규칙기반 문서 분류기를 이용한 XML 문서의 자동생성

○
김효정*, 민미경*
*서경대학교 컴퓨터과학과

Automatic Generation of XML Documents Using Rule-Based Document Classifier

○
Hyo-Jung Kim*, Mee-Kyung Min*
* Dept. of Computer Science, Seokyeong University

요 약

인터넷 중심의 정보화 사회가 되면서 기존의 문서는 대부분 전자 문서로 대치되어 가고 있다. 전자 문서간의 호환과 표준화를 위하여 XML(eXtensible Markup Language)이 웹 문서의 표준으로 지정되었으나, 현재까지 사용되고 있는 문서들이 XML 형태의 문서가 아니므로 이를 수동으로 변환해야 하는 어려움이 있다. 본 논문에서는 규칙기반 문서 분류기(Rule-Based Document Classifier)를 설계하여 다양한 형태의 문서를 자동으로 분류하고 그룹화한다. 그룹화된 문서를 이용하여 자동으로 DTD(Document Type Definition)를 생성하고, 자동 생성된 DTD를 이용하여 XML 형태의 문서로 자동 변환할 수 있는 자동 XML 변환기를 제시한다. 이러한 방법은 문서들을 자동으로 분류하고, 문서의 형태에 변화가 있을 때에도 유사한 문서로 분류할 수 있을 뿐만 아니라 문서를 재분류할 때 DTD의 중복 생성을 줄일 수 있는 등의 장점을 갖는다.

1. 서론

전자 문서를 다루는 전자도서관과 같은 기관에서 해결해야 하는 과제는 인터넷 중심의 정보화 사회가 되면서 기존의 다양한 형태의 전자 기술문서들이 어떻게 정의되고 얼마만큼 구조적인 문서 정보를 제공해 줄 수 있느냐 하는 것이다. 지금까지는 여러 규격의 문서를 정의하여 사용하였으나 이를 하나의 통합된 문서 구조를 갖는 전자 문서로 제공해줘야 한다.

문서에서 논리구조와 내용구조를 기술하기 위한 메타언어인 SGML(Standard Generalized Markup Language)은 플랫폼에 독립적인 문서 구조를 저장할 수 있으므로 다양한 응용에 사용될 수 있지만 너무 복잡하고 인터넷을 기반으로 하고 있지 않아 인터넷상에서 서비스를 제공하기가 어렵다. 이에 1996년 W3C(World Wide Web Consortium)의 XML 워킹 그룹은 기존에 사용하던 HTML(HyperText Markup Language)의 한계를 극복하고 SGML의 복잡함을 해결하기 위해 XML을 전자문서의 표준으로 제정했다.[1]

전자문서를 사용자에게 제공하기 위해 사용하고 있는 방법들은 다음과 같다. 첫째, 도서관 출판자료를 스캔한 이미지로 관리, 저장하여 사용자에게 이미지 형태의 파일로 제공하는 방법과 둘째, 기존에 출판된 문서자료를 SGML[7] 또는 XML[4]과 같은 표준화된 문서형태로 재 입력하는 방법이 있다.

두 가지 방법 모두 각각 장·단점을 갖고 있다. 첫 번째 방법은 두 번째 방법보다 인력·비용 면에서 효율적이지만, 이미지 파일이 가지는 편집상의 어려움, 통신상에서의 속도 문제, 디스플레이 문제 등의 어려움이 있다. 이에 비해 두 번째 방법은 앞의 방법보다 통신상의 속도가 빠르고, 웹 브라우저를 통해서 표준화된 문서를 보다 쉽게 디스플레이 할 수 있지만, 기

존의 출판자료나 이미 만들어져 있는 문서를 표준화된 형태의 SGML이나 XML로 만들기 위해서는 일일이 변환하는 과정이 필요하다.[6]

본 논문에서는 문서를 자동으로 분류하여 XML 문서를 자동 생성할 수 있는 규칙기반 문서 분류기를 제시한다. 이를 이용하여 다양한 형태의 전자문서를 자동 분류하여 그룹화한다. 그룹화된 문서의 정보를 기반으로 하여 자동으로 DTD를 생성하고, 자동 생성된 DTD를 이용하여 XML 형태의 문서로 자동 변환할 수 있다.

