수량 연관규칙 탐사를 위한 빈발구간 항목집합 생성방법[1)]

박원환*, 유기형**, 박두순*, 손진곤***

* 순천향대학교 정보기술공학부
** 통계청 통계정보국 전산개발과
*** 방송통신대학교 전산학과

A Large-Interval Itemsets Generation Method for Mining Quantitative Association Rules

Won-Hwan Park*, Gi-hyong Ryu**, Doo-Soon Park*, Jin-Gon Shon***

* Division of Information and Technology, Soonchunhyang University
** System Development Division, National Statistical Office
*** Division of Computer Science, Korea National Open University

요 약

대용량의 데이터베이스로부터 연관규칙을 발견하고자 하는 연구가 활발하며, 수량 데이터의 항목에도 적용할 수 있도록 이들 방법을 확장하는 연구가 최근에 소개되고 있다. 본 논문에서는 수량 데이터 항목을 이진 항목으로 변환하기 위하여 빈발구간 항목집합을 생성할 때, 수량 데이터 항목의 정의 영역 내에서 특정 영역에 집중하여 발생하는 특성인 지역성을 이용하는 방법을 제안한다. 이 방법은 기존의 방법보다 많은 수의 세밀한 빈발구간 항목들을 생성할 수 있을 뿐만 아니라 세밀의 정도를 판단하여 활용할 수 있는 생성순서 정보도 포함하고 있어, 원 데이터가 가지고 있는 특성의 손실을 최소화할 수 있는 특징이 있다. 성능평가를 통하여 기존의 방법보다 우수함을 보였다.

1. 서 론

대용량화된 데이터베이스에는 사용자가 미처 파악하지 못하는 중요한 정보 또는 지식이 포함되어 있을 수 있으나, 기본적으로 데이터베이스는 일정한 형태의 질의에 빠른 응답을 위한 시스템이므로 실세계(real world)에서 나타나는 다양한 규칙성(regularity)을 발견하기에는 한계가 있다.

데이터베이스 모델링에 반영되지 못하여 감추어져 있는 규칙성을 발견하는 데이터 마이닝(data mining)에

1) 본 연구는 정보통신부의 ITRC 사업에 의해 수행된 것임

대한 연구가 최근에 활발히 진행되고 있다. 이들의 한 분야인 연관규칙(association rules) 탐사에 관한 문제는 [1]에서 처음 제안된 이후 수많은 연구가 있었고, 최근에도 이 분야의 연구가 활발하다[2, 3].

연관규칙 탐사에는 항목의 발생 유무의 정보로 탐사하는 이진 연관규칙(binary association rules) 탐사와 항목의 수량도 고려하여 탐사하는 수량 연관규칙(quantitative association rules) 탐사로 구분된다. 과거에는 이진 연관규칙 탐사에 대한 연구가 많았다. 그러나 최근에는 이들 방법들을 수량항목에 확장하는 연구도 소개되고 있다[4, 5].

본 논문에서는 수량 연관규칙을 탐사하기 위하여 수량 데이터 항목의 정의영역, 즉 도메인을 여러 개의 빈발구간 항목집합[2]으로 변환할 때, 데이터 발생의 지역성(locality)을 고려하는 방법을 제안한다.

본 논문의 구성은 제2장에서 연관규칙탐사의 정의와 수량 데이터 항목 대한 빈발구간 항목집합 생성하는 기존의 방법에 대하여 알아본다. 3장에서는 빈발구간 항목집합을 생성하는 새로운 방법을 제안하고, 그 예를 보인다. 4장에서는 성능을 평가하고, 마지막으로 결론 및 향후과제를 밝힌다.

2. 연관 규칙 탐사

2.1 연관 규칙 정의

항목들의 집합 $I = \{i_1, i_2, i_3, \cdots, i_m\}$이라 하면, 트랜잭션 T는 I의 부분집합이다. 데이터베이스 D에는 n개의 T가 저장되어 있다. 여기서 연관규칙을 탐사한다고 하자.

항목들의 집합 X, Y에 대한 연관규칙 탐사의 결과는 사용자에 의하여 주어지는 최소지지도(minimum support, S_{min}) 이상의 지지도와 최소신뢰도(minimum confidence, C_{min}) 이상의 신뢰도를 갖고, 다음의 성질을 갖는 연관규칙 $R : X \rightarrow Y$의 집합이다.

