PDF 문서에 대한 XML 변환 소프트웨어 구현

*박민규 **정주용 ***김용한 ****박재홍

서울시립대학교 ****㈜넷앤티비

***yhkim@uos.ac.kr

An implementation of PDF-to-XML conversion software

*Park, MinKyu **Jeong, JuYong ***Kim, Yong Han ****Park, Jaehong

The University of Seoul ****Net&TV, Inc.

요약

PDF는 미국 어도비 시스템즈(Adobe Systems) 사에서 만든 전자 문서 파일 포맷이다. PDF에는 일반 문서 및 문자, 도형 그림, 글꼴 등을 포함할 수 있으며, 동일한 문서를 컴퓨터 운영체제와 관계없이 모니터, 프린터 등의 출력 장치로 같은 모습으로 출력하기 위한 목적으로 개발되었다. 현재 거의 모든 운영체제에서 PDF 문서를 읽거나 인쇄할 수 있으며, 온라인 및 오프라인 환경에서도 쉽게 PDF 문서를 공유할 수 있으며, 보안성이 높아 공공기관, 연구소 등에서 자료를 배포할 때 많이 사용되고 있다. XML은 사람이 쉽게 읽을 수 있고 컴퓨터와 같은 기계가 쉽게 판독할 수 있는 형식으로 부호화된 문서를 작성하기 위한 일련의 규칙을 정의하는 데 사용하는 마크업 언어로서 W3C에서 정의하였다. 현재 XML은 인터넷 상에서 데이터를 표현하거나 교환, 저장, 검색할 때 가장 널리 쓰이고 있다. 본 논문에서는 PDF 문서를 인터넷 상에서 쉽게 활용하도록 도와주며, XML 데이터 처리가 가능한 응용 프로그램에서 PDF 문서를 활용할 때 쉽게 사용할 수 있도록 도와주는 PDF - to-XML 변환 소프트웨어를 구현하였다.

1. 서론

PDF(Portable Document Format)[1]는 1993년에 미국의 어도비 시스템즈 사가 어크로뱃(Acrobat) 1.0 버전과 함께 발표하면서 등장하였다. 처음 발표 당시에는 어크로뱃 리더(Acrobat Reader)는 유료였다. 하지만 후에 이를 무료로 배포하고 PDF 포맷을 적극적으로 지원하면서 현재 온라인 상의 인쇄용 문서의 실질적 산업 표준으로 자리 잡게 되었다. 특히 컴퓨터 운영체제와 관계없이 동일한 PDF 문서에 대해 같은 모습으로 출력되기 때문에 문서를 배포할 때에도 널리 사용되고 있다. PDF 문서 형식은 공개되어 있으며, 지난 2008년 7월에 ISO에 의해 국제 표준(ISO 32000-1:2008)[1]으로 발간되었다.

XML(Extensible Markup Language)[2]은 HTML(Hypertext Markup Language)과 같은 마크업 언어로서 1996년 W3C(World Wide Web Consortium)에 처음 제안된 후, W3C는 1998년 2월에 차세대 인터넷 표준 언어로서 XML 1.0 표준을 발표하고 권장 규격으로 채택하였다. XML은 웹 문서에 쉽게 응용하기 위해 설계되었지만, 배우기 쉽고, 응용 프로그램에서도 쉽게 구현될 수 있도록 설계되었기 때문에 현재는 웹뿐만 아니라 전자상거래, 전자문서 교환, 검색 등과 같은 다양한 분야에서 활용되고 있다.

본 논문에서는 PDF 문서를 온라인 상에서 활용하는 데 도움을 주고, 응용 프로그램에서 PDF 문서를 쉽게 분석하고 활용하는 데 도움을 줄 수 있도록 하는 PDF - to-XML 변환 소프트웨어를 구현하였다.

※본 논문은 중소기업청에서 지원하는 2014년도 산학연 협력사업(No. C0248831)의 연구수행으로 인한 결과물임을 밝힙니다.

