상대 지지도를 이용한 의미 있는 희소 항목에 대한 연관 규칙 탐사 기법

(Mining Association Rules on Significant Rare Data using Relative Support)

하 단 심 † 황 부 현 ††
(Danshim Ha) (Buhyun Hwang)

요 약 최근의 데이타베이스 연구 분야에서는 대규모의 데이타베이스에 저장된 데이타를 분석하여 데이타베이스에 존재하지만 쉽게 드러나지 않는 암시적인 지식을 탐사하는 기술인 데이타 마이닝이 각광받고 있다. 본 논문에서는 이러한 데이타 마이닝의 기법 중의 하나인 연관 규칙 탐사 기법을 연구하며 비록 데이타베이스에서는 희소하게 나타나는 데이타이지만 임의의 데이타와 높은 비율로 동시에 나타나는 의미있는 희소 데이타를 고려한 연관 규칙 탐사 기법을 제안한다. 또한 이러한 희소 항목의 탐사에 대하여 기존의 연관 규칙 탐사 알고리즘과 제안한 알고리즘의 성능을 비교하여 평가한다.

Abstract Recently, data mining, which is analyzing the stored data and discovering potential knowledge and information in large database, is a key research topic in database research area. In this paper, we study methods of discovering association rules which are one of data mining techniques. And we propose a technique of discovering association rules using the relative support to consider significant rare data which have the high relative support among some data. And we compare and evaluate existing methods and the proposed method of discovering association rules for discovering significant rare data.

1. 서 론

대규모의 데이타의 수집과 저장 기술의 발달, 데이타베이스 관리 시스템(DBMS)과 바코드 사용의 일반화와 같은 정보 기술을 토대로 축적된 대규모의 데이타베이스 구축과 데이타 웨어하우스의 구축은 데이타 마이닝(data mining)을 용이하게 하였다[1,2,3,4].

정보 추출 방법론인 데이타 마이닝은 의사 결정에 필요한 새롭고 유익한 정보와 지식을 얻기 위하여 축적된 대규모의 데이타들을 분석하여 기업의 이윤 추구에 도움이 될 수 있는 정보와 지식을 획득할 수 있는 기술로써 통신, 은행, 소매, 유통 등 다양한 산업 분야에서 장바구니 분석, 사기 행위 색출(fraud detection), 고객 분류 등에 널리 사용되고 있다[1,2,3,4].

데이타 마이닝의 기법에는 연관 규칙 탐사(association rule), 항목 분류(classification), 클러스터링(clustering), 요약(summarization), 순차 패턴 탐사(sequential pattern) 등이 있다[1,2,3]. 연관 규칙 탐사는 많은 데이타 마이닝 기법 중 가장 활발한 연구가 이루어지는 분야로써 데이타들의 동시 발생 가능성을 탐사하는 기법이다. 항목 분류 기법은 감독자에 의해 미리 구분이 된 데이타의 집합에 대하여 그 안에 숨겨진 공통 특성을 탐사하여 구분하는 기법이며 클러스터링은 각각의 데이타 사이의 유사성을 바탕으로 데이타들을 그루핑(grouping)하는 기법이다. 데이타 요약은 대규모의 데이타베이스의 레코드들을 적은 양의 일반화된 대표적인 표현으로 축약하는 응용에 사용되며, 순차 패턴 탐사는 경향 분석에 이용되는 기법으로써 일정 시간동안의 데이타 분석을 통해 시간에 따른 패턴 유형을 탐사하는 기법으로 이들 기법은 각각의 응용에 따라 적절

· 이 연구는 1999년도 한국과학재단 특정기초연구비 지원에 의한 결과임 (과제번호 : 1999-2-303-006-3)
† 정 회 원 : LG전자 연구원
dsha@lge.com
†† 종신회원 : 전남대학교 컴퓨터정보학부 교수
bhhwang@chonnam.chonnam.ac.kr
논문접수 : 2001년 1월 19일
심사완료 : 2001년 9월 15일

하게 이용될 수 있다[1,2,3].

본 논문에서는 데이타 마이닝의 많은 기법 중에서 데이타 사이의 연관성을 탐사하는 기법인 연관 규칙 탐사 기법에 대하여 연구하고, 비록 데이타베이스에서는 희소하게 나타나지만 데이타의 전체 발생 빈도수 면에서 특정 데이타와 상대적으로 높은 비율로 동시에 나타나는 연관성을 탐사할 수 있는 연관 규칙 탐사 기법을 제안한다.

기존의 연관 규칙 탐사 기법은 사용자가 지정한 지지도와 신뢰도라는 수치를 만족하는 데이타들 사이에서 발생할 수 있는 연관성을 탐사하였다[5,6,7]. 지지도(support)는 통계적인 중요성을 의미하는 수치이며 신뢰도(confidence)는 규칙의 강도를 나타내는 수치이다[5,6,7]. 기존의 연관 규칙 탐사 과정에서 적용되는 지지도는 전체 데이타베이스에 대해서 유일한 값으로 적용되며, 이러한 방법은 데이타베이스에 존재하는 각각의 데이타들은 모두 유사한 발생 빈도로 나타남을 전제한다[8]. 그러나 실제로는 데이타베이스를 구성하는 데이타들은 데이타베이스의 특성에 따라 상대적으로 빈번하게 나타나는 데이타도 존재하고 반대로 그렇지 못하는 데이타도 존재한다. 그리고 데이타베이스에서 드물게 나타나는 희소한 데이타에도 유용하게 사용될 수 있는 의미 있는 정보가 존재할 수 있다. 그러나 기존의 연관 규칙 탐사 기법은 데이타의 발생 빈도 형태 등을 고려하지 않고 일괄적으로 동일한 지지도를 이용하여 연관 규칙을 탐사하기 때문에 기존의 방법에서의 희소항목에 대한 규칙의 탐사는 중복되고 불필요한 규칙의 생성을 초래하게 된다.

본 논문에서는 특정 데이타들과 상대적으로 높은 비중을 가지고 동시에 나타나는 의미 있는 희소 데이타들에 대한 연관 규칙 탐사 기법의 연구를 목적으로 한다. 이 논문에서는 데이타베이스에 존재하는 데이타들의 상대적인 빈도수에 대한 비율인 상대 지지도를 이용하여, 데이타베이스에서는 희소한 데이타일지라도 의미 있는 연관성에 대한 정보를 포함하고 있는 연관 규칙을 탐사할 수 있는 방법을 제안하고 그에 대한 구현과 평가를 통하여 제안한 방법의 유용성을 보이고자 한다. 제안하는 희소 데이타를 고려한 연관 규칙 탐사 기법은 마케팅, 판매 전략 등의 여러 응용에 적용될 수 있다.

