사용자 구분에 의한 지역적 연관규칙의 유도

박세일* , 이수원
숭실대학교 컴퓨터학과
lunacy@valentine.ssu.ac.kr, swlee@computing.ssu.ac.kr

Deriving Local Association Rules by User Segmentation

Se-Il Park, Soo-Won Lee
School of Computing, Soongsil Univ.

요 약

연관규칙 탐사기법은 트랜잭션들을 대상으로 항목간, 또는 속성간의 연관관계를 발견하는 방법으로, 데이터 집합의 구조를 쉽게 통찰할 수 있다는 장점으로 인하여 활발히 연구되어져왔다. 그러나 현재까지의 연구들은 전체 사용자 중 공통적인 특성을 지닌 사용자 그룹이 존재할 경우, 그러한 그룹별 연관규칙을 찾아낼 수 없다는 한계점을 지닌다. 본 논문에서는 이러한 점을 해결하기 위하여, 속성선택 및 사용자 구분 기법을 이용하여 사용자를 부분집합으로 구분하고, 그 부분집합별로 연관규칙을 발견한다. 또한 위와 같이 얻어진 연관규칙이 전체 사용자를 대상으로 한 연관규칙보다 해당 부분집합에 더욱 적합하다는 사실을 여러 연관규칙 평가치를 이용하여 평가한다.

1. 서론

데이터마이닝이란, 대규모의 데이터베이스로부터 숨겨진 지식이나 패턴, 새로운 지식 등을 발견하고자 하는 작업이다. 데이터마이닝 기법들 중 가장 활발히 연구되어지고 있는 연관규칙 탐사기법은 트랜잭션들을 대상으로 항목간, 또는 속성간의 연관관계를 발견하는 방법이며, 구해진 연관규칙은 장바구니 분석, 교차판매 전략, 카타로그 디자인, 상품 배치, 기만 탐지 등의 목적으로 활용된다.

연관규칙 탐사기법 알고리즘의 기본이라고 할 수 있는 Apriori 알고리즘이 개발된 이래로, 일반화된 연관규칙 탐사와 수치 연관규칙 탐사, 그리고 제약조건을 이용한 연관규칙 탐사 등, 연관규칙에 관련된 많은 연구가 진행되어왔다[1,2,3]. 그러나, 현재까지의 연구들은 전체 사용자 중 공통적인 특성을 가지는 사용자 그룹들이 존재할 경우, 그러한 그룹별 연관규칙을 찾아낼 수 없다는 한계점을 지닌다. 예를 들어, 도서대여 패턴에서 <표.1>과 같은 연령, 성별별 패턴들이 존재할 경우, 기존의 연관규칙 탐사 기법으로는 원하는 결과(예: 성인 남성 고객에게 특화된 도서 연관규칙)를 찾아 낼 수 없다.

즉, 전체 트랜잭션(D)의 크기($|D|$) 가 1000이고, 성인 남성 고객 집단(C_3) 의 크기($|C_3|$)가 300인 경우, r_1이라는 특정 연관규칙이 C_3에서만 발견되고, 그 빈도($f(r_1)$)가 100이라고 가정한다면, 지지도 임계치가 30% 일 때 전체 집합을 대상으로 한 연관규칙 탐사에서는 $f(r_1) / |D| = 100 / 1000 = 10$ (< 30 %)이므로 r_1이 발견되지 않는다. 하지만, C_3에서 연관규칙 탐사를 할 경우, $f(r_1)/ |C_3| = 100 / 300 = 33$ (> 30 %)이므로 r_1이 발견될 수 있다.

표 1. 도서대여 패턴

고객 분류	주 대여 패턴
유아 고객 : C_1	동화책, 유아용 만화책
10~20대 여자 고객 : C_2	로맨스, 공포 소설
성인 남성 고객 : C_3	무협 소설, 성인 만화

이러한 배경에서, 본 논문에서는 전처리 형태로 고객 트랜잭션을 특성에 따라 구분한 후, 각 집단별로 연관규칙 탐사기법을 시행하여 특정 집단에 적합한 지역적 연관규칙을 발견한다. 부분 집단에 지역적으로 적합한 연관규칙은 앞에서 언급한 장바구니 분석, 교차판매 전략 등, 연관규칙의 활용 시에 고객 구분 작업을 통하여 고객 별로 최상의 서비스를 제공하는데 도움을 줄 수 있다.

