사용자 구분에 의한 지역적 연관규칙의 유도
(Deriving Local Association Rules by User Segmentation)

박 세 일 † 이 수 원 ††
(Seil Park) (Soowon Lee)

요 약 연관규칙 탐사기법은 트랜잭션들을 대상으로 항목간 또는 속성간의 연관관계를 발견하는 방법으로, 데이타 집합의 구조를 쉽게 통찰할 수 있다는 장점으로 인하여 활발히 연구되어 왔다. 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 지역적 연관규칙이 전체 사용자를 대상으로 한 전역적 연관규칙보다 해당 부분집합에 더욱 적합하다는 사실을 여러 연관규칙 평가치를 이용하여 평가한다

키워드 : 지역적 연관규칙, 사용자 구분, 연관규칙 평가치

Abstract Association rule discovery is a method that detects associative relationships between items or attributes in transactions. It is one of the most widely studied problems in data mining because it offers useful insight into the types of dependencies that exist in a data set. However, most studies on association rule discovery have the drawback that they can not discover association rules among user groups that have common characteristics. To solve this problem, we segment the set of users into user-subgroups by using feature selection and the user segmentation, thus local association rules in each user-subgroup can be discovered. To evaluate that the local association rules are more appropriated than the global association rules in each user-subgroup, derived local association rules are compared with global association rules in terms of several evaluation measures.

Key words : Local Association Rules, User Segmentation, Association Rules Juteresting Measures

1. 서 론

데이타마이닝은 대규모의 데이타베이스로부터 숨겨진 지식이나 패턴, 새로운 지식 등을 발견하고 이를 실제 비즈니스 의사결정 등을 위한 정보로 활용하고자 하는 작업이다. 데이타마이닝 관련 연구들 중 가장 활발히 연구되어지고 있는 연관규칙 탐사기법은 트랜잭션들을 대상으로 항목간, 또는 속성간의 연관관계를 발견하는 방법이며, 획득된 연관규칙은 교차판매 전략, 상향판매 전략, 카타로그 디자인, 상품 배치, 기만 탐지 등의 목적으로 활용된다. 연관규칙 탐사기법 알고리즘의 기본이라고 할 수 있는 Apriori 알고리즘[1]이 개발된 이래로, 개념구조를 사용하여 보다 포괄적인 의미를 가지는 규칙을 찾아내는 일반화된 연관규칙 탐사[2], 수치 요소와 범주형 요소를 동시에 고려하는 수치 연관규칙 탐사[3], 그리고 반드시 필요한 조건이나 반드시 포함되지 않아야 할 조건을 제약조건으로 지정하여 원하는 규칙만을 결과로 고려하는 제약조건을 이용한 연관규칙 탐사[4] 등 연관규칙에 관련된 많은 연구가 진행되어왔다.

그러나, 현재까지의 연구들은 데이타의 내부 분포를 고려하지 않기 때문에 전체 트랜잭션(고객) 집합 중 공통적인 특성을 가지는 부분집합들이 존재할 경우 그러한 부분집합별 연관규칙을 찾아낼 수 없다는 한계점을 지닌다. 예를 들어, 도서대여 패턴에서 표 1과 같은 연령, 성별별 패턴들이 존재할 경우, 기존의 연관규칙 탐

표 1 도서 대여 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

† 학생회원 : 숭실대학교 컴퓨터학부
lunacy@valentine.ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
swlee@computing.soongsil.ac.kr
논문접수 : 2001년 8월 6일
심사완료 : 2001년 10월 10일

사기법을 이용해서는 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 발생하고, 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.

이러한 문제점을 해결하기 위하여, 본 논문에서는 전처리 개념으로 고객 트랜잭션을 특성에 따라 구분한 후, 각 부분집단별로 연관규칙 탐사기법을 시행하여 각 부분집단에 지역적으로 적합한 연관규칙을 유도한다. 부분집단에 지역적인 연관규칙은 앞에서 언급한 교차판매 전략, 상향판매 전략, 카타로그 디자인, 상품 배치, 기만 탐지 등의 연관규칙의 활용 시에 고객 구분작업을 통하여 고객별로 최상의 서비스를 제공하는데 도움을 줄 수 있다.

2장에서는 본 연구에 대한 이론적 배경에 대해 설명하고, 3장에서는 제안된 기법의 전체적인 구조와 세부적인 사항을 설명하며, 4장에서는 기존의 연관규칙 탐사기법과 제안된 연관규칙 응용기법의 비교실험을 통하여 제안된 기법의 유용성을 고찰한다. 5장에서는 결론 및 향후과제에 대해 언급한다.

2. 이론적 배경

2.1 연관규칙

2.1.1 정의 및 형식화

I가 항목(item)의 집합이고, 각각의 트랜잭션 T가 $T \subseteq I$를 만족하는 항목들의 집합일 때, 연관규칙은 $A \rightarrow B$의 형식으로 표현되며, $A, B \subset I, A \cap B = \varnothing$의 특성을 갖는다. 그리고, 이 규칙의 의미는 데이타베이스의 A를 포함하는 트랜잭션은 B를 같이 포함하는 경향을 보인다는 것이다. 일반적으로, $A \cup B$를 연관규칙 $A \rightarrow B$의 항목집합(itemset)이라고 하며, k개의 항목으로 이루어진 항목집합을 k-항목집합(k-itemset)이라고 한다. 즉, $[X] \in A$ 이고 $[Y] \in B$ 일 때 연관규칙 $[X] \rightarrow [Y]$의 항목집합인 $\{X, Y\}$는 2-항목집합이다. 각각의 규칙은 규칙 내 항목이 전체 트랜잭션에서 차지하는 비중을 나타내는 지지도(support)와 그 규칙의 강도를 나타내는 신뢰도(confidence)를 가진다. 전체 트랜잭션 수가 N이고, T_{AB}가 항목집합 A와 B내의 모든 항목들을 동시에 포함하는 트랜잭션이며, $f(T_{AB})$가 T_{AB}의 빈도일 때, 연관규칙 $A \rightarrow B$의 지지도와 신뢰도는 식1, 식2와 같이 표현된다. 단, 단일항목 집합 A의 지지도 $support(A)$는 $f(A) / N$이다.

$$support(A \rightarrow B) = f(T_{AB})/N \quad (1)$$

$$confidence(A \rightarrow B) = support(A \rightarrow B)/support(A) \quad (2)$$

A와 B의 발생 빈도를 이용하면, 이 두 항목을 포함하는 데이타 집합은 표 2와 같은 2×2의 발생 빈도표로 요약된다.

표 2 2×2 발생 빈도표

	B	$\neg B$	
A	f_{11}	f_{10}	f_{1+}
$\neg A$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

f_{ij}는 각 경우에 해당하는 빈도를 나타내며, $f_{i+} = f_{i1} + f_{i0}$이고, $f_{+j} = f_{1j} + f_{0j}$이다. 즉, f_{i+}와 f_{+j}는 각각 A와 B가 발생한 빈도이므로, 발생 빈도표를 이용하여 식1과 식2를 표현하면 다음과 같다(식3, 식4).

$$support(A \rightarrow B) = f_{11}/N \quad (3)$$

$$confidence(A \rightarrow B) = (f_{11}/N)/(f_{1+}/N) = f_{11}/f_{1+} \quad (4)$$

2.1.2 연관규칙 탐사

연관규칙 탐사문제는 사용자가 지정한 지지도 임계치(support threshold : θ_s)와 신뢰도 임계치(confidence threshold : θ_c)이상의 지지도와 신뢰도를 가지는 모든 연관규칙들을 발견하는 문제이다. 이 문제는 다음과 같이 크게 두 부분으로 나뉘어진다.

