수량적 연관규칙탐사를 위한 효율적인 고빈도항목열 생성기법

최 영 희[†] · 장 수 민[††] · 유 재 수[†††] · 오 재 철[††††]

요 약

본 논문은 기존의 수량적 연관규칙탐사를 위한 고빈도항목열 생성방법이 가지고 있는 문제점을 해결하는 효율적인 고빈도항목열 생성기법을 제안한다. 제안된 생성기법은 분할과정에서 최소분할지지율을 이용하여 분할간격을 유동적으로 결정하는 방법과 데이타의 집중도를 우선 순위로 하여 분할된 소간격을 병합하는 방법을 사용한다. 본 논문에서 제안한 방법은 기존의 방법보다 세밀한 고빈도항목열을 생성할 수 있는 것과 데이타들의 특성을 잃어버리지 않는 특징을 갖는다. 성능평가를 통하여 제안된 방법이 기존의 방법에 비해 보다 효율적인 고빈도항목열을 생성함을 보인다.

Generating Large Items Efficiently For Mining Quantitative Association Rules

Young-Hee Choi[†] · Su-Min Jang[††] · Jae-Soo Yoo[†††] · Jae-Chul Oh[††††]

ABSTRACT

In this paper, we propose an efficient large item generation algorithm that overcomes the problem of the existing algorithm for making large items from quantitative attributes. The proposed algorithm splits dataset into variable size of intervals by min_split_support and merges the intervals according to the support of each interval. It reflects characteristic of data to generated large items and can generate finer large items than the existing algorithm. It is shown through the performance evaluation that our proposed algorithm outperforms the existing algorithm.

1. 서 론

최근 데이터베이스가 여러 분야에 응용됨에 따라 데이터베이스의 규모가 대용량화되고 있다. 데이터베이스가 대용량화됨에 따라 데이터베이스 모델링에서는 현실세계(real world)의 모든 규칙성(regularity)을 반영하지 못하고 있다. 이런 규칙성을 발견하기 위해 데이타베이스 모델링에 반영되지 못하고 감추어져 있는 규칙성을 발견하는 데이터베이스 마이닝에 대한 연구[1][2][3][4]가 활발히 시작되고 있다.

데이타 마이닝에서 발견된 규칙은 업무 분야의 특성을 효과적으로 대변하여 의사결정(decision making)에 유용한 정보를 제공한다. 예를 들어 슈퍼마켓의 경우 상품 판매가 카운터에서 트랜잭션 별로 기록되어진다. 이런 트랜잭션을 바탕으로 95%의 고객이 상품 A,B를 구입하면 상품 C,D를 구입한다는 연관 규칙이 생성될 수 있다. 일반적인 규칙을 쉽게 상식적으로 접근할 수 있다. 또한 데이타마이닝의 연관 규칙 탐사는 상식적

※ 본 논문은 정보통신부의 정보통신 우수시범학교지원사업에 의하여 수행된 것입니다.
† 정 회 원 : 호원대학교 전기전자정보공학부 교수
†† 준 회 원 : 충북대학교 정보통신공학과
††† 정 회 원 : 충북대학교 전기전자공학부 교수
†††† 정 회 원 : 순천대학교 컴퓨터과학과 교수
논문접수 : 1998년 9월 28일, 심사완료 : 1999년 9월 10일

으로 접근할 수 없는 규칙을 찾아낼 수 있다. 최근에 한 슈퍼에서 아기기저귀를 사는 사람이 맥주를 함께 산다는 규칙을 발견하였는데 이처럼 대용량의 데이타 자체가 가지고 있는 모든 규칙을 탐사할 수 있다. 연관규칙탐사는 트랜잭션이 발생한 업무 분야의 특성을 효과적으로 대변하여 의사 결정에 유용한 정보를 제공하기 때문에, 대용량의 트랜잭션으로 구성된 데이타베이스로부터 연관 규칙을 탐사하기 위한 연구가 최근 활발히 이루어지고 있다[1][2].

대부분의 연관규칙탐사에 대한 연구는 탐사의 대상이 되는 데이터를 이진항목형태로 변환하여 탐사하는 방법이 대부분이다. 연관규칙탐사는 우선 데이타가 항목 중에 사용자가 정의한 최소지지율을 넘는 항목을 1차 고빈도항목열로 생성하고, 하나의 항목에 대한 고빈도항목열의 정보를 바탕으로 항목수를 확장해가면서 n차 고빈도항목열을 생성한다. 이와 같이 생성된 고빈도항목열은 연관규칙을 내포하고 있기 때문에 고빈도항목열 생성이 연관규칙탐사의 중요한 역할을 한다.

대부분의 고빈도항목열 생성에서 사용되고 있는 데이터항목은 이진항목인데 반해 수량적 데이터는 취할 수 있는 값의 범위가 크다. 예를 들어 사람의 연령과 같은 수치 데이터항목은 1~100 정도의 범위에서 하나의 값을 취한다. 수량적 데이타들은 각각의 수치데이타의 값을 그대로 항목으로 설정하면 고빈도항목열을 생성할 수 없는 문제점을 갖는다.

이와 같은 문제점을 해결하기 위해서 수량적데이타의 연관규칙탐사 방법은 일정한 소간격으로 분할하고 이웃하는 소간격들을 병합하는 방법을 이용하여 수량적 데이타를 이진항목의 데이타로 변형하여 고빈도항목열을 생성한다. 이처럼 수량적데이타에 대한 연관규칙탐사는 수량적 데이타에서 적절한 방법을 사용하여 어떻게 고빈도항목열을 생성하느냐가 성능에 중요한 영향을 미친다. 그러나 수량적 데이터를 일정한 소간격으로 분할하고 이웃하는 소간격들을 병합하는 방법을 이용한 고빈도항목열 생성에서는 일정한 간격으로 분할하기 때문에 불필요한 분할이 많이 발생하고 데이타들이 심하게 편중되어 있으면 고빈도항목열을 생성하는데 문제점을 갖는다.