본 논문의 구성은 다음과 같다. 먼저, 2장에서는 문서 생성기의 전체구조와 규칙기반 문서 분류기에 대해서 기술하고 3장에서는 자동 분류기를 이용한 XML 문서의 자동 변환 방법을 기술한다. 마지막으로 4장에서는 결론을 맺는다.

2. 규칙기반 문서 분류기와 문서의 자동 분류

2.1 전체구조

본 논문에서 제시하고자 하는 규칙기반 문서 분류기 및 XML 문서 자동 생성기의 구조는 아래의 그림과 같다.

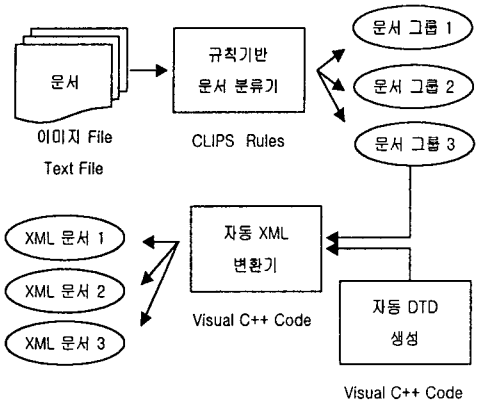


그림1. 전체구조

이미지나 텍스트 형태의 문서를 파일로 인식하여 분류할 수 있도록 전문가 시스템(expert system) 언어인 CLIPS[2] 규칙을 사용하여 규칙기반 문서 분류기를 설계한다. 설계된 문서 분류기는 유사한 형태의 문서를 같은 그룹으로 자동 분류한다. 분류된 문서 그룹의 구성 정보를 이용하여 자동으로 DTD를 생성하고, 자동으로 XML 문서로 변환해 준다.

2.2 규칙기반 문서 분류기

규칙 설계자는 CLIPS를 사용하여 규칙을 정의한다. 본 논문에서 대상으로 한 문서의 종류는 논문 형태이다. 논문 형태의 문서는 주제(title), 저자(author), 요약(abstract), 본문(body), 참고문헌(reference) 등의 특별한 형식을 가지므로 제안된 규칙기반 문서 분류기를 이용하여 유사한 형식을 갖는 논문별로 자동 분류하기에 용이하다. 분류된 논문의 형태가 기존에 정의되어 있는 문서의 형태일 경우에는 지정된 DTD를 사용하고 새로운 문서의 형태를 갖는 경우에는 새로운 형태로 분류하여 독립된 DTD를 새로 생성해준다. 입력 문서는 어떠한 형태의 논문이든지 가능하다. 예를 들면 아래와 같은 2가지 형태가 있다.

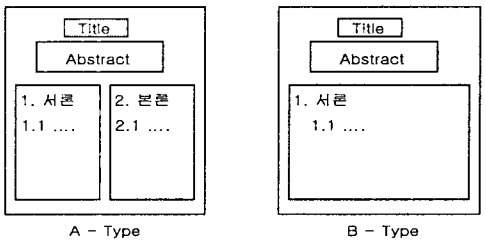


그림2. 분류하는 논문 형태의 예

CLIPS를 이용하여 규칙을 정의하면, 우선 문서의 형태를 인식하기 위해서 문서에 나타나는 주제, 저자, 요약, 본문, 참고문헌을 키워드로 검사한다. 키워드가 있으면 화면에 디스플레이하고 키워드가 없으면 다음 키워드를 읽는다. 이렇게 문서 구성 정보를 하나하나 검사하면서 논문 문서의 형태를 분류한다.

다음은 문서 분류기에서 사용되는 CLIPS 규칙의 예이다.

```
(rule (if englishtitle is yes)
      (then superroot is englishtitle))
(rule (if englishtitle is no)
      (then superroot is noenglishtitle))
(question englishtitle is "english-title?")

(rule (if superroot is englishtitle and
       koreantitle is yes)
      (then root is koreantitle))
(rule (if superroot is englishtitle and
       koreantitle is no)
      (then root is nokoreantitle))
(question koreantitle is "korean title?")
```

그림3. 규칙 정의 예

논문 문서의 구성 정보는 아래의 그림과 같은 형태를 갖는다.

