DEIM Forum 2010 C7-4

数式の構造を反映した検索法

高田　真澄†　　村尾　裕一††

† 電気通信大学大学院電気通信学研究科情報工学専攻　〒 182-8585 東京都調布市調布ヶ丘 1-5-1
†† 電気通信大学電気通信学部情報工学科　〒 182-8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †mas83024@jed.uec.ac.jp, ††murao@cs.uec.ac.jp

あらまし　近年、MathML の登場により数式を再利用可能な形で Web ページに含めることが可能となってきた。MathML で表現された数式の検索を実現すれば、内容に基づいた検索が可能になると考えられる。本研究では、数式の詳細な構造を反映した検索法を提案し、実験的にシステムとして実装する。演算子や変数といった数式の構成要素をキーワードとみなし、木構造におけるキーワードの位置および適用範囲を構造的な特徴と定義する。提案手法では、この構造的な特徴を適合条件に加えることで、数式を入力キーとした文書検索の精度向上を目指す
キーワード　数式の構造　情報検索　MathML

Masumi TAKADA† and Hirokazu MURAO††

† Department of Computer Science, Graduate School of Electro-Communications, University of Electro-Communications
†† Department of Computer Science, Faculty of Electro-Communications, University of Electro-Communications
E-mail: †mas83024@jed.uec.ac.jp, ††murao@cs.uec.ac.jp

Abstract　Recently, the advent and the use of MathML for mathematical expressions has made it possible to include them as reusable entities in Web pages. We expect that representing mathematical expressions in MathML with retaining mathematical structures can realize content-based retrieval. In this study, we propose a retrieval method that counts detailed structures of mathematical expressions, and implement an experimental system. We regard the constituents of mathematical expressions such as operators and variables as keywords, and treat the positions and the scopes of keywords within trees as structural features. The proposed method aims at improving the accuracy of retrieval based on mathematical expressions as input keys by adding those features to conditions for matching or ranking.
Key words　structure of mathematical expression, information retrieval, MathML

1. はじめに

数式は理工学を始め、あらゆる分野で用いられ、Web 上の文書においても重要な情報である場合が多い。しかし、Google のようなテキストベースの検索エンジンでは正確に数式の構造を捉えきれず、数式を対象とした検索は困難であった。近年、数式を XML で表現する MathML [1] の登場により数式データを再利用可能な形で Web ページに含めることが可能となり、検索性が高まってきている。MathML で表現された数式の検索を実現すれば、内容に基づいた検索が可能になると考えられる。

数式検索の研究事例としては数式を構成する要素に注目した手法 [2] と数式の木構造に注目した手法 [3] がある。前者は演算子・変数・数値といった各要素を単語として、全文検索方式を適用するものである。後者はルートから各葉ノードまでのパスを XPath で表記し、一致するパスの多さで類似度を決めている。これらに共通しているのは、詳細な構造すなわち数式を構成する各要素間の関係を考慮していない点である。ユーザの曖昧な情報要求に応えるためには、全体の構造だけでなく部分的な構造を考慮する必要性が出てくる。最近では数式の構造的な特徴を捉える研究も盛んに行われてきており、独自の問い合わせ言語により木構造マッチングを行う [4] や木構造の類似度を計量して適合性を調べる手法 [5] のような試みもあるが、確立された手法が存在しないのが現状である。

本研究では、数式を含む文書を検索対象として、数式の詳細な構造を反映した検索法を提案し、実験的にシステムの実装を行う。演算子や変数といった数式の構成要素をキーワードとみなし、木構造におけるキーワードの位置と演算子適用範囲を構造的な特徴と定義する。提案手法では、この構造的な特徴を適

合条件に加えることで数式を入力キーとした文書検索の精度向上を目指す。

本稿では、まず 2 章で数式データのフォーマットである MathML について紹介し、数式の再利用性について述べる。次に 3 章で数式検索に関する過去の研究事例を紹介し、数式検索における情報要求についてテキストベースの検索と対比させながら説明する。4 章で提案手法の解説をし、5 章でシステムの設計方針や構成について述べる。6 章では再利用性の低いデータ形式への対応法の検討を行う。7 章で実装したシステムでの実験・評価結果を示し、最後に 8 章でまとめ及び課題を述べる。