- $X \subseteq I$인 X에 대해, $X \subseteq T$이면, T는 X를 만족한다고 정의하고, D에서 X를 만족하는 트랜잭션의 수를 $freq(x)$로 표기한다.
- $X, Y \subseteq I$이고, $X \cup Y = \phi$ 이면, 규칙 $X \rightarrow Y$의 지지도(S) = $freq(X \cup Y)/n$ 과 신뢰도 $C = freq(X \cup Y)/freq(X)$를 갖는다.
- 최소지지도(S_{min}) 이상을 갖는 항목열 $X \subseteq I$를 빈발항목집합(large itemset)이라 정의한다. 이 때 X의 부분집합도 빈발항목집합이다.

이상과 같은 연관규칙의 탐사는 기본적으로 2단계로 이루어진다.

- 단계 1 : 빈발 항목집합(large itemsets)을 찾는 단계
- 단계 2 : [단계1]에서 생성된 빈발 항목집합을 사용하여 연관규칙을 생성하는 단계

연관규칙 탐사 과정의 성능은 단계1에서 결정되며, 1단계는 데이터베이스의 고려대상 항목 수의 증가에 따라 처리시간과 메모리 요구는 기하급수적으로 증가하기 때문이다. DHP[2], Sampling Approach[3] 등 연관규칙 탐사 알고리즘들 대부분이 이러한 문제의 해결에 중점을 두고 있다.

2.2 기존의 빈발구간 항목집합 생성방법

수량 데이터 항목을 이진 항목형태로 변환하여 기존의 탐사 알고리즘으로 연관규칙탐사가 가능하므로 수량 항목을 이진 형태로 변환하는 기존의 연구[4, 5]가 있었다.

Skriant[4]는 수량항목의 도메인을 일정한 범위의 소구간으로 일괄분할(partition)한 후, 이웃한 소구간 분할을 병합(merge)하여 최소지지도를 만족하는 빈발구간 항목집합을 생성한다.

이 경우에는 수량항목의 정의 영역에 데이터가 골고루 분포된 경우에는 효과적이지만, 일부 영역에 집중된 경우는 비효율적인 면이 있으므로 이의 해결 방법으로 분포도에 따라 분할하는 유동적 분할법[5]이 발표되었다.

이들 방법은 2 단계의 절차를 거쳐 최소지지도를 만족하는 빈발구간 항목집합을 생성한다. 첫째 단계에서는 소간격 분할을 생성하며, 두 번째 단계에서는 최소지지도를 만족할 때까지 이웃 소구간을 병합한다. 병합에 사용되는 기준은 일정범위 분할법에서는 최소지지도만 사용하고, 유동적 분할법

2) 수량항목의 빈발항목집합(large itemsets)을 이진항목의 경우와 구분하기 위하여 빈발구간 항목집합(large interval itemsets)이라 정의한다.

에서는 분할에 최소분할지지도, 합병에 최소지지도를 사용하여 분할 및 병합을 실시한다.

따라서 일정범위 분할법은 데이터의 분포를 고려한지 못하는 점이 약점이고, 유동분할법은 최소분할지지도라는 또 다른 분할기준을 사용함으로써 분할을 위한 부수적인 비교가 필요할 뿐만 아니라 사용자가 임의로 최소분할지지도을 설정해야하므로 최적의 기준설정이 어렵다는 문제점이 있다.

3. 빈발구간 항목집합 생성방법

3.1 데이터 발생의 지역성

수량 항목은 그림1와 같이 다양한 유형으로 나타나고 있으며, 이들 항목들의 정의 영역(도메인) 또한 좁은 영역, 넓은 영역 등 다양하다.

- 나 이 : □□□세
- 통근통학 소요시간 : □시 □□분
- 주택의 연건평 : □□□평(㎡) 등

그림1. 인구주택총조사 항목들 중 수량항목(예)

이와 같은 수량 항목은 적정 간격의 빈발 구간 항목으로 변환하여 연관규칙을 탐사한다. 이 때 빈발구간 항목을 판단하는 기준인 최소지지도는 해당 단위구간 항목의 빈도(frequency)에 따라 결정되므로. 실세계(real world)의 데이터는 그림2와 같이 빈도가 높은 영역의 주변에 치우치는 특성이 있으므로 이를 이용할 수 있다. 그림2는 그 예이며, (a)는 인구주택총조사의 대전지역 나이별 인구분포, (b)주택의 연건평별 가구분포를 나타낸다.

