PDF문서를 EPUB3.0 포맷으로 변환을 위한 효과적 색 추출 및 상호작용 효과삽입기법

이남희*, 김강석*, 김재훈*, 변계섭**
*아주대학교 대학원 지식정보공학과
**주식회사 오렌지디지트코리아
e-mail : {namsseng0, kangskim, jaikim}@ajou.ac.kr, louis@orangedigit.com

An effective color extraction and interactive insertion technique for converting PDF documents to EPUB3.0 format

Namhui Lee*, Kangseok Kim*, Jai-Hoon Kim*, Louis Byun**
*Dept. of Knowledge Information Engineering, Ajou University
**Orange Digit Korea Inc.

요 약

기존 책 문서인 PDF 문서를 전자책에서도 억세스 할 수 있도록 전자책의 표준 형태로 변환하는 과정이 필요하다. PDF 문서를 전자책의 대표적인 표준 형태인 EPUB3.0 으로 변환할 때, 인쇄 색상 표현방법인 CMYK를 디지털 색상 RGB 형태로 변환하는데 형태의 차이로 인하여 색감이 제대로 변환되지 못하는 문제점이 있다. 본 연구에서는 변환 시 색감을 잃지 않도록 ICC 프로파일을 이용한 변환 연구를 수행하였다. 또한 전자책 독자들을 위한 상호 작용적인 시각적인 효과를 제공하기 위하여, 많은 부분의 텍스트 중 특정 부분을 인식하여 효과 코드를 넣는 알고리즘을 제안하였다.

1. 서론

출판 시장의 흐름은 종전에 종이에 인쇄하여 책과 같은 형태의 출판물을 가지고 다니던 시대에서 벗어나, 최근 IT 기술과 통신기술의 발달로 개인 스마트폰이나 태블릿, 노트북 등을 이용해 앱 스토아 같은 북 스토아에서 온라인 디지털 형태로 구매해 읽는 추세로 바뀌어 가고 있다. 이러한 흐름에 국제전자출판포럼(IDPF : International Digital Publishing Forum)에서는 전자책 표준화작업을 진행하고 있고 2012년 EPUB 3.0 표준을 제정하였다[1]. 이러하여 각 전자책 출판 회사들이 속속 나타나고 있으며 시장은 확대되고 있다.

그러나 현재까지 출판된 책들을 다시 전자책으로 만드는 과정은 PDF나 종이로 된 책을 사람이 수작업을 통하여 제작하고 있다. 그렇기 때문에 기존에 책을 제작했던 작업 시간과 인건비용이 재투자 되어 낭비되는 부분이 많다. 이러한 문제를 해결하기 위하여 현재 PDF에서 EPUB 3.0의 형태로 변환하는 도구가 개발되고 있다. 그러나 폰트가 추출 되지 않는 문제, 글자 누락, 색감 저조, 표 추출 누락 등 완벽히 변환되지 않아 아직은 미흡한 실정이다.

본 논문에서는 PDF에서 EPUB 3.0 으로의 변환을 좀 더 만족스럽게 처리하기 위하여 다음과 같이 두 가지 연구를 진행하였다.

※ 본 논문은 미래창조과학부의 2015년 고용계약형 SW석사과정 지원사업을 지원받아 수행한 결과입니다.

- 색감변환: 기존의 PDF 문서에서 EPUB 3.0 형태로 변환하는 데에 있어 문제점인 색감처리 부분을 개선하기 위하여 인쇄 색상 기준인 CMYK에서 웹 형태의 RGB 색상으로 ICC Color 파일을 이용해 변환하는 방법에 대한 연구
- EPUB 3.0의 특징인 상호작용 효과를 단순히 고정적인 글자에서 각 페이지 마다 시각적인 효과를 주기 위해 제목과 같은 강조할 문구를 찾아 효과를 줄 수 있는 알고리즘 연구

위의 두 가지 연구를 통하여 인쇄 색감에 동일한 색을 표현을 하였고 문서 변환 시 전체 페이지에 특정 글자 부분에 상호작용 효과를 주어 시각적 효과를 도출하였다.