2. 관련 연구

2.1 연관 규칙의 정의

데이타 마이닝 기법 중 가장 활발히 연구되고 있는 분야는 데이타 사이의 연관성을 찾는 연관 규칙 탐사에 관한 분야로써 한 데이타 항목들의 그룹과 다른 데이타 항목들의 그룹 사이에 존재하는 연관성 탐사에 응용된다[5,7]. 연관 규칙의 표현은 데이타베이스의 데이타 항목 집합 I에 대한 부분 집합 X, Y ($X \subset I$, $Y \subset I$, $X \cap Y \neq \emptyset$)에 대한 연관 규칙이 존재할 때 $X \rightarrow Y(c\%)$로 표현하며, 그 규칙은 신뢰도가 c%임을 의미한다. X에 대한 지지도는 전체 트랜잭션에서 X를 접근하는 트랜잭션들의 비율로써 sup(X)로 표현하며 규칙 $X \rightarrow Y$의 신뢰도는 X를 포함하는 트랜잭션 중에서 X와 Y를 동시에 포함하는 트랜잭션의 비율로 conf($X \rightarrow Y$)로 표현한다[5]. 사용자는 규칙 탐사 시 지지도와 신뢰도에 대하여 탐사 기준의 임계값으로 최소 지지도(minsup)와 최소 신뢰도(minconf)를 지정하며, 규칙의 신뢰도와 지지도는 이 최소 지지도와 최소 신뢰도를 만족해야 채택된다.

즉, 규칙 R이 $sup(R) \geq minsup$ 이고 $conf(R) \geq minconf$의 조건을 만족할 때 R에 대하여 연관 규칙이 존재한다고 정의한다. 지지도를 만족하는 데이타 항목들은 빈발하다(large, frequent)라고 하며 빈발항목들에 의해 구성된 규칙들의 신뢰도를 검사하여 최종적인 연관 규칙을 탐사한다.

2.2 연관 규칙 탐사 기법

데이타 사이에 존재하는 연관성을 탐사하는 연관 규칙 탐사 기법에는 AIS[5], SETM[9], Apriori[6] 등의 기법이 연구되었다. 각 알고리즘 마다 특징은 다르나 기본적인 골격은 다음의 2단계로 구성된다[6].

- **1단계** : 사용자가 정의한 최소 지지도를 만족하는 데이타 항목의 집합을 탐사하는 단계이다. 이 단계에서는 각각의 데이타 항목에 대한 지지도를 계산하여 미리 정의된 최소 지지도를 만족하는 데이타 항목들을 추출해낸다.
- **2단계** : 연관 규칙을 생성하는 단계이다. 이 단계에서는 1단계에서 탐사된 데이타 집합을 이용하며, 데이타의 부분 집합에서 생성된 규칙 중 사용자가 정의한 최소 신뢰도를 만족하는 규칙들을 최종 규칙으로 탐사한다.

연관 규칙은 다양한 방법으로 응용되고 연구되고 있는데, 최근 연관 규칙 탐사는 복수의 지지도를 사용하는 연관 규칙의 탐사[8], 일정 주기 상에서 나타나는 순환적인 연관 규칙 탐사[10], 특정한 시간 구간에 대해서 강한 연관성을 가진 한시적 연관 규칙의 탐사[11], 공간 연관 규칙 탐사[12], 일정한 시간을 간격으로 동시에 나타나는 순차 패턴의 탐사[13] 등으로 응용되어 활발하게 연구되고 있다. 이 장에서는 연관 규칙 탐사 기법의 전형적인 방법인 Apriori 알고리즘[6]과 MSApriori[8] 알고리즘을 기술한다.

2.2.1 Apriori 알고리즘

연관 규칙 탐사 알고리즘의 가장 전형적인 방법인 Apriori 알고리즘은 [6]에서 소개되었으며 AprioriTid, AprioriHybrid [6] 등으로 확장되어 연구되었다.

Apriori 알고리즘은 후보항목을 구성하고, 구성된 후보항목집합에서 빈발항목집합을 탐사하는 과정으로 구성된다. Apriori 알고리즘은 후보항목 생성 시 모든 데이타베이스에서의 데이타 항목에 대한 생성이 아닌, 전 단계의 빈발항목집합을 대상으로 후보항목을 구성하는 점이 가장 큰 특징이다[6]. 빈발항목집합은 사용자가 정한 최소 지지도 minsup에 대하여 데이타 항목 집합 X의 지지도 sup(X)와 minsup와의 관계가 sup(X) ≥ minsup를 만족하는 항목들의 집합이며, 항목의 구성이 k개의 데이타로 구성된 k-빈발항목집합을 L_k로 표현한다. 후보항목집합은 빈발항목집합의 원소가 될 가능성이 있는 항목들로 구성된 집합으로, 빈발항목집합을 탐사하기 위해 사용되는 집합이다. k-후보항목집합은 C_k로 표현한다.

Apriori 알고리즘은 전 단계에서의 빈발항목집합을 이용하여 현재 단계의 후보항목집합을 구성하고 난 후, 데이타베이스의 스캔을 통해 지지도를 계산하여 최소 지지도와 비교한 후 현 단계의 빈발항목집합을 구성한다. Apriori 알고리즘의 단계의 진행은 후보항목을 구성하는 데이타의 개수의 순차적 증가에 따라 반복적으로 진행되므로 k 단계에서의 Apriori의 빈발항목 탐사는 k-1단계의 빈발항목집합으로부터 생성된 k-후보항목집합에 대하여 각각의 지지도를 계산한 후, 이들 중에서 지지도를 만족하는 항목의 탐사를 통해 이루어진다. Apriori는 더 이상의 후보항목을 생성할 수 없을 때까지 반복되어 실행된다.

후보항목집합의 구성은 전 단계의 빈발항목집합의 조인(join) 연산과 전지(pruning) 과정을 통해 구성된다. 조인 연산은 두 집합의 곱집합을 구하는 것과 같으며 전지 과정은 조인을 통해 생성된 후보항목집합의 부분 집합이 전 단계의 빈발항목집합의 원소가 아닌 경우, 그 항목을 삭제하는 과정이다. 그 이유는 전 단계에서 빈발하지 못하는 항목은 다음 단계에서도 빈발하지 못하기 때문이다. 전지 과정은 불필요한 후보항목의 수를 줄여 데이타베이스를 읽는 회수를 감소시키기 위하여 추가된 과정이다.

Apriori 알고리즘은 단계마다 후보항목을 생성하고 생성된 후보항목에 대하여 데이타베이스를 스캔하여 지지도를 계산하여 빈발항목을 추출하는 방법으로 요약할 수 있다.

