4N-6

数式データベースを用いた曖昧な数式の発見

Detecting Ambiguous Formulae Using Mathematical Expression Database

大瀬戸 良輔 [1]　甲斐 博 [1]
R.Osedo[1]　H.Kai[1]
（愛媛大学大学院理工学研究科 [1]）

1. はじめに

数式表記用の XML として知られる MathML には，数式の表記を記述するための Presentation Markup と数式の数学的意味を記述するための Content Markup が定義されている．Content Markup は数式処理システムにかけることで計算させることが可能だが，Presentation Markup は表示に適しており，広く一般的には Presentation Markup が普及している．

このような背景から，Presentation Markup から Content Markup への変換が可能になれば，Presentation Markup の再利用性が向上するといえる．

しかしながら数式表記には，複数の意味対応が存在するものがあるため，Presentation Markup から Content Markup への変換を施す際，一意の変換ができない．このように複数の意味対応が存在する数式を，本研究では曖昧な数式として定義する．

本研究では，Presentation Markup から Content Markup への変換を考える前段階として，曖昧な数式の発見を試みた．特に，Wolfram Functions Site[1] で公開されている MathML を収集し数式データベースを構築し，これを分類することを検討した．

2. Presentation Markup の曖昧性

Presentation Markup は数式の数学的意味を考えていないため，計算には不向きである．例えば，仕様書や設計書などに記述された数式を計算し，別の文書を作成するなどといった状況においては，すでに記述された数式 (Presentation Markup) を計算しやすい形式 (Content Markup) に変換できることが望ましい．

ただし，Presentation Markup では，一つの数式が様々な数学的意味を記述することができる．代表的なものとしては $|a|$ といった表現は，絶対値・行列式・濃度という複数の意味が考えられる．また，意味が異なるため対応する Content Markup も異なる．表 1 に $|a|$ を表す Presentation Markup と，対応する Content Markup をそれぞれ記述する．ここで，`&#10072;` は Unicode における"|" を表す数値文字参照である．

表 1　$|a|$ を表現する MathML

Presentation Markup	<mo>❘</mo> <mi>a</mi> <mo>❘</mo>
Content Markup1	<apply> <abs/> <ci>a</ci> </apply>
Content Markup2	<apply> <determinant/> <ci type="matrix">a</ci> </apply>
Content Markup3	<apply> <card/> <ci>a</ci> </apply>

また，$sinx$ という表記は変数 s，i，n，x の積と，三角関数 sin を適用した x のどちらを表すのかが不明瞭といった，数式の構造による曖昧さもある．

これら以外にも曖昧な数式の例は多く存在し，数式情報を異なる形式に変換する際は常に問題として現れてくる．

MathML の変換に関する研究では，Naylor らのメタスタイルシートを用いた変換に関する研究 [2] が挙げられる．これは独自に定義した XML をスタイルシートにより Presentation Markup と Content Markup へと変換する研究であり，複数の形式を任意に選択するための方法を検討している．

また，Presentation Markup から Content Markup への直接変換について述べられた研究としては，石山らの研究 [3] が挙げられる．この論文では，曖昧な数式に関する考察が行われており，スクリプトを用いた変換を検討している．

清水らの研究 [4] では曖昧な数式表記の解釈方法について述べられており，特にトークン間の位置関係から結合度を算出し，それに基づき意味的関連付けを行うことで曖昧な数式の意味を厳密に定めることに成功している．

但し，表 1 のような同一の表記で複数の意味が考えられる記号がどのような場合に表れるかについては網羅されていない．

3. 曖昧な数式の発見のための補助ツール

数学記号の曖昧性を判定する際は数式表記を実際に人の目で見て，表記のもつ意味を推測し，複数の意味対応が考えられると認識する必要がある．しかしながらこの手法で曖昧な数式を発見するためには，膨大な時間と MathML の知識が求められる．

そこで，本研究では曖昧な数式の発見を手助けするツールを開発した．まず，予め Presentation Markup とそれに対応する Content Markup を数式データベースへと格納しておき，それぞれで出現するタグやテキスト値を抽出し，これに対して検索を行うことで曖昧な数式を発見することを行った．本論では，具体的には以下のような手順で曖昧な数式を発見する．ここで扱う数式データベースには，

- 数式テーブル（数式 ID, Presentation Markup, Content Markup，URL）
- 表記テーブル（数式 ID）
- 意味テーブル（Content Markup 中のテキスト）

