데이터마이닝 기법 적용을 위한 공용 XML 구조 추출 알고리즘

장민석* · 방현진**

Common XML Structure Extracting Algorithm for Applying Data Mining Techniques

MinSeok Jang* · HyunJin Bang**

Dept. of Computer Information Science, Kunsan National University

E-mail : msjang@kunsan.ac.kr*, bbangfamily@hanmail.net**

요 약

현재 구조화된 데이터 표현의 표준으로 XML 언어가 일반화되고 있는 경향으로 인해 데이터 마이닝 대상으로서의 XML의 중요성이 점증하고 있는 실정이다. 특히 XML 문서에 연관규칙(association rule)을 적용함으로써 원하는 정보를 추출하는 연구가 진행되어 왔다. 하지만 마이너가 유사한 XML 문서들로부터 효율적으로 정보를 얻어내는 방법에 대한 문제에 대해서는 별 진전이 없었다. 본 연구에서는 다양한 XML Schema를 적용하는 유사한 XML 문서들로부터 공용 XML 구조를 추출하는 방법을 제안하고자 한다. 이러한 공용 XML Schema는 다양한 XML 구조를 단일화함으로써 우리가 원하는 정보를 정확하고 효율적으로 얻어낼 수 있도록 도와주는 데이터 마이닝의 사전 작업으로서 중요하다고 판단된다. 본 논문에서는 다양한 XML Schema를 적용하는 유사한 XML 문서들로부터 공용 XML 구조를 추출하는 방법을 제시한다.

ABSTRACT

Importance of XML as a target of Data Mining is growing because XML is used generally as a standard markup language for describing structured data. Especially researches have been done about extracting wanted informations by applying association rules to XML documents. But there are few development about solving the problems of method for efficiently obtaining informations from similar kinds of XML documents. To solve the problem this paper tries to suggest the method by which common XML structure is extracted form the same kinds of XML documents having a various XML schemas. The resulted schema structure is supposed to be important one as a preliminary job because it helps us to acquire the useful informations from various kinds of documents by unifying their structures.

키워드

XML, Data Mining, XML Schema, 공용구조, 연관규칙

Ⅰ. 서 론

최근 XML이 인터넷 문서 교환의 표준으로 채택 되면서 XML을 효율적으로 관리하기 위한 연구가 계속되고 있다. 더불어 XML은 데이터 마이닝의 대상으로서 부각되고 있다. XML은 문서교환의 표준으로서 많은 분량의 정확한 데이터를 산출할 수 있다. 반면에 사용자가 임의로 엘리먼트[1]를 정의할 수 있는 특징으로 인해 데이터 산출에 어려움이 있다. XML은 엘리먼트 아래 하위 엘리먼트를 가지는 계층 구조를 가짐으로써 데이터 마이닝 분야 뿐만 아니라 정보검색, 문서관리 시스템에 커다란 영향을 미쳐왔다. 따라서 XML 관련 저장 기법, 인덱싱 기법, 질의어, 그리고 질의 최적화에 이르기까지 많은 연구가 진행되고 있다[2].

그러나 이러한 연구들은 유사한 문서를 대상으로 하나의 XML Schema를 공유하는 것이 대부분이다. 따라서 다양한 구조를 가진 문서를 대상으로 하는 경우, 바로 마이닝을 위한 준비 자료로서

설계할 것이 아니라 문서들의 구조의 공유 정도를 미리 파악하는 일은 매우 중요하다. 또한 문서의 유사한 정도에 따라 문서를 분류하거나 클러스터링 하고자 하는 경우에도 XML의 큰 장점인 내재된 구조를 특징으로 하여 문서간의 공유 특징을 추출[3]하고 이를 토대로 기준을 갖춘 XML Schema를 적용하여 정제된 XML 문서로 구조 변환함으로써 데이터 마이닝을 위한 사전작업이 되어야 한다.

