http://dx.doi.org/10.14400/JDC.2014.12.11.257

효율적인 데이터베이스 마케팅을 위한 데이터마이닝 전처리도구에 관한 연구

이준석
동남보건대학교 경영학과

A Study on the Data Mining Preprocessing Tool For Efficient Database Marketing

Jun-Seok Lee
Dept. of Business Administration, Dongnam Health University

요 약 효율적인 데이터베이스 마케팅을 위하여 고객들을 세분화하고, 새로운 지식을 탐색할 수 있는 데이터마이닝의 필요성이 증대되고 있다. 데이터마이닝 도구를 구축하기 위해서는 단계별 구현이 요구되어 지는데, 본 연구에서는 데이터마이닝 을 위한 분산 환경에 적응 가능한 데이터 전처리 도구를 구성하였다. 기존의 데이터마이닝 도구인 앤서 트리, 클레멘타인, 엔터프라이즈 마이너, 캔싱턴, 웨카의 전처리 부분을 고찰하고, 분산 환경에서 효율적으로 사용할 수 있는 데이터 마이닝 전처리 도구를 구성하였다. 새로이 제안된 시스템은 엔터프라이즈 자바 빈즈와 XML을 기반으로 하였다.

주제어 : 빅데이터, 데이터마이닝, 데이터마이닝 도구, 전처리, 데이터베이스 마케팅

Abstract This paper is to construction of the data mining preprocessing tool for efficient database marketing. We compare and evaluate 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 evaluated preprocessing of data mining tools are Answer Tree, Climentine, Enterprise Miner, Kensington, and Weka. We propose a design principle for an efficient system for data preprocessing for data mining on the distributed networks. This system is based on Java technology including EJB(Enterprise Java Beans) and XML(eXtensible Markup Language).

Key Words : Big data, Data Mining, Data Mining Tool, Preprocessing, Database Marketing

1. 서론

현대사회는 지금까지의 산업화 시대와는 성격이 완전히 다른 경영환경의 급격한 변화, 끊임없는 신기술의 등장, 심화되는 경쟁 환경 등으로 특징지어지는 초 경쟁 환경이라고 정의했다. 이러한, 초 경쟁 환경과 정보 폭발은 지식경영이라는 새로운 개념을 필요로 하고 있다.

기업 활동을 수행함에 있어 여러 가지 분석을 위하여

* 본 연구는 2009년도 동남보건대학교 연구비 지원에 의하여 수행된 것임
Received 16 September 2014, Revised 26 October 2014
Accepted 20 November 2014
Corresponding Author: Jun-Seok Lee(Dongnam Health University)
Email: jslee@dongnam.ac.kr

ISSN: 1738-1916

처리하여야 하는 데이터는 기하급수적으로 증가하고 있다. 저장하는 데이터 양이 증가할수록 가치 있는 정보를 추출할 수 있는 가능성도 함께 높아진다.

정보기술의 발달로 고객과의 거래가 전자적으로 이루어지는 경우가 많아져 고객 데이터가 풍부해졌으며 고객 데이터베이스에 대한 접근이 용이해지고 활용도 쉬워졌다. 따라서 기업은 고객의 구매데이터 뿐 아니라 인구통계학적 정보나 심리학적 정보를 포함한 다양한 데이터를 통해 고객의 특성을 파악할 수 있다[1].

계속적으로 다양화되고 개성화 되는 고객들의 요구에 대하여 유효적절하고 신속한 대응이야말로 기업 간 경쟁력의 척도가 된다. 지속적인 경쟁우위 확보를 위하여 효과적이고 합리적이며, 신속한 전략이나 의사결정이 중요한 의미를 지니므로, 각 기업들은 최적의 전략이나 의사결정을 지원할 수 있는 정보의 필요성이 증대되었다.

기업이 다양한 경로를 통해 수집된 고객 정보를 데이터베이스화하고, 이러한 정보를 기업의 마케팅 전략에 활용하여, 고객 개개인과의 장기적 관계를 구축하고자 하는 제반 마케팅활동을 데이터베이스 마케팅(database marketing)이라고 하는데 기업의 경쟁 환경과 마케팅의 중요성은 그 기업의 데이터베이스 마케팅 수준과 연관이 있다. 즉, 기업이 속한 업계의 경쟁이 치열할수록 기업의 마케팅 의사결정이 전략적으로 중요할수록, IT와 마케팅의 전략적 통합 정도가 높을수록, 데이터베이스 마케팅에 대한 조직 구성원들의 이해도가 높고 데이터베이스 마케팅에 대한 명확한 목표를 조직구성원들이 공유할수록 그 기업의 데이터베이스 마케팅 수준은 높다고 말할 수 있다[2].