2장에서는 본 연구에 대한 이론적 배경에 대하여 설명하고, 3장에서는 제안된 기법의 전체적인 구조와 그 세부적인 사항을 설명하며, 4장에서는 기존의 연관규칙 탐사 기법과 제안된 연관규칙 응용 기법의 비교실험을 통하여 제안된 기법의 유용성을 고찰한다.

2. 이론적 배경

2.1 연관규칙[4]

• 정의 및 형식화

연관규칙은 $X \rightarrow Y$의 형식으로 표현되며, I가 항목들의 집합일 때, $X, Y \subset I$, $X \cap Y = \varnothing$의 특성을 갖는다. 여기서 X와 Y를 각각 이 규칙의 전건(antecedent)과 후건(consequent)이라고 한다. 또한, 이 규칙은 데이터베이스의 X를 포함하는 트랜잭션은 Y를 같이 포함하는 경향을 보인다는 의미를 가진다. 또한, 항목들의 집합을 항목집합(itemset)이라고 하며, k항목 수를 가지는 항목집합을 k-항목집합(k-itemset)이라고 한다.

• 지지도와 신뢰도

각각의 규칙은 규칙 내 항목이 차지하는 비중을 나타내는 지지도(support)와 그 규칙의 강도를 나타내는 신뢰도(confidence)를 가진다. $f(X)$가 항목 X의 빈도이고, 전체 트랜잭션 수가 N일 때, 지지도와 신뢰도는 다음과 같이 표현된다.

$$support\ (X \rightarrow Y) = f(X \cap Y) / N$$
$$confidence\ (X \rightarrow Y) = support(X \cap Y) / support(Y)$$

X와 Y가 이진 값을 가질 경우, 이 두 항목을 포함하는 데이터 집합은 다음과 같은 2×2의 contingency table로 요약된다.

	Y	¬Y	
X	f_{11}	f_{10}	$f_{1\cdot}$
¬X	f_{01}	f_{00}	$f_{0\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 0}$	N

f_{ij}는 각 경우에 해당하는 빈도를 나타내며, $f_{i\cdot}$와 $f_{\cdot j}$는 각각 i행과 j열들에 대한 빈도의 합을 뜻한다. 즉, $f_{i\cdot} = f_{i1} + f_{i0}$이다.

contingency table을 이용하여 지지도와 신뢰도를 표현하면 다음과 같다.

$$support\ (X \rightarrow Y) = f_{11} / N$$
$$confidence\ (X \rightarrow Y) = (f_{11} / N) / (f_{1\cdot} / N) = f_{11} / f_{1\cdot}$$

2.2 연관규칙에 대한 평가[5,6,7,8,9]

연관규칙 탐사로 도출되는 규칙의 수가 상당히 많은 경우, 방대한 양의 규칙들을 흥미도 평가치(interestingness measure)를 이용하여 가지치기하고 분석하는 추가적 작업이 필요하며, 규칙 수가 상대적으로 적은 경우에도 어떤 규칙이 더욱 중요한 규칙인지를 평가하는 작업이 필요하다.

• 주관적 평가(subjective measure)와 객관적 평가(objective measure)

연관규칙의 평가 방법은 크게 두 가지 접근방법이 존재하는데, 그것은 주관적 평가방법과 객관적 평가방법이다. 주관적 평가방법에서 규칙의 가치가 규칙을 평가할 사용자에 의하여 결정(user-driven)되는데 반하여, 객관적 평가방법에서의 규칙 가치는 규칙과 규칙 발견과정에서 사용되는 원본 데이터의 구조와 특성에 의하여 결정(data-driven)된다.

주관적 평가방법은 문제 영역에 비교적 적합한 평가를 내릴 수 있다는 장점을 가지지만, 규칙 평가의 공정성이 떨어질 수 있다. 객관적 평가 방

법은 이와는 반대로, 문제 영역에 독립적인 평가가 이루어질 수 있다는 장점을 가지기 때문에 본 논문에서는 객관적 평가방법을 이용하여 규칙의 흥미도와 유용성을 평가한다.

