퍼지 일반화 계층을 이용한 일반화된 퍼지 정량 연관규칙 마이닝

Mining Generalized Fuzzy Quantitative Association Rules with Fuzzy Generalization Hierarchies

한상훈[0] · 손봉기 · 이건명
Sang-Hoon Han[0], Bong-Ki Son, and Keon-Myung Lee
충북대학교 컴퓨터과학과, 첨단정보기술 연구센터
likelamb@aicore.chungbuk.ac.kr

Abstract

연관규칙 마이닝은 트랜잭션 데이터를 이루고 있는 항목간의 잠재적인 의존관계를 발견하는 데이터 마이닝의 한 분야이다. 정량 연관규칙이란 부류적 속성과 정량적 속성을 모두 포함한 연관규칙이다. 정량 연관규칙 마이닝을 위한 퍼지 기술의 응용, 정량 연관규칙 마이닝을 위한 일반화된 연관규칙 마이닝, 사용자의 관심도를 반영한 중요도 가중치가 있는 연관규칙 마이닝 등에 대한 연구가 이루어져왔다. 이 논문에서는 중요도 가중치가 있는 일반화된 퍼지 정량 연관규칙 마이닝의 새로운 방법을 제안한다. 이 방법은 부류적 속성의 퍼지 개념 계층과 정량적 속성의 퍼지 언어항 일반화 계층을 일반화된 연관규칙을 추출하기 위해 이용한다. 이것은 속성들의 수준별 일반화 계층과 속성의 중요도 가중치를 이용함으로써 사용자가 보다 융통성 있는 연관규칙을 마이닝할 수 있게 해준다.

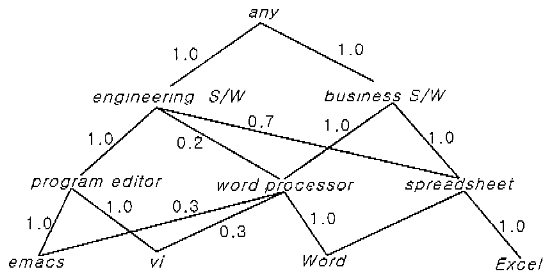
Keywords : 연관규칙(association rule), 퍼지 연관규칙(fuzzy association rule), 일반화된 연관규칙(generalized association rule), 정량 연관규칙(quantitative association rule), 중요도 가중치(importance weight)

1. 서 론

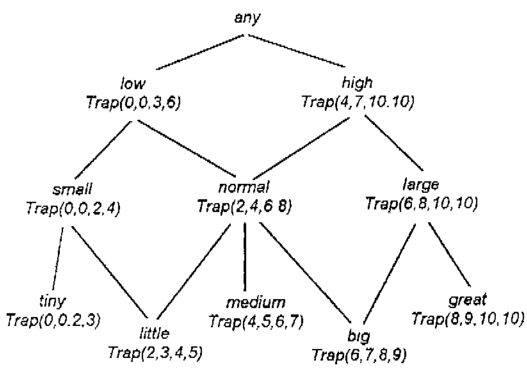
데이터 마이닝(data mining)은 데이터베이스로부터 감추어진 잠재적 유용한 정보를 발견하는 방법을 연구하는 분야이다. 연관규칙(association rule)이란 데이터 마이닝에서 중요한 연구분야의 하나이다. $X \rightarrow Y$로 표현되는 연관규칙은 "항목집합(itemset) X를 포함하는 트랜잭션(transaction)에 항목집합 Y도 포함된다."와 같이 트랜잭션 데이터베이스의 패턴을 의미한다. 예를 들어 "빵을 포함하는 1%의 트랜잭션 중에 오렌지를 포함하는 트랜잭션이 20%이다"라는 연관규칙이 있을 때, 20%는 신뢰도라 하고 1%는 지지도라 한다.

개념 계층은 개념의 영역 사이에 일반적인 관계를 정의하는데 사용된다. 개념 계층은 개념들 사이의 명백한 일반화 관계를 나타낸다. 개념 계층은 비순환 유향 그래프 (N, A)로 표현되는데, N은 개념 노드의 집합이고 A는 일반화 관계를 나타내는 간선 (n_i, n_j)의 집합이다. 간선 (n_i, n_j)는 n_i가 n_j의 일반화 개념이라는 것을 의미한다. 개념 계층에서 일반화 관계를 나타내는 모든 간선들은 명백하다.

