수량적 속성과 시계열 분석에 의한 연관규칙 탐사

양신모*0, 정광호*, 김진수*, 최성용*, 이정현**
인하대학교 전자계산공학과*, 인하대학교 컴퓨터공학부**
{sinmo*0, khjung*, kjspace*, sychoi* }@nlsun.inha.ac.kr, jhlee@inha.ac.kr**

Discovery of Association Rules Based on Data of Quantitative Attribute and Time Series

Shin-Mo Yang*0, Kwang-Ho Jung*, Jin-Su Kim*, Seong-Young Choi*, Jung-Hyun Lee**
{Dept.*, School**}of Computer & Science Engineering, Inha University

요 약

연관규칙은 데이터 안에 존재하는 항목들간의 종속 관계를 찾아 내는 것이다. 기존의 연구에서는 연관규칙 탐사 과정에서 발견항목 자체에만 관심을 두고 연구되어 왔다. 즉, 연관규칙 생성을 위한 후보 항목은 수량을 배제한 항목 대 수량비가 1:1인 상태에서 규칙을 발견하는 연구였다. 이것은 항목의 구매 수량에 관계없이 같은 가중치로 규칙을 발견하는 문제점을 갖고 있다. 두 번째 문제점은 연관규칙은 시간적 연장선상에서 발견되는 규칙이라 할 수 있다. 즉, 규칙을 발견하는 과정에서 모든 자료를 동일한 시간적 가중치를 두어 취급하는 것이다. 본 논문에서는 각각의 아이템을 (아이템,수량)의 묶음 단위로 후보항목을 만들어 수량적 속성이 포함된 아이템 대 수량 비 1:n의 관계에서 규칙을 발견하는 방법을 제안한다. 또한 과거의 자료들을 이용하여 예측할 때 모든 자료를 동일하게 취급하기 보다는 최근의 자료에 더 큰 비중을 주는 예측법을 사용하여 연관규칙 발견의 신뢰성을 높인다. 성능평가는 기존의 알고리즘과 비교하여 제안한 알고리즘의 성능향상 및 타당성을 보인다.

1 서 론

바코드 기법의 발전 및 기업과 정부의 전산 업무 환경이 급속도로 향상되며 변해 감에 따라 많은 양의 데이터가 생성, 수집, 저장되고 있다. 데이터베이스가 대용량화 되면서 데이터베이스 모델링에서는 현실 세계의 모든 규칙성을 반영하지 못하고 있다. 이처럼 감추어져 있긴 하나 잠재적 사용가치가 큰 패턴이나 흥미로운 규칙들을 발견하는 연구가 활발히 진행되어져 왔다. 대용량 데이터베이스에서의 지식 발견(knowledge discovery in databases)이라고 정의되는 데이터 마이닝(data mining)은 최근 들어 시장전략 수립, 수요예측, 의료진단, 상품진열 등 광범위한 분야에서 유용한 정보를 제공하기 시작했다[1].

대용량의 데이터베이스에서 어떤 사건들이 함께 발생하거나, 또는 하나의 사건이 다른 사건을 암시하는 것과 같은 사건간의 상호관계를 나타내는 연관규칙을 발견하는 문제는 가장 중요한 데이터 마이닝 문제들 중의 하나이다. 기존의 연관규칙을 이용한 데이터 마이닝은 주로 항목 자체에만 흥미를 두고 연구되었기 때문에 " 빵 2개와 우유 3개 이상을 구입하는 고객들의 75%가 휴지1개를 구입한다" 와 같은 수량적 속성이 포함된 연관규칙에 대해서는 관심을 두지않았다. 또한, 연관규칙의 발견에서 데이터의 시간적인 관련성을 배제하고 과거의 자료와 현재의 자료를 동일하게 취급하는 문제점으로 인해 관련이 없는 자료까지 규칙발견에 적용되어 나타나는 문제점이 있다.

본 논문에서는 기존의 Apriori알고리즘에서 트랜잭션 내의 항목의 수량적 속성을 배제하므로 항목간에 같은 가중치가 부여되는 문제점을 착안하여 수량적 가중치를 적용한 연관규칙 탐사 방법을 제시한다. 또한, 이때 시간적인 속성을 고려하여 현시점에 가까운 자료에 높은 비중을 두어 연관규칙 탐색을 한다.