(1) θ_s 이상의 지지도를 가지는 모든 빈발항목집합(large itemset)을 발견

(2) 구해진 빈발항목집합들을 이용하여 규칙들을 생성하고, 규칙들 중 신뢰도가 θ_c이상의 것을 선택

(2)의 문제는 메모리 내에서의 간단한 작업에 의하여 이루어질 수 있으나, (1)의 빈발항목집합의 발견은 해결하기 어려운 문제이다. 그 이유는 항목의 수가 m일 때, 가능한 빈발항목집합의 수는 2^m이므로, m이 커질수록 즉, 데이타베이스가 대용량이 될수록 디스크 I/O의 부하가 커지기 때문이다. 그러므로 데이타베이스의 접근 횟수를 최소한으로 줄여 효과적으로 빈발항목을 구하는 알고리즘에 대한 연구가 활발히 진행되어왔으며 그 대표적인 것으로는 항목집합 레벨단위의 점진적 알고리즘인 Apriori[1]가 있다. Apriori 알고리즘은 빈발항목을 효과적으로 구하기 위한 기법으로, k-항목집합의 pruning

을 통하여 검색해야할 다음 단계 k+1-항목집합의 수를 줄여나가는 기법이다. 그 외의 연관규칙탐사 알고리즘으로는 DIC[5], Partition[6], DHP[7], sampling[8]등이 있다. 본 연구에서 연관규칙의 유도를 위하여 사용하는 것은 Apriori 알고리즘이며, 이를 선택한 이유는 알고리즘의 구현과 변형이 용이하고 Apriori을 확장한 많은 연구결과가 존재하기 때문이다.

2.2 연관규칙에 대한 평가

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

연관규칙 탐사로 도출되는 규칙의 수가 상당히 많은 경우, 방대한 양의 규칙들을 평가하여 가지치기하고 분석하는 추가적 작업이 필요하다. 연관규칙의 평가방법은 크게 두 가지 접근방법이 존재하는데, 그것은 주관적 평가방법과 객관적 평가방법이다[9]. 주관적 평가방법에서 규칙의 가치가 규칙을 평가할 사용자에 의하여 결정(user-driven)되는데 반하여, 객관적 평가방법에서의 규칙 가치는 규칙과 규칙 발견과정에서 사용되는 원본 데이타의 구조와 특성에 의하여 결정(data-driven)된다. 규칙의 주관적 평가방법은 문제 영역에 비교적 적합한 평가를 내릴 수 있다는 장점을 갖지만, 지나치게 한 문제 영역에 편중되므로 다른 문제 영역의 적용을 위하여 매번 평가기준을 마련해야하며, 규칙 평가의 공정성이 떨어질 수 있다[10]. 이와는 반대로, 규칙의 객관적 평가 방법은 문제 영역에 독립적인 평가가 이루어질 수 있다는 장점을 가지기 때문에 본 논문에서는 객관적 평가방법을 이용하여 규칙의 흥미도와 유용성을 평가한다.

2.2.2 객관적 평가 방법

▪ 흥미도(interest)

신뢰도의 경우, 연관규칙의 중요도 평가치 자체로의 활용은 적합하지 않은데, 그 이유는 표 3과 같은 반례가 존재하기 때문이다.

표 3 신뢰도 반례

	B	$\neg B$	
A	20	10	30
$\neg A$	60	10	70
	80	20	100

θ_m, θ_c가 각각 5%, 30%이고, A와 B의 빈도가 위와 같은 경우, 규칙 $A \rightarrow B$는 67%의 높은 신뢰도를 가진다. 하지만, B의 지지도가 80%이기 때문에 사실상 이 규칙은 무의미한 규칙이라고 볼 수 있다. 이러한 이유는 신뢰도를 구할 때 후건부 B의 지지도를 고려하지 않았기 때문이며, 이 점을 보완한 연관규칙의 평가치가 흥미도[5]이다. 규칙 $A \rightarrow B$의 신뢰도가 $\frac{support(A \rightarrow B)}{support(A)}$로 구해지는데 반하여, 흥미도 $I(A \rightarrow B)$는 다음과 같이 구해진다(식5).

$$I(A \rightarrow B) = \frac{\frac{support(A \rightarrow B)}{support(A)}}{support(B)} = \frac{support(A \rightarrow B)}{support(A) * support(B)} = \frac{f_{11} * N}{f_{1+} * f_{+1}} \quad (5)$$

▪ 확신도(conviction)

식5를 이용하여 규칙 $A \rightarrow B$와 $B \rightarrow A$의 흥미도를 구하면, 두 규칙의 흥미도가 결국 같은 수식에 의하여 계산된다는 사실을 알 수 있다. 이러한 점을 보완하기 위한 평가치가 확신도[11]이며, 규칙 $A \rightarrow B$의 확신도 $conviction(A \rightarrow B)$는 식6과 같이 구해진다.

$$conviction(A \rightarrow B) = \frac{P(A) * P(\neg B)}{P(A, \neg B)} = \frac{f_{1+} * f_{+0}}{f_{10}} \quad (6)$$

확신도가 식6과 같이 표현될 수 있는 이유는 $A \rightarrow B \equiv \neg(A \wedge \neg B)$이므로, 규칙 $A \rightarrow B$의 강도를 직접 평가하지 않고 $A \wedge \neg B$의 강도로 대신하여 평가하기 때문이다. A와 $\neg B$의 강도는 $\frac{P(A, \neg B)}{P(A) * P(\neg B)}$으로 평가할 수 있으므로, $\neg(A \wedge \neg B)$의 부정을 처리하기 위하여 역수를 취한다.

▪ 피어슨 상관계수(Pearson's coefficient)

피어슨 상관계수 평가[12]는 두 변수간의 공분산(covariance)과 표준편차를 이용하여 규칙의 선형성(linearity)을 평가한다. 즉, 공분산 $Cov(A,B) = E(AB) - E(A)E(B)$일 때, $\rho_{AB} = \frac{Cov(A,B)}{\sigma_A \sigma_B}$이다.

