International Journal of Fuzzy Logic and Intelligent Systems, vol. 2, no. 3, December 2002 pp. 210-214

Generalized Fuzzy Quantitative Association Rules Mining with Fuzzy Generalization Hierarchies

Keon Myung Lee

School of Electrical and Computer Engineering, Chungbuk National University and Advanced Information Technology Research Center(AITrc)

Abstract

Association rule mining is an exploratory learning task to discover some hidden dependency relationships among items in transaction data. Quantitative association rules denote association rules with both categorical and quantitative attributes. There have been several works on quantitative association rule mining such as the application of fuzzy techniques to quantitative association rule mining, the generalized association rule mining for quantitative association rules, and importance weight incorporation into association rule mining for taking into account the users interest. This paper introduces a new method for generalized fuzzy quantitative association rule mining with importance weights. The method uses fuzzy concept hierarchies for categorical attributes and generalization hierarchies of fuzzy linguistic terms for quantitative attributes. It enables the users to flexibly perform the association rule mining by controlling the generalization levels for attributes and the importance weights for attributes.

Key Words : association rule, fuzzy association rule, generalized association rule, quantitative association rule, importance weight

1. Introduction

Data mining, also known as knowledge discovery in databases, is a study to find hidden, potentially useful information from databases. Inducing association rules is one of the important research issues in data mining. An association rule, expressed as $X \rightarrow Y$, indicates a pattern in transaction databases where the transactions containing the items X also contain the items Y. An example of an association rule is: "20% of transactions that contain bread also contain orange; 1% of transactions contain both items". Here 20% is called confidence and 1% is called support.

The problem of mining association rules was first introduced by Agrawal et al.[1], for databases consisting of only categorical attributes. In categorical association rules $X \rightarrow Y$, both antecedent X and consequent Y indicates the appearance of corresponding items.

To handle databases with both categorical and quantitative attributes (i.e., numeric, interval values), a quantitative association rule mining method was proposed by Srikant et al.[2]. The method finds association rules by partitioning the quantitative attribute domain, combining adjacent partitions, and then transforming the problem into binary one (i.e., categorical association rule mining problem). In quantitative association rule mining, records of a database table are considered as transactions, where a pair of an attribute and its corresponding value plays the role of an item in categorical association rule mining. Therefore, quantitative association rules come to have the pairs of an attribute and its value like <Age: 20..30> and <Sex: Female> $\rightarrow$ <Number of friends: 2..5>.

This work was supported by the Korea Science and Engineering Foundation(KOSEF) through AITrc.

As an approach to take into account the relative importance of attributes, the concept of weighted items was incorporated in the association rule mining[6]. The item weighting enables users to control the mining results by assigning larger weights to attribute in which they are more interested rather than others. There have been some studies to integrate fuzzy techniques into association rule mining [3,4,5,7,9].

Instead of using crisp intervals in quantitative association rule mining, some methods have been proposed to make use of fuzzy intervals defined by membership functions [3,4,5]. They label the fuzzy intervals with fuzzy linguistic terms such as *young*, *high*, and so on. It makes it possible to produce more descriptive association rules. Some methods have been proposed to exploit fuzzy techniques for quantitative attribute domain partitioning and attribute weighting for reflecting importance of attribute[7,9].

Sometimes it is more meaningful to have association rules with generalized items rather than base items appearing in transactions. There have been some works to produce generalized association rules[10].

In this study, we are interested in mining generalized fuzzy quantitative association rule with weighted items. We propose a mining method to use fuzzy concept hierarchies and generalization hierarchies of fuzzy linguistic terms for categorical attribute generalization and quantitative attribute domain partitioning, to allow users to weigh the importance of attributes.

This paper is organized as follows: Section 2 describes fuzzy concept hierarchies for categorical attributes and generalization hierarchies of fuzzy linguistic terms for quantitative attributes, which are used in generalizing attribute values. Section 3 introduces some notations and measures that are used in the proposed association rule mining method. Section 4 presents the proposed generalized fuzzy quantitative association rule mining method. Section 5 gives an example to apply the proposed method, and then Section 6 draws conclusions.

2. Fuzzy Concept Hierarchies and Generalization Hierarchies of Fuzzy Linguistic Terms

Concept hierarchies are commonly used to define generalization relationships among domain concepts. Conventional concept hierarchies represent only crisp generalization relationships between concepts. A concept hierarchy is represented by an acyclic digraph (N, A), where N and A are a set of concept nodes and a set of edges (n_i, n_j) to represent the generalization relationships, respectively. An edge (n_i, n_j) indicates that n_j is a generalized concept of n_i. In conventional concept hierarchies, all edges describing generalization relationship are crisp.

