중요지지도를 고려한 연관규칙 탐사 알고리즘

김 근 형[†]·황 병 웅[††]·김 민 철[†]

요 약

데이터마이닝 기법중의 하나인 연관규칙 탐사는 데이터베이스상에서 빈번하게 나타나는 데이터들 중 서로 연관성이 강한 데이터들을 탐색 대상으로 한다. 그러나, 빈번하게 나타나지 않는 희소한 데이터들이라 할 지라도 가중치가 높은 중요한 데이터이면서 서로 연관성이 강할 경우 비즈니스정보로서 중요한 가치가 있다. 본 논문에서는 데이터베이스 상에서 희소하게 나타나지만 중요한 의미를 갖고 또한 서로 연관성이 높은 데이터들을 탐사할 수 있는 연관규칙 탐사 알고리즘을 제안한다. 제안한 알고리즘의 성능을 시뮬레이션을 통하여 평가한 결과 희소하면서도 중요한 데이터들 사이의 연관규칙을 효율적으로 탐사함을 알 수 있었다.

Algorithm mining Association Rules by considering Weight Support

Keun Hyung Kim[†]·Byung Woong Whang[††]·Min Chul Kim[†]

ABSTRACT

Association rules mining, which is one of data mining technologies, searches data among which are frequent and related to each other in database. But, although the data are not frequent and rare in database, they have the enough worth of business information if the data ares important and strongly related to each other, In this paper, we propose the algorithm discovering association rules that consist of data, which are rare but, important and strongly related to each other in database. The proposed algorithm was evaluated through simulation. We found that the proposed algorithm discovered efficiently association rules among data, which are not frequent but, important.

키워드 : 데이터마이닝(Data Mining), 연관규칙(Associaton Rule), 희소데이터(Rare Data), 중요지지도(Importance Support)

1. 서 론

데이터마이닝은 축적된 대규모의 데이터들을 분석하여 기업의 이윤 추구에 도움이 될 수 있는 정보와 지식을 획득할 수 있는 기술이다. 데이터마이닝 기법 중 가장 활발하게 연구되고 있는 기법은 연관규칙 탐사 분야이다. 연관규칙은 동시에 자주 나타나는 데이터들에 대한 연관성을 규칙의 형태로 표현한 것으로 이미 발생한 트랜잭션들에 대하여 데이터 사이의 연관성을 발견하여 이를 바탕으로 고객들의 구매 패턴을 분석할 수 있고 상품들의 시장성 예측 등에도 그대로 적용할 수 있다. 연관규칙 탐사는 응용성이 높아 기업의 마케팅, 판매전략, 고개지원 등에 유용하게 이용되고 있다[1].

기존의 연관규칙 탐사 기법은 데이터베이스에서 사용자가 지정한 정도의 통계적인 가치로 정의되는 수치를 만족하는 데이터들 사이에서 발생할 수 있는 연관성을 탐사하였다. 이러한 방법은 데이터베이스에 존재하는 각각의 데이터들은 모두 유사한 발생 빈도로 나타남을 전제하여 연관규칙을 탐사하는 방식이다. 그러나, 실제로는 데이터베이스를 구성하는 데이터들은 데이터항목의 특성에 따라 상대적으로 빈번하게 나타나는 데이터도 존재하고 그렇지 못한 데이터도 존재한다. 그리고, 빈번하게 나타나지 않는 데이터에도 경우에 따라서는 의미있고 중요한 정보가 존재할 수 있다. 예를 들어, 할인마트의 경우 값 싼 식료품과 같이 상대적으로 빈발하게 팔리는 제품과 고가의 전자제품과 같이 식료품보다는 드물게 팔리는 제품들이 있는데, 판매되는 횟수가 실질적인 이윤의 절대조건은 아니다. 즉, 전자제품이 팔리는 횟수는 적지만 실질적인 이익을 가져다 줄 수 있다.

기존 대부분의 연관규칙 탐사 기법들은 빈번하게 나타나는 데이터들만을 탐사 대상으로 하였다. 희소한 데이터들을 탐사대상으로 하는 알고리즘들도 있으나 이 알고리즘들은 희소한 데이터들사이의 연관성 정도만을 고려하였기 때문에 의미없는 연관규칙을 탐사하게 되는 문제가 있었다. 빈번하게 나타나지 않는 데이터를 탐사대상으로 하는 이유는

† 정 회 원 : 제주대학교 경영정보학과 교수
†† 준 회 원 : 제주대학교 경영대학원 경영정보학과
논문접수 : 2003년 9월 2일, 심사완료 : 2004년 1월 6일

비록 희소하게 나타나지만 중요성이 있는 데이터들 사이의 연관성을 찾으려고 하는 것이다. 따라서, 희소하게 나타나는 데이터들을 대상으로 연관규칙을 탐사할 때는 데이터들의 중요성 정도를 파악할 필요가 있다.

본 논문에서는 희소한 데이터들을 대상으로 연관규칙을 탐사할 때 데이터의 중요도를 고려함으로써 의미있는 연관규칙을 보다 효율적으로 탐사할 수 있는 알고리즘을 제안한다. 본 논문에서 제안하는 연관규칙 탐사 기법은 불필요한 데이터들을 배제함으로써 기존의 방법보다 처리속도가 빠르면서 의미있고 중요한 연관규칙을 탐사할 수 있다.

