

# 시나리오 기반의 데이터 마이닝 도구 XM-Tool/Miner 설계 및 구현

이창호, 이남근, 이승희, 이병엽, 김주용\*

## A Design and Implementation of the Scenario-based Data Mining Tool named XM-Tool/Miner

Chang Ho Lee, Nahm-Guhn Rhee, Seung-Hee Lee, Byoung-Yup Lee, Ju-Young Kim\*

### 요약

정보기술이 발달하면서 자료의 흔적들이 체계화된 데이터베이스에 저장되고, 더불어 데이터베이스의 규모는 점점 커지고 있다. 데이터 마이닝은 이런 방대한 자료의 분석을 통해, 그 속에 숨어있는 의미를 찾는 과정이라고 볼 수 있다. 본 논문에서는 대우정보시스템(주)에서 개발된 사용자지향 데이터 마이닝 도구인 XM-Tool/Miner의 개발을 대상으로 하고 있다. 개발된 XM-Tool/Miner은 문제 중심적 마이닝 도구를 목표로 하였으며, 대표적인 마이닝 알고리즘을 적용하였고, 또한 사용의 편의성에 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 CRM 마케팅, 동향분석 및 예측 등에 유용한 정보를 추출하는 도구로 사용할 수 있을 것이다.

**Keywords:** Data mining, Data mining tool

### 1. 서론

데이터 마이닝은 정보기술의 발달과 더불어 발전되어 왔다. 특히 데이터베이스기술의 발달과 Data Warehousing, 인공지능(Artificial Intelligence)의 전문가 시스템(Expert System)과 기계학습(Machine Learning) 등이 데이터 마이닝을 발달시키는데 주요한 요인을 제공하였다. 기업경영에서 마케팅 정보에 대한 요구와 기

대가 달라진 것도 데이터 마이닝을 확산시키는데 중요한 요인이라고 볼 수 있다. 데이터 마이닝은 현재 광범위한 영역에서 활용되고 있으며, 기업에서 데이터베이스 마케팅을 포함한 CRM(Customer Relationship Management)이나 부정사용방지(Fraud Detection) 등에서 많이 적용되고 있다. 현재는 기업에서 뿐만 아니라,

---

\* 대우정보시스템(주) 기술연구소 1부

공공기관, 특히 검찰이나 경찰, 생명공학 등 보다 복잡한 정보분석이 요구되는 기관에서는 예외 없이 데이터 마이닝을 활용하고 있다. 최근 전자상거래의 활성화와 인터넷의 대중화에 따른 트랜잭션(transaction)이 늘어나면서, 인터넷상에서 사용자의 행동특성을 분석하는 Web Mining 과 자료검색 기능을 보다 확장한 Text Mining 기법 등을 통해, e-Business 를 위한 제반 마케팅 정보와 소비자 행동 모형들이 만들어지고 있다. 이런 추세라면 데이터 마이닝 기법과 그 분야는 앞으로도 계속 확장될 전망이다.

기존의 대표적인 데이터 마이닝 도구들은 기법의 다양성과 범용성에 초점을 맞추는 반면, 개별 마이닝 기법들에 대한 사용 용이성과 자동화에는 소홀한 실정이다. 기존의 도구들은 다양한 도구들에 대한 해박한 지식을 갖는 전문가적 입장에서 매우 훌륭한 도구이지만, 일반적인 사용자는 사용하기에 어려운 실정이다.

따라서 본 연구에서 연구대상으로 하는 XM-Tool/Miner 는 일반적이고 대표적인 마이닝 기법들을 선정하여 기법의 다양성보다 채택된 기법들을 일반 사용자 입장에서 용이하게 다룰 수 있는 기법의 자동화에 초점을 맞추었고, 도메인 중심의 마이닝 도구를 목표로 하였다.