퍼지 개념 계층은 퍼지 간선 $(n_i, n_j, \gamma_{n_i n_j})$로 표현되는데, $\gamma_{n_i n_j}$는 n_i의 n_j로의 일반화 정도이다. 퍼지 간선 $(n_i, n_j, \gamma_{n_i n_j})$의 n_j는 $\gamma_{n_i n_j}$정도로 n_i를 부분적으로 일반화한 개념이라는 것을 의미한다. [그림 1]은 부류적 속성 *software*에 대한 퍼지 개념 계층를 보인 것인데, 간선에 기입된 숫자는 해당 개념들 사이의 일반화 정도를 나타낸다. 예를 들어, *emacs*는 1정도로 *program editor*에 속하고, 0.3정도로 *word processor*에 속한다.



[그림 1] *software*의 퍼지 개념 계층



[그림 2] 정량적 속성에 대한 퍼지 언어항의 일반화 계층

정량적 속성을 일반화하기 위해서는 구간값(intervals) 대신에 소속함수(membership degree)에 의해 정의되는

본 연구는 첨단정보기술연구센터(AITrc)를 통해 한국과학재단 지원으로 수행된 것임.

퍼지 언어항을 사용할 수 있다. [그림 2]는 수치 속성값에 대한 퍼지 언어항을 보인 것인데, Trap(0,0,2,4)은 퍼지 언어항 *small*을 정의하는데 사용되는 사다리꼴 퍼지 숫자(trapezoidal fuzzy number)이다.

이 논문은 항목에 대한 사용자의 중요도 가중치(importance weight)를 고려함과 동시에 부류적 속성(categorical attribute)과 정량적 속성(quantitative attribute)을 구분하여 부류적 속성에는 퍼지 개념 계층(fuzzy concept hierarchics)을, 정량적 속성에는 퍼지 언어항의 일반화 계층(generalization hierarchies of fuzzy linguistic terms)을 이용하여 연관규칙을 마이닝하는 방법을 제안한다.

2. 일반화된 퍼지 정량 연관규칙 마이닝을 위한 척도

이 절에서는 제안하는 일반화된 퍼지 정량 연관규칙 마이닝에서 사용하는 척도에 대해 소개한다.

2.1 표기법

설명의 편의성을 위하여 다음과 같은 표기법들을 사용한다.

- $I=\{i_1, i_2, \ldots, i_n\}$: 부류적 속성이나 정량적 속성으로 이루어진 항목 i_k 의 집합.
- T : 트랜잭션($t=\{t.i_1, t.i_2, \ldots, t.i_n\} \in T$)들을 포함하는 트랜잭션의 집합. 여기서, $t.i_k$는 트랜잭션 t에 포함하는 항목 i_k의 값.
- $X=\{x_1, x_2, \ldots, x_n\} \subset I$: 항목집합.
- $A=\{f_{x_1}, f_{x_2}, \ldots, f_{x_n}\}$: 각 항목 x_i에 대응하는 값 f_{x_i}의 집합. f_{x_i}은 항목 값이나 정량적 속성의 퍼지 언어항과 부류적 속성의 일반화 정도.
- $F_{x_i}(\nu) = \bigcup_{k=1}^{d_i} F_{x_i}^{k}(\nu)$: 항목 x_i가 갖는 퍼지 일반화 계층에서의 일반화된 값과 값 ν의 집합. 여기서 d_i는 항목 x_i가 퍼지 일반화 계층에서의 위치한 높이(height)를 가리키며 $F_{x_i}^{k}$는 일반화 수준 k에 값 ν의 일반화된 값들의 집합을 가리킨다($F_{x_i}^{1}$는 트랜잭션에 나타나는 항목들로 이루어진 가장 낮은 수준에서의 항목 값들의 집합이다.).
- $f_{x_i}(t.x_i)$: $t.x_i$로의 f_{x_i}의 일반화 정도. x_i가 부류적 속성이라면 $f_{x_i}(t.x_i)$은 개념 $t.x_i$로의 개념 f_{x_i}의 일반화 정도로 정의되며, x_i가 정량적 속성이라면 $f_{x_i}(t.x_i)$은 퍼지 언어항 f_{x_i}에 대한 정량값 $t.x_i$의 만족 정도로 정의된다.