2. 数式データの形式

2.1 MathML

MathML は XML ベースの数式記述用マークアップ言語であり、描画用の Presentation Markup と内容記述用の Content Markup の 2 種類の記述方式が用意されている。

```
<math>
  <mrow>
    <msup>
      <mi>b</mi><mn>2</mn>
    </msup>
    <mo>-</mo>
    <mn>4</mn>
    <mi>a</mi><mi>c</mi>
  </mrow>
</math>
```

図 1　Presentation Markup

```
<math>
  <apply><minus/>
    <apply><power/>
      <ci>b</ci><cn>2</cn>
    </apply>
    <apply><times/>
      <cn>4</cn>
      <ci>a</ci><ci>c</ci>
    </apply>
  </apply>
</math>
```

図 2　Content Markup

Presentation Markup は数式の描画のためのマークアップであり、主に Web ブラウザでの数式表示に用いる。Content Markup は式の意味構造を正確に記述するためのマークアップであり、数式処理・計算ソフトの間での受け渡しを可能にするものである。図 1,2 に両方式による記述例を示す。

2.2 TeX との比較

数式を表現する手段としては、TeX が古くから利用されている。TeX は可読性に優れ、人間が直接記述するのに適した方式である。一方、MathML はどれが変数でどれが演算子かといった事を正確に表現しており、コンピュータが処理するのに向いた方式である。

```
b^2-4ac
```

図 3　TeX

TeX は Web 上においても Wikipedia を始め多くのサイトで使われている。ただし、それらは直接 HTML に数式を埋め込んでいるのではく、一旦サーバ側の変換プログラムに渡して画像へ変換してから表示している。画像形式となった数式は再利用性が低く、データの蓄積・検索には向いていない。この点では、直接 Web ページに埋め込む事が可能な MathML の方が有利である。

2.3 数式の再利用性

ここでは、数式の再利用性という観点から MathML の 2 種類の方式を比較する。Presentation Markup については、図 1 を見ると、記号を表示のために並べただけであり、演算子の適用順序が明確ではない事が分かる。一方、Content Markup については累乗や加算等が記号ではなく意味を持った演算子として、適用を繰り返す形で記述されている。適用を表すタグに続いて演算子を記述し、その後ろに引数を並べる形で記述するため、適用順序に曖昧さが存在しないのである。データの蓄積・検索向きの再利用性の高い方式と言える。これを踏まえ、以降では Content Markup 形式の数式を主な対象として議論していく。

2.4 数式構造と S 式

Content Markup の特徴は、演算子適用を表す”apply”を節ノード、演算子・変数・数値を葉ノードとした前置表現の木である事である。”apply”ノードを頂点とした部分木に注目すると、１番目の子に適用する演算子を持ち、２番目以降の子にはその演算子の引数が並んでいる。引数には葉ノード以外に部分木を持つことができ、演算子適用の入れ子構造を形成している。この考え方は、Lisp や Scheme で用いられる S 式に近い。そこで本研究では、演算子を先頭要素とし引数に後続の要素を並べたリストとする前置表現の S 式で数式の内容を表す。

3. 数式を検索するという事

3.1 研究事例

過去の研究事例を見ると、数式検索の手法は大きく 2 つに分けられる。１つは数式を構成する要素に注目した手法であり、[2] がその代表である。この研究では、MathML のタグ情報に注目し、Latent Semantic Indexing(LSI) を利用した数式検索を提案している。数式の特徴である演算子等を取り出してベクトル空間を生成し、次元を縮退して近似するというものである。特徴を見ているため、「sin,cos の式である」といった意味的な情報をとらえることはできるが、数式の構造は全くとらえることができない。

2 つ目は数式の木構造に注目した手法 [3] である。この研究では、ルートから各葉ノードまでのパスを生成して索引化を行い、検索の際には一致するパスの多さで類似度を決めている。構造的な特徴をパスで表現することにより、木構造の概形が類似した数式をとらえることが可能となっている。しかし、パスがちょうど一致しなければならないため、「sin の子孫に π を持つ」といった曖昧な特徴には応えることができない。