또한, 기업은 서로 다른 성격을 가진 고객을 일정한 집단으로 분류하여 고객의 특성을 활용할 수 있는 고객세분화 전략을 사용하고자 한다[3].

그러므로, 기업의 경영 활동에 사용되어 부가가치를 창출할 수 있는 것은 모두 지식이라 할 수 있다. 이러한 지식은 사용되어지기 위하여 필요할 때에 찾아 낼 수 있어야 하며, 용도에 맞게 활용될 수 있도록 관리되고 제공되어야 한다. 이러한 대용량 데이터에서 유용한 정보와 관계를 탐색하고, 모형화하여 지식을 발견하고자 하는 일련의 과정을 데이터마이닝이라고 한다[4].

데이터마이닝은 은행이나 기업에서 고객집단 분류, 사기행위 탐색, 고객가치 관리, 마케팅(marketing), 통신 산업, 보험업, 유통업과 같은 여러 산업분야에서 효과적으로 적용할 수 있다.

그러므로, 각 데이터 항목이나 속성간의 내재된 관계나 기존의 통계학적 기법을 통하여 추출해내기에는 복잡한 관계를 탐색하고, 탐색된 데이터항목이나 속성간의 관계를 가지고, 미래를 예측하는 기술이므로, 데이터마이닝은 잠재적으로 유효하고, 새롭고 타당성 있으면서 궁극적으로 데이터에서 이해할 수 있는 어떤 패턴을 탐색하는 일련의 과정이라고 할 수 있다.

이러한 데이터마이닝을 통해 전자상거래 쇼핑몰에서 고객을 대상으로 세분화작업을 실시하여 타겟 고객에게 차별화된 서비스를 제공하는 고객 맞춤형 웹서비스의 요구가 증대되고 있다[5].

따라서 데이터마이닝을 효율적으로 수행하기 위하여 클러스터링(clustering), 시계열 분석 등 각종 통계기법과 데이터베이스 기술 뿐 만 아니라 산업공학, 신경망, 인공지능, 전문가시스템, 퍼지이론, 패턴인식, 기계적 학습, 불확실성 추론, 정보검색에 이르기까지 각종 정보기술과 기법들을 사용하여야 데이터마이닝을 통하여 데이터베이스에 숨어있는 전략적인 정보를 발견할 수 있고, 이를 데이터베이스 마케팅에 활용할 수 있다.

데이터마이닝을 수행하기 위해서는 여러 단계를 거쳐야 한다[6].

첫째, 데이터 준비(data preparation)단계로 데이터마이닝의 목적을 정하고, 마이닝의 대상이 되는 데이터와 데이터베이스를 선택하여 준비한다.

둘째, 데이터 필터링(data filtering)단계로 잘못되거나 손상된 데이터를 처리하고, 모형을 이용한 예측에 정확성을 증대시킬 수 있는 데이터의 부분집합의 변환과 선택과정을 수행한다. 데이터 정제과정(data cleansing)이라고도 한다.

셋째, 데이터마이닝 단계로 실제 데이터마이닝 도구를 적용하여 마이닝 작업을 수행한다. 의사결정나무(decision tree)와 같은 데이터마이닝 기법들을 이용하여 분석 가능한 데이터의 구조, 경향이나 관계를 생성한다. 데이터마이닝 기법들은 기계학습이나 통계학적 기법 등을 이용한다.

넷째, 해석과 평가(interpretation & evaluation)단계로 생성된 지식이 유용하면, 문제에 적용하여 사용한다. 그렇지 않으면, 이전의 단계를 다시 수행해야 한다.

다섯째, 실행(take action)단계로 경영상의 유용한 결정을 하기 위하여 위와 같은 과정을 통해 얻어진 지식을 사용한다.

이러한 데이터마이닝의 5가지 단계에서 실제로 데이터마이닝 도구를 구현하기 위하여 기본적으로 데이터 준비 단계가 필수적이며, 실제 구현에도 가장 많은 기간과 비용이 소요되는 부분이다. 또한, 데이터 준비단계에서 중요한 부분은 바로 데이터마이닝도구가 데이터베이스와 연결하여 데이터베이스에 저장된 데이터를 마이닝도구에서 사용될 수 있도록 데이터 억세스(data access)와 전처리(preprocessing)하는 과정이다.