2.2.2 MSApriori (Multiple Support Apriori) 알고리즘

[6]에서 제안한 Apriori 알고리즘은 최소 지지도와 최소 신뢰도를 만족하는 모든 데이타들을 찾는 방법으로 전체 데이타베이스에 대하여 유일한 최소 지지도가 사용된다. 즉 Apriori 알고리즘은 유일한 최소 지지도를 지정하여 규칙을 탐사하므로 데이타베이스에 존재하는 모든 데이타들은 유사한 빈도수를 가지고 있는 것으로 가정하고 연관 규칙을 탐사하는 방법이다. 그러나 실세계의 많은 응용에서 모든 데이타들이 유사한 발생 빈도수를 가지고 나타나는 경우보다는 각각 상이한 빈도수로 나타나는 경우가 일반적이다[8].

Apriori 알고리즘에서 최소 지지도를 큰 값으로 설정한 경우, 희소한 데이타들에 대한 규칙은 탐사할 수 없다. Apriori 알고리즘에서 상대적으로 희소한 빈도를 갖는 데이타들에 대한 연관 규칙을 탐사해야 하는 경우, 사용자는 최소 지지도를 낮게 설정하여야 한다. 그러나 최소 지지도를 작은 값으로 설정한다면 의미있는 희소한 데이타뿐만 아니라 그 최소 지지도를 만족하는 빈발하는 데이타들로 구성된 모든 규칙들이 부가적으로 탐사되어 작은 최소 지지도를 만족하는 모든 데이타들이 서로 연관성을 가지고 있는 것으로 탐사하는 희소 데이타 문제(rare data problem)가 발생한다[8]. [8]에서는 데이타들의 빈도 형태를 획일적으로 가정하는 Apriori 알고리즘을 개선하여 희소 데이타도 탐사할 수 있는 방법으로 MSApriori(Multiple Support Apriori) 알고리즘을 제안하고 있다.

MSApriori 알고리즘은 데이타들의 빈도수를 고려하기 위하여 데이타 각각의 최소 지지도인 MIS(minimum item support)를 사용한다. MIS는 각각의 데이타에 지정된 최소 지지도이며 데이타 i의 MIS는 MIS(i)로 표현하며 MIS를 설정하는 식은 아래와 같다[8].

$$\begin{aligned} MIS(i) &= MI(i), && \text{if } (MI(i) > LS) \\ &= LS, && \text{otherwise} \\ MI(i) &= \beta \times f(i) && (0 \le \beta \le 1) \end{aligned}$$

MSApriori는 Apriori의 최소 지지도처럼 사용자 임계값으로 정의된 LS(Least Support)를 만족하는 데이타를 탐사 대상으로 하며 β라는 값을 각각의 데이타의 빈도수 f(i)에 곱한 MI(i)와 LS를 비교하여 데이타들의 MIS를 구하게 된다.

MSApriori 알고리즘에서의 최소 지지도는 규칙을 구성하는 데이타 항목 중에서 가장 낮은 MIS값이 그 규칙의 최소 지지도로 이용된다. 즉 규칙 $i_1, i_2, \ldots, i_k \rightarrow i_{k+1}, \ldots, i_r$에 대한 최소 지지도는 $\min(MIS(i_1), MIS(i_2), \ldots, MIS(i_r))$이다. MSApriori 알고리즘은 MIS를 이용하기 때문에, 규칙을 구성하는 항목들이 빈발항목들로만

구성되었다면 그 규칙은 비교적 높은 최소 지지도에 의해 탐사되며, 반대로 희소한 데이타들이 포함된 경우에는 비교적 낮은 최소 지지도가 적용되어 규칙을 탐사한다. MSApriori는 이러한 방법으로 Apriori 알고리즘에서의 일괄적인 지지도 적용으로 생기는 문제점을 해결하고자 했다. MSApriori가 희소 데이타를 탐사하는 예는 예제 1과 같다.

예제 1: MSApriori 탐사 과정

데이타베이스에 bread, shoes, clothes라는 데이타들이 존재하고 공식에 의해 지정된 이들의 MIS는 아래와 같다면,

MIS(bread) = 2% MIS(shoes) = 0.1%

MIS(clothes) = 0.2%

지지도가 0.15%인 clothes → bread인 규칙에 적용되는 최소 지지도는 min(MIS(clothes), MIS(bread)) = 0.2이므로 이 규칙은 최소 지지도를 만족하지 못하는 규칙이다. 그러나 같은 지지도 0.15%인 clothes → shoes의 경우, 이 규칙의 최소 지지도는 min(MIS(clothes), MIS(shoes)) = 0.1이므로 최소 지지도를 만족하는 규칙임을 알 수 있다.□

MSApriori 알고리즘은 각각의 데이타에 대한 빈도형태를 고려하여 데이타마다 다른 지지도인 MIS를 사용하여 연관 규칙을 탐사하는 방법이며 이러한 방법으로 Apriori 알고리즘에서의 일괄적인 지지도 적용으로 생기는 문제점을 보완할 수 있다. 그러나 MSApriori 알고리즘은 모든 데이타에 대하여 MIS를 지정하는 과정이 필요하며, MIS 설정 시 β의 값에 따라 탐사되는 규칙의 차이가 큰 문제가 있다. MSApriori는 데이타의 빈도수를 고려하기는 했지만 실질적인 탐사의 기준은 데이타 각각의 빈도수보다는 β의 설정에 의해 크게 좌우되며 적절한 β의 설정 기준이 연구되어야 한다.

3. 의미 있는 희소 데이타에 대한 연관 규칙 탐사 기법

3.1 문제 정의

실세계의 데이타들은 상대적으로 빈발하게 나타나는 데이타가 존재하는가 하면 반대로 상대적으로 드물게 나타나는 데이타들도 존재한다. 기존의 연관 규칙 탐사 기법은 발생 빈도수가 기준이 되어 규칙을 탐사하지만, 발생 빈도수만으로 데이타가 구성하는 규칙의 중요성을 구분하는데 부족한 점이 있다. 빈도수는 적지만 발생 빈도 중 높은 비율로 특정 데이타와 동시에 나타나는 특수한 경우와 같이 지지도는 작지만 신뢰도가 높은 경우, 규칙의 특수한 특성으로 탐사 가치가 있지만 기존의 연관 규칙에서는 지지도를 만족하지 못하면 탐사될 수 없다.

기존의 Apriori 알고리즘에서 이러한 성격의 희소한 데이타들이 포함된 연관 규칙을 탐사하고자 하는 경우, 사용자는 최소 지지도 값을 낮게 설정해야 한다. 그러나 낮은 최소 지지도는 최소 지지도 값을 만족하는 모든 빈발하는 데이타들이 빈발항목집합의 원소가 되어 중복되고 쓸모 없는 규칙을 대량으로 발생시키는 문제가 발생한다[8]. 즉 Apriori에서 희소항목 중에서도 높은 신뢰도로 동시에 나타나는 항목을 찾기 위해서는 최소 지지도를 낮추어야 하며, 낮은 최소 지지도는 데이타들 사이에 강한 연관성이 존재하지 않아도 연관성이 있다고 탐사하는 문제가 발생한다.