본 논문의 구성은 다음과 같다. 2장에서는 PDF 문서 구조에 대해 간단히 살펴본 후, 3장에서 본 논문의 변환 소프트웨어에서 변환할 PDF 오브젝트(object)의 범위 및 XML 변환 방법에 대해서 설명한다. 4장에서는 본 논문의 PDF - to-XML 변환 소프트웨어 구현에 대해서 설명하고, 5장에서는 변환 소프트웨어에 대해 실험을 진행하여 성능을 확인한다. 마지막으로 6장에서는 본 논문에 대한 결론을 맺는다.

2. PDF 문서의 구조

PDF 파일은 <그림 1>과 같이 크게 4가지 요소들로 구성된다. PDF 규격의 버전을 알리는 헤더(header), PDF 문서를 구성하는 PDF 오브젝트를 포함하는 바디(body), 간접 PDF 오브젝트에 관한 정보를 가지고 있는 상호 참조표(cross-reference table), 그리고 상호 참조 위치와 바디 내 특정 오브젝트의 위치 정보를 가지고 있는 트레일러(trailer)로 이루어진다. 이후 파일이 업데이트 될 경우, 추가적인 요소가 파일 끝에 붙으면서 수정될 수 있다.

PDF 문서는 PDF 오브젝트들로 구성된다. PDF 오브젝트에는 페이지 오브젝트, 디렉토리(dictionary) 오브젝트, 스트림(stream) 오브젝트, 폰트(font) 오브젝트, 이미지(image) 오브젝트 등이 있다.

PDF 문서는 PDF 파일의 바디 요소 안의 오브젝트들에 대한 계층적 구조체로 생각할 수 있다. 바디 요소는 카탈로그(catalog) 오브젝트로 표현되며, 페이지 오브젝트들이 이곳에 포함된다. <그림 2>는 카탈로그 오브젝트의 계층을 나타낸다. 페이지 오브젝트는 해당 페이지를 실질적으로 구성하는 콘텐츠(contents) 오브젝트, 리소스(resources)

오브젝트 등으로 구성된다. 리소스 오브젝트에는 폰트 오브젝트, XObject 등이 포함된다. XObject에는 이미지 오브젝트, 폼(Form) 오브젝트 등이 포함된다. 콘텐츠 오브젝트는 실제 화면에 렌더링될 내용에 대한 스트림 오브젝트로 구성되며 스크립트(script)가 이에 포함된다. 이 스크립트는 리소스 오브젝트에 포함되어 있는 오브젝트들과 연계된다. PDF 문서는 실제로 이 스크립트를 통해 화면에 출력되며, 스크립트에는 그래픽 요소들이 기능적으로 정의된다. 더 자세한 내용에 대해서는 PDF 표준 문서[1]를 참조하도록 한다.

