LK-002

MathMLで記述された数式コンテンツ検索システムの構築

Development of Search Engine for MathML-based Mathematical Descriptions

宮崎佳典* 井口義秀†

Yoshinori MIYAZAKI Yoshihide IGUCHI

1 はじめに

Web 上で数式を扱うケースは数多く存在し、例えば e-Learning 用数学コンテンツ、インターネット掲示板におけるやりとり、数学ブログなどが挙げられる。さらに、数学の分野に限らず、数式は物理学、経済学を始めとした様々な分野で頻繁に使用されている。

しかしながら、表示、検索なども含めて、これらを満足に実現する方法は存在せず、一次元方向の表現しかできないテキストによる代替や、あるいは数式表現を画像に変換してからコンテンツ内に埋め込む、などの措置が取られている（その他、プラグインを用いるケースなどもある）。後述するように、それらには見栄えや面倒くささなどの複数の問題点が存在し、ユーザーは多くの妥協を強いられた上で仕方なく数学コンテンツを作成せざるを得ないのが現状である。

より深刻に著者が危惧することは、このような方法で作成された数式コンテンツは、検索が困難であるという点である。データに対して検索エンジンが正しく検索結果を反映できないと言う環境では、価値のあるコンテンツはインターネットの大海原に埋没したも同然で、利用価値は大幅に低下してしまう。

一方で、Web 上での数式表現を目的に開発されたマークアップ言語 MathML がある。タグを用いての意味づけにより検索も可能であり、また MathML を実装するブラウザに限るが綺麗な数式の出力を実現している。

そこで著者らは MathML によって記述された数式データの検索システム構築を研究の目的としている。これが完成すれば、例えばある定理の名前は忘れてしまっても、定理の特徴（例：行列の分解定理で右辺が行列の積の形になっている、など）から対象ページを絞り込むことが可能となる。つまりある種のあいまい検索が備わり、ユーザーへの大きなサービスに繋がる。

Web 上で、数学関連のコンテンツがより充実すれば、学習者にとって身近なものとなり、昨今問題視されている数学離れに対しても一助となるのでは、という副次的効果もねらいの一つとして挙げられる。

2 Web 上数式表現の現状

Web 上でなんらかの数式を書きたい場合、次の 2 種類の方法で実現せざるを得ない現状がある。

1 つ目は、プレーンテキストを使って、数式を擬似的に表現する方法である。しかしながら、プレーンテキストの場合、左から右に進むような表記法しかできない。一方で、数式には 2 次元的な表現方法が頻繁に出てくる。左右だけでなく、上下の動きであり、例えば連分数などがこれに当たる。

例として、$1-\cfrac{1}{1-\cfrac{1}{1-\cfrac{1}{1-x}}}$ をプレーンテキストで無理矢理表現すると 1-1/(1-1/(1-1/(1-x))) となる。これを一瞥してもとの数式を思い浮かべることはなかなか難しい。逆にこの数式を間違えずに編集するのは容易ではないし、また検索も困難である。

残念ながら、HTML には文字や記号に対して、その色や大きさ、配置といったレイアウトを決めたり、図やリンクの情報を付加することができる一方で、数式を表現するためのレイアウトをほとんど備えてはいない（上付 \sup や下付 \sub 命令がある程度）。

もう 1 つは、数式表現を画像ファイルとして変換したものを貼り付けることで代替する方法である。この方法であれば、数式の見た目の美しさは損なわれない。しかし、画像ファイルによる数式は、周りのテキストと大きさや濃さが異なる場合があり、レイアウトが崩れる可能性がある。ページ全体の拡大・縮小などにも対応できない。また、基本的には検索処理も行えない（Alt タグなどで数式のソースを入れておくことは一方法であるが面倒であることこの上ない）、修正を加える

*静岡大学情報学部
†静岡大学情報学研究科 M1

にはもとデータがなければならないなどの不便も多い。

以上の点で、コンテンツ作成者が容易に数式を Web 上で取り扱っているとは言い難く、製作者ならびに利用者側双方で大きな妥協を余儀なくされている現状は打破しなければならない。

3 MathML とその利用

MathML(Mathematical Markup Language) は Web 上で数式を表現でき、その内容を符号化することを目的に作成されたマークアップ言語である。W3C(World Wide Web Consortium) によって、1999 年 7 月に MathML 規格バージョン 1.01、2003 年 10 月にバージョン 2.0 が勧告され、現在利用されている MathML の多くはバージョン 2.0 である。さらに、2006 年 6 月 28 日には、より多言語へのサポートやレイアウトの精微性の向上、諸技術の向上への対応などを目的としたバージョン 3.0 の規格の制定ためのワーキング・グループが発足している [1]。

