Mining Weighted Association Rules for Fuzzy Quantitative Items

Attila Gyenesei

Turku Centre for Computer Science

TUCS Technical Reports

No 346, May 2000

Mining Weighted Association Rules for Fuzzy Quantitative Items

Attila Gyenesei

University of Turku, Department of Computer Science,
Lemminkisenkatu 14, FIN-20520 Turku, Finland

Turku Centre for Computer Science
TUCS Technical Report No 346
May 2000
ISBN 952-12-0659-4
ISSN 1239-1891

Abstract

During the last ten years, data mining, also known as knowledge discovery in databases, has established its position as a prominent and important research area. Mining association rules is one of the important research problems in data mining. Many algorithms have been proposed to find association rules in large databases containing both categorical and quantitative attributes. We generalize this to the case where part of attributes are given weights to reflect their importance to the user.

In this paper, we introduce the problem of mining weighted quantitative association rules based on fuzzy approach. Using the fuzzy set concept, the discovered rules are more understandable to a human.

We propose two different definitions of weighted support: with and without normalization. In the normalized case, a subset of a frequent itemset may not be frequent, and we cannot generate candidate k-itemsets simply from the frequent (k-1)-itemsets. We tackle this problem by using the concept of *z-potential frequent subset* for each candidate itemset.

We give an algorithm for mining such quantitative association rules. Finally, we describe the results of using this approach on a real-life dataset.

Keywords: data mining, association rules, basket data, support, confidence, linguistic terms, weighted items

TUCS Research Group
Algorithmics Research Group

1 Introduction

The goal of data mining is to extract higher level information from an abundance of raw data. Mining association rules in one of the important research problems in data mining [1]. Given a set of transactions where each transaction is a set of items, an association rule is an expression of the form $X \Rightarrow Y$, where X and Y are sets of items. A well-known application of association rules is in market basket data analysis.

The problem of mining association rules was first introduced in [2], and later broadened in [3], for the case of databases consisting of categorical attributes alone. Categorical association rules are rules where the events X and Y, on both sides of the rule, are the appearance of given categorical items. In this case, we wish to find all rules with confidence and support above user-defined thresholds (minconf and minsup). Several efficient algorithms for mining categorical association rules have been published (see [3], [4], [5] for just a few examples).

A variation of categorical association rules was recently introduced in [6]. Their new definition is based on the notion of weighted items to represent the importance of individual items.

In practice the information in many, if not most, databases is not limited to categorical attributes, but also contains much quantitative data. Unfortunately, the definition of categorical association rules does not translate directly to the case of quantitative attributes. It is therefore necessary to provide a definition of association rules for the case of a database containing quantitative attributes. The problem of mining quantitative association rules is introduced and an algorithm has been proposed in [7]. The algorithm finds the association rules by partitioning the attribute domain, combining adjacent partitions, and then transforming the problem into binary one. An example of a rule according to this definition would be: "10% of married people between age 50 and 70 have at least 2 cars".

Other work on this problem includes [8] where clustering methods are used to improve the partitioning of the quantitative attributes in the algorithm. [9] and [10] also worked on the quantitative association problem. However, their work is related to a different version of the problem and is focused more on prediction, rather than association rules.

In [11], we showed a method to handle quantitative attributes using a fuzzy approach. We assigned each quantitative attribute several fuzzy sets which characterize it. Fuzzy sets provide a smooth transition between a member and non-member of a set. The fuzzy association rule is also easily understandable to a human because of the linguistic terms associated with the fuzzy sets. Using the fuzzy set concept, the above example could be rephrased e.g. "10% of married old people have several cars".

In this paper, we introduce a new definition of the notion of weighted itemsets based on fuzzy set theory. In a marketing business, a manager may want to mine the association rules with more emphasis on some itemsets in mind, and less emphasis on other itemsets. For example, some itemsets may be more interesting for the company than others. This results in a generalized version of the quantitative association rule mining problem, which we call weighted quantitative association rule mining.

The paper is organized as follows. In the next section, we will present the definition of mining quantitative association rules using a fuzzy approach. Then we will introduce the problem of weighted quantitative association rules in section 3. In section 4, we give a new algorithm for this problem. In section 5 the experimental results are reported, followed by a brief conclusion in section 6.

