

レイアウト知識を用いた PDF 形式論文からの情報抽出システムの試作

富田陽明 大園忠親 新谷虎松

名古屋工業大学 知能情報システム学科

e-mail: {tomida, ozono, tora}@ics.nitech.ac.jp

1 はじめに

PDF 形式の論文のレイアウトを解析して分類する場合、内部のレイアウト表現に非構造的な記述が多く、機械的に解析をすることが難しい。そこで、論文のレイアウトを構造的な表現に再構築する手法をとる。このレイアウト表現に基づいて分類をする際、多くの訓練データを必要とし、分類に対する計算コストが増大する。本論文では、その論文のレイアウトに対する背景知識をあらかじめ与えることによってレイアウトのゆらぎを吸収し学習コストを軽減させ、少ない訓練データで分類のためのルールを生成し、それに基づいて分類と情報抽出を行う手法を提案する。本研究では本手法による試作システムを実装し、その実験結果を示して本手法の有用性を検証する。

2 論文ファイルからの情報抽出

PDF 形式で作られた文書より目的の情報を抽出する手法を考える。PDF 形式の文書を閲覧するソフトウェア Adobe Acrobat には、PDF の内容をテキストデータとして出力する機能が備わっている。プレーンなテキストで PDF の内容を出力することや、スタイルシートを用いて表示の再現性を重視した出力形式などがある。プレーンテキストでの出力では文書の構造がわからず、情報抽出に支障が出るため文書の構造をタグで表した XML 形式で出力させる。しかし、文書の見た目は同じでも扱う PDF ファイルによりタグ構造が異なる場合が存在するので、タグ構造を利用した XML Wrapper を制作しても PDF ファイルによっては情報抽出が不可能となる。

そこで、PDF ファイル内部の構造に相違があっても視覚的なレイアウトに大きな相違がないことに注目し、視覚的レイアウトの情報から文書のレイアウト構造を再構築する。構築されたレイアウト毎に抽出テンプレートを作成し、各 PDF 論文をそれらのテンプレートにマッチするように分類する。一般的に文書の分類を行う

Implementing Information Extracting System from Papers with PDF Technology using Layout Knowledge
Haruaki Tomida, Tadachika OZONO, Toramatsu SHINTANI

Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya 466-8555 JAPAN

場合、classifier を作成する際に訓練データが大量に必要となり、その計算コストが大きくなる。しかし、論文のレイアウトによる分類という絞ったドメインにおいては、その論文のレイアウトに対する背景知識を予め与えておけば、その背景知識を見つけるための計算を減らすことができ、その結果用意する訓練データも減らすことができると考えられる。本論文では、classifier の作成の際に予め背景知識を与え、少ない訓練データで分類をするシステムを試作する。

3 レイアウト知識に基づいた classifier の作成と情報抽出

本論文において classifier とは、PDF 論文ファイルからの情報抽出を行う際のテンプレートとしての働きをするオブジェクトである。classifier は attribute と relation という 2 つの情報を持つ。attribute はそのレイアウト中に存在する矩形領域（文字列、画像が収まっている領域）に関する知識の集合であり、その矩形領域のラベル、大きさ、水平位置寄せ、内包する文字列、抽出の可否が記述される。relation は矩形領域同士の位置関係を記した知識の集合であり、矩形同士の水平位置関係、垂直位置関係が記述される。relation を図に書き出すと、縦筋を中心としてそこから横に枝が伸びる、葉脈のような図になる。

classifier 作成時に与える初期知識には、訓練データに共通していそうな attribute と relation を記述する。初期知識の記述例を図 1 で示す。relation について、Left は author1 が author2 より左側にある、Upper は source が title より上側にある、ということを表している。attribute について、TextLabel は位置寄せが左側で、文字フォントの種類と高さが任意であり、その領域内

```
//relation
Left(author1,author2)
Upper(source,title)

//attribute
TextLabel(abst_label,left,*,*,Abstract)
```