MathML は確かに Internet Explorer(IE) では実装されていないが、その他多くのブラウザ (Mozilla や Netscape、Amaya など) で動作し、さらに IE 上であっても BlackBoard や eCollege、Maple TA などの LMS (Learning Management System, 学習管理システム) では MathML をサポートしているため、Web 上での教育支援システムにおいて有効なものだと言える [2]。

3.1 MathML vs. LaTeX

さらに、数学表現のマークアップ言語の代表格である LaTeX と MathML とを比較してみる。まずは

$$\begin{bmatrix} 3x & 3 \\ -1 & y \end{bmatrix} \tag{1}$$

という数式を例に取り、LaTeX と MathML とでソースを記述してみると、次のような違いがある。

- LaTeX を利用した場合の記述例

```
\begin{eqnarray}
\left[
\begin{array}{cc}
3x & 3 \\
-1 & y \\
\end{array}
\right]
\end{eqnarray}
```

- MathML を利用した場合の記述例

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
<mrow><mtable align="right" width="80%">
<mtr><mtd><mrow><mrow><mo>[</mo>
<mtable><mtr><mtd columnalign="center">
<mrow><mn>3</mn><mi>x</mi></mrow></mtd>
<mtd columnalign="center"><mrow><mn>3</mn></mrow>
</mtd></mtr><mtr><mtd columnalign="center">
<mrow><mo>-</mo><mn>1</mn></mrow></mtd>
<mtd columnalign="center"><mrow><mi>y</mi></mrow>
</mtd></mtr></mtable><mo>]</mo></mrow></mrow>
</mtd></mtr></mtable></mrow></math>
```

一見すればわかるように、LaTeX の方が入力量は少ないため、入力には適していると言えよう。一方で、MathML は、同様にマークアップ言語である LaTeX よりもデータの意味付けにおいて有利になるように設計されている。しかし、それは同時に、そのデータ構造が LaTeX と比較して複雑で、相対的に多くのメタデータを必要とする。つまり、入力には適していない（中間言語としての意味合いが強い）が、検索時には複雑な処理が行えることを意味している。例えば、MathML は単項演算子の '+' と二項演算子のそれを区別できるが、LaTeX ではできないだろう。

4 数式検索システムの開発

本研究では MathML ベース数式検索システムを作成する。開発環境は Web アプリケーション開発用に Java サーブレットそして JSP を利用する。Java サーブレットの開発環境に Java 2 Platform Standard Edition(J2SE)、サーブレットを動作させるための Web コンテナとして Apache Tomcat を利用し、Java によるプログラミングの総合開発環境として Eclipse を用いる。

4.1 準備：検索対象データ生成

今回システムの検索対象とする数学コンテンツとして、元々LaTeX で作成された線形代数コンテンツ IMED Linear Algebra[4] を利用する。本研究では、MathML 形式の数式データを検索するという目的から、LaTeX ソースをまず MathML へと変換する作業を行う。変換には Try out TtM[5] を用いる。今回は検索対象の範囲を [4] 内の 1 章分にあたる 11 ファイル及び、テスト用の 10 ファイルの計 21 のファイルに絞った。

4.2 インタフェース部

本システムは以下の 3 つの数式検索用インタフェースと検索結果画面から構成される。

・ **検索数式入力画面**(図 1-①)