2. 관련 연구

2.1 연관규칙의 개요

연관규칙 탐사 과정은 100% 정확히 동시에 나타나는 규칙이 아닌, 데이터베이스에서 사용자가 지정하는 정도의 통계적 수치를 만족하는 규칙을 발견한다. 지지도(support)와 신뢰도(confidence)라는 통계적 척도를 적절하게 이용하여 데이터베이스에서의 데이터 사이의 연관성을 발견한다[2]. 지지도는 데이터베이스에서 탐사할 데이터들이 관심 있을 정도로 빈발하게 나타나는 항목을 고려하기 위한 데이터의 통계적인 중요성에 대한 척도이며 신뢰도는 그 규칙의 강도를 나타내는 척도이다. 지지도, 신뢰도, 최소지지도, 최소신뢰도의 정의는 <표 2-1>과 같다. 연관규칙은 규칙 R에 대하여 sup(R) ≥ minsup이고 conf(R) ≥ minconf를 만족하면 성립된다. 지지도를 만족하는 데이터항목들은 빈발하다(large)라고 하며 빈발항목들 또는 빈발항목 집합들에 의해 구성된 규칙들의 신뢰도를 검사하여 연관규칙이 탐사된다.

2.2 연관규칙 탐사 관련 선행연구

연관규칙 탐사는 사용자가 정의한 임계값인 최소지지도, 최소신뢰도를 만족하는 규칙이어야 하며 연관규칙의 탐사는 다음의 2단계로 구성된다[8].

단계 1 : 빈발항목집합들(large itemsets)을 찾아낸다. 미리 결정된 최소지지도 minsup 이상의 트랜잭션 지지도를 가지는 항목집합들의 모든 집합들을 빈발항목집합들이라 한다.

단계 2 : 데이터베이스로부터 연관규칙을 생성하기 위하여 빈발항목집합을 사용한다. 모든 빈발항목집합 L에 대하여 L의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 그러한 부분집합 A에 대하여, 만약 sup(A)에 대한 sup(L)의 비율이 적어도 최소신뢰도 minconf 이상이면 (sup(L)/sup(A) ≥ minconf), A→(L-A)의 형태의 규칙을 출력한다. 이 규칙의 지지도는 sup(L)이고 신뢰도는 sup(L)/sup(A)이다. □

연관규칙 탐사의 전체 성능은 첫 번째 단계에서 결정된다. 먼저 빈발항목집합을 확인한 후에 해당되는 연관규칙을 단계 2의 방법으로 쉽게 유도할 수 있다. 모든 알고리즘이 고려하는 기본적인 방법은 후보(candidate)라 지칭하는 빈발가능성이 있는 항목집합들의 생성을 포함한다. 이들 후보 항목집합들 중에 실제로 빈발한 항목을 찾기 위해서는 데이터베이스를 읽어나가면서 각 후보항목들에 대한 지지도가 계산되어야 한다. 후보 항목집합의 발생빈도를 계산하는 것은 상당량의 프로세싱 시간과 메모리를 요구하기 때문에 연관규칙 탐사 알고리즘의 성능은 후보들의 수에 비례한다.

〈표 2-1〉 연관규칙 탐사와 관련된 용어들의 정의

구 분	내 용
지 지 도	데이터집합 X에 대하여 전체 트랜잭션에서 X가 차지하는 비율로 X의 지지도는 sup(X)로 표시한다. $sup(X) = \frac{X\text{를 포함하는 트랜잭션의 수}}{\text{전체트랜잭션의 수}}$
신 뢰 도	규칙에 대한 강도를 나타내는 척도로써 규칙 X→Y로의 규칙이 존재한다면 X를 포함하는 트랜잭션 중에서 Y를 동시에 포함하는 트랜잭션의 비율로 conf(X→Y)로 표시한다. $conf(X \rightarrow Y) = \frac{X\text{와 } Y\text{를 동시에 포함하는 트랜잭션의 수}}{X\text{를 포함하는 전체트랜잭션의 수}}$
최소 지지도	사용자에 의해 정해진 지지도의 임계값으로 minsup로 표기한다.
최소 신뢰도	사용자에 의해 정해진 신뢰도 임계값으로 minconf로 표기한다.
빈발항목 집합	사용자가 정한 최소지지도 minsup에 대하여 데이터항목 집합 X의 지지도 sup(X)와 minsup와의 관계가 sup(X) ≥ minsup를 만족하는 X를 빈발하다(large)라고 정의한다. 빈발항목은 빈발항목집합의 원소에 포함되는 항목을 의미한다. k-빈발항목집합은 k개의 데이터항목들로 구성된 빈발항목집합으로 L_k로 표현한다.
후보항목 집합	빈발항목집합의 원소가 될 가능성이 있는 항목들로 빈발항목집합을 탐사하기 위해 사용되는 집합이다. k-후보 항목집합은 k-개의 데이터항목들로 구성된 후보항목집합을 말하며 C_k로 표현한다.
k-항목 집합	항목집합의 원소가 k개의 데이터항목들로 구성된 항목집합이다.

