Proceedings of Joint Conference of The Korean Data Analysis Society and Korean Data And Information Science Society, October 17-18, 2003, pp. 197-213

Relation for the Measure of Association and the Criteria of Association Rule in Ordinal Database

Hee Chang Park[1], *Ho Soon Lee*[2]

Abstract

One of the well-studied problems in data mining is the search for association rules. The goal of association rule mining is to find all the rules with support and confidence exceeding some user specified thresholds. In this paper we consider the relation between the measure of association and the criteria of association rule for ordinal data.

Keywords : data mining, association rule, measure of association

1. 서론

연관성 규칙은 각 항목간의 연관성을 반영하는 규칙으로서 둘 또는 그 이상의 품목들 사이의 지지도(support), 신뢰도(confidence), 향상도(lift)를 기반으로 하여 미리 결정된 최소지지도 및 최소 신뢰도 이상의 의미 있는 규칙을 찾아내는 데이터마이닝(data mining) 기법 중의 하나이다. 연관성 규칙은 Agrawal 등(1993)에 의해 처음 소개된 이후, Agrawal 등(1994)은 후보 항목 집합을 구성하고, 발생 빈도 수를 계산하고 난 후에 사용자가 정의한 최소 지지도를 가지고 빈발 항목 집합들을 결정하는 Apriori 및 AprioriTid 알고리즘을 제안하였다. Park 등(1995)은 데이터베이스를 중복되지 않는 크기로 분할하고 한번에 한 개의 분할 영역만을 고려하여 그 안에서 빈발 항목 집합을 생성하는 partitioning 알고리즘을 제안하였으며, Tovivonen (1996)은 무작위로 선정된 표본을 가지고 빈발 항목 집합들을 찾은 후에 그 결과를 데이터베이스의 나머지 부분을 가지고 증명하는 sampling 알고리즘을 제안하였다. 또한 Cheung 등 (1996)은 갱신된 데이터베이스에서 이전에 빈발 항목으로 다루어 졌던 항목 집합에 대해서 데이터베이스 스캔 과정을 생략하는 FUP(fast update) 알고리즘에 대한 연구를 하였고, Sergey 등(1997)은 데이터가 데이터베이스 전체에 골고루 퍼져있을 경우 적절한 간격을 이용한 알고리즘이 연구되었다. Liu 등(1999)은 후보 항목 집합들을 효율적으로 작게 구하여 이것을 기초로 전체 트랜잭션(transaction)의 크기와 개수를 줄여나가는 DHP(direct hashing and

[1]Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea, E-mail : hcpark@sarim.changwon.ac.kr

[2]Graduate Student Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea

pruning) 알고리즘을 제안하였다. Saygin 등(2002)은 트랜잭션에 있는 데이터를 unknown으로 대치시키고 이때의 최소 지지도와 최대 지지도의 범위를 조정하면서 고려할 항목 집합의 수를 줄이는 방법을 제안하였다. 이들 연구들은 주로 대용량 데이터베이스에서 효율적인 연관성을 찾아내고자 하는 처리속도 향상을 위한 알고리즘을 중심으로 연구가 진행되었다.

한편, 연관성 규칙에서 카이제곱 통계량의 제안은 Silverstein 등(1997)에 의해 이루어졌으며, 일반적인 연관성 규칙이 두 항목 집합의 동시발생만을 고려하여 지지도와 신뢰도를 계산함으로써, 두 항목 집합의 비발생에 대한 고려를 하지 않아서 발생할 수 있는 문제를 지적하기도 하였다. Song(2002)은 기존의 연관성 규칙에서 연관의 근거로 삼고 있는 최소 신뢰도를 통계적 관점에서 접근하여 파악함으로써 보다 객관적인 연관성 규칙의 연관 기준값을 제안하였다. 기존의 연구 결과는 신뢰도가 1 이상인 경우에 한해서 의미 있는 연관성 규칙으로 보며 이를 통해 관련성 여부를 파악했다. 그러나 기존의 연구에서는 관련성 여부는 파악할 수 있었으나 어느 정도의 관련성을 가지고 있는지의 여부는 파악하기 힘들뿐만 아니라 여러 항목들간의 연관성의 비교가 불가능하다는 한계가 있었다.

본 연구에서는 순위형 자료에서의 연관성 측도와 연관성 규칙의 평가 기준과의 관계를 제시함으로써 관련성의 정도가 어느 정도인지를 통계적 관점에서 접근하여 파악함으로써 보다 객관적인 연관성 규칙의 관련성 정도를 제시하고자 한다.

본 연구의 2절에서는 기존 데이터 마이닝에서의 연관성 규칙을 간략하게 설명하며 3절에서는 순위형 자료에서의 연관성 측도의 종류와 그 의미에 대해서 파악하며, 4절에서는 순위형 자료에서의 연관성 측도와 연관 규칙의 평가기준과의 관계를 규명하고자 한다. 5절에서는 모의 실험을 통해 본 연구에서 제시한 연관성 측도와 연관 규칙의 평가기준과의 관계를 비교 분석하며, 마지막 6절에서는 본 연구의 결론 및 향후 연구 과제에 대해서 언급하였다.

2. 연관성 규칙의 평가 기준

연관성 규칙의 평가기준에는 지지도(support), 신뢰도(confidence), 향상도(lift) 등이 있다. 지지도는 항목 집합 X와 항목 집합 Y가 동시에 발생한 거래의 비율을 의미하며, 전체 거래수를 항목 집합 X와 항목 집합 Y가 동시에 발생한 거래의 수를 나누어서 구한다. 지지도는 식 (2.1)과 같이 정의된다.

$$S_{(X \Rightarrow Y)} = \frac{X\text{와 } Y\text{를 동시에 구매한 거래수}}{\text{전체거래수}} = P(X \cap Y) \tag{2.1}$$

신뢰도는 항목집합 X가 포함된 거래 비율 중 항목 집합 X와 Y가 동시에 포함된 거래의 비율을 말한다. 신뢰도는 식 (2.2)와 같이 정의된다.

$$C_{(X \Rightarrow Y)} = P(Y|X) = \frac{P(X \cap Y)}{P(X)} \tag{2.2}$$

향상도는 실제거래발생 확률을 각 항목집합의 거래가 독립적일 경우 그 거래가 동시에 발생할 예상기대확률로 나눈 것을 의미한다. 향상도는 식 (2.3)과 같이 정의된다.

$$L_{(X \Rightarrow Y)} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X)P(Y)} \tag{2.3}$$

향상도의 값이 1보다 크면 두 아이템이 동시에 발생한 거래확률이 예상확률보다 더 크므로 향상도의 값이 1이상인 경우에 의미 있는 관련성 규칙이라고 할 수 있다. 그 값은 많은 관련성 규칙의 조합 중에서 의미 있는 관련성규칙을 발견하기 위한 기준으로서 많이 사용된다.