2 Fuzzy Association Rules

In [11], an algorithm for mining quantitative association rules using a fuzzy approach was proposed. We summarize its definitions in what follows.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be the complete item set where each i_j $(1 \le j \le m)$ denotes a categorical or quantitative (fuzzy) attribute. Suppose $f(i_j)$ represents the maximum number of categories (if i_j is categorical) or the maximum number of fuzzy sets (if i_j is fuzzy), and $d_{i_j}(l, v)$ represents the membership degree of v in the l^{th} category or fuzzy set of i_j. If i_j is categorical, $d_{i_j}(l, v) = 0$ or $d_{i_j}(l, v) = 1$. If i_j is fuzzy, $0 \le d_{i_j}(l, v) \le 1$.

Let $t = \{t.i_1, t.i_2, \ldots, t.i_m\}$ be a transaction, where $t.i_j$, $(1 \le j \le m)$ represents a value of the j^{th} attribute and can be mapped to

$$\Big\{\Big(l, d_{i_j}\left(l, t.i_j\right)\Big) \mid \text{ for all } l, 1 \le l \le f(i_j)\Big\}.$$

Given a database $D = \{t_1, t_2, \ldots, t_n\}$ with attributes I and the fuzzy sets associated with attributes in I, we want to find out some interesting, potentially useful regularities.

Definition 1. *A **fuzzy association rule** if of the form*

$$\textit{If } X = \{x_1, x_2, \ldots, x_p\} \textit{ is } A = \{a_1, a_2, \ldots, a_p\} \textit{ then } Y = \{y_1, y_2, \ldots, y_q\} \textit{ is } B = \{b_1, b_2, \ldots, b_q\},$$

where

$$a_i \in \{\textit{fuzzy sets related to attribute } x_i\},$$
$$b_j \in \{\textit{fuzzy sets related to attribute } y_j\}$$

and X, Y are itemsets.

X and Y are ordered subsets of I and they are disjoint i.e. they share no common attributes. A and B contain the fuzzy sets associated with the corresponding attributes in X and Y. As in the binary association rule, "X is A" is called the antecedent of the rule while "Y is B" is called the consequent of the rule.

If a rule is interesting, it should have enough support and a high confidence value. We define the terms of support and the confidence as in [11]: the fuzzy support value is calculated by first summing all votes of each record with respect to the specified itemset, then dividing it by the total number of records. Each record contributes a vote which falls in $[0, 1]$. Therefore, a fuzzy support value reflects not only the number of records supporting the itemset, but also their degree of support.

Definition 2. *The **fuzzy support value** of itemset $\langle X, A\rangle$ in transaction set D is*

$$FS_{\langle X,A\rangle} = \frac{\sum_{t_i\in D}\Pi_{x_j\in X} d_{x_j}(a_j, t_i.x_j)}{|D|}$$

Definition 3. *An itemset $\langle X, A\rangle$ is called a **frequent itemset** if its fuzzy support value is greater than or equal to the minimum support threshold.*

We use the discovered frequent itemsets to generate all possible rules. If the union of antecedent $\langle X, A\rangle$ and consequent $\langle Y, B\rangle$ has enough support and the rule has high confidence, this rule will be considered as interesting.

When we obtain a frequent itemset $\langle Z, C\rangle$, we want to generate fuzzy association rules of the form, "If X is A then Y is B", where $X \subset Z$, $Y = Z - X$, $A \subset C$ and $B = C - A$. Having the frequent itemset, we know its support as well as the fact that all of its subsets will be also frequent.

Definition 4. *The **fuzzy confidence value** of a rule is as follows:*

$$\begin{aligned} FC_{\langle\langle X,A\rangle,\langle Y,B\rangle\rangle} &= \frac{FS_{\langle Z,C\rangle}}{FS_{\langle X,A\rangle}} \\ &= \frac{\frac{\sum_{t_i\in D}\Pi_{z_j\in Z} d_{z_j}(c_j, t_i.z_j)}{|D|}}{\frac{\sum_{t_i\in D}\Pi_{x_j\in X} d_{x_j}(a_j, t_i.x_j)}{|D|}} \\ &= \frac{\sum_{t_i\in D}\Pi_{z_j\in Z} d_{z_j}(c_j, t_i.z_j)}{\sum_{t_i\in D}\Pi_{x_j\in X} d_{x_j}(a_j, t_i.x_j)}, \end{aligned}$$

where $Z = X \cup Y$, $C = A \cup B$.