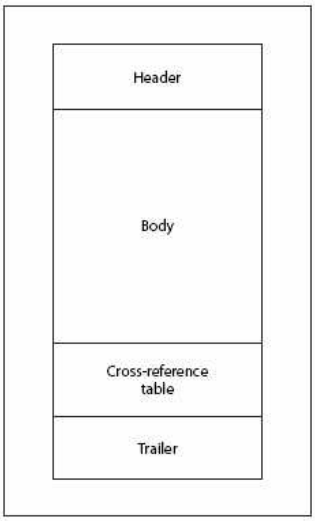

그림 1. PDF 파일의 구성

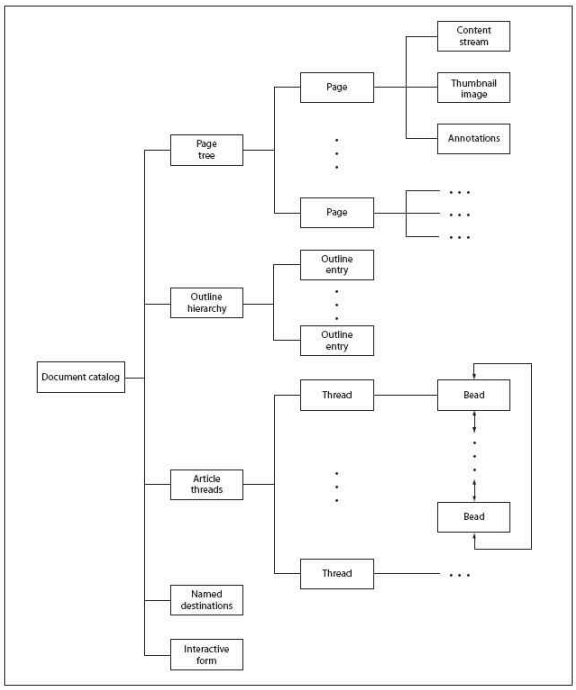

그림 2. PDF 문서의 카탈로그 오브젝트의 계층 구조

3. 변환할 PDF 오브젝트 범위 및 XML 변환 방법

본 논문의 변환 소프트웨어는 PDF 문서의 대부분의 그래픽 요소들을 XML로 변환하도록 구현하였다. 그래픽 요소에는 텍스트(text), 벡터 그래픽스(vector graphics), 이미지, 멀티미디어(multimedia) 콘텐츠 등이 있다. 본 논문에서는 비디오, 오디오 등 멀티미디어 콘텐츠에 대한 변환은 제외하였다. 더불어 PDF 문서 내의 문서 구조에 표(table) 객체가 포함되어 있다면, 해당하는 표의 내용을 단순 그래픽으로 출력하지 않고, 이에 맞게 표를 XML로 출력한다.

변환 소프트웨어가 출력하는 XML 문서의 형식은 그래픽 요소들을 화면에 가장 잘 표현할 수 있고 웹 브라우저가 해석 가능한 XHTML[3] 스키마(schema)를 따르도록 하였다.

PDF의 벡터 그래픽스 요소는 현재 거의 모든 웹 브라우저가 해석 가능한 SVG(Scalable Vector Graphics) 스키마를 따르도록 XML로 변환하여 XHTML 문서에 포함하였다. PDF에 내장되어 있는 이미지 스트림은 파일로서 저장한다. 이 이미지 파일에 대한 정보는 XHTML의 img 태그(tag)로 변환하였으며, 텍스트는 단어 단위로 XHTML의 span 태그로 변환하였다. 마지막으로 표 객체는 XHTML의 table 태그로 변환하였다.

변환되는 모든 PDF 객체들의 위치 및 모양 정보는 스타일(style) 속성으로 설정하였다. 만약 텍스트에서 사용하는 폰트가 PDF 문서 내에 PDF 문서에 내장된 폰트 스트림을 사용할 경우, 이를 파일로 저장한 후 이에 대한 정보를 스타일 객체의 @font-face로 변환하였다.

<그림 3>은 본 논문의 PDF - to-XML 변환 소프트웨어를 이용하여 샘플 PDF 문서를 위에 설명한 방법대로 XHTML 스키마를 따르는 XML로 변환한 소스 내용의 일부이다.

```
▼<html xmlns="http://www.w3.org/1999/xhtml">
  ▶<head>…</head>
  ▼<body style="width:1280px;height:720px;border:1px solid;margin:0px;">
    ▼<div name="layer">
      ▼<div name="layer">
        ▼<div name="svg_area" style="position:absolute;">
          ▶<svg width="1280.000000" height="720.000000">…</svg>
        </div>
      </div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
      ▶<div name="layer">…</div>
    </div>
    ▼<div id="p2_b1" name="block" style="position:absolute;left:223px;top:115.283px;">
      ▼<div id="p2_t1" name="line" style="position:absolute;left:0px;top:0px;">
        <span style="position:absolute;font-family:Arial;font-size:24px;color:#000000;width:210px;height:
        22.2px;letter-spacing:-0.054px;left:0px;top:0px;" id="p2_w1" name="word">
                    Abcde
                    </span>
      </div>
    </div>
    ▼<div id="p2_b2" name="block" style="position:absolute;left:223px;top:171.779px;">
      ▼<div id="p2_t2" name="line" style="position:absolute;left:0px;top:0px;">
        <span style="position:absolute;font-family:Arial;font-size:24px;color:#000000;width:169px;height:
        22.2px;letter-spacing:0px;left:0px;top:0px;" id="p2_w2" name="word">
                    Abcd
                    </span>
      </div>
    </div>
    ▶<div id="p2_b3" name="block" style="position:absolute;left:223px;top:229.379px;">…</div>
    ▼<div id="p2_b4" name="block" style="position:absolute;left:578px;top:303.995px;">
      ▼<div id="p2_t4" name="line" style="position:absolute;left:0px;top:0px;">
        <span style="position:absolute;font-family:Arial;font-size:24px;color:#000000;width:219px;height:
        22.2px;letter-spacing:-0.054px;left:0px;top:0px;" id="p2_w4" name="word">
                    abcdef
                    </span>
      </div>
    </div>
  </body>
</html>
```

