JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 23, 233-242 (2007)

Short Paper

Online Generation of Association Rules under Multi-dimensional Consideration Based on Negative-Border*

CHING-YAO WANG, SHIAN-SHYONG TSENG, TZUNG-PEI HONG+
AND YIAN-SHU CHU
Institute of Computer and Information Science
National Chiao Tung University
Hsinchu, 300 Taiwan
+*Department of Electrical Engineering*
National University of Kaohsiung
Kaohsiung, 811 Taiwan

Recently, some researchers have developed incremental and online mining approaches to maintain association rules without having to re-process the entire database whenever the database is updated or user specified thresholds are changed. However, they usually can not flexibly obtain association rules or patterns from portions of data, consider problems with different aspects, or provide online decision support for users. We earlier developed an online mining approach for generation of association rules under multidimensional consideration. The multidimensional online mining approach may, however, get loose upper-bound support of candidate itemsets and thus cause excessive I/O and computation costs. In this paper, we attempt to apply the concept of a negative border to enlarge the mining information in the multidimensional pattern relation to help get tighter upper-bound, and thus reduce the number of candidate itemsets to consider. Based on the extended multidimensional pattern relation, a corresponding online mining approach called Negative-Border Online Mining (NOM) is proposed to efficiently and effectively utilize the information of negative itemset in the negative border. Experiments for heterogeneous datasets are also performed to show the effectiveness of the proposed approach.

Keywords: Apriori algorithm, association rule, data mining, incremental mining, multi-dimensional mining, negative border

1. INTRODUCTION

Developing mining association rules from transactional databases has been one of the most interesting and popular research topics in data mining [2, 7]. Since the process of mining association rules is rather costly and time-consuming, some famous ap-

Received October 14, 2004; revised December 17, 2004; accepted February 14, 2005.
Communicated by Chin-Teng Lin.

* This is a modified and expanded version of the paper "A three-phased online association rule mining approach for diverse mining requests," presented at The Fourth International Conference on Electronic Business, Beijing, China, 2004. This research was supported by the National Science Council of Taiwan, R.O.C., under grand No. NSC 93-2752-E-009-006-PAE.

proaches, such as Apriori [3], DIC [5], DHP [18], Partition [19] and Sampling [16], have been proposed to reduce computation time and improve performance. However, these approaches process the data in a batch way. They do not utilize previously mined patterns for later maintenance, and may require considerable computation time to obtain the updated set of association rules or patterns [8]. Incremental data mining [8, 9, 14, 20] and online mining [1, 4, 13] have thus drawn much attention in recent years.

Although incremental data mining and online mining approaches are rather efficient, most of them focus on finding association rules or patterns in a specified part of a database. Some contexts (circumstance information) such as region, time and branch are usually ignored in mining requests. However, decision makers usually consider problems at different contexts and aspects [10-12]. They may need to analyze market demands, customer preferences, localities, and short-term/long-term trends.

Assume that data under decision support consideration evolve in a systematic way. For example, data may be inserted or deleted in a block during an interval of a month. In our previous study [21], we proposed the *multidimensional pattern relation* to structurally and systematically store the additional context information and mining information for each inserted block of data, and then developed a *multidimensional online mining* approach for online generation of association rules under multidimensional consideration. The multidimensional online mining approach may, however, get loose upper bound supports of candidate itemsets if they do not appear in the tuples of the multidimensional pattern relation, and their supports are thus estimated as the predefined minimum support [21]. This will cause a larger number of candidate itemsets to be considered. In this paper, we attempt to apply the concept of *negative border* to increase the mining information in the multidimensional pattern relation, calculate tighter upper bound supports of candidate itemsets, and then reduce the number of candidate itemsets. Based on the extended multidimensional pattern relation, we then develop an online mining approach called *Negative-Border Online Mining* (NOM) to utilize the information of a *negative itemset* in the negative border.

2. REVIEW OF INCREMENTAL MINING APPROACHES

In real-world applications, a database grows over time with the result that existing association rules may become invalid or that new rules may appear. To overcome these problems, some researchers have developed incremental mining algorithms to maintain association rules without having to reprocess the database. Considering an original database and newly inserted transactions, the four cases illustrated in Fig. 1 may arise.

Cheung *et al*. proposed an incremental mining algorithm, called FUP [8, 9], to efficiently cope with these four cases by pre-storing the previously mined large itemsets from the original database. It first calculates the large itemsets from the newly inserted transactions as candidates, and compares them with the pre-stored large itemsets. According to the comparison results, FUP can efficiently handle Cases 1, 2 and 4, and reprocess only the itemsets without sufficient information in Case 3 against the original database if necessary.