이러한 과정에서는 데이터베이스가 지역적 데이터베이스 뿐 만 아니라 네트워크로 연결된 원격 데이터베이스도 접속하여 연결 가능하여야 하므로 이에 따른 새로운 기술이 요구된다. 그리고, 데이터베이스가 저장된 서버(server)와 데이터마이닝 도구가 수행되는 클라이언트(client)사이에 컴퓨터의 기종이나 운영체제가 상이하여 발생하는 문제도 있으므로, 이에 대한 해결방법도 요구되고 있다.

Enterprise Miner, Clementine, 그리고 Answer Tree는 ODBC(Open Database Connectivity)를 통하여 데이터베이스에 접근한다. 따라서 이러한 데이터마이닝 도구들은 ODBC드라이버가 지원된다면 어떤한 지역 데이터베이스에도 접근이 가능하다. 하지만 이러한 데이터마이닝 도구들은 원격 데이터베이스에 접근하는데 있어 각기 다른 제약을 가지고 있다[7].

제안된 전처리 시스템은 엔터프라이즈 자바 빈즈(Enterprise Java Beans) 와 XML(eXtensible Markup Language)을 이용하여 분산객체의 구현에 적용가능하며, 이기종간의 데이터 교환이나 처리에 적합하며, 쉬운 네트워크 환경 구축이 가능하다도록 하였다. 이를 기반으로 기존의 데이터마이닝 도구보다 분산환경에서 데이터마이닝이 용이하게 구현될 수 있도록 하는 데이터마이닝 전처리 시스템을 구현하였다.

2. 데이터마이닝을 위한 전처리 과정

데이터마이닝을 위한 전처리과정은 데이터를 처리하고 분석하기 전에 얻고자 하는 정보를 정의하고, 데이터를 선택, 정제하는 과정이다.

실제 데이터마이닝에 사용되는 데이터베이스나 데이터 웨어하우스는 불완전하고, 잡음이 있으며, 상호 모순이 있는 데이터를 포함하고 있다. 이러한 데이터에서 누락된 데이터를 채워 넣고, 손상된 데이터를 복구하며, 모순이 있는 데이터를 수정하는 작업이 데이터 정제이다.

데이터 정제 작업 중에서 결측 값을 처리하는 방법은 결측 값이 있는 데이터를 무시하는 방법, 결측 값을 수작업으로 채우는 방법, 결측 값이 있는 필드에 'unknown'과 같이 동일한 내용을 채우는 방법, 결측값을 속성의 평균으로 채우는 방법, 결측값을 같은 클래스에 속하는 모든 표본에 대하여 속성 값의 평균으로 채우는 방법, 베이지안 분석이나 의사결정나무 분석과 같은 기법을 이용하여 가능한 값을 채우는 방법 등 여러 가지가 있는데 특성에 따라 전처리 시스템 설계자가 선택하여 사용할 수 있다.

데이터의 통합은 여러 곳에 나누어 저장된 데이터를 결합하여 분석하기 위해 필요하다. 실제 여러 곳에 저장된 데이터를 어떠한 기준으로 통합시킬 것인가에 대한 방법과 통합 시 데이터의 크기가 커지는 문제가 있어 이에 대한 해결이 필요하다.

데이터 변환은 집락화나 회귀분석(regression)을 이용하여 데이터로부터 잡음을 제거하는 평활화(smoothing)가 있고, 데이터의 일반화나 정규화, 속성 추가 작업이 있다.

데이터의 크기를 줄이는 작업으로 축소된 데이터를 이용하여 마이닝 작업을 수행하면 더욱 효율적일 뿐 만 아니라 원래 데이터를 이용한 분석과 비교하여도 같은 결과를 얻을 수 있다[8].

데이터 축소 방법은 데이터 큐브 집합(data cube aggregation), 차원 축소, 데이터 압축 등의 방법이 있다.

파일이나 데이터베이스에서 마이닝을 하기 위한 데이터를 선택하여 사용 가능하도록 데이터마이닝 도구에 입력하는 작업을 수행하는 데이터 억세스과정과 입력된 데이터나 데이터베이스에 불완전하며, 오류가 있는 데이터가 존재할 수 있기 때문에 데이터의 무결성과 질을 높이기 위한 작업을 수행하는 데이터 전처리과정으로 세분할 수 있는데 데이터 억세스에서는 데이터를 저장하는 파일이나 데이터베이스에 접근하는 방법은 파일이나 데이터베이스가 데이터마이닝 도구와 동일한 컴퓨터에 있는 경

우에는 기존의 여러 응용프로그램이나 운영시스템의 객체를 사용함으로서 데이터의 크기만 문제가 될 뿐 별 다른 문제점이 없으나, 파일이나 데이터베이스가 동일한 컴퓨터가 아닌 원격지에 위치한 컴퓨터에 저장되어 있는 경우는 네트워크로 접속하여 데이터에 접근하는데 많은 기술적 지원이 필요하다.