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

### 2.1 데이터 마이닝

최근 수년동안 데이터 마이닝에 대한 많은 연구가 이루어졌는데, 그동안 제안된 다양

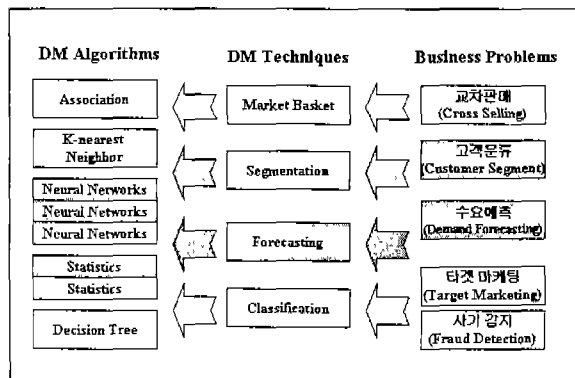
한 데이터 마이닝 기법들은 어떤 형태의 지식을 발견하고자 하는가, 어떤 종류의 데이터베이스에 적용될 수 있는가, 어떤 분야의 기술에 바탕을 두고 있는가 등의 기준에 의하여 데이터 집합의 일반적 특성을 분석하는 특성화(characterization), 특정 인자(factor)에 의해서 구분되는 데이터베이스 상의 다른 클래스에 대한 차별적인 특성을 추출하는 분류화(classification), 사전정보가 주어지지 않은 상태에서 데이터베이스 내에서 유사한 특성을 갖는 데이터들을 묶는 군집화(clustering), 여러 개의 트랜잭션들 중에서 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 데이터 베이스 내의 명시된 패턴을 찾는 패턴 분석(pattern analysis)등으로 분류 될 수 있다.[1]

### 2.2 CRM

CRM(Customer Relationship Management)은 고객전략(strategy)을 정의하고, 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 CRM 은 고객과 관련된 기업의 내외부 자료를 분석, 통합하여 고객특성에 기초한 마케팅 활동을 계획하고, 지원하며, 평가하는 과정이다. 또한 CRM 은 데이터 마이닝을 통해 고객데이터의 세분화를 실시하여 신규고객획득, 우수고객 유지, 고객가치증진, 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.

데이터 마이닝 알고리즘별 CRM 적용 가능 분야를 다음 <그림 1>에 보이고 있다. 제시된 그림은 기업의 CRM 활동의 문제(problem)을 어떠한 데이터 마이닝 기술에 의해 구현할 수 있고, 또한 이러한 데이터 마이닝 기술은 구체적으로 어떤 데이터 마이닝 알고리즘에 의해서 실행되는지를 예시하고 있다. [2]

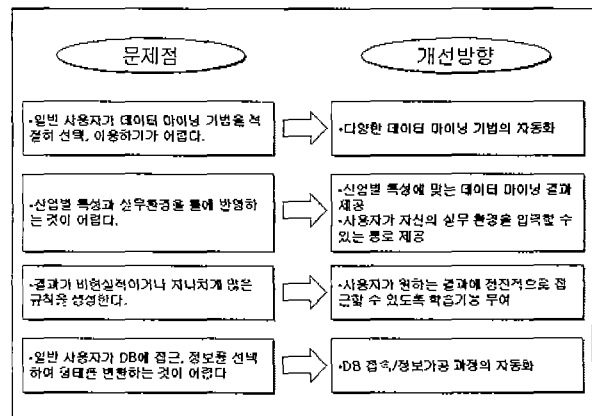


<그림 1> 데이터 마이닝 알고리즘별 CRM 적용분야

### 3. 기존 데이터 마이닝 도구의 문제점 및 개선방향

기존의 상품화 된 많은 데이터 마이닝 도구들은 통계를 기초로 하느냐, 인공지능을 기초로 하느냐에 따라서 수많은 제품이 출시되어 있다. 그러나 이러한 데이터 마이닝 도구들을 데이터 마이닝에 대한 지식이 없는 현업의 마케팅 담당자들이 사용하기엔 무척 까다롭다. 그 이유를 본 논문에선 두가지로 보고