연관성 규칙은 항목 집합간의 지지도와 신뢰도를 계산하여, 미리 분석자에 의해서 정해진 최소 지지도 및 최소 신뢰도를 모두 만족하는 두 항목 집합의 규칙을 강한 연관 규칙을 가진 것으로 판단한다. 연관성 규칙을 해석할 때 신뢰도의 값이 크면 좋지만 신뢰도가 크다고 모두 최선의 연관성 규칙은 아니다. 두 항목의 집합의 기본적인 지지도가 어느 정도 수준 이상일 경우만 고려해야 한다. 또한 신뢰도와 지지도는 자주 발생되는 항목 집합에 대해서는 연관성 때문이 아니라 우연하게 높게 나올 수도 있으므로 향상도를 잘 관찰할 필요가 있다. 그러나 이러한 연관성 규칙은 두 항목 집합간의 관련성의 여부만을 판단할 수 있다. 어느 정도의 관련성이 있는지, 그리고 $X \Rightarrow Y$의 연관성과 $X \Rightarrow Z$의 연관성 중 어느 것이 더 연관성이 높은지에 대해서는 판단하기 힘들다. 그러나 연관성의 측도는 관련성뿐만 아니라 수치로서 연관성의 정도를 객관적으로 나타내준다. 그러므로 $X \Rightarrow Y$의 연관성과 $X \Rightarrow Z$의 연관성 중 어느 것이 더 연관성이 높은지 판단할 수 있을 뿐만 아니라 관련성의 강함과 약함의 정도를 객관적으로 파악할 수 있다.

그러므로 본 논문에서는 순위형 자료에서 사용할 수 있는 연관성 측도와 연관성 규칙의 평가기준과의 관계를 규명하여 연관성 규칙의 관련성의 정도를 통계적 기법을 사용하여 보다 객관적인 연관성 규칙의 연관성 정도를 제시하고자 한다.

3. 순위형 자료에서의 연관성 측도

연관성 측도(Measure of Association)는 두 변수간의 관련성뿐만 아니라 두 변수간의 연관성의 정도를 측정하기 위한 측도를 말한다. 분할표에서 두 변수가 모두 순위형인 경우에는 다음과 같이 정의되는 일치쌍과 불일치쌍의 수에 의해서 그 연관성의 강도와 방향이 결정된다. 두 변수의 연관성은 연관성의 측도의 값이 -1에 가까울수록 두 변수가 음의 상관을 갖고 +1에 가까울수록 두 변수가 양의 상관을 갖는 것으로 해석한다. 본 절에서는 다음과 같은 2×2 분할표(contingency table)를 고려한다.

<표 1> 2×2 분할표

		Y		계
		L	T	
X	L	a_{LL}	a_{LT}	X_L
	T	a_{TL}	a_{TT}	X_T
계		Y_L	Y_T	S_{XY}

본 절에서는 일치쌍의 수를 N_c, 불일치쌍의 수를 N_d, X에 대한 동순위 쌍의 수를 N_X, Y에 대한 동순위 쌍의 수를 N_Y, X 및 Y에 대한 동순위쌍의 수를 N_{XY}, 전체 쌍의 수를 N, 전체 데이터 수를 S_{XY}로 정의한다. 위의 분할표에서 N_c, N_d, N_X, N_Y, N_{XY}, N을 기술하면 다음과 같다.

$$\begin{aligned} N_c &= a_{LL} \times a_{TT} \\ N_d &= a_{LT} \times a_{TL} \\ N_X &= a_{LL} \times a_{LT} + a_{TL} \times a_{TT} \\ N_Y &= a_{LL} \times a_{TL} + a_{LT} \times a_{TT} \\ N_{XY} &= N - N_c - N_d - N_X - N_Y \\ N &= \frac{S_{XY}(S_{XY}-1)}{2} = N_c + N_d + N_X + N_Y + N_{XY} \end{aligned} \tag{3.1}$$

두 변수의 연관성은 연관성 측도의 값이 -1에 가까울수록 두 변수가 음의 상관을 갖고 +1에 가까울수록 두 변수가 양의 상관을 갖는 것으로 해석된다. 순위형 자료에서의 연관성 측도에는 켄달의 타우 a, 켄달의 타우 b, 켄달의 타우 c, 굿맨-크루스칼의 감마, 소머스의 D 등이 있다.

켄달의 타우 a는 일치쌍의 수와 비일치쌍의 수의 차로 측정되어진다. 타우 a는 계산하기 가장 쉽고 이해하기 쉬우나 실제로 연관성 측도로 자주 사용되지는 않는다. 이는 타우 a는 일치쌍과 비일치쌍의 수의 차로만 측정되어지므로 동순위의 쌍이 대부분인 분할표에서의 연관성 측도로 사용하기에는 부적절한 단점을 가지고 있기 때문이다. 타우 a는 식 (3.2)와 같이 표현된다.

$$\tau_a = \frac{(N_c - N_d)}{N}, \tag{3.2}$$

여기서 $-1 \le \tau_a \le 1$이다.

켄달의 타우 b는 타우 a의 단점인 동순위를 고려하지 않은 것을 보완한 것으로 행변수와 열변수가 같은 척도를 가지고 있는 경우에 적합하다. 타우 b는 식 (3.3)과 같이 표현된다.

$$\tau_b = \frac{N_c - N_d}{\sqrt{N_c + N_d + N_X}\sqrt{N_c + N_d + N_Y}}, \tag{3.3}$$

여기서 $-1 \le \tau_b \le 1$이다.

켄달의 타우 c는 타우 b와 타우 a가 가지고 있던 단점을 보완한 것으로 행변수와 열변수가 다른 척도를 가지는 경우 타우 b보다 더 적절한 연관성 측도이다. 타우 c는 식 3.4와 같이 표현된다.

$$\tau_c = \frac{N_c - N_d}{S^2_{XY}[(m-1)/2m]}, \tag{3.4}$$

여기서 m은 min(R,C), $-1 \le \tau_c \le 1$이다.