그림 3. PDF - to-XML 소프트웨어의 출력 XML 소스 예

4. PDF - to-XML 변환 소프트웨어 구현

본 논문의 PDF - to - XML 변환 소프트웨어는 오픈소스 라이브러리인 Xpdf[4](버전 3.0.1)을 활용하여 작성되었으며, 개발에 사용된 운영체제는 마이크로소프트 윈도우즈 7이고, 프로그램 개발 도구는 마이크로소프트 비주얼 스튜디오 2012이다.

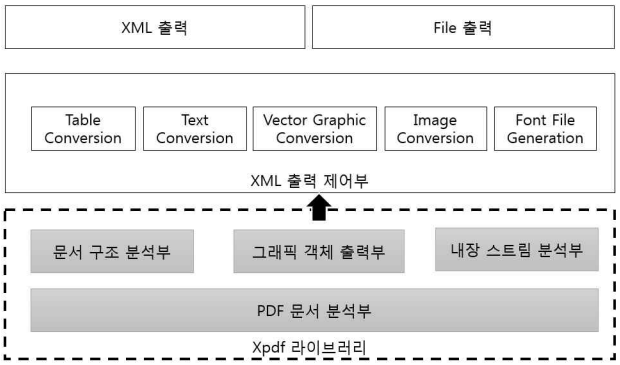

그림 4. PDF - to - XML 변환 소프트웨어의 구조

<그림 4>는 본 논문에서 구현한 PDF - to - XML 변환 소프트웨어의 구조를 나타낸다. 입력된 PDF 문서 파일은 먼저 Xpdf 라이브러리의 PDF 문서 분석부에서 분석되며, 페이지 순서에 따라 페이지 출력 시작 이벤트(event)가 발생한다. 이벤트 발생 후 페이지에 대한 PDF 콘텐츠 객체의 스트림 객체로부터 그래픽 객체 출력부를 통해서 정보가 출력된다. 이를 PDF - to - XML 변환 소프트웨어의 XML 출력 제어부에서 각 그래픽 객체에 맞는 XML 변환부를 통해 XML로 변환한다.

텍스트 객체를 XML로 변환 시 사용하는 폰트가 PDF에 내장된 폰트 스트림을 사용할 경우, 폰트 파일 생성부에서 Xpdf 라이브러리의 내장 스트림 분석부를 통해 폰트 파일을 생성하고 이에 대한 정보를 XML로 저장한다.

Xpdf 라이브러리에서 페이지 출력 종료 이벤트가 발생할 때마다 XML 출력 제어부는 문서 구조 분석부를 통해 표가 존재하는지 확인하고, 만약 존재한다면 표 변환부를 통해 XML로 변환하고, 해당 영역에 대해서 벡터 그래픽스 요소를 제거한다. 최종적으로 XHTML 스키마를 따르도록 XML로 구성하고 이를 파일로 출력한다.

5. 실험 및 성능 확인

본 장에서는 본 논문의 PDF - to-XML 변환 소프트웨어에 대해 실험을 진행하고 성능을 확인한다. 출력 결과물인 XHTML 스키마를 따르는 XML 문서를 웹 브라우저를 통해 출력하고, 이를 원본 PDF와 눈으로 비교함으로써 성능을 확인한다.

<그림 5>와 같이 PDF - to-XML 변환 소프트웨어는 입력된 PDF의 페이지 수만큼의 XML 문서 파일과 각 페이지에 포함된 이미지나 폰트 파일을 지정된 폴더 안으로 출력한다.

