Journal of Korea Multimedia Society Vol. 7, No. 9, September 2004(pp. 1329-1338)

데이터준비를 위한 XML 기반의 분산 MDR 검색 시스템 설계

고석범[†], 윤성대[††]

요 약

데이터마이닝은 방대한 데이터로부터 다차원적인 정보를 추출하는 것이다. 방대하게 구축되어 있는 데이터베이스에서 임의의 테이블의 컬럼에 대해 참조 할 수 있는 정보는 단순하게 컬럼명과 자료형 혹은 간단한 주석 정도이다. 그러한 비구조적이고 빈약한 내용만으로는 데이터마이닝을 위한 자료수집 및 자료탐색 단계에서 컬럼의 용도와 특성 및 스키마를 파악하여 데이터를 정제하고 수집하는 것이 난해 할 뿐만 아니라 너무 많은 시간이 소요된다. 이러한 문제를 해결하기 위해 본 논문에서는 관계형 데이터베이스 환경에서 데이터준비 단계에 대부분의 시간을 소요하는 문제를 해결하기 위한 방안을 제안한다. 즉, 데이터준비단계에서 유용한 요소들을 메타데이터의 표준인 ISO/IEC : 11179 MDR (MetaData Registry) 규격에 맞는 표준 메타데이터를 제안하고, 이기종 및 이질 DBMS간에 호환 가능한 XML 기반의 분산 MDR 검색 시스템 구조를 제안한다.

A Design of XML-Based Distributed MDR Retrieval System for Data Preparation

Sucbum Ko[†], Sungdae Youn[††]

ABSTRACT

The purpose of data mining is to extract multi-dimensional information from a large database. The only information that we can extract from a large database is the column name, data type or simple comments included in the columns of database tables. With such unstructured and scarce information, it is very difficult and time taking to collect and to cleanse data by analyzing the purpose, characteristic and schema of the column during the data preparation step. In order to solve this problem, we propose solutions for reducing the time spent data preparation step in a relational database environment in this paper. That is, we propose useful elements to be considered during the data preparation step and then these elements are organized to constitute MDR(Metadata Registry) which is becoming the international standard of ISO/IEC : 11179. Finally, we propose a XML-based distributed MDR retrieval system that is convertible among heterogeneous systems and heterogeneous DBMSs.

Key words: Data Mining(데이터마이닝), Data Preparation(데이터준비), Metadata Registry(메타데이터 레지스트리), XML, Distributed Database System(분산 데이터베이스 시스템)

※ 교신저자(Corresponding Author) : 고석범, 주소 : 부산시 남구 대연3동 599-1(608-737), 전화 : 051)620-6398, FAX : 051)620-6398, E-mail : sbko@kma.ac.kr
접수일 : 2003년 12월 24일, 완료일 : 2004년 3월 19일
[†] 준회원, 육군사관학교 전자계산학과 전임강사
[††] 정회원, 부경대학교 전자계산학과 교수
(E-mail : sdyoun@pknu.ac.kr)

1. 서 론

데이터마이닝(Data mining)은 대량의 데이터베이스에서 잠재되어 있는 유용한 정보를 추출하는 기법 및 과정이다. 일차원적인 질의에 대한 일차원적인 결과를 산출하는 기존의 데이터베이스의 질의시스

템과는 달리, 데이터마이닝을 통한 잠재적인 정보의 추출은 현재의 데이터를 통한 통계적인 분석을 통해 미래를 예측하거나 데이터간의 상관관계를 분석하는 등의 일을 통해 다차원적인 결과를 산출하는 것이 가능하다[1,18].

데이터마이닝의 처리과정은 크게 데이터마이닝 자료를 수집 및 가공하기 위한 데이터준비단계(Data Preparation Step)와 준비된 데이터로부터 잠재적인 정보를 추출하는 데이터마이닝단계(Data Mining Step)로 수행된다. 실제 두 단계별 소요 시간 및 중요도는 데이터준비단계에 편중되어 있다. 즉, 데이터준비단계에서의 소요시간 및 중요도는 각각 80%와 95%로서 데이터마이닝 단계에서의 20%와 5%에 비해 월등히 높다[2,3]. 데이터준비단계에서의 소요시간은 데이터베이스 스키마 분석, 데이터 수집 및 샘플링, 오류 데이터의 처리, 데이터변환을 위한 시간이 포함된다. 데이터준비단계가 데이터마이닝 단계보다 중요한 이유는 데이터준비단계에서의 올바른 데이터의 수집 및 변환이 데이터마이닝 단계에서 올바른 결과를 도출하고, 수행시간을 단축시키는데 종속적인 영향을 미치기 때문이다. 여기에서, 올바른 결과의 도출이란 오류(Missing)나 왜곡(Distortion)된 데이터를 통해 잘못된 데이터마이닝 결과를 초래하는 것이고, 수행시간의 단축은 분산 환경에서 지역 사이트(Local Site)에 존재하는 각 데이터 스키마 분석 및 데이터의 유형 분석을 위해 소비되는 시간과, 데이터의 수집 및 정재를 하는 시간을 단축하는 것이다. 이를 위해 데이터준비단계에 필요한 정보를 참조할 수 있는 특성화된 메타데이터(Meta data)의 설계가 필수적이다. 메타데이터 관련 국제표준으로는 현재 GILS, CSDGM, DC 등이 있으나, 이것을 하나로 통합하는 것이 ISO/IEC 11179 : MDR(Meta Data Registry)이다.