On the other hand, fuzzy concept hierarchies contain fuzzy edges (n_i, n_j, $\gamma_{n_in_j}$) where $\gamma_{n_in_j}$ is the generalization degree of n_j to n_i. A fuzzy edge ($n_i, n_j, \gamma_{n_in_j}$) indicates that n_j is a partially generalized concept of n_i with the degree $\gamma_{n_in_j}$. In fuzzy concept hierarchies, a concept may have partial generalization relationships with several generalized concepts. Since fuzzy concept hierarchies are flexible to represent generalization relationship between concepts, fuzzy concept hierarchies are used to generalize categorical attributes in our new fuzzy association rule mining method. Figure 1 shows a fuzzy concept hierarchy for a categorical attribute *software*, where numbers labeled on edges indicates the generalization degree between corresponding concepts. For example, *emacs* belongs to *program editor* with the degree 1 and it belongs to *word processor* with the degree 0.3.

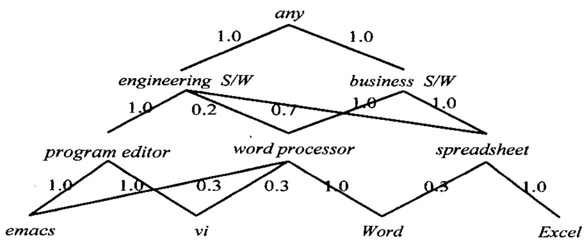


Figure 1. Fuzzy concept hierarchy for *programs*

For quantitative attributes, we can make fuzzy partitions with various granularities. It means that we can partition an attribute domain into various numbers of fuzzy intervals, for instance, with 10 fuzzy intervals, 5 fuzzy intervals, or 2 fuzzy intervals. As the smaller number of fuzzy intervals there are, the more general the fuzzy intervals are. That is to say, a fuzzy linguistic term set defined by 5 fuzzy intervals is more general than that by 10 fuzzy intervals. A generalization hierarchy of fuzzy linguistic terms on a quantitative attribute is a hierarchy structure in which upper level nodes represent more general fuzzy linguistic terms (i.e., more larger fuzzy intervals) than lower level nodes do. The fuzzy linguistic terms are organized into a hierarchy according to their coverage over the domain. For two fuzzy linguistic terms A and B, we say that A is more general concept than B when the membership function $\mu_A(x)$ of A contains that $\mu_B(x)$ of B (that is, $\forall x, \mu_A(x) \geq \mu_B(x)$).

The membership functions for fuzzy intervals making fuzzy partitions can be determined in various ways. One is by expert opinion or by people's perception. Another way is by statistical methods. Fuzzy clustering based on self-organized learning also is used to generate membership functions. Here we assume that the membership functions for fuzzy partitions with various granularities have been already determined by some way and the trapezoidal fuzzy number *Trap(a, b, c, d)* is used to represent the fuzzy linguistic terms (i.e., fuzzy intervals) for fuzzy partitions. Figure 2 shows a generalization hierarchy of fuzzy linguistic terms making fuzzy partitions on a quantitative domain.

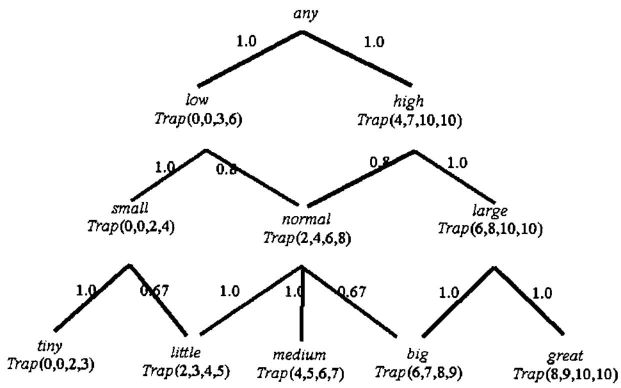


Figure 2. Generalization hierarchy of fuzzy linguistic terms for a quantitative attribute

3. Measures for Generalized Fuzzy Quantitative Association Rule Mining

This section describes some notations and some interestingness measures to be used in the description of the proposed method.

3.1 Notations

For the convenience of description, the following notations are used.

