数理解析研究所講究録
第1568巻 2007年 176-181

InftyEditor with InftyReader – pdf2latex を目指して

藤本 光史
MITSUSHI FUJIMOTO
福岡教育大学
FUKUOKA UNIVERSITY OF EDUCATION*

1 はじめに

Adobe Systems 社によって考案された PDF(Portable Document Format) は、Web 上での配布用文書フォーマットとして広く利用されている。Acrobat 5.0 からは視覚障害者のためのアクセシビリティも考慮されていて [1]、スクリーンリーダによる読み上げにも対応している。しかし、数式については全く読み上げることができない。

2 次方程式の解の公式

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

例えば、上のような文書を Microsoft Office Word 2003 と LaTeX 2_ε で作成し、それぞれ Acrobat 7 と dvipdfmx で PDF 化したものからテキストを抽出すると以下のようになる。

```
2 次方程式の解の公式
a
x b b ac
2
- ± 2 - 4
=
```

図 1: Word 2003 + Acrobat 7

```
2 次方程式の解の公式
x = -b ± √ b2 - 4ac
2a
```

図 2: LaTeX 2_ε + dvipdfmx

このため、視覚障害者に数式の入った文書を読んでもらうためには、TeX ソースファイルを渡すのがよいとされている[1)]。しかし、Web 上で公開されている PDF ファイルのソースを入手することは困難である。また、運良く著者からソースファイルが提供されたとしても、TeX ソースに複雑なマクロやスタイルファイルが使用されていると、著者以外の者がソースを直接読むことは難しくなる。

これをクリアするには、PDF から複雑なマクロやスタイルファイルが使用されていないプレーンな TeX ソースが生成できればよい。本稿では、数式 OCR を利用して、いわゆる pdf2latex を実現するソフトウェア InftyReader について紹介する。

*fujimoto@fukuoka-edu.ac.jp

1) TeX ソースファイルにはページ情報がないので、晴眼者との整合性を保つためには、PDF ファイルと共に配布する方がよい。

2 数式OCRについて

数式OCRとは、理工系の分野における数式入り文書画像を対象にしたOCRのことで、画像中の文字や記号を認識すると同時に数式の構造を解析・認識して、最終的にXMLやLaTeXなどの数学記述言語に変換することを目的とする分野である。

岡本 [2, 3]、Fateman[4]、鈴木 [5, 6] らによって、実用に耐えうる数式OCRアルゴリズムが発表されている。また、行列認識においては、金堀 – 鈴木 [7]、Sexton – Sorge[8] の仕事がある。

数式OCRの重要な要素としては、数式領域の抽出、2次元構造解析、数学文字認識、接触文字／分離文字の認識などがあり、パターン認識研究の重要な一分野として、活発な研究が行われている。

3 数式OCRソフト InftyReader

InftyReader[9] は、Infty Project[10] で開発されている数式OCRソフトである。印刷文書を600dpi(または400dpi)の白黒2値[2)]でスキャンした画像を認識対象とし、認識結果をIML(Infty XML) / LaTeX 2_ε/ HTML / MathML / HrTEX に変換することが可能である。InftyReaderはコマンドライン版とGUI版(InftyReaderPlus)が用意されており、対応する画像形式はTIFF/BMP/GIF/PNGで、GUI版のみPDFファイルの読み込みにも対応している。よって、InftyReaderPlusを利用すれば、本稿の目的であるpdf2latexを実現することができる。

