Journal of The Korea Society of Computer and Information
Vol. 20 No. 7, pp. 1-7, July 2015
www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2015.20.7.001

A Study of Main Contents Extraction from Web News Pages based on XPath Analysis

Bok-Keun Sun*

Abstract

Although data on the internet can be used in various fields such as source of data of IR(Information Retrieval), Data mining and knowledge information servece, and contains a lot of unnecessary information. The removal of the unnecessary data is a problem to be solved prior to the study of the knowledge-based information service that is based on the data of the web page, in this paper, we solve the problem through the implementation of XTractor(XPath Extractor). Since XPath is used to navigate the attribute data and the data elements in the XML document, the XPath analysis to be carried out through the XTractor. XTractor Extracts main text by html parsing, XPath grouping and detecting the XPath contains the main data. The result, the recognition and precision rate are showed in 97.9%, 93.9%, except for a few cases in a large amount of experimental data and it was confirmed that it is possible to properly extract the main text of the news.

▸ Keyword : Main Text Extraction, Web News Page, XPath Grouping, HTML Parsing

Ⅰ. Introduction

인터넷 상의 데이터는 정보의 검색, 데이터 마이닝, 지식정보서비스 등 다양한 분야의 원천 데이터로 사용하게 된다. 인터넷의 발달이 위와 같은 분야의 응용기술 범위를 상당히 넓히는데 기여하였으며, 새로운 지식 정보기반 서비스를 탄생시키고 있다.

HTML5와 같은 새로운 인터넷 언어 규약이 분야별로 지속적으로 논의되고 있는 시점에서 이러한 지식정보기술의 진보는 날로 더해갈 것으로 사료된다. 그러나 HTML5를 통한 인터넷 언어의 표준화는 현재에도 진행 중이며, 완전한 표준화가 되기까지는 상당한 시일이 걸릴 것으로 예측된다. 2000년대 중반 web 2.0의 태동과 함께 웹을 통한 콘텐츠 제공과 더불어 응용프로그램을 구동할 수 있는 플랫폼으로써의 기능을 수행할 수 있다는 가능성을 보여주었으나, 웹 브라우저의 특성에 따라 서로 호환이 되지 않는 문제점이 대두되었으며, XHTML의 사용상 어려움등과 함께 이러한 문제가 HTML5 표준 규약의 제정에 가장 큰 배경이 되었다[1,2].인터넷 상의 데이터를 활용한 지식 정보기반 서비스 개발의 관점에서 볼 때, 인터넷 상의 콘텐츠는 방대한 지식 정보의 보고이지만, 동시에 그만큼의 많은 불필요한 정보를 포함하고 있다[3].

콘텐츠를 담고 있는 웹 페이지의 구조는 해당 페이지의 운영 목적에 따라 모두 다르며, 페이지의 내용은 주 내용과 더불어 주 내용과는 상관없는 불필요한 이미지, 광고, 스크립트, 카테고리 정보, 댓글 등이 넘쳐나고 있다. 이러한 현상은 웹 페이지의 데이터를 기반으로 하여 만들어지는 다양한 지식정보기반 서비스의 연구에 선행되어 해결해야 할 문제이다. 텍스트 및 데이터 마이닝, 정보검색 응용, 지식정보서비스 분야에서 웹 페이지 상의 데이터를 활용하기 위해서는 해당 페이지의 콘텐츠 중 정확한 정보와 내용의 추출이 선행되어야 한다. 이를 위해 Wrapper의 활용, 온톨로지 등의 지식모델 활용 등 다양한 연구가 진행되고 있다. 본 논문에서는 웹 페이지의 주 데이터 추출을 위해 보다 간단하고, 새로운 접근 방법을 제시하고자 한다. 개인 블로그, 뉴스, 게시판, 포럼 등 다양한 웹 페이지가 존재하지만, 본 연구에서는 웹 페이지 중 뉴스 서비스를 제공하는 사이트의 데이터 추출로 범위를 제한하였다. 직접 구현한 html 파서를 활용하여 요소트리를 구성하고, pcdata(printable character data)를 포함 한 노드의 구조를 분석하여 페이지의 주 데이터를 텍스트 형태

• First Author : Bok-Keun Sun
*Bok-Keun Sun(bksun@hoseo.edu) Dept. of Computer Engineering, Hoseo University
• Received: 2015. 04. 29, Revised: 2015. 05. 30, Accepted: 2015. 06. 29.