$P(A)$를 p라고 할 경우, $p = f_{1+}/N$으로 표현되고, $\sigma_A = \sqrt{p(1-p)}$이므로, 위의 수식을 발생 빈도표를 이용하여 바꾸면 다음과 같다(식7).

$$\rho_{AB} = \frac{f_{11} f_{00} - f_{10} f_{01}}{\sqrt{f_{1+} f_{0+} f_{+1} f_{+0}}} \quad (7)$$

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

λ-상관계수 평가[12]는 한 변수를 이용하여 다른 변수의 존재를 예상할 경우, 그 오차가 두 변수가 의존적인 경우보다 그렇지 않은 경우에 더욱 커질 것이라는 이론을 기반으로 한다(식8).

$$\lambda_A = \frac{P(E_A) - P(E_A|B)}{P(E_A)} \quad (8)$$

다른 배경정보가 없는 경우, A값을 추측할 때의 정확

도 A'는 A에 관하여 알려진 확률 중 최고의 값이므로, A'는 $arg(max_k\ P(A_k))$이다. 또한, 이 추정을 이용할 때의 오차$(E_A) = 1 - P(A')$이므로, $P(E_A) = 1 - max_k\ P(A_k)$이다. $B = B_l$이라는 추가적인 정보를 고려하여 A를 추정할 경우의 A'는 앞에서와 같이 $arg(max_k\ P(A_k|B_l))$이고, 마찬가지로 $P(E_A|B_l) = 1 - max_k\ P(A_k|B_l)$이다. 그리고, B가 주어진 경우 A의 예측 오차의 평균 $P(E_A|B)$는 B값의 전체 범위를 고려해야하므로, $1 - \sum_j \max_k P(A_k, B_j)$ 이다. 이를 이용하여 λ-상관계수 평가식을 다시 정리하면 다음과 같다(식9).

$$\lambda_{AB} = \frac{\sum_j \max_k f_{jk} + \sum_k \max_j f_{jk} - \max_j f_{+k} - \max_j f_{j+}}{2N - \max_j f_{+k} - \max_j f_{j+}} \quad (9)$$

2.3 속성선택

2.3.1 속성선택의 개념

결정트리와 같은 대부분의 연역적 학습기법들은 예측해야 하는 목적속성과 관련성이 없는 속성들이 많이 존재할수록 그 성능이 떨어지게 된다. 그러므로, 학습모델의 성능을 개선하고, 학습시간을 단축하기 위하여 부적절한 속성들을 제외하고 적절한 속성 만을 고르는 작업이 필요한데, 이러한 작업을 속성선택이라고 한다[13]. 속성선택에 관련된 대부분의 연구는 목적속성이 지정되어 있는 교사학습을 기반으로하여 이루어지고 있으나, 계층적 군집화 기법에 전처리 개념으로 속성선택 기법을 적용하여 군집화 성능을 높인 연구[14]등 비교사학습을 기반으로 한 속성선택 관련연구도 있다.

2.3.2 속성선택 기법 및 분류

속성선택 작업은 개념적으로 크게 필터적 개념(filter approach)과 랩퍼적 개념(wrapper approach)으로 구분되어진다[15]. 필터적 개념 기법은 전처리 개념으로서, 속성이 활용될 주 학습기와는 독립적인 별도의 속성선택 기법을 통해 속성을 선택하는 방법이고, 랩퍼적 개념은 속성선택 시에 속성선택을 위한 탐사, 속성 평가, 연역적 학습기를 동시에 동작시켜 학습을 통해 속성을 선택하는 개념이다. 필터적 개념은 비교적 빠르고 쉽게 동작하지만, 선택된 속성이 학습기에 미치는 영향을 고려하지 않는다는 단점을 갖는다.

이 외에 정보검색에서 문서 검색이나 군집화 방법에서 주로 사용되는 속성선택 기법에는 DF(Document Frequency), TFIDF(Term-Frequency Inverse Document Frequency), MI(Mutual Information), Information Gain 등이 있다[16]. DF는 어느 한 속성(이 경우에는 단어)이 발생하는 전체 문서의 개수를 뜻하는 것으로, 일반적으로 전역적으로 많은 곳에서 발생하는 속성은 의미없는 속성이라고 전제한다. 이와는 반대로, TFIDF는 DF개념에 한 문서 내에서 어떤 속성이 많이 발생할수록 그 속성의 중요도가 증가한다는 사실을 이용한다. 또한, MI는 목적속성과 속성간의 연관성을 측정하기 위한 기법이며 Information Gain은 기계학습 분야에서 엔트로피를 이용하여 속성의 목적속성에 대한 분류성능을 바탕으로 속성의 정보량을 측정하는 기법이다.

3. 지역적 연관규칙의 유도

3.1 지역적 연관규칙

지역적(local) 연관규칙이란 전체 트랜잭션 중 일부분에서 구해진 연관규칙으로, 전체 트랜잭션에서 구해진 전역적(global) 연관규칙에 대비되는 개념이다(그림 1).

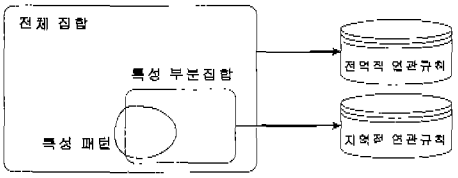


그림 1 전역적 연관규칙과 지역적 연관규칙

대부분의 연관규칙 활용에서는 데이타 분포를 고려하지 않고 전체 트랜잭션(사용자)들을 대상으로 연관규칙 탐사를 수행하여 획득된 전역적 연관규칙들을 이용한다. 이와 같은 방식은 획득된 연관규칙들이 전체 사용자에게 전반적으로 유용하다. 그러나, 데이타 내에 공통특성을 가지는 사용자 부분집합들이 존재 할 경우 그러한 부분집합 구성원들에 대한 차별적인 처리가 불가능하므로 그 구성원에게 적합한 규칙들을 제공할 수 없다는 단점을 지닌다. 즉, 전역적 연관규칙의 탐사에서는 특정 사용자들에게는 충분히 유용하지만, 전체 사용자 측면에서 볼 때 전반적으로 유용하지 못한 규칙들이 탈락될 수 있다. 이러한 규칙들은 그 특정 사용자들을 대상으로 하는 지역적 연관규칙 탐사로 발견된다.

예를 들어 그림 2와 같은 데이타가 존재한다고 가정하자. id는 사용자 구분자이고, 성별(Sex), 결혼여부(Marriage), 직업보유(Job), 도시거주(City) 등은 사용자 정보이며, Product는 사용자가 선택한 항목들의 리스트이다. 전체 데이터를 대상으로 θs, = 0.5, θc = 0.8인 연관규칙 탐사를 할 경우, 항목 {F, H}로 이루어진 규칙 [H] → [F]가 발견된다. 그런데 데이타를 살펴 보면, {Marriage = Y}인 사용자들의 경우 대부분 항목 {A, B}를 선택하였음을 알 수 있다. 즉, {A, B}는 전체에서의 지지도가 임계치보다 작기 때문에 전역적인 연관규

칙 탐사에서 발견되지는 않았지만 성인 사용자들에게는 중요한 항목들이라는 사실을 알 수 있다.

id	Sex	Marri-age	Job	City	Product
1	M	N	Y	N	C, F, H
2	M	N	N	N	F, G, H
3	F	N	N	Y	F ,H, N
4	M	N	Y	Y	F, H, L, M
5	F	N	Y	Y	C, E, F, J
6	F	Y	N	N	B, F, H, L
7	M	Y	N	N	B, F, H
8	M	Y	N	Y	A, B, C, F, H
9	F	Y	Y	N	A, B, C, F
10	F	Y	Y	N	A, B, C, D, G

⇒ [H] → [F]

그림 2 전역적 연관규칙

하지만, 기혼 사용자들을 대상으로 지역적 연관규칙 탐사할 경우(그림 3), 항목 {A, B}의 지지도가 0.6으로 θ_s보다 크고, 이 항목들로 이루어진 규칙 [A] → [B]의 신뢰도가 θ_c보다 크기 때문에 연관규칙으로 발견된다. 그러므로 서비스를 적용할 대상 사용자가 기혼자인 경우에 이러한 지역적 연관규칙들은 유용하게 사용되어 질 수 있다.

id	Sex	Marri-age	Job	City	Product
6	F	**Y**	N	N	**B**, F, H, L
7	M	**Y**	N	N	**B**, F, H
8	M	**Y**	N	Y	**A, B**, C, F, H
9	F	**Y**	Y	N	**A, B**, C, F
10	F	**Y**	Y	N	**A, B**, C, D, G

⇒[A]→[B]

그림 3 지역적 연관규칙

3.2 전체 시스템 구조

본 논문에서 지역적 연관규칙을 유도하기 위하여 제안한 시스템은 그림 4와 같은 구조를 가진다.