2 관련 연구

2.1 연관규칙(Association Rules)의 문제점

연관규칙을 탐사하는 방법은 기본적으로 먼저 최소지지도(minimum support)이상의 항목들인 빈발 항목집합(frequency item sets)을 찾는 과정과 다음 단계로 이 빈발 항목집합에서 항목들 사이의 신뢰도(confidence)를 측정하여 주어진 값 이상일 때만 연관규칙을 만들어 낸다. 그러므로, 연관규칙을 탐사하는 문제는 트랜잭션 데이터베이스에서 빈발 항목집합을 찾아내는 문제와 이 빈발 항목집합에서 연관규칙을 찾아내는 문제로 나뉘어 진다. 일반적으로 첫번째 경우 문제를 효율적으로 탐사 하는 연구가 활발히 이루어지고 있는 실정이다. 빈발 항목집합을 찾기 위해서는 먼저 후보 항목집합을 생성하고 전체 트랜잭션을 검색하여 각 후보 항목집합을 포함하는 트랜잭션 수를 결정한다. 이러한 트랜잭션의 수를 해당 후보 항목집합의 지지도라 하고, 이 지지도가 사용자가 미리 정한 최소 지지도보다 크거나 같을 때 이 후보 항목집합은 빈발 항목집합이 된다. 이 과정에서 연관규칙 탐사는 각각의 후보 항목집합을 생성할 때에 항목의 수량적인 가중치는 배제하고 단순히 하나의 항목으로 간주하여 연구되어왔다[1,2]. 다만 기존의 Apriori 알고리즘의 전단계를 모두 수용하면서 수량적 속성만을 추가한 연구가 진보된 형태라고 하겠다[3]. 즉, 지금까지는 항목들 간의 수량적 연관성은 제시되지 않고 단일 항목으로 연구 되고 있다.

2.2 시계열 분석(Time Series)

시계열이란 한 사상(event) 또는 여러 사상에 대하여 시간이 흐름에 따라 일정한 간격으로 이들을 기록한 자료를 말한다. 시계열 자료는 시간의 흐름에 따라 어떤 규칙이나 우연성에 의해 변화되어진다. 주어진 시계열 모형이 많은 경우에 지엽적으로는 타당하나 시간이 경과함에 따라 시계열이 생성되는 시스템에 변화가 있을 수 있다. 따라서 과거의 자료들을 이용하여 예측 등에 이용하고자 할 때 모든 자료를 동일하게 취급하기 보다는 최근의 자료에 더 큰 비중을 주는 예측법을 사용하는 것이 합리적일 것이다. 이는 어떤 시점의 자료는 과거 시점의 자료 값에 의존한다는 전제에 의해 이루어진다. 한 시점에서 관측된 시계열 자료는 그 이전까지의 자료들을 의존하게 된다. 실제로 어떤 시간 간격t마다 X값을 측정하여 얻은 시계열 데이터를 $X(0),X(t),X(2t),\cdots$등으로 정의하자. 전체 운동의 배후에 어떤 결정론적인 법칙이 있다면 어떤 시간에서의 $X(2t)$값은 과거의 $X(0),X(t)$값들에 의존할 것이다. 얼마나 많은 X값들에 의존할 것인지는 원래 자료가 가지고 있는 특성과 공간의 차원 값과 관계가 있다[4].

3 수량적 속성과 시계열 분석에 의한 연관규칙 탐사

3.1 수량적 속성에 기반한 긴밀성(Intimacy)

긴밀성이란 한 트랜잭션에서 아이템 간의 일대일 대응을 만들 수 있는 경우의 수와 비례하는 것이며, 아이템의 개수의 차이가

클수록 긴밀성은 감소한다[5]. 긴밀함의 정도는(degree of Intimacy)는 긴밀계수로 측정되고, 긴밀계수는 단일 아이템 또는 아이템과 아이템간의 전체 트랜잭션에서 차지하는 비중을 나타낸다. 본 논문에서는 수량적 정보를 이용하여 긴밀계수를 표현한다. 하나의 트랜잭션 $T_k = \{(A_1,n_1),(A_2,n_2),\ldots,(A_i,n_i)\}$이 주어졌을 때 아이템($A_i$)에 대한 아이템 집합 $X=\{A_1,A_2,\ldots,A_k\}$의 긴밀계수는 식(1)에 의해 정의한다.