II. 본 론

마이닝을 위한 기존 XML 데이터들은 이질적인 데이터 구조로 인해서 애플리케이션에 적용하기가 어려웠다. 데이터 마이닝은 주어진 자료를 분석해서 마이닝의 결과를 보여준다. 마이닝에 적합한 자료는 같은 포맷을 유지하고 있어야 한다. 이전의 마이닝 자료로서의 문서들은 유사하지만 이질적 구조로 이루어져 있는데 이러한 자료들로부터 효율적인 마이닝에 대한 결과를 기대하기는 어려웠다.

XML은 다른 비구조적인 문서와는 달리 구조정보를 내재하고 있다. 기존 연구에서는 XML구조간의 최대 유사 경로를 구하기 위해 연관규칙(association rule)[4][5][6]을 적용하는 경우가 많았다. 유사한 문서들에 대해서 XML의 엘리먼트의 경로를 추출하는 과정이 있다. 매핑테이블에 의해 추출된 경로들을 최대 유사 경로로 묶어주는 작업이지만 결국 유사한 경로를 묶어 줌으로 서로 유사 정도만 비교하였을 뿐이다. 본 논문에서는 그런 유사한 경로들을 가진 문서들에 대해서 마이닝에 적합한 형태인 XML문서로 변환 하는 과정을 연구한다.

표 1. 매핑테이블

element	mapping	element	mapping
book	book, books	publisher	publisher, pub
bookinfo	bookinfo, info	isbn	isbn, number
title	title	review	riview, riviews
author	author, writer	pubdate	pubdate, date
firstname	firstname,fname	section	section, sec
surname	surname,sname	price	price
editor	edtior, edtion	chap	chap, chapter
copyright	copyright, copy	para	para
legalnotice	legalnotice, legal	revhistory	revhistory, revision

III. 공용 XML 구조 추출 알고리즘

XML은 애플리케이션의 종류와 상관없이 데이터를 공유할 수 있는 방법을 제공한다. 그것은 업계에서 데이터를 교환하기 위해 공통의 XML Schema를 사용하기로 동의하기 때문이다. 자신이 개발한 애플리케이션에서 다른 애플리케이션의 데이터를 사용하고자 할 때 받아들일 데이터 검증을 위해 표준 XML Schema를 이용할 수 있을 것이다. 또한 자신의 데이터를 검증하기 위해서도 XML Schema를 이용할 수 있을 것이다. 마이닝을 위해서 자신의 데이터들을 XML Schema를 이용해서 변환 할 필요가 있다. 이런 정제된 데이터를 통해서 효율적인 마이닝이 이루어질 것이다.

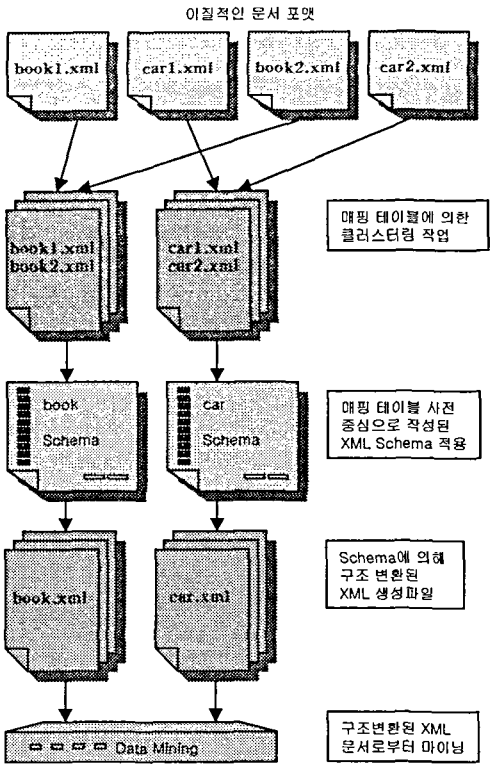

그림 1. 문서구조변환과정

그림 1은 이질적인 문서들을 정제된 공용의 XML문서로 변환하는 하나의 과정이다. 총 4단계를 거쳐 마이닝을 위한 준비가 완료된다. 먼저 각기 다른 이질적인 XML 문서들로부터 매핑테이블(공용으로 사용될 수 있는 항목에 대해 만들어 놓은 사전)을 이용한 클러스터링 작업을 한다. 이 클러스터링 작업은 크게 분류하여 BOOK과 CAR에 대하여 분류했다. 이 작업은 XML 문서의 태그 경로 중 첫 번째 경로인 book과 car에 대해서 분류되었다. 매핑 테이블에서 book을 books로 볼 수 있다.