The performance of the FUP algorithm will get degraded if a lot of candidate itemsets from the newly inserted transactions belong to Case 3. As a result, the concept of

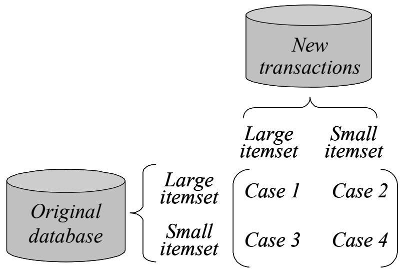


Fig. 1. Four cases arising from adding new transactions to existing databases.

negative border [17] was used to enlarge the amount of pre-stored mining patterns in incremental mining for further improving the maintenance performance at the expense of storage space [20]. A negative border is defined as follows.

Definition 1 Let R be a set of items, and L be a subset of the power set of R, which is closed with respect to the set inclusion relation. The *negative border* $NB(L)$ for L is a set that consists of the minimal itemsets $X \subseteq R$ and $X \notin L$.

For example, If $R = \{A, B, C, D\}$ and $L = \{\{A\}, \{B\}, \{AB\}\}$, then $NB(L) = \{\{C\}, \{D\}\}$.

Definition 2 A *negative itemset* for L is an element of $NB(L)$.

3. THE EXTENDED MULTIDIMENSIONAL PATTERN RELATION

The *extended multidimensional pattern relation* is conceptually similar to the construction of a data warehouse for OLAP [6, 15]. Both systematically preprocess the underlying data, integrate related information, and store the results in a centralized structured repository for later use and analysis. For providing ad-hoc, query- driven and online mining support, the extended multidimensional pattern relation consists of two major types of information. One is the *context information* used to represent the contexts of each block of data which are gathered together from a specific business viewpoint, such as region, time and branch, for multidimensional consideration. The other is the *mining information* used to record the available information mined from each individual block of data by a batch mining algorithm, such as the number of transactions, the number of mined large itemsets, and the set of previously mined large itemsets and negative itemsets with their supports, for efficient online mining. The definitions of the extended multidimensional pattern relation and its schema are as follows.

Definition 3 An *extended multidimensional pattern relation schema EMPR* with n_1 context attributes and n_2 content attributes can be represented as $EMPR(ID, CX_1, CX_2, \ldots, CX_{n_1}, CN_1, CN_2, \ldots, CN_{n_2})$, where ID is an identification attribute, CX_i, $1 \le i \le n_1$, is a context attribute, and CN_i, $1 \le i \le n_2$, is a content attribute.

Definition 4 An *extended multidimensional pattern relation* including tuples $\{t_1, t_2, \ldots, t_m\}$ is an instance of the given $EMPR(ID, CX_1, CX_2, \ldots, CX_{n_1}, CN_1, CN_2, \ldots, CX_{n_2})$. A tuple $t_i = (id_i, cx_{i1}, cx_{i2}, \ldots, cx_{in_1}, cn_{i1}, cn_{i2}, \ldots, cn_{in_2})$ in an extended multidimensional pattern relation indicates that for the block of data under the contexts of $cx_{i1}, cx_{i2}, \ldots, cx_{in_1}$, the mining information contains $cn_{i1}, cn_{i2}, \ldots, cn_{in_2}$.

The *frequent pattern se*t and the *negative pattern set* are two essential content attributes which are defined as follows.

Definition 5 A *frequent pattern set* (*fps*) for a block of data D is the set of all previously mined large itemsets with their supports for D. Assume that the minimum support is s and the number of large itemsets discovered from D is l. A frequent pattern set can be represented as $fps = \{(x_i, s_i) \mid s_i \geq s \text{ and } 1 \leq i \leq l\}$, where x_i is a large itemset and s_i is its support.

Definition 6 A *negative pattern set* (*nps*) for a block of data D is the set of all previously mined negative itemsets with their supports from NB (*fps*) for D.