또 다른 연관 규칙 탐사 기법으로 MSApriori의 경우, 데이타 각각의 빈도수를 고려하여 최소 지지도를 설정하여 희소항목을 탐사하고자 하지만, MIS는 데이타의 빈도수보다 β의 설정이 탐사되는 규칙을 좌우하나, 희소 데이타 중 상대적으로 높은 지지도로 동시에 나타나는 항목들간의 연관성 탐사를 위한 적절한 β의 설정이 필요하다는 단점이 있다.

이 장에서는 희소 데이타 항목 중 특정 데이타와 높은 지지도로 동시에 나타나는 연관성을 탐사하는 척도로써 상대 지지도를 이용하여 연관 규칙을 탐사하는 방법인 RSAA(Relative Support Apriori Algorithm)을 제안한다. 또한 제안한 RSAA의 의미 있는 희소 항목 탐사의 경우, 기존의 Apriori, MSApriori와 각각 비교하여 성능 면에서 우수함을 보이고자 한다.

3.2 RSAA

3.2.1 상대 지지도

이 장에서는 상대 지지도를 이용하여 희소 데이타 중에서도 임의의 데이타와 높은 신뢰도를 갖는 희소 데이타를 포함한 연관 규칙을 탐사하는 RSAA(Relative Support Apriori Algorithm) 방법을 제안한다. 의미 있는 희소 데이타의 정의는 정의 1과 같다.

정의 1. 의미 있는 희소 데이타

의미 있는 희소 데이타란 데이타베이스에서 발생 빈도수가 최소 지지도를 만족하지 못하지만 데이타의 발생 빈도수 중 높은 비율로 특정 데이타들과 연관되어 나타나는 데이타이다.□

Apriori와 같은 기존의 연관 규칙 탐사 알고리즘에서 의미 있는 희소 데이타를 탐사하기 위해서는 낮은 최소 지지도를 설정하여 이를 만족하는 빈발항목집합을 생성한 후, 이 빈발항목집합으로 생성될 수 있는 모든 규칙에 대하여 신뢰도를 적용하여야 한다. 실제로 빈발항목 탐사과정에서 이러한 의미 있는 희소 데이타를 탐사하

지 못하는 예는 아래와 같다.

예제 2: 의미 있는 희소 데이타를 탐사하지 못하는 경우

데이타 a, b, c 가 데이타베이스에 존재하고 a, b, c 각각은 데이타베이스에서 a = 20%, b = 40%, c = 25%의 지지도로 나타나고, 사용자가 최소 지지도를 30%라고 설정했다면 a 와 c 는 최소 지지도를 만족하지 못하므로 빈발항목집합의 원소가 될 수 없다. 그러나 만약 {a, b, c}가 18%의 지지도를 갖는다면 a는 발생빈도수 중 90%가 b, c와 동시에 나타남을 알 수 있다. 그러나 {a ,b, c}는 기존의 방법에서는 30%인 최소 지지도를 만족하지 못하므로 탐사될 수 없다.□

기존의 방법으로 탐사하기 힘든 이러한 의미 있는 희소 데이타를 탐사하기 위하여 RSAA에서는 2개의 최소 지지도가 설정되어 규칙탐사에 이용된다. 이 두 지지도는 각각 1차 지지도, 2차 지지도라 정의되며 각각의 정의는 정의 2와 같다.

정의 2. 지지도

- 1차 지지도: 빈발항목 탐사과정에 사용되는 사용자가 정의한 지지도의 임계값
- 2차 지지도: 희소항목 탐사과정에 사용되는 사용자가 정의한 지지도의 임계값. □

1차 지지도와 2차 지지도의 설정은 1차 지지도〉 2차 지지도를 만족하도록 설정해야 한다. 그렇지 않을 경우, 중복된 규칙이 생성되거나 희소항목에 대한 탐사를 하지 못하기 때문이다.

RSAA는 지지도 이외에 또 다른 척도로써 데이타 사이의 상대적인 빈도수를 고려하여 연관 규칙을 탐사할 수 있는 상대 지지도를 사용한다. 상대 지지도는 1차 지지도는 만족하지는 못하지만 2차 지지도를 만족하는 희소항목에 대한 임의의 다른 데이타 사이의 상대적인 지지도이며, 상대 지지도를 이용해 의미 있는 희소한 데이타들을 판별한다. 상대 지지도의 정의는 정의 3과 같다.

정의 3. 상대 지지도 (Relative Support)

데이타베이스가 데이타의 집합 $I = \{i_1, i_2, \ldots, i_m\}$ 와 같이 구성되고 데이타 항목 i의 지지도가 sup(i)로 표현된다면, 데이타 사이의 상대적인 지지도를 의미하는 상대 지지도 $Rsup(i_1, i_2, \ldots, i_k)$의 후보항목집합 $\{i_1, i_2, \ldots, i_k\}$에서의 상대지지도는

$$Rsup(i_1, i_2, \ldots, i_k) = \max(sup(i_1, i_2, \ldots, i_k)/sup(i_1), sup(i_1, i_2, \ldots, i_k)/sup(i_2), \ldots, sup(i_1, i_2, \ldots, i_k)/sup(i_k))$$

이다. □

상대 지지도는 $0 \leq Rsup \leq 1$인 값이며 데이타 항목을 구성하는 각각의 항목들이 후보항목과 이루는 신뢰도를 비교하여 이 중 가장 큰 값을 상대 지지도로 지정한다. 이는 후보항목을 구성하는 데이타 항목 $i_1, i_2, \ldots, i_k$ 각각이 후보항목$\{i_1, i_2, \ldots, i_k\}$에 대하여 상대적으로 얼마만큼의 비율의 지지도로 나타나는지에 대한 척도이며 사용자는 탐사할 데이타의 상대 지지도의 임계값인 최소 상대 지지도(minimum relative support)를 입력하여 규칙을 탐사한다. 최소 상대 지지도 값이 높을수록 사용자는 동시에 나타나는 비율이 큰 항목을 선택함을 의미한다. 그리고 2차 지지도를 만족하는 항목 집합에서 최소 상대 지지도를 만족하는 항목의 집합을 준빈발항목집합이라고 정의한다.

RSAA에서의 상대 지지도의 사용 과정은 그림 1과 같다. 알고리즘의 각 단계는 더 이상의 후보항목을 생성할 수 없을 때까지 반복된다.