본 논문은 이러한 문제점을 해결하기 위해서 수량적 데이터의 연관규칙탐사를 위한 효율적인 고빈도항목열 생성기법을 제안한다. 제안된 생성기법은 분할과정에서 최소분할지지율을 이용하여 분할간격을 유동적으로 결정하는 방법과 데이타의 집중도를 우선 순위로 하여 분할된 소간격을 병합하는 방법을 사용한다. 본 논문에서 제안한 방법은 기존의 방법보다 세밀한 고빈도항목열을 생성할 수 있는 것과 데이타들의 특성을 잃어버리지 않는 특징을 갖는다.

본 논문의 구성은 제2장에서는 연관규칙탐사의 기본정의와 관련연구에 대해서 알아보고 제3장에서는 기존 수량적 연관규칙탐사를 위한 고빈도항목열 생성기법의 문제점을 제시한다. 제4장에서는 본 논문에서 제안하는 방법을 소개한다. 제5장에서는 성능평가를 통하여 제안된 방법이 효율적으로 고빈도항목열을 생성함 보인다. 마지막으로 제6장에서는 결론과 향후 연구방향을 밝힌다.

2. 연관 규칙 탐사

2.1 연관규칙 기본정의

항목들의 집합을 $I=\{i_1, i_2, i_3, i_4, \ldots, i_m\}$이라 하면, 사건은 I의 부분집합으로 표현된다. N개의 사건 T가 저장된 데이타베이스 D에서 연관규칙을 탐사한다고 하자. 항목열 X,Y에 대한 탐사의 결과는 지지도 임계치(minimum support) ms이상의 지지도와 신뢰도 임계치(minimum confidence) mc이상의 신뢰도를 갖고, 아래와 같은 성질을 갖는 연관규칙 $X \Rightarrow Y$ 의 집합이다.

- $X \subseteq I$ 인 X에 대해, $X \subseteq T$이면 T는 X를 만족한다고 정의하고 D에서 X를 만족하는 사건수를 freq(x)로 표기한다.
- $X, Y \subseteq I$ 이고 $X \cup Y = \varnothing$이면, 규칙 $X \Rightarrow Y$는지지도 $S = \frac{freq(X \cup Y)}{N}$ 와 신뢰도 $C = \frac{freq(X \cup Y)}{freq(X)}$ 를 갖는다.
- ms이상의 지지도를 갖는 항목열 $X \subseteq I$ 를 고빈도항목열이라 정의한다. 이때 X의 모든 부분집합도 고빈도 항목열이다.

연관규칙의 탐사작업은 고빈도항목열 탐색과정과 연관규칙 생성과정으로 이루어진다. 탐색과정은 가능한 후보 항목열을 생성하여, 데이타베이스에서 ms이상의 지지도를 갖는 고빈도 항목열을 탐색한다. 생성과정에서는 고빈도 항목열($X \cup Y$)에 대해, 부부집합(X)과 그 여집합(Y)을 규칙의 조건부와 결과부에 대응시켜 mc이상의 신뢰도를 갖는 연관 규칙 $X \Rightarrow Y$을

생성한다.

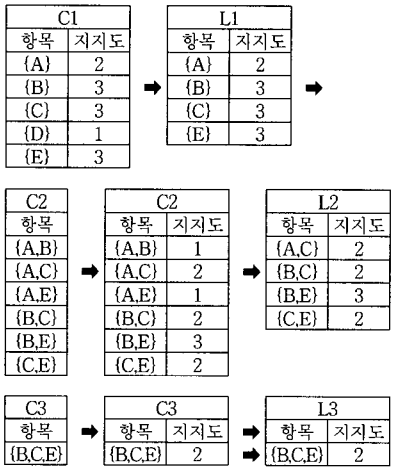
2.2 연관규칙탐사기법

연관규칙탐사를 위해 가장 먼저 제안된 기법은 Apriori 알고리즘이다. 물론, 그 이후에 많은 연구들을 통해 훨씬 높은 성능을 보이는 많은 방법들이 제안되었지만 연관규칙탐사의 기본 전략을 이해하는데 도움이 되므로 여기에서 간단히 예시한다. 그림 1과 같이 트랜잭션 데이타베이스가 주어진다고 가정하자. 이때 사건항목의 전체집합 I는 {A, B, C, D, E}라고 하자.

트랜잭션번호	사건항목
101	A, C, D
102	B, C, E
103	A, B, C, E
104	B, E

(그림 1) 연관규칙 탐사를 위한 예제 트랜잭션 데이타베이스

연관 규칙의 탐사작업은 크게 두 단계로 구성된다. 먼저 일단계 작업은 높은 지지도를 갖는 사건항목집합들을 식별하는 작업이다. 다음 단계는 얻어진 사건항목집합들로부터 높은 신뢰도를 갖는 연관 규칙을 도출하는 작업이다. (그림 2)는 일단계 작업을 보여 준다. 먼저 데이타베이스의 각 트랜잭션을 하나씩 조회하면서 원소가 한 개로 구성된 사건항목집합을 만들고 그것이 포함된 트랜잭션의 개수를 구하면 C1이 된다. C1으로부터 최소한 지지도가 40%이상 즉, 2개이상의 트랜잭션에 의해서 지지되는 사건항목집합만을 선별하면 L1이 된다. L1으로부터 원소 개수가 두 개인 모든 가능한 사건항목집합을 구성하면 C2가 된다. 다시 데이타베이스의 각 트랜잭션을 조회하면서 C2에서 얻어진 사건항목집합 각각에 대한 지지도를 구하면 두 번째 C2가 된다. 이 중 역시 40%이상의 지지도를 갖는 사건항목집합을 선별하면 L2가 된다. 같은 과정을 계속 반복하면 최종적으로 L3를 얻음으로써 알고리즘이 종료한다. 결과적으로 L1, L2, L3가 일 단계 과정의 최종 결과물이 된다. 이것으로부터 두 번째 단계인 신뢰도가 높은 연관 규칙을 발굴하는 것은 간단한 작업이다. 예를 들어 {B, C, E}의 지지도가 2라는 정보와 {B, C}의 지지도가 2라는 정보를 이용하여 연관 규칙 {B, C} => {E}의 신뢰도는 100%라는 사실을 알 수 있다. 위에서 예시한 Apriori 알고리즘은 원리가 간단한 장점은 있지만, 사건항목집합의 크기를 하나씩 늘려갈 때마다 전체 데이타베이스를 조회해야 한다는 문제점이 있다. 고전적인 인공지능분야와는 달리 데이타마이닝 분야에서는 디스크 접근 횟수가 성능에 중요한 영향을 미치는 요소가 된다. 따라서 수행시간이 불필요하게 많이 소요되는 단점이 있다. 이와 같은 단점을 극복하기 위해서 많은 변형 알고리즘[8][9]들이 제시되고 있다.