본 연구에서는 위의 문제들을 해결하기 위해, 데이터준비를 위한 XML 기반 메타데이터 레지스트리 분산 검색 시스템을 설계한다. 설계의 주요 목표는 데이터준비단계에서 관계형 데이터에 대한 유용한 메타 정보를 사용자에게 제공함으로써, 데이터 분석 및 수집 시간을 단축하고 작업의 용이성을 제공하는 것이다. 또한, 메타데이터를 국제 표준인 MDR에 호환하도록 설계하여 조직 내외적으로 자유로운 데이터 교환, 확장 및 활용이 가능하도록 한다. 그리고 데이터 소스간의 이질성, 이기종 환경 및 이종 DBMS의 상이점을 극복하기 위해 XML 기반 분산 검색 시스템을 설계한다.

본 논문의 구성은 2장에서 GILS, CSDGM, DC 이를 통합하는 MDR에 대한 메타데이터 표준들에 관하여 기술한다. 그리고 3장에서는 데이터준비단계에서 특화되고 유용한 속성정보를 정의하여 XML 기반 MDR을 정의한다. 4장에서는 정의된 MDR을 XML 기반으로 분산 운용 및 검색하기 위한 시스템 구조도를 제안하고, 5장에서는 적용 예제를 기술한다.

2. 메타데이터 표준

메타데이터는 데이터의 구조 및 특성을 설명하는 데이터이다. 메타데이터를 참조함으로써 데이터를 분석하는 시간을 단축시킬 수 있을 뿐만 아니라, 데이터베이스에 관한 비전문가도 용이하게 데이터의 구조 및 특성을 파악 할 수 있다. 현재 메타데이터의 표준으로는 크게 GILS, FGDC, DC가 있으며 이를 통합하는 국제표준이 ISO/IEC11179 : MDR이다. 이들 표준의 용도와 구성에 관하여 간략하게 기술한다.

2.1 GILS(Government Information Locator Service)

GILS는 1994년 미국 연방정부가 정부의 각 기관들이 보유하고 있는 정보를 일정 표준에 따라 공공에게 공개하고 이를 접근할 수 있는 체계를 만들기 위한 프로젝트에서 시작되었다. GILS는 분산 네트워크(Distributed Networking) 환경이 지원되며, 프로파일(Profile)에는 최종 정보에 대한 소재정보 메타데이터인 소재 안내 레코드(Locator Record)의 핵심 요소(Core Element)를 규정하고 전반적인 명세를 제공한다[6,17].

2.2 CSDGM(Content Standard for Digital Geospatial Metadata)

1994년부터 FGDC(Federal Geographic Data Committee)에서 제정한 지리공간 메타데이터의 표준이다. CSDGM은 디지털 지리공간데이터의 용어와 정의를 명시하고, 데이터요소(Data Element)와 복합요소(Compound Element)와 가능한 값, 데이터 생성 규칙, 사용방법 등을 제시하였다. 데이터요소는 데이

터 값의 유형과 데이터 내용에 대한 정보를 제시하고, 복합요소는 데이터 요소의 집합 또는 다른 복합요소들로 구성된다[7].

2.3 DC(Dublin Core)

1995년부터 워크숍(Workshop)을 통해 자원의 신속한 검색 및 데이터의 호환성을 목표로 단순한 구조의 메타데이터 형식을 개발하였다. DC는 도서정보 및 e-book에 관한 표준 메타데이터로 활용되고 있으며, 15개의 기본 데이터요소로 구성된다. 또한 15개 요소가 포괄적이므로 요소에 특정한 한정어(Qualifier)와 인코딩 스킴(Encoding Scheme)을 2000년 7월 DC 이용 위원회에 의해 발표되었다[8].

2.4 ISO/IEC: 11179 Meta data Registry(MDR)

MDR은 최근에 메타데이터를 통합하는 국제 표준으로서 미국 환경청의 EDR(Environmental Data Registry), 미국 교통부의 ITS(Intelligent Transportation System) 등 정부기관을 중심으로 메타데이터 시스템 구축의 핵심으로 응용되고 있다.