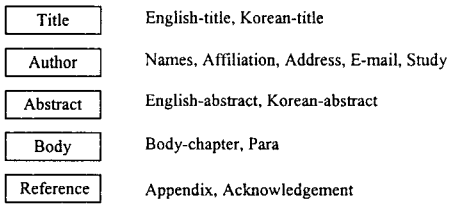


그림4. 논문 문서의 구성 정보

논문 문서의 구성 정보는 위의 그림에서처럼 문서 내의 제목, 저자, 요약, 본문, 참고문헌 등으로 구성된다. 이러한 논문 문서의 구성 정보를 규칙에 적용하여 주어진 키워드를 분류한다.

CLIPS로 정의된 규칙을 사용하여 규칙기반 문서 분류기를 실행하면 아래의 그림과 같다.

```
Defining deftemplate: rule
Defining defrule: propagate-goal +j+j
Defining defrule: goal-satified =j+j+j
Defining defrule: remove-rule-no-match +j+j
Defining defrule: modify-rule-match =j+j
Defining defrule: rule-satisfied =j+j
Defining defrule: ask-question-no-legalvalues +j+j+j+j
Defining defrule: ask-question-legalvalues +j+j+j
Defining deffacts: knowledge-base
TRUE
CLIPS> (reset)
CLIPS> (run)
english-title
korean-title
author
names
affiliation
field-of-study
address
email
english-abstract
korean-abstract
bodyies
body-chapter
arabic-chapter
sub-arabic
journal-name
appendix
Acknowledgement
This is an A-type PAPER.
CLIPS>
```

그림5. 규칙기반 분류기의 실행화면

위의 실행화면에서 알 수 있듯이 논문 문서에서 나타나는 키워드를 분류할 수 있다. 분류된 키워드를 이용하여 논문 문서의 형태를 결정한다.

3. 자동 XML 변환기

자동 XML 변환기의 구조는 다음과 같다.

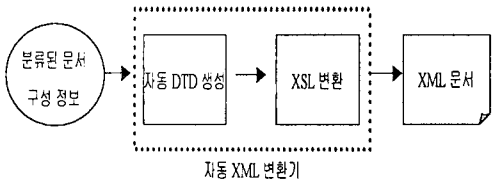


그림6. 자동 XML 변환기 구조

기존의 변환기는 DTD를 자동 생성하기보다는 문서에 적당한 DTD를 미리 정의해 놓고, 주어진 문서를 정의된 DTD에 알맞게 변환했으나, 자동 XML 변환기는 자동 분류된 논문 문서들에 알맞은 DTD를 자동으로 생성한다. 유사한 문서가 입력될 때에는 기존에 생성된 DTD를 재사용함으로써 논문 문서가 입력될 때마다 DTD를 생성하지 않아도 된다.

다음은 분류된 문서 구성 정보를 이용하여 자동으로 DTD를 생성한 화면이다.

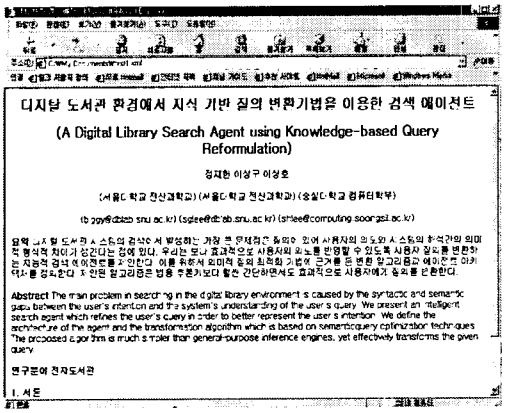
```
<!ELEMENT      TITLE          (english-title,korean-title)>
<!ELEMENT      english-title  (#PCDATA)>
<!ELEMENT      korean-title   (#PCDATA)>
<!ELEMENT      AUTHOR         (names,affiliation,address,email)>
<!ELEMENT      names          (#PCDATA)>
<!ELEMENT      affiliation    (#PCDATA)>
<!ELEMENT      address        (#PCDATA)>
<!ELEMENT      email          (#PCDATA)>
<!ELEMENT      BODIES         (body-chapter)>
<!ELEMENT      body-chapter   (#PCDATA)>
<!ELEMENT      REFERENCE      (journal-name,appendix,acknowledgement)>
<!ELEMENT      journal-name   (#PCDATA)>
<!ELEMENT      appendix       (#PCDATA)>
<!ELEMENT      acknowledgement (#PCDATA)>
```