(그림 2) Apriori 알고리즘을 이용한 연관규칙탐사과정

Apriori 알고리즘을 변형시킨 알고리즘으로 AprioriTid 알고리즘에서는 항목열의 크기를 1씩 증가시키면서, 각 항목열에 대해 만족하는 사건을 색인 관리하여, 데이타베이스를 직접 검색하지 않고 색인화일에 의해 항목열의 지지도를 계산한다. 대부분의 응용분야에서는 후보 2-항목열의 수가 가장 많으며 2-항목열을 만족시키는 사건수가 많음을 알 수 있다. AprioriTid 알고리즘에서는 후보 2-항목열을 만족시키는 사건의 색인화일이 커지는 문제점이 있다. AprioriTid를 개선한 AprioriHybrid 알고리즘은 고빈도 항목열의 탐색과정을 항목열의 크기가 2이하인 경우에는 Apriori를, 3이상의 경우에서는 AprioriTid를 사용한다. 또한 Partition 알고리즘에서는 데이타베이스를 일정크기의 영역(partition)으로 분할하여 2번의 데이타베이스 검색으로 고빈도 항목열이 탐색되는 기법을 제안하였다. 각 영역에서 탐색된 고빈도 항목열을 후보 항목열로 병합하여, 데이타베이스의 재검색으로 지지도를 계산하여 전체 데이타베이스에 대한 고빈도 항목열을 탐색한다.

이와 같이 고빈도항목열의 탐색과정은 후보 항목열

의 생성방법에 따른 탐색공간의 축소기법과 항목열의 지지도 계산방법에 따른 데이타베이스 검색기법의 연구가 진행되고 있다. 연관 규칙의 생성과정에 대한 연구는 다른 규칙에서 유추 가능한 형태의 중복성을 배제하면서, 의사결정자에게 규칙의 의미를 명료하게 전달할 수 있는 표현기법에 대한 연구가 있었다. 항목 분류간의 계층정보는 항목을 일반화시킴으로써 데이타베이스의 특성과 분류의 규칙을 탐사할 수 있다. 항목 분류상의 계층정보를 도입하여 일반화 연관규칙탐사기법에 대한 연구[14]도 있었다. 상위계층의 항목은 하위계층의 항목보다 큰 지지도를 갖음으로 하위계층에서 발견되지 않은 연관 규칙이 상위계층으로 일반화시켜 발견될 수 있다. 또한, 분류계층별로 지지도 임계치를 다르게 적용하는 다단계 연관 규칙의 탐사기법도 제안되었다.

앞에서 언급한 대부분의 연구에서는 항목의 발생 여부만을 고려하는 이진 연관 규칙의 탐사에 관하여 언급하고 있다. 그러나 이진연관규칙 이외에도 항목의 값을 임의의 실수값으로 확장시킨 수량 연관 규칙도 중요하다. [5]에서는 항목의 정의영역을 작은 범위의 소간격으로 분할한 후 이웃한 소간격을 병합하면서 지지도 임계치를 만족하는 간격을 생성한다. 생성된 간격을 단위 항목으로 고려하여 Apriori 알고리즘에 의한 수량 연관 규칙을 탐사한다. Apriori 알고리즘에서는 간격들 간의 연관성을 고려하지 않으므로, 간격의 수에 따라 탐색공간이 커지는 문제점이 발생된다. 또 수량적 연관규칙탐사의 연구로 이진 연관규칙을 확장하여 항목별 발생횟수를 고려하는 수량연관규칙의 한 형태인 단방향 수량적 연관규칙탐사에 대한 연구[17]가 있다.

3. 기존 고빈도항목열 생성기법

2장에서 언급한 바와 같이 연관규칙탐사는 대부분 이진 연관규칙탐사에 대한 연구이다. 이진 연관규칙탐사는 항목의 발생여부를 고려하여 연관규칙을 찾아낸다. 가장 잘 알려진 Apriori기법도 이진 연관규칙탐사기법이다. Apriori의 변형인 AprioriTid와 AprioriTid를 개선한 AprioriHybrid 또한 이진 연관규칙탐사기법으로 분류되어진다. 그러나 실세계의 데이터는 다양한 형태의 데이터를 가지고 있기 때문에 항상 이진항목형태로 되어 있는 것은 아니다.

