식물의 생태 정보를 위한 XML 스키마 자동 생성 기법

김석철[O†] 남윤영[†] 황인준[††]
아주대학교 정보통신전문대학원[†], 고려대학교 전자컴퓨터공학과[††]
{radiopd[O],youngman}@ajou.ac.kr[†], ewhang04@korea.ac.kr[††]

A Method of Automatic XML Schema Generation for Botanical Ecology

Suck-Chul Kim[O†] Yunyoung Nam[†] Een-Jun Hwang[††]
Graduate School of Information and Communication, Ajou University[†]
Department of Electronics and Computer Engineering, Korea University[††]

요 약

XML은 문서를 구조적으로 표현 가능한 언어이며, 주로 인터넷 상에서 데이터를 교환하는 목적으로 널리 쓰이고 있는 표준이다. 이러한 장점으로, RDBMS에 저장되어 있는 다양한 데이터를 XML 형태로 변환하기 위한 연구가 진행되어 왔다. 본 논문에서는, RDBMS에 저장된 데이터를 XML로 변환하기 위해 XML 스키마를 자동으로 생성하는 기법을 제안한다. XML로 변환하기 위해, 식물의 생태 정보 특징을 이용하여 XML 스키마를 자동으로 생성하였으며, 생성된 스키마를 통해 식물 정보를 XML로 변환하였다. 또한, XML 문서 생성 시에 사용자가 원하는 스키마를 선택적으로 이용할 수 있도록 하여, 동적으로 XML 문서를 생성할 수 있도록 하였다. 또한, RDBMS에 저장된 데이터의 XML 문서 전환에 소요되는 시간을 줄이기 위해 XML 문서에서 엘리먼트를 선택적으로 생성할 수 있도록 하였다.

1. 서 론

XML(eXtensible Markup Language)은 SGML(Standard Generalized Markup Language)과 HTML(Hypertext Markup Language)의 장점을 가진 마크업 언어로써 인터넷 상에서 정보를 교환하기 위한 목적으로 널리 사용되는 마크업 언어이다. 또한 국제 표준이고 플랫폼에 독립적이며 분산된 데이터를 통합하는데 유용하게 이용할 수 있는 장점을 가진다. 하지만 현재 대부분의 데이터는 MS-SQL Server, Oracle와 같은 RDBMS(Relational Database Management System)에 저장되어 있기 때문에 XML로 변환하는 과정이 필수적이다. 지금까지, RDBMS에 저장되어 있는 데이터를 XML 문서로 변환하기 위해 많은 연구가 있었으며, 이러한 변환을 위해, DTD나 XML 스키마가 이용될 수 있다[1]. 그러나 기존의 RDBMS에 저장되어 있는 데이터를 XML 문서로 변환하기 위해 스키마를 사용자가 일일이 정의하는 것은 많은 문제점이 발생한다. 또한, 식물 정보와 같이 사용자의 필요에 따라 선택되는 정보가 다를 수 있는 데이터의 경우 미리 정의된 스키마로 XML 문서를 생성할 때 사용자가 필요로 하는 정보가 누락되거나 불필요한 정보가 XML 문서에 포함될 수 있다.

본 논문에서는 XML 문서를 보다 빠르고 효율적으로 생성하기 위하여 RDBMS에 저장되어 있는 식물의 특징을 이용하여 XML 스키마를 자동적으로 생성하고, 이 XML 스키마를 기반으로 XML 문서를 선택적으로 생성할 수 있는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 살펴보며, 3장에서는 XML 스키마 및 XML 문서의 생성에 대해 서술하였으며, 마지막으로 4장에서는 결론 및 향후 연구방향에 대해서 제시한다.

※ 본 연구는 과학기술부 국책연구 개발 사업인 유전자원지원 활용사업단의 연구비 (no. BDM0100211)의 지원에 의해 수행되었습니다.