2. PDF(Portable Document Format)와 EPUB 3.0

PDF는 아도비 시스템즈에서 개발한 전자 문서 형식이며 문자, 도형, 그림, 글꼴 등을 표현할 수 있다. 대부분의 문서를 표현할 수 있고 압축 및 암호화를 통해 변조가 용이하지 않고, 아도비사에서 PDF 문서를 볼 수 있는 아크로벳 리더를 무료로 배포하여 기업에서 외부 문서나 개인은 제출 문서 형태로 이용하고 있다. 사실상의 표준 문서로 자리매김 하였다. 최근에는 동영상, 음악 등 미디어를 재생할 수 있고 자바 스크립트까지 지원 가능하다. 하지만 최근 지원하는 부분을 일반 사용자는 제작하기 어려워 사용하지 못하고 있다.

EPUB 3.0 은 전자책 제작 업체마다 서로 다른 포맷으로 제작하여 사용자가 각각 포맷에 맞는 뷰어를 설치해야 했었다. 이에 IDPF에서 EPUB 전자책 표준 포맷을 개발하였다. EPUB 2.0 표준이 2010년 9월 제정되었고 EPUB 3.0 표준이 2011년 10월에 최종 승인되면서 현재 전자출판 시장에 사실상의 표준으로 되어 가고 있다. EPUB 3.0 은 다양한 기기에 맞도록 최적화된 콘텐츠를 볼 수 있게 자동공간조정(Reflowable), 고정레이아웃(Fixed Layout)을 지원한다. 또한 HTML5, CSS3.0, Javascript가 가능하여 다양한 언어, 동영상, 음악, 상호작용효과, 폰트내장, MathML 등을 표현하고 기능 구현이 가능하다[2].

3. 연구 개요

본 연구에서는 PDF 포맷에서 요소들을 추출하기 위해 PDF 라이브러리 중 하나인 PDFBOX 라이브러리를 이용하였고, 색감을 표현하고 변환하기 위해 JapanColor 2001 Coated ICC Profile과 sRGB ICC Profile을 이용하였다.

3.1 CMYK 에서 RGB 색 변환

색의 혼합에는 가산혼합과 감산혼합으로 나뉠 수 있다. 가산혼합은 빨강, 초록, 파랑색을 이용해 색상을 표현하는 RGB를 말하며 감산혼합은 시안(Cyan), 마젠타(Magenta), 노랑(Yellow)을 이용해 색상을 표현하는 CMYK를 말한다. 색상을 언급하는 이유는 기존 출판물의 형태인 PDF에서 쓰이던 색의 형태가 인쇄 때문에 CMYK를 이용하였고 EPUB 문서는 HTML, XML, CSS 등으로 구성되어 있어서 웹 색상 방식인 RGB 색상으로 구성된다. 이에 따라 PDF 문서가 CMYK 색상의 문서 일 경우 RGB 형태로 바꿔줘야 하는데 변환공식으로 계산 할 경우 색이 맞지 않는다.

CMYK에서 RGB 변환 공식[3]

R = 255 × (1-C) × (1-K)
G = 255 × (1-M) × (1-K)
B = 255 × (1-Y) × (1-K)

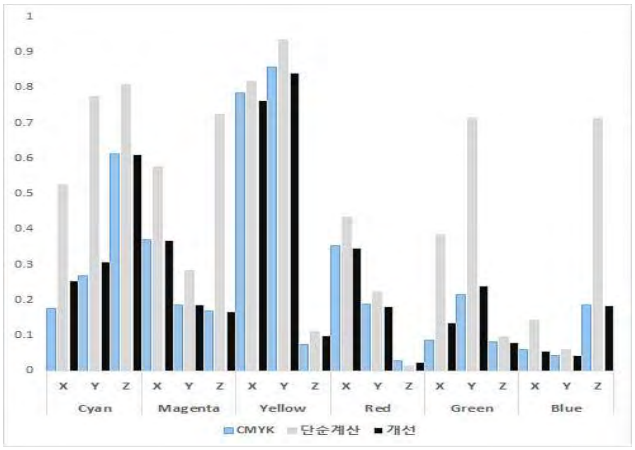
<표 1> 각 색상을 단순계산 했을 시 값

색상	CMYK (%)	단순계산 (R,G,B)
Cyan	(100,0,0,0)	(0,255,255)
Magenta	(0,100,0,0)	(255,0,255)
Yellow	(0,0,100,0)	(255,255,0)
Red	(0,100,100,0)	(255,0,0)
Green	(100,0,100,0)	(0,255,0)
Blue	(100,100,0,0)	(0,0,255)