감마는 동순위를 이루는 쌍을 제외한 쌍 중에서 일치쌍의 수와 비일치쌍의 수의 확률의 차로 측정되어진다. 감마 또한 타우 a가 가지고 있던 문제를 동순위의 쌍을 제거함으로써 해결했다. 감마는 서로 대칭인 순위형 변수에 적용하며 가장 일반적으로 쓰이는 연관성 측도이다. 감마는 식 (3.5)와 같이 표현된다.

$$\gamma = \frac{(N_c - N_d)}{(N_c + N_d)}, \tag{3.5}$$

여기서 $-1 \le \gamma \le 1$이다.

소머스의 D는 켄달의 타우 b의 수정으로 비대칭적 관계를 고려한 연관성 측도이다. 소머스 D는 식 (3.6)과 같이 표현된다.

$$\begin{aligned} D(Y|X) &= \frac{N_c - N_d}{N_c + N_d + N_X} \\ D(X|Y) &= \frac{N_c - N_d}{N_c + N_d + N_Y} \end{aligned} \tag{3.6}$$

여기서 $D(Y|X)$는 변수 X를 기초로 변수 Y를 예측한 것이고, $D(X|Y)$는 변수 Y를 기초로 변수 X를 예측한 것이다.

4. 연관규칙의 평가 기준과 연관성 측도와의 관계

4.1 기본 가정

본 절에서는 동시발생 빈도와 연관성 측도와의 관계, 그리고 연관규칙의 평가기준과 연관성 측도와의 관계를 규명하기 위해 다음과 같은 2×2 분할표(contingency table)를 가정한다.

<표 2> 기본 가정을 위한 2×2 분할표

		Y		합
		L	T	
X	L	a	$x_1 - a$	x_1
	T	$y_1 - a$	$t-(x_1+y$	$x_0 = t - x_1$
합		y_1	$y_0 = t - y_1$	t

각 셀은 다음 조건을 만족해야 한다.

$$\begin{aligned} &0 \le a \le t \\ &0 \le x_1 - a \le t \\ &0 \le y_1 - a \le t \\ &0 \le t-(x_1+y_1)+a \le t \\ &0 \le a \le x_1 \\ &0 \le a \le y_1 \end{aligned} \tag{4.1}$$

식(4.1)은 다음과 같이 다시 표현될 수 있다.

$$\begin{aligned} 0 \le a \le x_1 \\ 0 \le a \le y_1 \\ (x_1+y_1)-t \le a \le x_1+y_1 \end{aligned} \tag{4.2}$$

<표 2>에서 t, x_1, y_1은 단 한번의 데이터베이스 스캔으로 알 수 있으므로, 사전에 이미 알려져 있다고 가정한다. 또한 위에서 t, x_1, y_1이 알려져 있는 경우 항목 집합 X와 Y의 동시 발생 빈도인 a 에 따라 다른 셀 들은 위의 <표 2>과 같이 계산되어 표현되어 질 수 있다. 여기서 동시 발생 빈도 a에 따른 다른 셀을 계산한 것은 연관성 규칙에서 관심을 가지는 것은 동시에 구매한 경우이므로 동시 발생 빈도를 a라고 정의했다.

본 연구에서는 t, x_1, y_1을 고정시키고, 항목 집합 X와 Y의 동시 발생 빈도 a와 순위형 자료에서의 연관성 측도와의 관계를 알아보고, 기존의 연관규칙으로는 제시할 수 없었던 연관성의 정도를 연관성 측도를 통해 객관적인 연관성 정도의 기준을 제시하고자 한다.

4.2 동시발생 빈도와 연관성 측도와의 관계

(1) 동시발생빈도와 켄달의 타우 a 와의 관계

동시발생빈도와 켄달의 타우 a와 관계는 식 (4.3)과 같이 표현된다.

$$\tau_a = \left(\frac{2}{t-1}\right)a - \frac{2x_1y_1}{t(t-1)} \tag{4.3}$$

식 (4.3)에서 보듯이 동시발생빈도와 켄달의 타우 a와의 관계는 선형 관계를 가짐을 알 수 있다.

(2) **동시발생빈도와 켄달의 타우 b와의 관계**

동시발생빈도와 켄달의 타우 b와 관계는 식 (4.4)와 같이 표현된다.

$$\tau_b = \left(\frac{t}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \right) a - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \tag{4.4}$$

식 (4.4)에서 보듯이 동시발생빈도와 켄달의 타우 b와의 관계는 선형 관계를 가짐을 알 수 있다.

(3) **동시발생빈도와 켄달의 타우 c와의 관계**

동시발생빈도와 켄달의 타우 c와 관계는 식 (4.5)와 같이 표현된다.

$$\tau_c = \frac{4}{t} a - \frac{4x_1 y_1}{t^2} \tag{4.5}$$

식 (4.5)에서 보듯이 동시발생빈도와 켄달의 타우 c와의 관계는 선형 관계를 가짐을 알 수 있다.

(4) **동시발생빈도와 굿맨-크루스칼의 감마와의 관계**

동시발생빈도와 굿맨-크루스칼의 감마와의 관계는 식 (4.6)과 같이 표현된다.

$$\gamma = \frac{(ta - x_1 y_1)}{2a^2 + (t - 2x_1 - 2y_1)a + x_1 y_1} \tag{4.6}$$

식 (4.6)에서 보듯이 동시발생빈도와 굿맨-크루스칼의 감마와의 관계는 비선형 관계를 가짐을 알 수 있다.

(5) **동시발생빈도와 소머스의 D와의 관계**

동시발생빈도와 소머스의 D와 관계는 식 (4.7)과 식 (4.8)과 같이 표현된다.

$$D(Y|X) = \left(\frac{t}{ty_1 - y_1^2} \right) a - \frac{x_1 y_1}{ty_1 - y_1^2} \tag{4.7}$$

$$D(X|Y) = \left(\frac{t}{tx_1 - x_1^2} \right) a - \frac{x_1 y_1}{tx_1 - x_1^2} \tag{4.8}$$

식 (4.7), 식 (4.8)에서 보듯이 동시발생빈도와 소머스 D와의 관계는 선형 관계를 가짐을 알 수 있다.

4.3 연관규칙의 평가기준과 연관성 측도와의 관계

연관성 규칙의 평가기준에는 지지도(support), 신뢰도(Confidence), 향상도(Lift) 세 가지 기준이 있다는 것을 앞장에서 설명했다. 이 장에서는 이러한 연관성 평가 기준과 연관성 측도와의 관계를 유도하고자 한다.