로 추출한다.

다음 장에서 정보의 추출과 관련된 연구를 살펴본 후, 이어서 데이터 추출 시스템 및 알고리즘에 대해 상세하게 살펴보고 평가해보도록 한다.

II. Related Works

정보검색 서비스 및 응용분야, 지식 정보서비스를 위한 원천 데이터로써 웹 문서를 사용하기 위해서는 전처리 과정을 통해 웹 데이터 추출이 이루어져야 하며, 정확한 데이터의 추출을 위해 자연어처리, Wrapper의 제작, 온톨로지, html 태그의 분석 등 다양한 관점에서 연구가 이루어지고 있다.

[4]은 템플릿을 활용하여 웹 포럼 콘텐츠를 추출하는 연구를 진행하고 있다. 웹 페이지의 구조변화에도 대응이 가능하면서 효율적으로 정보추출이 가능하도록 DOM 트리를 활용하여 데이터를 추출하는 기법을 사용한다.

DANA[5]는 영어를 사용하지 않는 국가의 웹 콘텐츠는 영어가 아니라는 점에 착안하여 인코딩 정보를 활용하여 영어로 되어있지 않은 사이트의 메인 콘텐츠를 추출하는 시스템으로 ASCII정보와 non-ASCII정보의 분류 후 연산을 통해 해당 페이지의 주 데이터를 추출 한다.

[6]는 HTML 파서를 활용해 DOM 트리를 구성하고 탐색하면서 필터링 알고리즘을 활용하여 해당 페이지의 메인 콘텐츠를 추출한다. [7]은 페이지 분석을 통해 콘텐츠가 포함하고 있는 하이퍼링크의 비율을 기반으로 불필요한 데이터를 제거하면서 메인 콘텐츠를 찾아낼 수 있는 방법을 제안하고 있으며, 링크구조의 밀도와 각 링크에 임계값을 설정하여 웹페이지의 본문을 추출하는 방안에 대해 제시하고 있다.

[8]은 HTML 트리를 만들고 그 중 블록 속성을 가진 태그의 구조를 재정의 한 후 이의 분석을 통해 중국어 웹페이지의 본문 콘텐츠를 추출한다. 분석과정에서 사용되는 파라메터로는 텍스트의 길이, 구두점의 개수, 하이퍼링크의 수 등이 활용되며, 임계값의 설정을 통해 해당 블록 안의 콘텐츠가 메인 데이터인지 노이즈 데이터인지 확인하게 된다.

FODEX[9]는 본문 및 관련된 쓰레드의 구조를 파악하여 웹 뉴스 포럼의 데이터 추출을 진행하였으며, Clearly[10], Readability[11]는 Chrome 및 안드로이드 확장 프로그램으로 Body 태그 이하 모든 노드를 검사하면서 본문의 후보가 될 노드의 점수를 산출하는 방식으로 본문을 선정하게 된다.

다량의 웹 페이지에 대한 통계조사 결과 뉴스 등을 포함하는 페이지의 경우, 광고 링크나 다량의 이미지가 뉴스의 본문과는 상관 없다는 것이 밝혀졌으며[7], HTML 노드의 구조 역시 다르게 나타난다[12]. 이에 기반 하여 본 논문에서는 뉴스페이지의 XPath 정보를 활용하여 본문을 추출하는 시스템을 구현한다.

III. 데이터 추출 시스템

그림 1은 뉴스 페이지의 한 예를 보여주며, 해당 페이지 중 응용분야에서 필요한 기사의 본문 이외에는 모두 불필요한 데이터라고 할 수 있다.

Fig. 1. Web News Page Sample(ex-media.daum.net)

대부분의 웹 뉴스 페이지는 그림 1과 유사한 구조로 이루어져 있으나, 각 뉴스 제공 사이트마다 HTML 및 웹 언어의 사용방식이 다르다. 표 1은 10개 뉴스 사이트를 대상으로 전체 텍스트 대비 메인 콘텐츠의 텍스트 길이를 조사한 결과이다. 사이트별로 100개의 뉴스 페이지를 조사한 후 평균값을 표 1에 나타냈으며, 이에 따르면 총 텍스트 중 10.6%~57.8%가 실제 뉴스 원문이며, 나머지 텍스트는 모두 기사와 상관없는 링크, 댓글이나 광고에 활용되는 텍스트임을 알 수 있다.