AIS[6]에서의 많은 수의 후보항목의 생성은 Apriori_gen이라는 새로운 후보항목집합의 생성전략을 개발케 하였으며 이는 Apriori가 연관규칙 탐사 분야에 기여한 중요한 부분이다. Apriori_gen은 후보항목집합의 수를 줄이는데 성공적이어서 그 이후 대부분의 알고리즘에서 사용하게 되었다. 이 방법은 조인단계(join)와 전지단계(prune)로 구성된다. 후보 (k+1)-항목집합은 단지 모든 k-부분집합이 빈발할 때만 선택되어질 것이다. (그림 2-1)에서 나타내는 바와 같이 Apriori_gen은 빈발항목집합 L_k를 입력으로 사용하고 그들의 (k-1)개의 같은 항목들을 갖는 쌍 a, b를 찾는다. (k-1)개의 공통된 항목들과 두 개의 다른 항목들을 찾아서 후보 (k+1)-항목집합을 형성하기 위하여 조인된다.

```
Algorithm Apriori_gen()
insert into C_{k+1} /*join step */
select a.item_1, a.item_2, ···, a.item_k, b.item_k
from L_k a, L_k b
where a.item_1 = b.item_1, ···, a.item_{k-1} = b.item_{k-1}, a.item_k < b.item_k
/* prune step : now prune rule with subsets missing in L_k */
for all itemset c ∈ C_{k+1} do
 for all k-subsets s of c do
if (s ∉ L_k) then delete c from C_{k+1}
```

(그림 2-1) 후보항목집합 생성 알고리즘

전지단계에서는 만들어진 후보 (k+1)-항목집합의 부분집합 k-항목집합들이 이미 L_k에 있는지를 검사하여 없으면 이 후보항목집합을 전지한다.

연관규칙은 다양한 방법으로 응용되고 연구되고 있다. 최근의 연관규칙탐사 관련 연구들은 회소데이터들 사이의 연관성을 탐사하기 위하여 복수의 지지도를 사용하는 연관규칙탐사[9], 회소데이터들 사이의 연관성을 탐사하기 위하여 상대지지도를 사용하는 연관규칙탐사[3], 일정 주기 상에서 나타나는 순환적인 연관규칙 탐사[10] 등으로 응용되어 활발히 연구되고 있다.

2.4 기존 주요연구들의 분석 및 문제제기

[8]에서 제안한 Apriori 알고리즘은 최소지지도와 최소신뢰도를 만족하는 모든 데이터들을 찾는 방법으로 전체 데이터베이스에 대해 단일한 최소지지도가 사용된다. 즉, Apriori 알고리즘은 데이터베이스에 존재하는 모든 데이터들은 유사한 빈도수를 가지고 있는 것으로 가정하고 연관규칙을 탐사하는 방법이다. 그러나, 실세계의 많은 응용에서 모든 데이터들이 유사한 발생 빈도수를 가지고 나타나는 경우는 드물며 상대적으로 많은 빈도를 가지고 나타나는 데이터들이 존재하는가 하면 그렇지 않고 상대적으로 회소한 빈도를 가지고 나타나는 데이터들도 있다. 회소한 빈도를 가지고 나타나는 데이터들 중에서도 중요하고 의미있는 정보가 존재할 수 있다. Apriori 알고리즘에서 상대적으로 회소한 빈도를 갖는 데이터들에 대한 연관규칙을 탐사해야 하는 경우, 사용자는 최소지지도를 낮게 설정하여야 한다. 그러나, 최소지지도를 작은 값으로 설정한다면 최소지지도를 만족하는 빈발한 데이터들이 너무 많게 되고 결과적으로 연관규칙 탐사의 1단계를 처리하는 시간이 길어지게 된다. 또한, 목적하는 회소 데이터뿐만 아니라 그 최소지지도를 만족하는 빈발하는 데이터들로 구성된 모든 규칙들이 부가적으로 탐사되어 작은 최소지지도를 만족하는 모든 데이터들이 서로 연관성을 가지고 있는 것으로 탐사되는 문제가 발생한다.

[9]에서는 데이터들의 빈도형태를 획일적으로 가정하는 Apriori 알고리즘을 개선한 방법인 MSApriori(Multiple Support Apriori) 알고리즘을 제안하고 있다. MSApriori 알고리즘은 데이터들의 빈도형태를 고려하기 위하여 데이터항목 각각의 최소지지도인 MIS(Minimum Item Support)를 사용한다. 데이터베이스를 구성하는 각각의 데이터항목들은 자신의 MIS를 가지고 있다. MSApriori 알고리즘에서의 최소지지도는 규칙을 구성하는 데이터항목 중에서 가장 낮은 MIS값이 그 규칙의 최소지지도로 이용된다. 따라서, 규칙을 구성하는 데이터항목들이 빈번하게 나타나는 데이터들로만 구성되었다면 그 규칙에는 비교적 높은 최소지지도가 적용되며, 반대로 회소한 데이터들로만 구성된 규칙의 경우에는 비교적 낮은 최소지지도가 적용되어 규칙을 탐사한다. 그러나, MSApriori 알고리즘은 데이터베이스에 존재하는 데이터항목 모두에게 MIS를 지정하기 위해서 모든 데이터들의 빈도형태를 연관규칙의 탐사 단계 이전에 파악하여야 하는 문제가 발생한다.

