수량 연관규칙 생성을 위한 데이터의 지역성을 고려한 효과적인 알고리즘 제안

이혜정*, 박원환**, 박두순*
* 순천향대학교 정보기술공학부
** 한국통계청

An Efficient Algorithm Using the locality of Data for Mining Quantitative Association Rules

Hye-Jung Lee*, Won-Whan Park**, Doo-Soon Park*
* Division. of Computer Science and Computer Engineering, SoonChunHyang University
** Korea National Statistical Office, Building 3 Government Complex,

요 약

최근 대용량의 데이터베이스로부터 연관규칙을 발견하여 이를 활용하는 단계에서 이러한 연관규칙을 수량항목에도 적용할 수 있도록 확장하는 연구가 소개되고 있다. 본 논문에서는 수량항목을 이진항목으로 변환하기 위하여 빈발구간 항목집합(Large Interval Itemsets)을 생성할 때 수량 항목이 특정 영역에 집중하여 발생하거나 골고루 분포되어 있지 않은 경우, 이러한 지역성(locality)을 고려하여 빈발구간 항목집합을 생성하는 방법을 제안한다. 이 방법은 기존의 방법보다 많은 수의 세밀한 빈발구간 항목들을 생성할 수 있을 뿐만 아니라 의미 있는 구간을 중심으로 빈발구간 항목들이 순서대로 생성되기 때문에 세밀도를 판단하여 활용할 수 있으며, 원 데이터가 가지고 있는 특성의 손실을 최소화할 수 있는 특징이 있다. 또한 인구센서스등 실 데이터를 사용한 성능평가를 통하여 기존의 방법보다 우수함을 보였다.

1. 서론

대용량의 데이터베이스에 내재된 유용한 지식을 탐사하는 기술을 데이터마이닝(data mining)이라 하며, 데이터마이닝은 크게 예측(prediction)과 지식탐사(knowledge discovery)로 구분한다[1]. 연관규칙(association rules) 은 지식탐사의 범주에 해당되며 이에 관한 문제는 Agrawal[2]이 처음 제안한 이후 수많은 연구가 있었고, 최근에도 이 분야의 연구는 활발하게 진행되고 있다. 현실세계에서 발생하는 트랜잭션은 그 특성을 나타내는 항목집합(itemset)의 단위로 데이터베이스에 기록되어 대용량화된다. 여기에 자주 발생하는 항목집합들 간의 상호관련성을 발견하는 작업을 연관규칙(association rule)이라 한다. 센서스 데이터와 같이 수량 속성이 강한 자료의 경우에도 이진 연관규칙 탐사가 가능하다. 그러나 연령, 자녀 수 등 센서스 데이터에 포함된 많은 수량 속성을 무시하고 연관규칙을 탐사한다는 것은 발견할 수 있는 규칙이 극히 제한적이거나 불가능하다. 반대로 수량항목의 정의영역 내의 각 수치를 그대로 항목으로 설정하여 빈발항목집합을 생성한다면 탐사공간이 너무 넓고, 발생하는 항목이 분산되어 있어 탐사가 불가능하다. 이와 같은 문제는 수량 항목의 정의영역을 여러 개의 소구간으로 분할하여 최소지지도를 만족할 때까지 병합하여 빈발구간 항목집합을 생성한다[7][9]. 빈발구간 항목은 최소지지도를 만족해야 하는데, 이의 판단기준은 해당 구간 내에 데이터의 발생빈도(frequency)에 따라 결정된다. 일반적으로 실세계 데이터들의 발생빈도는 특정 영역에 치우치는 경향, 즉 지역성(locality)를 갖는다. 본 논문에서는 센서스 데이터와 같이 수량적 속성 데이터의 항목을 여러 개의 구간으로 분할할 때 데이터 발생의 지역성을 고려하여 빈발구간 항목집합을 생성하는 방법을 제안한다. 이 방법은 밀도 높은 구간을 중심으로 빈발구간 항목을 생성하므로 원래의 데이터가 가진 특성의 손실을 최소화할 수 있는 특징이 있다.