Table 1: Part of Database Containing Membership

$\langle Balance, medium\rangle$	$\langle Credit, high\rangle$	$\langle Income, high\rangle$
0.5	0.6	0.4
0.8	0.9	0.4
0.7	0.8	0.7
0.9	0.8	0.3
0.9	0.7	0.6

Example 1. We use an example to illustrate the computation of the fuzzy support value and the fuzzy confidence value. Let $X = \{Balance, Income\}$, $A = \{medium, high\}$ and $Y = \{Credit\}$, $A = \{high\}$ and a part of database shown in Table 1.

The fuzzy support of itemset $\langle X, A\rangle$ is

$$FS_{\langle X,A\rangle} = \frac{0.2 + 0.32 + 0.49 + 0.27 + 0.54}{5} = 0.364.$$

The fuzzy confidence value of the rule, "If $Balance$ is $medium$ and $Income$ is $high$ then $Credit$ is $high$" is as follows.

$$FC_{\langle\langle X,A\rangle,\langle Y,B\rangle\rangle} = \frac{0.12 + 0.288 + 0.392 + 0.216 + 0.378}{0.2 + 0.32 + 0.49 + 0.27 + 0.54} = 0.766.$$

3 Weighted Quantitative Association Rules

Let $\langle X, A\rangle$ the *itemset-fuzzy set pair*, where X is set of attributes x_j and A is the set of fuzzy sets a_j. We assign a weight $w_{(x,a)}$ for each 1-itemset $\langle X, A\rangle$, with $0 \leq w_{(x,a)} \leq 1$, to show the importance of the item.

Generalizing Definition 2, we can define the weighted fuzzy support for the weighted itemset as follows:

Definition 5. *The* ***weighted fuzzy support*** *of an itemset $\langle X, A\rangle$ is*

$$\begin{aligned} WFS_{\langle X,A,w\rangle} &= \left(\Pi_{x_j\in X} w_{(x_j,a_j)}\right) \cdot FS_{\langle X,A\rangle} \\ &= \frac{\sum_{t_i\in D} \Pi_{x_j\in X} w_{(x_j,a_j)} d_{x_j}(a_j, t_i.x_j)}{|D|} \end{aligned}$$

Notice the difference from [6] where the sum of weights is used, instead of the product. There are still other ways to define the combined weight. For example, the minimum of item weights would also make sense, and result in a simple algorithm. Product and minimum share the important property that if one item has zero weight, then the whole set has zero weight.

Similar to [11], a support threshold and a confidence threshold will be assigned to measure the strength of the association rules.

Definition 6. *A itemset $\langle X, A\rangle$ is called a* ***frequent itemset*** *if the weighted fuzzy support of such itemset is greater than or equal to the (user defined) minimum support threshold, or*

$$WFS_{\langle X,A,w\rangle} \geq minsup$$

Definition 7. *The* ***weighted fuzzy confidence*** *of the rule "If X is A then Y is B" as follows:*

$$WFC_{\langle\langle X,A,w\rangle,\langle Y,B,w\rangle\rangle} = \frac{WFS_{\langle Z,C,w\rangle}}{WFS_{\langle X,A,w\rangle}}$$

where $Z = X \cup Y$, $C = A \cup B$.

Definition 8. *An fuzzy association rule "If X is A then Y is B" is called an* ***interesting rule*** *if $X \cup Y$ is a frequent itemset and the confidence, defined in definition 7, is greater than or equal to a (user defined) minimum confidence threshold.*

Table 2: Part of (user defined) Weights

Item	Weigth
$\langle Balance, medium\rangle$	0.6
$\langle Credit, high\rangle$	0.1
$\langle Income, high\rangle$	0.9

Example 2. Given the database in Table 1, Table 2 shows their sample weights. If the value of minsup is 0.15, then the itemset $\langle\langle Balance, medium\rangle$, $\langle Income, high\rangle\rangle$ will be a frequent itemset since

$$(0.6 \cdot 0.9) \cdot 0.364 = 0.197 \geq 0.15.$$

The weighted fuzzy confidence value of the rule, "If *Balance* is *medium* then *Income* is *high*" is

$$\frac{(0.6 \cdot 0.9) \cdot 0.364}{0.6 \cdot 0.760} = 0.432.$$

The frequent itemsets in the weighted approach have the important property, that all its subsets are also frequent. Thus, we can apply the traditional bottom-up algorithm, tailored to fuzzy sets in [11].

4 Normalized Weighted Quantitative Association Rules

There is one possible problem with our definition. Even if each item has a large weight, when the number of items in an itemset is large, the total weight may be very small.

In this section, we deal with the mining of weighted association rules for which the weight of an itemset is normalized by the size of the itemset.