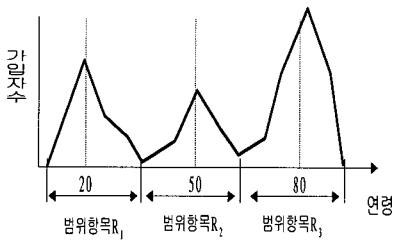
수량적 데이터에 대한 연관규칙탐사는 우선 수량적 데이타를 이진 연관규칙탐사기법으로 사용할 수 있도록 이진항목형태로 매핑해야 하는 문제점을 가지고 있다. 연관규칙을 탐사하고자 하는 수량적 데이타에서 유일하게 구분할 수 있는 수치데이타 값들의 범위를 도메인이라 한다. 도메인이 적을 경우에 이진항목으로 매핑은 아주 쉽지만 많을 경우에는 문제가 달라진다. 도메인 범위가 클 경우에는 수량적 데이타가 갖는 수치데이타의 값을 그대로 항목으로 매핑하려면 그 수치데이타가 갖는 유일한 값의 도메인만큼 항목이 생성되어, 각각의 항목이 갖는 지지율은 아주 낮은 지지율을 갖게되어 사용자가 정의한 임의의 최소지지율을 만족하는 항목을 찾을 수 없을 것이다. 수량적 데이타의 유일한 값의 도메인이 많을 경우에 생기는 이러한 문제점을 해결하기 위해서 (그림 3)과 같이 수량적 데이타를 소간격으로 분할하고 분할된 소간격을 병합하여 범위항목열을 생성하여 이진항목으로 매핑하여 수량적 데이타에 대한 연관규칙을 탐사한다. 우선 나이의 수치데이타를 일정한 소간격으로 나누면서 지지수와 지지율을 구한다. 예제에서는 16~20, 21~25처럼 5살 간격으로 소간격을 구성하였다. 그런 다음 인접하는 소간격을 병합하면서 범위항목열을 만들어 간다. 범위항목중에 사용자최소지지율보다 높은 항목은 고빈도항목열이 된다. 예제에서는 16~25와 36~45가 고빈도항목열이다.

나이	지지수	지지율
실제데이타		
16	1	2%
19	2	4%
21	2	4%
23	4	8%
24	7	14%
26	5	10%
29	2	4%
32	1	2%
35	1	2%
39	3	6%
40	5	10%
41	10	20%
42	7	14%
소간격으로 분할		
16~20	3	6%
21~25	13	26%
26~30	7	14%
31~35	2	4%
36~40	8	16%
41~45	17	34%
소간격 병합으로 범위항목열 생성		
16~25	16	32%
26~35	9	18%
36~45	25	50%

(그림 3) 수량적데이타를 이진 항목 형태로 매핑

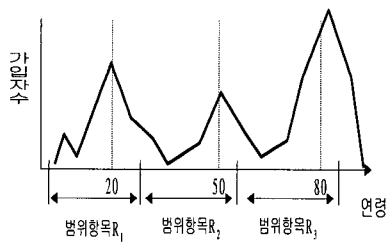
그러나 임의의 범위로 나누어 범위항목열로 생성하여 연관규칙탐사를 하는 데에는 두 가지 문제점을 갖고 있다. 첫째로 범위항목수가 많을 경우 수량적 데이타의 유일한 값의 도메인이 많을 경우와 같이 사용자가 정의한 임의의 최소지지율을 만족하는 항목을 찾을 수 없다. 둘째로 데이타를 일정한 간격으로 분할하여 범위항목을 생성하는 과정에서 사용자 최소지지율과 사용자 최소신뢰도를 만족하지 못하는 경우가 있다. 어떠한 규칙은 특정 범위항목에 변환하지 않은 상태에서는 사용자가 정의한 최소신뢰도를 만족하는 반면에 범위항목으로 변환하면서 최소신뢰도를 만족하지 못하는 경우가 발생한다. 이처럼 정보손실은 범위항목간격이 커질수록 늘어난다. 또한 수량적 연관규칙탐사도 데이타 마이닝 작업의 한 분야이기 때문에 작업의 분석 대상은 최소 수천 건 혹은 수만 건 이상의 대용량 데이타이다. 이처럼 대용량 데이타를 분석하기 때문에 수행복잡도가 지수형(exponential)이거나, 고차수의 다항식형(high-degree polynomial)으로 나타나는 알고리즘은 적용하기 곤란하다.

범위항목을 생성하기 위한 수량적 연관규칙탐사에 대한 연구[5]는 항목들을 일정한 소간격으로 분할하여 이웃하는 소간격을 병합하면서 사용자가 정의한 최소지지율을 만족할 경우 고빈도항목열로 본다. 그러나 이러한 방법은 분산이 잘 된 데이타에서는 잘 적용되지만 심하게 편향적인 데이타에 적용하는데 알맞지 않다. 또한 고정적인 일정한 간격으로 분할하기 때문에 불필요한 분할을 하는 단점을 가지고 있다. 또한 집중된 데이타와 분산된 데이타들간의 특성을 반영하지 않고 분할과 병합을 하기 때문에 집중된 데이타영역과 분산된 데이터영역이 서로 병합되어 데이타의 특성을 잃게 된다. (그림 4)는 수량적 데이타에서 연령층이 20대와 50대, 그리고 80대에서 가장 두드러지게 어떤 보험에 가입한 데이타에 대해서 수량적 데이타가 가지고 있는 특성을 잘 반영한 범위항목열을 생성한 예이다. 반대로 데이타의 특성을 반영하지 않고 범위항목열을 생성할 경우 아무리 사용자의 최소지지율을 만족하는 범위항목열을 생성하였다 하여도 중요한 정보를 손실하게 된다.



(그림 4) 데이타특성을 고려한 범위항목열생성

(그림 5)는 데이타가 갖고 있는 특성을 반영하지 않고 범위항목열을 생성한 경우이다. 데이타 특성을 무시한 형태의 범위항목열을 가지고 연관규칙탐사를 할 경우 데이타의 특성을 반영하지 않기 때문에 연령층이 20대와 50대, 그리고 80대에서 가장 두드러지게 어떤 보험에 가입한 정보를 무시한 결과를 가져온다. 이처럼 소간격으로 분할과 분할된 소간격을 병합할 때 어떠한 방법으로 분할·병합해야 하는가가 중요하다.



