韓國 컴퓨터情報學會 論文誌
第9卷 第3號, 2004. 9.

2004-9-3-1-14

이기종 분산환경에서 데이터마이닝을 위한 데이터준비 시스템 구현

이 상 희*, 이 원 섭**

Implementation of Data Preparation System for Data Mining on Heterogenious Distributed Environment

Lee sang hee*, Lee won sup**

요 약

본 논문에서는 데이터 마이닝을 위한 데이터 준비 과정에 대하여 기존의 데이터 마이닝 도구들의 효율성을 비교하고, 새로운 효율적인 데이터 준비 시스템 설계 기준을 제안하고자 한다. 지역 및 원격 데이터베이스 접근방법, 이기종 컴퓨터간의 정보 교환을 기준으로 기존의 데이터마이닝 도구들의 기능을 비교하였다. 본 논문에서는 앤서트리, 클레멘타인, 엔터프라이즈 마이너, 웨카를 비교하였다. 또한, 본 논문에서는 분산 네트워크 상에서 데이터 마이닝을 위한 효율적인 데이터 준비 시스템을 위한 설계기준을 제안한다.

Abstract

This paper is to investigate the efficiency of the process of data preparation for existing data mining tools, and present a design principle for a new efficient data preparation system . We compare the often used data mining tools based on the access method to local and remote databases, and on the exchange of information resources between different computers. The compared data mining tools are Answer Tree, Clementine, Enterprise Miner, and Weka. We propose a design principle for an efficient system for data preparation for data mining on the distributed networks.

▸ Keyword : 데이터마이닝, 이기종 분산 환경

• 제1저자 : 이상희
• 접수일 : 2004.08.28, 심사완료일 : 2004.09.14
* 청강문화산업대학 컴퓨터소프트웨어과 조교수
** 인덕대학 컴퓨터전자전공 조교수

Ⅰ. 서 론

데이터 마이닝(data mining)은 데이터준비(data preparation), 데이터 정제(data cleansing), 데이터 마이닝, 해석(interpretation) 및 평가(evaluation), 실행(action)의 5가지 단계로 이루어진다[1]. 데이터 준비 단계가 데이터 마이닝에 있어서 가장 먼저 진행되고, 가장 시간과 노력이 많이 투입되는 단계로 알려져 있다. 마이닝을 위한 기초 데이터는 대부분 다양한 이기종 컴퓨터에 축적 운영되며 지역적으로 분산되어 있는 것이 일반적이다. 또한, 마이닝 도구는 데이터베이스와는 기종이나 운영체제가 다른 별도의 컴퓨터에 저장되며, 마이닝을 직접 수행하는 마이너는 대부분 데이터가 저장되어 있는 컴퓨터나 마이닝 도구가 설치되어 있는 컴퓨터와는 별도의 다른 컴퓨터를 활용하는 경우가 많다. 따라서 데이터마이닝을 위한 데이터 준비과정은 단순한 데이터 접근하여 데이터를 획득하는 것뿐만 아니라 지역적으로 분산되어 있는 여러 시스템에 산재된 데이터에 접근하고, 획득하고, 통합, 관리하는 과정이다.

또한, 데이터 준비 시스템에는 다양한 형식을 표준화된 형식으로 변환하는 기능과 마이닝 도구가 저장된 컴퓨터와 데이터를 교환하는 기능, 다양한 이기종 환경에서 접근하는 여러 사용자를 지원하는 기능이 있어야 한다.

본 논문에서는 2장에서 기존 데이터마이닝 도구들이 데이터 접근을 살펴보고, 지역 및 원격 데이터베이스 접근, 그리고 이기종 컴퓨터에 존재하는 정보자원간의 데이터 교환을 지원하는 기능에 따라 이들 도구들을 분석, 비교하고 평가한다. 3장에서는 효율적인 데이터준비과정을 제시하고, 이를 기반으로 분산 환경 하에서 데이터마이닝을 위해 구현된 새로운 데이터준비시스템을 제안한다.