2. 관련 연구

2.1. AutoMed

AutoMed[3]는 기존의 DTD나 XML 스키마와는 다른 독창적인 XMLDSS(XML DataSource Schema)라는 스키마 언어를 정의하였다. 그림 1에서 보는 바와 같이 각각의 데이터 소스는 로컬(local) 스키마로 표현이 되고, 이 local 스키마는

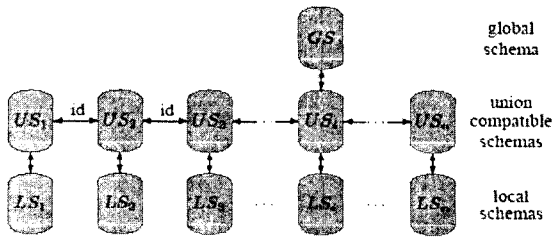

그림 1. AutoMed의 시스템 구조도

union-compatible 스키마로 전환이 된다. union-compatible 스키마는 다시 최종적인 작업을 통해서 글로벌(global) 스키마로 전환되는 구조를 가지고 있다. AutoMed는 XML 문서에서 스키마를 생성하는 과정이 반자동적이라는 조건에서 스키마의 재구성은 자동적으로 이루어지는 구조를 가지고 있다.

2.2. XClust

XClust[4]는 DTD를 기반으로하는 스키마 클러스터링을 구현하고 있다. XClust는 그림 2에서 보는 바와같이 DTD 문서들 사이의 유사도를 기준으로 DTD를 통합한다. 유사도를 계산하는 기준에는 DTD 문서들 간의 엘리먼트의 이름(약어, 동의어, 유의어), 엘리먼트의 path, DTD 문서의 기본구조와 같은 것들이 있다.

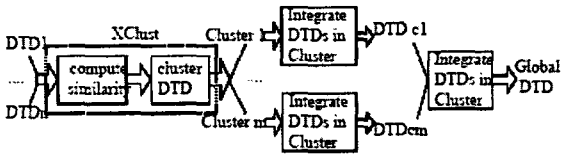

그림 2. XClust의 시스템 구조도

2.3. 식물 정보의 특징

식물 정보는 계층적인 특징을 가지고 있다. 기존의 생물자원 정보의 표준화 작업의 한 예를 보면, 식물정보는 종정보와 내용정보로 나뉘는데 각각의 정보의 단위를 구성하고 있는 XML DTD의 설계의 기본 단위들이 계층적인 특징을 가지고 있다[5]. 예를 들어서 Species 2000[6]과 ITIS[7]에서 제공하고 있는 종 정보 검색 내용의 DTD 설계를 보면 생물종 DTD는 kingdom, subkingdom, phylum, subphylum, genus, subgenus 등의 XML DTD 내에서의 계층적인 특징을 가지고 있다.

따라서 본 논문에서는 식물의 계층적인 특징을 이용하여 RDBMS에 저장되어 있는 데이터를 사용자 정의된 스키마를 생성하지 않고 자동적으로 XML 스키마를 생성하고, 생성된 스키마를 바탕으로 사용자가 원하는 XML 문서를 선택적으로 생성할 수 있게 구현하였다.

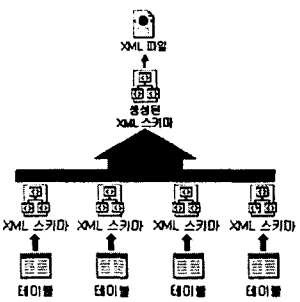

그림 3. XML 생성 과정

3. 식물 정보의 XML 생성 시스템의 구현

3.1. XML 문서 생성 과정

XML 문서 생성 과정의 전체 모습은 그림 3과 같다. 먼저 RDBMS에 저장되어 있는 테이블에서 자동적으로 테이블 단위의 XML 스키마를 생성한 후, 시스템에서 생성된 XML 스키마 사이의 일정한 기준을 바탕으로 XML 스키마를 클러스터링한다. 마지막으로 클러스터된 스키마를 바탕으로 사용자가 원하는 형태의 XML 문서를 선택적으로 생성할 수 있게 한다. 본 논문에서는 XDR 스키마[1]를 사용하였다.