전체 시스템은 크게 두 단계로 구분되는데, 첫 번째 단계에서는 공통특성을 가지는 사용자 부분집합들을 구하기 위하여 전체 사용자에 대한 데이타베이스를 대상으로 사용자를 적절히 구분하는 속성들을 선택하고 그 속성에 따라 사용자를 구분한다. 속성선택 방법은 항목을 목적속성(target attribute)으로 하여 결정트리에서 사용되는 엔트로피 개념을 확장하여 사용하였다. 이 때,

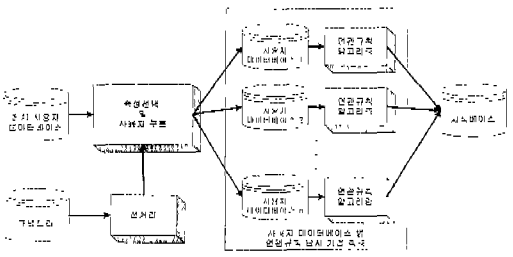


그림 4 시스템 구성도

목적속성으로 사용되는 항목들의 수가 지나치게 많은 경우 효과적인 속성선택이 불가능하므로, 항목을 항목들의 개념트리에서 도출된 적절한 레벨의 항목 개념들로 상향치환하여 사용한다. 선택된 속성에 의하여 분할된 하위 부분집합들이 지정된 종료조건을 만족시키지 않을 경우, 하위 부분집합들에 대하여 이 과정을 반복적으로 수행한다. 두 번째 단계에서는 이전 단계에서 구해진 각각의 사용자 데이타베이스를 대상으로 연관규칙 탐사 알고리즘을 적용하여 각각의 사용자 그룹에 적합한 지역적 연관규칙을 탐사한다. 표 4는 위의 모든 단계들을 정리한 것이다. 단, k는 목적속성으로 사용될 항목 수의 임계치이다.

표 4 지역적 연관규칙의 유도

지역적 연관규칙의 유도 :
① 목적속성으로 사용되는 항목들의 수 > k 인 경우 :
  개념트리를 이용한 목적속성 상향치환;
② 목적속성과 사용자 속성들로 부터 분류속성선택;
③ 선택된 분류속성을 이용한 사용자 구분;
④ 다음의 종료조건을 만족시키지 않는 각각의 사용자 그룹들에 대하여 ②실행;
  종료조건 :
  - 더 이상 선택 가능한 속성이 없다.
  - 사용자 그룹의 크기가 일정이하이다.
  - 사용자 그룹의 순도가 일정이상이다.
⑤ 사용자 그룹 별 연관규칙 탐사;
⑥ 지식베이스 구조화;

3.3 전처리 및 목적속성 상향치환

전체 사용자의 데이타베이스 중에서 고유의 특성을 보유한 사용자 그룹들을 구하기 위하여 양질의 속성을 선택하는 단계이다. 이를 위하여 본 연구에서는 사용자가 선택한 항목의 항목 클래스를 속성선택 시에 목적속성으로 사용하는데, 항목 클래스란 항목 개념트리에서 항목의 상위 개념이다(그림 5).

즉, 항목 집합이 {Atoz, Pride, Tico, ...}라면 이들의

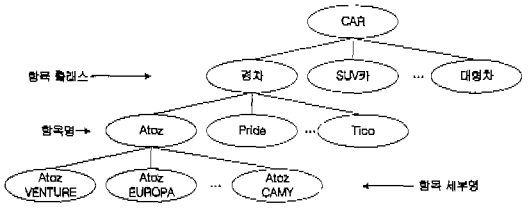


그림 5 개념트리

항목 클래스는 [경차]가 된다. 개념트리 상에서 위쪽으로 향할수록 일반적인 개념이고, 아랫쪽으로 향할수록 구체적인 개념을 가지므로, 그림 5에서와 같이 여러개의 자동차 항목들 {Atoz, Pride, Tico, ...}는 자동차 클래스 [경차]라는 하나의 일반 개념으로 상향 치환될 수 있다 [17]. 항목 자체를 속성선택에 직접 고려하지 않는 이유는 일반적으로 목적속성이 될 항목의 수가 너무 많을 경우 엔트로피 기반의 속성선택 기법이 불가능하거나 정확도가 떨어지기 때문이다.

3.4 속성선택 방법

하위 그룹으로 항목 클래스를 명확히 구분시키는 속성을 선택하기 위하여 각 속성들에 대한 평가가 필요한데, 속성의 평가에는 항목 클래스 엔트로피를 사용한다. 즉, 한 속성(a_n)을 선택하여 항목 클래스를 하위집단 ($C_1, C_2, \ldots, C_j$)으로 구분한 다음, 모든 항목 클래스에 대하여 각 하위집단 별 빈도를 계산하고, 하위집단의 크기를 고려하여 항목 클래스 엔트로피를 계산한다.

$C_* = C_1 \cup C_2 \cup \ldots \cup C_j$이고, $f(C_* : i)$가 항목집단 i가 C_*에서 발생한 빈도이며, $|C_*|$가 하위집단 C_*의 크기일 때, 하위집단별 항목 클래스 i의 발생비율 $P(C_* : i) = f(C_* : i) / |C_*|$이므로, 이를 이용하여 속성 a_n의 평가치 $S(a_n)$을 수식으로 표현하면 식10과 같다. 단, i는 항목 클래스, j는 분류 가능한 속성 값이다.

$$S(a_m) = -\sum_i \sum_j P(C_j; i) \log_2 P(C_j; i) * \mathrm{Prob}(i) \quad (10)$$

여기서 엔트로피에 곱해지는 $Prob(i)$는 항목 클래스 i가 전체 항목들 중에서 차지하는 비중을 뜻하며, 엔트로피의 가중치 역할을 한다.

3.5 지식베이스

위의 전체 단계를 간단히 나타내면 그림 6과 같다.

우선, 입력 데이타 내에서 클래스 속성으로 사용될 항목 클래스를 선정하고, 사용자 정보와 항목 클래스를 이용하여 속성선택과 사용자 구분 단계를 점진적으로 적용하여 위와 같은 결정트리를 구축한다. 또한 결정트리의 단말노드들에 해당하는 사용자 그룹들에 대하여 연관규칙 탐사를 시행하여 지역적 연관규칙을 발견한다.