Ⅱ. 기존 데이터마이닝 도구들의 데이터 접근 방법 비교

본 논문에서는 일반적으로 많이 사용하는 3개의 데이터 마이닝 도구와 자바기술을 활용하여 개발한 한 개의 도구를 비교하고자 한다. 3개의 기존도구들은 SPSS에서 개발한 의사결정나무 알고리즘[2]인 앤서트리(Answer Tree), 역시 SPSS에서 개발한 클레멘타인(Clementine), SAS의 엔터프라이즈마이너(Enterprise Miner)[3]이다. 자바로 작성된 도구는 Waikato 대학교에서 개발한 데이터마이닝 도구로 공개소프트웨어인 웨카(Weka)[4]이다.

본 논문에서 4가지 데이터마이닝 도구들에 대해 지역 및 원격 데이터베이스 접근 기능, 이기종 컴퓨터에 위치한 정보자원간의 데이터 교환지원 기능의 관점에서 간단히 알아보겠다.

앤서트리, 클레멘타인, 엔터프라이즈 마이너는 ODBC(Open Database Connectivity)를 기반으로 데이터베이스에 접근한다. 따라서 지역 데이터베이스에의 접근에 있어서는 ODBC 드라이버가 설치되어 있는 어떤 데이터베이스와도 데이터 접근이 가능하다. 하지만 원격 데이터베이스에 접근하는데 있어서는 각기 다른 제약이 있다. 먼저 앤서 트리는 원격 데이터베이스에 연결하는 기능을 제공하지 않는다. 클레멘타인과 엔터프라이즈 마이너는 별도의 소프트웨어를 이용하여 원격지 컴퓨터에 접속한다. 클레멘타인과 엔터프라이즈 마이너는 별도의 소프트웨어를 이용하여 원격 컴퓨터에 접속한 후에야 데이터 접근을 위해 ODBC로 원격 데이터베이스와 데이터 준비 시스템을 연결할 수 있다. 클레멘타인은 원격 컴퓨터와 접속을 위해 이기종 컴퓨터에 접속하여 통합사용을 지원하는 Exceed를 사용하며, 엔터프라이즈 마이너는 SAS에서 제공하는 SAS Connect 기능을 이용하여 원격지 컴퓨터에 접속한다.

자바로 구현된 웨카는 의사결정나무기법, 선형회귀, 모형트리생성 등의 전통적인 통계 분석방법뿐만 아니라 다양한 마이닝 기법을 가지고 있으며, JDBC(Java Database Connectivity)를 통해 지역 데이터베이스와 연결이 가능하다. 원격 데이터베이스와의 연결을 위하여 자바 환경에서 컴퓨터나 프로그램, 객체 사이에 통신할 수 있는 기능을 제공하는 RMI(Remote Method Invocation)기법을 사용한다. 따라서 웨카를 사용하면 플랫폼에 독립적으로 수행되는 자바의 특성을 기반으로 하여 이기종 정보자원간의 데이터의 교환이 가능하다. 하지만 입력파일이 ARFF형식으로 지정되어 있어 데이터베이스에서 가져온 입력데이터로 마이닝 과정을 실행하기 위해서는 ARFF형식으로 변환해야하는 단점이 있다.

III. 효율적인 데이터 준비 시스템의 제안 및 구현

앞에서 언급한 기존의 데이터 마이닝 도구들은 데이터 접근, 데이터 교환 등 데이터 준비 과정에서 많은 제약이 있다. 최신 소프트웨어 기술과 네트워크 기술을 활용하여 기존 데이터 마이닝 도구들이 가지고 있는 제약점들을 보완할 수 있다. 제약점을 보완할 수 있는 기술로는 다음과 같은 것들이 있다.

첫째, 3-티어(tier) 구조를 데이터 준비 시스템에 적용이다. 대부분의 데이터 준비 시스템이 클라이언트/서버의 2-티어 구조를 사용하고 있었으나 3-티어 구조는 2-티어 구조에 비해 클라이언트가 간단한 사용자 인터페이스 기능만을 수행하게 된다. 그러므로, 3-티어 구조는 2-티어 구조의 문제점인 클라이언트에서 비즈니스 로직 (business logic)의 수행으로 인한 시스템 성능 저하 문제와 클라이언트 소프트웨어 배포를 위한 비용 문제를 해결할 수 있다.

둘째, 데이터 준비 시스템은 클라이언트와 서버의 기종이나 운영체제가 상이할 경우, 다양한 데이터베이스로부터 데이터를 통합할 수 있어야 한다. 또한, 이기종 컴퓨터간의 데이터 교환을 지원하는 기능을 가져야 한다.