• 객관적 평가 방법

연관규칙의 객관적 평가치로는 다음과 같은 것들이 있다.

- 흥미도(interest)

신뢰도의 경우, 연관규칙의 중요도 평가치 자체로의 활용은 적합하지 않은데, 그 이유는 다음과 같은 반례가 존재하기 때문이다.

	Y	¬Y	
X	20	10	30
¬X	60	10	70
	80	20	100

지지도와 신뢰도 임계치가 각각 5 %, 30%이고, X와 Y의 빈도가 위와 같은 경우, 규칙 $X \rightarrow Y$는 67%의 높은 신뢰도를 가진다. 하지만, Y의 지지도가 80% 이기 때문에 사실상 이 규칙은 무의미한 규칙이라고 볼 수 있다. 이러한 이유는 신뢰도를 구할 때 후건부 Y의 지지도를 고려하지 않았기 때문이며, 이 점을 보완한 연관규칙의 평가치가 흥미도이다.

규칙 $X \rightarrow Y$ 의 신뢰도가 $\frac{support(X \cap Y)}{support(X)}$ 로 구해지는데 반하여, 흥미도 $I(X \rightarrow Y)$는 다음과 같이 구해진다.

$$I(X \rightarrow Y) = \frac{\frac{support(X \cap Y)}{support(X)}}{support(Y)} = \frac{support(X \cap Y)}{support(X) * support(Y)} \quad \text{식.1}$$

이를 contingency table에서의 빈도로 표현하면,

$$I(X \rightarrow Y) = \frac{f_{11} * N}{f_{1\cdot} * f_{\cdot 1}} \quad \text{식.2}$$

이다.

- 확신도(conviction)

앞에서 살펴본 바와 같이, 규칙 $X \rightarrow Y$ 와 $Y \rightarrow X$ 의 흥미도 수식을 정리하면 같아지므로, 이 경우 두 규칙의 흥미도가 같은 수식에 의하여 계산되며, 결과적으로 $X \rightarrow Y$와 $Y \rightarrow X$의 흥미도가 같다는 잘못된 평가를 한다는 사실을 알 수 있다. 이러한 점을 보완하기 위한 평가치가 확신도이며, 규칙 $X \rightarrow Y$의 확신도 $conviction(X \rightarrow Y)$는 (식.3)과 같이 구해진다.

$$conviction(X \rightarrow Y) = \frac{P(X) * P(\neg Y)}{P(X, \neg Y)} = \frac{f_{1\cdot} * f_{\cdot 0}}{f_{10}} \quad \text{식.3}$$

이와 같은 수식이 가능한 이유는 $X \rightarrow Y \equiv \neg(X \wedge \neg Y)$이므로, 규칙 $X \rightarrow Y$의 강도를 직접 평가하지 않고 $X \wedge \neg Y$의 강도로 대신 평가한다. X와 $\neg Y$의 강도는 $\frac{P(X, \neg Y)}{P(X) * P(\neg Y)}$ 으로 평가할 수 있으므로, $\neg(X \wedge \neg Y)$의 부정을 처리하기 위하여 역수를 취한다.

- χ^2-독립성 검증

χ^2-독립성 검증은 각각의 변수들이 상호간에 독립적이라는 초기 가정하에 예상되는 사건 빈도와 실제 관찰된 사건 빈도를 비교하여 두 사건 간에 존재하는 의존도를 파악하는 방법이다.

$$\chi^2 = \sum_{j,k} \frac{(f_{jk} - E(f_{jk}))^2}{E(f_{jk})} \quad \text{식.4}$$

특정사건 f_{jk}이 일어날 예상 빈도 $E(f_{jk}) = P(f_{jk}) * N$이고, $E(f_{jk}) = f_{j\cdot}/N * f_{\cdot k}/N * N$이므로, 위의 공식을 이전과 같은 2×2의 contingency table에서의 빈도로 표현하면 다음과 같다.

$$\chi^2 = \frac{N(f_{11} f_{00} - f_{01} f_{10})^2}{f_{1\cdot} f_{0\cdot} f_{\cdot 1} f_{\cdot 0}} \quad \text{식.5}$$

χ^2 검증은 항목간의 독립성 여부를 판단할 수 있을 뿐, 규칙 내 항목간의 상관관계의 세기를 나타내지는 않는다. 그러므로, χ^2 검증은 문제의 탐색공간을 줄이는 역할은 할 수 있지만, 규칙간의 순위를 정하기 위한 평가치로는 적당하지 않다.