그림7. 자동 DTD 생성 화면

XML 문서는 문서의 구조적인 정보만을 저장하고 외양에 대한 정보는 저장하지 않기 때문에 XML 문서를 보기 위해서는 문서의 외양에 대한 정보를 갖는 스타일 시트 (style sheet)가 필요하다.[3] 가장 일반적인 스타일 시트로 DSSSL(Document Style Semantics and Specification Language), CSS(Cascade Style Sheet)가 있고 XML에서는 CSS 이외에 XSL(eXtensible Style Language)을 사용한다. XSL은 DSSSL에서 많은 것을 가져와서 CSS와는 비슷하지만 더 다양한 기능을 가진 것이 XSL이다.[5] 본 논문에서는 XSL을 이용하여 문서의 외양을 표현

했다.

다음은 XSL를 적용한 XML 문서의 예이다.


디지탈 도서관 환경에서 지식 기반 질의 변환기법을 이용한 검색 에이전트

(A Digital Library Search Agent using Knowledge-based Query Reformulation)

그림8. XSL 적용 후의 XML 문서 예

4. 결론

현재 대부분의 전자문서 정보 검색 시스템에서는 기술 문서의 저장, 관리 방법에 있어서 텍스트 또는 이미지 형태, 표준화 문서(XML) 형태, 또는 혼합 형태를 사용하고 있다. 기술문서들을 표준화 문서 형태로 변환하기 위해서는 텍스트 또는 이미지 형태의 문서들을 XML 문서로 변환할 때 수동으로 변환해야 하는 어려움이 있다. 수동으로 XML 문서로 변환할 경우에는 사용자가 XML 문서 형태로 직접 만들어야만 하지만 논문 문서를 XML로 자동 변환해 줌으로써 보다 효율적인 변환이 가능하다.

본 논문에서는 규칙기반 문서 분류기를 이용하여 지금까지 생성되어진 다양한 형식의 전자문서를 표준화된 XML 문서로 자동 변환할 수 있는 XML 문서 생성기를 제시하였다.

전자 기술문서를 파싱 과정을 거쳐 자동으로 분류하고, 자동으로 DTD를 생성하여 사용자에게 보다 구조화된 전자문서를 제공할 수 있다.

규칙기반 문서 분류기를 이용함으로써 가지는 장점은 다음과 같다. 첫째, 다양한 형태의 전자 문서를 특성에 맞는 그룹으로 자동 분류할 수 있다. 둘째, 문서의 형태에 변화가 있을 시에도 가장 적합한 그룹을 검색하여 모호한 형태의 문서를 가장 유사한 문서 그룹으로 분류하여 XML 문서로 처리할 수 있다. 셋째, 유사한 DTD를 갖는 문서를 재분류할 경우에는 기존에 생성된 DTD와 상이한 부분만을 추가하여 DTD를 재생성 할 수 있으므로 DTD의 중복 생성을 최대한 줄일 수 있다.

[참고문헌]

[1] James Clark, "Comparison of SGML and XML", *http://www.w3.org/TR/NOTE-SGML-xml.htm.*

[2] Joseph Giarratano and Gary Riley, *Expert Systems : Principles and Programming*, pp.373-499, 1989.

[3] Steven Holzner , *알기쉬운 XML*, 정보문화사, 1999.

[4] W3C, "Extensible Markup Language 1.0", *http://www.w3c.org, Recommendation*, 1998.

[5] *http://www.microsoft.com/XML/xsl/tutorial.*

[6] 박건일, 김유성,"멀티미디어 기술문서를 위한 자동 XML 변환기 개발", *한국정보과학회 추계 학술발표 논문집*, 제 26권 2호, 1999.

[7] 정회경, *XML 가이드*, 그린, 1998.