2. 연관규칙 탐사

연관 규칙을 탐사하기 위하여 데이터베이스에 있는

모든 항목들의 지지도를 계산하여 빈발 항목집합(large itemsets)을 찾고, 이로부터 주어진 신뢰도를 바탕으로 실제의 규칙을 탐사하는 과정으로 연관 규칙 탐사의 전체성능은 빈발 항목집합들(large itemsets)을 찾는 단계에서 결정된다. Apriori[3], AprioriTID[3,10], AprioriHybrid[3], DHP[4], Partition[5], DIC[11], Direct Sampling[5], Sampling Approach[6]등 연관규칙 탐사 알고리즘들 대부분이 이러한 문제의 해결에 중점을 두고 있다. 수량 데이터의 연관규칙 탐사 방법으로 수량항목을 이진항목으로 변환하여 기존의 탐사 알고리즘으로 연관규칙 탐사를 고려할 수 있다. Skriant[7]는 수량항목의 정의영역, 즉 도메인을 일정한 범위의 소구간으로 일괄분할(partition)한 후, 이웃한(adjacent) 소구간 분할을 병합(merge)하여 최소지지도를 만족하는 빈발 구간 항목집합을 생성한다. 이 경우에는 수량 항목의 정의 영역에 데이터가 골고루 분포된 경우에는 효과적이지만, 일부 영역에 집중된 경우는 비효율적인 면이 있으므로 이의 해결 방법으로 분포도에 따라 분할하는 유동적 분할법이 발표되었다[8]. 이들 두 가지 방법은 2 단계의 절차를 거쳐 최소지지도를 만족하는 빈발구간 항목집합을 생성한다. 첫째 단계에서는 소간격 분할을 생성하며, 두 번째 단계에서는 최소지지도를 만족할 때까지 이웃 소구간을 합병한다. 분할과 병합에 사용되는 기준은 일정범위 분할법에서는 최소지지도만 사용하고, 유동적 분할법은 최소지지도와 최소분할지지도를 사용하여 분할 및 병합을 실시하였다. 따라서 일정범위 분할법은 데이터의 분포를 고려하지 못하는 점이 약점이고, 유동분할법은 최소분할지지도라는 또 다른 분할기준을 사용함으로써 분할을 위한 부수적인 비교가 필요할 뿐만 아니라사용자가 임의로 최소분할지지도을 설정해야 하므로 최적의 기준설정이 어렵다는 문제점이 있다.

3. 지역성을 고려한 빈발구간 항목집합 생성 방법

본 논문에서 제안하는 방법은 최빈수(mode)의 단위구간을 중심으로 수량 항목의 정의영역을 이진항목으로 변환하여 빈발구간 항목집합을 생성하는 방법이다. 데이터 발생의 밀집도가 높은 영역이 최빈수의 단위구간이다. 최빈수 단위구간은 손실하면 않되는 의미 있는 구간으로, 이를 이용하는 방법이다. 빈발구간 항목집합을 생성하기 위하여 필요한 기호를 다음과 같이 정의한다.

- D는 유한한 범위의 수량 항목이 포함된 트랜잭션 들의 집합으로 Lq, f(Lq)를 포함한다.
- Lq는 { lq1, lq2, ... , lqn-1, lqn }이며, lqi (1≤ i ≤ n)는 단위구간 항목(item)으로 이산(discrete)하다.
- f(Lq)는 {f(lq1), f(lq2), ... , f(lqn-1), f(lqn)}이며, f(lqi) (1≤ i ≤n)는 단위구간에서 데이터 발생빈도이다.
- Max_lq는 최빈수(mode)의 단위구간이다.
- FL은 빈발구간 항목집합(large interval itemsets)이며, m개의 원소들{fl1, fl2, … ,flm }로 구성되고, 각각은 최소지지도(Smin)를 만족한다.
- lq_ti(1≤ i ≤n)는 해당 단위구간 lqi의 사용 가능여부를 표기한다.

