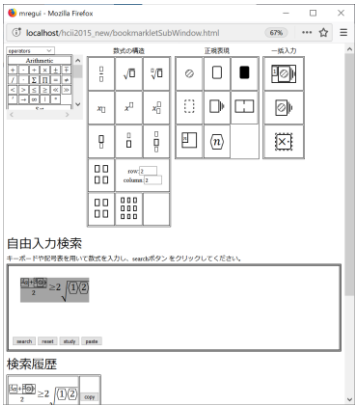

図 1 数式検索システムの入力インタフェース

数式検索を行うには，まずインタフェース下部の「自由入力検索」の枠内に，キーボードにて英数字や記号の入力を行う．また入力の補助機能として，インタフェース上部のパレットから数式上の記号（左）や数式構造（中央）及び正規表現（右）をクリックで追加できる．利用可能な正規表現の例を表 1 に示す．

表 1 本システムで利用できる正規表現の例

	任意の 1 文字 パターン: .		入力文字列の 1 回以上の繰り返し パターン: (…)+
	内部入力文字のいずれか パターン: […]	n	後方参照用ラベル（n） パターン: (?<cn>…)
	内部入力文字以外 パターン: [^…]	⟨n⟩	n（番）を後方参照 パターン:¥k<cn>

キーボードの入力とパレットを組み合わせて検索クエリを入力し終えた後，枠内下部の「search」ボタンを押すことでマッチングされ，web ページ内の該当する数式の部分に黄色いマーキングが施される．

このシステムには，基本となる上記の数式検索システムの他に，応用機能として公式検索，変形依拠公式提示機能も備えられている．

3. 数式検索の手法と特徴

当システムでは数式と検索クエリを正規表現として表した上でマッチングを行う．その前段階として検索対象の数式の構造管理と正規化及び対象の数式と検索クエリの正規表現への変換の処理を行う．

このシステムの表記として利用する MathML はタグにより数式上の文字の持つ意味や構造を示す．例えば，タグ<mi>は識別子，<mo>は演算子，<mn>は数を表す．また MathML は入れ子による階層構造を持つ．当システムでは Web ページから数式を読み

取った際，まずタグによる入れ子構造を木構造に変換して階層を管理する．この時見た目が同一の式でも，タグの区切りや種類が違うことで表記揺れが生じることがある．当システムでは検索の前に表記揺れを 1 つの記法に統一させる処理（正規化）を行い，検索に用いる MathML を統一した表記に変換する．

本システムでは数式の検索に正規表現エンジン Onigmo を用いたマッチングを行う．このため，木構造の数式とクエリが出揃ったところで，対象の数式の MathML と検索クエリを Onigmo パターンに変換する．ここで，数式構造を表すタグが文字として識別子と同様にマッチングされることを防ぐため，以下のようなパターンに構造を変換する：

/構造節名/{統合節 1/}[/{統合節 2/}[/{統合節 3/}]

この内パターン中の “[…]”内部はその節が存在する構造にのみ記述される．また構造節名は種類によって表 2 に示すような形で変換される．

表 2 構造節名変換の対応例

数式構造の構造節名	変換先
sup(上付き文字)	:
sqrt(平方根)	:::
frac(分数)	::::
mstyle(マーキング)	::::::::::

正規表現に変換された両者はマッチングが行われ，マッチした部分を表 2 の mstyle 構造で囲んでおく．その後マッチング後の数式を MathML に逆変換して表示を置き換える．マッチング箇所の構造は <mstyle>の設定で黄色背景を付けマーキングする．

4. 括弧による入れ子構造処理の正常化

数式には括弧が入れ子構造で複数階層に用いられるものがある．従来のシステムでは，そのような数式を検索した際に括弧の位置が変化する不具合が発生していた．以下に検索する数式と不具合の起きた結果の例を図 2 に示す．

$$(x(n(bc(sq)uo)rt)xyz)$$

$$(x(n(bc(sq)),u,o))rtxyz$$

図 2 数式の例と検索クエリが “(sq)”の際の結果

これは当システムの MathML が括弧の表現に利用しているタグ<mfenced>の構造の区切りの動作によるものである．このタグは内部構造を<mrow>にて囲む．これが<mstyle>により中断されるとき，終端に直後の</mrow>や</mfenced>を付けるようになっている．これらのタグが異常な位置から引用されることで，括弧の構造がずれるようになっていた．

これを解決するために，括弧の表現として正規化するタグを<mfenced>から<mi>に変更した．<mi>の<mrow>を用いない単純な文字の構造では括弧を入れ子構造で用いた数式でも，数式の表示を維持したまま正常なマッチングが実現できた．また公式検索等の本システムの応用機能においても，修正後も修正前と同様の動作ができたことを確認した．

5. 内部入力文字処理の修正

2 節で示した「内部入力文字のいずれか」，「内部入力文字以外」の正規表現は，数式的意味が同一な文字（以外）を一括で検索する用途のものである．従来のシステムでは，これらを用いた検索クエリにて検索を行うと，結果の表記が Onigmo パターンになることがある不具合が発生していた．以下に検索する数式と不具合の起きた結果の例を図 3 に示す．

$$\cos\left(\pi+\frac{\pi}{6}\right)$$

$$cos(\pi+/:::/\{\pi/\}/\{6/\})$$

図 3 数式の例と検索クエリが 6 の際の結果

これは内部入力文字（以外）の処理範囲に 3 節で示したような “/構造節名”, “/{”, “/}”といった数式構造用の制御文字列が含まれるためである．構造を示すにはこれらの文字列が途中で分断されずに連続する必要がある．しかし，内部入力文字（以外）は 1 文字単位でマッチングやマーキングを行うため文字列が分断される．これを防ぐため，検索クエリ内の内部入力文字の正規表現 “[…]” に制御文字列を対象から取り除く処理を加える必要がある．対処法として「否定先読み」“(?!…)”と「否定後読み」“(?<!…)” のパターンを従来の正規表現の前後に挿入した．先読みや後読みにより処理範囲に条件を足すことで数式構造のパターンを 1 文字ずつ除外する．これにより図 3 と同様のマッチングでも，数式を維持しながら，かつ内部入力文字以外として正常な動作を実現した．しかし，例外的に数式上の括弧として “{}” を用いる場合は処理範囲の除外が正常ではなくなる．これは制御文字列との区別のため正規表現上で“¥¥{”のように表記することで一部が先読みや後読みにて除外されるためである．よってこの正規表現は不完全であり，今後の改善が課題のひとつとなる．

6. まとめ・今後の展望

MathML の正規化や内部入力文字処理の修正により従来よりも複雑な数式への正常な対応が実現できた．しかし，修正には不完全なものがある他，上記のもの以外にも不具合は幾つか発見しているため，今後はこれらの修正をしつつ，それを機にした新たな機能の拡張を目指していきたい．

参考文献

(1) 渡部 孝幸, 宮崎 佳典, 正規表現を用いた数式検索手法の提案, 情報処理学会論文誌, Vol.56, No.5, pp.1417-1427 (2015)

(2) 脇 弘太, 宮崎 佳典, 代数的変形に対応した変形依拠公式提示ツールの開発, 情報処理学会第 82 回全国大会, pp. (4)-747-748 (2020)