해외학술 프로시딩 DB 구축 프로세스 개선 및 구현

Improvement on foreign proceedings DB building process

여일연, 김선태, 김순영, 한희준, 윤희준, 예용희
한국과학기술정보연구원

Yeo il-yeon, Kim sun-tae, Kim soon-young, Han hee-jun, Yoon hee-jun, Yae yong-hee
Korea Institute of Science and Technology Information

요약

본 논문은 한국과학기술정보연구원의 해외저널 및 프로시딩 논문에 대한 국가과학기술전자도서관 NDSL에서 서비스 중인 IEEE 컨퍼런스 프로시딩 자료에 대한 DB 구축 프로세스를 개선하고, 그에 따른 시스템을 구현하였다. 각 출판사에서 제공하는 원시데이터 포맷이 각각 상이하고 수시로 변경됨에 따라 NDSL에서는 그에 따른 분석과 프로그램 수정이 매번 필요하게 되고, 그 시간만큼의 서비스 지연이 발생할 수밖에 없다. 본 논문에서는 원시데이터 포맷의 변화에 따른 프로그램의 의존도를 최소화하기 위해 프로그램을 모듈화 하여 원시데이터 포맷이 변경되어도 프로그램 상에서 변화되는 부분을 최소화하였다.

Abstract

This thesis improved the DB building process for IEEE conference proceedings that is being serviced in National Digital Science Library(NDSL) of KISTI and implemented the resulting system. Raw data format provided by the each publisher is different and change from time to time, so NDSL have to analyze and fix the programming. And it causes the service delays. This paper modularize the program to minimize dependence on the program due to the changes of raw data format.

I. 서 론

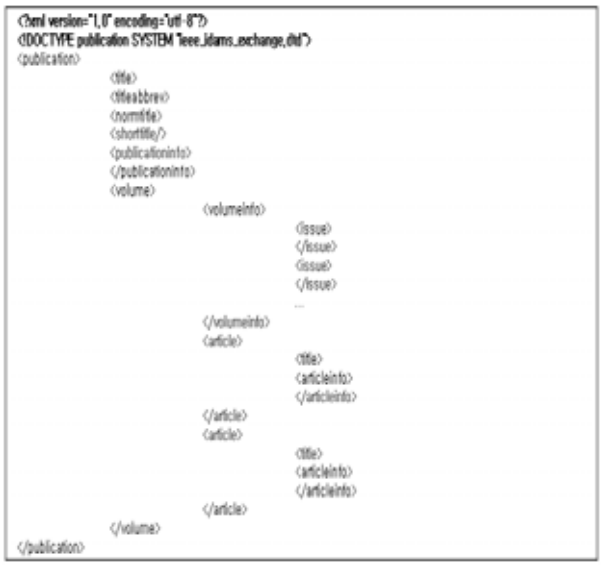
1. 개발배경

한국과학기술정보연구원에서 운영하고 있는 국가과학기술전자도서관 NDSL은 국내 학계, 연구계, 산업계의 모든 연구자를 위한 해외 학술 저널 및 프로시딩 포탈로서 56,000여 종의 학술저널과 188,000여 종의 프로시딩을 서비스하고 있다[1]. 그중 IEEE에서 출간하는 자료는 전자, 전기분야에서 가장 핵심이 되는 중요한 자료라고 볼 수 있다.

기존 IEEE에서 발간하는 컨퍼런스 프로시딩 전자자료의 경우 주간 단위로 발송되며 약 5천여 건의 데이터가 일반 태그 데이터 형태로 입수가 되었다. 그림 1참조. 하지만 인터넷 사용 환경이 발전함에 따라 원시데이터의 형태가 구조화된 태그 데이터인 XML형태로 추세가 흘러감에 따라 IEEE에서 발간하는 전자자료도 XML형태로 입수되게 되었다. 그림 2참조. 그림 2에서 보는 바와 같이 신규 입수된 XML 형태의 원시데이터는 하나의 저널 정보가 나열되고, 그 밑에 그 저널의 권, 호 정보가 표시된다. 그리고 그에 따른 여러 건의 기사정보가 반복적으로 나열되는 구조를 가지고 있다. 이렇게 입수되는 자료가 기존의 포맷에서 변화됨에 따라 IEEE 컨퍼런스 프로시딩의 전자도서관 서비스가 신규데이터를 반영하지 못하고 있는 실정이었다. 이러한 현상은 비단 IEEE자료에 국한되지 않고 모든 발행기관의 원시데이터에 해당되는 사항이다. 전자도서관과 같이 원시데이터를 생성하지 않고 그 데이터를 받아서 서비스 하는 업무의 특성상 원시데이터를 생성하는 각 발행기관에 종속적일 수밖에 없다. 이에 본 논문에서는 이러한 현상을 조금이나마 극복하고자 구조화된 XML 데이터를 가지고 데이터에 독립적인 DB 구축 프로그램을 구현하고자 한다.

```
|IELDVD000|10508|33281|1574063|5|2005||
(201) Procurement and Supply of Safety Related Systems
(265) 0-86341-577-6
(261) 0537-9989
(305) IEE
(110) 2005
(120) 9 Nov. 2005
(010) Journal Paper
(210) Blank page
(415) 0_2
(416) 0_2
```

▶▶ 그림 1. 기존 원시데이터



▶▶ 그림 2. 신규 원시데이터

II. 본론

1. 프로그램 구조

본 프로그램은 Java-API를 이용하여 원시데이터 가공프로그램을 설계하였다. 자바에서 제공하는 DOM이나 SAX를 이용하여 구현할 수도 있으나, 문헌정보에 사용되는 데이터의 특성상 추출한 엘리먼트들에 대한 후처리, 가공단계가 너무나도 많이 요구되기 때문에 본 프로그램에서는 일반적인 태그추출법을 사용하여 엘리먼트들의 값을 취하도록 하였다.

