데이터 마이닝 도구 XM-Tool/Miner 개발에 관한 연구

이남근*, 이창호*, 김주용*, 이병엽*, 이승희*

*대우정보시스템 기술연구소
e-mail: ngrhee@disc.co.kr

A Study on the Development of Data Mining Tool named XM-Tool/Miner

Nahm-Guhn Rhee*, Chang-Ho Lee*, Ju-Young Kim* ,Byung-Yup Lee*, Seung-Hee Lee*
*Dept of Information Technology, Daewoo Information Systems Co., Ltd.

요약

정보기술이 발달하면서 자료의 흔적들이 체계화된 데이터베이스에 저장이 되고, 더불어 데이터베이스의 규모는 점점 커지고 있다. 데이터 마이닝은 이런 방대한 자료의 분석을 통해, 그 속에 숨어있는 의미를 찾는 과정이라고 볼 수 있다. 본 논문에서는 대용량 데이터베이스에 존재하는 여러 유용한 지식을 추출하는 방법으로서 데이터 마이닝을 분류화, 클러스터링, 요약규칙, 시간에 따른 분석 및 예측등으로 분류하여 요약, 제시하였고, 이렇게 추출된 패턴, 정보, 지식들의 유용성을 측정하는 지표를 정리하였다. 개발된 XM-Tool/Miner은 문제 중심적 마이닝 도구를 목표로 하였으며, 대표적인 마이닝 알고리즘을 적용하였고, 또한 사용의 편이성에 초점을 맞추었다. 더 나아가 데이터 마이닝 기법뿐만 아니라 데이터의 샘플링과 성능향상을 통하여 방대한 데이터로부터 다양한 지식탐사가 가능해지고, 발견된 규칙 또는 지식의 유용성 측정을 통하여 업무 분야의 특성에 따라 효과적으로 반영되며 의사결정 및 CRM마케팅, 동향분석 및 예측 등에 유용한 정보를 추출하는 도구로 사용할 수 있을 것이다.

1. 서론

데이터 마이닝은 정보기술의 발달과 더불어 발전되어 왔다. 특히 데이터베이스기술의 발달과 Data Warehousing 기법, 인공지능(Artificial Intelligence)의 전문가 시스템(Expert System)과 기계학습(Machine Learning), EDA(Exploratory Data Analysis) 등이 데이터 마이닝을 발달시키는데 주요한 요인을 제공하였다. 기업경영에서 마케팅 정보에 대한 요구와 기대가 달라진 것도 데이터 마이닝을 확산시키는데 중요한 요인이라고 볼 수 있다.

데이터 마이닝은 현재 광범위한 영역에서 활용되고 있으며, 가장 고전적인 활용 영역은 기업에서 데이터베이스 마케팅을 포함한 CRM(Customer Relation Management)이나 부정사용방지(Fraud Detection) 등에서 많이 적용되고 있다. 또한 다양한 부분에 구축되어 있는 자료들을 통합해서 보다 상위수준의 전략적 지식을 도출하기 위해서도 많이 이용되고 있다. 현재는 비단 기업에서 뿐 아니라, 공공기관, 특히 검찰이나 경찰, 생명공학 등 보다 복잡한 정보분석이 요구되는 기관에서는 예외 없이 데이터 마이닝을 활용하고 있다. 더욱이 인터넷상의 Transaction이 늘어나면서, 인터넷상에서 사용자의 행동특성을 분석하는 Web Mining과 자료검색 기능을 보다 확장한 Text Mining 기법 등을 통해, e-Business를 위한 제반 마케팅 정보와 소비자 행동 모형들이 만들어지고 있다. 이런 추세라면 데이터 마이닝 기법과 그 분야는 앞으로도 계속 확장될 전망이다.