MDR은 데이터에 대한 정확한 의미의 정의와 그 의미에 대한 손쉬운 접근을 목표로 하는 통합 저장소이다[10-14]. MDR은 데이터 요소(Data element)와 값 도메인(Value domain)으로 구성되며 이 두 요소와 관련된 여러 속성들이 정의되어 있다. 이들 속성은 필수속성(Main Property)과 조건부속성(Conditional Property) 그리고 선택속성(Optional Property)으로 나뉘어 진다[13]. 필수속성은 MDR에 반드시 포함되어야 하는 것이고, 조건부속성은 다른 속성의 정의 여부에 따라 종속적으로 포함여부가 결정되는 속성이며, 선택 속성은 독립적으로 선택 가능한 속성이다. MDR에서 정의한 속성은 총 63개이며 메타데이터 기술을 위한 속성 44개와 메타데이터 관리를 위한 속성 19개로 이루어져 있다. 모든 속성에 관한 자세한 사항은 MDR 표준 스펙에 명시되어 있다[10,11].

3. 데이터준비를 위한 XML 기반 MDR (DP-XMDR)정의

본 장에서는 데이터준비단계에서 필요한 요소들을 제안하고, 표준적인 참조가 가능하도록 MDR 스펙에 따라 표준 메타데이터를 정의한다. 또한, 이기종 시스템(Heterogeneous System) 및 이질 DBMS (Heterogeneous DBMS) 환경에서 상호운용(Interoperability)이 가능하도록 표준 메타데이터를 XML 스키마로 정의한다. 본 연구에서 메타데이터 표준들 중에 MDR을 적용하는 이유는, MDR이 메타데이터의 표준을 통합하는 포괄적인 메타데이터의 명세이기 때문이다.

3.1 속성 정보의 정의

성공적인 데이터마이닝 결과의 도출과 분석시간의 단축을 위해, 데이터준비단계에서 컬럼에 관한 유용한 속성 정보들을 데이터분석가에게 제공해야 한다. 데이터준비단계와 관련된 속성 정보는 관계형 데이터베이스(Relational Database)의 컬럼(Column)이 가지는 특성으로서 다음과 같이 두 종류로 정의한다.

- **정의 1 : 정적 속성 정보(Static Property Information)**

데이터베이스 설계단계에서 결정되는 속성의 집합으로서, 인스턴스(Instance)에 추가, 변경, 삭제에 독립적이며 변하지 않는 속성 정보

정의 1에 대한 정적 속성 정보의 예로서는 컬럼에 허용되는 최대값 및 최소값, 허용되는 값 종류(Distinct Value), 내포하는 단위(Measurement) 등이 있다.

- **정의 2 : 동적 갱신 속성 정보 (Dynamic Updated Property Information)**

인스턴스의 추가, 변경, 삭제에 따라 변동하는 속성의 집합으로서, 인스턴스에 종속적인 정보

정의 2에 대한 동적 갱신 속성 정보의 예로서는 컬럼에 대한 인스턴스의 최대값 및 최소값, 인스턴스의 오류데이터 개수, 표준편차(Standard Deviation) 등이 있다.

3.2 정적속성 정보기반 MDR의 정의

본 논문에서 제안하는 시스템에서는 정적 속성 정보로서 메타데이터를 구성하며 표 1에 제안하였다. 데이터마이너(Data Miner)가 데이터준비단계에서는 느끼는 가장 큰 고충은 방대한 데이터의 정보를

분석해야 하는 부담이다. 그러므로, 컬럼에 관한 속성정보를 제공함으로써 지역 사이트의 관계형 스키마와 레코드를 일일이 분석하지 않고도 컬럼의 용도와 특성을 쉽게 파악 할 수 있어야 한다. 본 논문에서는 지속적인 갱신(Update)과 시스템 구조의 복잡도(Complexity)가 가중되는 동적 갱신 속성 정보는 배제하고, 정적 속성 정보로 접근이 용이하고 단순한 구조를 가지는 메타데이터를 표 1과 같이 정의하였다.

표 1. 데이터준비를 위한 정적속성 정보와 MDR 속성

정적속성 정보	MDR속성(카테고리.속성명)
컬럼명	*Identifying.name*
출처	Identifying.context name
허용최대값	Data element.maximum size
허용최소값	Data element.minimum size
심볼릭	Value meaning.value meaning description
단위	Value domain.unit of measure name
자료형	*Value domain.data type name*
데이터부류	Relational.classification scheme name
주석	Administrative.comments
표현형식	Data element.layout of representation
허용 값 형태	Permissible value.value

표 1에서 MDR 속성 항목은 정적속성 정보에 대해 ISO/IEC : 11179에서 기술한 MDR 속성에 해당하는 항목을 추출한 것이다. 기술 형태는 해당하는 속성의 카테고리(Category)명과 속성명을 '카테고리.속성명' 형태로 기술하였다. 그 외에, ISO 11179에서 기술한 속성집합에서 필수 정의 속성에 대해서는 참고문헌[11]을 참조하고, 본 논문에서는 생략하였다. 표 1에 기술한 정적 속성 정보들에 관한 설명과 데이터준비단계에서 참조되는 정보의 예는 다음과 같다.