뉴스 원문과 상관없는 불필요한 데이터를 제거하고 순수하게 원문에 해당하는 데이터를 추출하기 위하여 본 논문에서 제시한 시스템은 입력으로 들어온 뉴스 페이지를 XPath를 활용한 트리 형태의 자료구조로 구성 한 후, 분석을 통해 해당페이지의 본문을 자동으로 추출할 수 있도록 구성하였다. 다음절부터 XPath를 활용한 트리구조 구성, 트리구조 기반 본문데이터 XPath 결정 및 본문추출과정에 대해 논한다.

그림 1과 같은 뉴스 원문을 포함한 페이지는 템플릿을 활용하여 작성되기 때문에 추출이 필요한 뉴스 도메인에서 사용하는 템플릿의 구조를 파악할 경우 그 도메인의 뉴스 원문을 추출할 수 있게 된다.

www.kci.go.kr

Table 1. News Source Text Length Comparison

사이트	총 텍스트	원문 텍스트
1	10874	1356(12.4%)
2	8674	3491(40.2%)
3	4944	2101(42.4%)
4	10830	1163(10.7%)
5	4424	1053(23.8%)
6	6594	702(10.6%)
7	13399	7293(54.4%)
8	6326	3136(49.5%)
9	2019	1169(57.8%)
10	2524	1051(41.6%)

뉴스 도메인마다 사용되는 템플릿의 구조는 모두 다르며, 이러한 구조를 파악하기 위해 [4][7]와 같은 연구가 이루어지고 있으나, 계산에 소모되는 비용이 크며, 템플릿 구조가 바뀔 때마다 새로운 비용이 발생하게 된다.

1. XPath 경로표현법을 활용한 트리 구성

XPath는 XML 문서의 요소데이터와 속성데이터의 탐색을 위해 사용되어지며, W3C의 XSLT 표준의 핵심 요소이다[13,14]. XPath는 XML 문서의 노드 또는 노드집합의 선택을 위한 경로의 표현방법으로 사용될 수 있으며, 문자열이나 숫자 등과 관련된 함수를 제공한다. 또한 XSLT, XPointer, XQuery 등 XML 문서를 다루는 다른 표준에서도 활용된다.

XPath는 XML문서의 특정한 요소나 속성까지 도달하기 위한 경로를 트리형태의 계층 구조를 이용해서 표현하며, 그림 2는 HTML파일의 XPATH 표현 예를 보여준다.

Fig. 2. XPath Expression of the Web Page

본 논문에서는 뉴스 원문이 포함된 HTML파일의 파싱을 통해 <HTML> tag를 root로 하는 요소(element)트리를 만들면서 트리의 각 요소마다 XPath를 포함하도록 구성하였다.

2. 본문 검출 알고리즘

논문에서 예로 제시한 그림 2의 요소트리 구조를 요약하면 그림 3과 같으며, 전체 요소트리 중 텍스트 데이터를 포함한 요소 수(N)는 총 907개이다. 수식(1)의 $XPATH$는 N개의 요소를 가진 중복집합(multiset)으로 표현되며, 연속되는 동일한 XPath(ID)의 중복도(multiplicity, C) 계산을 통해 요소집합으로 구분하였다. 텍스트가 포함된 요소를 수식(1)과 같이 중복집합으로 구분할 경우, n개의 요소블록 그룹이 검출되며, 그림 2의 예에서 볼 때, 69개의 요소집합이 생성된다.

$$\begin{cases} \begin{cases} XPATH = \{ID_N, \{ID_1, C_1\}, \{ID_2, C_2\}, ..., \{ID_n, C_n\}\} \\ ID : XPATH \rightarrow N \geq 1 = \{1,2,3,...N\} \end{cases} \\ N = \sum_{i=1}^{n} C_i \end{cases} \quad (1)$$

검출 된 요소블록(ID_1-ID_n)은 모두 뉴스 원문을 포함하는 XPath의 가능성을 가지고 있다고 볼 수 있으며, 이중 하나를 선정하여 뉴스 원문을 가진 요소블록으로 판정하게 된다. 판정을 위해 수식(2)와 같이 각 ID가 가지고 있는 문자의 길이를 모두 합산하여 가장 길이가 긴 데이터를 가지고 있는 ID를 뉴스 원문을 가진 XPath로 선정하게 된다.