단순한 변환 공식에 대입하여 단순 계산 했을 경우 <표 1> 과 같이 실질적으로 보는 색감이 같지 않다. 이러한 부분 때문에 국제 컬러협회(International Color Consortium (ICC))가 공표한 표준, 즉 색 입력 장치 또는 색 출력 장치의 특성을 구현하는 데이터 집합인 ICC 프로파일을 이용한다. 프린터, 모니터 등 각종 입력장치나 출력장치는 정의한 ICC 프로파일이 있는데[4] 본 연구에서는 PDF를 개발한 어도비 시스템에서 문서를 열 수 있는 Acrobat Reader에서 CMYK 색상을 표현하기 위해 기본으로 설정 되어있는 “JapanColor2001Coated” ICC 프로파일과 윈도우즈 디스플레이 색 시스템의 기본인 sRGB ICC 프로파일을 이용하여 구현하였다. 두 프로파일을 이용하여 색을 표현하기 위해서는 공통된 색 공간을 거쳐야 하는데 각 프로파일들은 CIE XYZ 색 공간 값을 계산할 수 있다. CIE XYZ 색공간은 인간 색채 인지에 대한 연구를 바탕으로 수학적으로 정의된 색 공간의 하나이다. <그림 1>은 색변환 코드로서, 먼저 추출하고자 하는 텍스트의 색이 CMYK 일 경우 해당 값을 “Japan Color 2001 Coated” ICC 프로파일로 구성된 색을 입력하여 CIE XYZ 값을 도출 한 뒤에 이 값을 sRGB ICC 프로파일에 입력하고 RGB 값을 추출하면 색의 손실 없이 변환 할 수 있다.

```
IF colorspace type of text == CMYK THAN
CIEXYZ = CMYK_ColorSpace.
                    toCIEXYZ(colorspace valueof text)
RGB value = sRGB_colorSpace.fromCIEXYZ(CIEXYZ)
```

<그림 1> 변환 의사코드



<그림 2> CIE XYZ값

<그림 2>에서 보면 색 공간 값인 CIE XYZ 값의 차이가 많이 나아진 것을 알 수 있다. 또한 RGB 색상 값에도 많은 차이가 나타나는데 <표 2>를 보면 Cyan 의 Green 수치는 무려 93이나 차이를 보였다.

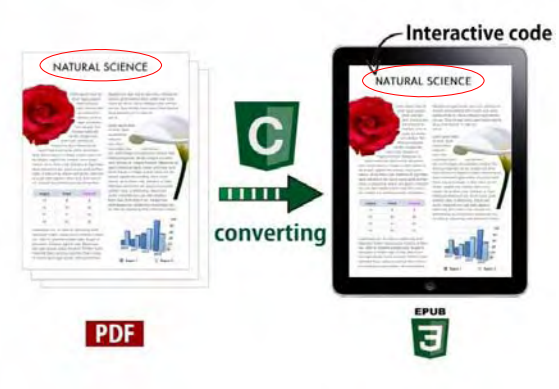
<표 2> ICC 프로파일 이용

색상	CMYK (%)	단순계산 (R,G,B)	ICC 적용 (R,G,B)
Cyan	(100,0,0,0)	(0,255,255)	(0,162,232)
Magenta	(0,100,0,0)	(255,0,255)	(228,3,129)
Yellow	(0,0,100,0)	(255,255,0)	(253,240,10)
Red	(0,100,100,0)	(255,0,0)	(229,24,35)
Green	(100,0,100,0)	(0,255,0)	(0,155,74)
Blue	(100,100,0,0)	(0,0,255)	(31,51,138)

3.2. 일괄적 상호 작용 효과 적용

기존 PDF 문서는 문서 기록에 초점을 맞추었기 때문에, 포함 되어 있는 미디어는 이미지에 국한 되어 있었다. 그러나 전자책 시장으로 오면서 책 넘김 효과, 동영상 추가, 상호 작용적 효과를 넣어 사용자와 교감을 이끌어 내는 방법이 연구되고 있다. 일례로 기존의 디지털 출판 도구를 이용하여 앱북을 만드는 연구도 있었고 쉽게 제작하는 부분[5]을 모색하고 있다.