他にも数式の構造に注目した研究事例をここでいくつか紹介しておく。MathML の数式検索ではないが、参考になる研究としては [6] がある。この研究では数式処理システムでの利用を想定した数学公式データベースを構築し、独自の検索言語による検索法を提案している。構造・性質・名前という 3 つの観点から公式を捉え、検索の際には各々の条件でふるいをかけている。構造のインデックスは関数名や定数等のキーワードのレベル (木構造でのノードの深さ) で表現される。このインデックスを利用することで、検索要求の構造に完全に一致しなくても、

同レベルにあるキーワードを入れ替えた近い構造をも適合可能にしている。

MathML の数式を対象とし、構造が曖昧である場合にも検索可能となるように試みたのが、論文 [7] である。この論文では、利用者がイメージする数式の構造を S 式で表現し、S 式の構造を適合条件として検索実験を行い、その有効性を確認している。以下がその一例であるが、これは「2 乗して足した式」という情報要求を表している。

```
(plus (power 2) (power 2))
```

構造的な特徴の表現の仕方は研究によって様々であるが、いずれにおいても適合条件を柔軟にし、どれだけ曖昧な情報要求に応えられるかがポイントとなっている。

3.2 情報要求の処理

入力に完全に一致するものを探すだけであるならば、文字列のマッチングでも事足りるであろうが、実際には様々な情報要求が存在する。数式検索の分野にも言えることで、例えばある公式について知りたいがおおまかな構造しか分からないため、曖昧な入力式が与えられた場合を考える。すると情報要求は、「入力 (構造) が特徴的に表れている文書」や「入力 (構造) に関連する文書」である。ここではシステムの方向性を明確にするために情報要求の処理法について、複数の段階に分けて考える。

- 数式の名前が存在し、その名前を知っている場合

公式名を検索クエリとして、テキストベースの検索を行えばよい

- 名前は分からないが、数式は正確に記述可能な場合

数式を入力として、完全に適合する数式を探せばよい

- 名前も数式自体も曖昧な特徴しか分からない場合

おそらくこのパターンが一番多いと考えられる。数式の構造的な特徴や数式に関連するであろう単語を入力として、最も適合性の高い文書を提示する必要がある。

以上の 3 パターンに応えるためには、検索システムの条件に「数式」、「公式名」、「構造的特徴」、「数式に関連した単語」を指定できれば良い。本研究においては、「公式名」と「数式に関連した単語」を単語入力として、「数式」と「構造的特徴」を数式入力とする。数式入力では完全な数式の形を要求せず、利用者のイメージとして入力された形から構造的な特徴を抽出し、検索を行う。ただし、数式入力が完全な形をしている場合には完全一致の検索としても振舞うようにする。

4. 構造を反映した検索法

数式は一般に演算子・変数・数値で構成されるが、これは S 式による表現の場合も同様である。下の数式 $b^2 - 4ac$ を表す S 式を例として構成要素に分類したのが表 1 である。

```
(minus (power b 2) (times 4 a c))
```

検索要求の数式の構造とは、すなわち S 式の入れ子構造であり、これら構成要素の関係である。例えば、「minus の子に power,times を持つ」といった関係が検索要求となる。しかし、検索の度に S 式の構造をいちいち見るわけにはいかないので、

表 1　数式の構成要素

演算子	minus,power,plus
変数	b,a,c
数値	2,4

構造を数値で表現する必要がある。

構造の数値化

S 式のリスト構造において各キーワードが出現する順番を位置情報として記録していく。さらに入れ子構造を反映するため、演算子はその適用範囲も記録する。なお変数については、数式内での出現順に番号付きの名前を与え、変数名依存の解消をはかる。図 4 に解析結果を木で表現した例を示す。各ノードの脇に記載されているのが位置情報である。

適合条件

入力に対し適合する条件は二つある。まず一つ目は、検索キーに含まれるキーワード全てを含むことである。二つ目は検索キーの S 式に近い構造を持つことである。入力式があるノードとノードの間に親子関係を持っている場合、対象データとしては先祖子孫関係を持っていれば近いとみなす。これは、検索要求の曖昧さに合わせて適合条件を緩くするためである。例でいえば、入力数式は「演算子 plus の子孫に変数を 2 つ持つ」と解釈され、$a + b$ のような完全に一致するものだけでなく $a + 3b$ のような構造的に近い数式も適合する。

表 2　位置情報の記録

キーワード	出現位置	適用範囲
minus	1	(1,8)
power	2	(2,4)
b	3	
2	4	
times	5	(5,8)
4	6	
a	7	
c	8	