또한, 데이터 전처리 과정에서는 수집된 데이터의 정확성을 높이기 위하여 데이터 내에 존재하는 오류 값이나 특이 값을 보정하고, 결측값(missing value)을 처리하며, 중복데이터를 제거하는 작업을 수행하며 데이터 정제, 또는 데이터 필터링이라고도 한다.

데이터 전처리는 입력되는 데이터베이스의 행 또는 열을 삭제하거나, 조건에 만족하는 데이터를 추출하는 단순한 질의어로도 처리 가능한 부분에서부터 데이터의 값이나 형을 변환하는 작업까지 다양한 종류가 있는데 필터링 기능을 기반으로 몇 가지로 나누어 볼 수 있다.

첫째, 데이터베이스 관리 부분으로, 데이터 테이블의 선택, 행이나 열의 삭제, 정렬이나 병합, 널(null) 값의 삭제와 같이 작업이다.

둘째, 표본 추출로 대용량의 데이터를 기반으로 하기 때문에 시스템의 효율이 떨어지거나 시간적인 비용이 요구될 수 있다. 이때 고려하여야 하는 과정이 바로 표본 추출로, 표본 추출이란 방대한 양의 데이터인 모집단에서부터 모집단을 닮은 작은 양의 표본 데이터를 추출하는 것이다.

셋째, 데이터 변환으로 분석자나 담당자의 요구에 의해 기존 변수를 이용하여 새로이 생성하고, 수정 변환작업이다. 예를 들어 각 고객의 연령정보는 있지만 너무 다양하여 10대, 20대, 30대 등으로 크게 나누어야 할 경우나 데이터마이닝 도구가 한글은 인식하지 못하는 경우 한글로 기록된 데이터 값을 영어나 수치데이터로 변환시켜야 하며, 신경망 기법 사용 시 모든 데이터 값을 0과 1 사이로 변환하는 것이 필요하다. 이러한 정보는 기존의 변수를 이용하여 새로이 생성해야 할 변수가 되며, 이러한 작업들을 수행하는 단계이다.

3. 데이터마이닝 전처리 도구의 구성

SPSS에서 개발한 앤서트리(Answer Tree)는 의사결정나무 알고리즘으로 모집단을 분할한 표본에 적합하도록 하기 위해 각각의 변수에 가중 값을 할당할 수 있다. 앤서 트리는 ODBC 드라이버가 설치되어 있는 어떤 데이터베이스와도 데이터베이스 질의를 수행할 수 있으며, 단계별로 연결된 데이터베이스의 테이블이나 필드 등을 선택해 사용할 수 있다. 연결하고자 하는 데이터베이스의 ODBC가 사용자 컴퓨터에 설치되어 있지 않으면 ODBC 데이터 원본 관리자를 이용하여 새로 지정할 수 있다.

클레멘타인은 1992년 SPSS에서 개발한 데이터마이닝 도구로 마이닝의 각 단계를 나타내는 아이콘을 이용하여 데이터 흐름의 단계를 시각적으로 나타내어, 마이닝작업을 수행할 수 있다. 클레멘타인을 이용한 데이터마이닝은 발견을 통한 프로세스(discovery -driven process)로, 이전의 과정에서 발견된 지식을 바탕으로 다음 과정을 선택하는데 사용자의 전문지식을 적용함으로써 결과를 구할 수 있다.

클레멘타인의 데이터 억세스는 ODBC로 원격 데이터베이스와 연결한 후 SQL을 이용한 질의어로 연결된 데이터베이스의 데이터를 가져와 데이터마이닝 작업을 수행한다. ODBC로 연결한 데이터베이스의 데이터이외에도 플랫(flat)형태의 파일이나 SPSS, SAS, 엑셀 또는 다른 데이터 소스와도 데이터 입출력이 가능하도록 하는 인터페이스가 있다.

또한, 데이터의 조작을 위하여, 표본 추출, 선택, 새로운 필드의 구성, 데이터의 결합을 실행할 수 있으며, 문자나 기호의 처리뿐 만 아니라 날짜나 시간의 처리에도 기초적인 수학이나 산술식을 사용할 수 있다. 또한, 시계열 자료의 처리와 어떠한 데이터의 순차적인 형태를 처리할 수 있다.