3.2. 각 RDBMS 테이블에서의 XML 스키마 자동 생성

XML 스키마 자동 생성 과정은 표 1과 같다. 먼저 각 테이블에 해당하는 모든 필드를 엘리먼트로 설정을 한다. 다음으로는 각 필드가 가진 이름의 특징을 기반으로 각 필드를 클러스터링한다. 예를 들어, 필드명이 학명, 국명, 이명, 수중엽, 수중엽_크기와 같은 형식으로 명명이 되어 있다고 가정을 하면 같은 엘리먼트 단위로 클러스터링되는 필드는 prefix, postfix가 같은 단어를 중심으로 (학명, 국명, 이명), (수중엽, 수중엽_크기)등과 같은 형태로 클러스터링 된다.

표 1. XML 스키마 자동 생성 과정

```
1: 각 테이블 단위의 모든 필드를 엘리먼트로 설정
2: Postfix, Prefix를 통한 필드의 클러스터링
3: 클러스터링된 필드를 통한 엘리먼트 및 속성의 설정
4: 루트 엘리먼트의 설정
```

다음으로는 클러스터링된 필드의 이름을 바탕으로 루트 엘리먼트와 속성으로 나눈다. 나누는 과정은 표 2와 같다. 루트 엘리먼트와 속성으로 나누는 대상은 클러스터링된 각각의 필드의 개수가 가장 많은 묶음을 대상으로 한다. 대상이 되지 않는 다른 필드들은 루트 엘리먼트의 하위 엘리먼트로 설정을 한다. 필드의 이름이 같은 prefix로 클러스터링된 경우에는 prefix를 제외한 필드명의 이름의 길이가 가장 작은 필드를 루트 엘리먼

표 2 엘리먼트, 속성 설정 알고리즘

```
1: 클러스터링된 필드들의 묶음을 클러스터링한 기준을 중심
   으로 묶음별로 나눈다
2: RootNode = FindMaxNumofClustered(각 묶음);
3: If(RootNode의 클러스터링 기준)==prefix
4: 묶음.BOF = 엘리먼트;
5: while(RootNode.EOF)
6: {(RootNode의 필드들==(primary key or min(묶음의prefix
   를 제외한 postfix의 길이))?엘리먼트:Attribute);}
7: else
8: 묶음.BOF = 엘리먼트;
9: while(묶음.EOF)
10:{(각 묶음의 필드들==(primary key or min(묶음의postfix
   를 제외한 prefix의 길이))?엘리먼트:Attribute);}
```

트로 설정한다. 또한, 필드의 이름이 같은 postfix로 클러스터링된 경우에는 가장 처음으로 읽힌 필드가 루트 엘리먼트가 되고 나머지 필드들은 루트 엘리먼트의 속성이 된다. 위 두 가지 경우, 필드들 가운데 기본키(primary key)가 있으면 그 필드가 루트 엘리먼트 필드가 되고 나머지 필드는 속성이 된다. 그림 4는 이와 같은 과정의 예를 보이고 있다.

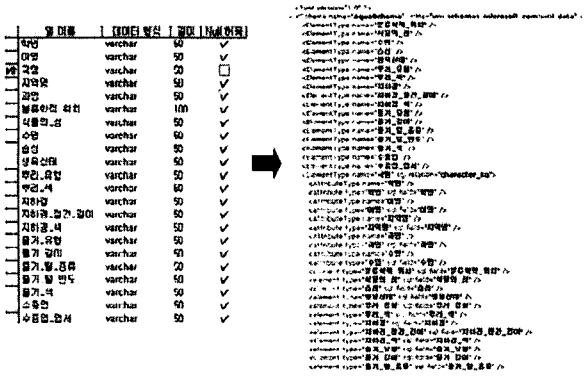

그림 4. XML 스키마 자동 생성의 예

3.3. XML 스키마 간의 전환

3.2와 같은 과정을 통해서 만들어진 각각의 테이블 스키마는 표 3과 같은 과정을 통해서 식물 정보에 대한 스키마로 전환이 된다.