[3]에서 제안한 RSAA(Relative Support Apriori Algorithm)는 1차지지도와 2차지지도를 설정하여(1차지지도 > 2차지지도) 1차지지도를 만족하는 데이터들에는 Apriori를 적용하고, 1차 지지도는 만족하지 않지만 2차지지도를 만족하는 데이터들에는 상대지지도를 적용하여 이를 만족하는 데이터집합을 빈발한 데이터집합으로 평가한다. 상대지지도는 신뢰도의 의미와 비슷한 개념으로 데이터집합내의 데이터들 사이의 연관성 정도를 의미한다. 2차지지도를 만족하는 데이터집합에 대하여 상대지지도를 다시 적용함으로써 최소지지도를 만족하는 빈발한 데이터들이 너무 많지 않게 되어 연관규칙 탐사의 1단계를 처리하는 시간이 Apriori에 비하여 짧아지게 된다.

그러나, 빈발하게 나타나지 않는 회소한 데이터들 사이의 연관성을 탐사하려는 이유는 그 회소한 데이터가 비즈니스적 가치가 높을 수도 있다는 기대 때문이다. 회소한 데이터의 중요성 정도가 높아야 비로소 회소한 데이터 사이의 연관성은 비즈니스적 가치가 있을 수 있다. 즉, 빈발한 데이

터는 빈번하게 나타나기 때문에 의미가 있는 것이고 희소한 데이터는 중요성 정도가 어느 정도 내포되어야 의미가 있는 것이다. 예를 들어, 할인매장에서 A제품들은 B제품들에 비하여 상대적으로 희소하게 판매되지만 판매 이윤이 B제품에 비하여 높기 때문에 A제품들 사이의 판매 연관성은 관심의 대상이 될 수 있다. 이에 반하여 C제품은 A제품만큼 희소하게 판매되면서 판매이윤도 별로 높지 않다면 C제품들 사이의 판매 연관성은 관심의 대상이 될 수 없다. 따라서, 희소한 데이터들을 대상으로 연관규칙을 탐사할 때는 희소한 데이터의 중요성 정도를 고려하는 것이 보다 합리적이고 현실적인 정보처리방법이 될 수 있다.

RSAA 알고리즘은 희소한 데이터들 사이의 연관성만을 고려하였고 희소한 데이터들 사이의 중요성 정도는 고려하지 않기 때문에 결과적으로 무의미한 연관규칙들이 생성된다.

3. 새로운 알고리즘의 제안

이 장에서는 희소한 데이터들을 대상으로 연관규칙들을 탐색할 때 데이터들 사이의 연관성뿐만 아니라 중요성도 동시에 고려하여 보다 효율적으로 연관규칙을 탐사할 수 있는 알고리즘을 제안한다.

3.1 중요가중치의 정의

데이터항목의 중요순위(Order of Importance)는 데이터항목의 매출 이익, 선호도, 심리적 가치등에 의해서 순차적으로 결정된 서열이다. 이때, 중요가중치(Weight of Importance)는 데이터항목의 중요순위를 기초로 데이터항목의 중요도를 가늠할 수 있는 척도이다. 중요가중치를 [정의 1]과 같이 정의한다.

[정의 1] 중요가중치(Weight of Importance)
데이터베이스의 트랜잭션 데이터는 데이터항목의 집합 $I = \{i_1, i_2, i_3, \cdots, i_m\}$를 포함하고 각 항목의 중요순위는 $i_1 < i_2 < i_3 < \cdots < i_m$일 때 데이터항목 i_k의 중요순위 값을 k라고 하자. 이때 i_k의 중요가중치 w_k는 다음과 같다.

$$W_k = \frac{k}{\sum_{a=1}^{m} a} \quad \square$$

예를 들어 7개의 데이터 항목들 $\{i_1, i_2, i_3, \cdots, i_7\}$이 있을 경우 i_2의 중요가중치 w_2는 $2/(1+2+\cdots+7) = 2/28 = 0.07$이 된다.

또한, 임의의 연관규칙을 구성하는 데이터항목 집합의 전체적인 중요성 정도를 나타낼 수 있는 중요지지도는 다음과 같다.

[정의 2] 중요지지도(Weight Support)
데이터항목 집합 $\{i_1, i_2, i_3, \cdots, i_m\}$의 중요지지도 $Wsup(i_1, i_2, i_3, \cdots, i_m)$은 아래와 같이 정의된다.

$$Wsup(i_1, i_2, i_3, \cdots, i_m) = MIN(w_1, w_2, w_3, \cdots, w_m) \quad \square$$

임의의 데이터항목 집합의 중요지지도는 $0 < Wsup < 1$인 값으로써 사용자가 정의하는 최소중요지지도 값을 0으로 설정하면 데이터의 중요성 정도를 고려하지 않고 연관규칙 탐사를 하게 되고 최소 중요지지도 값이 1에 가까울수록 보다 중요한 데이터들을 대상으로만 연관규칙 탐사를 하게 된다.

3.2 새로운 알고리즘

제안하는 알고리즘은 2차지지도와 상대지지도 그리고 중요지지도라는 척도를 사용하여, 비록 데이터베이스에서는 희소하지만 높은 비율로 동시에 발생하는 중요한 데이터항목들 사이의 연관성을 추출해 낼 수 있다.

<표 3-1>은 제안하는 알고리즘에서 사용하는 자료구조와 함수를 나타내고 있다.