셋째, 데이터 준비 시스템은 분산 환경을 지원하는 언어로 작성되어야 한다. 분산환경에 적합한 개발 언어는 여러 종류가 있지만 이식성면에서 자바가 있다. 자바는 플랫폼에 독립적 수행되어 클라이언트/서버 환경에서 이기종 컴퓨터간의 사용할 수 있다. 또한, 자바는 이식성이 높으며, 높은 수행능력과 동적 환경을 제공한다. 그러므로, 원격 데이터베이스와 데이터 접근 작업을 수행하기 위해서 자바를 이용하는 것이 바람직하다. 또한, 3-티어구조에서 자바로 미들티어를 구연할 수 있으며, JDBC는 미들티어에서 데이터베이스에 접근하는 기능을 가지고 있다.

본 논문에서는 이러한 점을 고려하여 바람직한 데이터 준비 시스템을 구성하기 위해 다음과 같은 소프트웨어 기술을 설계기준으로 제시한다.

첫째, 분산 환경에 능동적으로 대처하기 위하여 어플리케이션 서버, 데이터베이스 서버와 클라이언트 서버를 3-티어 방식으로 구성하여 가용성과 확장성, 효율성을 높인다.

둘째, 3-티어의 구성요소 중 어플리케이션 서버는 EJB(Enterprise Java Beans)를 사용하여 구연한다. EJB는 자바언어로 작정된 컴포넌트로 분산 환경에 접합하고, 멀티 티어로 확장할 수 있고, 보안성이 높으며, 하드웨어 플랫폼이나 운영체제에 관계없이 이식 가능하므로 어플리케이션 서버 구현에 적합한 기술이다[5].

셋째, 클라이언트와 서버간의 데이터 전송은 XML을 이용한다. XML은 서로 다른 데이터베이스의 호환이 가능하고, 분산처리가 가능하며, 표현 양식을 선택할 수 있어 데이터 교환이 용이하다. 이진코드로 표현되는 데이터의 경우 각 시스템에서 데이터의 표준이 정해져 있지 않으면 교환을 할 수 없지만, XML을 이용하게 되면 데이터를 교환하는데 있어서 각 시스템에 맞춰서 보내는 것이 아니라 파싱한 XML문서를 각자의 시스템에 변환만 해주면 된다.

또한, XML은 다양한 데이터베이스를 검색하여 데이터를 쉽게 통합할 수 있다[6].

위와 같은 설계기준을 기반으로 하여 실제적인 데이터 준비 시스템을 구현하였다. 이 시스템은 EJB로 구성한 어플리케이션 서버, 어플리케이션 서버에 SQL(Structured Query Language)로 질의하고, 질의결과를 XML문서로 전달받는 클라이언트, 그리고 데이터베이스로 구성된다.

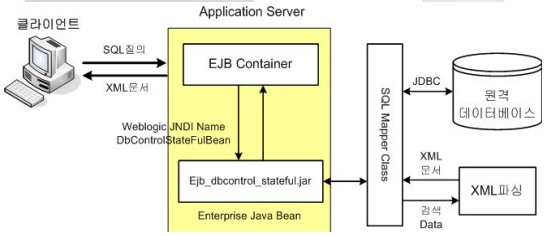

그림 3-1. 데이터 준비 시스템의 구성도

이 시스템에 의해 클라이언트가 필요한 원격 데이터베이스를 접근하는 과정은 다음과 같다.

① 클라이언트는 IP, 포트, SID, 아이디, 패스워드와 같은 해당 데이터베이스에 접속하는데 필요한 정보와 함께 EJB 컨테이너에 원격 데이터베이스와의 연결이 요청한다. 어플리케이션 서버에 설치되어 있는 EJB 컨테이너는 서버측을 담당하는 EJB를 호출하여 클라이언트의 요청을 수행한다.

② 어플리케이션 서버의 EJB에서 JDBC 응용프로그램 인터페이스 SQLMapper 클래스를 사용하여 원격 데이터베이스와 연결한다.
③ 원격데이터베이스와 연결된 어플리케이션 서버는 클라이언트가 SQL문으로 질의하면 SQLMapper 클래스를 통하여 데이터베이스에 질의를 전송한다.
④ SQLMapper 클래스는 데이터베이스에서 검색된 데이터를 XMLParseUtil 클래스를 이용하여, XML, XSL문서로 변환한다. 변환된 XML문서는 EJB 컨테이너에 의하여 클라이언트로 전송된다.
⑤ 클라이언트는 서버로부터 전송된 XSL 파일을 파싱하여 HTML 파일로 만든 후 화면에 표시한다.