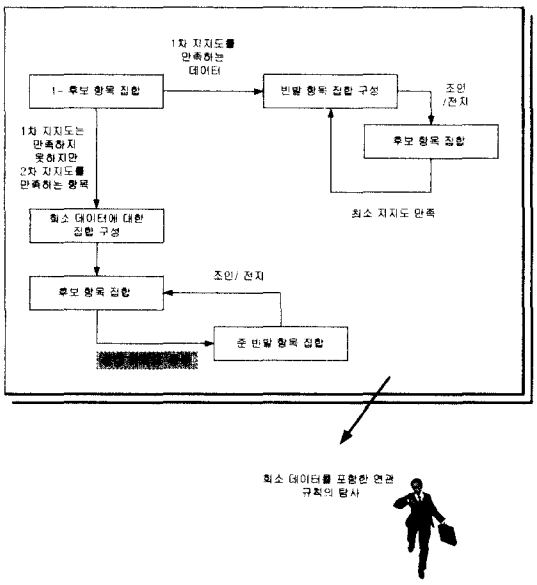

그림 1 상대 지지도의 사용 과정

상대 지지도를 정의하여 연관 규칙을 탐사한다면 데이타베이스를 구성하는 항목들의 상이한 빈도수가 반영된 연관 규칙을 탐사할 수 있다. 예를 들어, 슈퍼마켓에서 소비자들의 food processor, cooking pan의 구매 횟수는 bread, milk의 구매 횟수보다 적지만 수익 면에서 더 많은 이익을 남기는 항목이다. 기존의 유일한 지지도를 사용하는 방식을 이용하여 food processor나 cooking pan에 관계된 연관성을 탐사하기 위해서는 낮은 수치의 지

지도를 사용해야 하며, 이 경우 낮은 지지도를 만족하는 빈발 항목들이 구성하는 수많은 불필요한 규칙이 생성되게 된다. 상대 지지도를 사용하면 빈발 항목들의 불필요한 규칙의 생성을 막으면서도 희소 항목들이 구성하는 의미 있는 연관 규칙을 탐사할 수 있다.

3.2.2 RSAA의 후보항목 생성 방법

RSAA의 후보항목 생성 방법은 희소 데이타를 포함한 후보항목을 구성할 수 있어야 한다. RSAA에서 후보항목은 2부분으로 나누어 구성이 된다.

RSAA에서의 후보항목 구성은 1차 지지도를 만족하는 빈발항목과, 1차 지지도를 만족하지는 않지만 2차 지지도를 만족하는 희소항목들로 나누어서 구성이 된다. 1차 지지도를 만족하는 원소들은 기존의 Apriori의 방식과 동일하다.

후보항목의 구성은 데이타베이스의 모든 데이타에 대해 지지도를 카운트하여 1차 지지도를 만족하는 빈발항목에 대한 1-후보항목집합인 C_1과 1차 지지도를 만족하지 못하지만 2차 지지도를 만족하는 희소 데이타 항목 집합의 후보항목집합 NC_1으로 각각 분리하여 생성한다.

분리된 희소 데이타 항목에 대한 후보항목집합의 구성은 NC_k(Not large Candidateset), NLC_k(Candidateset generated by joining Not large and Large item)의 두 그룹으로 나누어 각각 생성된다. k는 각 항목을 구성하는 데이타의 개수이다. NC_k는 2차 지지도를 만족하는 데이타들로만 구성된 k-후보항목집합이며 NLC_k는 1차 지지도를 만족하는 1-후보항목과 2차 지지도를 만족하는 데이타로 생성된 k-후보항목집합이다. 2-후보항목집합은 NC_2와 NLC_2로 표기하며, 각각은 다른 조인 방법을 거쳐 생성된다. NC_2는 NL_1(Not Large itemset)의 조인으로 생성되며 NLC_2는 1-빈발항목집합 L_1(Large itemset)과 1-준빈발항목집합 NL_1의 조인으로 생성한다. L_1은 첫 번째 데이타베이스 스캔 시에 얻어지는 1차 지지도를 만족하는 데이타의 집합이며, NL_1은 1차 지지도를 만족하지 못하지만 2차 지지도를 만족하는 데이타의 집합이다. NC_k와 NLC_k는 각각 전 단계의 NL_{k-1}, NLL_{k-1}의 조인을 통해 생성된다. NL_k(semi-Large itemset generated by joining Not frequent item)는 NC_k에 대하여 2차 지지도와 상대 지지도를 평가하여 생성되고, NLL_k(semi-Large itemset generated by joining Large itemset and not large itemset)는 NLC_k에 대하여 같은 방법으로 생성된다. k-준빈발항목집합은 NL_k와 NLL_k를 의미한다. 알고리즘 1은 RSAA의 후보항목 생성 알고리즘이다.

알고리즘 1 희소 데이타에 대한 후보항목 구성 알고리즘(rsaa-gen)

```
if (k=2) then
    insert into NC2
        select p.item1 , q.item1 from NL 1p, NL1 q
    insert into NLC2
        select p.item1 , q.item1 from NL 1p,L1 q
else
    insert into NCk

        select p.item1 , p.item2 , ... , p.itemk-1 , q.itemk-1
        from NLk-1 p , NLk-1 q
        where p.item1 = q.item1, p.item2 = q.item2, ... ,
         p.itemk-1 = q.itemk-1, p.itemk-1 < q.itemk-1

    insert into NLCk
        select p.item1 , p.item2 , ... , p.itemk-1 , q.item1
        from NLLk-1 p , NLLk-1 q
        where p.item1 = q.item1, p.item2 = q.item2, ... ,
         p.itemk-1 = q.itemk-1, p.itemk-1 < q.itemk-1
```

후보항목에서의 전지기법은 [6]에서의 전지기법과 동일하다. 즉, 이전 단계의 빈발항목집합의 원소가 아닌 후보항목은 다음 단계의 빈발항목집합의 원소가 될 수 없으므로 삭제한다[6]

3.2.3 RSAA의 자료 구조

RSAA에서는 상대 지지도를 계산하기 위하여 매 단계마다 준 빈발항목 탐사에 이용되는 모든 데이타들의 지지도를 필요로 한다. RSAA에서는 최초의 데이타베이스 스캔 시 얻어낸 데이타 항목에 대한 지지도를 유지하며, 1차 지지도를 만족하는 항목들로 구성된 L_1과 1차 지지도는 만족하지 못하나 2차 지지도를 만족하는 항목들의 집합인 NL_1를 각각 유지한다. 각각의 구조는 그림 2와 같다.