x_i가 부류적 속성일 때, 일반화 정도는 다음과 같이 계산되어진다.

$$f_{x_i}(t.x_i) = \bigoplus_{\forall p: t.x_i \rightarrow f_{x_i}} \; \bigotimes_{\forall (a,b) \in p} \mu_{ab} \tag{1}$$

단, p는 퍼지 개념 계층에서 x_i로부터 f_{x_i}로 가는 경로(path), (a,b)는 경로상의 하나의 간선(edge), μ_{ab}는 개념 a로의 개념 b의 일반화 정도, $\otimes$는 최소값이나 곱셈 등과 같은 T-norm 연산자(operator), $\oplus$는 합집합 등과 같은 T-conorm 연산자이다.

x_i가 부류적 속성일 때, 일반화 정도는 다음과 같이 계산되어진다.

$$f_{x_i}(t.x_i) = \mu_{f_x}(t.x_i) \tag{2}$$

여기서 $\mu_{f_x}(\ )$는 퍼지 언어항 f_{x_i}의 소속 함수(membership function)로 정의한다.

- σ_{tA}은 항목 값들의 집합 A에 대한 트랜잭션 t 의 지지도

$$\sigma_{tA} = \bigotimes_{\forall f_x \in A} w(x_i) f_{x_i}(t.x_i) \tag{3}$$

- $\sum count$ 연산자는 T에 있는 트랜잭션들의 연관 관계를 나타내는 정도의 합을 계산하는데 사용된다.

$$\sum_{t \in T} count(\sigma_{tA}) = \sum_{t \in T} \; \bigotimes_{\forall f_x \in A} w(x_i) f_{x_i}(t.x_i) \tag{4}$$

- $w(x_i)$: 항목 x_i의 중요도 가중치.
- σ : 최소 지지도.
- χ : 최소 신뢰도.
- C_r : r 개의 항목들로 이루어진 후보 항목들의 집합.
- L_r : r 개의 항목들로 이루어진 빈발 항목집합(large itemsets).

2.2 척도

연관규칙 마이닝 알고리즘에서는 마이닝된 결과의 흥미도(interestingness)를 결정하기 위해 여러 가지 척도(measure)를 사용한다. 지지도(support degree)와 신뢰도(confidence degree)는 이러한 전형적인 척도이다. 이 척도의 기본적인 역할은 트랜잭션의 지지도가 후보 항목집합(candidate itemset)에 포함될 수 있을지를 평가하는 것이다. 이 논문에서는 기본적인 지지도와 신뢰도에 각 항목에 대한 중요도 가중치와 일반화된 개념이나 값들의 일반화 정도를 함께 고려한 방법을 사용한다.

- 항목집합 A에 대한 가중치 지지도

$$w-support(A) = \frac{\sum_{t \in T} count(\sigma_{tA})}{|T|} \tag{5}$$

- 연관규칙 A→B에 대한 가중치 신뢰도

$$w-confidence(A \rightarrow B) = \frac{w-support(A \bigcup B)}{w-support(A)} \tag{6}$$

- R-interest 척도는 중복되는 규칙을 제거하기 위한 방법으로 사용된다. R-interest 척도에서 말하는 흥미로운 규칙이란 그 규칙의 기대되는 지지도나 신뢰도가 R배 이상인 것을 가리킨다[10].

$A=\{f_{a_1}, f_{a_2}, \ldots, f_{a_n}\}$와 $B=\{f_{b_1}, f_{b_2}, \ldots, f_{b_n}\}$로 이루어진 어떤 규칙 $A \rightarrow B$가 있고 $\hat{A}=\{\hat{f}_{a_1}, \hat{f}_{a_2}, \ldots, \hat{f}_{a_n}\}$와 $\hat{B}=\{\hat{f}_{b_1}, \hat{f}_{b_2}, \ldots, \hat{f}_{b_n}\}$는 A와 B 각각의 상위 개념(ancestor)이라고 가정하자. 어떤 규칙의 집합에서, $A \rightarrow B$의 상위 개념은 $A' \rightarrow B'$이고 $A' \rightarrow B'$의 상위 개념은 $\hat{A} \rightarrow \hat{B}$이라고 할 때, $A' \rightarrow B'$ 규칙이 존재하지 않는다면, 규칙 $\hat{A} \rightarrow \hat{B}$ 가 규칙 $A \rightarrow B$의 가까운 상위 개념이라고 말할 수 있다. $w-support_{E(\hat{A} \cup \hat{B})}(A \bigcup B)$은 주어진 $\hat{A} \bigcup \hat{B}$