Definition 9. *The **normalized weighted fuzzy support** of itemset $\langle X, A \rangle$ is given by*

$$NWFS_{\langle X,A,w \rangle} = \left(\Pi_{x_j \in X} w_{(x_j,a_j)} \right)^{1/k} \cdot FS_{\langle X,A,w \rangle}$$

where $k =$ size of the itemset $\langle X, A \rangle$.

Notice that we actually use the geometric mean of item weights as the combined weight. This is a direct analogy to [6], where the arithmetic mean is applied in normalization.

Definition 10. *A k-itemset $\langle X, A \rangle$ is called a **frequent itemset** if the normalized weighted fuzzy support of such an itemset is greater than or equal to the minimum support threshold, or*

$$NWFS_{\langle X,A,w \rangle} \geq minsup$$

Example 3. Let $X = \{Balance, Credit\}$ and $A = \{medium, high\}$ and a database is shown in Table 1 and Table 2, then

$$NWFS_{\langle X,A,w \rangle} = (0.6 \cdot 0.1)^{1/2} \cdot 0.586 = 0.144.$$

If the value of minsup is 0.10, then itemset $\langle \langle Balance, medium \rangle, \langle Credit, high \rangle \rangle$ will be a frequent itemset. However, it is not necessarily true for all subsets of a frequent itemset to be frequent. In this example, $\langle Credit, high \rangle$ is a subset of the frequent itemset $\langle \langle Balance, medium \rangle, \langle Credit, high \rangle \rangle$, but it is not a frequent itemset.

In [11], we generated frequent itemsets with increasing sizes. However, since the subset of a frequent itemset may not be frequent, we cannot generate candidate k-itemsets simply from the frequent $(k-1)$-itemsets. For all k-itemset, which may contribute to be subsets of future frequent itemsets, will be kept in candidate generation process.

We can tackle this problem by using a new *z-potential frequent subset* for each candidate itemset.

Definition 11. *A k-itemset* $\langle X, A \rangle$ *is called a* ***z-potential frequent subset*** *if*

$$\left(\Pi_{x_j \in X} w_{(x_j, a_j)} \cdot \Pi_{y_j \in Y} w_{(y_j, b_j)}\right)^{1/z} \cdot FS_{\langle X,A,w \rangle} \geq minsup$$

where z *is number between* k *and the maximum possible size of the frequent itemset, and for each* $\langle Y, B \rangle$, $Y \neq X$, *is the remaining itemset with maximum weights.*

Example 4. Referring to Tables 1 and 2, the 2-potential frequent subset for the itemset $\langle Credit, high \rangle$ is

$$((0.1) \cdot 0.9)^{1/2} \cdot 0.760 = 0.228 \geq 0.10,$$

which implies that it may be the subset of any frequent 2-itemsets (notice that 0.9 is the maximum remaining weight).

4.1 Algorithm for Mining Normalized Weighted Association Rules

A trivial algorithm would be to solve first the non-weighted problem (weights=1), and then prune the rules that do not satisfy the weighted support and confidence. However, we can take advantage of weights to prune non-frequent itemsets earlier, and thereby cut down the number of trials.

An algorithm for mining normalized weighted quantitative association rules has the following inputs an outputs.
Inputs: A database D, two threshold values *minsup* and *minconf.*
Output: A list of interesting rules.
Notations:

D	the database
D_T	the transformed database
w	the itemset weights
F_k	set of *frequent k-itemsets* (have k items)
C_k	set of *candidate k-itemsets* (have k items)
I	complete item set
$minsup$	support threshold
$minconf$	confidence threshold

Main Algorithm $(minsup, minconf, D)$

I = Search(D);
(C_1, D_T, w) = Transform(D, I);
$k = 1$;
(C_k, F_k) = Checking($C_k, D_T, minsup$);
while ($|C_k| \neq \emptyset$) do
begin
    $inc(k)$;
    if $k == 2$ then
        C_k = Join1(C_{k-1})
    else C_k = Join2(C_{k-1});
    C_k =Prune(C_k);
    (C_k, F_k) = Checking($C_k, D_T, minsup$);
    $F = F \cup F_k$;
end
$Rules(F, minconf)$;

The subroutines are outlined as follows:

1. **Search**(D): The subroutine accepts the database, finds out and returns the complete item set $I = \{i_1, i_2, \ldots i_m\}$.

   For example, $I = \{Balance, Credit, Income\}$ for the database given in Table 1.