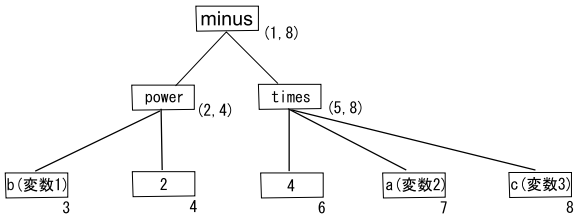

図 4　木による表現

5. 数式検索システムの概要

5.1 設計方針

以下に本システムの設計方針を掲げる。

- 本文および数値化した数式の構造を索引に登録
- 数式入力には完全な形を要求せず、構造的な特徴を抽出
- 数式の構造・大きさを元にした順位付けを行う

5.2 システムの全体構成

本システムの全体構成を図 5 に示す。文書フィルタでは、本文から形態素解析により単語の集合を生成する。数式データは親が演算子で子が引数となる純粋な S 式の形に変換し、各キーワードの位置情報をカウントする。インデクサはその解析結果をデータベースの対応するテーブルに登録していく。検索プログラムでは、検索キーに指定された S 式または単語を元にクエリを構築し、問い合わせを行う。

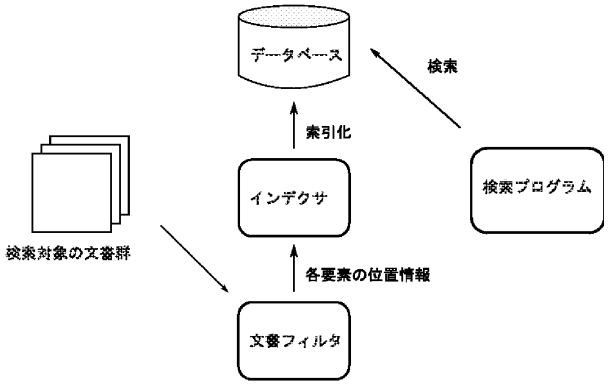

図 5　システム構成

5.3 文書の解析と索引化

文書フィルタではまず、文書群から本文と MathML(Content Markup) の形式で埋め込まれた数式データを抽出する。本文の中で必要な情報はテキストノードであるので、木を再帰的に探索して全てのテキストノードを取り出す。続いてテキストノードの集合を形態素解析し、本文を構成する各単語の集合を生成する。この単語の集合がテキストベースの検索における索引である。

次に数式データの解析についてであるが、まず行うのが S 式への変換である。変換は XML の S 式表記である SXML [8] という規格に従って行う。しかし、SXML のままでは数式の特徴とは関係ないノードが存在し、構造も捉えにくいため、親が演算子で子が引数となる純粋な形に簡略化しておく (図 6)。

```
(math (eq D (minus (power b 2) (times 4 a c))))
```

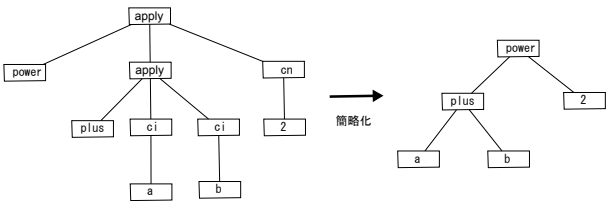

図 6　構造変換による数式表現の簡略化

変換が済んだら、S 式における各構成要素の位置情報をカウントする。位置情報のデータ群は添字をキーワード、値を位置情報とした連想配列として定義される。

> ’キーワード’: (開始位置, 終了位置, 階層レベル)

開始位置はキーワードの出現する位置であり、終了位置はキーワードの適用範囲内において最後に出現するキーワードの位置である。階層レベルは、数式の木構造における深さを表している。インデクサはこの位置情報をデータベースの対応するテーブルに登録していく。

5.4 データベースの構成

ここでは、解析結果を索引化するためのデータベースの構成を説明する。本データベースのスキーマを以下に規定する。検索の際にベースとなるのは、キーワードの位置テーブルと単語の位置テーブルである。キーワードの位置テーブルでは、数式テーブルの数式番号、キーワードテーブルのキーワード番号と紐付き、各数式のキーワードの位置を表現する。同様にして、単語の位置テーブルでは文書テーブルの文書番号、単語テーブルの単語番号と紐付き、各文書の単語の位置を表現する。