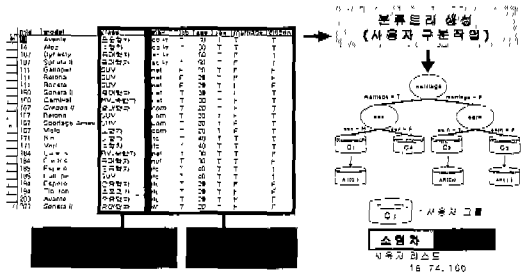


그림 6 전체 수행과정

위의 결정트리에서 단말노드로 표시된 C_1, C_2, C_3, C_4는 특성 부분집합으로 간주되는데, 그 이유는 결정트리가 단말노드의 엔트로피를 최대한 낮추도록하는 성질을 자체적으로 가지기 때문이다. 만약 최적의 경우라면 단말노드의 엔트로피가 0이 되겠지만, 목적속성이 여러 개일 경우에는 그렇지 못하다. 위의 예에서 C_3의 사용자들은 소형차의 항목들을 주로 선택하는 경향이 있음을 알 수 있으며, 만약 C_3 내에서 연관규칙을 탐사할 경우, 소형차에 속하는 항목들로 이루어진 규칙들이 발견될 확률이 높을 것이라는 사실을 추측할 수 있다.

발견된 지역적 연관규칙들은 각각의 사용자 그룹에 해당하는 그룹 속성과 같이 지식화되므로, 지식베이스 내에는 다음과 같은 형태의 규칙이 존재한다(그림 7).

{Marriage = F} & {Earn = F} ⇒ C_3
C_3 : [Tico] → [Atoz]
[Matiz] → [Pride]
[Atoz] → [Tico] & [Matiz]
⋮

그림 7 지식베이스 내의 규칙 형태

이 규칙에서 {Marriage = F} & {Earn = F}는 사용자 그룹의 분류속성이고, [Tico] → [Atoz] 등은 그룹 내에서만 사용되는 지역적 연관규칙이다. 지역적 연관규칙의 활용 시에는 서비스 대상이 되는 사용자의 사용자 속성을 확인하고, 그 사용자 유형에 적합한 사용자 그룹에서 구해진 지역적 연관규칙 집합을 선택하여야 한다.

4. 실험 및 분석

4.1 실험 데이타

4.1.1 실험 데이타1

실험 데이타1은 자동차 관련 쇼핑몰에서 얻어진 데이타 집합으로, 700명의 사용자에 대한 사용자 정보와 선

호하는 자동차 및 자동차의 클래스 정보를 포함한다. 사용자 정보로는 성별, 결혼 여부, 수익직업 여부, 도시거주 여부 등 21개가 있으며, 전체 트랜잭션의 크기는 3000개이다. 실험 데이타1의 스키마는 표 5와 같다.

표 5 실험 데이타1 스키마

속성명	내용	속성명	내용
Mid	사용자 ID	Mail2	신차소개 메일링 (T/F)
Model	선호 자동차 항목	Mail3	레져, 여행 메일링 (T/F)
Class	자동차의 항목 클래스 (6개항목)	Mail4	자동차 스포츠 메일링 (T/F)
Email	이메일 (nominal)	Mail5	자동차 기술 메일링 (T/F)
Job	수익직업 보유 (T/F)	Mail6	신상품 안내 메일링 (T/F)
Age	연령대 (nominal)	Mail7	차량품질정보 메일링 (T/F)
Sex	성별 (T/F)	Diesel	Diesel 선호 (T/F)
Marriage	결혼 (T/F)	LPG	LPG 선호 (T/F)
Citizen	도시거주 (T/F)	A_T	자동 변속기 선호 (T/F)
Area	거주지역 (nominal)	M_T	수동 변속기 선호 (T/F)
Mail1	최신 자동차 뉴스 메일링 (T/F)		

4.1.2 실험 데이타2

실험 데이타2는 비디오 대여점에서 얻어진 데이타 집합으로, 2400명의 사용자에 대한 사용자 정보와 대여한 비디오 및 비디오의 클래스 정보를 포함한다. 사용자 정보로는 성별, 거주지역, 휴대폰 보유, 회원등급 등 8개가 있으며, 전체 트랜잭션의 크기는 36500개이다. 실험 데이타2의 스키마는 표 6과 같다.

표 6 실험 데이타2 스키마

속성명	내용	속성명	내용
Mid	사용자 ID	Age	연령대 (nominal)
Tapename	대여 비디오 명	Hp	핸드폰 보유 (T/F)
Class	비디오 항목 클래스(6개 항목)	Value	회원등급(nominal)
Sex	성별 (T/F)	Delay	지연 내역 (T/F)
Area	거주지역 (nominal)	Marriage	결혼 여부 (T/F)

4.2 실험목적 및 실험방법

실험의 목적은 전체 사용자 집합 U 내에서, 공통적인 특성을 가지는 사용자 그룹 C_n들이 존재할 때, 이 C_n의 사용자들을 구분하고, U에서의 전역적 연관규칙 $AR(U)$와 C_n에서의 지역적 연관규칙 $AR(C_n)$들 중 어느 연관규칙 집합이 현재 C_n의 사용자들과 앞으로 C_n에 속할 사용자들에게 더욱 적합한지를 파악하는 것이다.

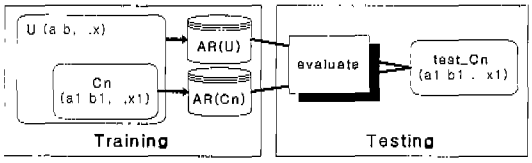


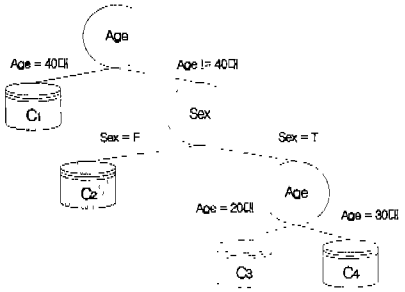
그림 8 실험과정

이를 위하여, 그림 8와 같이 전체 데이타를 트레이닝 집합(80%)과 테스트 집합(20%)로 분할하고, 트레이닝 집합 U 에서 사용자의 데이타와 사용자가 선택한 항목/클래스를 이용하여 사용자를 적절히 구분하는 속성을 선택한 후, 그 속성에 따라 사용자를 구분하여 사용자의 부분집합 C_n을($\subset U$) 구한다. U와 C_n들에 대하여 각각 전역적 연관규칙인 $AR(U)$와 지역적 연관규칙인 $AR(C_n)$을 발견한다. 그리고, 특정 부분집합 C_n에 해당하는 $test_C_n$에서 $AR(U)$와 $AR(C_n)$을 여러 평가치를 이용하여 평가한다.