以下がInftyReaderPlusのユーザインターフェースである。

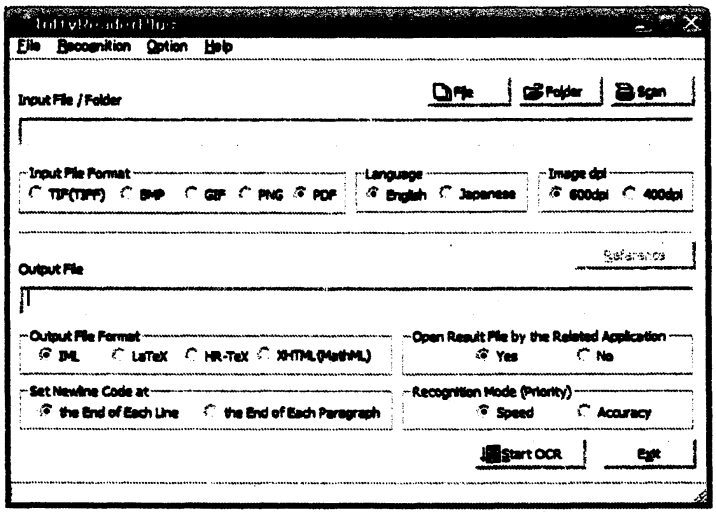

図 3: InftyReaderPlus のユーザインターフェース

[File]から認識させたいPDFファイルを選択し、[Output File Format]として[LaTeX]を選択し、Start OCRボタンをクリックすると認識作業が開始され、認識作業が完了すると自動的に結果が表示されるようになっている。以下は、InftyReaderPlusでPDFファイルを認識させる場合の動作の概略である。

1. PDFファイルの選択

---

2)カラー画像やグレースケール画像は非サポート。

2. 各ページの TIFF ファイルを生成[3)]
3. 各ページのテキスト情報を PDF ファイルから取得[4)]
4. TIFF ファイルからノイズを除去
5. 図領域、表領域、テキスト領域（数式を含む）に分類
6. テキスト領域から数式領域の切り出し
7. 非数式領域の認識[5)]
8. 数式領域の認識/構造解析
9. 表領域の認識
10. 認識結果を IML に出力
11. LaTeX/MathML などにコンバート
12. 認識結果のファイルに関連付けられたソフトを起動

PDF ファイルの認識では、PDF 文書内に格納されているテキスト情報 (Step 3) を参照しながら認識を行うため、テキスト情報を持つ PDF ファイルの認識では、通常の印刷文書を認識した場合と比較して、テキスト部の認識率が向上する。

Infty Project で開発された数式エディタ InftyEditor[11] がインストールされている場合、[Output File Format] として [IML] を選択すると、認識完了後、自動的に InftyEditor が起動し、認識結果を確認しながら編集することが可能である。以下は、InftyEditor で認識結果を開いた状態である。認識結果と原画像を同時に表示することが可能であり、InftyEditor 内のカーソル位置に合わせて原画像の表示位置も移動するようになっている。

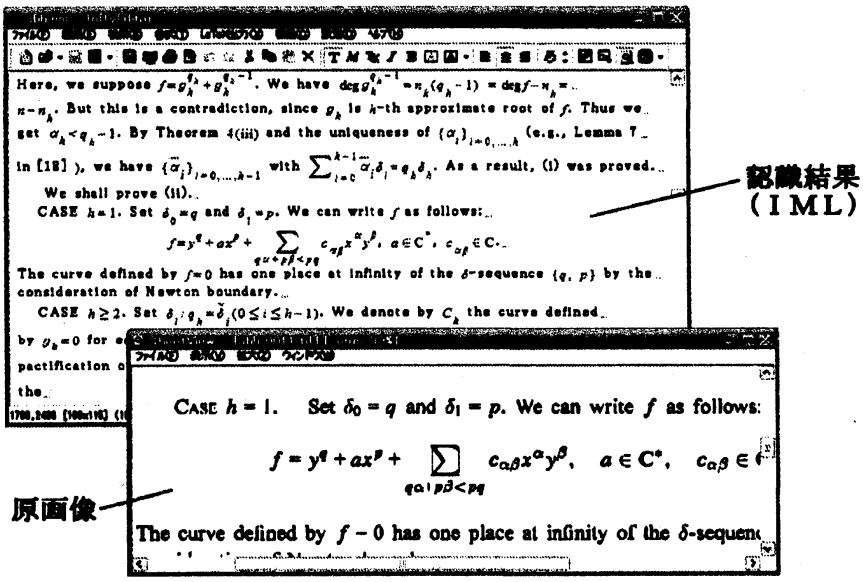

図 4: 認識結果の InftyEditor での確認・編集

[3)] Xpdf project の pdftoppm と、libtiff ツールの ppm2tiff を利用。
[4)] Xpdf project の pdftotext を利用。
[5)] 3 の結果と東芝から提供された OCR エンジンを利用。

4 行列と表の認識

InftyReader は行列の認識はもちろん、数式を含んだ表の認識も可能である。以下にその例を挙げる。

4.1 行列の認識例

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$
$$B = \begin{pmatrix} a_0 & & * \\ 0 & a_1 & \\ \vdots & & \ddots \\ 0 & & & a_n \end{pmatrix}$$