있다. 첫째는, 통계학 및 인공지능에 대한 사전 지식의 부족에서 비롯된 문제로, 통계학이나 인공지능은 각각의 측정치, 알고리즘에 따라 사용될 변수들 뿐만 아니라 측정치 산출이나 알고리즘 수행 자체에 대한 많은 입력치(parameter) 들이 존재하며, 이 때문에 사전지식이 없는 일반인이 사용하기 어렵다. 둘째는 데이터 마이닝의 전체 프로세스가 항상 정규적인 것이 아니라 도메인(domain)에 따라 틀리다는 것이다. 즉, 사용자는 자신이 적용하고자 하는 도메인에 가장 적합한 데이터 마이닝 프로세스를 시행착오를 통하여 찾아야 한다.



<그림 2> 기존 데이터 마이닝 도구의 문제점 및 개선방향

본 논문의 대상인 XM-Tool/Miner의 구현에선 이러한 기존 데이터 마이닝 도구의 문제점을 <그림 2>에서와 같이 분류하고 이에 대해 제시된 개선방안을 해결하고자 하였다.

### 4. 사용자지향의 데이터 마이닝 도구 XM-Tool/Miner 소개

XM-Tool/Miner은 일반 사용자, 기업의 마케팅 담당자, 기획 담당자를 대상으로 CRM 분야에 최적화된 데이터 마이닝 툴의 개발을

목적으로 대우정보시스템(주) 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.

#### 4.1 XM-Tool/Miner의 특징

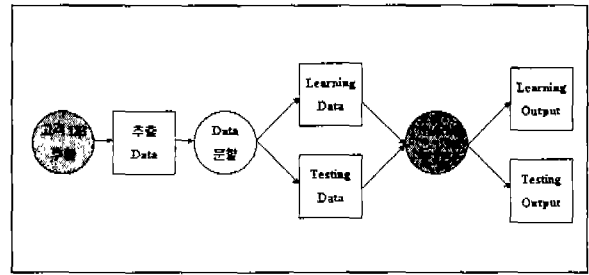
##### 플랫폼 독립적인 웹 어플리케이션

XM-Tool/Miner은 Java를 이용하여 플랫폼 독립성을 가지고 있으며, 서버와 클라이언트가 분리되어 인터넷상의 어디서나 접속이 가능하다. 더하여 CORBA 컴포넌트로 구성되어 있기 때문에 다중사용자를 지원한다. 즉, 사용자는 사내 또는 외부에 XM-Tool/Miner 서버를 두고 동시에 여러 부서에서 접속, CRM 활동을 전개할 수 있다.

##### 시나리오(Scenario)의 사용

XM-Tool/Miner은 CRM 관련 주요 도메인별 시나리오를 제공함으로써, 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.

다음 <그림 3>은 Neural Network을 사용하여 고객프로파일링(Customer Profiling)을 하는 시나리오의 예를 보이고 있다.



<그림 3> 시나리오 예시

시나리오는 데이터 마이닝 프로세스를 표현하는 노드(node)와 프로세스간의 데이터의 흐름을 표현하는 아크(arc)로 이루어져 있다.

##### 대용량 데이터베이스 지원

기존의 데이터 마이닝 도구들이 미리 준비된 데이터 마이닝용 데이터를 입력하는 반면에, XM-Tool/Miner은 직접 Legacy 데이터 베이스에 연결하여 대용량의 데이터를 실시간으로 처리할 수 있도록 구현되었다. 이는 실질적인 기업의 CRM 과정에 소요되는 데이터가 통계자료에 사용되는 것처럼 정제된 형태가 아니고 이를 준비하는 시간이 많이 소요되므로 실질적인 활용가치가 높은 기능이라 볼 수 있다.