표 1과 같이 매핑테이블은 book과 books가 같다고 본다. 이런 매핑테이블은 사용자-정의 사전을 구축하여 자동화할 수 있다. 애플리케이션에서 훨씬 많은 매핑 사전을 등록 할수록 문서 분류와 변환이 더욱 효율적이 될 수 있다. 두 번째

로 매핑테이블 사전 중심으로 작성된 XML Schema로 클러스트링된 문서들을 스키마에 맞게 구조 변환시킨다. 그러면 마이닝에 사용할 수 있는 XML 결과 문서가 생성된다. 생성된 XML결과 문서를 통해 효율적인 데이터 마이닝을 할 수가 있다. 다음 장에서는 이질적 XML문서 포맷을 경로 추출후 매핑테이블에 어떻게 적용하는지 설명한다.

IV. 공용 XML 포맷 전처리

XML 문서 포맷의 경로는 트리 구조로 되어 있다고 생각하면 쉽게 추출된다. 그림 2를 보면 book 엘리먼트 뒤에는 bookinfo와 그 뒤에는 author가 나온다. bookinfo와 author는 book의 하위노드라고 생각하면 된다. 따라서 book1.xml 파일의 author의 경로를 간단하게 표기하기 위해 book.bookinfo.author 라고 하자. 그럼 두개의 문서를 비교해 보자. book1.xml과 book2.xml문서는 유사한 듯 보이지만 엘리먼트가 이질적이다. 표 1의 매핑 테이블에 의해 첫 번째 경로 book과 books는 같다고 볼 수 있다. 그리고 book2.xml 문서는 bookinfo라는 엘리먼트가 없다. 하지만 book1.xml에서 author에 매핑되는 엘리먼트인 writer가 존재한다. 결국 author와 writer는 같은 엘리먼트이므로 경로로 표시하면 book.bookinfo.author는 books.writer와 같다. 이런 이질적 문서이지만 내용은 매핑테이블에 의해서 하나로 묶을 수가 있다. 한 가지 더 생각해야 할 점은 엘리먼트 경로가 다른 문서들을 같은 구조의 스키마로 어떻게 변환 할 것인가라는 것이다.

표 2는 book1.xml과 book2.xml 파일의 경로를 간단히 나열해보았다. 그림 2의 문서들을 그대로 경로로 나타내 보았는데 이런 간단한 예를 보면 쉽게 이해 할 수 있다. 두 문서를 기준 스키마 포맷에 맞춰서 하나의 공통된 문서로 바꿀 수가 있다. 표 1의 매핑 테이블과 기준 스키마는 이질적 두 문서를 공용으로 사용할 수 있도록 해준다.

엘리먼트의 이름과 다른 경로구조를 가지는 두 문서를 같은 이름의 구조로 변환하는 과정에 대해서 다음 장에서는 두 문서를 기준 스키마를 통한 xml결과 문서로 생성되는 과정을 설명한다.

V. XML 결과 문서 생성

서로 다른 문서 간에서 공용할 수 있는 엘리먼트를 통해 하나의 구조로 나타낼 수 있음을 전 장에서 설명하였다. 그럼 이번 장에서는 그 구조를 통일 시켜주는 스키마에 대해서 살펴보고 마이닝에 효율적인 XML 결과 문서를 설명하겠다.

```
<?xml version="1.0"?>
<book>
  <bookinfo>
    <author>
      <firstname>Jorge</firstname>
      <surname>Godoy</surname>
    </author>
    <editor>
      <firstname>Jorge</firstname>
      <surname>Godoy</surname>
    </editor>
    <copyright>
      <year>2000</year>
      <holder>Conectiva</holder>
    </copyright>
    <title>Network and Systems
    Administration Using Linux</title>
    <legalnotice>
      <para>The contents of this book can
      be freely used and distributed as far
      as the source is mentioned as a reference
      that is, its bibliography.</para>
    </legalnotice>
```