〈표 3-1〉 알고리즘에서 사용되는 자료구조와 함수들

이 름	의 미
minWsup	최소중요지지도 : 사용자에 의해 정해지는 중요지지도 임계값
support1	1차지지도
support2	2차지지도
minRsup	최소상대지지도 : 사용자에 의해 정해지는 상대지지도 임계값
Wsup	데이터항목 집합의 중요지지도
Rsup	데이터항목 집합의 상대지지도 : 데이터항목 집합 $X = \{i_1, i_2, i_3, \cdots, i_k\}$일 때 $Rsup(i_1, i_2, i_3, \cdots, i_k) = \max(sup(i_1, i_2, i_3, \cdots, i_k)/sup(i_1), sup(i_1, i_2, i_3, \cdots, i_k)/sup(i_2), \cdots, sup(i_1, i_2, i_3, \cdots, i_k)/sup(i_k))$
$semiL_k$	k-준빈발항목집합 : 1차지지도는 만족하지 않지만 2차지지도, 상대지지도, 중요지지도를 만족하는 k-항목집합
$semiC_k$	k-준후보항목집합 : (k-1)-준빈발항목집합들로부터 생성된 k-항목집합
Apriori_gen()	(k-1)-준빈발항목집합들로부터 k-준후보항목집합들을 생성하는 함수

(그림 3-1)의 알고리즘에서 (1), (2)행은 빈발 데이터와 희소데이터를 구분하는 과정이고, (3)행을 통하여 빈발 데이터에 대해서는 Apriori 알고리즘을 적용하고 (3)행 이후에서는 희소 데이터에 대하여 연관규칙을 탐사하는 과정이다.

구체적으로 살펴보면, (그림 3-1)의 (1), (2)행에서는 데이터베이스에 있는 각 데이터항목들의 지지도를 계산하여 1차지지도를 만족하는 것들은 1-빈발항목집합 L_1에, 1차지지도는 만족하지 않지만 2차지지도를 만족하는 데이터항목들은 1-준후보항목집합인 $semiC_1$에 포함시킨다.

(3)행에서는 1-빈발항목집합 L_1을 Apriori 알고리즘에 적용하여 k-후보항목집합, k-빈발항목집합들을 생성하면서 연

관규칙 탐사를 수행한다.

(3)행 이후부터는 희소데이터들이 포함된 $semiC_1$에 대하여 k-준후보항목집합들과 k-준빈발항목집합들을 생성하면서 연관규칙 탐사를 수행한다.

(4)행의 for 반복문 블록에 의하여 k-준후보항목집합이 공집합이 될 때까지 연관규칙 탐사과정은 반복 수행된다. (6)행에서 1-준후보항목집합 $semiC_1$끼리를 조인한 결과와 $semiC_1$과 L_1을 조인한 결과를 유니온하여 2-준후보항목집합 $semiC_2$를 생성한다.

```
Proposed Algorithm
I ←{i1, i2, i3, ···, im} / *데이터베이스의 모든 항목 */
while each item i ∈ I do
(1) if (i.support ≥ support1) then i ∈ L1
(2) else (i.support ≥ support2) then i ∈ semiC1
end
each item ∈ L1
(3) do Apriori Algorithm
end
(4) for (k ← 2 ; semiCk ≠ ∅ ; k++)do
(5) if (k = 2) then
(6) semiC2 ← Apriori_gen(semiC1, semiC1)
 ∪ Apriori_gen(semiC1, L1) ;
 else
(7) semiCk ← Apriori_gen(semiCk-1, semiCk-1) ;
 end if
(8) while (a ∈ semiCk) do
(9) if ((a.support ≥ support2) and
 (a.Rsup ≥ minRSup) and (a.Wsup ≥ minWsup))
 then a ∈ semiLk
end if
end while
end for
(10) Answer ← ∪k semiLk
end
```

(그림 3-1) 제안하는 알고리즘

$semiC_1$과 L_1을 조인한 결과는 2-후보항목집합은 될 수 없지만(왜냐하면, $semiC_1$이 빈발항목집합이 아니기 때문) 2-준후보항목집합은 될 수 있다. (7)행은 k값이 증가하는(단, $k>2$) for 반복 시마다 k-준후보항목집합을 생성하는 과정이다. (8), (9)행에서는 while 반복을 수행하면서 k-준후보항목집합의 각 요소에 대하여 2차지지도, 상대지지도, 중요지지도를 적용하여 k-준빈발항목집합을 생성한다. (10)행에서는 반복문들을 통하여 생성된 모든 k-준빈발항목집합($k \geq 1$)들을 유니온하여 최종적인 준빈발항목집합을 생성함으로써 연관규칙 탐사과정의 1단계를 완성한다.

3.3 중요지지도를 고려한 연관규칙 탐사의 예

이 절에서는 제안한 알고리즘이 연관성이 높은 희소한 데이터중에서도 중요한 데이터 항목만을 대상으로 마이닝을 수행하는 예를 고찰한다. (그림 3-2)는 트랜잭션 데이터베이스의 예를 나타내고 있다. 트랜잭션에 포함되는 데이터 항목들은 A, B, C, D, E, F, G이고 데이터항목들의 중요순위는 A<B<C<D<E<F<G라고 가정한다. 연관규칙탐사를 위하여 1차지지도는 40%, 2차지지도는 20%, 최소상대지지도는 0.7, 최소중요지지도는 0.1로 설정한다.