$$\begin{cases} Length\,of\ ID_n = \sum_{i=1}^{C_n} textLength\,of\ ID_n \\ Candidate\ ID = \max(ID_1, ID_2, ...ID_n) \end{cases} \quad (2)$$

3. XTractor 시스템의 구현

뉴스 원문 검출 시스템(XTractor)은 네트워크상의 HTML 문서를 입력으로 하여 원문을 검출하게 되며, 중간에 요소트리, XPath 그룹화, 원문 XPath의 3가지 중간 데이터를 산출한다. 시스템의 구현 및 실행은 javaSE 1.8.0_40 SDK와 실행환경을 활용하였다.

3.1 요소 및 요소트리

입력으로 들어온 HTML 문서는 파서를 통해 요소 트리로 구성된다. 직접 구현한 파서는 HTML문서의 태그별 토큰화(tokenization)를 통한 요소 생성과 트리구조화를 반복하면서 요소트리를 생성하게 된다. 그림 3은 파싱관련 클래스 및 요소트리에 해당하는 Element 클래스 다이어그램의 요약을 나타낸다.

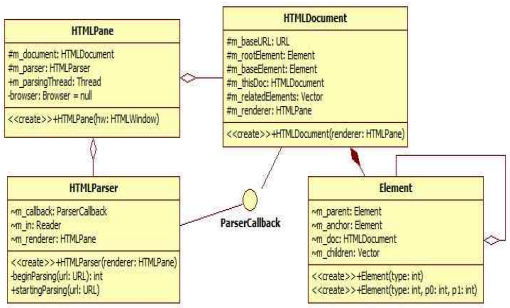

Fig. 3. A Summary of the Class Associated with the HTML Parsing

www.kci.go.kr

HTMLParser 클래스는 입력된 HTML파일의 태그를 토큰화하여 ParserCallback 인터페이스에 전달한다. ParserCallback 인터페이스는 HTML 태그를 root로 하는 Element의 생성을 시작으로 하위 Element를 생성하여 트리구조화를 수행하며, 파싱이 끝나게 되면 요소트리가 생성된다. 그림 4는 요소트리 중 pcdata를 가진 요소트리의 개념도를 나타낸다.

요소트리 생성은 문서의 모든 태그에 대해 이루어지며, 파싱이 끝나게 되면 요소트리 탐색을 통해 그룹화와 본문후보의 XPath를 선택하게 된다.

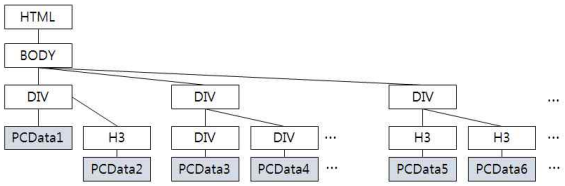

Fig. 4. Schematic Veiw of the Element Tree with the pcdata

3.2 요소트리 탐색방법 및 엑스패스 그룹핑 알고리즘

요소트리의 각 요소는 고유의 XPath를 가지고 있으며, pcdata를 가지고 있는 요소트리만 선별하여 연속적으로 발견되는 동일 XPath를 가진 요소별로 그룹화가 가능하다.

그림 3에 나타난 pcdata의 XPath는 다음과 같다.

- PCData1 : HTML/BODY/DIV/
- PCData2 : HTML/BODY/DIV/H3
- PCData3 : HTML/BODY/DIV/DIV
- PCData4 : HTML/BODY/DIV/DIV
- PCData5 : HTML/BODY/DIV/H3
- PCData6 : HTML/BODY/DIV/H3

3장에서 예로 든 페이지의 경우 위와 같은 pcdata를 포함하는 XPath가 907개 생성된다. 생성된 요소트리 탐색은 root 요소부터 시작하여 pcdata를 포함하는 최하위 노드 요소까지 깊이우선 탐색(depth-first search)하면서 XPath의 그룹화를 진행한다.