- 피어슨 상관계수(Pearson's coefficient)

상관계수 평가는 두 변수간의 공분산(covariance)과 표준편차를 이용하여 규칙의 선형성(linearity)을 평가한다. 즉, 공분산 $Cov(X,Y) = E(XY) - E(X)E(Y)$일 때, ρ_{XY}는

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad \text{식.6}$$

과 같이 표현된다.

$P(X)$를 p라고 할 경우, $p = f_{1\cdot}/N$이라고 표현되고, $\sigma_X = \sqrt{p(1-p)}$ 이므로, 위의 수식을 contingency table을 이용하여 바꾸면 다음과 같다.

$$\rho_{XY} = \frac{f_{11} f_{00} - f_{10} f_{01}}{\sqrt{f_{1\cdot} f_{0\cdot} f_{\cdot 1} f_{\cdot 0}}} \quad \text{식.7}$$

- λ-상관계수(Goodman and Kruskal's λ-coefficient)

λ-상관계수 평가는 한 변수를 이용하여 다른 변수의 존재를 예상할 경우, 그 오차가 두 변수가 의존적인 경우보다 그렇지 않은 경우에 더욱 커질 것이라는 이론을 기반으로 한다. 이를 수식으로 표현하면 다음과 같다

$$\lambda_X = \frac{P(E_X) - P(E_X|Y)}{P(E_X)} \quad \text{식.8}$$

다른 배경정보가 없는 경우, X값을 추측할 때의 정확도 X'는 X에 관하여 알려진 확률 중 최고의 값이므로, X'는 $arg(max_k\ P(X_k))$이다. 이 추정을 이용할 때의 오차(E_X) = $1 - P(X')$이므로, $P(E_X) = 1 - max_k\ P(X_k)$이다.

$Y = Y_j$이라는 추가적인 정보를 고려하여 X를 추정할 경우의 X'는 앞에서와 같이 $arg(max_k\ P(X_k|Y_j))$이고, 마찬가지로 $P(E_X|Y_j) = 1 - max_k\ P(X_k|Y_j)$이다. Y가 주어진 경우 X의 예측 오차의 평균은 Y값의 전체 범위를 고려해야하므로, 다음과 같이 표현된다.

$$P(E_X|Y) = 1 - \sum_j \max_k P(X_k, Y_j) \quad \text{식.9}$$

위의 수식을 이용하여 λ-상관계수 평가식을 다시 정리하면 다음과 같다.

$$\lambda_{XY} = \frac{\sum_j \max_k f_{jk} + \sum_k \max_j f_{jk} - \max_k f_{\cdot k} - \max_j f_{j\cdot}}{2N - \max_k f_{\cdot k} - \max_j f_{j\cdot}} \quad \text{식.10}$$

3. 지역적 연관규칙의 유도

3.1 시스템의 전체적인 구조

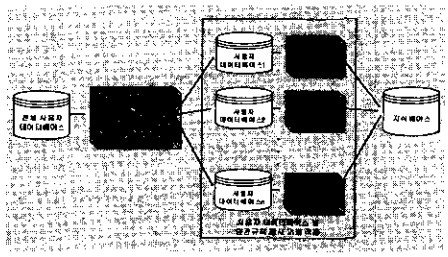

그림 1 . 시스템 구성도

본 논문에서 지역적 연관규칙을 유도하기 위하여 제안한 시스템은 <그림1>과 같은 구조를 가진다. 첫 번째 단계에서는 전체 사용자에 대한 데이터베이스를 대상으로 사용자를 잘 구분하는 속성을 찾고 그 속성에 따라 사용자를 구분하며, 두 번째 단계에서는 이전 단계에서 구해진 각각의 사용자 데이터베이스를 대상으로 연관규칙 탐사 알고리즘을 적용하여 각각의 사용자 그룹에 적합한 연관규칙을 유도한다.