2. 데이터 마이닝(Data Mining) 과 CRM(Customer Relationship Management)

2.1 데이터 마이닝

최근 수년동안 데이터 마이닝에 대한 많은 연구가 이루어졌는데, 그동안 제안된 다양한 데이터 마이닝 기법들은 어떤 형태의 지식을 발견하고자 하는가, 어떤 종류의 데이터베이스에 적용될 수 있는가, 어떤 분야의 기술에 바탕을 두고 있는가 등의 기준에 의하여 데이터 집합의 일반적 특성을 분석하는 특성화(characterization), 특정 인자(factor)에 의해서 구분되는 데이터베이스 상의 다른 클래스에 대한 차별적인 특성을 추출하는 분류화(classification), 사전 정보가 주어지지 않은 상태에서 데이터베이스 내에서 유사한 특성을 갖는 데이터들을 묶는 군집화(clustering), 여러 개의 트랜잭션들 중에서 동시 발생하는 트랜잭션의 연관관계를 발견하는 연관규칙탐사(association), 시계열 데이터(주식, 물가, 판매량, 과학적 실험 데이터)들이 시간 축으로 변하는 전개과정을 특성화하여 동적으로 변화하는 데이터의 분석을 수행하는 경향분석(trend analysis), 대용량 데이터베이스 내의 명시된 패턴을 찾는 패턴 분석(pattern analysis)등으로 분류 될 수 있다.[1]

2.2 CRM

CRM(Customer Relationship Management)은 고객전략을 정의하고(strategy), 경영자원의 분배를 최적화하며(administration), 고객과 관련된 모든 부분에서 최상의 서비스를 제공하기 위해(service), 고객의 욕구와 행동 그리고 고객의 수익성을 규명해 가는(profitability), 지속적이고 자동화된 프로세스(automation)로서 CRM은 고객과 관련된 기업의 내외부 자료를 분석, 통합하여 고객특성에 기초한 마케팅활동을 계획하고, 지원하며, 평가하는 과정이다. 또한 CRM은 데이터 마이닝을 통해 고객데이터의 세분화를 실시하여 신규고객획득, 우수고객 유지, 고객가치 증진, 잠재고객 활성화, 평생 고객화와 같은 사이클을 통하여 고객을 적극적으로 관리하고 유도하며 고객의 가치를 극대화시킬 수 있는 전략을 통하여 마케팅을 실시한다. 이러한 CRM의 정의로 볼 때 CRM에서의 데이터 마이닝은 필수 불가결한 도구라고 볼 수 있다.

3. XM-Tool/Miner 개발

XM-Tool/Miner은 일반 사용자, 기업의 마케팅 담당자, 기획 담당자를 대상으로 CRM 분야에 최적화된 데이터 마이닝 툴의 개발을 목적으로 대우정보시스템㈜ 기술연구소의 주도하에 개발되었다. XM-Tool/Miner은 사용하기 쉬운 인터페이스, 고객 프로파일링(customer profiling), 장바구니 분석(market basket analysis) 등 CRM의 주요 이슈들에 대한 시나리오를 제공한다. 또한 XM-Tool/Miner은 Pure Java로 작성되어 플랫폼에 독립적이며, 서버와 클라이언트가 분리되어 사용자가 어느 곳에서든 클라이언트 프로그램을 다운로드 받아서 이전 작업환경을 그대로 이어서 작업할 수 있는 어플리케이션이다.

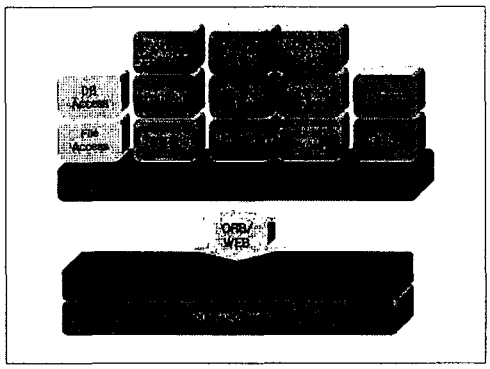
3.1 개발 목적

XM-Tool/Miner은 비즈니스 영역에 범용화 도구보다는 특화된 문제 중심적인 마이닝 도구의 개발과 기존 마이닝 툴의 문제점들을 개선하여 다양한 마이닝 기법의 자동화, 산업별 특성에 맞는 마이닝 결과의 제공, Java/Corba를 활용한 Internet의 연동, 그리고 업무 유형별 시나리오에 기반 하여 데이터 마이닝 과정의 자동화 등의 개발 목적을 두고 있다.