에 대한 $A \bigcup B$의 지지도의 기대값이라고 정의하고 $w-confidence_{E(\hat{A} \rightarrow \hat{B})}(A \rightarrow B)$은 주어진 규칙 $\hat{A} \rightarrow \hat{B}$에 대한 규칙 $A \rightarrow B$의 신뢰도의 기대값이라 정의한다.

■ $w-support_{E(\hat{A} \cup \hat{B})}(A \bigcup B)$

$$= \prod_i \frac{\sum_{t \in T} f_{a_i}(t.a_i)}{\sum_{t \in T} \hat{f}_{a_i}(t.a_i)} \times \prod_i \frac{\sum_{t \in T} f_{b_i}(t.b_i)}{\sum_{t \in T} \hat{f}_{b_i}(t.b_i)} \times w-support(\hat{A} \bigcup \hat{B}) \quad (7)$$

■ $w-confidence_{E(\hat{A} \rightarrow \hat{B})}(A \rightarrow B)$

$$= \prod_i \frac{\sum_{t \in T} f_{a_i}(t.a_i)}{\sum_{t \in T} \hat{f}_{a_i}(t.a_i)} \times \prod_i \frac{\sum_{t \in T} f_{b_i}(t.b_i)}{\sum_{t \in T} \hat{f}_{b_i}(t.b_i)} \times w-confidence(\hat{A} \bigcup \hat{B}) \quad (8)$$

어떤 규칙이 상위 개념을 가지고 있지 않거나 흥미로운 상위 개념들 중의 가까운 상위 개념에 대하여 R-interesting 관계가 성립될 때 규칙 $A \rightarrow B$는 흥미롭다라고 말할 수 있다.

3. 일반화된 퍼지 정량 연관규칙 마이닝 알고리즘

이 논문에서 제안하는 일반화된 퍼지 정량 연관규칙 마이닝 알고리즘에서 연관규칙 마이닝을 위해 빈발 항목 집합을 생성하는 방법은 Apriori algorithm[1]에 기초를 두고 있다.

Procedure Mining_Algorithm

Input

트랜잭션 데이터 베이스 T

부류적 속성을 위한 퍼지 개념 계층

정량적 속성을 위한 퍼지 언어항의 개념화 계층

각 항목에 대한 중요도 가중치 w

최소 지지도 σ

최소 신뢰도 χ

R-interest 척도를 위한 R 값

Output

지지도와 신뢰도를 만족하는 일반화된 퍼지 정량 연관규칙

Begin

각 트랜잭션 $t=\{t.i_1, t.i_2, \ldots, t.i_n\}$을 $t'=\{V_{i_1}(t), V_{i_2}(t), \ldots, V_{i_n}(t)\}$로 변환. 단, $V_{i_k}(t)=\{(i_k, f_{i_k}, f_{i_k}(t.i_k)) \mid f_{i_k} \in F_{i_k}(t.i_k)\}$

지지도를 만족하는 1-빈발 항목집합 L_1을 트랜잭션 데이터 베이스로부터 생성.

Do

후보 k-빈발 항목집합 C_k을 k-1 빈발 항목집합 L_{k-1}으로부터 생성.

k-항목집합에 포함된 각 항목에 대해 가중치 지지도를 식 (4),(5)에 의해 계산.

최소 지지도 σ 보다 적은 지지도를 갖는 k-항목 집합을 C_k로부터 제거.

제거되지 않은 k-항목집합을 빈발 항목집합 L_k에 포함시킴.

Until 빈발 항목집합이 생성되지 않음

생성된 빈발 항목집합으로 퍼지 정량 연관규칙을 만듬.

식 (6)을 이용하여 각 규칙의 가중치 신뢰도를 계산.

최소 신뢰도 χ 보다 적은 신뢰도를 갖는 규칙을 제거.

식 (7),(8)을 이용하여 각 규칙의 기대되는 지지도와 신뢰도를 계산.