4.3 실험결과 및 분석

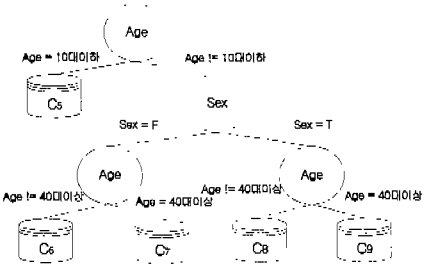
알고리즘의 종료조건을 하위집단의 크기가 모집단의 1/5 이하이거나 엔트로피가 0.85 이하일 경우로 하여 실험 데이타1과 실험 데이타2를 대상으로 3장의 속성선택 및 사용자 구분 단계를 수행하면 실험 데이타1에서는 그림 9, 실험 데이타2에서는 그림 10과 같은 분류규칙이 생성된다. 실험 데이타1의 경우 전체 사용자 집합은 C_1, C_2, C_3, C_4의 네 그룹으로 구분되지만, C_2의 크기가 너무 작으므로 C_2를 제외한 다른 그룹에서 각각 연관규칙을 발견하면, $AR(C_1)$, $AR(C_3)$, $AR(C_4)$의 지역적 연관규칙이 발견된다. 집합의 크기가 작은 경우에도 임계치의 조정을 통하여 규칙을 발견할 수는 있으나, 집합들 간의 비교를 위하여 임계치를 낮은 수치로 통일하였기 때문에 규칙이 발견되지 않는다. 실험 데이타2에서도 전체 사용자가 C_5, C_6, C_7, C_8, C_9의 다섯 그룹으로 구분되지만, 앞에서와 같은 이유로 C_7를 제외한 나머지 집합들에서 $AR(C_5)$, $AR(C_6)$, $AR(C_8)$, $AR(C_9)$를 발견하였다.

C_1, C_3, C_4, C_5, C_6, C_8, C_9에서 구해진 연관규칙 $AR(C_1)$, $AR(C_3)$, $AR(C_4)$, $AR(C_5)$, $AR(C_6)$, $AR(C_8)$, $AR(C_9)$들과, U_1과 U_2에서 구해진 연관규칙 $AR(U_1)$, $AR(U_2)$를 각각의 테스트용 부분집합에 대하여 평가하면 다음과 같다(표 8 ~ 14). 결과 테이블에서 규칙수는 평가할 연관규칙들 중 C_n를 대상으로 한 지지도와 신뢰도가 각각 θ_s, θ_c이상인 규칙들의 수이며, 그 외의 다른 평가수치들은 선별된 연관규칙들에 대하여 U와 C_n의 트랜



- {Age = 40대} ⇒ C_1
- {Age != 40대} & {Sex = T} & {Age = 20대} → C_3
- {Age != 40대} & {Sex = T} & {Age = 30대} → C_4

그림 9 실험 데이타1의 분류규칙



- {Age = 10대이하} → C_5
- {Age != 10대이하} & {Sex = F} & {Age != 40대이상} ⇒ C_6
- {Age != 10대이하} & {Sex = T} & {Age != 40대이상} → C_8
- {Age != 10대이하} & {Sex = T} & {Age = 40대이상} ⇒ C_9

그림 10 실험 데이타2의 분류규칙

재션을 대상으로 항목 발생빈도를 조사하여 발생 빈도표를 구성하고 2장에서 설명한 평가식에 따라 규칙을 평가한 후, 그들의 평균값을 계산한 수치들이다(표 7). 단, 아래의 테이블에서 U_1와 U_2는 각각 실험 데이타1과 실험 데이타2에서의 전체집합을 나타내며, U는 실험 데이타1, 2를 고려하지않은 개념적인 전체집합을 나타낸다.

표 7 결과 테이블 예

지지도 \ 임계치	θ_s=0.01, θ_c=0.01	
	$AR(C_1)$	$AR(U_1)$
규칙수	292	62
흥미도	3.7498	1.7552
확신도	55.1629	36.7943
p상관계수	0.4005	0.1731
λ상관계수	0.1653	0.0322

⇒C_1과 U_1에서 실험 임계치 θ_s=0.01, θ_c=0.01일때 발견한 $AR(C_1)$와 $AR(U_1)$를 $test_C_1$에서 각각 평가한 수치를 나타낸다. 이 경우 $AR(C_1)$에서는 292개의 연관규칙이, $AR(U_1)$에서는 62개의 연관규칙이 $test_C_1$에 적합한한 것으로 판명되었다.

표 8 실험 데이타1 : test_C_1에서 $AR(C_1)$과 $AR(U_1)$ 평가

지지도 \ 임계치	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01	
	$AR(C_1)$	$AR(U_1)$	$AR(C_1)$	$AR(U_1)$	$AR(C_1)$	$AR(U_1)$
규칙수	292	62	60	28	36	16
흥미도	3.7498	1.7552	1.9458	1.4973	1.7316	1.6552
확신도	55.1629	36.7943	38.9605	34.535	37.1679	37.1426
p상관계수	0.4005	0.1731	0.2092	0.1367	0.1796	0.1741
λ상관계수	0.1653	0.0322	0.0448	0.0311	0.0367	0.0544

표 9 실험 데이타1 : $test_C_3$에서 $AR(C_3)$와 $AR(U_1)$ 평가

지지도 \ 임계치	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01	
	$AR(C_3)$	$AR(U_1)$	$AR(C_3)$	$AR(U_1)$	$AR(C_3)$	$AR(U_1)$
규칙수	70	62	40	30	8	6
흥미도	2.4159	2.2803	1.9719	1.8736	1.9294	1.645
확신도	53.3263	53.4042	49.6355	49.6648	50.4683	48.1696
p상관계수	0.2328	0.228	0.1907	0.1896	0.2358	0.1955
λ상관계수	0.0233	0.0238	0.0095	0.0108	0	0

표 10 실험 데이타1 : $test_C_4$에서 $AR(C_4)$와 $AR(U_1)$ 평가

지지도 \ 임계치	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01	
	$AR(C_4)$	$AR(U_1)$	$AR(C_4)$	$AR(U_1)$	$AR(C_4)$	$AR(U_1)$
규칙수	260	162	60	34	14	12
흥미도	2.3419	1.9878	1.6973	1.6831	1.6115	1.6765
확신도	107.900	102.329	101.590	100.324	104.511	106.961
p상관계수	0.1497	0.1359	0.1266	0.1369	0.1642	0.1822
λ상관계수	0.0075	0.0062	0.0061	0.0067	0.0162	0.0189

표 11 실험 데이타2 : $test_C_5$에서 $AR(C_5)$와 $AR(U_2)$ 평가

지지도 \ 임계치	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01		θ_s=0.04, θ_c=0.01	
	$AR(C_5)$	$AR(U_2)$	$AR(C_5)$	$AR(U_2)$	$AR(C_5)$	$AR(U_2)$	$AR(C_5)$	$AR(U_2)$
규칙수			224	26	38	8	10	4
흥미도			4.6003	2.4039	3.6175	2.5833	3.9555	2.5
확신도			62.9107	46.1513	55.5789	48.4167	58.4667	47.8333
p상관계수			0.4917	0.2983	0.4016	0.323	0.4342	0.3127
λ상관계수			0.2218	0.0818	0.1656	0.0977	0.1897	0.0955