を格納する．ここで URL は Wolfram Functions Site 上の数式の存在する URL を示す．

（1）Presentation Markup に出現すると予想される表記（テキスト）を入力する．以降はここで入力した表記に関して，対応する意味を探していく．
（2）入力された表記をもつ Presentation Markup に対応する数式 ID を表記テーブルに追加する．
（3）表記テーブルの先頭にある数式の URL および Content Markup を取得し，これらを GUI で表示する．
（4）ユーザは表示された数式画像と Content Markup を見比べ，入力された表記に対応する値 (例えば，Content Markup の演算子タグや ci 要素などのテキスト文字) を，表示された Content Markup から探し出す．探し出したテキストは意味テーブルに格納する．
（5）探し出したテキストと同じテキストを数式（Content Markup）に含む数式 ID は全て表記テーブルから削

除する．
(6) (3) へ戻り，表記テーブルのデータが空になるまで処理を続ける．
(7) 最終的に意味テーブルへ格納されたテキスト値が，その表記のもつ意味を表すデータの集合である

ツールの使用例を図 1，図 2 に示す．本ツールは Java 言語を用いて開発した．処理の状況をコンソール上に表示し，Content Markup や数式画像の表示，並びに意味の登録は GUI 上で行う．

図 1 では表記 C に対応する意味が意味テーブルに 3 つ格納されており，このテーブルをもとに表記テーブルを再構成する過程を示している．図 2 では GUI で表示された画像をもとに，表記 C に対応する意味の決定とその値の意味テーブルへの登録を行っている．

この方法の問題点は，

- (1) で入力した表記と (4) で得られたテキストの 1 対 1 の対応がとれているかどうかが保証されていない
- 1 つの Content Markup 中に (1) で入力した表記に対応する複数の意味が含まれる場合，必要な意味を見逃してしまう可能性がある

ことである．これらを改善する方法は今後の課題である．

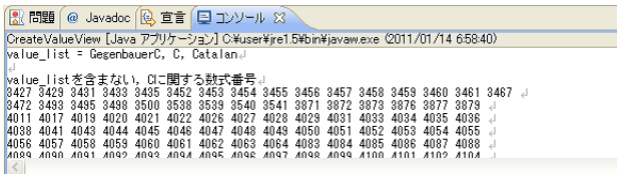

図 1　ツールの使用例 (コンソール)

図 2　ツールの使用例 (GUI)

4. 実験

本研究では，Wolfram Functions Site より取得した 151,626 の数式について調査を行った．これらの数式は semantics タグで表記されており，全て Presentation Markup と Content Markup が併記されている．これらを PostgreSQL に格納して調査を行う．

今回は特に多くの異なる意味をもっていると考えられる数式表記 C について，得られた意味を以下の表 2 に示す．

表 2　C と表記する数式

表記	意味	Content Markup
C	変数 C	<ci>C</ci>
C	カタラン数	<ci>Catalan</ci>
$C(x)$	フレネル積分	<apply> <ci>FresnelC</ci> <ci>x</ci> </apply>
C_n^m	ゲーゲンバウア多項式	<apply> <ci>GegenbauerC</ci> <ci>n</ci> </apply>
$C_p(z)$	円分多項式	<apply> <ci>Cyclotomic</ci> <ci>p</ci> <ci>z</ci> </apply>

MathML Content Markup では数式の意味を表現する演算子タグが定義されていない場合は，表 2 のように ci 要素で表記する．また，csymbol 要素や semantics 要素などの外部定義を参照する要素や，属性の指定などで明示的に意味を記述することもある．今回の実験で見つかったのは表 2 の通りであるが，このほかにも，C という数式表記には組合せ ${}_nC_m$ などの意味が考えられる．

5. 結論

本研究では複数の意味対応をもつ可能性のある数式を曖昧な数式として定義し，これを探し出すことを手助けするツールを開発し，曖昧な数式の発見を試みた．

但し，本研究では曖昧な数式をヒューリスティックに探し当てているため，限られた数の曖昧な数式しか得られていない．今後の課題として，まず曖昧な数式の検出を自動化する方法の検討が挙げられる．また，Wolfram Functions Site 上にはない数式も存在するので，検索対象の拡大も検討していく必要がある．

参考文献

[1] Wolfram Functions Site, http://functions.wolfram.com/
[2] Bill Naylor and Stephen Watt: Meta-Stylesheets for the Conversion of Mathematical Documents into Multiple Forms. Annals of Mathematics and Artificial Intelligence, Vol.38,No 1-3,pp.3-25, 2003
[3] 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(506), pp.23-30, 2001
[4] 清水智巨, 陳ウン, 岡田稔: 曖昧さに注目した数式構造理解, 電子情報通信学会, PRMU, パターン認識・メディア理解 Vol.99(182), pp.1-8, 1999