우리가 개발한 데이터 준비 시스템은 자바로 작성되어, 기존에 자바로 구현된 도구에 탑재하여 사용할 수 있다. 또한, 이 시스템은 리눅스를 서버의 운영체제로 하고, 어플리케이션 서버로 웹로직 5.1을 사용하였다. 데이터베이스관리 시스템으로는 오라클 8i[8]를, 그리고 클라이언트 프로그램은 자바로 작성하였다. 새로운 데이터 준비 시스템의 클라이언트 부분은 다차원 데이터를 위한 시각화 도구인 DAVIS[7]에 탑재하였다. DAVIS는 자바로 작성되어 어떠한 플랫폼에서도 실행 가능한 공개 소프트웨어이다. 본 논문에서는 원격 데이터베이스에 저장되어 있는 데이터를 이용하여 마이닝에 사용하는 실제 예를 제시한다.

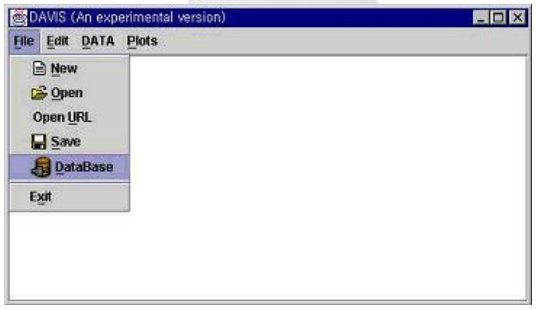

그림 3-2. DAVIS 실행창

먼저 DAVIS를 실행한다(그림 3-2). 원격 데이터베이스에 연결하기 위해서는 DAVIS 실행창에서 "DataBase" 메뉴를 선택하면 (그림3-2)과 같은 "ConnectManager" 대화창이 나타나고, 이 대화창에. 원격 데이터베이스의 IP주소, 데이터베이스 이름, 데이터베이스 사용자의 ID와 패스워드를 입력한다.

그림 3-3. Connect Manager 대화창

원격 데이터베이스와 연결되면 (그림 3-4)와 같은 창이 생성되는데 이 화면은 사용자가 입력한 원격 데이터베이스와 연결되어 질의를 할 수 있는 화면으로 "DataBase Editor" 대화창이다. 좌측창은 계층적 구조로 현재 연결된 데이터베이스의 모든 테이블의 이름을 제시하며, 테이블을 선택하면 우측창에 해당 테이블의 데이터가 나타난다. 생성된 대화 창은 연결된 데이터베이스에 저장된 데이터를 사용하기 위하여 질의 창에서 SQL문을 입력하여 데이터베이스에 대한 질의 작업을 수행할 수 있다. 좌측의 테이블 뷰어 창에서 데이터를 가져오기를 원하는 테이블을 선택하면 선택된 테이블은 XML문서로 변환되어 전송된다. 전송된 데이터의 스타일은 XSL로 구성하여 데이터를 파싱하여 보여준다.

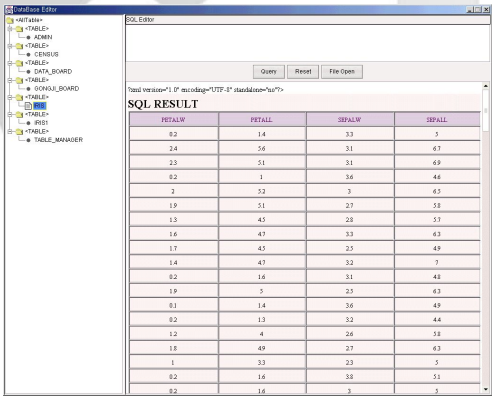

그림 3-4. 데이터베이스 관리창

현재 테이블에 조건에 의한 데이터 필터링을 하려면 SQL 명령문을 SQL Editor창에 입력하면 된다. 이 결과를 사용하여 DAVIS에서 마이닝 작업을 수행하기 위해서는 여기서 만들어진 결과는 DAVIS에서 데이터를 관리하는 "DataTable"클래스로 전달하여 클래스 내부에서 정의된 벡터에 저장하여 사용한다. 이러한 과정을 수행함으로서 DAVIS의 다른 기능들을 수행할 수 있다. 그림3-5는 데이터 테이블에 입력여부를 알려주는 "ViewData"윈도우이다.