- **컬럼명(Column Name)**
  관계형 데이터베이스의 스키마에 정의된 컬럼명
  예) name : 사원의 이름을 나타내는 컬럼은 'name'
- **출처(Source)**
  해당 컬럼이 어느 데이터베이스에 포함되어 있는지에 관한 정보를 나타낸다.
  예) employee : 사원의 이름을 나타내는 'name' 컬럼은 employee 데이터베이스에 속한다.
- **허용최대값(Upper Bound)**
  수치형 컬럼에 대해 인스턴스에 허용되는 최대값
  예) 100 : 사원의 저축률을 나타내는 'savingRate'의 허용최대값은 100이다.
- **허용최소값(Lower Bound)**
  수치형 컬럼에 대해 인스턴스에 허용되는 최소값
  예) 0 : 사원의 저축률을 나타내는 'savingRate'의 허용최소값은 0이다.
- **심볼릭(Symbolic)**
  각 심볼릭에 대응되는 내용에 대한 기술
  예) (m : 남자), (f : 여자) : 성별을 나타내는 컬럼 'gender'에서 'm'은 남자, 'f'는 여자, 그 외의 값은 오류를 의미한다.
- **단위(Measurement)**
  수치형 컬럼이 내포하는 단위
  예) 길이(m) : 신장을 나타내는 컬럼 'height'의 단위는 미터이다.
- **자료형(Data type)**
  스키마 설계시 컬럼에 대한 자료형으로서 SQL 99의 정의를 따른다.
  예) INTEGER : 컬럼 'height'의 자료형은 integer이다.
- **데이터부류(Data Classification)**
  컬럼에 기입되는 인스턴스 값들의 특성에 관한 정보로서 샘플링계획 및 데이터마이닝 단계에서 모형설계에 대한 중요한 정보가 된다. 데이터부류의 내용은 연속형(Continuous Type), 이산형(Discrete Type), 증가형(Incremental Type), 순서형(Ordinal Type), 계층형(Classical Type), 구간형(Interval type), 명칭형(Nominal Type), 비율형(Ratio Type)이 있다[3].
  예) 이산형 : 성별을 나타내는 컬럼 'gender'는 이산형이므로 로지스틱 회귀모형으로 예측한다.
- **표현형식(Presentation Form)**
  컬럼에 기입되는 인스턴스의 표현 형태를 나타낸다.
  예) n(3)-n(3) : 우편번호를 나타내는 'zipcode' 컬럼은 3자리 숫자, 구분 문자 '-', 3자리 숫자로 표기된다.
- **허용값 형태(Allowance Range)**
  인스턴스로 가능한 허용값의 범위 및 형태를 기입

한다.

예) 000-000~999-999 : 우편번호를 나타내는 'zip-code' 컬럼의 인스턴스의 각 자리수는 0에서 9이다.

3.3 XML 스키마 변환(XML Schema Transformation)

현재 구축되어 있는 대부분의 데이터는 이기종 환경 및 이질 DBMS에 저장 되어있을 뿐만 아니라, 메타데이터에 접근하는 클라이언트의 환경도 다양하다. 따라서 어느 환경에서나 상호운용 가능한 XML 표현 기법이 유용하다. 그림 1은 설계된 MDR을 XML 스키마로 변환한 것이다.

```
<?xml version="1.0" encoding="euc-kr"?>
<xsd:schema
  xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<xsd:complexType name="DataElement">
<xsd:sequence>
  <xsd:element name="Name" type="xsd:string"/>
  <xsd:element name="MaximumSize"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="MinimumSize"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="ContextName"
  type="xsd:string"/>
  <xsd:element name="ValueMeaningDescription"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="UnitOfMeasureName"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="DataTypeName"
  type="xsd:string"/>
  <xsd:element name="ClassificationSchemeName"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="LayoutOfRepresentation"
  type="xsd:string"/>
  <xsd:element name="Value" type="xsd:string"/>
  <xsd:element name="Comments" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>
</xsd:schema>
```

그림 1. 데이터준비를 위한 XML 기반 MDR(DP-XMDR)

표 1의 정적속성 정보는 대응되는 MDR에서 명시한 필수속성과 선택속성이 있다. XML의 occurrence는 기본값이 minOccurs="1", maxOccures="1"이므로 인스턴스를 반드시 한번 기술해야 하므로, MDR에서는 필수속성이 된다. 따라서 선택속성을 나타내기 위해서는 그림 1과 같이 선택속성에 대해서는 minOccurs="0"을 추가해야 한다. 그리고 DP-XMDR과 다른 목적의 MDR과의 스키마 통합이 용이하도록, 요소(Element)의 명칭을 MDR에서 정의한 명칭과 동일하도록 정의함으로써, DP-XMDR이 확장성을 가지도록 정의하였다.