(그림 5) 데이타 특성을 고려하지 않은 범위항목열생성

4. 제안하는 고빈도항목열 생성기법

본 논문에서 제안하는 고빈도항목열 생성기법은 일정한 간격으로 분할하는 방법이 가지고 있는 문제점을 해결하기 위해서 데이터가 지닌 특성에 따라 유동적인 간격으로 분할하는 방법과 데이터의 특성을 최대한 반영하기 위해서 데이타의 집중도를 고려하여 병합하는 방법을 사용한다. 기존의 방법보다 세밀한 고빈도항목열을 생성할 수 있는 것과 데이타 클러스터링 효과[18]를 유도하여 데이타들의 특성을 잃어버리지 않는 특징을 갖는다.

4.1 유동적 분할

일정한 간격으로 분할하는 방법이 갖는 문제점은 수량적 데이타가 잘 분산되어 있으면 분할에 별문제가 되지 않지만 수량적 데이타의 범위가 넓으면서 한 두 군데로 집중되어 있으면 구간은 넓으면서 지지율이 낮

은 구간을 일정한 간격으로 분할하는 불필요한 분할을 한다는 것이다. 일정한 간격으로 분할하는 방법에서 생기는 불필요한 분할을 줄이기 위해서 본 논문에서는 일정한 간격으로 분할하는 방법이 아닌 데이터가 지닌 특성에 따라 유동적인 간격으로 분할하는 방법을 사용한다. 우선 전체 데이타를 한 덩어리로 보고 두 개로 분할하여 최소분할지지율 미만인 부분은 분할하지 않고 최소분할지지율 이상인 부분을 분할한다. 여기서 최소분할지지율은 어떤 구간을 분할을 할 것인지 하지 않을 것인지 구분하는 지지율이다. 최소분할지지율은 항상 사용자최소지지율 보다 적은 지지율을 설정한다. 분할과정은 모든 부분이 최소분할지지율 미만인 간격이 될 경우 끝낸다.

(그림 6)은 (그림 3)에 있는 수량적 데이타를 사용자 최소지지율은 30%이고 최소분할지지율을 10%로 설정하였을 경우에 유동적으로 분할하는 과정을 보여주고 있다. 우선 모든 데이타를 하나의 구간으로 보고 첫 번째 분할에서는 나이의 최소값 16과 최대값 42의 중간값 29를 기준으로 두 개의 구간으로 분할한다. 분할된 두 개의 구간 16~29와 30~42의 범위가 차지하는 지지율을 최소분할지지율과 비교하면 두 개구간 다 최소분할지지율 10% 이상이므로 다시 각 범위의 중간값을 기준으로 2차 분할한다. 계속해서 분할이 되다가 2차분할 중에 30~36의 구간처럼 최소분할지지율 10% 미만인 구간은 더 이상 분할하지 않는다. 이러한 분할과정을 통하여 모든 구간이 최소분할지지율인 10% 미만이면 분할을 끝낸다. 이처럼 분할간격을 최소분할지지율을 기준으로 분할하기 때문에 구간이 넓으면서 지지율이 낮은 구간에서 불필요한 분할을 하지 않는다. 그리고 분할과정에서 분할이 되는 구간은 분할된 횟수의 값을 유지한다. 분할횟수의 값은 나중에 병합할 때 유용한 정보가 된다. 마지막 5차분할에서 나타나는 분할횟수의 값을 보면 데이타가 집중된 곳은 분할횟수의 값이 4에서 5의 비교적 높은 값을 갖게 된다.

(그림 7)은 (그림 6)과 같이 수량적 데이타를 유동적으로 분할하는 과정을 보여준다. 분할을 지지율에 따라 분할하기 때문에 수량적 데이타가 유일한 도메인이 크게 구성되어 있다하여도 분할횟수는 어느 정도 일정하다. 일정한 구간으로 분할하는 방법에서는 도메인의 크기에 따라 분할 회수가 비례하게 된다. 반면에 유동적으로 분할하는 방법은 일정한 간격으로 분할하는 방법보다 분할횟수는 적고, 일정한 간격으로 분할할 경우에 생기는 데이타 특성을 반영하지 못하는 문제점을 해결한다. 그러나 매 분할마다 최소분할지지율과 비교해야 하는 단점이 있다.

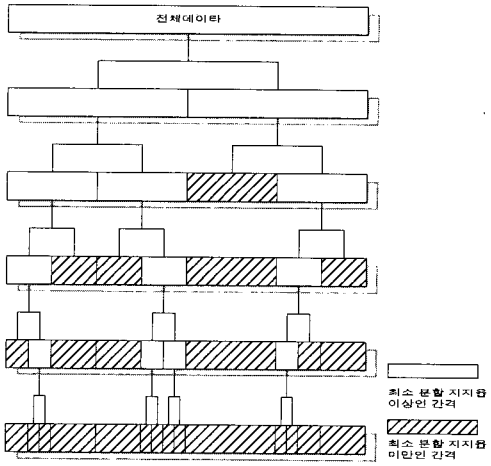
나이	지지수	지지율	분할횟수
1차분할			
16~29	23	46%	1
30~42	27	54%	1
2차분할			
16~22	5	10%	2
23~29	18	36%	2
30~36	2	4%	2
37~42	25	50%	2
3차분할			
16~19	3	6%	3
20~22	2	4%	3
23~26	16	32%	3
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~42	22	44%	3

(I)

나이	지지수	지지율	분할횟수
4차분할			
16~19	3	6%	3
20~22	2	4%	3
23~24	11	22%	4
25~26	5	10%	4
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~41	15	30%	4
42~42	7	14%	4
5차분할			
16~19	3	6%	3
20~22	2	4%	3
23~23	4	8%	5
24~24	7	14%	5
25~25	0	0%	5
26~26	5	10%	5
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~40	5	10%	5
41~41	10	20%	5
42~42	7	14%	4

(II)