빈발구간 항목집합 생성 방법은 1차 최빈수의 단위구간(lqi)을 선택한 후, 이를 기준으로 인접(좌~우) 단위구간(lqi-1, lqi+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 상한한계로 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 진행하며, 더 이상 단위구간이 존재하지 않을 때 중단한다.

그림 3, 4는 이러한 절차를 코드로 표기한 것이다. 그림 3은 최빈수의 단위구간을 선정하여 Gen-FL함수에 그 값을 전달한다. 그림 4에서는 전달된 단위구간을 기준으로 최소지지도(Smin)를 만족할 때까지 좌~우의 단위구간을 병합하는 절차를 수행한다.

```
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 = ∪ flk// Answer
    Max_lq = 0
end
```

그림 3. 최빈수를 사용한 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 andlq_ti+jis tagged
                 return// flk is not large
    end
Return
```

그림 4. 단위구간 병합함수

그림 5는 앞에서 언급된 그림2의 (b) 즉, 연건평별 가구수 데이타를 제안한 방법에 의해 반발구간 항목집합을 생성하는 절차를 도식한 것이다. 그림 6의 (a)는 기존 방법에 의해 빈발구간 항목집합이 생성된 결과를 도식한 것이고, (b)는 기존방법과 제안된 방법에 의해 생성된 빈발구간 항목에 대해 병합된 구간의 지지도를 비교한 도표이다.

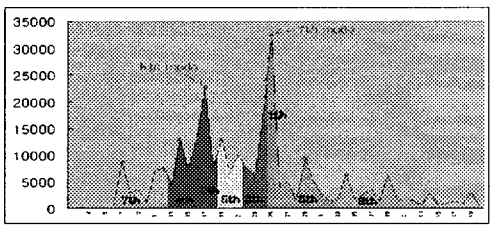

그림 5 지역성을 고려한 빈발구간 항목 생성과정

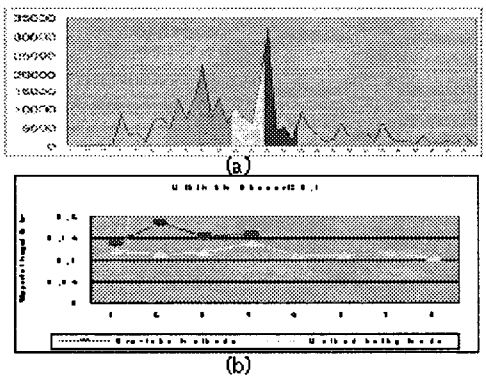

그림 6 지역성을 고려한 방법과 기존방법의 비교

최빈수의 단위구간을 기준으로 생성된 빈발구간 항목은 그림5 에서 보듯이 1차에서 k차로 진행될수록 빈발구간 항목의 폭이 넓어지는 특징이 있다.

반면, 그림 6의 (a)는 앞서 설명된 기존의 방법들로 모두 4개의 단위구간을 일정하게 병합하여 빈발항목 집합을 생성하는 것을 볼 수 있다, 또한 (b)에서와 같이 제안된 방법의 병합된 구간의 지지도는 지정된 최소지지도를 기준으로 편차가 크지 않은 반면 기존의 방법들은 지정된 최소지지도 보다 월등히 높은 지지도를 보이고 있다. 이는 지역성을 전혀 고려하지 않았기 때문이며, 본 데이터가 가지고 있는 특성이나, 의미 있는 데이터에 대한 손실이 클 수 있다는 것을 의미한다.

4. 성능평가

본 논문에서 제안된 방법의 성능평가 위하여 일정범위 분할 및 병합방법(M1), 유동적 분할 및 병합방법(M2) 그리고 데이터의 지역성을 이용한 빈발구간 항목집합 생성방법(M3)을 사용하여 생성되는 빈발구간 항목 수와 생성구간의 평균간격을 비교한다. 성능평가에는 다음의 3가지 데이터를 사용한다.

i) 인구주택총조사 중 대전광역시의 연령별 인구데이터 1,214,327 레코드, ii) 사업체기초통계조사 중 대전광역시의 종사자 규모별 사업체 데이터 88,869 레코드, iii) 지역성이 없는 모의 데이터 37,000 레코드