먼저, 연관성 규칙을 동시발생빈도와의 관계식으로 나타내면 식 (4.9)와 같이 정의된다.

$$\begin{aligned} S_{(X \Rightarrow Y)} &= \frac{a}{t} \\ C_{(X \Rightarrow Y)} &= \frac{a}{x_1} \\ L_{(X \Rightarrow Y)} &= \frac{ta}{x_1 y_1} \end{aligned} \tag{4.9}$$

(1) 연관규칙의 평가기준과 켄달의 타우 a 와의 관계

연관규칙의 평가기준과 켄달의 타우 a와 관계는 식 (4.10)과 같이 표현된다.

$$\tau_a = \begin{cases} \left(\frac{2t}{t-1}\right) S_{(X \Rightarrow Y)} - \frac{2x_1 y_1}{t(t-1)} \\ \left(\frac{2x_1}{t-1}\right) C_{(X \Rightarrow Y)} - \frac{2x_1 y_1}{t(t-1)} \\ \frac{2x_1 y_1}{t(t-1)} L_{(X \Rightarrow Y)} - \frac{2x_1 y_1}{t(t-1)} \end{cases} \tag{4.10}$$

식 (4.10)에서 보듯이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 a와 선형관계를 가짐을 알 수 있다.

(2) 연관규칙의 평가기준과 켄달의 타우 b와의 관계

연관규칙의 평가기준과 켄달의 타우 b와의 관계는 식 (4.11)과 같이 표현된다.

$$\tau_b = \begin{cases} \frac{t^2 S_{(X \Rightarrow Y)}}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \\ \frac{tx_1 C_{(X \Rightarrow Y)}}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \\ \frac{x_1 y_1 L_{(X \Rightarrow Y)}}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} - \frac{x_1 y_1}{\sqrt{(ty_1 - y_1^2)(tx_1 - x_1^2)}} \end{cases} \tag{4.11}$$

식 (4.11)에서 보듯이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 b와 선형

관계를 가짐을 알 수 있다.

(3) 연관규칙의 평가기준과 켄달의 타우 c와의 관계

연관규칙의 평가기준과 켄달의 타우 c와의 관계는 식 (4.12)와 같이 표현된다.

$$\tau_c = \begin{cases} 4S_{(X \Rightarrow Y)} - \dfrac{4x_1y_1}{t^2} \\ \dfrac{4x_1}{t} C_{(X \Rightarrow Y)} - \dfrac{4x_1y_1}{t^2} \\ \dfrac{4x_1y_1}{t^2} L_{(X \Rightarrow Y)} - \dfrac{4x_1y_1}{t^2} \end{cases} \tag{4.12}$$

식 (4.12)에서 보듯이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 켄달의 타우 c와 선형 관계를 가짐을 알 수 있다.

(4) 연관규칙의 평가기준과 굿맨-크루스칼의 감마와의 관계

연관규칙의 평가기준과 굿맨-크루스칼의 감마와의 관계는 식 (4.13)와 같이 표현된다.

$$\gamma = \begin{cases} \dfrac{(t^2S_{(X \Rightarrow Y)} - x_1y_1)}{2t^2S^2_{(X \Rightarrow Y)} + (t - 2x_1 - 2y_1)tS_{(X \Rightarrow Y)} + x_1y_1} \\ \dfrac{(tx_1C_{(X \Rightarrow Y)} - x_1y_1)}{2x_1^2C^2_{(X \Rightarrow Y)} + (t - 2x_1 - 2y_1)x_1C_{(X \Rightarrow Y)} + x_1y_1} \\ \dfrac{(y_1x_1L_{(X \Rightarrow Y)} - x_1y_1)}{\dfrac{2\,x_1{y_1}^2}{t^2} L^2_{(X \Rightarrow Y)} + \dfrac{(t - 2x_1 - 2y_1)x_1y_1}{t} L_{(X \Rightarrow Y)} + x_1y_1} \end{cases} \tag{4.13}$$

식 (4.13)에서 보듯이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 굿맨-크루스칼의 감마와 비선형 관계를 가짐을 알 수 있다.

(5) 연관규칙의 평가기준과 소머스 D와의 관계

연관규칙의 평가기준과 소머스 $D(Y|X)$와의 관계는 식 (4.14)와 같이 표현된다.

$$D(Y|X) = \begin{cases} \dfrac{t^2S_{(X \Rightarrow Y)}}{ty_1 - y_1^2} - \dfrac{x_1y_1}{ty_1 - y_1^2} \\ \dfrac{tx_1C_{(X \Rightarrow Y)}}{ty_1 - y_1^2} - \dfrac{x_1y_1}{ty_1 - y_1^2} \\ \dfrac{x_1y_1L_{(X \Rightarrow Y)}}{ty_1 - y_1^2} - \dfrac{x_1y_1}{ty_1 - y_1^2} \end{cases} \tag{4.14}$$

연관규칙의 평가기준과 소머스 $D(X|Y)$와의 관계는 식 (4.15)와 같이 표현된다.

$$D(X|Y)=\begin{cases}\dfrac{t^2S_{(X\Rightarrow Y)}}{tx_1-x_1^2}-\dfrac{x_1y_1}{tx_1-x_1^2}\\ \dfrac{tx_1C_{(X\Rightarrow Y)}}{tx_1-x_1^2}-\dfrac{x_1y_1}{tx_1-x_1^2}\\ \dfrac{x_1y_1L_{(X\Rightarrow Y)}}{tx_1-x_1^2}-\dfrac{x_1y_1}{tx_1-x_1^2}\end{cases} \tag{4.15}$$

식 (4.14), (4.15) 에서 보는 바와 같이 연관규칙의 평가기준인 지지도, 신뢰도, 향상도 모두 소머스 D와 선형 관계를 가짐을 알 수 있다.

기존의 연구 결과는 신뢰도가 1 이상인 경우에 한해서 의미 있는 연관성 규칙으로 보며 이를 통해 관련성 여부를 파악했으나 본 연구 결과를 통해 어느 정도의 관련성을 가지고 있는지의 여부를 파악할 수 있을 뿐만 아니라 $X\Rightarrow Y$의 연관성과 $X\Rightarrow Z$ 의 연관성의 비교가 가능하며 양과 음의 관련성 여부도 파악할 수 있다.