図 1: 与える初期知識の例

```
Comform(抽出対象 PDF 論文レイアウト  $l$ , classifier  $c$ )
```

```
if  $c$  の各 attribute のラベルが張り終わっている:
```

```
if  $l$  と  $c$  の relation が整合している:
```

```
return  $l$ ;
```

```
else:
```

```
return null;
```

```
foreach  $c$  の各 attribute  $attr$ :
```

```
foreach  $l$  の各矩形領域  $r$ :
```

```
if  $p$  と  $r$  がマッチする:
```

```
 $l_{clone} = p$  の複製;
```

```
 $l$  のラベルを  $r$  のラベルに張り替
```

```
える;
```

```
 $l_{next} =$ 
```

```
Comform( $l_{clone}$ ,  $c$ );
```

```
if  $l_{next} \neq$ 
```

```
null;
```

```
return
```

```
 $l_{next}$ ;
```

```
end if; end foreach;
```

```
 $l =$  Comform( $l$ ,  $c$ ); end foreach;
```

```
return  $l$ ;
```

図 2: レイアウトと classifier の照合アルゴリズム

の文字列が "Abstract" である矩形領域に "abst_label" というラベルを貼る, ということを表している. 訓練データの内, この初期知識にマッチするものを正事例とし, マッチしないものを負事例とみなす.

まず, 訓練データのクラス分けを行う. 訓練データとなる PDF 論文ファイルのレイアウト構造を解析し, その概形をユーザに提示し, 抽出対象となる矩形領域にラベルを付加する. その後, 獲得したレイアウトと初期知識とのマッチングを行い, 正事例と考えられるレイアウトであればそのレイアウトに初期知識を付加し, attribute と relation を解析する. この作業を全ての訓練データに対して行う.

訓練データの解析が全て終わったら, classifier の作成をする. 各訓練データのレイアウトを比較して, 共通する部分を探し, その attribute と relation を classifier に格納する. classifier は分類するレイアウトの形式毎に作成する.

情報抽出は, 作成した classifier に基づいて行う. 抽出対象の PDF ファイルのレイアウトと classifier を照合し, そのレイアウトとマッチングする classifier を探す. 図 2 のようなアルゴリズムを用いて, マッチングしたらその classifier の attribute の情報に基づき抽出すべき領域を決定し, 抽出結果を出力する.

	1 classifier	3 classifiers
適合率	0.683	0.823/0.892/0.627 ²
再現率	0.952	0.952/0.812/0.372
F 尺度	0.795	0.882/0.888/0.466

表 1: 抽出実験結果

4 評価実験

本論文では, ScienceDirect¹ で閲覧できる Artificial Intelligence 誌 (以下 AI 誌) のうち 1999 年から 2002 年にかけて掲載された PDF ファイル 273 本について抽出実験を行った. このファイルの中には, 論文ファイルの他に, 論説のファイルや書籍のレビューが書かれたファイルなど, 論文以外のデータが 34 本含まれている. 本研究では, それぞれの PDF 論文ファイルについて, 表題, 著者, 論旨の 3 つの項目を抽出する. classifier 作成にあたっては, 各 5 ファイルずつの訓練データと, "Abstract" と書かれた矩形領域, また表題と著者の位置関係を記した初期知識ファイルを用いる.

実験では, 著者の矩形領域が 1 つである classifier のみを作成した抽出と, 著者の矩形領域がそれぞれ 1, 2, 3 個ある classifier をそれぞれ作成した抽出について, 適合率 (Precision), 再現率 (Recall), $\alpha = 0.5$ とする F 尺度 (F measure) を用いた. その結果を, 表 1 に示す. 少ない訓練データ数にも関わらず高い F 尺度値を提示している. また, classifier の数が多いほど精度が高くなると思われる.

5 おわりに

本論文では, PDF 論文ファイルの分類のためにあらかじめ背景知識を与えることにより少ない訓練データで分類が可能であることを示した.

参考文献

- [1] Esposito, F., Malerba, D., Lisi, A., F.: Machine Learning for Intelligent Processing of Printed Documents, 2000.
- [2] 富田陽明, 大園忠親, 新谷虎松: "ルールに基づく PDF 形式の論文からの情報抽出," 平成 14 年度電気関係学会東海支部連合大会講演論文集, pp.259, 電気関係学会東海支部連合大会, 2002 年 9 月, ,

¹ <http://www.sciencedirect.com>

² 左から, 著者の領域がそれぞれ 1, 2, 3 個の場合