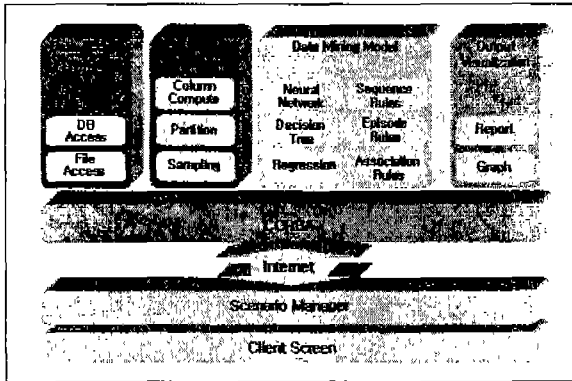
#### 4.2 구현 및 실행환경

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

#### 4.3 시스템 구성 및 주요기능

XM-Tool/Miner의 구조는 다음 <그림 4>에서 보는 것처럼 CORBA를 기반으로 각 서버 컴포넌트(Component)가 서버를 구성하고 있으며, 클라이언트는 시나리오 매니저(Scenario

Manager)가 담당하고, 서버와 클라이언트는 CORBA-IIOP 를 통해서 통신하게 된다.



<그림 4> XM-Tool/Miner 구조도

#### 4.3.1 Scenario Manager

시나리오 매니저(Scenario Manager)는 사용자가 시나리오를 생성하고 이를 유지, 관리해주는 역할을 한다. 해당 시나리오는 서버에 각 사용자 계정에 보관되며, 템플릿 기능을 이용하여 미리 정의된 시나리오를 상속받아서 새로운 시나리오를 구성할 수도 있다.

#### 4.3.2 Data Access/Transform

XM-Tool/Miner 의 Data Access 는 크게 파일과 데이터베이스로 나뉘어 진다.

- Data Access
  - File Extraction : 구분자(delimiter)로 구분된 파일
  - DB Extraction : MS SQL Server, Oracle DBMS
- Transform
  - Unique Value Filtering
  - Binary Mapping
  - Date 타입 변환

#### 4.3.3 Data Preparation

개발된 XM-Tool/Miner 의 Data Preparation

은 크게 세가지 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
- Column Computing
  - Computed Column, Summarization Column 생성 기능 제공

#### 4.3.4 Data Mining Algorithms

##### (1) Decision Tree

의사결정 트리는 데이터의 클러스터링(Clustering)이나 분류(Classification), 결과 예측을 위해 자주 사용되는 데이터 마이닝 알고리즘이다. XM-Tool/Miner 에선 Quinlan 의 C4.5[3]를 사용하였다.

##### (2) Association Rule

연관 규칙이란 한 항목 그룹과 다른 항목 그룹 사이에 존재하는 강한 연관성을 찾아내는 마이닝 기법이다.[4] 주로 검색알고리즘에 대한 연구가 많이 수행되고 있으며, XM-Tool/Miner 에선 Agrawal 의 Apriori 알고리즘[5]을 사용하였다.

##### (3) Neural Network

매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형(flexible nonlinear model)의 하나로, 이러한 신경망 모형은 고객의 신용평가, 불량거래의 색출, 의료진단예측, 우량고객의 선정, 타겟 마케팅의 여러 주제(DM, TM)등을 비롯한 여러 분야에 적용될 수가 있는데, 주

로 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) 알고리즘[6]을 사용하였다.

(4) Sequence Rule

트렌드를 식별해 내기 위해 일정시간 동안의 레코드를 분석하여 순서 패턴을 찾아내는 데이터 마이닝 기법이다.[7] Association Rule의 응용의 한가지이며, 역시 Apriori 알고리즘을 사용하였다.