클레멘타인은 특히, 원격 데이터베이스와 연결하기 위해 ODBC 노드를 이용하는데, ODBC 노드는 ODBC 연결을 위한 속성 명세, 데이터 소스로부터 연결, 또는 연결 끊기와 같은 옵션, 연결된 데이터베이스의 테이블 보기 기능이 있고, 연결된 데이터베이스에 대하여 질의어를 이용한 질의가 가능하다. 이러한 ODBC 연결 요소들은 원격 데이터베이스를 관리하기 위하여 필요하며, 요소들은 ODBC 대화 창에서 프로퍼티 버튼으로 제어 가능하다.

SAS에서 개발한 엔터프라이즈 마이너(Enterprise

Miner)는 기본적으로 제공하는 기능별 작업 도구를 작업의 순서와 동일하게 드래그 앤 드롭(drag and drop)방식으로 배열함으로서 PFD(process flow diagram)를 구성하고 작업의 전체 과정을 한 화면에서 제어, 관리할 수 있는 마이닝 도구이다.

모형화 기법도 신경망, 의사결정나무분석, 회귀분석과 같은 전통적인 통계 분석 방법 뿐 만 아니라 최신의 다양한 마이닝기법을 제공하고, 사용에 따라서는 사용자가 정의한 모형을 이용할 수도 있다. 마이닝과정에서 구축된 두 개 이상의 모형을 그래프화하여 시각적으로 비교, 평가할 수 있는 여러 가지 챠트를 제공하며, 이를 이용하면, 구축된 모형 중 성능이 가장 좋은 모형을 손쉽게 선택할 수 있다.

엔터프라이즈 마이너는 클라이언트/서버 환경을 지원하여 대용량 데이터 모형화의 원격지 수행이 가능하다. 인포믹스, 오라클, 사이베이스, DB2와 같은 다양한 데이터베이스와 데이터 웨어하우스에 접근하여 데이터를 추출한다.

Input Data Source 노드를 이용하여 입력 데이터를 받는 부분으로 클라이언트/서버 환경에서는 서버에 대한 프로파일이 필요하다. 서버에 대한 프로파일은 연결하고자 하는 호스트의 IP, 연결을 위한 프로토콜, 연결 포트, 연결하고자 하는 데이터의 위치를 저장한 파일이다.

입력 데이터가 지정되면 출력 데이터의 이름, 간단한 설명, 자료의 역할, 관찰 치의 수, 변수의 수 등에 대한 정보가 데이터 대화상자에 나타나고, 변수의 특성, 요약 통계량 등의 정보가 얻어진다. 여기서 자료의 역할이란 선택된 데이터가 분석 과정에서 어떻게 사용될 것인지를 설정해 주는 것이다.

엔터프라이즈 마이너는 데이터 입출력이 용이하며, 어떠한 종류의 데이터베이스에도 연동될 수 있다. 또한, 데이터 처리에 한계가 없고, 병렬처리 알고리즘을 통해 처리능력을 향상시켰다. 다양한 데이터마이닝 기법들을 제공한다. 로지스틱 회귀모형, 의사결정나무, 신경 망 기법, 연관성(association) 측정, 집락화와 같은 다양한 모형화 기법을 지원한다.

켄싱턴은 InforSense사의 마이닝 도구로 인터넷과 같은 네트워크를 이용한 데이터분석을 위한 지식경영 플랫폼으로 설계되었다. 그러므로, 해당 조직내의 내부 데이터베이스를 포함하여, 온라인 데이터베이스와 같은 어떠한 데이터 소스라도 탐색이나 접근이 가능하도록 설계되었다.

켄싱턴은 분산 클라이언트/서버 실행모형로 이식성, 확장성, 네트워킹 기능을 자바로 구현하여, 분산 컴퓨팅이 가능하며 컴포넌트(component)의 멀티-티어 구조를 통해 원격 데이터베이스에 접근할 수 있다. 그러므로, 조직내의 인트라넷에서의 내부 데이터베이스를 포함하여, 온라인 데이터 소스와 같은 어떠한 데이터 소스라도 탐색이나 접근이 가능하도록 설계되었다. 켄싱턴의 데이터마이닝 데이터 억세스는 원격 데이터베이스와 연결 후 데이터소스를 전송 받기 위하여 데이터베이스의 위치에 대한 명세가 필요하다. 접근할 수 있는 인터넷 주소와 데이터베이스의 로그-인(log-in)에 대한 설명을 표현하기 위해 켄싱턴에서는 북마크(bookmark)라는 개념을 사용하였다.

북마크 윈도우에서는 연결한 데이터베이스의 타입을 명세한다. 명세할 내용은 데이터베이스 자체의 이름, IP 주소(데이터베이스의 위치, 포트번호), 사용자 이름, 암호이다. 또한, 추가로 데이터베이스의 별명을 명기할 수 있다.