표 12 실험 데이타2 : *test_C6*에서 $AR(C_6)$와 $AR(U_2)$ 평가

임계치 / 지지도	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01		θ_s=0.04, θ_c=0.01	
	$AR(C_6)$	$AR(U_2)$	$AR(C_6)$	$AR(U_2)$	$AR(C_6)$	$AR(U_2)$	$AR(C_6)$	$AR(U_2)$
규칙수	6270	1490	660	250	184	68	66	26
흥미도	3.8634	3.2073	2.8637	2.6788	2.4327	2.2447	2.2932	2.2037
확신도	200.602	186.555	183.374	179.126	174.651	169.300	174.368	175.837
ρ상관계수	0.2541	0.2242	0.2363	0.226	0.2299	0.2185	0.2441	0.2404
λ상관계수	0.0265	0.0183	0.0167	0.0157	0.0118	0.0088	0.0097	0.0167

표 13 실험 데이타2 : *test_C8*에서 $AR(C_8)$과 $AR(U_2)$ 평가

임계치 / 지지도	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01		θ_s=0.04, θ_c=0.01	
	$AR(C_8)$	$AR(U_2)$	$AR(C_8)$	$AR(U_2)$	$AR(C_8)$	$AR(U_2)$	$AR(C_8)$	$AR(U_2)$
규칙수	48064	7354	3314	664	594	150	188	40
흥미도	3.9953	3.218	3.1272	2.6362	2.7258	2.4052	2.4532	2.4959
확신도	294.951	276.142	287.813	257.985	272.045	253.085	260.919	262.608
ρ상관계수	0.2687	0.253	0.2834	0.2621	0.2956	0.2772	0.299	0.3118
λ상관계수	0.0359	0.0328	0.0427	0.0267	0.0379	0.0247	0.0282	0.0307

표 14 실험 데이타2 : *test_C9*에서 $AR(C_9)$과 $AR(U_2)$ 평가

임계치 / 지지도	θ_s=0.01, θ_c=0.01		θ_s=0.02, θ_c=0.01		θ_s=0.03, θ_c=0.01		θ_s=0.04, θ_c=0.01	
	$AR(C_9)$	$AR(U_2)$	$AR(C_9)$	$AR(U_2)$	$AR(C_9)$	$AR(U_2)$	$AR(C_9)$	$AR(U_2)$
규칙수					86	42	22	16
흥미도					2.4203	2.1762	1.9545	1.9875
확신도					32.5795	31.0595	27.8182	30.375
ρ상관계수					0.3851	0.3509	0.2786	0.3182
λ상관계수					0.1331	0.12	0.0909	

거의 모든 경우, $AR(U)$보다 $AR(C_n)$에서 많은 수의 연관규칙이 발생하였으며, 평가치 역시 대부분의 경우 $AR(U)$보다 $AR(C_n)$의 수치가 높은 것을 알 수 있다. $AR(U)$에 속하지 않으면서 $AR(C_n)$에만 속하는 규칙들은 본 논문에서 제안한 기법으로만 발견할 수 있는 지역적 연관들이다. 앞에서 밝힌 바와 같이, 결과 테이블들의 지원 규칙수를 살펴보면 모든 경우 $AR(C_n)$에서 더욱 많은 연관규칙들이 발견되었으므로 모든 사용자 집합들에서 고유의 지역적 연관규칙들이 발견되었음을 알 수 있다. 또한 이러한 지역적 연관규칙들의 평가치를 살펴보면, 전역적으로 구해진 연관규칙들에 비하여 우수하다는 사실을 알 수 있으므로, 실험 데이타1과 2에서 부분집합 C_n(혹은 *test_C*$_n$)에 적합한 연관규칙의 집합은 $AR(U)$가 아니라 $AR(C_n)$라는 것을 알 수 있다. 그러나 이러한 평가는 *test_C*$_n$을 기반으로 이루어진 것들이기 때문에, 지역적 연관규칙들이 전체집합에 대하여 전반적으로 우수하다는 것은 아니다.

실험 수행 시에는 θ_s만을 변화시키고 θ_c는 변화시키지 않았는데, 그 이유는 θ_s가 연관규칙 탐사 과정 중에 발생되는 항목집합의 구성에 직접적으로 관여하기 때문이다. 즉, θ_s를 변화시키는 경우 결과로 얻을 수 있는 연관규칙 집합이 큰폭으로 변화하므로, 다양한 시점에서의 실험결과를 얻을 수 있다. 이에 반하여 θ_c는 연관규칙 탐사 단계에서 구성된 규칙 선택을 위한 강도평가 만을 담당하므로 θ_s에 비하여 상대적으로 적은 결과의 변화만을 일으킨다. 실제로 θ_c를 변화시키며 실험 한 경우에도 위의 결과들과 큰 차이를 보이지 않았으므로 본문에는 수록하지 않았다. 또한, 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.

표 15는 표 8의 결과 테이블에서 θs, θc가 0.03 일 때 발견된 규칙들이다. C_1은 40대 이상의 사용자들의 집합인데, 이 사람들은 대부분 Van이나 SUV와 같은 가정용, 다목적 차량에 많은 관심을 가진 것으로 조사되었

표 15 실험 데이타1의 C_1과 U_1에서 발견된 연관규칙

θ_s = 0.03, θ_c = 0.01	
$AR(C_1)$	$AR(U_1)$
규칙수 : 36	칙수 : 16
Sonata II] → [Carnival], [Carnival] → [Sonata II] [Sonata II] → [Musso], [Musso] → [Sonata II] [Verna] → [Avante], [Avante] → [Verna] [Verna] → [EF Sonata], [EF Sonata] → [Verna] [Santamo] → [Starex], [Starex] → [Santamo] [Musso] → [Magnus], [Magnus] → [Musso] [Trajet] → [Carnival], [Carnival] → [Trajet] [Musso] → [Starex], [Starex] → [Musso] [Musso] → [Korando], [Korando] → [Musso] [Rezzo] → [Verna], [Verna] → [Rezzo] [Santamo] → [Musso], [Musso] → [Santamo] [Santamo] → [Carstar], [Carstar] → [Santamo] [Musso] → [Carnival], [Carnival] → [Musso] [Carnival] → [Carstar], [Carstar] → [Carnival] [Carens] → [Rezzo], [Rezzo] → [Carens] [Pride] → [Carnival], [Carnival] → [Pride] [Musso] → [Grandeur XG], [Grandeur XG] → [Musso] [Chairman] → [Grandeur XG] [Grandeur XG] → [Chairman]	[Carnival] → [Musso] [Musso] → [Carnival] [EF Sonata] [Trajet] [Trajet] → [EF Sonata] [Trajet] → [Carnival] [Carnival] → [Trajet] [Avante] → [Verna] [Verna] → [Avante] [EF Sonata] → [Grandeur XG] [Grandeur XG] → [EF Sonata] [Carens] → [Rezzo] [Rezzo] → [Carens] [Trajet] → [Rezzo] [Rezzo] → [Trajet] [Carnival] → [Rezzo] [Rezzo] → [Carnival]