<그림 6>은 실험에 사용된 샘플 PDF 문서들 중 하나를 어도비 리더(Adobe Reader)로 출력한 모습이며, <그림 7>은 이 샘플 PDF를 본 논문의 PDF-toXML 변환 소프트웨어를 이용하여 출력한 결과를 담은 XML 문서를 크롬 웹 브라우저를 통해 확인한 모습이다. 결과에서 확인할 수 있듯이 변환된 XML 문서가 원본 PDF 문서와 거의 같은 형태로 크롬 웹 브라우저를 통해 출력되었다. <그림 7>의 오른쪽 부분은 크롬 웹 브라우저의 '소스 보기' 기능을 사용해 현재 보고 있는 페이지에 해당하는 XML 문서의 내용을 나타낸 것이다.

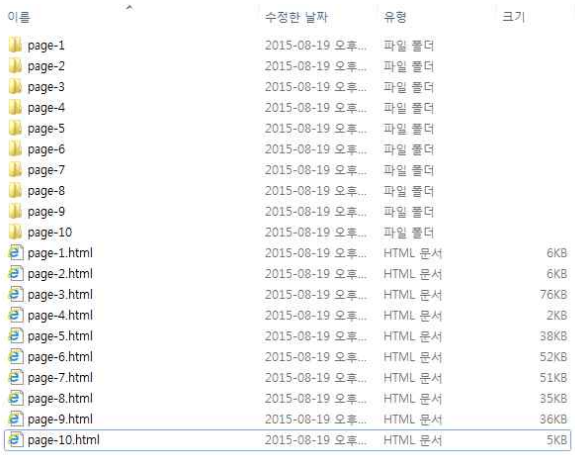

그림 5. PDF - to - XML 변환 소프트웨어의 결과 출력 예

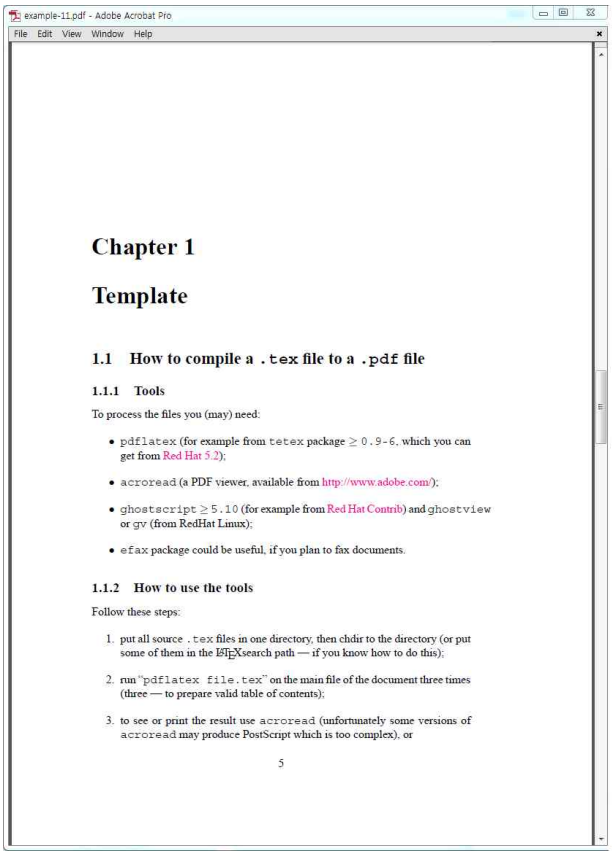

그림 6. 샘플 PDF를 Adobe Reader로 출력한 모습

<그림 8>은 실험을 위해 마이크로소프트사의 파워포인트 2012를 이용하여 벡터 그래픽스 요소, 텍스트, 이미지, 표 객체 등 다양한 객체들이 함께 한 페이지 내에 존재하도록 시험용 문서를 작성한 후에 PDF로 변환하여 샘플 PDF 문서를 얻고 이를 활용하여 PDF-to-XML 변환 소프트웨어의 성능을 실험한 화면이다. 실험 결과 원본 PDF와 거의 같은 형태로 출력됨을 확인하였다.