(5) Episode Rule

시간의 순서에 따라 발생한 이벤트 리스트를 일정한 순서로 자주 발생하는 순서 패턴을 찾아내는 데이터 마이닝 기법이다.[8]

(6) Regression

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

4.3.5 Output Visualization

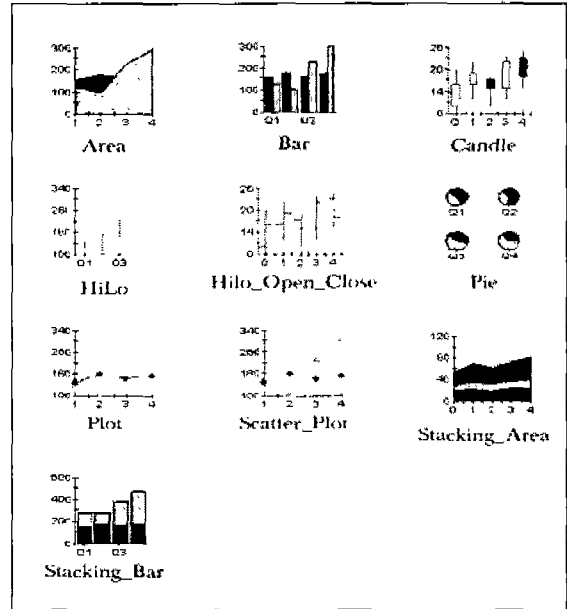
(1) Report

XM-Tool/Miner는 기본적으로 데이터 마이닝 데이터에 대해서는 테이블뷰와 기술통계치를 제공하며, 데이터 마이닝 알고리즘에 대해서는 생성된 물을 볼 수 있는 기능을 제공한다. 향후 사용자가 정의한 보고서를 생성하는 부분도 추가할 예정이다.

(2) Graph

XM-Tool/Miner는 데이터 마이닝 데이터에 대한 테이블뷰와 기술통계치 정보를 바

탕으로 10 가지의 차트(chart)를 통하여 보여주는 기능을 제공한다. 지원되는 차트는 다음 <그림 5>와 같다.

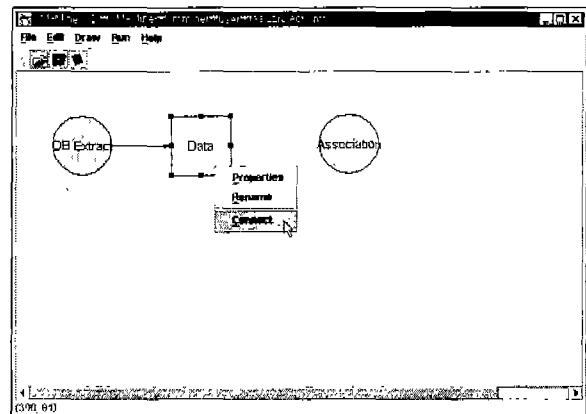


<그림 5> 지원되는 차트 예시

4.4 실행화면

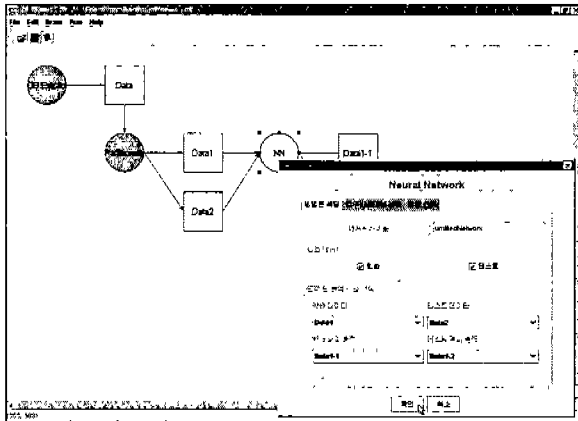
개발된 XM-Tool/Miner의 실행화면은 클라이언트 영역의 시나리오 매니저의 화면으로 구성된다.

시나리오는 생성하는 화면을 다음 <그림 6>에 보이고 있다.



