レイアウト情報とテキスト情報を用いた学術論文の構造化

杉木 健二*，松原 茂樹，吉川 正俊 (名古屋大学)

XML Conversion of Electronic Technical Papers using Layout Information and Text Information

Kenji Sugiki, Shigeki Matsubara, Masatoshi Yoshikawa (Nagoya University)

1 はじめに

電子ジャーナルや機関リポジトリなど，近年，学術論文の電子的な流通環境が整備されつつある．しかし，その多くは PDF や PS 形式で論文を保存しており，検索や抽出など利用環境の高度化には限界がある．

そこで本稿では，PDF 形式の学術論文を XML 形式に変換する手法を提案する．レイアウト情報とテキスト情報をもとに作成した構造化ルールを用いて論文構造タグを付与する．国際会議録を用いて学術論文の構造化実験を実施した．

2 学術論文の構造化

本手法では，PDF 文書を，構造情報が付与された XML 文書に変換することにより，学術論文の構造化を実現する．本原稿の変換により生成される XML 文書の例を図 1 に示す．

本手法の流れは以下の通りである．

1. PDF 形式の論文を，テキストの各行の位置情報とフォント情報を含むテキストデータに変換する．
2. レイアウト情報とテキスト情報に基づく構造化ルールを適用し，表題，節，段落などの構造タグを付与する．

2.1 レイアウト情報とテキスト情報

レイアウト情報とは，論文におけるそれぞれのテキスト位置やフォントサイズであり，テキスト情報とは，それぞれのテキスト中に含まれる特定のパターンや正規表現である．本研究では，これらの情報を組み合わせてルールを作成する．ルールの例を以下に示す．

- 文字列 t のフォントサイズが 18pt 以上，かつ，1 ページ目の冒頭に出現するならば，t はタイトルである．
- 文字列 t が，パターン *Introduction* にマッチすれば，t は第 1 節のタイトルである．
- 文字列 t のフォントサイズが第 1 節のタイトルと同一であり，かつ，正規表現“ $[1-9]¥.?_*[A-Z1-9][a-zA-Z1-9]+$ ”にマッチすれば，t は節のタイトルである．

2.2 タグ付け

作成したルールを用いて，論文にタグを付与する．本研究では，表題，著者，所属，E メール，抄録，節，段落，引用，リスト，表，図，注釈，文献をタグ付けの対象とした．

3 評価実験

本手法による XML 変換の性能を評価するために、学術論文の構造化実験を行った．実験データとして VLDB 2004[1] の論文 30 文書を用いた．テストデータへのタグ付けには人手で作成した 30 個のヒューリスティックルールを使用した．PDF ファイルのテキストデータへの変換には，pdftohtml[2] の XML 出力機能を用いた．評価は，正解の XML 文書を作成し，各タグの精度および再現率により行った．

```
<conference-paper filename="toukaishibu2005.xml">
  <head>
      <title>レイアウト情報とテキスト情報を用いた学術論文の構造化</title>
      <title-en>XML Construction of for Electronic Technical Papers using Layout Information
                and Text Information</title-en>
      <author>杉木 健二</author>
      <author>松原 茂樹</author>
      <author>吉川 正俊</author>
      <affiliation>名古屋大学</affiliation>
      <author-en>Kenji Sugiki</author-en>
      <author-en>Shigeki Matsubara</author-en>
      <author-en>Masatoshi Yoshikawa</autho-en>
      <affiliation-en>Nagoya University</affiliation-en>
  </head>
  <body>
    <section>
      <st>はじめに</st>
      <p>電子ジャーナルや機関リポジトリなど，近年，学術論文の電子的な流通環境が整備されつつある．
         しかし，その多くは PDF や PS 形式で論文を保存しており，検索や抽出など利用環境の高度化には
         限界がある．</p>
      <p>そこで本稿では，PDF 形式の学術論文を XML 形式に変換する手法を提案する．レイアウト情報と
         テキスト情報をもとに作成した構造化ルールを用いて論文構造タグを付与する．国際会議録を
         用いて学術論文の構造化実験を実施した．</p>
    </section>
    <section>
      <st>学術論文の構造化</st>
      <p>本手法では，PDF 文書を，構造情報が付与された XML 文書に変換することにより，学術論文の
         構造化を実現する．本原稿を変換により結果生成される XML 文書の例を図 1 に示す．</p>
      <p>本手法の流れは以下の通りである．
          <ol>
            <li>PDF 形式の論文を，テキストの各行の位置情報とフォント情報を含むテキストデータに
                変換する．</li>
            <li>レイアウト情報とテキスト情報に基づく構造化ルールを適用し，表題，節，段落などの
                構造タグを付与する．</li>
          </ol>
      </p>
      <subsection>
        <sst>レイアウト情報とテキスト情報</sst>
    ....
```

Fig. 1: 本原稿の変換結果の例

Table 1: VLDB2004(30 文書) を用いた構造化実験の結果

精度 (%)	再現率 (%)
98.0(1535/1566)	82.8(1535/1853)

$$\text{精度 (\%)} = \frac{\text{正解タグ数}}{\text{付与したタグ数}} \times 100$$

$$\text{再現率 (\%)} = \frac{\text{正解タグ数}}{\text{付与すべきタグ数}} \times 100$$

実験結果を表 1 に示す．精度，再現率ともに高く，PDF 形式の論文の構造化による XML 文書への変換可能性を確認した．

4 おわりに

本稿では，PDF 形式の学術論文を XML 形式に変換する手法を提案した．評価実験により，PDF 形式の論文の構造化による XML 文書への変換可能性を確認した．今後は，多くの論文集を対象とした，より汎用的なシステムを作成する予定である．

文献

(1) VLDB http://www.informatik.uni-trier.de/ ley/db/conf/vldb/

(2) pdftohtml http://pdftohtml.sourceforge.net/