5. 모의 실험

본 절에서는 4절에서 논의된 관계식을 기반으로, <표 3>을 이용하여 모의 실험을 실시하였다. 모든 실험은 t, x_1, y_1을 $t=100, x_1=45, y_1=30$으로 고정한 후 결과를 살펴보았다. 첫 번째 실험은 카이제곱 검정 결과가 유의한 경우에서의 동시발생 빈도와 연관성 측도와의 관계를 규명하고, 두 번째 실험은 연관 규칙의 평가기준과 각각의 연관성 측도와의 관계를 규명하고자 한다. 세 번째 실험은 연관규칙의 각각의 평가기준과 연관성 측도와의 관계를 규명하고자 한다.

<표 3> 모의 실험을 위한 2×2 분할표

		Y		합
		T	L	
X	T	a	45 - a	45
	L	30 - a	25 + a	55
합		30	70	100

먼저 동시발생빈도, 연관 규칙의 평가기준, 그리고 연관성 측도값을 계산하면 <표4>와 같이 얻어진다.

<표 4>에서 보는 바와 같이 $t=100, x_1=45, y_1=30$ 인 경우, a가 취할 수 있는 정수 값의 범위는 식 (5.1)과 같다.

$$0\le a\le 30 \tag{5.1}$$

또한 유의수준 $\alpha=0.05$ 하에서 $\chi^2(1)=3.84146$이고, 위의 <표 4>에서 카이제곱 검정 결과가 유의한 경우의 a 의 범위는 식 (5.2)와 같다.

<표 4> 동시발생빈도, 연관 규칙의 평가기준과 연관성 측도

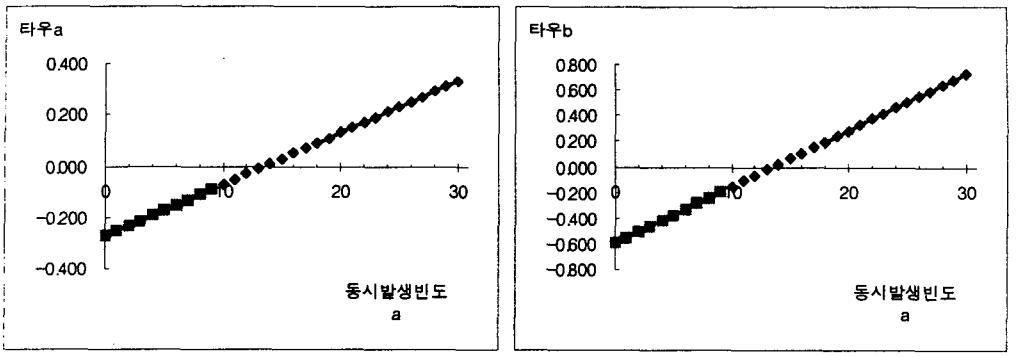
동시발생 빈도	S(X=>Y)	C(X=>Y)	L(X=>Y)	타우 a	타우 b	타우 c	감마	소머스 D(Y|X)	소머스 D(X|Y)
0	0	0.000	0.000	-0.273	-0.592	-0.540	-1.000	-0.643	-0.545
1	0.01	0.022	0.074	-0.253	-0.548	-0.500	-0.960	-0.595	-0.505
2	0.02	0.044	0.148	-0.232	-0.504	-0.460	-0.914	-0.548	-0.465
3	0.03	0.067	0.222	-0.212	-0.461	-0.420	-0.862	-0.500	-0.424
4	0.04	0.089	0.296	-0.192	-0.417	-0.380	-0.804	-0.452	-0.384
5	0.05	0.111	0.370	-0.172	-0.373	-0.340	-0.739	-0.405	-0.343
6	0.06	0.133	0.444	-0.152	-0.329	-0.300	-0.668	-0.357	-0.303
7	0.07	0.156	0.519	-0.131	-0.285	-0.260	-0.592	-0.310	-0.263
8	0.08	0.178	0.593	-0.111	-0.241	-0.220	-0.510	-0.262	-0.222
9	0.09	0.200	0.667	-0.091	-0.197	-0.180	-0.424	-0.214	-0.182
10	0.1	0.222	0.741	-0.071	-0.154	-0.140	-0.333	-0.167	-0.141
11	0.11	0.244	0.815	-0.051	-0.110	-0.100	-0.240	-0.119	-0.101
12	0.12	0.267	0.889	-0.030	-0.066	-0.060	-0.145	-0.071	-0.061
13	0.13	0.289	0.963	-0.010	-0.022	-0.020	-0.048	-0.024	-0.020
14	0.14	0.311	1.037	0.010	0.022	0.020	0.048	0.024	0.020
15	0.15	0.333	1.111	0.030	0.066	0.060	0.143	0.071	0.061
16	0.16	0.356	1.185	0.051	0.110	0.100	0.235	0.119	0.101
17	0.17	0.378	1.259	0.071	0.154	0.140	0.325	0.167	0.141
18	0.18	0.400	1.333	0.091	0.197	0.180	0.410	0.214	0.182
19	0.19	0.422	1.407	0.111	0.241	0.220	0.490	0.262	0.222
20	0.2	0.444	1.481	0.131	0.285	0.260	0.565	0.310	0.263
21	0.21	0.467	1.556	0.152	0.329	0.300	0.635	0.357	0.303
22	0.22	0.489	1.630	0.172	0.373	0.340	0.698	0.405	0.343
23	0.23	0.511	1.704	0.192	0.417	0.380	0.755	0.452	0.384
24	0.24	0.533	1.778	0.212	0.461	0.420	0.806	0.500	0.424
25	0.25	0.556	1.852	0.232	0.504	0.460	0.852	0.548	0.465
26	0.26	0.578	1.926	0.253	0.548	0.500	0.892	0.595	0.505
27	0.27	0.600	2.000	0.273	0.592	0.540	0.926	0.643	0.545
28	0.28	0.622	2.074	0.293	0.636	0.580	0.955	0.690	0.586
29	0.29	0.644	2.148	0.313	0.680	0.620	0.980	0.738	0.626
30	0.3	0.667	2.222	0.333	0.724	0.660	1.000	0.786	0.667

$$a \leq 9.0317, \quad a \geq 17.9683 \tag{5.2}$$

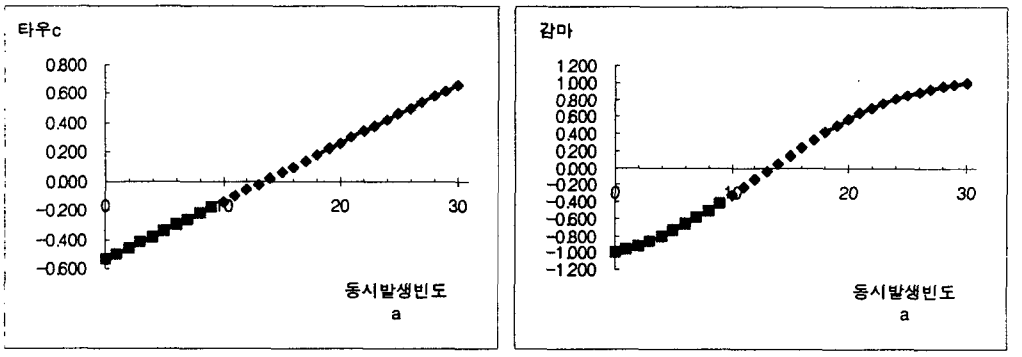
따라서 조건 (5.1)과 조건 (5.2)에서 a는 정수값만 허용되므로 a가 취할 수 있는 값은 식 (5.3)과 같다.