2. **Transform**(D, I): This step generates both a new transformed (fuzzy) database D_T from the original database by user specified fuzzy sets and weights for each fuzzy set. At the same time, the *candidate 1-itemsets* C_1 will be generated from the transformed database. If a 1-itemset is *frequent* or *z-potential frequent subset* then it will be kept in C_1, else it will be pruned. For example, $C_1 = \{\langle Balance, high\rangle, \langle Balance, low\rangle, \langle Balance, medium\rangle, \langle Credit, high\rangle, \langle Credit, medium\rangle, \langle Credit, low\rangle\}$ is the complete set of *candidate 1-itemsets.*

3. **Checking**($C_k, D_T, minsup$): In this subroutine, the transformed (fuzzy) database is scanned and the weighted fuzzy support of candidates in C_k is counted. If its weighted fuzzy support is larger or equal than $minsup$, we put it into the frequent itemsets F_k.

4. **Join1**(C_{k-1}): This Join step generates C_2 from C_1 as follows:
   insert into C_2
   select $\langle X, A\rangle, \langle Y, B\rangle$
   from $\langle X, A\rangle, \langle Y, B\rangle$ in C_1
   where $X \neq Y$

   For example, after this Join step C_2 will be $C_2 = \{\langle\langle Balance, high\rangle, \langle Income, high\rangle\rangle, \langle\langle Balance, medium\rangle, \langle Income, low\rangle\rangle, \cdots\}$, but $C_2 \neq \{\cdots, \langle\langle Balance, low\rangle, \langle Balance, medium\rangle\rangle, \cdots\}$.

5. **Join2**(C_{k-1}): This Join step generates C_k from C_{k-1} as in [11].

6. **Prune**(C_k): During the prune step, the itemset will be pruned in either of the following cases:
   - A subset of the candidate itemset in C_k does not exist in C_{k-1}.
   - The itemset cannot be a *z-potential frequent subset* of any frequent itemset.

7. **Rules**(F): Find the rules from the *frequent itemsets* F, similar to [2].

5 Experimental Results

We assessed the effectiveness of our approach by experimenting with a real-life dataset. The data had 5 quantitative attributes: monthly-income, credit-limit, current-balance, year-to-date balance, and year-to-date interest. There were 500,000 records in the data.

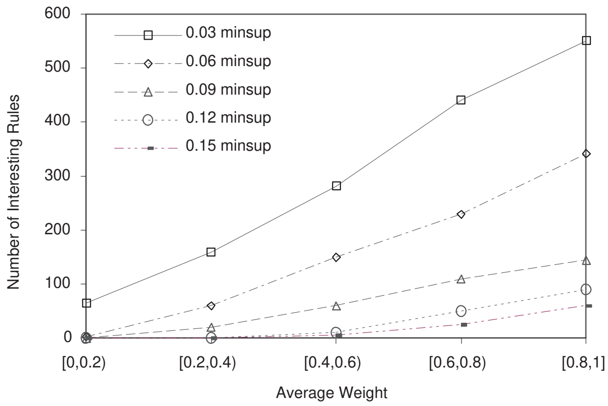

Figure 1: Random Weight Intervals

The experiments will be done applying only the normalized version of the algorithm (as discussed in Section 4), due to reasons explained above. We use the above five quantitative attributes where three fuzzy sets are defined for each of them.

Figure 1 shows the increase of the number of rules as a function of average weight, for five different support thresholds. The minimum confidence was set to 0.25. We used five intervals, from which random weights were generated. The increase of the number of rules is close to linear with respect to the average weight.

In the following experiments we used a random weight between 0 and 1 for each fuzzy set.

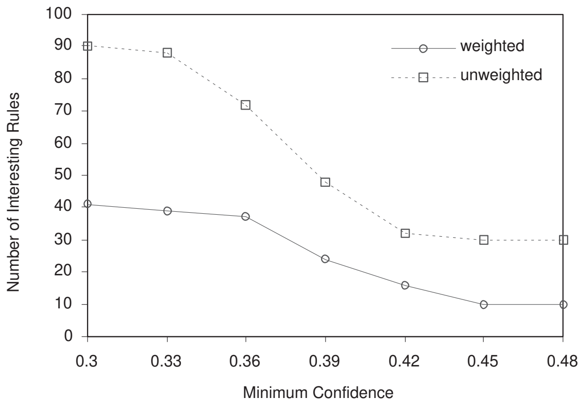

Figure 2: Number of Interesting Rules

Figure 2 shows the number of generated rules as a function of minimum confidence threshold, for both weighted and unweighted case (as in [11]). The minimum support was set to 0.1. The results are as expected: the numbers of rules for the unweighted case are larger, but both decrease with increasing confidence threshold.