(그림 6) 유동적인 분할과정



(그림 7) 유동적 간격으로 분할하는 과정에 대한 도식도

4.2 데이타 집중도를 고려한 병합

기존의 방법은 병합과정에서 어떤 특정한 우선 순위가 없이 병합하게 되는데, 이러한 병합방법은 데이타의 특성을 고려하지 않은 결과를 가져온다. 본 논문에서는 이러한 단점을 보완하기 위해서 유동적 소간격 분할방법을 통하여 생성된 소간격의 분할횟수와 각 구간이 갖는 지지율의 정보를 가지고 데이타의 집중도를 고려하여 병합하는 방법을 사용한다. 최소분할지지율을 임계치로하여 분할된 소간격을 병합함으로써 고빈도항목을 생성한다. 기존의 방법과 달리 이웃하는 소간격을 병합할 때에는 각 간격의 분할횟수와 병합되는 간격이 갖는 지지율이 높은 것부터 병합한다. 간격이 갖는 분할횟수와 지지율은 데이타의 집중도를 나타낸다. 집중된 간격부터 우선적으로 병합함으로써 데이타의 클러스터링효과를 유도하면서 병합횟수를 줄여 병합한다. 병합과정은 병합된 간격이 사용자가 정의한 최소지지율을 넘을 경우 고빈도항목으로 설정하고 그 구간은 더 이상 병합을 하지 않는다. 그리고 더 이상 병합할 소간격이 없을 때 병합과정을 마친다.

(그림 8)은 (그림 6)에서 최소분할지지율을 10%로 설정하여 유동적인 분할과정을 통하여 생성한 소간격을 병합하는 과정을 보여준다. 여기서 사용자최소지지율은 30%이다. 소간격을 병합할 경우 우선 분할횟수의 최대값인 5인 구간들을 우선 병합의 대상에 포함한다. 병합대상으로 분류된 구간들을 서로 이웃하는 구간끼리 두 개의 구간을 병합하여 지지율이 가장 높은 구간부터 병합을 한다. 병합된 구간은 하나의 구간으로 설정하고 병합되기 전의 두 개 구간의 분할횟수중에 큰 값에서 1을 뺀 값을 분할횟수값으로 유지한다. (그림 8)에서는 40~41의 구간이 가장 데이타 집중도가 높기 때문에 우선적으로 병합하고 분할횟수의 값은 5에서 1을 뺀 4가 된다. 그리고 그 다음으로 높은 지지율을 갖는 구간 23~24의 구간이 병합된다. 이처럼 병합을 계속하다가 더 이상 병합대상이 없을 경우 분할회수의 최대값이 5보다 하나 적은 4이상의 구간으로 병합대상을 확대한다. 그리고 병합과정에서 사용자최소지지율을 넘는 구간은 고빈도항목열로 보고 다음 병합단계부터는 병합대상에 포함시키지 않는다. 제1차 병합에서 구간 40~41이 사용자최소지지율 이상이므로 고빈도항목열로 처리되고 병합대상에서 제외된다. 이러한 과정을 통하여 최종적으로 사용자최소지지율 30%이상인 고빈도항목열 23~26과 40~41구간이 생성된다.

나이	지지수	지지율	분할횟수
유동적인 분할을 통해 생성된 소간격			
16~19	3	6%	3
20~22	2	4%	3
23~23	4	8%	5
24~24	7	14%	5
25~25	0	0%	5
26~26	5	10%	5
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~40	5	10%	5
41~41	10	20%	5
42~42	7	14%	4
제1차병합			
16~19	3	6%	3
20~22	2	4%	3
23~24	11	22%	4
25~26	5	10%	4
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~41	15	30%	4
42~42	7	14%	4
제2차병합			
16~19	3	6%	3
20~22	2	4%	3
23~26	16	32%	3
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~41	15	30%	4
42~42	7	14%	4

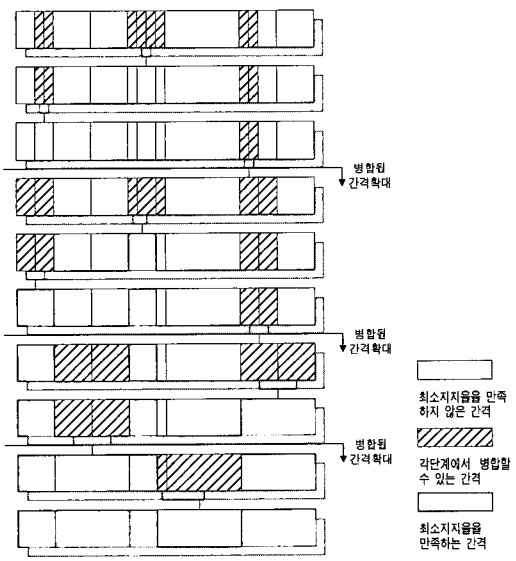
(I)

나이	지지수	지지율	분할횟수
제3차병합			
16~22	5	10%	2
23~26	16	32%	3
27~29	2	4%	3
30~36	2	4%	2
37~39	3	6%	3
40~41	15	30%	4
42~42	7	14%	4
제4차병합			
16~22	5	10%	2
23~26	16	32%	3
27~29	2	4%	3
30~39	5	10%	1
40~41	15	30%	4
42~42	7	14%	4
제5차병합			
16~22	5	10%	2
23~26	16	32%	3
27~39	7	14%	2
40~41	15	30%	4
42~42	7	14%	4

(II)

(그림 8) 데이타 집중도를 고려한 병합과정

(그림 9)는 데이터 집중도를 고려한 병합과정을 보여준다. 데이타 집중도를 고려한 병합의 특징은 우선 최대분할횟수와 지지율이 높을수록 데이타의 집중도가 크다. 이러한 정보를 이용하여 데이타의 집중도가 높은 것부터 병합함으로써 클러스터링효과를 갖는다. 또한 보다 세밀한 고빈도항목열을 생성한다.