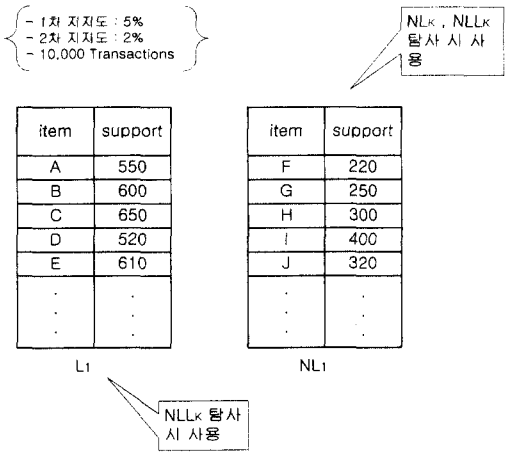

그림 2 L1과 NL1의 구조

3.2.4 RSAA의 실행 예

이 장에서는 RSAA의 실행 예를 설명한다. 예에 사용될 트랜잭션은 표1과 같고 각각의 아이템들의 지지도는 표 2와 같다. 이 예에서는 1차 지지도는 40%, 2차 지지도는 20%, 최소 상대 지지도는 0.7로 설정하여 의미 있는 희소 항목을 탐사하고자 한다.

표 1 데이타베이스의 트랜잭션의 예

Transaction ID	항목	Transaction ID	항목
1	1 2 3	6	1 2
2	3 4	7	3 4 6 7
3	1 2 3 4 5 6 7	8	1 2
4	6 7	9	1 3 4 5 6
5	3 4 5 6	10	1 2

표 2 데이타들의 지지도

item	support	item	support	item	support
1	6	4	5	6	5
2	5	**5**	**3**	**7**	**3**
3	6				

표 2에서 {5, 7}은 1차 지지도를 만족하지 못하지만 2차 지지도를 만족하는 희소 데이타이다. 이 예에서는 5와 같은 희소 데이타가 3, 4, 6과 100%의 높은 비율로 동시에 나타나는 데이타를 탐사함을 보임으로써, 5와 관련된 의미 있는 연관 규칙을 탐사함을 보이고자 한다.

RSAA의 실행 결과, 빈발항목집합은 {1, 2, 3, 4, 6}이며 희소항목집합은 {5, 7} 이다. 희소항목에 대한 2-후보항목집합인 NC_2, NLC_2는 각각 {{5, 7}}, {{5, 1},{5, 2},{5, 3},{5, 4},{5, 6},{7, 1},{7, 2},{7, 3},{7, 4},{7, 6}}이다. RSAA는 이 NC_2와 NLC_2를 이용하여 2-준빈발항목집합 NL_2와 NLL_2를 탐사하기 위하여, NC_2와 NLC_2의 항목에 대하여 상대 지지도를 계산한다. 상대 지지도를 계산한 결과, 2-준빈발항목집합은 {{3, 5}, {4, 5}, {5, 6}, {6, 7}}이며 이 들을 이용하여 3-후보항목집합인 NLC_3를 구성한다. NLC_3는 {{3, 4, 5}, {3, 5, 6}, {4, 5, 6}, {5, 6, 7}}이며 각각의 상대 지지도를 계산하여 NLL_3를 탐사하며 탐사된 3-준빈발항목집합은 {{3, 4, 5}, {3, 5, 6}, {4, 5, 6}}이다. 이 과정은 더 이상의 후보항목을 생성할 수 없을 때까지 반복되어 실행된다.

이 예를 통하여 비록 5는 1차 지지도를 만족하지 못하는 희소 데이타이지만 {3, 4},{3, 6},{4, 6}과 항상 동시에 일어남을 알 수 있으며, 제안하는 RSAA는 빈발항목에 포함되지 못하더라도 상대적인 빈도수 측면에서 의미 있는 희소 데이타에 대한 탐사를 할 수 있음을 알 수 있다.

3.2.5 알고리즘

이 장에서는 RSAA의 알고리즘을 소개한다. 알고리즘을 통해 L_k, NL_k NLL_k,를 구하는 것이 최종 목적이다. L_k는 1차 지지도를 만족하는 원소들만으로 후보항목을 구성하여 탐사된 빈발항목집합이며, NL_k는 2차 지지도를 만족하는 원소들로만 후보항목을 구성하여 2차 지지도와 상대 지지도를 만족하는 빈발항목집합을 구성한 것이다. NLL_k는 1차 지지도를 만족하는 원소와 2차 지지도를 만족하는 원소로 후보항목을 구성하여 2차 지지도와 상대 지지도를 만족하는 집합으로 빈발항목집합을 구성한 것이다. 이 알고리즘에서는 L_k를 구성하는 방식은 Apriori의 방식과 동일하므로 탐사과정을 생략하였다. NL_k와 NLL_k는 2-후보항목의 생성만 다를 뿐 나머지 과정은 동일하므로 NL_k의 탐사과정만 소개한다. 알고리즘에 사용된 함수와 변수들은 표3과 같으며 RSAA의 알고리즘은 알고리즘2와 같다.

표 3 RSAA에서 사용되는 자료 구조와 함수들

$support_1$	1차 지지도
$support_2$	2차 지지도
minRsup	최소 상대 지지도
Rsup	데이타 항목의 상대 지지도
k-itemset	k개의 아이템으로 구성된 집합
NC_k , NLC_k	희소항목에 대한 k-후보항목집합
NL_k , NLL_k	k-준빈발항목집합
rsaa-gen	준빈발항목집합에 대한 후보항목 생성 함수
subset	트랜잭션에 후보항목이 존재하는지 검사하는 함수

알고리즘 2 RSAA알고리즘

```
D : Database(a set of transactions)
I = { i1, i2, ... , ik } : a set of data items

for (all ik ∈ I)
    if (ik.support ≥ support1) then ik ∈ L1
    else if (ik.support ≥ support2) then ik ∈ NL1
end

for Li
    do Apriori Algorithm
end
```

```
if (k=2)
    NC2 = rsaa-gen(NL1, NL1 );
    NLC2 = rsaa-gen(NL1 , L1);
end

for (k=3; NLk-1 ≠ Ø or NLLk-1 ≠Ø;k++) do
    NCk = rsaa-gen(NLk-1, NLk-1 );

    for all transaction t ∈ D do
        NCt = subset(NCk , t);

        for all candidates nc ∈ NCt do
            nc.count++;
        end
        if nc.count ≥ supportc then
            each itemik in nc
                nc.Rsup = max(sup(i1,i2, ... ,ik)/sup(i1),
                              sup(i1,i2, ... ,ik)/sup(i2), ...,
                          sup(i1,i2, ... , ik)/sup(ik) )
        end
        if nc.Rsup ≥ minRsup then
            NLk = {nc | nc ∈ NCt and nc.Rsup ≥ minRsup}
        end
    end
Answer = ∪kNLk ∪ ∪kNLLk
end
```

4. 성능평가

이 장에서는 RSAA를 기존의 연관 규칙 탐사 방법인 Apriori와 MSApriori를 각각 RSAA에 대하여 비교하여 평가한다. 성능평가를 통해 제안한 RSAA가 의미 있는 희소 항목을 탐사 시에 기존의 두 방법보다 효과적으로 수행함을 보이고자 한다.

