N-032

PDF 文書中の数式情報抽出に関する研究
Expression information in PDF document extraction

松島 一平 †
Ippei Matsushima

古賀 雅伸 †
Masanobu Koga

1. はじめに

近年、論文等の文書を作成する際のファイルフォーマットとして PDF が多く用いられている。これは PDF には PC 上で閲覧する際に非常に見やすい、様々な OS 上で見ることができる、ファイルサイズが小さいなどの利点があるからである。

PDF の構造 [1] は簡単に述べるとオブジェクトの集合体であり、各オブジェクトは PDF を読み取り表示するツールがアクセスしやすいように配置されている。また、テキスト及び画像を表示する部分には適宜圧縮がかけてある。そのため、ファイルの中身を直接エディタ等で見たところで内容は非常にわかりにくいものとなっている。したがって、PDF を解析するためには専用の処理が必要である。

近年、PDF 文書中の情報を利用しようとする研究も多く、テキストや画像、表などを抽出するといった試みがある。理工系の論文には多くの数式が含まれているが、PDF からテキストなどを抽出できるツール群、例えば Adobe Acrobat でも数式に関してはテキストとしては保存できているが、数学的な意味は失われた形となる。PDF 文書中に存在する数式情報を数学用アプリケーションで利用できる形式で出力できれば、その論文の数学的データベースとしての価値が高まる。しかし、PDF の構造及びテキスト表示の特性が非常に複雑なため、PDF から数式情報を直接読み取ることは困難である。

数式を記述している部分は TEX の様に数式の始点と終点を指定するシーケンスがないので、数式とテキストの区別は非常につきにくい。このため、PDF 中の文章を数式として認識することは非常に難しい問題である。

そこで、本研究では PDF ファイル中に含まれる数式情報を MathML[2] に変換できる、PDF2MathML を開発した（図 1）。

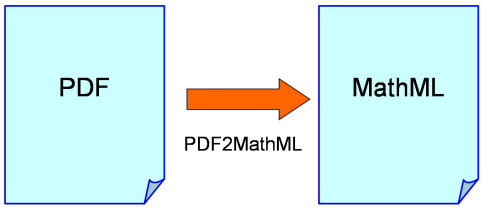

図 1: PDF2MathML

†九州工業大学

2. 研究で使用した技術

2.1 MathML

本研究は、PDF から抽出した数式情報を MathML の形式で出力する。MathML とは XML[3] ベースの言語であり、数学的な情報を表すことに特化した言語である。そして MathML の表記には 2 種類あり、数式の表示に関する情報を保持する Presentation Markup と数学的な意味を保持する Content MarkUp がある。以下に MathML の例を示す。表現したい式が $x^2 + 2x = y_i$ とすると、Presentation Markup ならば

Presentation Markup

```
<math xmlns=
"http://www.w3.org/1998/Math/MathML">
  <msup>
    <mi>x</mi>
    <mn>2</mn>
  </msup>
  <mo>+</mo>
  <mn>2</mn>
  <mi>x</mi>
  <mo>=</mo>
  <msub>
    <mi>y</mi>
    <mi>i</mi>
  </msub>
</math>
```

と表記する。

2.2 JPedal

本研究では、PDF を解析ために JPedal[4] を用いた。JPedal は、Java 言語で PDF を処理することができ、GPL 版がフリーで使用できるソフトウェアである。JPedal を利用することで、PDF ファイル中のテキスト表示に関する以下のような情報を抽出する。

- テキスト
- テキストを配置する座標
- テキストのフォント
- テキストの文字サイズ
- 図形として描かれている直線の始点、終点の座標

3. PDF2MathML

PDF2MathML は、実行時の引数に数式情報を抽出したい PDF ファイルを指定することで、含まれている数式情報を数式 1 つにつき MathML ファイルを 1 つ生成する。また、任意の範囲のページを指定できる。