이와 같은 특성은 센서스 데이터 외에도 관심 지역, 연령 등에 따라 나타날 수 있으므로 음반판매, 영화관 관람자 등과 같은 데이터에서도 나타날 수 있다.

이 특성을 데이터 발생의 지역성(locality)라 정의하며, 이의 존재는 특정구간인 최빈수(mode) 단위구간을 기준으로 좌-우로 확장하여 고려대상 구간을 설정하였을 때, 고려대상 구간이 차지하는 비율(A)과 빈도가 차지하는 비율(B)의 비, 즉 B/A의 값으로 판단한다.

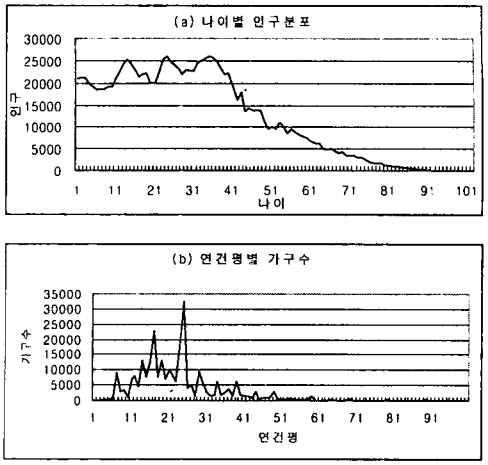

그림2. 데이터 발생의 지역성(예)

3.2 빈발구간 항목 생성방법

본 논문에서 제안하는 방법은 최빈수(mode)의 단위구간[3]을 중심으로 수량항목의 정의영역을 이진항목으로 변환하는 방법이다.

3.2.1 기호설명

데이터 발생의 지역성을 고려한 빈발구간 항목집합을 생성하기 위하여 필요한 데이터베이스(D), 수량항목(Lq), 발생빈도(f(Lq)) 등의 기호를 다음과 같이 정의한다.

- D는 유한한 범위의 수량 항목이 포함된 트랜잭션들의 집합으로 Lq, f(Lq)를 포함한다.
- $Lq = \{ lq_1, lq_2, \ldots, lq_{n-1}, lq_n \}$이며, lq_i $(1 \le i \le n)$는 단위구간 항목(item)으로 이산(discrete)적 이다.
- $f(Lq) = \{f(lq_1), f(lq_2), \ldots, f(lq_{n-1}), f(lq_n)\}$이며, $f(lq_i)$ $(1 \le i \le n)$는 단위구간에서 데이터 발생빈도이다.
- Max_lq는 최빈수(mode)의 단위구간의 빈도, 또는 병합구간의 빈도이다.
- F_L = 빈발구간 항목집합(large-interval itemsets)이며, m개의 원소들$\{fl_1, fl_2, \ldots, fl_m\}$로 구성되고, fl_i,(i=1, …, m)는 S_{min}을 만족한다.
- $lq_l_i(1 \le i \le n)$는 해당 단위구간 lq_i의 사용 가능여부를 표기한다. □

3) 수량항목 정의영역의 기본단위를 단위구간으로 정의한다.

3.2.2 빈발구간 항목집합 생성 알고리즘

데이터 발생의 빈도가 가장 높은 영역이 최빈수(mode)의 단위구간이다. 우수한 결과를 얻을 수 있는 방법으로 이 단위구간을 기준으로 빈발구간 항목집합 생성하는 방법을 제안한다.