$$Intico_x = \min\left(\frac{n_{1,2,\ldots,l}}{n_1}, \frac{n_{1,2,\ldots,l}}{n_2}, \ldots, \frac{n_{1,2,\ldots,l}}{n_k}\right) \times (n_1 + n_2 + \ldots + n_l) \quad (1)$$

n_i : 트랜잭션에서 발생한 A_i 아이템의 수, $i = 1, 2, \ldots, k$

$n_{1,2,\ldots,l}$: 트랜잭션에서 만들 수 있는 아이템들의 순서쌍$(A_1, A_2, A_3, \ldots, A_k)$ 의 수

긴밀계수의 첫번째 특성으로는, 아이템 n개의 긴밀계수 값과 서로 다른 n 개 아이템의 긴밀계수가 같은 값을 갖는 것이다. 즉, 아이템 내 긴밀계수 산출이나 아이템 간 긴밀계수 산출 방법에는 편차가 없다. 이러한 특성은 트랜잭션 내에서 발생한 총 수량이 포함 되었기 때문에 나타나는 것으로, 수량적 정보에 의한 대표적인 긴밀계수의 특성이다. 두번째 특성으로는, 아이템 간의 수량적 차이가 클수록 긴밀계수 값은 작아지는 것이다. 따라서, 하나의 트랜잭션에서 아이템 내의 긴밀도는 아이템 간의 긴밀도에 영향을 미친다. 이는 하나의 아이템이 트랜잭션 내에서 발생할 확률을 나타낸다.

3.2 수량적 속성의 긴밀율(Intimacy Ratio)과 긴밀 아이템 집합(Intimacy Itemset)

긴밀율은 전체 데이터베이스에서 아이템 내의 긴밀 정도에 비해 아이템 간의 긴밀 정도가 어느 정도인지를 나타내는 지표이다[5]. 아이템간의 긴밀율이 높다는 것은 상호 의존도가 높아서 구매 트랜잭션의 경우에 구매 상승효과를 기대할 수 있는 아이템 집합을 의미한다. 긴밀율 측정을 위해서는 최소 2개 이상의 아이템을 필요로 한다. 따라서, 1-아이템 집합의 긴밀율은 식(2)에 의해 별도로 구한다. 식(2)는 각 아이템들이 전체 트랜잭션에서 얼마나 자주 발생하는지를 나타내는 빈도율을 정의한다.

$$Intimacy\_ratio_{(A_i)} = \frac{Intimacy_i}{|D|} \quad (2)$$

$Intimacy_i$: 1-아이템 집합 A_i 긴밀계수, i=1,2,…,k

$|D|$: 트랜잭션의 총 개수

그리고 아이템 집합 $X=\{A_1,A_2,\ldots,A_k\}$에서 1-아이템 집합 이상의 긴밀율은 식(3)과 같이 정의 한다.

$$Intimacy\_ratio_x = \frac{Intimacy_{1,2,\ldots,k}}{Intimacy_1 + Intimacy_2 + \ldots + Intimacy_k} \quad (3)$$

for $k \geq 2$

$Intimacy_{1,2,\cdots,k}$: 아이템 집합 $\{A_1,A_2,\ldots,A_k\}$의 긴밀계수

$Intimacy_i$: 1-아이템 집합$\{A_i\}$의 긴밀계수, i =1,2,…,k

3.3 지수 평활법(Exponential Smoothing Method) 에 의한 시계열 자료 분석

지수 평활법은 과거의 관측값으로 미래의 값을 예측할 때 최근의 자료에 더 많은 가중값을 부여하여 예측하는 방법이다. 본 논문에서는 트랜잭션을 n개의 파티션(partition)으로 나누어 현재를 기준으로 과거로 갈수록 가중치를 적게 부여하는 방법을 적용하였다. 이는 미래의 값을 예측하는데 필요한 정보는 최근의 자료에 더 많이 포함될 수 있으며, 모든 자료를 동일하게 취급하기 보다는 최근의 자료에 더 큰 비중을 주는 시계열 특징을 나타낸다. 지수 평활법은 과거로 갈수록 가중값의 크기를 지수적으로 줄여나가 상대적으로 최근 관측값에 더 큰 비중을 두는 방법이다. 본 논문에서는 최소지지도와 신뢰도의 가중치를 지수 평활법을 이용하여 실험에 적용한다. 지수 평활법은 다음 식(4)에 의해 정의한다.