<그림 6> 실행화면 예시 - 시나리오 편집

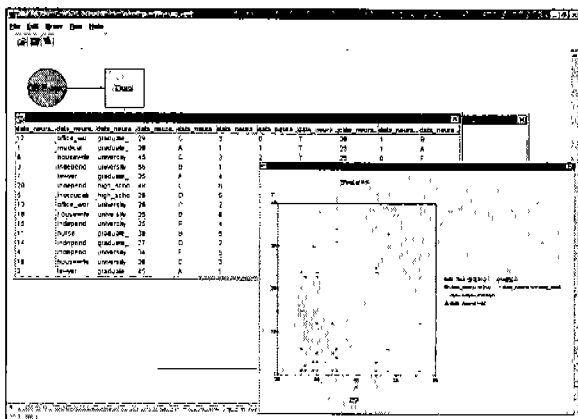
이어서 생성된 시나리오 상에서 각 노드에 대한 입력치(parameter)를 입력하는 작업을 수행해야 하는데, 그 예시가 <그림 7>이다.



<그림 7> 실행화면 예시 - 입력치(parameter) 설정

입력치가 설정되면 실행(run) 기능을 통해 데이터 마이닝을 수행하게 되고, 이의 결과를 확인하게 된다.

수행된 데이터 마이닝 결과를 확인 하는 화면은 <그림 8>과 같다.



<그림 8> 실행화면 예시 - 수행결과 확인

## 5. 결론

대표적인 데이터 마이닝 도구들은 기법의 다양성과 범용성에 초점을 맞추는 반면, 개별 데이터 마이닝 기법들에 대한 사용 용이성과 자동화에는 소홀한 실정이다. 이를 해결하고자 대우정보시스템(주)에서 개발된 XM-Tool/Miner은 일반 사용자 입장에서 용이하게 다룰 수 있는 기법의 자동화에 초점을 맞추고, 기업의 기능적 분야 또는 기업의 유형을 우선 결정하는 도메인 중심의 데이터 마이닝 도구를 구현하기 위하여 도메인별 최적화된 데이터마이닝 프로세스를 추상화한 시나리오(Scenario)를 사용하여 데이터 마이닝 도구를 개발하였다. 개발된 XM-Tool/Miner은 업무 분야의 특성에 따른 시나리오를 통하여 효과적인 데이터 마이닝 프로세스를 수행할 수 있으며, 의사결정 및 CRM 마케팅, 동향분석 및 예측 등에 유용한 정보를 추출하는 도구로 사용할 수 있을 것이다.

향후 대우정보시스템(주) 기술연구소는 XM-Tool/Miner를 최적화하고 수행성능을 향상시키며, 도메인별 적용을 통하여 최적의 시나리오 템플릿을 구축하여 제품화할 계획으로 있다.

## 참고문헌

- [1] 김정자, 이도현 “데이터 마이닝 기술 및 연구동향”, 정보과학회지, 16 권 9 호, pp. 7-14, 1998.
- [2] SAS Korea, “SAS/Enterprise Miner 를 이용한 데이터마이닝”, SAS Korea Ltd. 1998.
- [3] J. Quinlan, "C 4.5 : Programs for Machine Learning", Morgan Kaufman, 1988.

- [4] R. Agrawal, T. Imielinski and A. Swami, Mining Association Rules between Sets of Items in Large Database, Proc. ACM SIGMOD, pp. 207-216, 1993.
  
- [5] R. Agrawal and R. Srikant, Fast Algorithm for Mining Association Rules", Proc. VMDB, pp. 487-499, 1994.
  
- [6] D. Rumelhart, G. Hinton and R. Williams, "Learning Internal Representation by Error Propagation : in Parallel Distributed Processing", The MIT Press, 1986.
  
- [7] R. Agrawal and R. Srikant, "Mining Sequential Patterns", Proc. ICDE, 1995.
  
- [8] H. Mannila, H. Toronen and A. Verkano, "Discovery of Frequent Episodes in Event Sequences", Series of Publications C, University of Helsinki, 1997.