빈발구간 항목집합 생성방법은 1차 최빈수의 단위구간(lq_i)을 선택한 후, 이를 기준으로 인접(좌-우) 단위구간(lq_{i-1}, lq_{i+1})을 최소지지도를 만족할 때까지 병합한다. 이때 좌-우 항목의 값(빈도 또는 지지도) 중에서 최소지지도 보다 높거나 같으면서 가장 근접하는 값을 취하여 최소지지도를 만족할 때까지 병합한다. 만약 하한한계 또는 상한한계[4]로 인하여 더 이상 진행을 할 수 없을 경우는 한 쪽 값만을 취하여 병합을 진행한다. 진행도중에 양쪽(좌-우) 모두 한계에 도달하면 비빈발이므로 빈발하지 않은 구간으로 설정하고, 다음 최빈수의 단위구간을 선정하여 동일하게 진행한다.

인접 단위구간을 병합하는 도중에 최소지지도를 만족하면 병합을 중단하고, 빈발구간 항목집합에 포함시키고, 동시에 빈발구간 항목영역으로 설정한다. 계속하여 잔여 단위구간 중에서 최빈수의 단위구간을 선택하여 동일한 방법으로 빈발구간 항목을 생성하며, 더 이상 단위구간이 존재하지 않을 때 중단한다.

그림3, 4는 이러한 절차를 코드로 표기한 것이다. 그림3은 최빈수의 단위구간을 선정하여 Gen-F_L함수에 그 값을 전달한다. 그림4에서는 전달된 단위구간을 기준으로 최소지지도(S_{min})를 만족할 때까지 좌-우의 단위구간을 병합하는 절차를 수행한다. 이때 상한과 하한 경계의 인접 여부에 따라 4가지 경우(case)를 고려하고 있다.

최빈수의 단위구간을 기준으로 생성된 빈발구간 항목은 그림5에서 보듯이 1차 최빈수를 기준한 항목의 구간영역이 k차 최빈수를 기준한 항목의 구간영역보다 좁거나 같다. 즉, 1차에서 k차로 진행될수록 빈발구간 항목의 폭이 넓어지는 특징이 있다. 이는 초기에 생성된 빈발구간 항목이 나중에 생성된 빈발구간 항목 보다 데이터 자체가 가지고

4) 상한과 하한 한계는 하한(lq_1) 및 상한(lq_n) 경계 또는 이미 사용한 영역의 단위구간 경계이다.

```
// 최소지지도(Smin)는 사용자가 지정
// DB를 검색하여 f(lq)를 생성
 FL = ψ

for (k=1 ; Lq≠ψ; k++) do begin
    Max_lq = MAX(f(lqi)), (not tagged, 1≤i≤n)
    flk merge lqi ;
    CALL Gen_FL
    FL = U flk                    // Answer
    Max_lq = 0
  // if sum of lqs between tagged < Smin
     then not large quantitative itemsets
    lqi = used tag (1≤i≤n)  // not large

end
```

그림3. 제안 빈발항목생성 알고리즘

```
Function Gen_FL

 for ( j=1 ; Max_lq ≥ Smin, j++ ) do begin

   case 1 : lq_ti-j, and lq_ti+j are not tagged

     if (f(lqi-j)+f(lqi+j)) ≤ (Smin-Max_lq)
        then Max_lq = Max_lq + f(lqi-j) +f(lqi+j);
             flk merge lqi-j, lqi+j ; lq_ti-j, lq_ti+j = tag
        else if (f(lqi-j)≤f(lqi+j) and
                 (Smin)-Max_lq)≤f(lqi-j))
                 then Max_lq = Max_lq + f(lqi-j);
                      flk merge lqi-j; lq_ti-j = tag
                 else Max_lq = Max_lq + f(lqi+j);
                      flk merge lqi+j; lq_ti+j = tag
              endif
     endif

   case 2 : lq_ti-j is not tagged, lq_ti+j tagged

            Max_lq = Max_lq + f(lqi-j) ;
            flk merge lqi-j; lq_ti-j = tag

   case 3 : lq_ti-j is tagged, lq_ti+j not tagged

            Max_lq = Max_lq + f(lqi+j) ;
            flk merge lqi+j; lq_ti+j = tag

   case 4 : lq_ti-j and lq_ti+j is tagged
            return          // flk is not large
 end

Return
```

그림4. 단위구간 병합함수

있는 특성의 손실이 적은 세밀한 빈발구간 항목임을 의미한다. 그러므로 생성되는 빈발구간 항목의 생성순서는 향후 규칙(rules)을 생성할 때 좋은 정보로 활용될 수 있다.