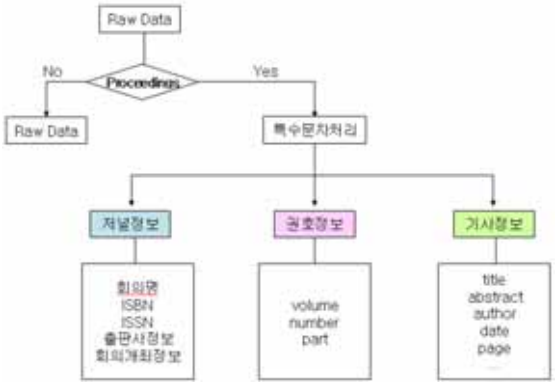
1.1 구성

프로그램은 크게 4부분으로 나뉜다. 첫 번째 부분은 입수된 원시데이터를 컨퍼런스 프로시딩자료와 그 외의 자료로 분리해 내는 과정이다. 매주 IEEE로부터 입수하는 자료에는 Journal Article과 Conference Proceeding Paper가 섞여서 들어오기 때문에, 입수된 데이터가 컨퍼런스 프로시딩 자료인지를 판단해 구분해 내는 과정이 필요하다.

두 번째 과정은 분리해낸 프로시딩 자료를 전처리 하는 과정이다. 즉 XML에 포함된 특수문자나 NDSL에서 서비스하기 위해 처리하여야 할 문자, 기호들을 처리하는 과정이다.

세 번째 부분은 입수된 원시데이터를 저널, 권호, 기사부분으로 분리하는 부분이다. 논문의 서론부분에서 밝혔듯이 IEEE로부터 입수되는 원시데이터는 하나의 파일에 하나의 저널정보, n개의 권호정보, m개의 기사정보로 나뉘어져 있다. 최종적으로 서비스될 기사정보에는 그 기사에 해당하는 저널정보와 권호정보가 필요로 하기 때문에 분리된 저널정보와 권호정보는 따로 저장 될 필요가 있다.

최종적으로 분리된 저널, 권호, 기사정보들의 엘리먼트들을 파싱해 내는 과정이 필요하다. 전체적인 흐름도는 그림 3에 표현되어져 있다.



▶▶ 그림 3. 프로그램 흐름도

1.2 오류사항

본 프로그램에서 에러로 처리한 항목들은 일반적인 문헌정보에서 오류로 처리되는 저널명 누락, 기사명 누락, 발행년도 누락, 저자명 누락, 권호정보 오류 등의 항목과 각종 표기 오류 등을 포함하였다.

1.3 모듈화

본 논문에서는 입수되는 원시데이터의 형식에 좀 더 독립적일 수 있는 프로그램을 구현하고자 하였다. 이에 각 출판사 마다 달라질 수 있는 데이터의 표현방법이나 데이터의 구조 등에 최소한의 영향을 받을 수 있도록 프로그램을 모듈화 하여, 데이터의 변경에 따른 프로그램의 영향을 최소화 하고자 하였다.

입수되는 원시데이터가 구조화된 XML문서라는 가정 하에 데이터를 저널정보, 권호정보, 기사정보로 분리하여 저장하고 각 기사정보들은 기사건수만큼의 루프를 거침으로써 IEEE와 같은 하나의 문서에 여러 건의 기사정보가 포함된 구조의 문서뿐만 아니라 하나의 문서에 한건의 기사가 포함된 문서에도 쉽게 적용이 가능하게 된다. 또한 각 엘리먼트를 추출하는 부분이나 엘리먼트의 값을 후처리 하는 부분도 개별 함수로 구분되어 있어 필요시에만 호출하여 사용할 수 있도록 하였다.

III. 결 론

문헌정보를 처리하는 프로그래밍은 입수되는 서지정보의 형태에 따라 매우 의존적일 수밖에 없다. 본 논문에서는 입수되는 원시데이터에 보다 독립적이고자 시스템을 구현하였다. 각 정보들을 처리하는 부분을 모듈화 하였고, 저널, 권호, 기사단

위로 나누어 처리, 저장함으로써 다른 출판사에서 입수되는 XML문서에 대해 보다 유동적으로 처리가능하게 하였다. 앞으로 좀더 연구개발해 나가야 할 부분은 본 논문에서 구현한 명령어 기반의 프로그램을 사용자가 보다 더 쉽게 운영할 수 있도록 GUI기반의 CS프로그램으로 개발하여 프로그램에 미숙한 문헌정보 처리 담당자들이 쉽게 사용할 수 있도록 하여야겠다.

▌참 고 문 헌▐

[1] http://www.ndsl.or.kr