웨카는 Waikato 대학교에서 개발한 데이터마이닝 도구로서 공개 소프트웨어이며, 자바로 작성되어 어떤 플랫폼에서도 실행가능하며, 직접 데이터를 사용하여 적용하거나, 사용자가 작성한 자바코드에서 호출하여 사용할 수 있다. 의사결정 나무기법, 선형회귀, 모형 트리 생성 등의 기능을 가지고 있으며, 집락화 기법과 연관성 규칙의 기능도 가지고 있다. 웨카는 원격 데이터베이스와 접속하기 위하여 RMI기법을 이용하는데, 제공된 클래스 중 weka.experiment 패키지내의 RemoteExperiment 클래스에서 RMI 브릿지로 원격 데이터베이스와 접속한다. 이때 원격 데이터베이스는 웨카의 클래스들이 접근할 수 있도록 지원되어야 한다. RemoteEngine 클래스는 RMI를 통하여 객체를 전송하기 위한 서버 클래스로서 접근할 원격 데이터베이스의 종류에 따라 각각의 데이터베이스에 해당하는 JDBC드라이버가 필요하다.

이러한 기존의 데이터마이닝 도구는 데이터베이스와 연결하기 위한 구조를 가지고 있으나 몇 가지 문제점이 있다.

첫째, 대부분의 데이터 억세스 시스템이 클라이언트/서버의 2-티어 구조를 사용하고 있어, 클라이언트의 사

용자 프로그램에서 원격 데이터베이스와 연결하는 작업을 수행한다. 이러한 구조는 사용자 프로그램 자체에 데이터베이스 전용의 클라이언트 소프트웨어, 해당 데이터베이스의 ODBC까지 설치되어야만 작동이 가능하다.

또한, 클라이언트에 작업 처리를 위한 소스 코드 구현 부분인 비즈니스 로직이 포함되어 있어 클라이언트 프로그램의 크기가 커져 전체 시스템의 성능에 영향을 줄 수 있다.

둘째, 위와 같은 구조에서는 시스템이 데이터베이스 자체나 데이터베이스 엔진을 변경 또는 업그레이드하려면 많은 클라이언트들의 환경 설정을 고려하여야 하고, 계속적인 클라이언트 프로그램의 버전 향상이 필요하다.

아울러, 이러한 방식으로 버전을 향상시키면 어떤 경우에는 데이터베이스 엔진의 버전이나 클라이언트 프로그램의 버전 차이 때문에 문제가 발생할 수 있다.

셋째, 연결하고자 하는 데이터베이스에 대한 ODBC 드라이버가 클라이언트에 설치되어 있어야 하며, 만약 ODBC 드라이버가 설치되어 있지 않은 데이터베이스와는 연결할 수 없다. 그러므로, 사용자가 별도로 설치하여야만 한다.

넷째, 원격 데이터베이스와의 연결을 위해서 별도의 프로그램이 필요한 경우도 있다. 자체적인 원격 연결 구조가 없는 경우 다른 프로그램으로 원격 서버와 연결한 후 데이터베이스와 접속해야 하는 경우가 있으며, 이는 진정한 의미의 클라이언트/서버 환경이라고 할 수 없다.

다섯째, 데이터베이스로부터 가져온 입력파일의 형태가 클라이언트 프로그램에서 사용 가능하도록 일정한 형태로 변환되어야 한다. 이러한 변환과정은 대량의 데이터를 사용하는 데이터마이닝 도구에서는 전체 시스템의 성능에 영향을 줄 수 있다.

이러한 문제점들을 해결하는 새로운 시스템을 구현하기 위하여 클라이언트/서버 방식의 2-티어 구조대신에 어플리케이션 서버가 추가된 멀티-티어 구조를 사용한다. 3-티어 구조를 사용하면 클라이언트 프로그램은 단순한 사용자 인터페이스 프로그램이 될 수 있다.

또한, 클라이언트와 서버의 기종이나 운영체제가 상이할 경우 데이터베이스에서 데이터를 클라이언트로 가져오는 기법이 이기종간의 데이터 교환이 적합하고, 서버와 데이터베이스, 데이터베이스와 클라이언트 사이에 데이터 상호 전달이 많이 발생하는 미들-티어에서 데이터들의 변환이 가능한 구조로 구성하여야 한다.