본 연구에서는 PDF에서 EPUB 3.0 형태로 변환에 있어서 글자 부분에 다양한 효과를 넣는 방법을 연구하였다. 기존 발표문서나 웹에 있어서 헤드라인을 강조하여 보는 이로 하여금 주목을 끌기 위하여 애니메이션 기능이나 나타내기 등과 같은 효과를 주기도 하였다. 기존에 변환 시에는 그러한 효과들을 일일이 작업자의 손으로 코드를 삽입해야 했고 그에 따른 시간과 비용은 만만치 않았다. 이러한 부분을 변환 툴에 삽입하여 사람이 일일이 하지 않고 한 번에 도구를 이용하여 자동으로 추출, 효과 삽입을 하였다. 그 방법은 문서 변환 시 제목 부분이나 중제목 부분을 찾아내고 그 부분에 애니메이션 기능이나 나타내기 효과를 삽입한다. <그림 3>는 변환학고 바뀌는 부분을 나타내는 예시이다.



<그림 3> 상호작용 효과 적용 참고

```
FOR inputText = first to last of allText DO
   IF size of iput text > size of maximum text THEN
      secondText  = maxText
      maxText = inputText
    ELSE IF size of inputText == size of maxText THEN
      IF colorSum() of inputText<
          colorSum() of inputText THEN
         secondText = maxText
         maxText = inputText
      ELSE
         secondText = inputText
return maxText
```

<그림 4> 추출 의사코드

```
<div class="fade_in">
<div class="fnt2 vp_textbox text2">Magenta</div>
</div>
```

<그림 5> 예시 적용코드

문서에 제목부분이나 강조하기 위해 글자를 평문보다 크게 설정하거나, 굵게 표시하거나, 검정색이 아닌 다른 색을 주었다. 그래서 위와 같은 부분을 <그림 4> 코드를 이용하여 찾아내고 따로 분류를 해서 그 부분에 효과를 삽입하도록 하였다. 물론 다양한 형태의 문서가 존재하지만, 그 부분은 변환을 하였다. 위 알고리즘으로 찾아낸 글자들을 XHTML 코드 내에서 <그림 5>의 예시처럼 효과를 주었다.

4. 결론

본 연구에서는 PDF 문서를 EPUB 3.0 형태로 변환하는 데에 있어서 사람의 손을 거치지 않고 보정 없이 바로 볼 수 있는 형태로 제작하는데 목적을 두었다. 기존의 수많은 책들을 변환하고 재가공, 판매하는 부분에서 재투자 비용을 절감할 수 있는 부분을 찾아내야 할 것이다. 기존의 전자책으로의 변환 시장은 인간이 일일이 손으로 변환하고 효과를 주어야 했지만 이러한 연구가 지속적으로 발전되어 간다면 머지않아 기존 출판 책들을 전자책 시장에서 만나볼 날이 올 것이다.

본 연구를 진행하면서 기존 문서들을 디지털화 하는데 향후 다음과 같은 연구를 진행할 예정이다. 첫 번째로는 기존의 책은 크기가 정해져 있지만 각종 장치들은 크기와 형태가 다양하다. 가로, 세로로 볼 수도 있고 해상도도 다양하다. 이에 따라서 다양한 기기들에 맞추기 위해 EPUB 3.0 의 장점인 자동공간조정(Reflowable) 형태로의 변환이 필요하다. 두 번째로는 글자나 이미지의 상호 작용효과 뿐만 아니라 상황이나 날씨, 지역에 따른 내용, 이미지 바꿈 효과를 넣어 좀 더 다양한 형태로 생성되어야 할 것이다.

향후 다양한 형태의 빅 데이터를 활용하여 전자책을 가공하고 온라인으로 연결을 한다면, 독창적인 형태, 획기적인 전자책이 많이 출시될 수 있을 것이며 침체된 출판 시장이 재도약할 기회가 될 것이다.

참고문헌

[1] 이경호 · 임순범 (2010). “전자책 포맷 기술 및 표준화” 「정보과학회지」, 28(10): 31-39.

[2] 정의석 (2012). “EPUB 3.0 표준소개” 「TTA 저널」, 144: 55-58

[3] rapidtables. “CMYK to RGB color conversion” http://www.rapidtables.com/convert/color/cmyk-to-rgb.htm

[4] 송경철 · 강상훈 (2004). “고품질 컬러인쇄물의 색 교정 시스템 개발에 관한 연구” 「한국인쇄학회지」, 22(2): 55-72.

[5] 문현숙 (2013). “디지털 퍼블리싱을 활용한 인터랙티브 앱북 제작” 「디지털디자인학연구」, 13(2): 441-449.