- 数式 (数式番号, 数式の内容)
- キーワード (キーワード番号, キーワード名)
- キーワードの位置 (数式番号, キーワード番号, 開始位置, 終了位置, レベル)
- 単語 (単語番号, 単語)
- 単語の位置 (文書番号, 単語, 出現位置)
- 文書 (文書番号,URL)

5.5 クエリ構築と問い合わせ

検索キーには S 式または単語を指定することが可能である。単語の指定があった場合は、単語の位置テーブルに問い合わせ、適合した文書の番号を取得する。S 式の指定があった場合は、その入力式についても各キーワードの位置情報をカウントする。解析結果は対象文書内の数式を解析した時と同様の連想配列である。処理の流れを以下に示す。

step1　ユーザが入力した数式を解析し、各キーワードの位置情報を取得する。

step2　各キーワードについて、キーワードテーブルに問い合わせ、キーワード番号を取得する。

step3　各キーワード番号およびキーワードの位置情報から、キーワードの位置テーブルへ問い合わせるためのクエリを構築する。(適合条件は 5. 章参照)

step4　構築したクエリを元に問い合わせを行う。入力数式の全キーワードについて適合条件を満たす数式が検索結果となる。結果のデータは、数式番号と数式内の全キーワードの位置情報が合わせて返される。

本節では、検索キーに単語が指定された場合と S 式が指定された場合を別々に説明したが、両方を同時に指定することも可能であり、この場合は両者の検索結果の積集合 (AND) をとる。

5.6 順位付け

適合した数式を順位付けする際の尺度として、今回は二点を採用する。

- ノード間の関係

親子関係にちょうど適合するような数式のスコアを高くするのが狙いである。キーワード間の構造における距離を階層レベル

の差と定義し、先祖子孫関係を数値的に表現する。入力数式と適合数式の全キーワード間の距離の差が、構造の近さを表す指標である。

- 式の大きさ

入力数式の大きさに近いほど良いスコアとする。これは、一般的に冗長な数式より簡潔に記述された数式の方が重要な場合が多いためである。数式の大きさ=キーワードの総数とみなせるので、キーワード総数の差分が数式の大きさがどの程度近いかを表す指標となる。

最終的には、最小値を基準に 0 から 1 の間で正規化し、値が大きいほど良いスコアとなるようにする。表 3 と表 4 にそれぞれの尺度で順位付けした例を示す。

表 3　ノード間の関係による順位付けの例

入力	適合した式	S 式表記	スコア
ab	$ab+c$	(plus (times a b) c)	高
	$a(b+c)$	(times a (plus b c))	低
$\sin x$	$\sin(a+b)$	(sin (plus a b))	高
	$\sin x^3+\cos x$	(sin (plus (cos x) (power x 3)))	低

表 4　数式の大きさによる順位付けの例

入力	適合した式	S 式表記	スコア
ab	$ab+c$	(plus (times a b) c)	高
	$x+ab+9y$	(plus x (times a b) (times 9 y))	低

6. 実　　験

本研究で実装したシステムの有効性を検証するための実験を行った。いくつかの入力データを用意して実験を行い、それぞれ精度を比較する。精度は、

$$\frac{\text{適合した文書}}{\text{検索された文書}}$$

と定義され、検索結果にどれだけノイズがあるかを表す指標となる。対象文書は MathML Test Suite(注1) 内にある Content Markup が埋め込まれたページ群である。Test Suite 内には単純な数式からやや複雑な数式までバランスよく存在するため、数式入力による検索による適合具合を知るのに適している。結果を表 5 に示す。

表 5　実 験 結 果

入力	ヒット件数	適合数	精度	適合例
ab	39	34	0.87	$-ab$
				$x(y+z)z$
				$x(a/b+c)-1$
$\sin x$	24	18	0.75	$\sin(a+b)$
				$1/\sin t$
				$\sin(\cos x+x^3)$
$\sin x/\cos x$	14	7	0.5	$(\sin x/\cos x)^2$
				$(1-\cos x)/\sin x$
				$\cos t/\sin t$