이들 데이터의 분포는 그림 2의 (a), (b)와 같으며, 사용한 최소지지도는 9가지(40%, 35%, 30%, 25%, 20%, 15%, 10%, 5%, 3%)를 사용하였다. M1에서 사용한 일정분할 간격은 2, M2의 최소분할지지도는 최소지지도의 1/2로 하였다.

4.1 생성 빈발구간 항목수 비교

각 방법의 성능평가 결과는 그림 7과 같다. 이들 도표를 보면 제안한 방법(M3)이 25% 미만부터는 보다 많은 수의 빈발구간 항목을 생성하고 있음을 알 수 있다. 각 데이터별 시험결과를 살펴보면, 첫째 그림 7 (a)의 경우는 최소지지도가 낮아질수록 제안한 방법이 보다 많은 빈발구간 항목을 생성함을 알 수 있고, (b)에서 M3는 최소지지도 15%부터 빈발구간 항목수가 증가한다. (c)는 10%이하에서 M3가 보다 많은 빈발구간 항목을 생성하고 있다.

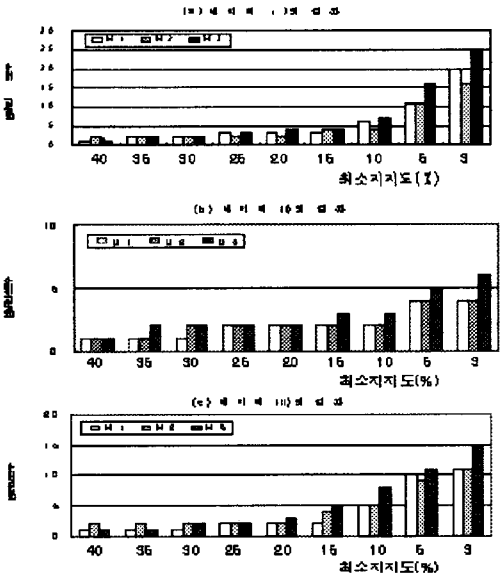

그림 7. 생성 빈발구간 항목수

이상과 같이 제안한 방법이 최소지지도 25% 미만에서는 기존의 방법에 비해 보다 많은 빈발구간 항목을 생성하므로 우수한 성능을 보여준다.

4.2 구간 평균간격 비교

각 방법으로 생성된 빈발구간 항목들의 구간들의 평균간격은 그림 8과 같다. 첫째 그림 8의 (a)는 M1의 구간 평균간격이 타 방법에 비하여 월등히 넓음을 알 수 있고, M3가 좁은 구간 평균간격의 빈발구간항목집합을 생성함을 알 수 있다. 둘째, 그림 8의 (b)는 M1이 최소지지도 30%에서 15%사이에 넓은 구간 평균간격이 보이고 있다. 이는 데이터 분포의 특성을 고려하지 않았기 때문이다. M3의 경우 '특이 부분'이라 표시한 부분에 유달리 넓은 구간 평균간격을 보이고 있다. 이는 앞에서 언급한 바와 같이 빈도가 극히 낮은 영역에서 생성된 빈발구간 항목으로 인하여 구간 평균간격이 넓게 나타나고 있다. 이러한 경우는 생성순서 정보를 활용하여 제거하거나 규칙생성 과정에서 활용유무 판단이 필요하다. M3'은 M3가 생성한 빈발구간 항목들 중에서 나중에 생성된 것을 제외하여 M2와 동일한 수의 빈발구간 항목에 대한 구간 평균간격이다. 그 결과 M3'는 M2와 거의 동일한 구간 평균간격을 보이고 있다. 마지막으로 그림 8의 (c)는 지역성이 없는 모의 데이터에 대한 구간 평균간격의 그래프이다. 이 데이터에 대해서는 세 가지 방법 모두의 구간 평균간격이 유사하게 나타나고 있다.

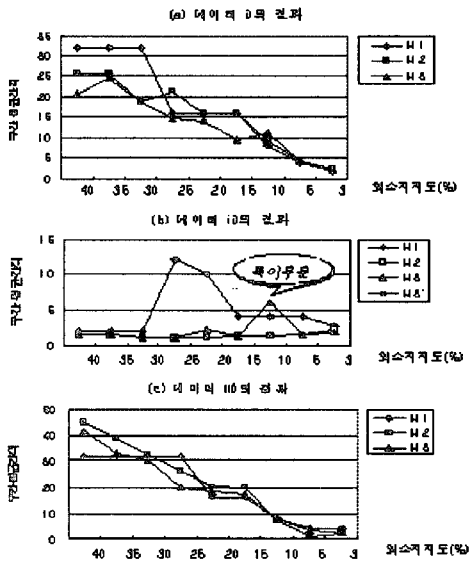

그림 8. 구간 평균간격

5. 결론

본 논문에서는 수량 항목의 정의영역을 이진항목 형태의 빈발구간 항목으로 변환하기위한 보다 효율적인 방법을 제안하였다. 그리고 인구주택총조사 등 실제 데이터를 사용하여 성능평가를 실시하여 제안한 방법이 기존의 방법보다 보다 많은 수의 세밀한 빈발구간 항목집합을 생성할 수 있으므로 성능의 우수함을 보였다. 제안한 방법은 탐사 대상 데이터가 가지는 특성, 즉 데이터 발생의 지역성을 고려하는 방법으로 최빈수를 사용하여 빈발구간 항목을 생성하였다. 최빈수를 사용함으로써 얻는 효과로는 보다 세밀한 빈발구간 항목을 생성함은 물론 빈도가 높아 의미 있는 데이터임에도 불구하고 손실하게 되는 확률을 극히 최소화 시킬 수 있으며, 그림 8에서와 같이 최빈수의 차수, 즉 생성순서에 따라 빈발구간 항목의 세밀도가 감소하는 특징이 있다.