3.2 개발 사양

XM-Tool/Miner은 Sun社의 JDK1.2.2, Inprise社의 CORBA서버인 Visibroker 4.0 에 기반 하여 만들어진 Pure Java 어플리케이션으로서 현재 Windows NT, 98 상에서 개발, 테스트되었다.

XM-Tool/Miner의 구조는 다음 <그림 1>에서 보는 것처럼 CORBA를 기반으로 각 서버 컴포넌트(Component)가 서버를 구성하고 있으며, 클라이언트는 시나리오 매니저(Scenario Manager)가 담당하고 각 서버와 클라이언트는 CORBA를 통해서 통신하게 된다.



<그림 1> XM-Tool/Miner의 구조

3.3 주요 기능

3.3.1 Data Access

XM-Tool/Miner 의 Data Access는 크게 파일과 데이터베이스로 나뉘어 진다. 데이터 베이스는 Oracle, MS SQL Server, JDBC 드라이버와 호환되는 모든 DBMS 지원 가능하다. 파일 데이터는 콤마로 구분되는 CSV 파일 외에 기타 구분자로 분리되는 모든 파일이 사용 가능하다.

3.3.2 Data Preparation

개발된 XM-Tool/Miner 의 Data Preparation은 크게 네 가지 기능들로 구성되어 있다. 각각의 기능들을 살펴보면 아래와 같다.

- Sampling
  - Random Sampling : 샘플링 숫자와 퍼센트에 따른 샘플링 제공.
- Partitioning
  - Random Sampling : 데이터 수와 퍼센트에 따른 파티셔닝 제공.
- Transform
  - Unique Value Filtering.
  - Binary Mapping.
  - Date 타입 변환.
- Column Computing
  - Computed Column, Summarization Column 생성 기능 제공.

3.3.3 적용 알고리즘

(1) Decision Tree

의사결정 트리는 데이터의 클러스터링(Clustering)이나 분류(Classification), 결과 예측을 위해 자주 사용되는 데이터 마이닝 알고리즘이다. XM-Tool/Miner에선 퀸란(Quinlan)의 C4.5[2]를 사용하였다.

(2) Association Rules

연관 규칙이란 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 강한 연관성을 찾아내는 마이닝 기법이다.[3] 주로 검색알고리즘에 대한 연구가 많이 수행되고 있으며, XM-Tool/Miner에선 Agrawal의 Apriori 알고리즘을[4] 사용하였다.

(3) Neural Network

매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형(Flexible nonlinear Model)의 하나로, 신경생리학과의 유사성 때문에 일반적으로 다른(통계적) 예측모형에 비해 흥미롭게 여겨지고 있다. 이러한 신경망 모형은 고객의 신용평가, 불량거래의 색출, 의료진단예측, 우량고객의 선정, 타겟 마케팅의 여러 주제(DM, TM)등을 비롯한 여러 분야에 적용될 수가 있는데, 주로 교사학습에 적용되어 목적변수(target)에 대한 예측(Prediction)이나 분류(Classification)를 목적으로 감춰진 패턴을 찾고 이를 일반화하는데 이용된다. XM-Tool/Miner에선 역전파(Back Propagation) 알고리즘과 코호넨망(Kohonen map, SOFM)[5]을 사용하였다.

(4) Sequence Rules

트렌드를 식별해 내기 위해 일정시간 동안의 레코드를 분석하여 순서 패턴을 찾아내는 데이터 마이닝 기법이다.[6] Association Rule의 응용의 한가지이며, 역시 Apriori 알고리즘을 사용하였다.

(5) Episode Rule Learning (Episode Rule Algorithm)

시간의 순서에 따라 발생한 이벤트 리스트를 일정한 순서로 자주 발생하는 순서 패턴을 찾아내는 데이터 마이닝 기법이다.[7]

(6) Regression

데이터들의 연속적인 수가 값을 가질 경우 이러한 데이터들을 선으로 연결하여 선상에 존재하는 값으로 존재하지 않는 데이터를 예측하는 통계기법이다.