표 3 스키마 간의 전환 과정

```
1: 전환될 스키마 최상위 노드 = 식물종 정보를 가진 스키마
2: While(각 스키마.EOF)
3: {If(각 스키마 최상위 엘리먼트 필드 이름.Contains(최상
   위 스키마 노드의 필드 이름))
4:     최상위 스키마.AddSubNode(각 스키마);
5:  Else
6:     최상위 스키마.AddDirectSubNode(각 스키마);}
```

먼저 전환이 된 스키마의 최상위 노드는 식물종의 정보를 가진 스키마로 선정을 한다. 식물종의 정보를 가진 스키마는 스키마의 정보를 횡적 구조로 표현하기에 적합하기 때문에 전환이 되는 스키마의 최상위 노드로 선정을 하였다[5]. 최상위 노드가 선정이 된 후, 만약 각각의 스키마에 해당하는 최상위 엘리먼트 필드의 이름이 최상위 노드가 가진 필드의 이름과 동일하다면 필드를 기준으로 동일한 이름을 가진 최상위 노드 스키마의 서브노드로 병합한다. 만약 일치하는 필드의 이름이 없을 경우에는 최상위 노드 스키마 루트 엘리먼트의 direct 서브노드로 스키마를 합친다. 이러한 과정을 통해서 스키마의 계층성이 표현 될 수 있다.

3.4. XML 문서의 선택적 생성

식물 정보 데이터는 일반적인 다른 데이터베이스와는 달리 정보의 질이 전문적인 정보와 일반적인 정보로 나뉠 수 있다. 따라서 시스템은 사용자가 원하는 XML의 문서를 선택적으로 추출할 수 있도록 지원을 해야 한다. 본 논문에서는 3.3과 같은 과정이 일어날 때마다 표시를 하여 XML 스키마를 바탕으로 XML 문서를 선택적으로 생성할 수 있도록 구현하였다. 그림 5에서와 같이 XML 스키마가 다른 XML 스키마로 전환이 될 때, 합쳐지는 최상위 노드의 엘리먼트들 이름을 하나의 배열로 저장을 하여 사용자가 그 스키마에 해당하는 대표 정보를 볼 수 있고 선택할 수 있도록 구현을 하였다. 예를 들어,그림 5에서 b, c, d의 순서로 전환이 이루어진다면 사용자는 각각 선택적으로 a~d 가운데 원하는 엘리먼트와 서브 엘리먼트를 선택할 수 있도록 구현함으로써 XML 문서의 재사용성에 효과적이다.

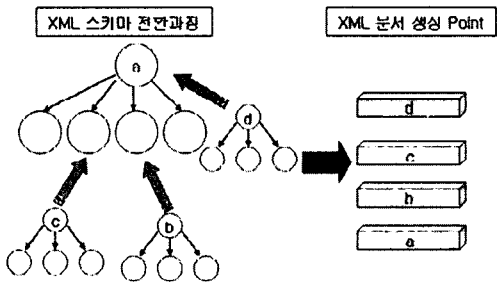

그림 5. 스키마 전환 시의 문서 생성 지점 설정의 예

4. 결론 및 향후 과제

본 논문에서는 사용자 정의된 DTD나 XML 스키마가 아닌 식물 정보의 계층성을 바탕으로 RDBMS에서 자동적으로 스키마를 생성하여, 이 스키마를 바탕으로 사용자 기반의 선택적 XML 문서의 생성을 할 수 있는 방법을 제안하였다. 또한 XML 스키마를 생성하는 데에 있어서 본 논문에서는 각 하나의 테이블을 단위로 생성을 하였다. 향후 과제로는 단일 테이블이 아닌, 서로 밀접한 관련이 있는 여러 테이블들이 있을 경우에, 테이블간의 관계를 통해 새로운 스키마 생성 기법을 개발하는 것이다.

5. 참고문헌

[1] http://msdn.microsoft.com
[2] R.Duckstein, K.Bohm, "Database Support for Species Extraction from the Biosystematics Literature - a Feasibility Demonstration," Technical Report
[3] L.Zamboulis, A.Poulovassilis, "Using AutoMed for XML Data Transformation and Integration," Springer-Verlag, LNCS Vol.3084, pp 58-69
[4] M.L.Lee, L.H.Yang, W.Hsu, X.Yang,"XClust: 클러스터링 ing XML Schemas for Effective Integration," ACM CIKM'02
[5] 이계준, 손강렬, 양진호, 윤희준, "컴포넌트를 이용한 XML 기반 생물자원정보관리시스템 설계," 지식정보인프라, 통권 10호, 2002
[6] Species 2000, "http://www.sp2000.org"
[7] Integrated Taxonomic Information System, "http://www.itis.usda.gov"