4.1 RSAA와 Apriori의 비교

RSAA에 대한 성능평가를 Apriori와 비교하여 평가하였다. 평가 환경은 Windows ME O/S 환경과 CPU Pentium Ⅲ 450, 128MB의 PC에서 수행하였으며 랜덤하게 생성된 10,000건의 데이타를 사용하여 RSAA와 Apriori의 성능평가를 수행하였다.

Apriori는 다음과 같은 방식으로 RSAA와 성능을 비교한다. Apriori에서 희소항목을 탐사하기 위하여 최소 지지도를 RSAA의 2차 지지도와 같은 값으로 설정하여 Apriori를 구동시킨 결과와 RSAA를 비교하였다. Apriori의 최소 지지도를 RSAA에서의 2차 지지도로 설정한 이유는 RSAA의 2차 지지도를 Apriori의 최소 지지도로 설정하고, Apriori의 신뢰도를 RSAA의 최소 상대 지지도로 설정하면 두 알고리즘은 같은 결과를 유도할 수 있기 때문이다.

이 성능평가에서는 Apriori에서 희소항목을 찾기 위해 지지도를 낮추어서 수행한 시간과 RSAA를 통한 희소항목의 탐사를 수행한 시간을 비교했고, 더불어 RSAA에서 1차 지지도와 2차 지지도의 편차가 큰 경우의 RSAA와 Apriori를 비교하여 평가했다. 실험에 사용된 지지도는 1차 지지도는 3%, 4%, 5%로 지정하였고 2차 지지도는 2%, 3%, 4%로 설정하였으며 최소 상대 지지도는 0.5로 설정하였다. 수행 시간의 비교 결과는 표 4와 같으며 비교 결과는 RSAA가 수행 시간 면에서 개선됨을 알 수 있다.

표 4 RSAA와 Apriori의 수행 시간 비교

Apriori		RSAA			
minsup	time	1st minsup	2nd minsup	minRsup	time
2 %	95400	3%	**2%**	0.5	50640
3 %	46190	4%	**3%**	0.5	32960
4 %	29940	5%	**4%**	0.5	24830

그림 3은 표4의 RSAA와 Apriori의 수행 시간의 비교를 도표로 표현한 것이다. 희소항목 탐사를 위하여 최소 지지도를 낮춘 경우, Apriori에서는 낮은 지지도를 만족하는 모든 후보항목이 생성되므로 많은 수행시간이 소요되었고, RSAA에서는 1차 지지도로 빈발항목과 희소항목을 구분한 후 희소항목은 2차 지지도와 최소 상대 지지도를 만족하는 항목들을 추출해내므로 불필요한 희소 데이타가 포함된 규칙을 많이 제거할 수 있었다.

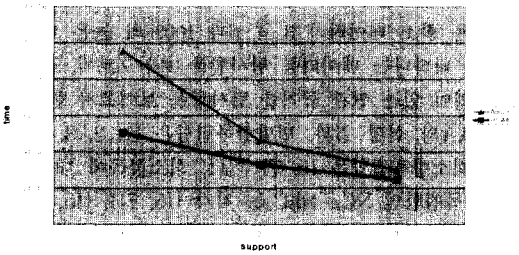

그림 3 RSAA와 Apriori의 수행 시간의 비교

표 5 지지도의 편차가 큰 경우의 RSAA와 Apriori의 수행 시간 비교

Apriori		RSAA			
minsup	time	1st minsup	2nd minsup	minRsup	time
2 %	95400	10%	**2%**	0.5	28010
3 %	46190	10%	**3%**	0.5	24210
4 %	29940	10%	**4%**	0.5	21850

또한 RSAA에서 1차 지지도와 2차 지지도의 편차가 큰 경우에 대해서도 평가하였다. 실험 결과는 표 5와 그림4에서 보는바와 같이 수행 시간 면에서 RSAA가 우수함을 알 수 있었다.

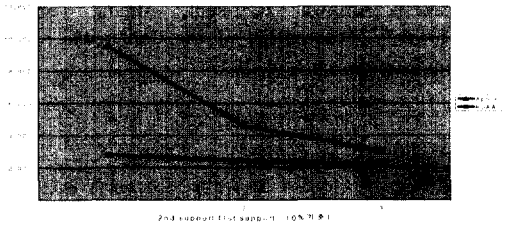

그림 4 지지도 편차가 큰 경우의 RSAA와 Apriori의 수행 시간 비교

성능 분석을 통하여 RSAA는 희소한 데이타에 대한 연관 규칙을 탐사하고자 할 때, 기존의 Apriori에서 지지도를 낮추어 희소항목의 연관 규칙을 탐사하는 것보다 개선된 수행시간을 보임을 알 수 있었다. 또한 1차 지지도와 2차 지지도의 편차가 큰 경우, RSAA가 우수한 성능을 보였다. 그리고, RSAA의 사용은 희소항목이 이루는 규칙 중에서도 강한 상관 관계가 있는 것들만 탐사함으로써 불필요한 규칙의 생성을 줄일 수 있었다.

4.2 RSAA와 MSApriori의 비교

두 번째 성능평가는 RSAA와 MSApriori를 비교하여 수행하였다. RSAA와 MSApriori는 모두 희소항목 탐사를 고려한 연관 규칙 탐사 기법이지만 두 방법은 기법면에서 상이하다. MSApriori는 후보항목을 구성하는 데이타 각각의 MIS값을 비교한 후, 그 중 가장 작은 MIS를 해당 후보항목에 적용할 최소 지지도로 설정하여 후보항목의 실제 지지도와 비교하는 방식이다. 반면 RSAA는 후보항목을 구성하는 데이타에 대한 상대 지지도를 구하여 그 상대 지지도가 최소 상대 지지도를 만족하는지를 비교하는 방식으로써 RSAA와 MSApriori 모두 희소 항목의 탐사를 고려하였다. RSAA는 희소 항목 중에서도 의미 있는 희소 항목을 탐사하고자 하는 방법이지만 MSApriori는 단순히 희소 항목도 탐사하기 위한 방법으로 두 방법에서 탐사되는 희소 항목과 규칙은 성격이 다름을 알 수 있다.