TID	항 목	TID	항 목
1	B C	6	D F
2	D E	7	C D F G
3	B C D E F G	8	A B
4	F G	9	C D E
5	C D E F	10	A B

(그림 3-2) 트랜잭션 데이터베이스

트랜잭션을 구성하는 데이터항목들에 대한 중요지지도는 (그림 3-3)과 같다. (그림 3-3)에서 $semiL_1$ = {A, E, G}는 준빈발항목집합으로 1차지지도는 만족하지 못하지만 2차지지도는 만족하는 데이터항목이다. L_1 = {B, C, D, F}는 빈발항목집합이다. (그림 3-4)는 $semiL_1 \times semiL_1$과 $L_1 \times semiL_1$을 수행하여 생성된 2-준후보항목집합의 지지도와 상대지지도, 중요지지도를 나타내고 있다. (그림 3-4)에서 2-준후보항목집합에 대하여 상대지지도가 0.7이상인 항목은 {{A, B}, {E, C}, {E, D}}이다. 이 중에서 최소중요지지도가 0.1 이상인 항목은 {{E, C}, {E, D}}가 된다. {A, B}는 중요지지도가 0.04이므로 상대지지도가 높지만 전지한다. 이상에서 알 수 있는 바와 같이 제안한 알고리즘은 상대지지도가 높은 희소한 데이터중에서 중요지지도가 높은 데이터항목들만을 대상으로 준빈발항목집합들을 생성함을 알 수 있다.

항 목	지지도	중요가중치	항 목	지지도	중요가중치
A	2	(1/28) = 0.04	F	5	(6/28) = 0.2
B	4	(2/28) = 0.07	G	3	(7/28) = 0.25
C	6	(3/28) = 0.1			
D	6	(4/28) = 0.14			
E	3	(5/28) = 0.18			

(그림 3-3) 데이터항목들의 지지도 및 중요지지도

항목집합	지지도	상대지지도	중요지지도
{A, B}	2	1.0	0.04
{A, C}	0	0.0	0.04
{A, D}	0	0.0	0.04
{A, F}	0	0.0	0.04
{E, B}	1	0.25	0.07
{E, C}	3	0.75	0.18
{E, D}	3	1.0	0.14
{E, F}	3	0.5	0.18
{G, B}	1	0.33	0.25
{G, C}	2	0.66	0.25
{G, D}	2	0.66	0.14
{G, F}	2	0.66	0.25
{A, E}	0	0.0	0.04
{A, G}	0	0.0	0.04
{E, G}	1	0.33	0.18

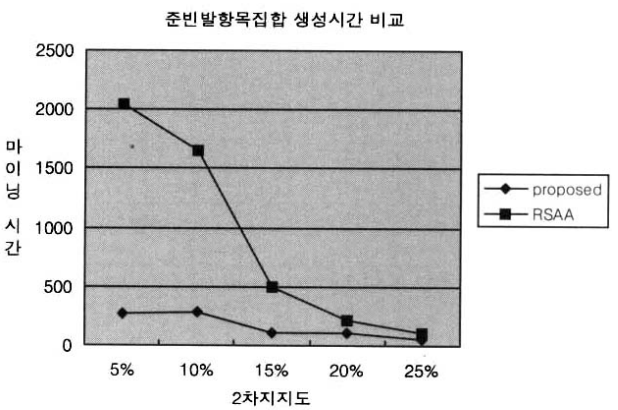
(그림 3-4) 준후보항목집합들의 지지도 및 중요지지도

4. 실험 및 결과 분석

이 장에서는 희소 데이터를 대상으로 연관규칙 탐사를 할 때 기존의 RSAA보다 제안한 알고리즘이 보다 효율적임을 보이고자 한다. RSAA와 제안한 알고리즘을 비교하기 위한 기준은 3가지로 설정하였는데 첫째, 전체 준빈발항목집합들을 생성하는 시간, 둘째 전체 준후보항목집합의 개수, 셋째 전체 준빈발항목집합의 개수이다[3]. 후보항목집합과 빈발항목집합은 RSAA와 제안한 알고리즘 모두 Apriori 기법을 이용하기 때문에 평가대상에서 제외하였다.

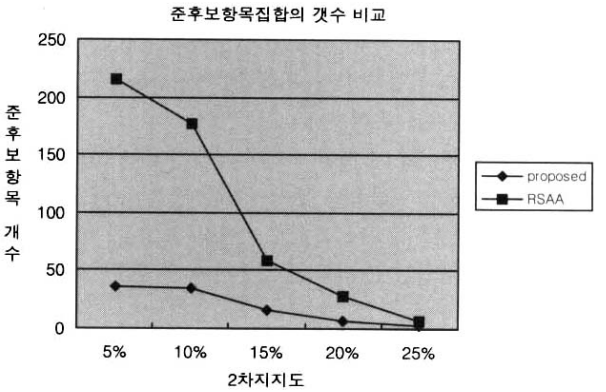
RSAA와 제안한 알고리즘은 C언어를 이용하여 구현되었으며 구현된 알고리즘은 펜티엄Ⅳ의 Windows-98 환경에서 실행 및 실험되었다.

임의의 트랜잭션 수는 10,000 개이며 데이터항목의 개수는 55개, 트랜잭션 당 최대 항목의 개수는 14개로 설정하였다. 트랜잭션 데이터는 랜덤함수를 이용하여 생성하였고 연관성이 있는 희소한 데이터 비율은 전체 데이터의 25% 수준으로 설정하였다. 연관규칙탐사의척도로서 상대지지도는 0.5, 중요지지도는 0.5, 1차지지도는 70%로 설정하고 RSAA와 제안한 알고리즘을 비교하였다.