図 5: 原画像

$$A = \begin{pmatrix} a_{11} & \rightarrow & a_{12} \\ a_{21} & \rightarrow & a_{22} \end{pmatrix},$$
$$B = \begin{pmatrix} a_0 & \rightarrow & \rightarrow & * & \rightarrow \\ 0 & \rightarrow & a_1 & \rightarrow & \rightarrow \\ \vdots & \rightarrow & \rightarrow & \ddots & \rightarrow \\ 0 & \rightarrow & \rightarrow & \rightarrow & a_n \end{pmatrix}$$

図 6: 認識結果 (IML)

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$
$$B = \begin{pmatrix} a_0 & & * \\ 0 & a_1 & \\ \vdots & & \ddots \\ 0 & & & a_n \end{pmatrix}$$

図 7: 認識結果 (MathML)

4.2 数式を含む表の認識例

	Rep.1	Rep.2	Rep.3
Integration	$\int \log x\,dx$	$\int x' \log x\,dx$	$x\log x - \int x \cdot \frac{1}{x}\,dx$
	$\int \tan x\,dx$	$\int \frac{\sin x}{\cos x}\,dx$	$-\log\lvert\cos x\rvert$
Derivation	$\frac{d}{dx}\log x^2$	$\frac{(x^2)'}{x^2}$	$\frac{2}{x}$
	$\frac{d}{dx}\tan x$	$\frac{d}{dx}\frac{\sin x}{\cos x}$	$\frac{1}{\cos^2 x}$

図 8: 原画像

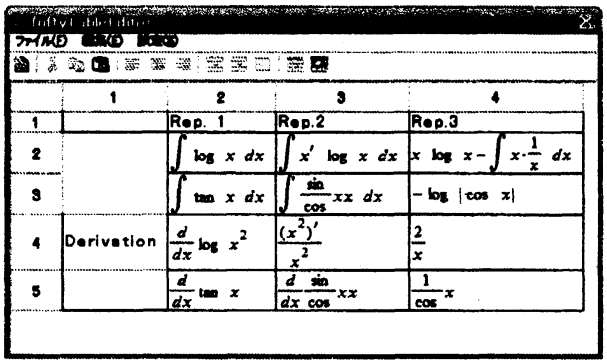

図 9: 認識結果 (IML)

数式を含む表の認識機能は現在ベータ版であり、図 9 のように誤認識がいくつか見られる。

5 アクセシビリティ

これまで見てきたように、InftyReader は数式が含まれた文書画像や PDF ファイルから LaTeX ソースを生成することが可能である。LaTeX ソースがあれば、スクリーンリーダを用いることで視覚障害者は数式文書にアクセスできる。しかし、LaTeX に精通していない視覚障害者にとっては、まだアクセシブルとは言えないだろう。そこで Infty Project では、HrTEX という簡易 TeX へのコンバートと IML 形式のファイルの読み上げソフトの開発という 2 つのアプローチを行った。

5.1 HrTEX

HrTEX とは、Human Readable TeX とも呼ばれている視覚障害者向けの TeX ライクな数式記述言語である [12]。University of Linz で策定されたもので、視覚障害者がスクリーンリーダで読むことを考慮されており、一種の簡易版 TeX と言える。

InftyReader では、認識結果を HrTEX 形式に出力できるようになっている。これによって、LaTeX に精通していない視覚障害者も数式文書にアクセスすることが可能となる。

5.2 ChattyInfty

ChattyInfty は、InftyEditor にスクリーンリーダ XPReader(95Reader) Ver6 (英語版は Microsoft 社の speechAPI) を用いた読み上げ機能を追加したソフトである [13]。マウスを使用せず、キーボードだけで利用することが可能であり、アルファベットの大文字や小文字なども区別する詳細読みモードと、それをしない簡易読みモードの切り替えが可能で、数式部とテキスト (非数式部) をトーンを切り替えて読み上げたり、分母・分子の読み上げ順を切り替えたりすることができる。

ChattyInfty は、音声で数式を読み上げるだけでなく、視覚障害者が音声を利用して数式を含む文章を作成・編集できる環境を提供する。開発メンバーには視覚障害者が含まれており、ソフトウェアの挙動についても視覚障害者の意見が反映された形で開発が進められている。