3.2 속성 선택 방법 및 시스템 전체 수행

전체 사용자의 데이터베이스 중에서 고유의 특성을 보유한 사용자 그룹들을 구하기 위하여 양질의 속성을 선택하는 단계이다. 이를 위하여 사용자가 선택한 항목의 상위개념인 항목 클래스를 고려하고, 하위 그룹으로 항목 클래스를 확실히 구분시키는 속성을 선택한다. 예를 들어, 1, 2, 3, ..., 10이 항목의 클래스이고, 속성 a_1과 a_2가 모집단을 <그림2>와 같이 분류한다고 가정하자.

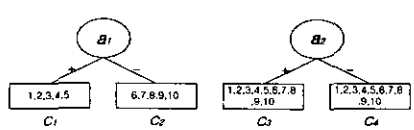

그림 2 . 하위집단으로 분류

<그림2>에서, a_1은 항목 클래스를 명확히 구분시키는데 반하여, a_2는

하위 집합 C_3, C_4으로 모호하게 항목 클래스를 구분한다. 그러므로, 이 경우 양질의 속성은 a_1이고, 이를 사용하여 사용자를 구분한다.

위와 같은 목적을 위하여 각 속성들에 대한 평가가 필요한데, 속성의 평가에는 항목 클래스 엔트로피를 사용한다. 즉, 한 속성(a_n)을 선택하여 항목 클래스를 하위 집단(C_1, C_2)으로 구분한 다음, 모든 항목 클래스에 대하여 각 하위집단별 빈도를 계산하고, 하위집단의 크기를 고려하여 항목 클래스 엔트로피를 계산한다. $C_n = C_1$ or C_2이고, $f(C_n : i)$가 항목집단 i가 하위집단 C_n에 발생한 빈도이며, $|C_n|$가 하위집단 C_n의 크기일 때, *하위집단별 항목 클래스 i의 발생비율* $P(C_n : i) = f(C_n : i) / |C_n|$ 이므로, 이를 이용하여 속성 a_n의 평가치 $S(a_n)$을 수식으로 표현하면 다음과 같다.

$$S(a_n) = -\sum_i \sum_j P(C_j:i) \log_2 P(C_j:i) * Prob(i) \quad \text{식.11}$$

(i는 항목 클래스, j는 분류 가능한 속성 값)

여기서 엔트로피에 곱해지는 $Prob(i)$는 항목 클래스 i의 확률로, 엔트로피의 가중치 역할을 한다. a_1, a_2, ..., a_n 중 S값을 최소로 하는 속성이 선택된다.

이전 단계에서 선택된 속성을 이용하여 모집단을 하위집단으로 구분한다. 각각의 하위집단들에 대하여, 종료조건을 검사하여 다음의 종료조건을 만족하지 않는 하위집단들은 이전의 속성선택 단계를 수행한다.

1. 더 이상의 속성이 없다.
2. 하위집단의 크기가 일정이하이다.
3. 하위집단의 엔트로피가 일정이하이다.

모든 노드의 수행이 끝난 경우, 종료된 각각의 하위집단들에 대하여 연관규칙 탐사를 수행한다. 본 연구에서 연관규칙의 유도를 위하여 사용하는 것은 Apriori 알고리즘[4]이며, 이를 선택한 이유는 알고리즘의 구현과 변형이 용이하고, 기존의 많은 연구가 Apriori를 통하여 이루어져 Apriori을 확장한 많은 연구결과가 존재하기 때문이다.

4. 실험 및 분석

4.1 테스트 데이터 및 실험 방법

실험데이터는 자동차 관련 쇼핑몰에서 얻어진 데이터 집합으로, 707명의 사용자에 대한 회원 정보와 선호하는 자동차 및 자동차의 클래스를 포함한다. 회원의 정보로는 성별, 결혼여부, 직업여부, 도시거주여부 등이 있으며, 전체 트랜잭션의 크기는 2598개이다.