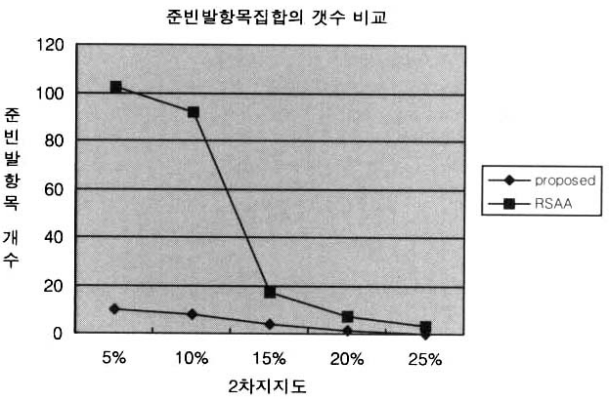
(그림 4-1) 마이닝 시간 비교

(그림 4-1)은 2차지지도 값에 따라 전체 준빈발항목집합들을 생성하는 시간이 얼마나 걸리는지를 나타내고 있다. 여기에서의 시간 단위는 CPU가 특정 명령집합을 처리하는 시간이다. 전체 준빈발항목집합을 생성하는 시간은 RSAA보다 제안한 알고리즘이 덜 걸리고 있음을 알 수 있다. 이것은 제안한 알고리즘의 경우 중요지지도가 0.5 미만인 데이터는 데이터마이닝에서 제외함으로써 처리해야 할 준후보항목집합 및 준빈발항목집합의 수가 줄어들었기 때문이다. 또한, 2차지지도 값이 커질수록 데이터마이닝 시간은 짧아지고 있음을 알 수 있다.



(그림 4-2) 준후보항목집합의 개수 비교

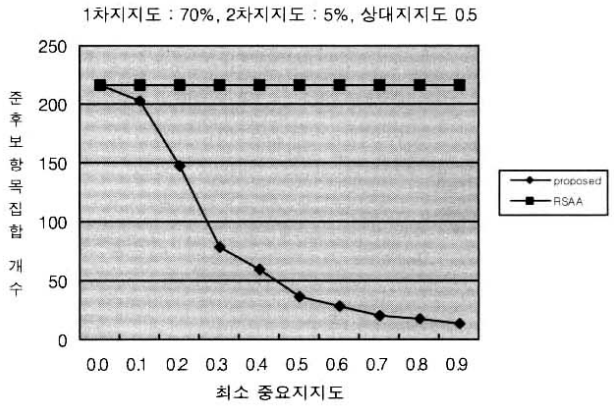
(그림 4-2)는 2차지지도 값에 따라 최종적으로 생성된 전체 준후보항목집합들의 개수가 얼마나 되는지를 보여주고 있다. 최종적으로 생성된 전체 준후보항목집합들의 수는 RSAA보다 제안한 알고리즘이 적음을 알 수 있다. 왜냐하면, 중요지지도가 0.5 이상인 중요한 데이터항목집합들만을 탐사 대상으로 하기 때문이다.



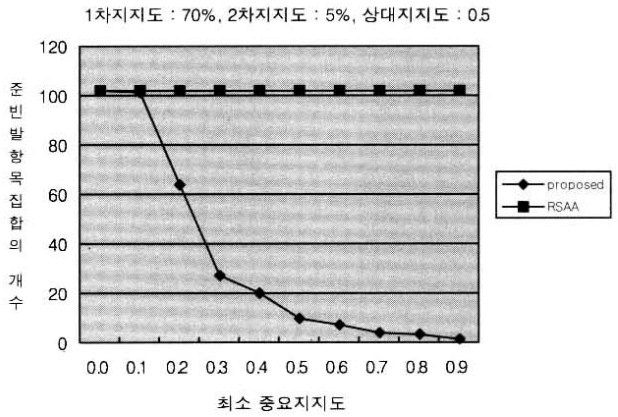
(그림 4-3) 준빈발항목집합의 개수 비교

(그림 4-3)은 2차지지도 값에 따라 최종적으로 생성된 전체 준빈발항목집합들의 개수가 얼마나 되는지를 보여주고 있다. 제안한 알고리즘은 중요한 데이터만을 탐사 대상으로 하므로 최종적으로 생성된 전체 준빈발항목집합들의 수는 RSAA보다 제안한 알고리즘이 적다.

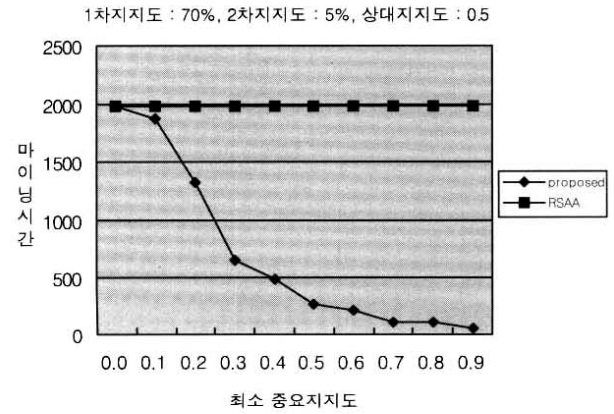
(그림 4-4), (그림 4-5), (그림 4-6)은 제안한 알고리즘과 RSAA 알고리즘에 대하여 1차 지지도를 70%, 2차 지지도를 5%, 상대지지도를 0.5로 설정하였을 때 중요지지도 변화에 따른 후보 및 빈발항목집합 개수의 변화, 마이닝 시간의 변화를 나타내고 있다. 제안한 알고리즘에서 최소 중요지지도가 0일 때는 중요지지도를 고려하지 않는 상황이 되기 때문에 RSAA 알고리즘과 동일한 결과를 나타내게 된다. 따라서, 제안한 알고리즘은 RSAA 알고리즘의 기능을 수용할 수 있는 보다 일반화된 알고리즘이라 할 수 있다.