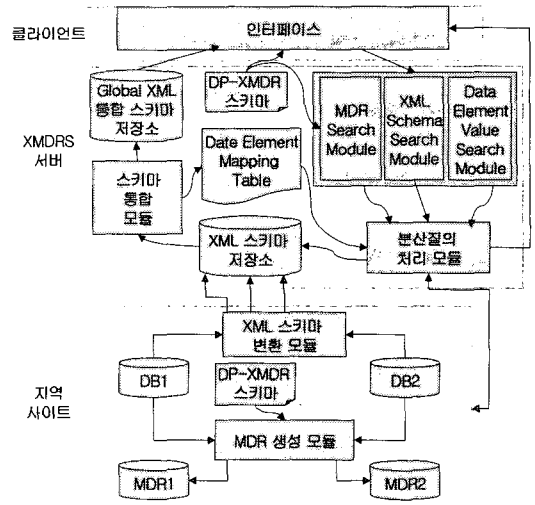

그림 2. DP-XMDRS 전체 구조도

4. XML MDR Distributed Retrieval System For Data Preparation (DP-XMDRS) 설계

본 장에서는 기존의 RDB를 XML 기반으로 분산검색하고, 3장에서 제시한 DP-XMDR의 정적속성 정보를 관리 및 검색 할 수 있는 DP-XMDRS를 설계한다. DP-XMDRS의 전체 구조도는 그림 2와 같으며, 지역 사이트, DP-XMDRS서버, 클라이언트로 구성된다.

그림 2와 같이 클라이언트에서는 인터페이스를 통해 MDR 및 관계형 데이터베이스의 스키마 및 인스턴스를 GUI 환경으로 검색 할 수 있다. 각 지역 사이트에는 관계형 DB스키마와 메타 정보를 포함하는 MDR이 존재한다. DP-XMDRS 서버에는 크게 스키마 저장 모듈, 스키마통합 모듈, 검색 모듈로 구성되어 있다. 클라이언트, 지역 사이트, DP-XMDRS 서버에서의 각 저장소 및 모듈에 관한 자세한 설명은 다음과 같다.

· 클라이언트

인터페이스

Global XML 통합 스키마를 사용자에게 제공하여,

web-based 및 GUI로 분산질의 환경을 제공한다. 제안된 시스템에서 제공되는 검색범위는 일반적인 분산 데이터 검색과 MDR 속성의 인스턴스의 검색을 포함한다.

· 지역사이트

① XML 스키마 변환 모듈(XML Schema Transfer Module)
Duckett[16]의 제안한 절차로 관계형 DB의 SQL 스키마를 XML 스키마로 변환하는 모듈이다.

② MDR 생성 모듈(MDR Creation Module)
DP-XMDR을 참조하여 관계형 DB의 컬럼에 대해 XML 문서로 MDR을 생성하는 모듈이다. MDR 생성 시기는 두 경우로 나눌 수 있다. 첫째는 관계형 스키마 생성시기에 스키마 설계자에게 DP-XMDR의 속성을 질의하여 생성한다. 둘째는 MDR에 생성되지 않는 스키마들을 데이터마이너가 분석할 때 DP-XMDR의 속성을 질의하여 생성한다.

· XMDRS 서버

① DP-XMDR 스키마(DP-XMDR Schema)
데이터준비단계에서 컬럼에 대한 유용한 정적속성 정보들이 MDR의 표준 스펙에 맞게 XML 스키마 형태로 정의되어 있다.

② XML 스키마 저장소(XML Schema Storage)
각 지역사이트로부터 전달받은 XML 스키마를 저장하는 곳으로, 각 DB에 저장된 관계형 스키마를 XML질의 및 인터페이스로 분산질의하기 위해 참조된다. 데이터마이너들은 제한된 통합스키마(Integrated Schema)가 반영된 인터페이스를 통하지 않고 XML 스키마 저장소를 액세스하여 관계형 데이터베이스의 스키마 검색이 가능하다.

③ 스키마 통합 모듈(Schema Integrating Module)
XML 스키마 저장소의 스키마는 Name conflict, Data Type conflict, Structural conflict, Constraints conflict가 발생한다. 이러한 충돌들을 해결하기 위해 스키마 통합 모듈에서는 전문가의 지식을 반영하여 스키마를 통합하는 방법을 적용한다[20].

④ Data Element Mapping Table
Global XML 통합 스키마의 각 element가 어느 MDR에 속하는 것인 지의 정보를 저장하는 테이블이다. 즉, 스키마 통합과정에서 변경되는 element의 명칭, 데이터 타입, occurrence constraints를 기록하는 테이블이다.