$Z_1, Z_2, \ldots Z_n$ 이 관측되었을때, Z_{n+l}의 예측값 표현

$$Z_n(l) = c\sum_{i=0}^{n-1}(1-w)^i Z_{n-i}$$
$$= c[Z_n + (1-w)Z_{n-1} + \Lambda + (1-w)^{n-1}Z_1] \quad (4)$$

c : 정규화 상수(normalizing constant)

3.4 수량적 속성과 시계열 분석을 위한 탐색 알고리즘

시간적인 가중치 부여를 위하여 일차적으로 수행하여야 할 것은 데이터베이스를 파티션하는 것이다. 제안된 알고리즘의 단계 1에서는 n개의 파티션으로 나누어 현재를 기준으로 가장 최근의 자료에 가중치를 부여하기 위한 과정이다. 이처럼 분할된 데이터베이스에 대하여 빈발 항목집합을 찾는다. 이때의 빈발 항목집합은 3.2절의 지수 평활법에 의한 수정된 최소 지지도를 사용한다. 이로써 현재의 파티션을 중심으로 지수 평활법을 적용하여 과거로 갈수록 가중치를 줄여줌으로써 각 파티션마다 다른 최소지지도 값을 가지게 된다. 이렇게 얻어진 최소지지도를 바탕으로 식(1)과 식(2)를 이용하여 긴밀 아이템 집합을 만든다. 단계 2에서는 각각의 분할된 데이터베이스에서 구해진 빈발 항목집합을 합하여 전체 후보 빈발 항목을 생성한다. 단계 3에서는 전체 후보 빈발 항목에 대하여 전체 데이터베이스를 순차적으로 검색함으로써 이들의 발생 수를 계산하여 계산된 발생수가 최소 지지도보다 큰 경우에 대하여 최종 빈발 항목으로 선택하게 된다. 본 논문에서 제안한 탐색 과정을 다음의 알고리즘1로 표현한다.

알고리즘1. 수량적 속성과 시계열을 적용한 연관규칙탐사 (Apri_QA_TS)

```
P : partition_database(DB)
n : Number of partitions

//단계 1 : DB를 시간을 기반으로 파티션한다.
for i=1 to n do begin
    read_in_partition(p_i ∈ P)
    L^i = gen_large_itemset(p_i)
end

//단계 2 : 각 파티션의 빈발 항목을 합병하여 전체 후보 빈발 항목을 생성한다.
for (i=2; L_j^i ≠ φ, j=1,2,Λ,n ; i++) do begin
    C_i^G = ∪_{j=1,2,Λ,n} L_j^i
end

//단계 3 : 순차적 DB 스캔을 통해 최종 빈발 항목을 선택한다.
for i=1 to n do begin
    read_in_partition(p_i ∈ P)
    for all candidates c ∈ C^G gen_count(c, p_i)
end
L^G = {c ∈ C^G | c.count ≥ min Sup}
Answer = L^G
```

함수 gen_large_itemsets은 분할 p_i를 입력으로 하여 출력으로 모든 길이의 지역 빈발 항목집합들 $L_1^i, L_2^i, \Lambda, L_l^i$를 생성한다. L_k^p는 분할 p에서 지역 빈발 k-항목집합들의 집합이고, C_k^p는 분할 p에서 지역 후보 k-항목집합들의 집합이다. C_k^p는 전역후보 k-항목집합들의 집합이고 C^G는 전체 전역 후보 항목 집합들의 집합이고, L^G는 전체 전역 빈발 항목집합들의 집합이다.