(그림 9) 데이타의 집중도를 고려한 병합과정의 도식도

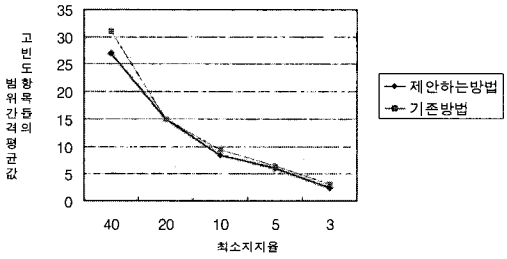
5. 성능평가

제안된 방법의 우수성을 보이기 위해 약 5만개의 레코드를 갖는 데이타 집합을 사용하여 성능평가를 수행한다. 이 데이터 화일에 있는 레코드들은 특정지역 사람들의 연령, 교육정보, 결혼유무, 자녀수 등과 같은 정보를 갖는다. 성능평가를 위해 다양한 정보중 수치적 항목인 그 지역에 사는 사람들의 연령항목을 이용하였다.

본 논문에서 제안한 기법을 이용한 고빈도항목열 생성은 일정한 간격으로 분할하여 이웃하는 간격을 병합하는 방법보다 데이타의 특성을 잘 반영하고 고빈도항목열로 생성된 범위항목들의 범위간격이 보다 세밀한 특징을 갖는다. 또한 분산된 소간격과 집중된 소간격간에 병합을 지양하기 때문에 데이타특성의 손실을 방지한다. 그러나 사용자가 정의한 최소지지율을 만족하는 고빈도항목열은 일정한 간격을 분할하여 이웃하는 간격을 병합하는 방법보다 적게 생성되는 경우가 있다. 적게 생성되는 이유는 편중되어 있지 않는 분산된 데이타를 병합의 대상에서 가장 나중에 고려하는데 분산된 데이타가 전체 데이타의 사이사이에 존재할 수 있기 때문이다. 그러나 이러한 데이타를 이웃하는 소간격을 우선 순위 없이 병합할 경우 데이타의 특성을 잃게된다. 일정한 간격으로 분할하는 방법은 수량적 데이타가 갖는 범위에 따라 분할간격이 변하게 된다. 수량적 데이타가 갖는 범위가 클수록 분할간격이 커진다. 그러나 최소분할지지율을 임계치로 분할하는 유동적인 분할방법은 수량적 데이타의 특성에 따라 분할간격이 다르게 설정되고 수량적 범위에 관계없이 분할된다. 병합측면에서도 이웃하는 분할간격을 우선 순위를 고려하지 않고 병합하는 방법은 분할된 간격이 갖는 특성을 고려하지 않고 병합하기 때문에 지지율이 낮은 분할간격과 높은 분할간격이 서로 병합되어 데이타 특성을 잃게 되고 고빈도항목열로 생성되는 범위항목의 간격이 크다. 그래서 병합과정에서 데이타의 집중도 따라 우선순위를 주어 병합함으로써 데이타의 특성을 유지하면서 고빈도항목열로 생성되는 범위항목의 간격을 보다 세밀하게 병합한다.

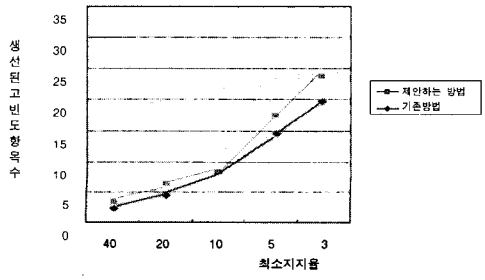
(그림 10)은 고빈도항목열로 생성된 범위항목의 간격평균값을 일정한 간격으로 분할하여 이웃하는 소간격을 병합하는 방법과 최소분할지지율을 이용하여 분할하고 우선 순위를 주어 병합하는 방법을 서로 비교한 것이다. 사용자 최소지지율이 40%인 경우 기존방법

에서는 생성된 고빈도항목열의 범위간격의 평균값은 32인 반면에 제안하는 방법은 27이다. 이처럼 제안하는 방법은 기존의 방법보다 세밀한 간격을 생성한다.



(그림 10) 생성된 고빈도항목의 세밀도 비교

본 논문에서는 최소지지율을 만족하면서 많은 고빈도항목열을 생성하고 고빈도항목열의 범위간격이 되도록 적게 하는데 목적이 있다. 범위간격이 넓을수록 그 고빈도항목은 데이타의 중요도가 적어진다. 예를 들어 전체 데이타중에서 최소지지율 40%와 최소신뢰도 60%를 만족하는 규칙을 탐사할 경우 만족하는 규칙이 나이가 20세에서 50세까지인 사람은 자동차를 1대이상 가지고 있다는 정보와 20세에서 30세까지인 사람이 자동자를 1대이상 가지고 있다는 정보는 후자가 더 중요한 정보가 될 것이다. 또한 최소지지율의 변화에 따라 생성된 고빈도항목수를 비교해보면 (그림 11)과 같이 최소지지율이 낮을수록 제안하는 방법이 생성하는 수가 많다. 그 이유는 생성되는 간격이 기존방법보다 더 세밀하기 때문이다. 제안하는 방법이 세밀하게 생성하지만 우선 순위에 맞추어 병합하기 때문에 병합대상에서 우선 순위가 높은 간격과 낮은 간격이 서로 심하게 섞여 있으면 고빈도항목열 생성과정에서 우선 순위가 낮은 간격은 고빈도항목열로 생성되지 못하게 된다. 고빈도항목열로 고려되지 않는 간격이 많으면 고빈도항목수는 기존방법보다 적게 생성되게 된다.