그림 5는 그룹화 알고리즘의 슈도코드를 나타내며, 그룹화 후에는 그림 3의 요소트리가 그림 6과 같이 그룹화 된다.

```
Void AnalizeFunc(Node node)
  IF PCDATA Node
    IF Node==new Node
      Save Previous Node group length
    ELSE continue to add length
  END IF
  ELSE
    WHILE(child node is Not NULL)
      Get child node
      AnalizeFunc recursive call
    END WHILE
  END ELSE
```

Fig. 5. Grouping Algorithm Pseudo code

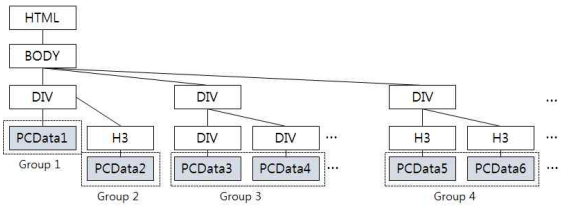

Fig. 6. Schematic View of the pcdata node grouping

요소트리 탐색을 마치게 되면 XPath 그룹이 만들어지게 되며, 3장에서 예로든 페이지의 경우 67개의 그룹이 생성된다. HTML 페이지의 모든 pcdata는 XPath 그룹에 매핑되며, 생성된 그룹 중 pcdata의 길이가 가장 긴 XPath를 선정하여 해당 pcdata를 뉴스 원문으로 선정하게 된다. 그림 7은 제시한 시스템의 분석화면이다. (1)부분이 pcdata를 포함한 XPath의 항목을 나타내며, (2)부분은 연속되는 XPath의 그룹을 나타낸다. 그림 5의 슈도코드에 따라 뉴스원문으로 선택된 XPath가 (3) 부분에 나타나며, (4) 부분이 해당 XPath의 뉴스 원문을 나타낸다.

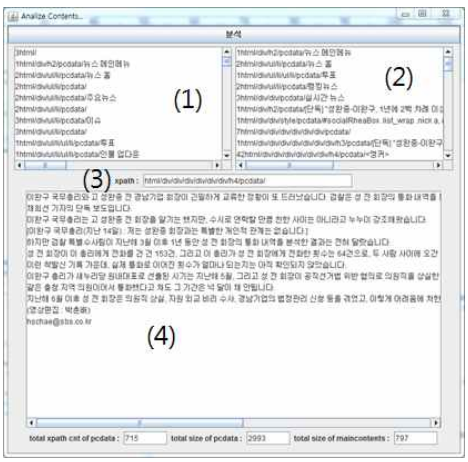

Fig. 7. Example of System Analysis Screen

www.kci.go.kr

IV. 평 가

본 논문에서 구현한 시스템은 뉴스 사이트의 기사를 추출하기 위한 것이므로 평가 시 실험에 사용한 웹 문서는 뉴스 원문이 포함되어있는 페이지로 제한하였다. 같은 이유로 각 사이트의 분류별 포탈 페이지 역시 실험에서 제외되었다.

시스템의 평가를 위해 뉴스 원문을 제공하는 10개의 사이트를 선정하고, 수식(3)과 같이 사이트 당 100개의 뉴스원문 페이지(N)를 대상으로 본문 인식률(R)과 정확률(P)을 측정하였으며, Readability에서 사용한 태그의 노드검사 방법과 비교분석을 통해 제안한 시스템의 평가를 진행하였다.

Table 2. Extract Performance Comparison Table

사이트 이름	문서 수(N)	본문 인식률(%)		평균 정확률(%)	
		XTractor	태그노드검사	XTractor	태그노드검사
다음뉴스	100	100	100	92.7	92.7
네이버뉴스	100	98	95	93.8	93.8
MBC	100	97	97	93.5	91.5
KBS	100	98	97	91.9	91.9
YTN	100	98	97	92.4	90.2
조선일보	100	95	100	93.6	93.6
중앙일보	100	96	96	95.7	95.7
경향신문	100	98	75	93.2	90.8
오마이뉴스	100	99	97	92.5	92.5
전자신문	100	100	100	100	93.1
계/평균	1,000	97.9	95.4	93.9	92.5

$$\begin{cases} \text{인식률}(R) = \dfrac{N-r}{N} \times 100 \\ \text{정확률}(P) = \dfrac{M}{M+G} \times 100 \end{cases} \quad (3)$$

인식률로 평가 시스템이 페이지의 원문에 해당하는 부분을 정확히 검출했는지 평가한다. 원문을 찾지 못하거나, 원문과 전혀 다른 내용을 원문으로 제시할 경우(r) 시스템이 인식하지 못한 것으로 평가하였다.