(그림 4-4) 중요도 변화에 따른 준후보항목집합 개수



(그림 4-5) 중요도 변화에 따른 준빈발항목집합 개수



(그림 4-6) 중요도 변화에 따른 마이닝 시간의 비교

5. 결 론

기존의 대부분의 연관규칙탐사 알고리즘들은 빈발하게 나타나는 데이터항목들만을 대상으로 연관규칙 탐사 작업을 수행하였고, 빈발하게 나타나지 않는 데이터항목들을 대상으로 연관규칙을 탐사하는 알고리즘들도 있으나 이러한 알고리즘들은 데이터항목 사이의 연관성만을 고려하기 때문에 중요성이 있는 희소데이터들에 대한 연관성을 탐사하려는 원래의 목적에 충실한 연관규칙을 탐사할 수 없었다.

본 논문에서는 데이터항목들에 중요순위를 부여하고 중요가중치를 기반으로 데이터항목들의 중요성 정도를 측정하여 어느 정도의 중요지지도를 만족하는 연관규칙을 탐사하는 알고리즘을 제안하였다. 본 논문에서 제안한 연관규칙 탐사 알고리즘을 활용함으로써 데이터베이스 상에 희소하게 나타나지만 중요한 데이터항목들에 대한 연관성 탐사를 보다 효율적으로 처리할 수 있다. 또한, 최소 중요지지도를 조절하여 기존의 RSAA 알고리즘의 기능을 수용할 수 있게 할 수 있으므로 본 논문에서 제안한 알고리즘은 기존의 RSAA 알고리즘을 일반화시켰다고 볼 수 있다.

기업활동의 결과로 생성된 데이터들은 갈수록 많아져서 데이터 과잉 문제가 발생하는 현실을 감안할 때 보다 중요한 데이터들만을 대상으로 연관규칙을 탐사할 수 있는 본 알고리즘은 향후 기업 데이터 분석에 보다 유용하게 적용될 수 있을 것으로 기대한다.

참 고 문 헌

[1] 김종훈, "데이터마이닝에 기초한 DB 마케팅 분석방법과 사례", 월간 경영과 컴퓨터, Oct., 2000.

[2] 박종수, 유원경, 홍기형, "연관규칙 탐사와 그 응용", 정보과학회지, 제16권, 1998.

[3] 하단심, 황부현, "의미있는 희소 데이터를 포함한 연관규칙 탐사 기법", 정보과학회논문지, 2001.

[4] 김정자, 이도헌, "데이터마이닝 기술 및 연구동향", 정보과학회지, 제16권 제9호, 1998.

[5] 남도원, 김성민, 이동하, 오재훈, 김성훈, 이전여, "한시적 연관규칙에서의 부분구간 탐사", 한국정보과학회 데이터베이스 학술대회, 1999.

[6] R. Agrawal, Tomasz Imielinski, Arun Swami, "Mining Association Rules between Sets of Items in Large Database," *Proc. of ACM SIGMOD*, 1993.

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

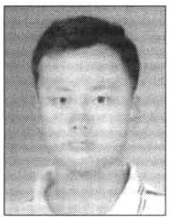
[8] R. Agrawal and R. Srikant, "Fast algorithms for Mining Association Rules," Proc. of VLDB, 1994.

[9] Bing Liu, Wynne Hsu, Yiming Ma, "Mining Association Rules with Multiple Minimum Supports," Proc. of ACM SIGKDD(KDD-99), 1999.

[10] B. Ozden, S. Ramaswamy, A. Silberschatz, "Cyclic Association Rules," Proc. of ICDE, 1998.

김 근 형

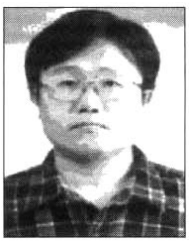
e-mail : khkim@cheju.ac.kr
1990년 서강대학교 컴퓨터학과(공학사)
1992년 서강대학교 컴퓨터학과(공학석사)
2001년 서강대학교 컴퓨터학과(공학박사)
1992년~1994년 현대전자 소프트웨어 연구소
2001년~현재 제주대학교 경영정보학과 조교수
관심분야 : 데이터베이스, 데이터마이닝, 데이터웨어하우스, MIS



황 병 웅

e-mail : musam@badaro21.net
1991년 탐라대학교 경영학과(경영학사)
2003년 제주대학교 경영정보학과 (경영학석사)
관심분야 : 데이터마이닝, MIS

김 민 철



e-mail : khkim@cheju.ac.kr
1991년 중앙대학교 경영학과(경영학사)
1995년 고려대학교 경영학과(경영학석사)
2000년 고려대학교 경영학과(경영학박사)
1995년~1996년 SK텔레콤 기획본부
2001년~현재 제주대학교 경영정보학과 조교수
관심분야 : 경영정보기술, MIS