전체 사용자 집합 U 내에서, 공통적인 특성을 가지는 사용자 그룹 Cn들이 존재할 때, 이 Cn의 사용자들을 구분하고, U에서의 연관규칙 $AR(U)$와 Cn에서의 연관규칙 $AR(Cn)$중 어느 연관규칙 집합이 Cn에 더욱 적합한지를 파악하도록 한다. 전체 데이터베이스에서 사용자의 데이터와 사용자가 선택한 자동차/클래스를 이용하여 사용자를 잘 구분하는 속성을 선택하고, 그 속성에 따라 사용자를 구분하여 사용자의 부분집합을 구한다. 구해진 각각의 부분집합들과 전체집합에 대하여 연관규칙 탐사기법을 적용하여 지역적 연관규칙과 전역적 연관규칙을 발견하고, 특정 부분집합에서 여러 평가치를 이용하여 평가함으로써 지역적 연관규칙이 특정 부분집합에 더욱 적합하다는 사실을 증명한다

알고리즘의 종료조건을 하위집단의 크기가 모집단의 1/5 이하이거나 엔트로피가 0.5이하일 경우로 하여, 주어진 데이터베이스를 대상으로 3장의 속성선택 및 사용자 구분 단계를 진행하면 첫 단계에서 sex가 먼저 선택되며, sex가 male인 경우, 다시 marriage가 선택되어 C_1과 C_2가 구해진다. 그러므로 {sex=male & marriage=T} => C_1, {sex=male & marriage=F} => C_2 이다. sex가 female인 C_3의 경우, 자체의 크기가 너무 작기 때문에 종료조건에 의하여 더 이상의 분기는 일어나지 않는다.

4.2 실험결과 및 분석

C_1, C_2, C_3에서 구해진 연관규칙 $AR(C_1)$, $AR(C_2)$, $AR(C_3)$들과, U에서 구해진 연관규칙 $AR(U)$를 각각의 부분집합에 대하여 평가하면 다음과 같다. (단 ms = 지지도 임계치, mc = 신뢰도 임계치)

표 2. $[C_1]$ $AR(C_1)$: $AR(U)$

임계 / 평가치	ms=0.01,mc=0.01		ms=0.03,mc=0.03		ms=0.05,mc=0.05	
	$AR(C_1)$	$AR(U)$	$AR(C_1)$	$AR(U)$	$AR(C_1)$	$AR(U)$
지원 규칙 수	180	56	18	10	2	4
흥미도 평균	2.6465	1.8186	1.9083	1.8166	1.7336	1.8207
확신도 평균	344.6226	304.9084	317.2983	309.2962	328.3176	322.2378
ρ 상관계수 평균	0.1309	0.0789	0.1424	0.1206	0.1795	0.1595
λ 상관계수 평균	0.0068	0.0017	0.0000	0.0000	0.0000	0.0000

표 3. $[C_2]$ $AR(C_2)$: $AR(U)$

계평 / 가치	ms=0.01,mc=0.01		ms=0.03,mc=0.03		ms=0.05,mc=0.05	
	$AR(C_2)$	$AR(U)$	$AR(C_2)$	$AR(U)$	$AR(C_2)$	$AR(U)$
지원 규칙 수	176	76	22	16	6	4
흥미도 평균	1.8710	1.5428	1.5455	1.9266	1.5645	1.7032
확신도 평균	484.8448	454.868	470.6269	500.7308	501.4521	526.3402
ρ 상관계수 평균	0.0878	0.0690	0.1105	0.1607	0.1750	0.2176
λ 상관계수 평균	0.0029	0.0007	0.0000	0.0000	0.0000	0.0000

표 4. $[C_3]$ $AR(C_3)$: $AR(U)$

임 계 / 평가치	ms=0.01,mc=0.01		ms=0.03,mc=0.03		ms=0.05,mc=0.05	
	$AR(C_3)$	$AR(U)$	$AR(C_3)$	$AR(U)$	$AR(C_3)$	$AR(U)$
지원 규칙 수	412	24	8	6	8	0
흥미도 평균	6.7784	2.5036	2.0351	2.0738	2.0351	x
확신도 평균	93.2411	49.0752	53.6125	48.8833	53.6125	x
ρ 상관계수 평균	0.5353	0.1628	0.2453	0.1755	0.2453	x
λ 상관계수 평균	0.2247	0.0000	0.0167	0.0000	0.0167	x