- $I = \{i_1, i_2, \ldots, i_n\}$ is a set of items where each i_k denotes a categorical or quantitative attribute.
- T is a transaction set which consists of transactions $t = \{t.i_1, t.i_2, \ldots, t.i_n\} \in T$. Here $t.i_k$ denotes the value of attribute i_k in transaction t.
- $X = \{x_1, x_2, \ldots, x_m\} \subset I$ is an itemset that is a subset of I.
- $A = \{f_{x_1}, f_{x_2}, \ldots, f_{x_n}\}$ is an item value set where f_{x_i}

denotes a value corresponding to the item (i.e., attribute) x_i. That is, it is an item value or its generalized item value for categorical attributes, or a fuzzy linguistic term for quantitative attributes.

- $F_{x_i}(v) = \bigcup_{k=1}^{d_i} F^k_{x_i}(v)$ is the set of value v and its generalized values in fuzzy generalization hierarchy (i.e., fuzzy concept hierarchy or generalization hierarchy of fuzzy linguistic terms) for attribute x_i. Here d_i is the height of fuzzy generalization hierarchy for attribute x_i. $F^k_{x_i}$ is the set of generalized values of value v at the generalization level k and $F^1_{x_i}$ is the set of item values at the base level (i.e., appearing in the transactions).
- $f_{x_i}(t.x_i)$ denotes the generalization degree of f_{x_i} to $t.x_i$. When x_i is a categorical attribute, $f_{x_i}(t.x_i)$ denotes the generalization degree of concept f_{x_i} for concept $t.x_i$. When x_i is a quantitative attribute, it denotes the satisfaction degree (i.e., membership degree) of quantitative value $t.x_i$ to the fuzzy linguistic term f_{x_i}.

When x_i is a categorical attribute, the generalization degree is calculated as follows:

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

Here, p denotes a path from x_i to f_{x_i} in the fuzzy concept hierarchy, (a,b) is an edge on the path, μ_{ab} is the generalization degree of concept b to concept a, $\otimes$ denotes a T-norm operator like minimum, multiplication, etc., and $\oplus$ denotes a T-conorm operator like union, etc.

When x_i is a categorical attribute, the generalization degree is calculated as follows:

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

Here, $\mu_{f_{x_i}}()$ denotes the membership function of fuzzy linguistic term f_{x_i}.

- $\sigma_{t.A}$ denotes the support degree of transaction t to the attribute value set A.

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

- $\Sigma count$ operator is used to sum up all degrees that are associated with the transactions in T.

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

- $w(x_i)$ denotes the importance weight for the attribute x_i.
- σ denotes the predefined minimum support value.
- χ denotes the predefined minimum confidence value.
- C_r denotes the set of candidate itemsets with r attributes (items).
- L_r denotes the set of large itemsets with r attributes (items).

3.2 Measures

The association rule mining algorithms use some measures to determine the interestingness of the mined results. The support and the confidence measures are the typical ones. A basic operation for these measures is to evaluate the support degree of a transaction (or record) to a candidate item set. In order to evaluate the support degree, we use a method to combine the importance weights of attributes and the generation degrees of generalized concepts or values to their corresponding attribute values appearing in transactions and to aggregate the combined degrees.

- Weighted support *w-support(A)* for an item value set A

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

Here $\sum_{t \in T} count(\sigma_{t.A})$ is computed by Eq.(4).

- Weighted confidence *w-confidence(A→B)* for an association rule $A \to B$

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

- R-interest measure is a way used to prune out redundant rules. The rules of interest, according to R-interest, are those rules whose degrees of support are more than R times the expected degrees of support or whose degrees of confidence are more than R times the expected degrees of confidence[10].

Suppose that there is a rule $A \to B$ where $A = \{f_{a_1}, f_{a_2}, \ldots, f_{a_m}\}$ and $B = \{f_{b_1}, f_{b_2}, \ldots, f_{b_n}\}$. Let $\widehat{A} = \{\widehat{f_{a_1}}, \widehat{f_{a_2}}, \ldots, \widehat{f_{a_m}}\}$ and $\widehat{B} = \{\widehat{f_{b_1}}, \widehat{f_{b_2}}, \ldots, \widehat{f_{b_n}}\}$ be ancestors (i.e., generalized concepts) for A and B respectively. Given a set of rules, $\widehat{A} \to \widehat{B}$ is called a close ancestor of $A \to B$ if there is no rule $A' \to B'$ such that $A' \to B'$ is an ancestor of $A \to B$ and $\widehat{A} \to \widehat{B}$ is an ancestor of $A' \to B'$. Let $w-support_{E(\widehat{A} \cup \widehat{B})}(A \cup B)$ denote the expected value of the support degree of $A \cup B$ given $\widehat{A} \cup \widehat{B}$. Let $w-confidence_{E(\widehat{A} \to \widehat{B})}(A \to B)$ denote the expected value of the support degree of $A \to B$ given $\widehat{A} \to \widehat{B}$.