3.1 数式情報の抽出

JPedal を用いて抽出した情報から、以下のような条件を用い数式とテキストを判別する。

1. 演算子を含む
2. 数式フォントが存在する
3. 数字を含む

これら 3 つの条件のうち、1. の条件に加え、2. または 3. の条件が成立すると、その文字群は数式であると判断する。

3.2 オブジェクト モデルの生成

そして、選別された数式情報を図 2 に示すようにオブジェクトモデルとして表現する。

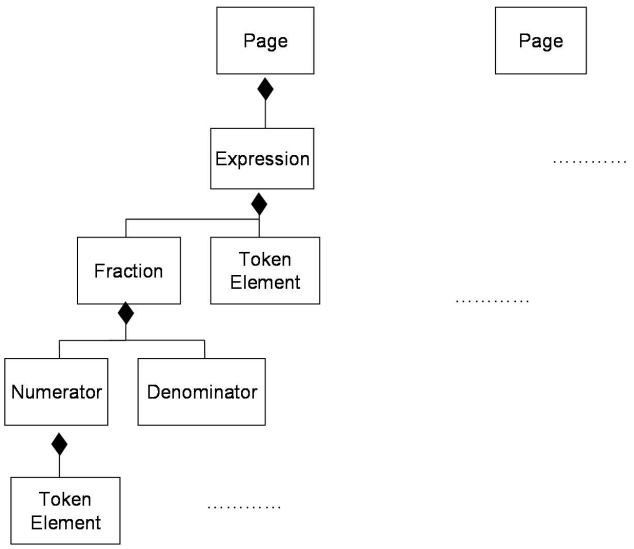

図 2: 数式のオブジェクトモデルの概念図

これは、MathML の Presentation Markup に準拠したオブジェクトモデルである。

3.3 MathML の生成

このオブジェクトモデルに沿って XML ドキュメントツリーを生成することで、全く同じ構造の Presentation Markup の MathML を得ることができる。XML ドキュメントツリーの生成には JDOM[5] を用いた。

4. 例題

本研究で作成したプログラムで、PDF ファイルから MathML を抽出する例を示す。まず数式を含む TEX 文書を作成する。TEX から生成された dvi ファイルを ps ファイルに変換する。次に Adobe Acrobat で ps ファイルを PDF ファイルに変換する。これにより作成された PDF ファイルを PDF2MathML の引数に与え、実行する。以下に例題として作成した PDF 文書中に含まれる数式を示す。

$$2C = \begin{bmatrix} a+c & 2 \\ 3 & \frac{1}{2} \end{bmatrix}$$

PDF2MathML により得ることができた MathML の一部を図 3 示す。図 3 は webMathematica[6] を用い、MathMLが正しく生成できたことを確認した様子を示す。

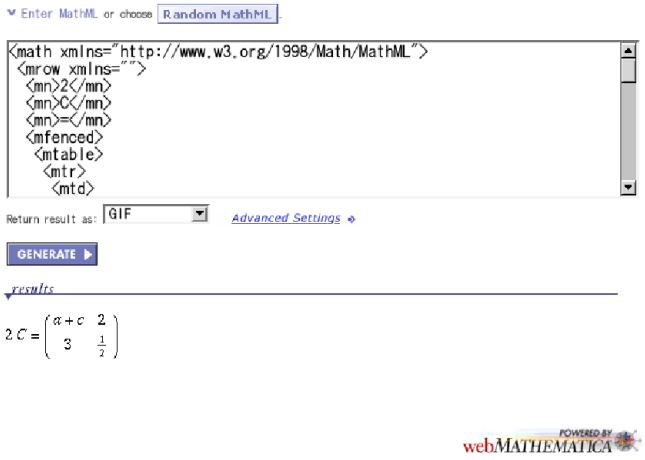

図 3: webMathematica による数式情報の確認

5. まとめ

本研究では、学会や Web 上で論文などを公開するために使われている PDF ファイルから数式情報を、MathML という形式で抽出できるツール PDF2MathML を開発した。本研究で開発した PDF2MathML では、現在 Presentation Markup の生成しか行うことができない。今後は、作成した Presentation Markup からの Content Markup へ変換する方法を検討する。

参考文献

[1] アドビシステムズ. PDF リファレンス 第 2 版. ピアソン・エデュケーション, 2001.

[2] W3C. *MathML* . http://www.w3.org/Math/.

[3] W3C. *XML* . http://www.w3.org/XML/.

[4] *JPedal* . http://www.jpedal.org/.

[5] *JDOM* . http://jdom.org/.

[6] *webMathematica* . http://www.wolfram.com/products/webmathematica/.