2-티어 구조 환경에서는 클라이언트에는 데이터베이스 엔진과 거기에 필요한 ODBC와 같은 드라이버들, 그리고, 데이터베이스 전용의 클라이언트 소프트웨어까지 설치되어야만 데이터베이스 프로그램이 제대로 작동할 수 있다. 그리고 데이터베이스 클라이언트 어플리케이션 안에는 여러 가지 일을 처리하기 위한 구현 부분인 소스 코드들이 많이 포함되어 있는데, 이러한 구조에서는 클라이언트 프로그램이 멀티 티어에 비하여 상대적으로 크기 때문에 클라이언트 컴퓨터의 성능이 전체 시스템의 성능을 결정한다.

또한, 시스템에서 클라이언트 프로그램 자체나 데이터베이스 엔진 혹은 데이터베이스 전용 클라이언트 소프트웨어를 업그레이드하려면 많은 클라이언트들의 환경 설정을 고려하여 업그레이드 방안을 마련해야 한다. 어떤 경우에는 데이터베이스 엔진의 버전이나 데이터베이스 클라이언트 소프트웨어의 버전 차이 때문에 예상치 못한 문제가 발생할 수도 있다.

이러한 문제를 해결하기 위해 멀티 티어 구조를 사용한다. 클라이언트/서버의 2-티어 구조에 데이터베이스 서버에 접근해서 데이터를 처리하는 중개자 혹은 어떤 업무에 필요한 비즈니스 로직들을 따로 별도의 어플리케이션에 저장하는 어플리케이션 서버를 추가하는 것이다.

이러한 방식을 이용하면 클라이언트 어플리케이션은 단순히 사용자 인터페이스만 가지고 있는 간단한 클라이언트가 될 수 있다. 클라이언트, 어플리케이션 서버, 데이터베이스 서버 등이 논리적으로 3개의 계층을 이루고 있다고 해서 3-티어 구조라고도 하고, 이런 방식으로 작성한 어플리케이션을 분산 처리 어플리케이션이라고 부른다. 여기서 어플리케이션 서버가 위치하는 곳을 논리적으로 미들 티어라고 한다.

데이터 전처리 도구는 데이터베이스와 연결, 관리하는 엔터프라이즈 자바 빈즈로 구성한 어플리케이션 서버와 어플리케이션 서버에 대한 SQL질의와 질의결과를 XML 문서로 전달받는 클라이언트로 구성된다.

어플리케이션 서버는 JDBC와 데이터 억세스 시스템의 한 부분인 SQLMapper 클래스를 이용하여 원격 데이터베이스와 접속하고, 검색된 데이터를 XML 파싱 도구를 이용하여 XML문서로 변환한다. 변환된 XML문서는 엔터프라이즈 자바 빈즈 콘테이너(container)에 의하여

클라이언트로 전송된다.

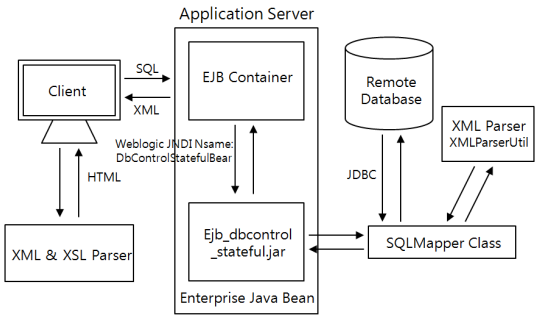
데이터 전처리 도구 구성을 위하여 클라이언트가 어플리케이션 서버에게 작업 수행을 요청할 때 이루어지는 서버의 동작은 다음과 같다.

첫째, 어플리케이션 서버는 클라이언트의 요청을 처리한다.

둘째, 어플리케이션 서버에 배치되어 있는 엔터프라이즈 자바 빈즈는 클라이언트의 요청을 받은 엔터프라이즈 자바 빈즈 콘테이너에 의해 호출되어 원격 데이터베이스의 여러 클라이언트의 요청을 실질적으로 수행하고 그 결과를 XML 문서로 파싱한다.

셋째, 원격 데이터베이스의 연결은 자바의 JDBC 응용프로그램 인터페이스를 사용하여 엔터프라이즈 자바 빈즈에서 수행한다.

다음의 [Fig. 1]은 클라이언트, 어플리케이션 서버, 데이터베이스간의 동작 원리에 대한 데이터 전처리 도구의 구성도이다.



[Fig. 1] Construction of data mining preprocessing tool

서버와 클라이언트의 데이터 전송은 XML 문서 양식을 사용한다. 본 논문에서 구현한 시스템에서는 문서 양식을 XSL을 이용하여 문서 통합이 용이하여, 문서 표준을 만들 수 있는 장점을 가지고 있다.