- $$w-support_{E(\widehat{A} \cup \widehat{B})}(A \cup B) = \prod_i \frac{\sum_{t \in T} f_{a_i}(t.a_i)}{\sum_{t \in T} \widehat{f_{a_i}}(t.a_i)} \times \prod_i \frac{\sum_{t \in T} f_{b_i}(t.b_i)}{\sum_{t \in T} \widehat{f_{b_i}}(t.b_i)} \times w-support(\widehat{A} \cup \widehat{B}) \tag{7}$$

- $$w-confidence_{E(\widehat{A} \to \widehat{B})}(A \to B) = \prod_i \frac{\sum_{t \in T} f_{a_i}(t.a_i)}{\sum_{t \in T} \widehat{f_{a_i}}(t.a_i)}$$

$$\times \prod_{i} \frac{\sum_{t \in T} f_{b_i}(t.b_i)}{\sum_{t \in T} \widehat{t_{b_i}}(t.b_i)} \times w-confidence(\widehat{A} \to \widehat{B}) \tag{8}$$

A rule $A \to B$ is called interesting if it has no ancestor or it is R-interesting with respect to its close ancestors among its interesting ancestors.

4. Generalized Fuzzy Quantitative Association Rule Mining Algorithm

In this study, we assume that categorical attributes have their own fuzzy concept hierarchy and quantitative attributes have their own generalization hierarchy of fuzzy linguistic terms. Meanwhile, to guide the mining direction for target association rules, the users are allowed to assign the importance weights to attributes.

The following shows the procedure of the proposed mining method for generalized fuzzy quantitative association rules: This procedure is based on the Apriori algorithm[1] to generate large itemsets for categorical association rule mining.

Procedure *Mining_Algorithm*
Input
  a database T to be mined
  fuzzy concept hierarchies for categorical attributes
  generalization hierarchies of fuzzy linguistic terms for quantitative attributes
  importance weights w for attributes
  minimum support degree σ
  minimum confidence degree χ
  R value for R-interest measure
Output
  generalized fuzzy quantitative association rules with their support and confidence degrees
Begin
  Transform each transaction $t = \{t.i_1, t.i_2, \ldots, t.i_n\} \in T$ into an augmented transaction
    $t' = \{V_{i_1}(t), V_{i_2}(t), \ldots, V_{i_n}(t)\}$ where
    $V_{i_k}(t) = \{(i_k, f_{i_k}, f_{i_k}(t.i_k)) | f_{i_k} \in F_{i_k}(t.i_k)\}$.
  Find the 1-large itemset L_1 by checking the generated database with respect to support degrees.
Do
  Produce candidate k-itemsets C_k from $k-1$ large itemsets L_{k-1}. On constructing candidate k-itemsets, keep two values defined on the same attribute from being included in the same itemset.
  Compute the weighted support degree for each itemset in k-itemsets by Eqs.(4)-(5).
  Delete the k-itemsets of which support is less than specified support degree σ from C_k.
  Insert the remaining k-itemsets into the large itemsets L_k.
Until no more large itemset is produced.
  Make fuzzy quantitative association rules with the found large itemsets.
  Compute the weighted confidence degree for each rule by Eq.(6).
  Remove the rules with the confidence degree less than the prespecified threshold χ.
  Compute the expected support and confidence degrees for each rule by Eqs.(7)-(8).
  Remove the rules which do not satisfy R-interestingness.
  Output the remaining fuzzy quantitative association rules with their support and confidence degrees.
End.

5. An Example

This section shows an example to apply the proposed generalized fuzzy association rule mining method. Table 1 shows an example data set for program usage hours of users. Suppose that we use the fuzzy concept hierarchy of Figure 1 for the attribute *program*, the generalization hierarchy of fuzzy linguistic terms of Figure 2 for the attribute *usage hour*, and the fuzzy concept hierarchy of Figure 3 for the attribute *user*. Suppose that the importance weights for *program*, *user*, and *usage hour* are 1.2, 0.8, 1.0, respectively.