book1.xml

```
<?xml version="1.0"?>
<books>
    <title>How to XML</title>
    <writer>
      <firstname>Sangeun</firstname>
      <surname>An</surname>
    </writer>
    <edition>
      <firstname>Sungkuk</firstname>
      <surname>Kim</surname>
    </edition>
    <copyright>
      <year>2005</year>
      <holder>KNU.XMLAB</holder>
    </copyright>
    <para>XML was introduced to overcome the
    limitations of HTML.</para>
```

book2.xml

그림 2. 이질적 문서

스키마는 이질적인 XML 문서를 구조화함으로 데이터교환에 아주 좋은 이점이 있다. 그럼 스키마는 본 논문에서 어디에 활용되는지 알아보자.

표2의 경로 테이블을 보면 엘리먼트 경로 구조가 두 문서 간에 이질적임을 알 수가 있다. 이질적인 두 문서를 공용 포맷으로 만들기 위해 바로 잡아주는 것이 필요할 것이다. 그때 필요한 것이 기준이 될 스키마이다. 경로 이름과 경로 구조가 다른 두 문서를 하나의 공용 포맷으로 만들기 위해서 사전 준비된 기준 스키마를 사용자가 작성한다. 그림 3은 사전 준비된 기준 스키마이다. 애플리케이션 프로그램에 의해서 이질적인 각각의

문서들은 표 1의 매핑테이블과 표 2의 경로 테이블에 의거하여 이 기준이 될 스키마에 맞게 구조 변환이 이루어진다.

표 2. 문서의 경로추출

Doc.	paths
book1.xml	book.bookinfo.author.firstname
	book.bookinfo.author.surname
	book.bookinfo.editor.firstname
	book.bookinfo.editor.surname
	book.bookinfo.copyright.year
	book.bookinfo.copyright.holder
	book.bookinfo.title
	book.bookinfo.legalnotice.para
book2.xml	books.title
	books.writer.firstname
	books.writer.surname
	books.edition.firstname
	books.edition.surname
	books.copyright.year
	books.copyright.holder
	books.para

에플리케이션에서는 사용자에 의해 매핑테이블을 작성 또는 매핑테이블을 사전 작성된 공용 규격 폼을 사용하며 애플리케이션은 불러온 문서들에 대해서 경로추출테이블을 작성하게 된다. 애플리케이션은 분석된 경로테이블에서 클러스터링 작업이 이루어진다. 그 후에 사전에 준비된 기준 포맷이 될 스키마에 의해서 그 문서들이 XML 문서 변환작업이 이루어지게 된다. 애플리케이션은 XML 결과 문서를 생성하고 마이닝의 준비물을 완료하게 된다.

그림 3의 기준 스키마에 의해 작성된 XML 결과 문서는 표 2의 경로테이블이 다음 표 3과 같이 경로 구조가 같게 나타난다. 결국 두 문서의 이질적인 경로 구조가 정해진 스키마에 의해 획일적으로 나타났다.

표 3. XML결과문서의 경로테이블

Doc.	paths
book1'.xml book2'.xml	book.bookinfo.title
	book.bookinfo.author.firstname
	book.bookinfo.author.surname
	book.bookinfo.editor.firstname
	book.bookinfo.editor.surname
	book.bookinfo.copyright.year
	book.bookinfo.copyright.holder
	book.bookinfo.legalnotice.para