정확률은 평가 시스템이 제시 한 원문이 실제 원문(M) 외에 불필요한 데이터(G)를 얼마나 포함하고 있는지 평가한다. 정확률이 높을수록 원문 외의 불필요한 데이터가 적다는 것을 나타낸다. 원문을 인식하지 못한 페이지의 정확률은 측정할 수 없으므로 표 2의 정확률은 평가 시스템이 원문을 인식한 페이지를 대상으로 평가하여, 평균 정확률을 나타내었다.

측정 및 평가결과 10개 사이트의 뉴스원문 추출 성능은 표 2와 같다. XTractor의 뉴스 원문 전체에 대한 평균 인식률은 97.9%로 태그의 노드검사 방법과 비교해서 높게 나타난다.

태그의 노드검사 방법은 특정 사이트의 전체 원문 데이터를 인식하지 못하는 경우가 발생하였으나, XTractor는 원문이 아닌 데이터를 뉴스원문으로 인식하는 현상이 발생하였다. 이러한 현상은 뉴스 원문의 길이가 광고, 댓글 등 불필요한 정보보다 짧은 경우에 주로 발생하였으며, 추후 시스템의 보완이 필요한 부분이다.

XTractor의 전체 평균 정확률은 93.9%로써 비교대상 시스템에 비해 높은 결과를 나타내며, 원문 추출결과에 포함된 FP(False Positive) 데이터는 저작권 정보, 미란다원칙 등 뉴스데이터의 마무리에 붙는 정보로 나타났다. 단, Readability는 태그의 노드검사를 통해 기사 원문을 사용자가 보기에 편리하게 재배치해주는 툴로써 이미지, 링크 처리 등에서 약간의 차이를 보였을 뿐 큰 차이는 나타내지 않았다.

표 3은 사이트 별 XPath 및 원문데이터의 크기 정보를 나타낸다. 전체 pcdata 중 원문 pcdata의 길이 평균은 약 31%정도로 나타난다. 이 비율은 전체 화면에서 뉴스 원문이 차지하는 비율이며, 크기가 낮은 것은 원문 정보보다 많은 불필요한 정보를 포함하고 있다는 의미가 된다. 원문을 제외한 모든 데이터는 정보처리의 관점에서 불필요한 데이터라고 할 수 있으나, 기사 활용 측면에서는 기사에 대한 댓글 등 의미 있는 자료도 함께 들어있는 경우가 있으므로 원문이 차지하는 비율이 낮다고 무조건 불필요한 정보가 많이 포함된 페이지라고 단정할 수 없다.

뉴스원문을 포함하는 페이지의 pcdata 중 69% 정도는 정보처리의 관점에서 불필요한 데이터라고 할 수 있다. 불필요한 데이터를 버리고 원문 데이터를 추출하기 위해 본 논문에서 제시한 XTractor를 사용할 경우, 평균 93.9%의 정확률로 원문을 추출할 수 있다고 할 수 있다. 본문과 함께 추출되는 6.1%에 해당하는 불필요한 데이터 또한 원문과 같은 XPath를 사용하는 저작권정보, 미란다원칙 등에 해당하며, 향후 충분히 해결할 수 있는 문제로 판단된다.

www.kci.go.kr

Table 3. Pcdata size analysis and site-specific XPath of main text

사이트 이름	pcdata XPath 수	전체 pcdata 길이	원문 pcdata 길이	대표적 원문 XPath
다음뉴스	873	4944	2101	html/body/div/div/div/div/div/div/div/div/div
네이버뉴스	811	10830	1163	html/body/table/tr/td/div/div/div
MBC	182	6326	3136	html/body/div/div/div/div/div/div/div
KBS	522	6594	702	html/body/div/div/div/div/ul/li/div
YTN	319	13399	7293	html/body/div/div/div/div/div/div
조선일보	98	2019	1169	html/body/div/div/div/div
중앙일보	74	2524	1051	html/body/div/div/div/div/div/div
경향신문	1295	10874	1356	html/body/div/div/div/div/div
오마이뉴스	524	8674	3491	html/body/div/div/div/div/div/div/div/div/div
전자신문	396	4424	1053	html/body/div/div/div/div/div/p

Ⅴ. Conclusion

인터넷을 통한 정보의 획득은 우리에게 여러 측면에서 유익한 점과 불편한 점을 가져오고 있다. 원하는 정보를 쉽게 찾을 수 있는 반면, 너무나 많은 정보에 노출되면서 발생하는 단점들 또한 극복해야 할 과제로 다가오고 있다.