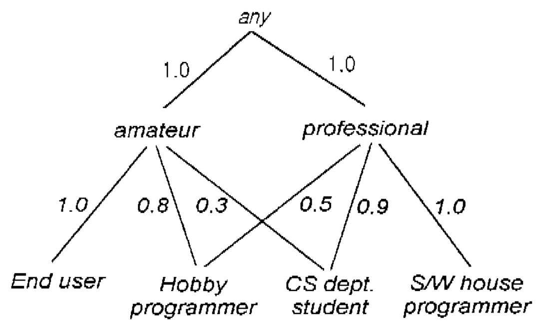
R-interestingness를 만족하지 않는 규칙을 제거.

제거되지 않는 퍼지 정량 연관규칙과 그 규칙의 지지도, 신뢰도를 출력.

End

4. 예 제

이 절에서는 이 논문에서 제안한 일반화된 퍼지 정량 연관규칙 마이닝 방법을 사용한 예제를 소개한다. [표 1]은 사용자에 따른 프로그램의 사용시간을 나타낸 것이다. 프로그램 속성은 [그림 1]의 퍼지 개념 계층을 이용하고, 사용시간 속성은[그림 2]의 퍼지 언어항의 일반화 계층을 이용하며, 사용자 속성은 [그림 3]의 퍼지 개념 계층을 이용하기로 한다. 또한, 프로그램, 사용자, 사용시간 속성의 중요도 가중치는 각각 1.3, 0.8, 1.0 으로 한다.



[그림 3] 프로그래밍 작업의 퍼지 개념 계층

[표 1]에 의해 후보 항목 집합 $A=\{program\ editor, professional, big\}$이 있다고 가정하고, 식 (5)에 의해 A의 가중치 지지도를 계산하면 다음과 같다. 단, 식 (4)의 $\otimes$의 위치에 곱셈 연산자가 있다고 가정한다.

$$count(\sigma_{T_1A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0, 1.0 \cdot 1.0\} = 1.04$$
$$count(\sigma_{T_2A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.9, 1.0 \cdot 0.5\} = 0.468$$
$$count(\sigma_{T_3A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.0, 1.0 \cdot 0.0\} = 0.0$$
$$count(\sigma_{T_4A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.5, 1.0 \cdot 0.0\} = 0.0$$
$$count(\sigma_{T_5A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9, 1.0 \cdot 0.0\} = 0.0$$
$$count(\sigma_{T_6A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0, 1.0 \cdot 1.0\} = 1.04$$

$$count(\sigma_{T_7A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.5, 1.0 \cdot 0.5\} = 0.26$$
$$count(\sigma_{T_8A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9, 1.0 \cdot 0.0\} = 0.0$$
$$w-support(A) = \frac{1.04 + 0.468 + 1.04 + 0.26}{8} = 0.351$$

	프로그램	사용자	사용시간
T_1	emacs	S/W house programmer	8
T_2	vi	CS dept. student	6.5
T_3	Excel	End user	1
T_4	Word	Hobby programmer	2
T_5	Word	CS dept. student	4
T_6	vi	S/W house programmer	7
T_7	emacs	Hobby programmer	6.5
T_8	Excel	CS dept. student	2

[표 1] 프로그램 사용 데이터

퍼지 정량 연관규칙 $\{program\ editor, professional\} \rightarrow \{big\}$의 가중치 신뢰도를 계산하여 보자. 단, 여기서 $B=\{program\ editor, professional\}$이고 $C=\{big\}$라고 하자.

$$count(\sigma_{T_1B}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0\} = 1.04$$
$$count(\sigma_{T_2B}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.9\} = 0.936$$
$$count(\sigma_{T_3B}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.0\} = 0.0$$
$$count(\sigma_{T_4B}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.5\} = 0.0$$
$$count(\sigma_{T_5B}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9\} = 0.0$$
$$count(\sigma_{T_6B}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0\} = 1.04$$
$$count(\sigma_{T_7B}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.5\} = 0.52$$
$$count(\sigma_{T_8B}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9\} = 0.0$$
$$w-support(B) = \frac{1.04 + 0.936 + 1.04 + 0.52}{8} = 0.442$$
$$w-confidence(B \rightarrow C) = \frac{0.351}{0.442} = 0.778$$