⑤ Global XML 통합 스키마 저장소
스키마 통합 모듈을 통해 XML 스키마 저장소의 XML 스키마 간에 발생하는 충돌문제를 해결하여 생성된 통합 스키마를 저장하는 곳이다.

⑥ MDR search module
인터페이스를 통해 탐색하고자 하는 컬럼에 대한 MDR 속성 값을 검색하는 모듈이다.

⑦ XML Schema Search Module
XML 스키마 저장소의 원본 XML 스키마들을 검색하기 위한 모듈로서, 지역 사이트의 관계형 스키마를 분석하거나 MDR의 정의 여부를 검색하기 위해 사용된다.

⑧ Data Element Value Search Module
분산질의 처리 과정을 통해, 관계형 DB에 저장되어 있는 컬럼의 값을 검색하는 모듈이다. 데이터마이너는 MDR search module을 통해 컬럼의 정적속성정보를 참조하여 데이터준비에 필요한 요소임을 결정하여 이 모듈을 통해 데이터를 수집한다.

표 2. employee 테이블

employee		
pk	empID	UNIQUE
	name	CHAR
	idNum	CHAR
	address	VARCHAR
	telNum	CHAR
	gender	INTEGER
	height	INTEGER
	weight	INTEGER

5. DP-XMDRS 저장 모듈별 예제

본 장에서는 그림 2에서 제안한 DP-XMDRS 구조도에서 각 저장 모듈에 포함되어 있는 XML 문서의 예를 나타낸다. 예제에 사용된 관계형 스키마는 표 2와 같이 사원의 인적사항을 나타내는 employee 테이블이다.

5.1 XML 스키마 저장소의 원본 XML 스키마

XML 스키마 저장소에는 각 지역 사이트에서 관계형 스키마를 XML 스키마 변환 모듈을 통해 생성

된 원본 XML 스키마가 그림 3과 같이 저장되어 있다.

```
xsd1_employee_inform.xsd
<?xml version="1.0" encoding="euc-kr"?>
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema">
<xsd:complexType name="employee">
<xsd:sequence>
  <xsd:element name="empId" type="xsd:string"/>
  <xsd:element name="name" type="xsd:string"/>
  <xsd:element name="idNum" type="xsd:string"/>
  <xsd:element name="address" type="xsd:string"/>
  <xsd:element name="telNum" type="xsd:string"/>
  <xsd:element name="gender" type="xsd:DataElement"/>
  <xsd:element name="height" type="xsd:string"/>
  <xsd:element name="weight" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>
<xsd:complexType name="DataElement">
<xsd:sequence>
  <xsd:element name="Name" type="xsd:string"/>
  <xsd:element name="MaximumSize"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="MinimumSize"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="ContextName"
  type="xsd:string"/>
  <xsd:element name="ValueMeaningDescription"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="UnitOfMeasureName"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="DataTypeName"
  type="xsd:string"/>
  <xsd:element name="ClassificationSchemeName"
  type="xsd:string" minOccurs="0"/>
  <xsd:element name="LayoutOfRepresentation"
  type="xsd:string"/>
  <xsd:element name="Value" type="xsd:string"/>
  <xsd:element name="Comments" type="xsd:string"/>
</xsd:sequence>
</xsd:complexType>
</xsd:schema>
```

그림 3. employee 테이블에 대한 원본 XML 스키마

원본 XML 스키마는 관계형 스키마의 구조를 설명하고 각 컬럼에 대해 MDR이 정의되어 있는지의 정보를 포함한다. 그림 3과 같이 루트 엘리먼트(Root Element)는 테이블명이고, 종속되는 엘리먼트는 해당하는 컬럼명들로 구성된다. 여기에서 컬럼에 대한 엘리먼트의 type이 DataElement인 것은 지역사이트에서 MDR이 정의되어 있음을 나타낸다. 즉, gender 컬럼을 제외한 모든 컬럼들은 MDR이 정의 되어 있지 않음을 나타낸다.

5.2 MDR 인스턴스 예

지역 사이트의 MDR에는 DP-XMDR에서 정의되어 있는 정적 속성 정보들이 그림 4와 같이 컬럼에 대해 저장되어 있다. 그림 4는 성별을 나타내는 employee 테이블의 gender 컬럼에 대한 MDR의 내용으로서 데이터준비단계에 필요한 유용한 정보를 제공한다.

```
xsi_employee_inform_table.xml
<?xml version="1.0" encoding="euc-kr"?>
<employee xmlns:xsi =
  "http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation=
  "xsd1_employee_inform.xsd">
<empId> not define </empId>
<name> not define </name>
<idNum> not define </idNum>
<address> not define </address>
<telNum> not define </telNum>
<gender>
  <name> gender </name>
  <ContextName> employee_inform </ContextName>
  <ValueMeaningDescription> 0:남자, 1:여자
    </ValueMeaningDescription>
  <DataTypeName> INTEGER </DataTypeName>
  <ClassificationSchemeName> 이산형
    </ClassificationSchemeName>
  <LayoutOfRepresentation> 0 - 1
    </LayoutOfRepresentation>
  <Comments> 성별을 나타냄 </Comments>
</gender>
<height> not define </height>
<weight> not define</weight>
</employee>
```

그림 4. employee 테이블의 MDR 인스턴스

그림 4와 같이 DP-XMDR에서 정의된 속성 중에 선택 가능한 속성에 대해서는 생략이 가능하다. 데이터 마이너는 MDR 인스턴스를 통해 지역 사이트의 RDB 레코드를 분석하지 않아도, gender 컬럼은 이산형으로서 '0'은 남자, '1'은 여자를 나타내는 컬럼임을 알 수 있다.