Finally, we examined how the performance varies with the number of records. This is confirmed by Figure 3, which shows the execution time as we increase the number of input records from 100,000 to 500,000, for five different support thresholds. The graphs show that the method scales quite linearly for this dataset.

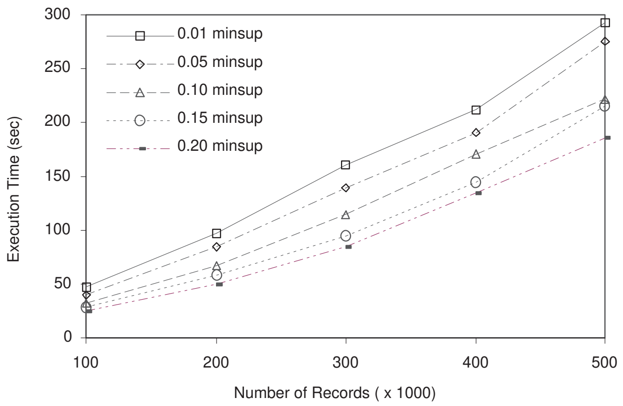

Figure 3: Scale-up: Number of Records

6 Conclusion

We have proposed generalized mining of weighted quantitative association rules based on fuzzy sets for data items. This is an extension of the fuzzy quantitative association mining problem. In this generalization, the fuzzy sets are assigned weights to reflect their importance to the user. The fuzzy association rule is easily understandable to a human because of the linguistic terms associated with the fuzzy sets.

We proposed two different definitions of weighted support: without normalization, and with normalization. In the normalized case, a subset of a frequent itemset may not be frequent, and we cannot generate candidate k-itemsets simply from the frequent (k-1)-itemsets. We tackled this problem by using the concept of *z-potential frequent subset* for each candidate itemset.

We proposed a new algorithm for mining weighted quantitative association rules. The algorithm is applicable to both normalized and unnormalized cases but we prefer the former, as explained above. Therefore, the performance evaluation has been done only for the normalized version.

The results show that by associating weight to fuzzy sets, non-interesting rules can be pruned early and execution time is reduced.

Acknowledgment

I wish to thank Jukka Teuhola for his comments and suggestions.

References

[1] Piatesky-Sshapiro, G., Frawley, W.J.: Knowledge Discovery in Databases. AAAI Press/The MIT Press, Menlo Park, California (1991)

[2] Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. Proceedings of ACM SIGMOD (1993) 207–216

[3] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. Proceedings of the 20th VLDB Conference (1994) 487–499

[4] Mannila, H., Toivonen, H., Verkamo, A.I.: Efficient Algorithms for discovering association rules. KDD-94: AAAI Workshop on Knowledge Discovery in Databases (1994) 181–192.

[5] Toivonen, H.: Sampling large databases for association rules. Proceedings of the 20th VLDB Conference (1996)

[6] Cai, C.H., Fu, Ada W.C., Cheng, C.H., Kwong, W.W.: Mining Association Rules with Weighted Items. Proceedings of IEEE International Database Engineering and Applications Symposium (1998) 68–77.

[7] Srikant, R., Agrawal, R.: Mining quantitative association rules in large relation tables. Proceedings of ACM SIGMOD (1996) 1–12.

[8] Zhang, Z., Lu, Y., Zhang, B.: An Effective Partitioning-Combining ALgorithm for Discovering Quantitative Association Rules. Proceedings of the First Pacific-Asia Conference on Knowledge Discovery and Data Mining (1997)

[9] Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama T.: Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms and Visualisation. Proceedings of the ACM SIGMOD Conference on Management of Data (1996)

[10] Yoda., K., Fukuda, T., Morimoto, Y., Morishita, S., Tokuyama T.: Computing Optimized Rectilinear Regions for Association Rules. Proceedings of KDD Conference (1997)

[11] Gyenesei, A.: A Fuzzy Approach for Mining Quantitative Association Rules. Turku Centre for Computer Science, Technical Report No. 336. (2000)

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.abo.fi

University of Turku
- **Department of Mathematical Sciences**

Åbo Akademi University
- **Department of Computer Science**
- **Institute for Advanced Management Systems Research**

Turku School of Economics and Business Administration
- **Institute of Information Systems Science**