PDF 표준에서는 표에 대한 정의가 문서 구조로서 정의되는데 실제로 표가 그려지는 부분은 벡터 그래픽스로 분리되어 그려진다. 그렇기 때문에 해당 영역이 표 영역이라는 것을 알고 있더라도 다른 벡터

그래픽스 요소가 그 영역에 포함되어 있으면 표를 인식하는 것이 거의 불가능하다. 또한 표에 배경색이나 선의 두께, 모양 등과 같이 스타일 요소가 적용되어 있다면, 이를 고려하여 표를 추출해야 하는데 이를 위해서 복잡한 그래픽스 분석 알고리듬이 필요하다. 그래서 본 논문에서는 표 위에 텍스트와 이미지 객체 외의 벡터 그래픽스 요소가 오버레이(overlay)되어 있을 경우나 표에 스타일이 적용되어 있을 경우에는 표로서 인식하지 않고 벡터 그래픽스로 인식하여 XML로 출력하도록 구현하였다.

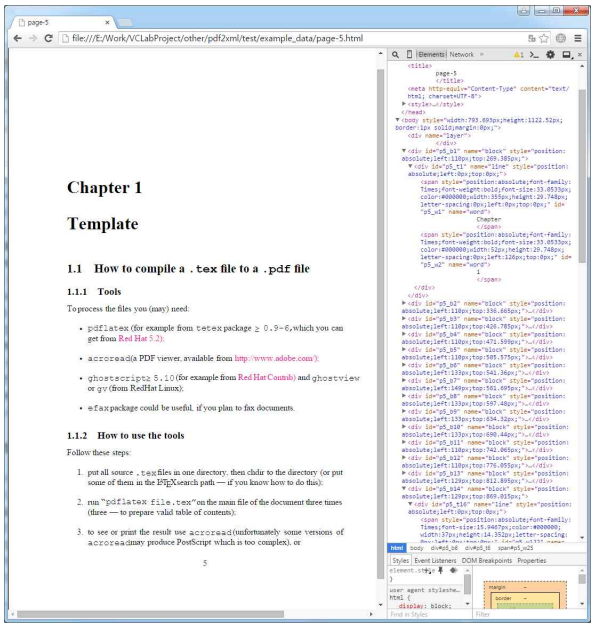

그림 7. 변환된 XML 문서를 크롬 웹 브라우저로 출력한 모습

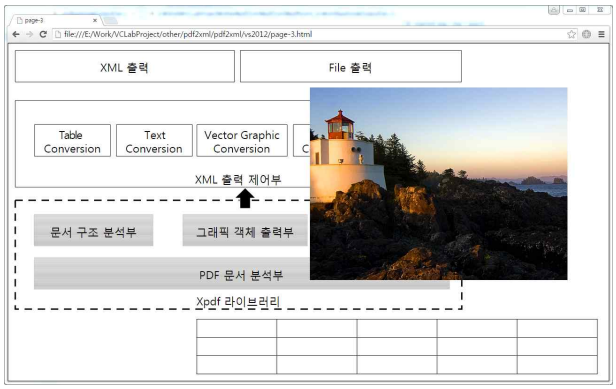

그림 8. 다양한 객체들이 함께 존재하는 샘플 PDF 실험 모습

6. 결론

본 논문에서는 Xpdf 오픈 소스 라이브러리를 활용하여 PDF-to-XML 변환 소프트웨어를 구현하였다. 구현한 소프트웨어를 이용하여 여러 PDF 문서들을 XML 문서로 변환하였고, 변환된 XML 문서를 웹 브라우저를 통해 화면에 출력한 결과를 원본 PDF 문서를 어도비 리더로 화면에 출력한 결과와 비교함으로써 성능을 확인하였다. 문서 저작, 검색 등의 분야에서 본 논문에서 개발한 PDF-to-XML 변환 소프트웨어를 활용할 수 있을 것이다.

참고문헌

[1] ISO/IEC 32000-1, Document management - Portable document format - Part1: PDF 1.7, International Organization for Standardization, 2008.

[2] World Wide Web Consortium, Extensible Markup Language (XML) 1.0 (Fifth Edition), http://www.w3.org/TR/2008/REC-xml-20081126, 2008.

[3] World Wide Web Consortium, XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition), http://www.w3.org/TR/xhtml1/, 2002.

[4] http://www.foolabs.com/xpdf.