6. 데이터베이스 아키텍처별 비교분석

본 장에서는 대표적인 분산데이터베이스 아키텍처인 연방 DBMS(federal DBMS), 데이터웨어하우스 시스템과 본 논문에서 제안한 DP-XMDRS를 9개의 항목을 선정하여 비교 분석한다.

표 3은 연방 DBMS, 데이터웨어하우스 시스템, DP-XMDRS간에 각 항목별로 비교평가 한 것이다. 표 3과 같이, 데이터웨어하우스 시스템에는 일반적이고 방대한 메타데이터를 보유하는 반면 DP-XMDRS는 데이터 준비를 위해 전문화 된 메타데이터만을 보유하고 있다. 그리고, 다른 DBMS간의 통합 및 운용의 용이성을 의미하는 상호운용성에서, 연방 DBMS의 경우에는 이질적인 요소가 많아 가장 난해한 반면 DP-XMDRS는 MDR 표준을 준수하고 XML 기반으로 설계되어 다른 시스템과의 상호운용이 용이하다. 데이터웨어하우스 시스템의 메타데이터는 DP-XMDRS에 비해 방대하고 포괄적이므로, 데이터마이너에게 데이터 준비를 위한 전문적인 메타정보를 제공하지 못한다. 로컬 DB 접근 빈도면에서는 메타데이터를 보유하지 않은 연방 DBMS에서 가장 높고, DP-XMDRS의 경우에는 스키마 분석, MDR 정의여부를 참조하기위해 로컬 DB를 접근하지 않아도 되므로, 데이터웨어하우스 시스템 보다는 로컬 DB 접근 빈도는 높지만, 연방 DBMS 비해 낮은 빈도를 가지는 특성이 있다. 데이터 준비를 위한 사용자 요구수준의 경우에, 메타정보를 제공하지 않는 연방 DBMS에서는 직접 각 로컬 DB에 접근하여 처리해야 하므로 DB 및 데이터마이닝에 대한 전문 지식이 요구되고, 데이터웨어하우스 시스템에서는 로컬 DB에 접근할 필요가 없지만, 방대한 메타정보에서 데이터마이닝과 관련된 요소들을 추출하여 분석해야 하므로 데이터마이닝에 대한 전문적인 지식이 필요하다. 반면에 DP-XMDRS를 이용하는 경우에는 데이터 준비단계에 필요한 요소들로 메타데이터를 구성하고 있으므로 비전문가도 전문화된 메타데이터를 참조하여 쉽게 데이터 준비단계를 수행 할 수 있다. 그러나, 연방 DBMS와는 달리 데이터웨어하우스 시스템과 DP-XMDRS에서는 메타정보를 운용하기 위한 부가적인 데이터베이스가 필요하고, 스키마의 변화에도 복잡하고 민감한 특성이 있다. 로컬 DBMS의 인스턴스의 변경에 따른 메타데이터 일치성을 만족시키기 위한 부담 측면에서, DP-XMDRS는 각 지역 DB의 인스턴스의 삽입, 변경, 추가에 대해 영향을 받지 않는 정적속성정보만으로 설계하여, 매번 메타정보를 변경해야하는 데이터웨어하우스 시스템에 비해 우수한 장점이 있다.

7. 결 론

본 논문에서는 데이터마이닝을 위해 관계형 데이터베이스로부터 데이터준비단계에서 데이터마이너에게 컬럼 및 DB 스키마에 대한 유용한 정보를 제공하여 분석 및 수집시간을 단축하기 위한 방안을 제시하였다. 제안한 DP-XMDRS를 통해 비전문가도 지역 사이트의 관계형 데이터베이스를 번거롭게 분석하지 않아도, 컬럼에 대한 정적 속성정보를 참조하여 데이터마이닝을 위해 필요한 데이터를 쉽게 수집 및 가공 할 수 있다. 그리고 DP-XMDR은 ISO/IEC 11179 표준을 준수하므로, 향후 표준 속성의 추가에 의한 확장이 가능하고 다른 목적의 MDR과도 상호통합이 용이하도록 설계하였다.