표 16 실험 데이타2의 C_2와 U_2에서 발견된 연관규칙

θ_s = 0.03, θ_c = 0.01	
$AR(C_2)$	$AR(U_2)$
규칙수 : 38	칙수 : 8
[식스센스] → [리베라메], [리베라메] → [식스센스] [반칙왕] → [동감], [동감] → [반칙왕] [동감] → [반칙왕, 식스센스], [반칙왕] → [동감, 식스센스] [식스센스] → [동감, 반칙왕], [동감, 반칙왕] → [식스센스] [식스센스, 동감] → [반칙왕], [식스센스, 반칙왕] → [동감] [동감] → [퍼펙트스톰], [퍼펙트스톰] → [동감] [동감] → [식스센스], [식스센스] → [동감] [미션임파서블2] → [엑스맨], [엑스맨] → [미션임파서블2] [미션임파서블2] → [동감], [동감] → [미션임파서블2] [반칙왕] → [식스센스], [식스센스] → [반칙왕] [으랏차차스모부] → [식스센스], [식스센스] → [으랏차차스모부] [미션임파서블2] → [할로우맨], [할로우맨] → [미션임파서블2] [에린브로코비치] → [식스센스], [식스센스] → [에린브로코비치] [블레스더차일드] → [식스센스], [식스센스] → [블레스더차일드] [미션임파서블2] → [퍼펙트스톰], [퍼펙트스톰] → [미션임파서블2] [미션임파서블2] → [포스트맨블루스] [포스트맨블루스] → [미션임파서블2] [퍼펙트스톰] → [미션임파서블2, 동감] [미션임파서블2] → [퍼펙트스톰, 동감] [동감] → [퍼펙트스톰, 미션임파서블2] [퍼펙트스톰, 미션임파서블2] → [동감] [퍼펙트스톰, 동감] → [미션임파서블2] [동감, 미션임파서블2] → [퍼펙트스톰]	[반칙왕] → [식스센스] [식스센스] → [반칙왕] [미션임파서블2] → [할로우맨] [할로우맨] → [미션임파서블2] [미션임파서블2] → [엑스맨] [엑스맨] → [미션임파서블2] [패트리어트] → [미션임파서블2] [미션임파서블2] → [패트리어트]

다. 때문에, 발견된 규칙들을 살펴보면 상당수의 Van, SUV 차량들의 연관규칙이 발생되었음을 알수 있다. 물론 $AR(C_1)$와 $AR(U_1)$ 자체에는 이외에도 많은 규칙들이 발견되었으나, 위의 내용은 $test_C_1$라는 채에 의하여 걸러져 나온 규칙들로, $test_C_1$에 적합한 규칙들만이 선택된 결과이다. 즉, $AR(C_1)$에서는 36개, $AR(U_1)$에서는 16개의 규칙이 발견되으며, $AR(U)$에 속하는 대다수의 규칙들이 $AR(C_1)$에 속하므로, $AR(C_1)$내의 $AR(U_1)$에 속하지 않는 규칙들이 제안된 기법을 통하여 얻을수 있는 새로운 규칙들이라는 사실을 알 수 있다. 그러므로 이 규칙들은 U_1전체에서는 전반적으로 유용성이 떨어지지만, C_1에 해당하는 사용자들에게는 유용한 지역적 연관규칙이다.

그리고 표 16은 표 11의 결과 테이블에서 θ_s, θ_c가 0.03 일 때 발견된 규칙들이며, 이 경우에도 많은 지역적 연관규칙들이 발생했음을 볼 수 있다.

5. 결론 및 향후과제

본 논문에서는 전체 사용자 데이타베이스 중에서 공통특성을 가지는 부분집합들에 대한 지역적 연관규칙을 추출하기 위한 방법을 제안하고, 제안한 방법을 통하여 유도된 연관규칙 집합이 전체 사용자를 대상으로 얻어진 연관규칙보다 특정 사용자 그룹에 더욱 적합하다는 사실을 실험적으로 평가하였다. 또한, 연관규칙은 그 적합도를 증명하기가 매우 어렵기 때문에 연관규칙에 관련된 많은 평가치들을 조사하고 활용하였다.

그러므로, 제안된 기법을 응용하여 그 성능을 직접적으로 평가하기 위한 작업이 추후 필요하다. 본 논문에서 공통특성을 가지는 사용자 그룹을 구분하기 위하여 활용한 속성선택 및 사용자 구분 기법은 개념트리 상에서의 목적항목 상향치환을 통한 결정트리 구성 기법이다. 따라서, 기존의 결정트리의 단점을 그대로 가질 수 있으므로 이에 대한 보완과 개선이 필요하다. 예를 들어, 결정트리의 단말노드로 사용자들을 구분하였는데 결정트리의 특성 상 유사한 공통 특성을 가지는 사용자들을 서로 다른 노드로 배정할 수 있다. 그러므로, 결정트리의 구축이 종료된 후 단말노드들을 대상으로 특성을 확인하고 비슷한 특성을 보이는 단말노드들의 경우 병합하는 등의 처리가 필요하다. 또한, 본 논문에서는 개념트리 상에서 임의대로 연관규칙을 추출할 레벨(자동차 명, 비디오 명)과 분류정보를 추출하기 위한 레벨(항목 클레스)을 선정하였으나, 개념트리 구조의 이용을 통하여 수행 레벨을 자동적으로 선정할 수 있도록 하는 연구가 필요하다.

참 고 문 헌

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

[2] 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.

[3] 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.

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

[5] 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.

[6] A. Sarasere, E. Omiecinsky, and S. Navathe. "An efficient algorithm for mining association rules in large databases". *In 21st Int'l Conf. on Very Large Databases (VLDB)*, ZTrich, Switzerland, Sept. 1995.

[7] J. S. Park, M. S. Chen, and P. S. Yu. "Efficient parallel data mining for association rules." *In Proc. 1995 International Conference on Information and Knowledge Management*, Baltimore, MD, November 1995.

[8] H. Toivonen. "Sampling large databases for association rules." *In Proc. 22nd VLDB*, 1996.

[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.

[10] 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.

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

[12] 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.
[13] Deogun, J., Choubey, S. "On Feature Selection and Effective Classifiers." *Journal of ASIS* 49, 423-434, May 1998.
[14] L. Talavera. "Feature selection as a preprocessing step for hierarchical clustering." *In Proceedings of the 16th International Conference on Machine Learning*, pages 389-397. Morgan Kaufmann, 1999.
[15] R. Kohavi, John G. "Wrappers for Feature Subset Selection." *In Artificial Intelligence journal, special issue on relevance*, Vol. 97, No. 1-2 (pp. 273-324), 1997.
[16] D. Lewis, "Feature selection and feature extraction for text categorization." *Proceedings of Speech and Natural Language Workshop* (pp. 212--217). San Francsico: Morgan Kaufman, 1992.
[17] 문홍기. "배경지식을 활용한 연관규칙 발견 및 확장." 숭실대학교 석사학위 논문, 2000.

박 세 일
1998년 숭실대학교 인공지능학과 학사. 2001년 숭실대학교 컴퓨터학과 석사. 관심분야는 Data Mining, CRM, Planning, Machine Learning, AI

이 수 원
1982년 서울대학교 자연과학대학 계산통계학과 학사. 1984년 한국과학기술원 전산학과 석사. 1984년 ~ 1987년 LG 중앙연구소 주임 연구원. 1994년 University of Southern California 전산학과 박사. 1995년 ~ 현재 숭실대학교 컴퓨터학부 조교수. 관심분야는 Data Mining, Agent, Machine Learning, CRM, Expert System, AI