연관규칙 $\{emacs, S/W\ house\ porgrammer\} \rightarrow \{big\}$의 가까운 상위개념이 $\{program\ editor, professional\} \rightarrow \{normal\}$라고 가정하고 $w-support_{E(\hat{P} \cup \hat{Q})}(P \cup Q)$를 계산하여 보자. 단, $P=\{emacs, S/W\ house\ porgrammer\}$, $Q=\{big\}$ 이고, $\hat{P}=\{program\ editor, professional\}$, $\hat{Q}=\{normal\}$이다.

$$w-support_{E(\hat{P} \cup \hat{Q})}(P \cup Q) = \frac{1+1}{1+1+1+1}$$
$$\times \frac{1+1}{1+0.9+0.5+0.9+1+0.5+0.9}$$
$$\times \frac{1+0.5+1+0.5}{0.75+1+0.5+0.75} \times 0.442 = 0.077$$

반면에, $w-support(P \cup Q)$은 0.125이다. 이때, 2-interesting(R=2) 이라면, 규칙 $P \rightarrow Q$은 $w-support(P \cup Q) < 2\times\ w-support_{E(\hat{P} \cup \hat{Q})}(P \cup Q)$이므로 흥미롭지 않은 규칙이 되는 것이다.

5. 결론 및 향후과제

이 논문에서 속성의 중요도 가중치를 고려한 일반화된 퍼지 정량 연관규칙 마이닝 방법에 대해 제안한다. 일반화를 위하여 부류적 속성에는 퍼지 개념 계층을 이용하고, 정량적 속성에 대해서는 퍼지 언어항의 일반화 계층을 이용한다. 마이닝된 연관규칙의 흥미도를 평가하기 위해서 속성의 중요도 가중치와 항목 값들의 일반화 정도를 고려한 몇 가지 척도를 소개한다. 또한, 일반화된 퍼지 정량 연관규칙을 생성하는 마이닝 절차에 대해서 제안하였다.

이 논문에서 제안된 마이닝 방법은 일반화된 퍼지 연관규칙을 생성하기에 유용하며, 속성의 중요도 가중치를 조절함으로써 마이닝 대상을 융통성 있게 변경할 수 있다. 향후 과제로는 제안된 마이닝 방법의 메모리 활용과 마이닝 속도를 향상시킴으로써 성능 향상을 위한 연구가 필요하다.

참 고 문 헌

[1] R. Agrawal, T. Imielinski, A. Swami, Mining Association Rules between Sets of Items in Large Databases, *Proc. of the ACM SIGMOD Conf*, Washington D.C., pp.207-216, 1993.

[2] R. Srikant, R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables, *Proc. of the ACM SIGMOD*, Montreal, Canada, 1996.

[3] C.M. Kuok, A. Fu, M.H. Wong, Mining Fuzzy Association Rules in Databases, *Proc. of SIGMOD-Record*, Vol.27, No.a, pp.41-46, 1998.

[4] T.P. Hong, C.-S. Kuo, S.-C. Chi, A Fuzzy Data Mining Algorithm for Quantitative Values, *Proc. of 3rd Int. Conf. on Knowledge-based Intelligent Information Engineering Systems*, Adelaide, Australia, pp.480-483, 1999.

[5] W. Zhang, Mining Fuzzy Quantitative Association Rules, *Proc. of IEEE Int. Conf. on Tools with Artificial Intelligence*, pp.99-102, 1999.

[6] C. H. Cai, W. C. Fu, C. H. Cheng, W. W. Kwong, Mining Association Rules with Weighted Items, *Proc. of IDEAS'98*, pp.68-77, 1999.

[7] J. Shu Yue, E. Tsang, D. Yeung, D. Shi, Mining Fuzzy Association Rules with Weighted Items, *proc. of IEEE*, pp.1906-1911, 2000.

[8] K.-M. Lee, A Database Summarization Using Fuzzy Concept Hierarchies and Fuzzy Linguistic Terms, *Proc. of Int. Conf. SCI/ISAS'99*, Orland, Florida, Vol.3, pp.514-519, 1999.

[9] A. Gyenesei, Mining Weighted Association Rules for Fuzzy Quantitative Items, *TUCS Technical Report No.346*, May 2000.

[10] R. Srikant, R. Agrawal, Mining Generalized Association Rules, *Proc. of the 21st Int'l Conference on Very Large Databases*, Zurich, Swizerland, Sep. 1995.