$$0 \leq a \leq 9, \quad 18 \leq a \leq 30 \tag{5.3}$$

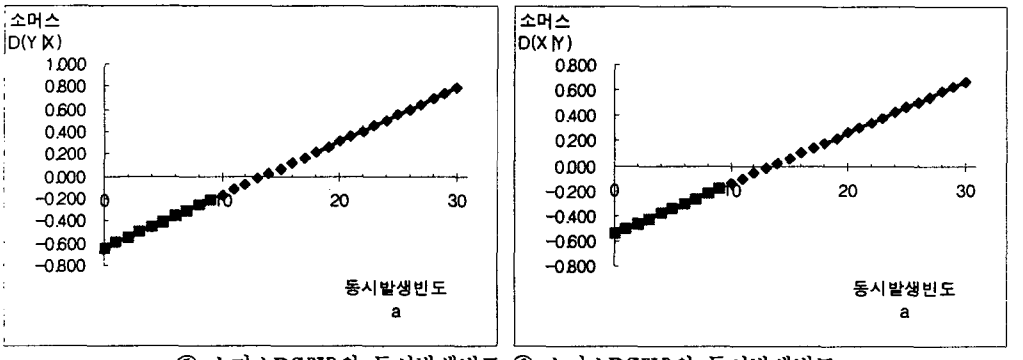
<표 4>에 대해 동시발생빈도와 연관성 측도와의 관계를 그림으로 나타내면 <그림2>와 같다. 여기서 연한 선으로 되어 있는 부분이 유의한 경우이다.



① 타우 a와 동시발생빈도 ② 타우 b와 동시발생빈도



③ 타우 c와 동시발생빈도 ④ 감마와 동시발생빈도

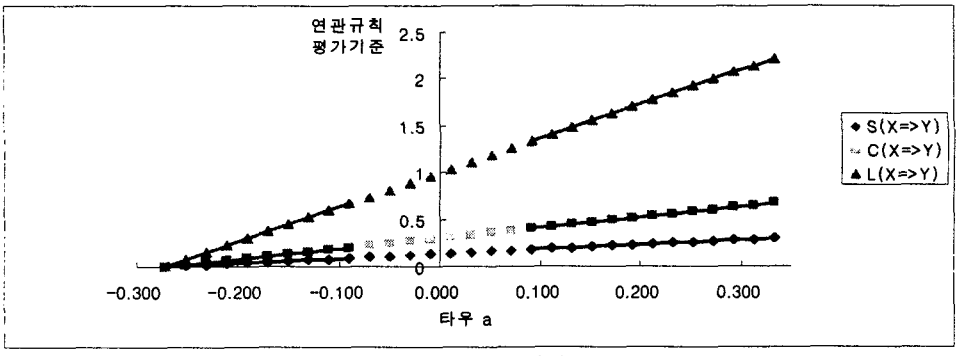


⑤ 소머스D(Y|X)와 동시발생빈도 ⑥ 소머스D(X|Y)와 동시발생빈도

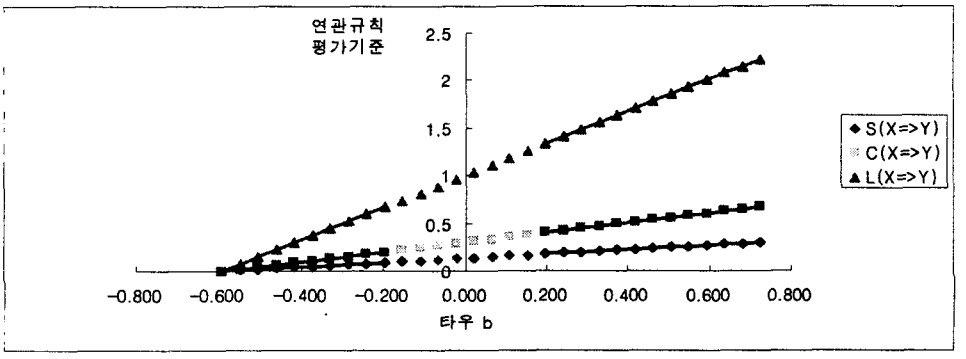
<그림 2> 동시발생빈도에 따른 연관성 측도값의 변화

위의 <그림 2>에서 보듯이 감마와 동시발생빈도와의 관계를 제외하고는 타우 a, 타우 b, 타우 c, 소머스 D(Y|X), 소머스 D(X|Y)와 동시발생빈도와의 관계는 선형관계를 가짐을 알 수 있다. 즉 동시발생빈도가 증가할수록 연관성 측도들의 값이 감마를 제외하고는 모두 선형적으로 증가함을 알 수 있다.

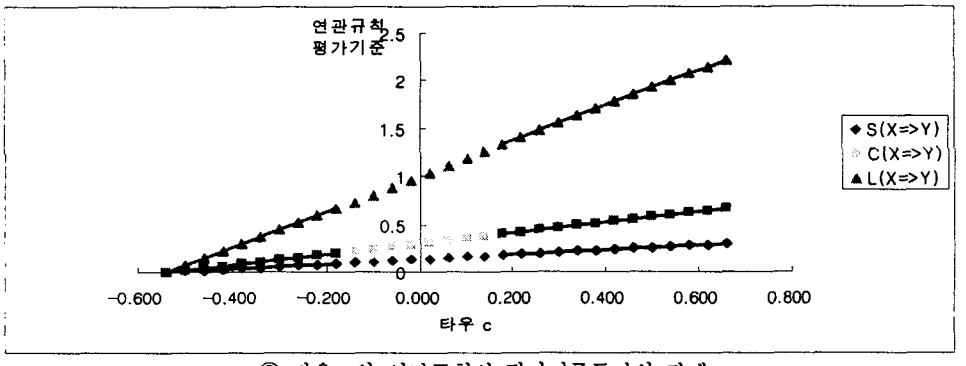
<표 4>에 대해 연관규칙의 평가기준과 연관성 측도와의 관계를 그림으로 표현하면 <그림3>과 같다. 여기서 연한 선으로 되어 있는 부분이 유의한 경우이다.