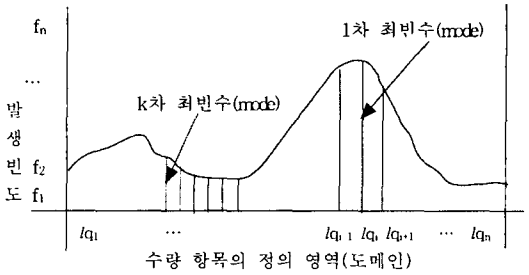

그림5. 최빈수를 이용하는 방법

나이	지지수(빈도)	단위구간 지지도(%)	최빈수 차수	병합구간 지지도	빈발유무
10	11	1.1		3.4	비빈발
11	23	2.3	10차		
12	25	2.5		11.4	빈발
13	28	2.8			
14	30	3.0			
15	31	3.1	8차		
16	33	3.3		10.8	빈발
17	38	3.8	5차		
18	37	3.7			
19	33	3.3	6차	12.8	빈발
20	32	3.2			
21	32	3.2			
22	31	3.1			
23	32	3.2	7차	11.9	빈발
24	24	2.4			
25	23	2.3			
26	19	1.9			
27	21	2.1			
28	14	1.4	11차	1.4	비빈발
29	21	2.1		10.1	빈발
30	39	3.9			
31	41	4.1	3차		
32	54	5.4		11.3	빈발
33	59	5.9	2차		
34	61	6.1	1차	10.9	빈발
35	48	4.8			
35	39	3.9	4차	11.0	빈발
37	35	3.5			
38	36	3.6			
39	28	2.8	9차	5.0	비빈발
40	22	2.2			

표1. 빈발구간 항목 생성과정(예)

데이터 발생의 지역성을 이용하여 빈발구간 항목집합을 생성하는 방법 진행절차의 예는 표1과 같으며, 10세부터 40세까지의 31개 단위구간에 대하여 데이터 1000건, 최소지지도 10%를 사용하였다.

그의 결과는 빈발구간 항목은 8개, 비빈발 항목은 3개가 생성되었다. 예에서 1차, 2차로 생성된 빈발 항목은 2개의 단위구간이 병합되었지만, 7차는 5개, 8차는 4개의 단위구간이 병합되는 것을 알 수 있다.

4. 성능평가

본 논문에서 제안한 데이터 발생의 지역성을 고려하여 빈발구간 항목집합을 생성하는 방법의 우수성을 보이기 위하여 3가지 방법, 즉 일정범위 분할 및 병합방법(M1), 유동적 분할 및 병합방법(M2) 그리고 제안한 방법(M3)을 사용하여 생성되는 빈발구간 항목수와 생성된 빈발구간 항목들의 구간 평균간격을 비교한다.

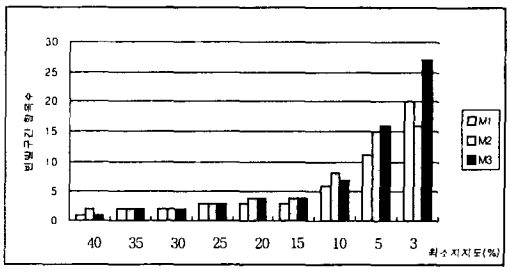

그림6. 생성 빈발구간 항목 수

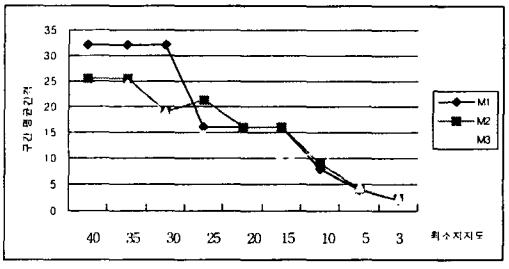

그림7. 생성 빈발구간 항목들의 평균간격

성능평가는 인구주택총조사 중 대전광역시의 연령별 인구 데이터(1,2143,327 레코드)에 대하여 최소지지도는 9가지(40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, 3%)를 사용하였다. 또한 M1에 대해서는 2개 구간을 일정하게 병합하였고, M2의 최소분할지지도는

최소지지도의 1/2로 하였다.