4 실험 과정 및 결과 분석

본 논문에서 제안한 알고리즘(Apri_QA_TS)의 실험 과정 모형을 그림1과 같이 구성하였다. 본 실험의 데이터 소스는 웹로그 파일, 웹페이지의 링크 정보 데이터베이스이다. 실험은 액세스 로그 파일을 이용하여 정제 과정을 거친 후 사용자 트랜잭션을 생성한다. 정제 과정은 액세스 로그파일에서 여러 필드 중 IP, 접속시간, 접속 페이지 필드(*.html, *.htm)만을 필터링한다. 그리고 사용자별로 방문을 결정하고, 한번의 방문 동안 이동한 경로를 세션으로 결정한다. 이후에 (웹문서,방문횟수)의 단위로 재정제하여 제안한 알고리즘을 적용한다. 발견된 연관규칙을 해석하는 과정에서 웹 사이트 링크 정보 데이터베이스를 참고하여 웹 사이트 구조 개선을 보인다.

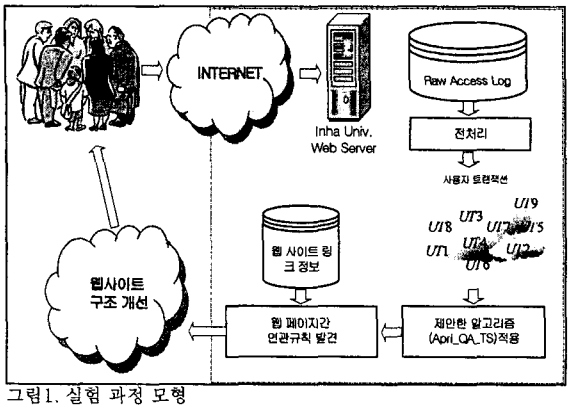

그림1. 실험 과정 모형

4.1 실험 평가 및 분석

실험을 위하여 2002년 8월 한달 동안의 인하대학교 대학원 홈페이지 로그 파일을 사용하였다. 성능평가는 두 가지로 나누어서 하였다. 첫번째 방법으로 기존의 Apriori알고리즘과 제안한 알고리즘(Apri_QA_TS)의 성능평가를 보인다. 시계열에 의한 파티션을 하나로 둔 상태에서 최소지지도(min_support)와 신뢰도(confidence)를 변화시켜 가면서 이루어진다. 이는 수량적 속성만 적용된 알고리즘의 성능평가를 위한 것이다. 두번째 방법으로 수량적 속성과 시계열에 의한 파티션의 개수를 늘려가면서 성능평가하였다. 이것은 파티션에 의한 실험의 의존도를 보여주기 위한 것이다. 평가 방법으로는 식(5)를 이용하여 재현율(recall ratio), 정확률(precision ratio), F-검정으로 표현한다.

재현율=검색된 적합 문헌수/적합문헌 총 수
정확률=검색된 적합 문헌수/검색문헌 총 수 (5)
F-검정=2×(재현율×정확률)/(재현율+정확률)

아래의 그림2와 그림3은 첫번째 방법으로 기존의 Apriori알고리즘과 제안한 알고리즘의 성능평가를 가시화하여 보여주고 있다.

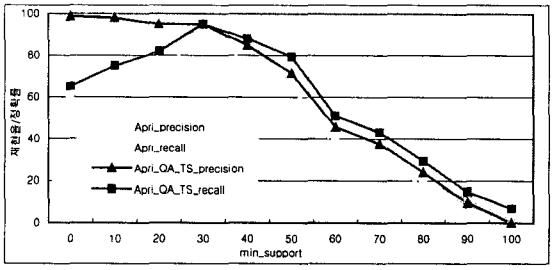

그림 2. 지지도 변화에 따른 재현율과 정확률 성능비교

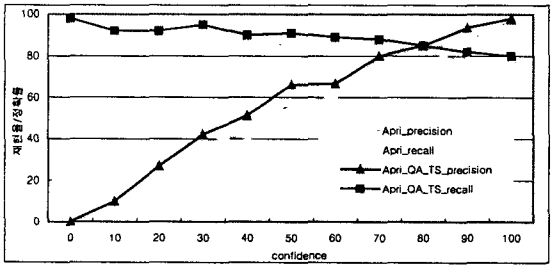

그림 3. 신뢰도 변화에 따른 재현율과 정확률 성능비교

각각의 지지도와 신뢰도를 변화시켜감에 따라 기존의 Apriori 알고리즘과 제안한 알고리즘(Apri_QA_TS)의 재현율, 정확률의 비교 평가 한다. 두번째 성능평가로는 데이터베이스 내의 트랜잭션을 시계열 모형화하여 파티션 수를 증가 시켜감에 따라 F-검정을 비교한다. 다음 표 1 과 그림 4 는 두번째 평가 방법에 대한 결과표이다.