[참고문헌]

[1]. Sholom M. Weiss, Nitin Indurkhya,: Predictive Data Mining, Morgan Kaufmann Publishers, Inc. Francisco, California.(1998)

[2]. 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.(1993) 207-216.

[3]. R. Agrawal and R. Srikant.: Fast Algorithms for mining association rules, In Proceedings of the 20th VLDB Conference. Santiago, Chile. Sept(1994).

[4]. J.S, park. M.S, Chen. And P.S, Yu.: An Effective hase-based algorithm for mining association rules, In Proceedings of ACM SIGMOD Con-ference on Management of Data. May(1995) 175-186.

[5]. A. Savasere, E. Omiencinsky and S. Navathe, "An efficient algorithm for mining rules in large databases", In proceedings of the 21st VLDB Conference (1995). 432-444.

[6]. J.S, Park. P.S, Yu and M. S, Chen.: Mining Association Rules with Adjustable Accuracy, In Proceedings of ACM CIKM 97, November (1997) 151-160.

[7]. R, Srikant and R, Agrawal.: Mining Quan-titative Association Rules in Large Relational Tables, Proceedings of the ACM SIGMOD Conference on Management of Data. (1996)

[8]. Young-Hee, Choi. Su-Min, Jang. Jae-Chul, OH.:Generationg Large Items Efficiently For Mining Quantitative Association Rules, Vol.6. KIPS (1999) 2597-2607.

[9]. Rajeev Rastogi and Kyuseok Shim.: Mining Optimized Association Rules with Categorical and Numeric Attributes, IEEE Transactions on Knowledge and Data Engineering, vol. 14, No.1, January/February (2002) 29-50.

[10]. S, Brin. R, Motwani. J.D, Ullman. and S, Tsur.: Dynamic Itemset Counting and ImplicationRules for Market Basket Data, In Proceedings of ACM SIGMOD Conference on Management of Data(1997) 255-264