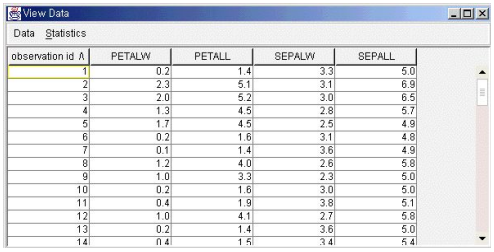

그림 3-5. 데이터 관리창

IV. 결 론

본 논문에서는 분산 네트워크 환경에서 데이터 마이닝 시스템을 구현하기 위하여 기존 데이터 마이닝 도구들의 특징과 데이터 접근 방법과 데이터 전처리 방법에 대하여 상호 비교하였다. 또한, 이러한 비교를 바탕으로 데이터 마이닝 시스템은 분산 환경에 적응 가능하여야 하며, 이에 따른 원격 데이터베이스와의 연결, 데이터의 전송문제가 대두되었다. 이러한 문제의 해결을 위하여 어플리케이션 서버를 EJB 기반의 3-티어 구조를 이용하여 설계 구현하였다. 그 결과 사용자가 네트워크로 연결된 어느 곳에서든지 데이터에 접근, 분석, 결과를 배치할 수 있도록 하는 데이터 마이닝을 위한 데이터 접근 시스템을 구현할 수 있었다.

3-티어 구조의 어플리케이션 서버 구축에 EJB로 사용함으로서 분산환경에 효율적이고도, 능동적으로 대처할 수 있도록 하였으며, 데이터의 전송은 XML을 이용하여 시스템 환경에 유용하게 대처할 수 있도록 하였다. 개발된 데이터 접근 시스템을 DAVIS에 탑재되어 사용될 수 있도록 하였으며, 실험결과 제대로 작동하는 것을 확인하였다.

본 논문에서 제시하는 데이터 접근 시스템은 자바로 작성된 데이터 마이닝 도구에는 어느 도구나 탑재가 가능하며, 이러한 데이터 접근 시스템을 탑재함으로서 적응력 높은 데이터 마이닝 도구가 되도록 지원할 수 있다.

참고문헌

[1] Huh, M. Y., Song, K. R., "The Prospect of the Structure of Data Mining Solution in the Future," International Conference on Data Mining, Visualization and Statistical System, KSS, 2000

[2] 김은수, 송강수, 이원돈, 송정길, "웹 마이닝을 이용한 개인광고기법에 관한 연구," 한국컴퓨터정보학회, 8권, 4호, 2003

[3] Georges, J., Potts, W., Wielenga, D., SAS EnterpriseMiner Software : Applying Data Mining Techniques Course Notes, SAS Software Korea, 1998

[4] Ian H. Written, Eibe Frank (2000), Data Mining Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaugmann Publishers.

[5] Sun Microsystems, Enterprise Java Beans Technology, http://java.sun.com/products/ejb/

[6] Zisman, A., An Overview of XML, Computing & Control Engineering Journal, Vol. 11, Issue 4.

[7] Huh, M. Y., Song, K. R., Nakano, J., Yamamoto, Y., Fuziwara, T., Kobayashi, I., "Davis-Jasp: a Datamining Solution by Combine Two Separate Java-based systems," Contributed Paper, ISI. 2001

[8] Banerjee, S., Krishnamurthy, V., Krishnaprasad, M., Murthy, R., "Oracle8i-the XML Enabled Data Management System," Data Engineering, Proceedings. 16th International Conference on 2000, 2000

저 자 소 개

이 상 희

1996년 ~ 현재 청강문화산업대학 컴퓨터소프트웨어과 교수

<관심분야> 분산데이터베이스, 데이터웨어하우스, 데이터마이닝

이 원 섭

1999년 ~ 현재 인덕대학 컴퓨터전자전공 교수

<관심분야> 분산데이터베이스, 데이터마이닝, XML