DP-XMDRS는 XML 기반으로 설계되어, 지역사이트의 데이터베이스와 MDR의 스키마 및 인스턴스를 인터페이스를 통해 검색 할 수 있고, 이질 및 이기종 DBMS 환경에서도 상호 호환 가능하도록 설계하

표 3. 데이터베이스 아키텍처별 비교분석

비교항목	연방 DBMS	데이터웨어 하우스 시스템	DP-XMDRS
메타데이터 보유	no	yes	yes
MDR 준수성	no	no	yes
상호운용성	difficult	medium	easy
데이터마이닝 정보 전문성	no	general	special
로컬 DB 접근빈도	high	low	medium
사용자 요구 수준	DB and data mining expert	data mining expert	non-expert
부가 DB 규모	none	very large	small
스키마 변경 부담	little	much	medium
인스턴스 변경 부담	little	much	little

였다. 또한 XSLT(XSL Transformation)와 같은 XML 관련 기술을 접목하여 다양한 플랫폼(Platform)으로의 출력과, XML질의어를 통해 인터페이스에서 제공하지 않는 다양한 검색이 가능하다. 또한, 본 논문에서 제안한 DP-XMDR을 탑재한 DP-XMDRS는 분산 데이터베이스 환경에서 데이터마이닝을 하기위해 데이터준비단계에서 메타정보를 검색하는 시스템 구축의 기반으로 응용 할 수 있다.

향후 과제로는 첫째, DP-XMDR의 속성값에 대해 CSDGM과 같이 데이터마이닝과 관련하여 체계적으로 사용 가능한 속성값의 유형과 내용에 대한 정의가 필수적이다. 예를 들어 데이터부류와 표현형식 속성정보에 대해 표준적인 속성값의 종류와 표기형식이 필요하다. 둘째, 3장에서 제시한 동적 속성정보를 MDR에 저장할 때 데이터마이닝 특성에 맞는 지역 데이터베이스와의 갱신주기를 최소화 할 수 있는 알고리즘에 대한 연구가 필요하다.

참 고 문 헌

[1] W. Klosgen and J.M. Zytkow, Handbook of Data Mining and Knowledge Discovery, Oxford University Press, New York, 2002.

[2] O.P. Rud, Data Mining Cookbook - Modeling Data for Marketing, John Wiley & Sons Publishers, New York, 2000.

[3] D. Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, San Francisco, 1999.

[4] K. Munroe and Y. Papakonstinou, "BBQ: A Visual Interface for Browsing and Querying XML", Visual Database Systems, pp. 123-126, May 2000.

[5] C. Baru, A. Gupta, B. Ludaescher, R. Marciano, Y. Papakonstantinou, P. Velikhov, "XML-Based Information Mediation with MIX", Exhibitions Program of ACM SIGMOD, pp. 597-599, June 1999.

[6] Government Information Locator Service (GILS), "http://www.gils.net".

[7] Federal Geographic Data Committee(FGDC), "http://www.fgdc.gov".

[8] Dublin Core Metadata, "http://dublincore.org".

[9] ISO/IEC 11179-1 : Framework for The Generation and Standardization of Data Elements.

[10] ISO/IEC 11179-2 : Classification of Concepts for Identification of Domains.

[11] ISO/IEC 11179-3 : Basic Attributes of Data Elements.

[12] ISO/IEC 11179-4 : Rules and Guidelines for The Formulation of Data Definitions.

[13] ISO/IEC 11179-5 : Naming and Identification Principle for Data Elements.

[14] ISO/IEC 11179-6 : Registration of Data Elements.

[15] A. Jain and D. Zongker, "Feature Selection : Evaluation, Application, and Small Sample Performance", IEEE Trans. Pattern Anal. Machine Intell., vol.2, pp.153-158, 1997.

[16] J. Duckett at. al, Professional XML Schemas, Wrox press, New York, 2001.

[17] 표준화밸리, "http://isv.kisti.re.kr".

[18] 최국렬외, 데이터마이닝 이론과 실습, 청구문화사, 서울, 2001.

[19] 이호경 외, Openning XML, 구민사, 서울, 2002.

[20] 나민영, 유진철, 김종화, 김재범, 고석범, "메타데이터 레지스트리와 XML 기술의 연계", 한국데이터베이스 진흥센터, 서울, 2003.

고 석 범

1996년 동서대 컴퓨터공학 학사
1998년 아시카가공대 정보시스템공학 석사
2001년 부경대 전자계산학 박사 수료
2001년~2003년 육군사관학교 전자계산학과 전임강사

관심분야 : 분산데이터베이스, 데이터마이닝, 통합 데이터베이스, XML

윤 성 대

1980년 경북대 컴퓨터공학 학사
1984년 영남대 전자계산학 석사
1997년 부산대 전자계산학 박사
1986년~현재 부경대 전자계산학과 교수

관심분야 : 병렬운영체제, 데이터마이닝, Multi-threaded architecture