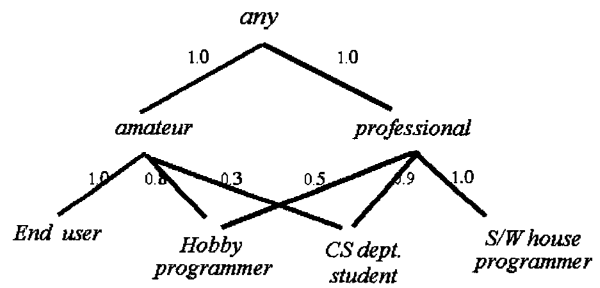


Figure 3. A concept hierarchies of *programming occupations*

Consider a candidate itemset A={*program editor*, *professional*, *big*} for Table 1. The weighted support for A is computed by Eq.(5) as follows: Here, product operator is used in the place of $\otimes$ in Eq.(4).

$count(\sigma_{T_1 A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0, 1.0 \cdot 1.0\} = 1.04$

$count(\sigma_{T_2 A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.9, 1.0 \cdot 0.5\} = 0.468$

$count(\sigma_{T_3 A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.0, 1.0 \cdot 0.0\} = 0.0$

$count(\sigma_{T_4 A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.5, 1.0 \cdot 0.0\} = 0.0$

$count(\sigma_{T_5 A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9, 1.0 \cdot 0.0\} = 0.0$

Table 1. Program usage data

	program	*user*	*usage hour*
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

$$count(\sigma_{T_6A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 1.0, 1.0 \cdot 1.0\} = 1.04$$
$$count(\sigma_{T_7A}) = \prod\{1.3 \cdot 1.0, 0.8 \cdot 0.5, 1.0 \cdot 0.5\} = 0.26$$
$$count(\sigma_{T_8A}) = \prod\{1.3 \cdot 0.0, 0.8 \cdot 0.9, 1.0 \cdot 0.0\} = 0.0$$
$$w-support(A) = \frac{1.04+0.468+1.04+0.26}{8} = 0.351$$

Let us compute the weighted confidence for the fuzzy quantitative association rule {*program editor, professional*} → {*big*}. Let B = {*program editor, professional*} and C = {*big*}.

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

To compute the R-interest measure for support degree, it is needed to compute the expected value of the support degree of an association rule given its close ancestor. Suppose that a close ancestor of an association rule {*emacs, S/W house programmer*} → {*big*} is the above association rule {*program editor, professional*} → {*big*}. Let P = {*emacs, S/W house programmer*}, Q = {*big*}, $\hat{P}$ = {*program editor, professional*}, and $\hat{Q}$ = {*big*}. Let's compute $w-support_{E(\hat{A} \cup \hat{B})}(A \cup B)$.

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

On the other hand, $w-support(P \cup Q)$ is 0.125. When we are interested in 2-interesting (i.e., $R=2$) rule, $P \to Q$ is ignored since $2 \cdot w-support_{E(\hat{A} \cup \hat{B})}(A \cup B) > w-support(P \cup Q)$.

6. Conclusions

In this study, we have tackled the problem to induce generalized fuzzy quantitative association rules with importance weights for attributes. For the generalization, we use fuzzy concept hierarchies for categorical attributes and generalization hierarchies of fuzzy linguistic terms for quantitative attributes. To evaluate the interestingness of mined association rules, we introduced some measures that consider generalization degrees of item values and importance weights of attributes simultaneously. We proposed a mining procedure to produce generalized fuzzy quantitative association rules.

The proposed mining method is a useful one to make it possible to produce generalized fuzzy association rules and to guide the mining direction by adjusting importance weights for attributes. Further works are still required to improve the performance speedup and memory utilization of the proposed mining method.

References

[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., USA, pp.207-216, 1993.

[2] R. Srikant, R. Agrawal, Mining Quantitative Association Rules in Large Relational Tables, *Proc. of 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.1, pp.41-46, 1998.

[4] T.P. Hong, C.-S. Kuo, S.-C. Chi, A Fuzzy Data Mining Algorithm for Quantitative Values, *Prof. 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.

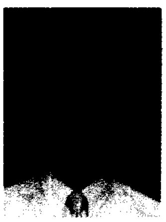
[6] C. H. Cai, W. C. FU, C. H. Cheng, W. W. Kwong, Mining Association Rules with Weighted Items, *Proc. of IDEAS98*, pp.68-77, 1998.

[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, Switzerland, Sep. 1995.



Keon Myung Lee
1990 BS Dept. of Comp. Sci., KAIST
1992 MS Dept. of Comp. Sci., KAIST
1995 Ph.d. Dept. of Comp. Sci., KAIST
1996~present, Assistant professor, Chungbuk national university
Research interests: soft computing, machine learning, bioinformatics, information security