6 数学論文の電子化

2005 年末に、Infty Project のメンバーが中心となって、視覚障害者の科学技術情報へのアクセシビリティ向上を目的とする特定非営利活動法人サイエンス・アクセシビリティ・ネット [14] を設立した。そこでは InftyReader に関する研究成果を応用して、数式を含む書籍の点字化事業や、数学論文の電子化事業を行っている。現在までに、Funkcialj Ekvacioj (917 論文、約 1 万 5 千ページ) [15]、Advanced Studies in Pure Mathematics (758 論文、約 2 万ページ)、Journal of the Mathematical Society of Japan (311 論文、アブストラクトのみ)、Hokkaido Mathematical Journal (1,046 論文、16,093 ページ) や京都大学数理解析研究所講究録 [16] などの電子化を行っている。

数学論文は雑誌毎にフォントや体裁が異なっていることが普通である。そこで数学論文誌の電子化においては、最初にターゲットとなる数学論文誌から数本の論文の認識を行い、その雑誌の特徴を把握し、認識率が向上するよう InftyReader をカスタマイズした上で、自動処理を行っている。柔軟なカスタマイズが可能な点も InftyReader の特徴の一つである。

7 今後の課題

本稿では、pdf2latex の必要性について述べ、それを実現する数式 OCR ソフト InftyReader の概要を紹介した。そして、InftyReader は数式文書に対するアクセシビリティを向上させる可能性があること、数学論文の電子化に応用できることを示した。

今後の課題としては、カラーやグレースケールでキャプチャされた文書画像への対応、図 (線画) 認識機能、文書毎に認識調整を行うカスタマイズ機能などが挙げられる。図の認識は、現在の InftyReader でも行っているが、ビットマップ画像として認識され編集はできない。線画についてはベクター形式で認識し、InftyEditor で編集後、SVG や EMF フォーマットで出力できるようにしたい。

参 考 文 献

[1] 渡辺哲也, Web アクセシビリティ調査, http://www.nise.go.jp/research/kogaku/twatanab/WebAccess/WebAccessJp.html

[2] M.Okamoto and A.Miyazawa, An Experimental Implementation of a Document Recognition System for Papers Containing Mathematical Expressions, Proc. of SSPR90 (1990) 335–350.

[3] 岡本正行, 東 裕之, 記号のレイアウトに注目した数式構造認識, 電子情報通信学会論文誌, Vol.J78-D2, No.3 (1995) 474–482.

[4] R.J.Fateman, T.Tokuyasu, B.P.Berman and N.Mitchell, Optical Character Recognition and Parsing of Typeset Mathematics, Journal of Visual Communication and Image Representation, Vol.7, No.1 (1996) 2–15.

[5] K.Inoue, R.Miyazaki and M.Suzuki, Optical Recognition of Printed Matheml Documents, Proc. of ATCM98, Springer-Verlag (1998) 280–289.

[6] Y.Eto and M.Suzuki, Mathematical Formula Recognition Using Virtual Link Network, Proc. of ICDAR2001, IEEE Computer Society Press (2001) 430–437.

[7] T.Kanahori and M.Suzuki, A Recognition Method of Matrices by Using Variable Block Pattern Elements Generating ectangular Areas, Graphics Recognition – Algorithms and Applications, Lecture Notes in Computer Science 2390, Springer (2002) 320–329.

[8] A.P.Sexton and V.Sorge, Semantic Analysis of Matrix Structures, Proc. of ICDAR2005, IEEE Computer Society Press (2005) 1141–1145.

[9] InftyReader, http://www.sciaccess.net/jp/InftyReader/

[10] Infty Project, http://www.inftyproject.org/

[11] InftyEditor, http://www.sciaccess.net/jp/InftyEditor/

[12] F.Burger, M.Batusic, K.Miesenberger and B.Stöger, Access to Mathematics for the Blind – Defining HrTEX Standard, Proc. of ICCHP'96, ACM (1996) 609–616.

[13] ChattyInfty, http://www.sciaccess.net/jp/ChattyInfty/

[14] サイエンス・アクセシビリティ・ネット, http://www.sciaccess.net/

[15] 高山信毅, 鈴木昌和, Funkcialaj Ekvacioj の遡及電子化中間報告, 京都大学数理解析研究所講究録 1463, 「紀要の電子化と周辺の話題」 (2006) 21–27.

[16] 高橋安司, 京都大学数理解析研究所講究録の公開について, 「数学ジャーナルの電子化および電子化後における諸問題とその解決に向けて」報告集, 東京大学数理科学研究科 (2007) 28pages.