표 1. 제안한 알고리즘의 파티션에 따른 성능비교

파티션 수 / F-검정	1	2	3	4	5
	6.211	10.493	5.259	3.814	3.667

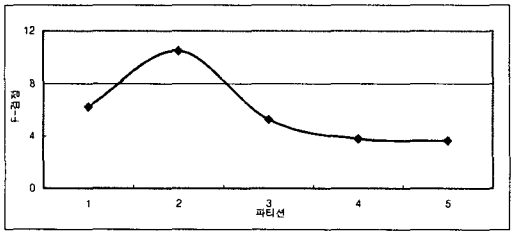

그림4. 파티션의 변화에 따른 F-검정 성능비교

첫번째 실험의 결과는 기존에는 사용자에 따라 임의로 정해졌던 지지도와 신뢰도를 실험을 통하여 데이터베이스에 맞는 임계값을 찾음을 보여준다. 두 알고리즘 모두 지지도가 30%일때 신뢰도가 82%일때 재현율과 정확률이 교차되고 있다. 이 시점이 본 실험의 가장 적합한 지지도와 신뢰도의 임계값이 된다. 또한, 제안한 알고리즘이 기존의 알고리즘 보다 나은 재현율과 정확률을 보이고 있다. 두번째 성능평가 결과 파티션의 개수가 두개일 때 가장 높은 F-검정을 보이고, 두개 이상의 파티션의 개수에서는 일정한 값에 수렴하고 있다. 이는 파티션이 증가할 때 마다 지수적으로 가중치를 줄여나감으로써 나타나는 현상이다. 위의 성능평가를 고려해 볼 때 항목간의 수량적 관련성 정보와 시계열로 연관규칙을 탐사한 결과가 기존의 Apriori알고리즘보다 월등한 수행평가를 보이고 있음을 증명해 준다.

5 결론

연관규칙은 아이템 수량에 관계없이 같은 가중치로 규칙을 발견하는 문제점과 규칙 발견 시 시계열특성에 대한 분석을 하지않음으로 인해 현재와 과거의 시간을 같은 가중치로 적용하여 규칙을 발견하는 문제점을 가지고 있다. 본 논문에서는 이러한 기존의 연관규칙의 문제점을 착안하여 항목간의 수량적 속성을 적용하여 항목간의 긴밀성을 측정하였다. 또한, 가장 최근의 사건이 현재의 사건에 더 밀접한 관계를 두고 있다는 시계열의 특성을 적용하여 현재를 기준으로 시간적 가중치를 점차 줄여 나가는 규칙 발견 알고리즘을 제안하였다. 본 논문에서 제안한 알고리즘을 이용하여 얻어진 정보들은 수량의 변화와 시간에 민감한 분야에 적용되어 전략적 예측에 유용하게 사용된다고 사료된다.

참고문헌

[1] Viet Phan-Luong, "The Representative Basis for Association Rules, " In Proc. of IEEE International Conference on Data Mining, pp. 639-640, 2001.

[2] M.S. Chen, J. Han and P.S. Yu, "Data Mining" An Overview from a Database Perspective. IEEE Transaction on Knowledge and Data Engineering, 8(6): pp. 866-883, 1996.

[3] R. srikan, R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables," In Proc. of ACM SIGMOD Conference on Management of Data, Montreal, Canada, 1996.

[4] Fu-lai Chung, Tak-chung Fu, Robert Luk and Vincent Ng, "Evolutionary Time Series Segmentation for Stock Data Mining," In Proc. of IEEE International Conference on Data Mining, pp. 83-99, 2002.

[5] 양신모,정경용,김진수,최성용,이정현, "아이템의 범주적 속성과 수량적 속성에 기반한 연관규칙 발견," 한국정보과학회 HCI연구회 제 12회 학술발표논문집(1), pp.456-461, 2003.