이처럼 book1.xml과 book2.xml 문서는 그림 4처럼 다시 작성이 될 것이다. 매핑테이블과 기준 스키마에 의해서 새로운 XML문서가 생성이 된다. XML결과 문서는 같은 구조를 가지고 각 엘리먼트에는 매핑테이블에 맞게 적재적소하게 될 것이다. 결국 내용은 유사하지만 구조와 엘리먼트 이름이 달라서 마이닝에 효율적이 못한 문서들을 마이닝을 위한 효율적인 문서로 바꾸는 과정을 설명하였다.

```
<xsd:schema xmlns:xsd="http://www.w3c.org/2001/XMLSchema">
<xsd:element name="book">
<xsd:complexType>
 <xsd:sequence>
  <xsd:element name="bookinfo">
   <xsd:complexType>
    <xsd:sequence>
     <xsd:element name="title" type="xsd:string"/>
     <xsd:element name="author" maxOccurs="unbounded">
      <xsd:complexType>
       <xsd:sequence>
    <xsd:element name="firstname" type="xsd:string"/>
    <xsd:element name="surname" type="xsd:string"/>
       </xsd:sequence>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="editor" maxOccurs="unbounded">
      <xsd:complexType>
       <xsd:sequence>
        <xsd:element name="firstname" type="xsd:string"/>
    <xsd:element name="surname" type="xsd:string"/>
       </xsd:sequence>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="copyright">
      <xsd:complexType>
       <xsd:sequence>
        <xsd:element name="year" type="xsd:string"/>
    <xsd:element name="holder" type="xsd:string"/>
       </xsd:sequence>
      </xsd:complexType>
     </xsd:element>
     <xsd:element name="legalnotice">
      <xsd:complexType>
       <xsd:sequence>
        <xsd:element name="para" type="xsd:string"/>
       </xsd:sequence>
      </xsd:complexType>
     </xsd:element>
```

그림 3. 기준 스키마

VI. 결 론

XML이 일반화되고 있는 가운데 데이터 마이닝의 대상으로서 XML이 공용될 수 있도록 서로 이질적인 문서들로부터 엘리먼트를 통일시켜 주고 스키마를 조절함으로 데이터마이닝에 효율적 결과를 나타낼 수 있도록 해보았다. 본 논문의 공용 XML 구조 추출은 마이닝을 위한 전처리 과정에서 사용될 예정이다.

```
<?xml version="1.0"?>
<book>
  <bookinfo>
    <title>Network and Systems
    Administration Using Linux</title>
    <author>
      <firstname>Jorge</firstname>
      <surname>Godoy</surname>
    </author>
    <editor>
      <firstname>Jorge</firstname>
      <surname>Godoy</surname>
    </editor>
    <copyright>
      <year>2000</year>
      <holder>Conectiva</holder>
    </copyright>
    <legalnotice>
      <para>The contents of this book can
      be freely used and distributed as far
      as the source is mentioned as a reference
      that is, its bibliography.</para>
    </legalnotice>
```

book1'.xml

```
<?xml version="1.0"?>
<book>
  <bookinfo>
    <title>How to XML</title>
    <author>
      <firstname>Sangeun</firstname>
      <surname>An</surname>
    </author>
    <editor>
      <firstname>Sungkuk</firstname>
      <surname>Kim</surname>
    </editor>
    <copyright>
      <year>2005</year>
      <holder>KNU.XMLAB</holder>
    </copyright>
    <legalnotice>
      <para>XML was introduced to overcome the
    limitations of HTML.</para>
    </legalnotice>
```

book2'.xml

그림 4. XML 결과 문서

감사의 글

본 연구는 한국과학재단 목적기초연구 *(R01-2004-000-10946-0)* 지원으로 수행되었음.

참고문헌

[1] W3C 웹사이트, http://www.w3.org
[2] 이정원, 이기호, "XML 공유 구조 발견을 위한 변형 순차패턴 마이닝 알고리즘", 2002 한국정보과학회 가을 학술발표논문집, Vol.29. No.2 pp1~3, 2002.
[3] Jung-Won Lee, Kiho Lee, Won Kim, "Preparations for Semantics-based XML Mining", In Proc. of the ICDM '01, Nov. 2001.
[4] Amnon Meisels, Michael Orlov and Tal Maor, "Discovering Associations in XML Data", In Proc.of the DASWIS'02, 2002.
[5] 이정원, 김호숙, 최지영, 김현희, 용환승, 이상호, 박승수, "데이타마이닝 알고리즘의 분류 및 분석", 한국정보과학회 논문지:데이타베이스, 제28권, 제3호, pp279~300, 2001년 9월.
[6] D Braga, A Campi, S Ceri, M Klemettinen, P Luca Lanzi, "A Tool for Extracting XML Association Rules", In Proc. of the ICTAI'02, pp57~64, 2002.