위의 테이블에서 지원 규칙의 수는 평가할 연관규칙들 중, C_n를 대상으로 한 χ^2-독립성 검증을 통과한 규칙들의 수이다. 본 실험에서는 유의수준 α=0.05에서 3.84이상의 값을 갖는 규칙들을 선택하였으며, 거의 모든 경우, $AR(U)$보다 $AR(C_n)$에서 많은 수의 연관규칙이 발생했음을 알 수 있다. 그 외의 다른 평가수치들은 선별된 연관규칙들에 대하여 U와 C_n의 트랜잭션을 대상으로 항목 발생빈도를 조사하여 contingency table을 구하고 2장에서 설명한 평가식에 따라 규칙을 평가한 후, 그들의 평균값을 계산하였다. 평가치의 평균값 역시 대부분의 경우 $AR(U)$보다 $AR(C_n)$의 수치가 높은 것을 알 수 있으므로, 결론적으로, 부분집합 C_n에 적합한 연관규칙의 집합은 $AR(U)$가 아니라 $AR(C_n)$라는 것을 알 수 있다. 전체 테이블 중에서 지지도와 신뢰도 임계치가 올라갈수록 $AR(U)$의 수치가 더 좋은 경우가 간혹 있는데, 그 이유는 전역적으로 높은 평가치를 가지는 한 두 개의 규칙이 존재하기 때문이다. 즉, 고려 대상이 되는 규칙들의 수가 많을 때는 수치의 평균값에 큰 영향을 미치지 못하지만, 규칙의 수가 작아질수록 수치에 큰 영향을 미치기 때문이다.

5. 결론 및 향후과제

본 논문에서는 전체 사용자 데이터베이스 중에서 공통특성을 가지는 부분집합들에 대한 지역적 연관규칙을 얻기 위한 방법을 제안하고, 제안한 방법을 통하여 유도된 연관규칙 집합이 전체 사용자를 대상으로 얻어진 연관규칙보다 특정 사용자 그룹에 더욱 적합하다는 사실을 증명하였다. 연관규칙의 직접적인 활용을 통하여 그 적합도를 증명하기가 매우 어렵기 때문에 연관규칙에 관련된 많은 평가치들을 조사하고 활용하였다. 그러므로, 제안된 기법을 응용하여 그 성능을 직접적으로 평가하기 위한 작업이 추후 필요하다. 또한, 본 논문에서 공통특성을 가지는 사용자그룹을 구하기 위하여 이용한 속성 선택 및 사용자 구분 기법의 분류 정확도 향상을 위한 추가적인 연구가 필요하며, 사용자의 배경정보가 부족한 경우 사용자 그룹을 구하기 위한 기법의 연구가 필요하다.

참고문헌

[1] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements", *Proc. of the Fifth Int'l Conf. on Extending Database Technology (EDBT)*, Avignon, France, March 1996.

[2] R. Srikant ,and R. Agrawal, "Mining quantitative association rules in large relational tables". *In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal*, Canada, June 1996.

[3] Ng. R. T. Lakshmanan, L. Han, J. "Exploratory mining and pruning optimizations of constrained association rules." *SIGMOD-98*, 1998.

[4] R. Agrawal and R. Srikant. "Fast algorithms for mining association rules in large databases". *In VLDB-94*, September 1994.

[5] A. Silberschatz and A. Tuzhilin. "On subjective measures of interestingness in knowledge discovery". *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, 1995.

[6] C. Silverstein, R. Motwani, and S. Brin. "Beyond market baskets: Generalizing association rules to correlations." *In SIGMOD*, 1997.

[7] S. Brin, and R. Motwani, "Dynamic itemset counting and implication rules for market basket data." *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26(2):255, 1997.

[8] P.N. Tan and V. Kumar, "Interestingness Measures for Association Patterns : A Perspective ." *TR00-036. ftp://ftp.cs.umn.edu/dept/users/kumar/WEB/*, 2000.

[9] R.J. Hilderman and H.J. Hamilton. "Knowledge discovery and interestingness measures: A survey." *Technical Report CS 99-04*, Department of Computer Science, University of Regina, October 1999.