이 장에서는 RSAA와 MSApriori에 대한 비교의 수행은 MSApriori에서도 RSAA에서와 같이 희소 항목 중에서도 상대적으로 특정 데이타들과 높은 비율로 동시에 나타나는 항목을 탐사할 수 있도록 매개 변수들을 조정하여 그 성능을 평가하였다.

MSApriori는 MIS 계산 시에 데이타 각각의 빈도수와 곱해지는 β값에 따라서 탐사되는 항목에 큰 차이가 있었으며 MSApriori의 결과가 RSAA의 결과를 포함하도록 β값을 조정하였다. 이 실험 평가에서는 10,000건의 트랜잭션을 대상으로 RSAA의 결과를 포함하도록 조정된 β값은 0.25였으며, RSAA의 2차 지지도와 MSApriori의 최소 지지도인 LS(Least Support)를 1.2%로 동일하게 설정하여 RSAA에서 탐사되는 특수한 항목을 MSApriori에서도 탐사할 수 있도록 하였다. RSAA의 1차 지지도는 2%, 최소 상대 지지도는 0.5로 설정하였으며, RSAA와 MSApriori를 탐사된 항목 수에 대하여 비교한 결과는 표6과 같다.

표 6 RSAA와 MSApriori의 비교

지지도	120 - 199	200 - 299	300 - 399	400 - 499	499 이상	수행시간
RSAA	15	1	27	6	3	70350
MSApriori	85	14	0	2	0	142040

이 실험 평가의 예에서는, 120-199 구간에 있는 데이타들이 희소 데이타 탐사 대상이다. 즉, RSAA에서 이 구간에 존재하는 15개의 의미 있는 희소항목을 MSApriori에서 탐사하기 위하여 β값을 조정한 결과, 0.25 이하의 낮은 값에서부터 탐사되었다. 또한 이 실험에서는 MSApriori는 탐사되는 항목은 85개로 더 많지만, 그 중 상대 지지도가 50%를 만족하는 의미 있는 희소항목은 15개에 불과하며, 15개의 의미 있는 희소항목을 탐사하기 위해서는 나머지 탐사된 많은 항목 중에서 추려내야 한다. 또한, MSApriori는 높은 지지도를 갖는 항목이라 할 지라도 β의 설정에 따라서 탐사되지 못하는 경우가 존재하며, 낮은 β의 설정에도 불구하고 작은 지지도를 갖는 항목들만이 주로 탐사되는 점이 관찰되었다.

이 실험 평가를 통하여 MSApriori는 RSAA에서와 같은 희소항목 중에서도 의미 있는 희소 데이타에 대한 탐사는 RSAA에서 탐사하고자 하는 의미 있는 희소항목 뿐만 아니라, 낮은 상대 지지도를 갖는 희소항목들도 같이 탐사되거나 데이타의 빈도수보다는 β값에 따라 탐사 대상의 변동이 심함을 알 수 있다. 따라서 단순한 희소항목이 아닌, 의미 있는 희소 데이타를 탐사하는 경우, MSApriori보다 RSAA가 불필요한 규칙을 생성하지 않고도 원하는 희소 데이타들을 탐사할 수 있음을 알 수 있다.

5. 결 론

최근 데이타 마이닝은 데이타베이스 연구 분야에서 각광받고 있는 분야로써 특히 데이타 사이의 연관성을 탐사하는 연관 규칙에 관한 연구가 활발하게 이루어지고 있다.

본 논문에서는 희소한 데이타의 발생 빈도 중 특정 데이타들과 상대적으로 높은 비율로 동시에 나타나는 강한 연관성을 탐사할 수 있도록 상대 지지도라는 새로운 척도를 이용하여 희소 데이타를 고려하여 연관 규칙을 탐사할 수 있는 기법을 제안하고 설계하였다. 그리고 의미 있는 희소 항목을 탐사하고자 하는 경우에 대하여 기존의 두 연관 규칙 탐사 알고리즘과 제안한 알고리즘의 성능을 비교하여 효율적으로 의미 있는 희소 항목을 탐사함을 보였다.

향후 연구 방향으로는 알고리즘의 효율적인 개선과 보다 세밀한 후보항목의 전지 기법에 대한 연구가 필요하다.

참 고 문 헌

[1] Michael J.A.Berry, Gordon Linoff, *Data Mining Techniques-For Marketing, Sales, and Customer Support*, Wiley Computer Publishing, 1997

[2] Pieter Adrians, Dolf Zantige, *Data Mining*, Addison Wesley, 1996

[3] 이도헌, "데이타 마이닝 : 개념 및 연구동향", 데이타베이스연구회지 제 13권, pp.122-137, 1998

[4] 김종훈, "데이타 마이닝에 기초한 DB 마케팅 분석 방법과 사례", 월간 경영과 컴퓨터 10월호, 2000

[5] Rakesh Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Database," Proceeding of ACM SIGMOD, pp.207-216, 1993

[6] Rakesh Agrawal, R.Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of VLDB Conference, pp.487-499, 1994

[7] 박종수, 유원경, 홍기형, "연관 규칙 탐사와 그 응용", 정보과학회지 제 16권, pp.37-44, 1998

[8] Bing Liu, Wynne Hsu, Yiming Ma, "Mining Association Rules with Multiple Minimum Supports," Proceedings of the ACM SIGKDD (KDD-99), pp.337-341, 1999

[9] M.Houtsma, Arun Swami, "Set-Oriented Mining for Association Rules," IBM Research Report, 1993

[10] B.Ozden, S.Ramaswamy, A.Silberschatz, "Cyclic Association Rules," Proceeding of ICDE, pp.412-421, 1998

[11] 남도원, 김성민, 이동하, 오재훈, 김성훈, 이전영, "한시적 연관규칙에서의 부분 구간 탐사", 한국정보과학회 데이타베이스 학술대회, pp.36-43, 1999

[12] Krzystof Koperski, Jiawey Han, "Discovery of Spatial Association Rules in Geographic Information Databases," Proceedings of SSD'95, pp.47-66, 1995

[13] Rakesh Agrawal, R.Srikant, "Mining Sequential Patterns," Proceedings of the Conference of Data Engineering, pp.3-14, 1995

하 단 심

1999년 전남대학교 전산학과(학사). 2001년 전남대학교 전산학과(석사). 2001년~현재 LG전자 근무. 관심분야는 데이타베이스, 데이타마이닝

황 부 현

1978년 숭실대학교 전산학과(학사). 1980년 한국과학기술원 전산학과(석사). 1994년 한국과학기술원 전산학과(박사). 1980년 ~ 현재 전남대학교 컴퓨터정보학부 교수. 2001년 ~ 현재 전남대학교 정보통신연구소장. 관심분야는 분산시스템, 분산데이터베이스, 객체지향시스템, 전자상거래