(注1)：http://www.w3.org/Math/testsuite/mml2-testsuite/index.html

精度は入れ子構造が少し複雑になると下がるものの、比較的良い値である。入力数式を ab とした検索結果には、$x(y+z)z$ のように一見すると不適合な数式が結果に含まれている。数式の外見を情報要求とした場合には、確かに適合しているとは言えない。しかし、本研究の想定する情報要求は数式の内容であるため、変数名が変わることや、変数を式で置き換えることがある。つまり a,b はパターン変数に近い役割を担っているのである。厳密に式変形をして比較しなくとも、構造の条件を緩くする事で変形した数式に適合することが確認できた。

7. 表示用形式への対応法の検討

提案手法には、対象文書に埋め込まれた数式が Content Markup 形式であるという前提が存在した。しかし、実際に文書に埋め込まれる数式の多くは Presentation Markup 形式や TeX 形式である。これら表示用マークアップに対して内容に基づいた検索を行うには、Content Markup への変換が必要となる。Content Markup 形式への変換を試みた事例として [9] があるが、その中で問題となっているのは Presentation Markup の記述方法の自由度の高さである。2. 章でも述べたように表示用マークアップには意味・構造的な曖昧さが存在してしまうので、1 対 1 の完全な自動変換は困難なのである。最も分かりやすいのは乗算を含む式である。数式 ac を Presentation Markup 形式で記述すると、乗算の演算子が省略されるため、変数 a と c の乗算であるか変数 ac であるかが読み取れない。演算子の省略以外にも、演算子の適用順序を明確にしない記述が可能であるといった問題もあり、これらの曖昧さが意味形式への自動変換を難しくしている。しかしながら、表示用マークアップのパターンから意味を予測することは可能である。意味予測により候補となる Content Markup を複数生成し、検索対象として含めるという方法が有効であると考えられる。ただし、Presentation Markup の記述パターンは非常に多様であるため、使用頻度の高いタグに絞る等の対策は必要であろう。

8. おわりに

本研究では、数式の構成要素に注目し、それらの先祖子孫関係を適合条件に取り入れた構造ベースの数式検索法を提案した。また、実験的な数式検索システムを実装し、提案手法の検証を行った。その結果、変形された式等の構造的に近い数式を適合可能であることが分かり、構造を条件とした検索の有効性を確認した。

今後の課題は意味的な類似性の考慮である。一般に数式同士の意味的な類似性を直接判定するのは困難だが、ここではアイディアを述べておく。数式の付近には本文というメタデータがあり、中には数式の意味を示す単語が含まれていることが多い。そのメタデータを元にあらかじめ対象文書群を意味的尺度でグループ分けしておけば、グループ内のどれかが適合した際に類似文書としてグループ内の他の文書を提示することができると考えられる。

また、数式入力の問題がある。今回実装したシステムの数式入力は S 式を採用しているため、数式を専門的に用いる人達に

とっては都合が良いが、一般の利用者には扱い難いと考えられる。今後は数式入力システムと連携し、GUI による入力も可能にする必要があるだろう。

文　　献

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

[2] 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).

[3] 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 (セッション 4: XML), 情報処理学会研究報告. 情報学基礎研究会報告, Vol. 2007, No. 54, pp. 55–59 (2007).

[4] 小田切健一, 村田剛志：MathML を用いた数式検索, 人工知能学会全国大会 (2008).

[5] 横井啓介, 相澤彰子：類似性を考慮した数式検索手法の提案, FIT2009 第 8 回情報科学技術フォーラム (2009).

[6] 三枝義則, 阿部昭博, 佐々木建昭, 増永良文, 元吉文男, 佐々木睦子：数式処理システム GAL における数学公式データベースのインデキシング手法 (新しいデータベース技術論文特集), 電子情報通信学会論文誌. D-1, 情報・システム. 1, 情報処理, Vol. 74, No. 8, pp. 577–585 (1991/08).

[7] 椎名正樹：数式を含む文書の検索方式, 修士論文, 電気通信大学大学院電気通信学研究科 情報工学専攻 (2007).

[8] SXML, http://okmij.org/ftp/Scheme/SXML.html.

[9] 石山寿子, 高野文子, 佐藤浩史, 原俊介, 大武信之：XML における数式の表示形式から意味形式への変換, 電子情報通信学会技術研究報告. ET, 教育工学, Vol. 101, No. 506, pp. 23–30 (20011208).

[10] 徳永健伸：情報検索と言語処理 (言語と計算), 東京大学出版会 (1999).