파서는 아파치사의 Xalan을 사용했는데, Xalan은 XML문서를 HTML, 텍스트 또는, 다른 XML 문서 형태로 만들어 주는 변환기능을 수행하며, 가장 많은 응용프로그램 인터페이스를 제공하고 다른 파서보다 한글에 관련한 문제 발생이 적다.

클라이언트는 자바로 구현된 데이터마이닝 도구에 탑재되어 사용할 수 있는데 서버에서 받아온 XML, XSL 파일을 파싱해서 HTML 파일로 만든 후 화면에 표시하며, 탑재한 도구와 데이터 교환이 가능이 하도록 하였다.

4. 결론

효율적인 데이터마이닝 도구는 분산 환경에 적응 가능하여야 하며, 이에 따른 원격 데이터베이스와의 연결, 데이터의 전송, 변환문제의 해결이 요구되며, 이러한 문제의 해결을 위하여 분산 컴포넌트의 멀티-티어 구조를 이용하여, 사용자가 네트워크로 연결된 어느 곳에서든지 데이터에 접근, 분석, 결과를 배치할 수 있도록 하는 데이터마이닝을 위한 데이터 전처리 도구를 구성하였다.

데이터 전송은 데이터의 관리나 시스템 환경에 적응성 있는 구조를 제공함으로서 이기종 서버간의 데이터 교환문제이나 데이터의 형식에 따른 변환문제의 해결이 필요하였는데, 멀티-티어 구조에서 어플리케이션 서버의 기능구축을 엔터프라이즈 자바 빈즈로 사용함으로서 분산환경에 효율적이고도, 능동적으로 대처할 수 있도록 하였으며, 데이터의 전송은 XML을 이용하여 시스템 환경에 유용하게 대처할 수 있도록 하였다.

어플리케이션 서버는 클라이언트의 요청을 처리하기 위하여, 클라이언트의 요청을 받은 엔터프라이즈 자바 빈즈 콘테이너에 의해 호출된 서버 측을 담당하는 엔터프라이즈 자바 빈즈는 원격 데이터베이스에 대한 클라이언트의 요청을 실질적으로 수행하고 그 결과를 XML 문서로 파싱한다. 원격 데이터베이스의 연결은 자바의 JDBC 응용프로그램 인터페이스를 사용하여 엔터프라이즈 자바 빈즈에서 수행하게 된다.

이러한 도구를 통하여 데이터베이스가 원격지에 위치하더라도 데이터베이스에 대한 몇 가지 정보만 사용자가 알고 있으면 네트워크상의 어떤 곳에서도 접속이 가능하도록 하였고, 클라이언트에서 질의어를 이용하여 데이터베이스에 대하여 데이터 선택, 입력, 갱신, 삭제 작업을 수행할 수 있도록 하였다.

ACKNOWLEDGMENTS

This study was supported by a grant of the Dongnam Health University.

REFERENCES

[1] T. H. Hong, E. M. Kim, Predicting the Response of Segmented Customers for the Promotion Using Data Mining, Information Systems Review, Vol. 12, No. 2, pp. 75-88, 2010.

[2] J. M. Lee, J. S. Park, J. B. Jang, An Investigation of the Factors Influence Database Marketing Sophistication, Journal of the Korea Industrial Information System Society, Vol. 6, No. 3, pp. 95-106, 2001.

[3] C. Ordonez, Z. Chen, Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis, IEEE Transactions on Knowledge & Data Engineering. Vol. 24, Issue 4, pp 678-691, 2012.

[4] Y. H. Jung, S. H. Eo, H. S. Moon, H. J. Cho, A Study for Improving the Performance of Data Mining Using Ensemble Techniques, Communications of the Korean statistical society, Vol. 17, No 4, pp. 561-574, 2010.

[5] J. A. Berry & G. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, John and Sons, 2004.

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

[7] Y. G. Choi, A Study for Improving the Performance of Data Mining Using Ensemble Techniques, Journal of Information Technology Applications & Management, Vol. 15, No. 2, pp. 1-14, 2008.

[8] X. Wu, X. Zhu, G. Q. Wu, IEEE Transactions on Knowledge & Data Engineering. Vol. 26, Issue 1, pp 97-107, 2014.

이 준 석(Lee, Jun Seok)

- 1988년 2월 : 청주대학교 전자계산학과(공학사)
- 1991년 8월 : 청주대학교 전자계산학과(공학석사)
- 1995년 2월 : 한남대학교 전자계산공학과(공학석사)
- 2001년 8월 : 성균관대학교 통계학과(통계학박사)
- 1994년 3월 ~ 현재 : 동남보건대학교 경영학과 교수
- 관심분야 : 빅데이터, 데이터마이닝
- E-Mail : jslee@dongnam.ac.kr