検索したい数式のタイプにより、4 つの入力形態を許す。組み合わせによって複雑な検索要求を発行できる。
(A). 英数記号・演算子
英数記号で記述できる数値や識別子、演算子の入力を行う。ここに入力された文字列は簡単な字句解析プログラムにより、MathML における数値・識別子・演算子用のタグが付加され出力される。
(B). 行列
次の項目 (C) に本来分類されるが、行数・列数の指定ができるよう独立して用意した ((B) をクリックすると画面 (B)' がポップアップする)。行列サイズを指定しない場合は行数・列数共に '0' と入力すればよい。
(C). 数学特有表記
いわゆる数学特有の 2 次元表記の入力を行う。GUI のボタンを複数用意することで実現する (例えばべき乗を含む数式を作成したい場合には、図 2（(C) クリックでポップアップ）の左から 2 番目のアイコンを利用)。
(D). 数学記号やギリシャ文字
ASCII 文字にはない数学記号やギリシャ文字の入力を扱う。これらの文字は HTML のバージョン等により、その表記方法が実体参照、10 進数文字参照、16 進数文字参照といった複数の表記法を持つ。本プログラムではそれに対応済みである。

・ **検索数式表示画面**(図 1-②)
検索のための数式を作成・表示するための画面である (図 1 は"=(行列)・(行列)"の型の数式検索入力例)。数式作成には、上の (A)-(D) より文字入力あるいは数式表現を選択・send ボタンを押すという操作を繰り返し、最後に"search"をクリックにより検索が開始する。数式は MathML や正規表現文字列でも表示することができ、その切り替えは図 1-③ で行う。

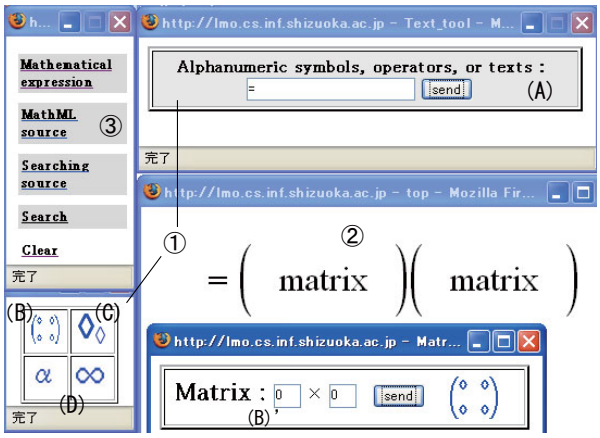

図 1. 検索システム実行画面

最後に図 2 と共に、(C) で扱うことのできる表現を簡単にまとめておく。図の左から $\Diamond_{\Diamond}$（例：下付文字)、$\Diamond^{\Diamond}$（例：上付文字、べき乗、行列転置)、$\Diamond_{\Diamond}^{\Diamond}$（例：上下付文字、総和、積分)、$\overset{\Diamond}{\underset{\Diamond}{\Diamond}}$（例：総和)、$_{\Diamond}\Diamond_{\Diamond}$（例：順列記号)、$\frac{\Diamond}{\Diamond}$（例：分数)、$\sqrt{\Diamond}$（例：平方根)、$\sqrt[\Diamond]{\Diamond}$（例：べき乗根)、$\overset{\Diamond}{\Diamond}$（例：上配置数式 ($\overrightarrow{AB}$ 等))、$\underset{\Diamond}{\Diamond}$（例：下配置数式 ($\lim_{x\to\infty}$ 等))、$\Diamond\Diamond$（積)。

図 2. 数学特有表現用ポップアップメニュー

4.3 検索処理部

検索には、正規表現を用いたマッチングを行う。以下の表 1 に J2SE がサポートする正規表現の中から主に利用している構文を示す。

表 1. 検索時に利用する J2SE の正規表現構文

正規表現	説明
.	任意の文字
?	直前の文字が、1 回または 0 回現れる
$(abc\|def)$	'|' で区切られた文字列のいずれか
$[\hat{}a]$	[ ] 内の文字 (左の場合は 'a') を含まない文字
$*?$	直前の文字を 0 回以上最小回数繰り返す

ここでは 3.1 項で挙げた (1) の検索を例にとり、その数式作成の手順と対応する MathML タグ、検索用正規表現文字列生成法を記述し処理部の説明を行う。
対象となる数式は 2 × 2 の行列であるので、まず 4.2 項の B の行列入力画面より行数・列数を指定し、要素不定の 2 × 2 の行列を生成する。その裏で MathML 形式として、行列表現に mtable タグ、mtr タグ、mtd タグ、行列式のための括弧表現に [ ]、および ( ) といった記号もしくは、mfenced タグが生成される。なお、[ ]、( ) は記号データであるので mo タグを利用する。対応する検索用正規表現文字列は以下のようになる：