(그림 11) 생성된 고빈도항목수 비교

6. 결론 및 향후방향

본 논문에서는 수량적 데이타를 포함하는 대용량의 관계형 테이블에 대한 연관규칙탐사의 고빈도항목열 생성과정에서 생기는 문제점을 제시하고 이를 해결하는 효율적인 고빈도항목열 생성기법을 제안하였다. 기존방법을 이용한 고빈도항목열 생성방법에는 데이타들이 심하게 편중되어 있으면 불필요한 분할이 발생하는 문제점과 분할·병합과정에서 데이타의 특성을 고려하지 않는 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해서 분할과정에서 최소분할지지율을 이용하여 분할간격을 유동적으로 분할하는 방법과 분할된 소간격의 데이타의 집중도를 우선 순위로 하여 데이타 특성을 고려한 병합방법을 사용하여 고빈도항목열을 생성하였다.

제안하는 방법의 특징은 보다 세밀한 간격의 데이타를 생성하는 것과 병합과정에서 데이타가 가지고 있는 특성을 잃어버리는 것을 최소화하였다는 것이다. 향후 연구방향으로는 단일항목만을 고려하여 고빈도항목열을 생성하였던 방법을 여러 개의 수량적 데이타를 고려하여 고빈도항목열를 효율적으로 생성할 수 있도록 확장하는 것이다.

참 고 문 헌

[1] R. Bayardo, "Efficiently Mining Long Patterns from Databases," To appear in Proc. of the ACM SIGMOD Conference on Management of Data, Seattle, Washington, June 1998.

[2] R. Srikant, Q. Vu, R. Agrawal, "Mining Association Rules with Item Constraints," Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

[3] K. Ali, S. Manganaris, R. Srikant, "Partial Classification using Association Rules," Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

[4] D. Gunopulos, R. Khardon, H. Mannila, H. Toivonen, "Data mining, Hypergraph Transversals, and Machine Learning," 16th ACM Symp.

on Principles of Database Systems (PODS), Tuscon, AZ, 1997.

[5] R. Srikant, R. Agrawal : "Mining Quantitative Association Rules in Large Relational Tables," Proc. of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996.

[6] R. Agrawal, J.C. Shafer, "Parallel Mining of Association Rules : Design, Implementation and Experience," IBM Research Report RJ 10004, January 1996. IEEE Transactions on Knowledge and Data Engineering, Vol.8, No.6, December 1996.

[7] R. Srikant, R. Agrawal : "Mining Generalized Association Rules," Proc. of the 21st Int'l Conference on Very Large Databases, Zurich, Switzerland, Sep. 1995.

[8] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen and A. I. Verkamo, "Fast Discovery of Association Rules," Advances in Knowledge Discovery and Data Mining, Chapter 12, AAAI/MIT Press, 1995.

[9] R. Agrawal, R. Srikant : "Fast Algorithms for Mining Association Rules," Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, Sept. 1994.

[10] Expanded version available as IBM Research Report RJ9839, June 1994.

[11] R. Agrawal, T. Imielinski, A. Swami : "Mining Associations between Sets of Items in Massive Databases," Proc. of the ACM SIGMOD Int'l Conference on Management of Data, Washington D.C., May 1993, 207-216.

[12] 이도헌, "데이타 마이닝 : 개념 및 연구동향", 데이타베이스연구회지 13권 4호, 1997

[13] 오명우, 박지웅, 한기준, "GIS데이타베이스를 위한 공간 데이타 마이닝", 데이타베이스연구회지 13권 4호, 1997

[14] 나민영, 최병갑, "데이타베이스 마이닝을 위한 지식기반 트리분류기", 데이타베이스 연구회지 12권 4호, 1996

[15] 윤유경, 윤종필, "대용량 데이타베이스에서 시간흐름에 따른 연관 항목 집합의 변화 탐색", 데이타베이스 연구회지 12권 4호, 1996

[16] 박종수, "대용량 데이타베이스상의 효과적인 관련 규칙 탐사를 위한 데이타 전지기법", 데이타베이스 연구회지 12권 4호, 1996

[17] 김의경, 이도헌, 김명호, 이윤준, "대용량 데이타베이스상의 단반향 수량 연관규칙탐사", 정보과학회 논문지(B) 24권 4호, 1997.4

[18] 김주형, 홍봉희, "효율적인 공간 질의 처리를 위한 Quadtree 클러스터링 알고리즘", 정보과학회논문지(B) 25권 1호, 1998.1

최 영 희

email : ironman@soback.kornet.nm.kr

1983년 광운대학교 전자공학과 졸업(공학사)

1986년 광운대학교 대학원 컴퓨터공학과 졸업(공학석사)

1987년~현재 호원대학교 전기전자정보공학부 교수

관심분야 : 데이터마이닝, 전자상거래, 멀티미디어 데이터베이스, 분산객체컴퓨팅 등

장 수 민

email : jsm@pretty.chungbuk.ac.kr

1997년 목포대학교 전산통계학과 졸업(이학사)

1999년 충북대학교 정보통신공학과졸업(공학석사)

관심분야 : 데이터마이닝, 실시간 데이타베이스, 네트워크게임, 전자상거래

유 재 수

e-mail : yjs@cbucc.chungbuk.ac.kr

1989년 전북대학교 컴퓨터공학과 졸업(학사)

1991년 한국과학기술원 전산학과 졸업(공학석사)

1995년 한국과학기술원 전산학과 졸업(공학박사)

1995년 3월~1996년 8월 목포대학교 전산통계학과 전임강사

1996년 9월~현재 충북대학교 전기전자공학부 조교수

관심분야 : 데이터베이스시스템, 멀티미디어 데이터베이스시스템, 정보검색, 분산객체컴퓨터

오 재 철

e-mail : ojc@sunchon.sunchon.ac.kr

1978년 전북대학교 전기공학과(공학사)

1982년 전북대학교 대학원 전기공학과(공학석사)

1988년 전북대학교 대학원 전기공학과(공학박사)

1984년~1986년 기전여자대학 전산과 전임강사

1986년~현재 순천대학교 컴퓨터과학과 교수

관심분야 : 지식탐사, 분산처리, 멀티미디어 데이터베이스 등