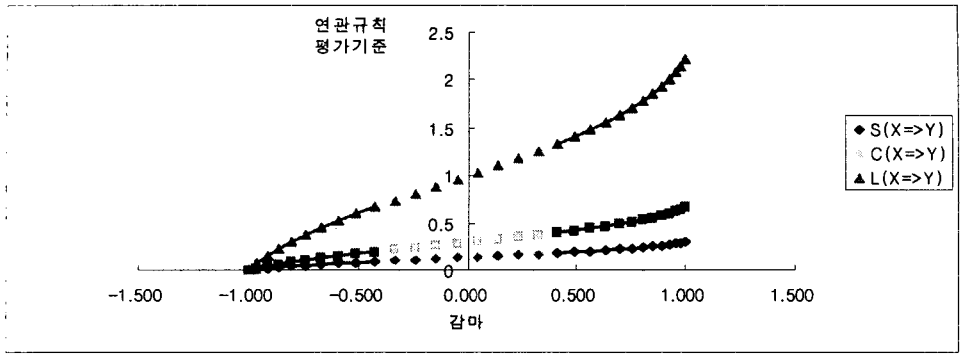
① 타우 a와 연관규칙의 평가기준들과의 관계



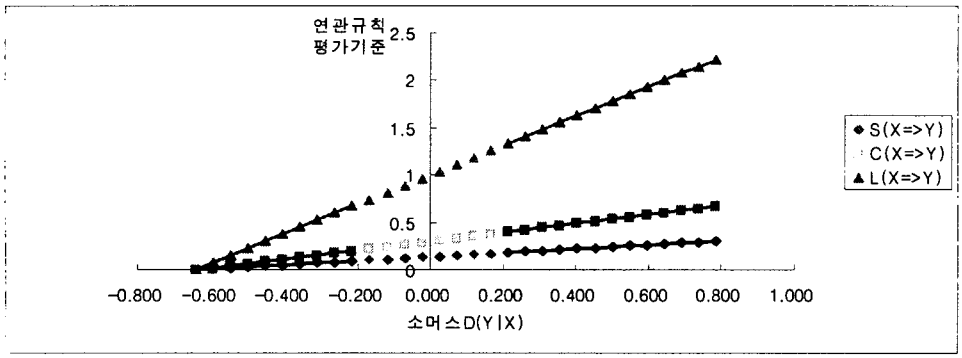
② 타우 b와 연관규칙의 평가기준들과의 관계



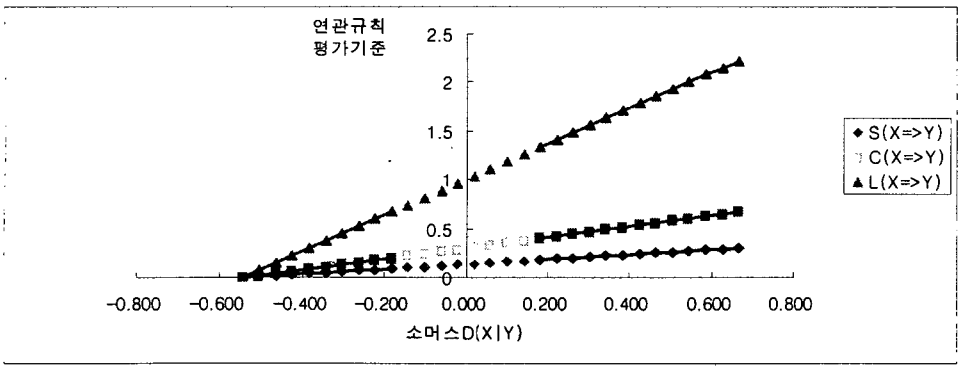
③ 타우 c와 연관규칙의 평가기준들과의 관계



④ 감마와 연관규칙의 평가기준들과의 관계



⑤ 소머스 D(Y|X)와 연관규칙의 평가기준들과의 관계



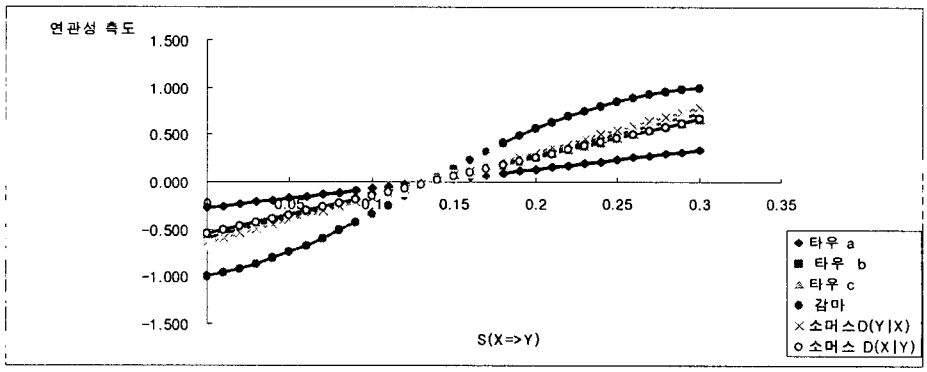
⑥ 소머스 D(X|Y)와 연관규칙의 평가기준들과의 관계

<그림 3> 연관성 측도에 따른 연관규칙의 평가기준값의 변화

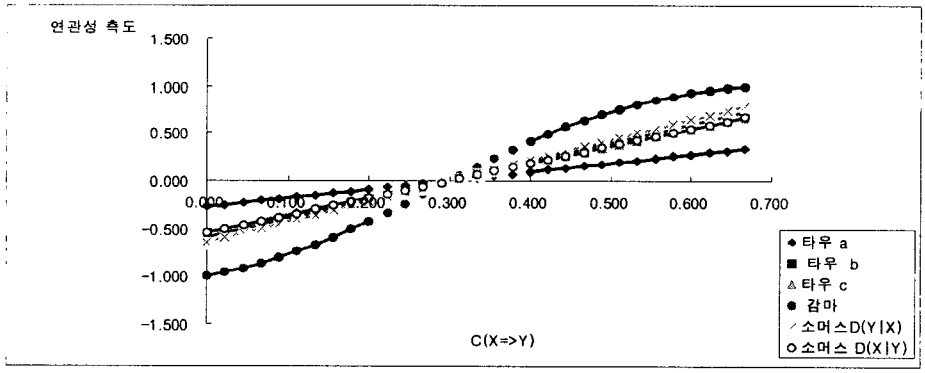
위의 <그림 3>에서 보듯이 감마와 연관규칙의 평가기준들과의 관계를 제외하고는 타우 a, 타우 b, 타우 c, 소머스 D(Y|X), 소머스 D(X|Y)와 연관규칙의 평가기준들과의 관계는 선형관계를 가짐을 알 수 있다. 즉 연관규칙의 평가기준인 지지도, 신뢰도, 향상도가 증가할수록 연관성 측도들의 값이 감마를 제외하고는 모두 선형적으로 증가함을 알 수 있다. 그러나 감마와 연관규칙의 평가기준들과의 관계에서도 선형 관계를 가지는 것은 아니지만 지지도, 신뢰도, 향상도가 증가할수록 감마

값도 증가함을 알 수 있다.

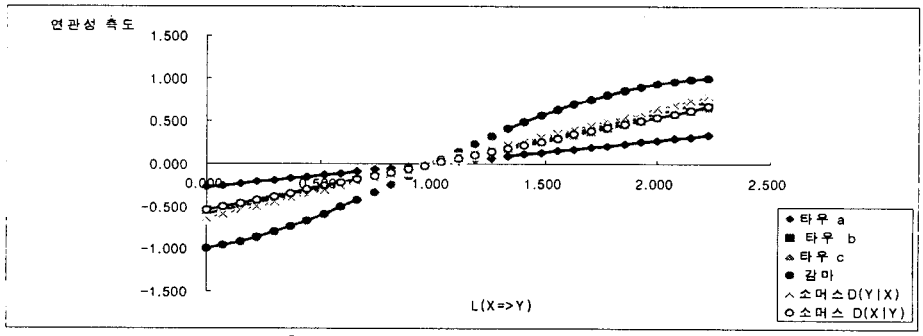
마지막으로 <표 4>에 대해 연관성 측도와 연관규칙의 평가기준과의 관계를 그림으로 표현하면 <그림4>과 같다. 여기서 연한 선으로 되어 있는 부분이 유의한 경우이다.



① 지지도와 연관성 측도들과의 관계



② 신뢰도와 연관성 측도들과의 관계



③ 향상도와 연관성 측도들과의 관계

<그림 4> 연관 규칙의 평가기준값에 따른 연관성 측도값의 변화

이 그림에서 보는 바와 같이 타우 a와 감마를 제외하고는 타우 b, 타우 c, 소머스 D(Y|X), 소머스 D(X|Y)는 거의 비슷한 값을 가짐을 알 수 있다.

6. 결론 및 향후 과제

연관성 규칙은 둘 또는 그 이상의 품목들 사이의 관련성을 발견하고 분석하는 방법이다. 그러나 기존의 연구에서는 관련성의 여부는 파악할 수 있었으나 어느 정도의 관련성이 있는지의 정도는 파악할 수 없었으며 $X \Rightarrow Y$ 연관성과 $X \Rightarrow Z$ 연관성을 비교 분석할 수 없었다. 그러나 본 연구에서는 기존의 연관성 규칙에서 사용하는 세 가지 평가기준과 순위형 자료에서의 연관성 측도를 관련시킴으로써 연관성 규칙에 대한 관련성 정도를 객관적으로 제시해주며 $X \Rightarrow Y$ 연관성과 $X \Rightarrow Z$ 연관성을 비교 분석이 가능하도록 하였다. 향후 연구 과제로는 N×N 분할표에서의 연관성 측도와 연관규칙의 평가기준과의 관계에 대한 연구가 필요하다.

참고문헌

[1] Agrawal, R., Imielinski, R., Swami, A. (1993). Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD Conference on Management of Data*, Washington, D.C.

[2] Agrawal, R. Imielinski, T. Swami, A. Mining Associations Between Sets of Items in Massive Databases, *Proceedings of the ACM SIGMOD,* Washington, DC, May 1993, pp. 207-216.

[3] Agrawal, R., John, C.S. (1996). Parallel Mining of Association Rules, *IEEE Transactions on Knowledge and Data Engineering,* Vol.8, No.6.