3.4 특징

3.4.1 플랫폼 독립적인 웹 어플리케이션

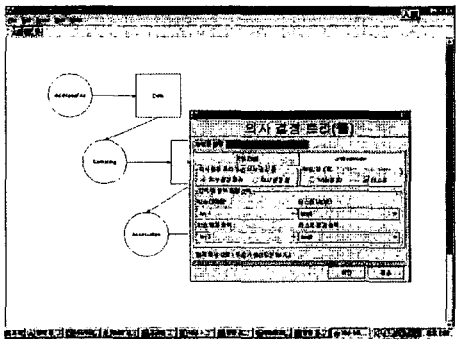
XM-Tool/Miner는 Java를 이용하여 플랫폼 독립성을 가지고 있으며, 서버와 클라이언트가 분리되어 인터넷상의 어디서나 접속이 가능하고, 클라이언트 화면을 Java Applet으로 이식하고 있다.

3.4.2 시나리오(Scenario)의 사용

XM-Tool/Miner는 CRM 관련 주요 도메인별 시나리오를 제공함으로서, 한번 작업된 데이터 마이닝 모델과 그 결과 이력을 기록, 관리함으로써 향후 데이터 마이닝 작업에서 재작업이 발생하는 것을 방지하고, 유사 문제에 쉽게 접근할 수 있도록 개발되었다. XM-Tool/Miner의 시나리오는 시나리오 매니저(Scenario Manager)가 관리한다.

3.5 실행화면

<그림 2>는 개발된 XM-Tool/Miner의 실행화면이다. 전체적으로 데이터의 추출, 변환, 모델링 등의 프로세스는 원형의 노드(node)로 표현되었고, 이의 결과로 발생하는 출력인 데이터는 사각형의 노드로 표현되었다. 사용자는 각각의 노드를 통해서 속성 정보를 편집하고 이를 실행할 수 있다.



<그림 2> XM-Tool/Miner 의 실행화면

4. 결론

기존의 대표적인 데이터 마이닝 도구들은 기법의 다양성과 범용성에 초점을 맞추는 반면, 개별 마이닝 기법들에 대한 사용 용이성과 자동화에는 소홀한 실정이다. 따라서 기존의 도구들이 다양한 도구들에 대한 해박한 지식을 갖는 전문가적 입장에서 매우 훌륭한 도구이지만, 일반적인 사용자는 사용하기에 어려운 실정이다. 따라서 대우정보시스템㈜에서 개발된 XM-Tool/Miner은 매우 일반적이고 대표적인 마이닝 기법들을 일차적으로 선정하여 기법의 다양성보다는 채택된 기법들을 일반 사용자 입장에서 용이하게 다룰 수 있는 기법의 자동화에 초점을 맞추어 개발되었고, 기업의 기능적 분야 또는 기업의 유형을 우선 결정하는 도메인 중심의 마이닝 툴을 목표로하였다. 본 연구를 통하여 개발된 XM-Tool/Miner은 업무 분야의 특성에 따른 시나리오를 통하여 효과적인 데이터 마이닝 프로세스를 수행할 수 있으며, 의사결정 및 CRM마케팅, 동향분석 및 예측 등에 유용한 정보를 추출하는 도구로 사용할 수 있을 것이다.

참고문헌

[1] 김정자, 이도헌 "데이터 마이닝 기술 및 연구동향", 정보과학회지, 16권 9호, pp. 7-14, 1998.

[2] J. Quinlan, "C 4.5 : Programs for Machine Learning", Morgan Kaufman, 1988.

[3] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Database", Proc. ACM SIGMOD, pp. 207-216, 1993.

[4] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules", Proc. VMDB, pp. 487-499, 1994.

[5] T. Kohonen, "The Self-Organizing Map", Proc. IEEE, 78, pp. 1464-1479, 1990.

[6] R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. ICDE, 1995.

[7] H. Mannila, H, Toronen and A. Verkano, "Discovery of Frequent Episodes in Event Sequences", Series of Publications C, University of Helsinki, 1997.