```
(<mo>\\(</mo>|<mo>\\[</mo>|<mfenced>)<mtable>
<mtr><mtd><mrow>.*?</mrow></mtd><mtd><mrow>.*?</mrow></mtd></mtr>
<mtr><mtd><mrow>.*?</mrow></mtd><mtd><mrow>.*?</mrow></mtd></mtr>
</mtable>(<mo>\\)</mo>|<mo>\\]</mo>|</mfenced>)
```

(※実際は必要なタグ情報間にスペース・タブ・改行記号を読み飛ばす正規表現が含まれているがここでは省略している。)
その後、4.2 項の A の入力画面により、行列の (1,1) 要素となる "3x" を入力する。'3' は数字であるため mn、

‘x’ は識別子にあたるため mi のタグが相当し、先の MathML データの字句解析を行った結果、上記検索用文字列の最初の “.*?” 部を <mn>3</mn><mi>x</mi> で自動置換する。

最後に数学特有表現の検索処理部の説明として、タグに対応した検索用正規表現文字列をまとめる（表 2）。

表 2. 数式特有表記用検索文字列

タグ名	検索文字列
msub	”<msub>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</msub>”
msup	”<msup>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</msup>”
msubsup	”<msubsup>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</msubsup>”
munderover	”<munderover>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</munderover>”
mmultiscripts	”<mmultiscripts>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<none[\t]*?/>[\t\n]*?<mprescripts[\t]*?>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<none[\t]*?/>[\t\n]*?</mmultiscripts>”
mfrac	”<mfrac>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</mfrac>”
msqrt	”<msqrt>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</msqrt>”
mroot	”<mroot>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</mroot>”
mover	”<mover>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</mover>”
munder	”<munder>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?<mrow>.*?</mrow>[\t\n]*?</munder>”

5 評価実験

システムの評価実験は、今回検索対象とした 11 ファイルに対して行った。具体的には、当該ファイルに含まれる個々の数式に対する検索文字列を図 1 の実行画面によって作成し、正しくヒットしたかを調べた。今回は対象となるデータを少なく制限したため、正誤判定は目視で確実に行った。その結果を下の表 3 に示す。

大方正しく検索できたことが観察される一方で、一部検索から漏れた数式の存在も確認された。これは、同じ数式でも MathML による異なる表記法が存在したためである。特に、元々のソースファイルが構造的に正しく作成されていない場合、MathML 形式に変換されても、不適当な構造が継続されてしまうため、結果として検索されなかったと言える。特に、今回は実験として、元々の LaTeX ファイルをコンバータを用いて MathML にしているため、コンバータの精度が検索結果に影響を与えている。今後のコンバータの精度向上が望まれる。

最後に検索結果の様子を一例だけ示す。下図 3 は図 1 で例に用いた、行列の積が等式の直右に現れる数式を出力させた結果である。多くの検索エンジンの出力レイアウトと同様に、ファイル名ならびに一致した前後の文字列がやはり綺麗な数式で一覧表示されていることがわかる。

表 3 システムの評価実験結果

#	1	2	3	4	5	6	7	8	9	10	11
数式数	17	18	2	7	22	37	5	10	13	55	45
検索成功数	13	18	2	4	16	18	2	10	8	53	39

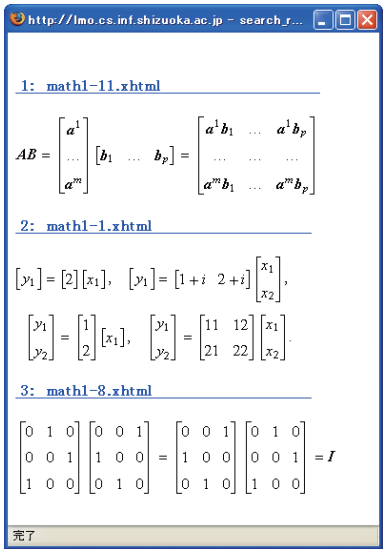

図 3. 検索システム実行結果（例）

6 おわりに

本稿では、MathML 形式で記述された数式コンテンツに対し、検索を行う Web アプリケーションの開発ならびに評価実験の結果を報告した。まだ完全なシステムに至ってはいないが、一部の数式を除き、正しく検索することを確認した。解決すべき課題としては、前述の不備を直すこと、そしてより大量のコンテンツを対象とした場合の高速検索処理実現などが挙げられる。

参考文献

[1] W3C：Mathematical Markup Language(MathML) Version 2.0 (Second Edition), (2003).

[2] Hans Cuypers, Karin Poels, Rikko Verrijzer, Olga Caprotti, Jouni Karhima, State of the Art in Mathematical E-learning, WebALT Consortium, pp. 9-11 (2005).

[3] Keiji EMI, Tsuneo IMAI, Developing E-learning and Communication Environment to Support Displaying Math Equations, International Symposium on Recent Trends in Global e-Learning and Collaboration, pp. 21-24 (2007).

[4] IMED Linear Algebra: http://128.97.76.119/linearalgebra/.

[5] TTM(Try out TtM): http://hutchinson.belmont.ma.us/tth/mml/ttmmozform.html.