인터넷을 통해 제공되는 언론사의 뉴스 페이지들은 광고와 링크, 자바스크립트를 활용한 동적 콘텐츠, 댓글등 뉴스 원문을 읽기에 적절하지 않은 콘텐츠를 동시에 포함하고 있으며, 이러한 데이터의 제거를 위해 블록 속성을 가진 태그의 분석, 기계학습 등 다양한 연구가 이루어지고 있다.

본 연구는 뉴스 원문데이터를 포함하는 XPath를 찾아내는 방법에 중점을 두었으며, 연구 결과는 비교 시스템에 비해 높게 나타났다. 문제점으로는 뉴스 원문의 길이보다 긴 불필요한 데이터가 같이 존재할 경우 불필요한 데이터를 원문으로 판단하며, 뉴스 원문의 마지막에 포함되는 저작권 정보와 같은 데이터가 포함되는 현상이 발생하였다. 첫 번째 문제의 경우, 패턴인식의 예외상황으로 표현될 수 있으며, 두 번째 문재의 경우는 휴리스틱을 적용하여 해결할 수 있을 것으로 기대된다.

향후 데이터 기반 서비스, 정보검색 및 응용분야, 지식 정보 서비스를 위해서는 양질의 원천 데이터가 필요하며, 이의 확보를 위해 본 논문에서 제시한 XTractor 시스템은 유용하게 활용될 수 있을 것으로 판단된다.

REFERENCE

[1] HTML5, http://www.w3.org/TR/html5/

[2] Myeong-Chul Park, Seok-Gyu Park, Hyun-Syug Kang, "Interactive Learning Tool Based on HTML5 Using Unplugged Contents", Journal of The Korea Society of Computer and Information, Vol.19, No.11, pp. 73-79, November 2014

[3] D. Shen, Q. Yang, Z. Chen, “Noise reduction through summarization for Web-page classification”, Information Processing and Management vol.43, pp.1735-1747, 2007.

[4] J. Si, W. Wang, “A Template-based forum posts content extraction method”, International Conference on ICECE, pp.38-41, 2011.

[5] H. Mohammadzadeh, T. Gottron, F. Schweiggert, G. Nakhaeiza, “A Fast and accurate approach for main content extraction based on character encoding”, 22nd International workshop on database and expert systems applications, pp.167-171. 2011.

[6] S.Gupta, G. Kaiser, D. Neistadt, and P. GS.Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-based content extraction of HTML documents", in WWW '03: Proceedings of the 12th International Conference on WWW, ACM, pp.207-214, 2003.

[7] R. Gunasundari, S. Karthikeyan, “A Study of content extraction from web pages based on links”, International Journal of Data Mining & Knowledge management Process(IJDKP) vol.2, No.3, 2012.

www.kci.go.kr

[8] B. Zhou, C. Wang, Q. Su, “Chinese web page content extraction based on page content analysis”, Journal of Computational Information Systems vol.5, No.6, pp.1861-1871, 2009.
[9] S.Pretzsch, K.Muthmann, A.Schill, “FODEX-Towards generic data extraction from web forums”, 26th International conference on advanced information networking and applications workshops, pp.821-826, 2012
[10] Clearly, *https://chrome.google.com/webstore/ detail /clearly/iooicodkiihhpojmmeghjclgihfjdjhj*
[11] Readability, https://www.readability.com/
[12] A. Arasu, H.Garcia-Molina, “Extracting structured adta from web pages”, SIGMOD ‘03:Proceedings of the 2003 ACM SIGMOD international conference on Management of data, ACM, pp.337-348, 2003.
[13] Sangyoon Oh, "X2RD: Storing and Querying XML Data Using XPath To Relational Database", Journal of The Korea Society of Computer and Information, Vol.14, No.3, pp. 57-64, March 2009.
[14] XPath, http://www.w3.org/TR/xpath/

Authors

Bok-Keun Sun received the B.S., M.S. and ph. D. degrees in computer engineering from Hoseo University, Korea, in 1999, 2001 and 2006, respectively.

Dr. Sun joined the facult of the Department of Computer Engineering at Hoseo University, Asan, Korea, in 2008. He is currently a professor in the Department of Computer Engineering, Hoseo University. He is interested in Data Mining, Embedded system Design and IoT Computing.

www.kci.go.kr