[4] A grawal, R., Srikant, R. (1994). Fast algorithms for mining association rules, *Proceedings of the 20th VLDB Conference,* Santiago, Chile.

[5] Bing, L., Wynne, H., Yiming, M. (1999). Mining Association Rules with Multiple minimum Supports, *Proceedings of ACM KDD-99.*

[6] Cheung, D.W., Han, J., Ng, V., Wong, C.Y. (1996). Maintenance of discovered association rules in large databases: An incremental updating technique, *Int'l Conference on Data Engineering*, New Orleans, Louisiana.

[7] Cheung, D.W., Han, J., Ng, V., Fu, A.W., Fu, Y. (1996). A Fast distribution algorithm for mining association rules, *Int's Conference on Parallel and Distributes Information System*, Miami Beach, Florida.

[8] Markus, Hegland. Algorithms for Association Rules, *Australian National University, Canberra ACT 0200, Australia.*

[9] Park, Hee Chang. Song, G.M (2002). Statistical Decision making of Association Threshold in Association Rule Data Mining, *Journal of Korean Data & Information Science Society* 2002, Vol. 13, No.2 pp. 115-128.

[10] Park, J.S., Chen, M.S., and Philip, S.Y. (1995). An effective hash-based algorithms for mining association rules, *Proceedings of ACM SIGMOD Conference on Management of Data.*

[11] Saygin, Y., Vassilios, S.V., Clifton, C. (2002). Using Unknowns to Prevent Discovery of Association Rules, *2002 Conference on Research Issues in Data Engineering.*

[12] Sergey, B., Rajeev, M., Jeffrey, D.U., Shalom, T. (1997). Dynamic itemset counting and implication rules for market data, *Proceedings of ACM SIGMOD Conference on Management of Data.*

[13] Silverstein, C., Brin, S., Motwani, R. (1998). Beyond Market Baskets: Generalizing Association Rules to Dependence Rules, *Data Mining and Knowledge Discovery,* No.2, P 39-68.

[14] Toivonen, H. (1996). Sampling Large Database for Association Rules, *Proceedings of the 22nd VLDB Conference,* Mumbai(Bombay), India.