각 방법의 성능평가 결과는 그림6, 7과 같다. 그림6을 살펴보면, 최소지지도 35%, 30%, 25%에서 동일한 수의 빈발구간 항목을 생성하고 있다. 40%에서는 M3가 적은 수를 생성하지만, 5%이하에서는 오히려 많은 수를 생성한다. 그림7의 평균 간격을 살펴보면, 최소지지도 10% 이하에서는 대등하지만 15% 이상에서는 M3가 간격이 좁게 나타나고 있다.

이상과 같이 생성 빈발구간 항목 수와 구간의 평균간격을 각각을 비교하여 보았다. 생성되는 빈발구간 항목 수는 제안하는 방법의 성능이 낮은 최소지지도에서는 우수한 반면, 높은 최소지지도에서는 다소 떨어지고 있다. 그러나 빈발구간 항목수와 평균간격을 종합하여 비교하여 보면 높은 최소지지도인 40%에서 생성되는 항목은 적지만 평균 간격이 영역의 좁게 나타나고 있고, 35%~15%에서는 생성되는 항목수는 비슷하지만 평균간격이 좁게 나타나고 있다. 10%이하에서는 평균간격은 비슷하나 생성 항목수는 많다. 따라서 제안한 방법이 기존의 방법보다 세밀한 빈발구간항목집합을 생성하고 있으므로 성능의 우수함을 보여준다.

5. 결론 및 향후과제

본 논문에서는 수량 항목을 포함하는 대용량의 데이터베이스에서의 연관규칙의 탐사를 위해 수량 항목의 정의영역을 빈발구간 항목으로 분할하여 이진항목으로 변환하는 보다 효율적인 방법을 제안하였다. 그리고 실제 데이터인 인구주택총조사 데이터를 사용하여 성능평가를 실시하여 제안한 방법이 기존의 방법보다 보다 세밀한 빈발구간 항목집합을 생성할 수 있음을 알았다.

제안한 방법은 탐사 대상 데이터가 가지는 특성, 즉 데이터 발생의 지역성을 고려하는 방법으로 최빈수를 빈발구간 항목생성을 위하여 사용하였다. 최빈수를 사용함으로써 얻는 효과로는 보다 세밀한 빈발구간 항목을 생성함은 물론 그림5에서와 같이 최빈수의 차수, 즉 생성순서에 따라 빈발구간 항목의 세밀도가 감소하는 특징이 있다. 이는 생성되는 빈발구간 항목의 순서에 따라 원 데이터가 가지고 있는 특성의 손실 정도가 다르다는 것이다. 즉 초기에 생성된 것이 나중에 생성된 것 보다 손실이 적은 특징을 가진다. 이 특징은 향후 연관규칙을 탐사할 때, 필요한 규칙의 질(quality)의 정도에 따라 사용자가 빈발구간 항목의 생성순서를 감안하여 활용유무를 조정할 수 있다. 다만, 이 때 고려하여야 할 사항은 최빈수가 지역성이 없는 영역에 존재하는 특이한 분포의 데이터일 경우는 초기에 생성된 빈발구간 항목이 앞서 언급한 특징을 갖지 못할 수 있다는 것을 감안하여야 한다.

따라서 향후의 연구 과제로 앞에서 언급한 바와 같이 빈발구간 항목의 생성순서를 연관규칙의 탐사에 효과적으로 활용하는 방법에 대한 연구이다.

[참고문헌]

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", *In Proc. of the ACM SIGMOD Conference on Management Data*, pp. 207-216, 1993.

[2] J.S. park, M.S. Chen and P.S. Yu, "An Effective hase-based algorithm for mining association rules", *In Proceedings of ACM SIGMOD Conference on Management of Data*, pp. 175-186, May 1995.

[3] M.J. Zaki, S. Parthasarathy, Wei Li, and M. Ogihara," Evaluation of Sampling for Data Mining of Association Rules", The Univ. Rochester Technical Report 617, May 1996.

[4] R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of the ACM SIGMOD Conference on Management of Data*, June 1996.

[5] 최영희, 장수민, 유재수, 오재철,"수량적 연관규칙 탐사를 위한 효율적인 고빈도항목열 생성기법", 한국정보처리학회 논문지 제6권 제10호, pp2597 - 2607, 1999.