4. ONLINE GENERATION OF ASSOCIATION RULES

The goal of online generation of association rules is to find the association rules satisfying the constraints in a mining request on line. The flexibility of mining requests allowed can increase through the usage of the proposed extended multidimensional pattern relation. Assume that an extended multidimensional pattern relation based on an initial minimum support s includes tuples $\{t_1, t_2, \ldots, t_m\}$. Given a mining request q with a set of contexts cx_q, a new minimum support s_q ($s_q \geq s$), and a new minimum confidence $conf_q$, the online generation of association rules can select the tuples from the relation satisfying cx_q, integrate the mining information in these tuples, and then derive the association rules simultaneously satisfying s_q and $conf_q$. Let s_x denote the support of an itemset x, t_i denote the i-th tuple in an extended multidimensional pattern relation, $t_i.trans$ denote the number of transactions kept in t_i, $t_i.fps$ denote the *frequent pattern set* for t_i, $t_i.nps$ denote the *negative pattern set* for t_i, and $t_i.s_x$ denote the support of x in t_i. Also, a tuple in an extended multidimensional pattern relation is called a *matched tuple* (*mt*) if it satisfies the given context constraints. The following lemma can be derived (the proofs are omitted here).

Lemma 1 If an itemset x satisfies a mining request q, there must exist at least a matched tuple t, such that $t.s_x$ satisfies s_q.

Lemma 2 If an itemset x satisfies a mining request q, it must belong to the candidate itemsets obtained by collecting the ones whose supports are larger than or equal to s_q in a matched tuple.

Lemma 3 If x is a candidate itemset, then $\forall x' \subset x$, x' is also a candidate itemset.

Definition 7 The *appearing count* $Count_x^{appearing}$ of a candidate itemset x is defined as $Count_x^{appearing} = \sum_{t_i \in mt \ \& \ x \in t_i.fps \cup t_i.nps} t_i.trans \times t_i.s_x$.

Definition 8 The not-appearing *upper bound count* $Count_{\bar{x}}^{UB}$ of a candidate itemset x is defined as $Count_{\bar{x}}^{UB} = \sum_{t_i \in mt \ \& \ x \notin t_i.fps \cup t_i.nps} \min(t_i.trans \times s - 1,\ t_i.trans * \min_{\forall x' \subset x}(t_i.s_{x'}))$.

Definition 9 The *upper bound support* s_x^{UB} of a candidate itemset x is defined as $s_x^{UB} = \dfrac{Count_x^{appearing} + Count_{\bar{x}}^{UB}}{Match\_Trans}$, where $Match\_Trans = \sum_{t_i \in mt} t_i.trans$ is the number of transactions in the matched tuples.

Note that in the multidimensional online mining approach [21], the upper bound support $s_x^{UB^{old}}$ of a candidate itemset x is:

$$s_x^{UB^{old}} = \frac{\sum_{t_i \in mt \ \& \ x \in t_i.ps} t_i.trans \times t_i.s_x + \sum_{t_i \in mt \ \& \ x \notin t_i.ps} (t_i.trans \times s - 1)}{\sum_{t_i \in mt} t_i.trans},$$

where $t_i.ps$ denotes the set consisting of all previously mined large itemsets with their supports in t_i. The following lemma can easily be derived to show that s_x^{UB} is tighter than $s_x^{UB^{old}}$.

Lemma 4 If x is a candidate itemset, then $s_x^{UB} \le s_x^{UB^{old}}$.

The following lemmas are important to the design of the proposed mining algorithm.

Lemma 5 If x is a candidate itemset, then $s_x \le s_x^{UB}$.

Lemma 6 If x is a candidate itemset, then $\forall x' \subset x, s_{x'}^{UB} \ge s_x^{UB}$.

Lemma 7 If a candidate itemset x is contained in all the matched tuples, then $s_x^{UB} = s_x$.

5. NEGATIVE-BORDER ONLINE MINING (NOM)

The NOM approach, which consists of three phases, *generation of candidate itemsets*, *reduction of candidate itemsets*, and *generation of association rules*, is described below.

The Negative-Border Online Mining (NOM) approach:
Input: an extended multidimensional pattern relation based on an initial minimum support s and a mining request q with a set of contexts cx_q, a minimum support s_q ($s_q \geq s$) and a minimum confidence $conf_q$.
Output: a set of association rules satisfying the mining request q.

Phase 1: Generation of candidate itemsets:
(a) Select the tuples satisfying cx_q from the extended multidimensional pattern relation.
(b) Collect the itemsets appearing in the matched tuples and satisfying s_q as the candidate itemsets for q.
(c) Calculate $Count_x^{appearing}$ and $Count_{\bar{x}}^{UB}$ for each candidate itemset x.

Phase 2: Reduction of candidate itemsets:
(a) Calculate s_x^{UB} for each candidate itemset x from

$$s_x^{UB} = \frac{Count_x^{appearing} + Count_{\bar{x}}^{UB}}{Match\_Trans}.$$

(b) Discard the candidate itemset x and its proper supersets from the candidate set if $s_x^{UB} < s_q$.
(c) Put x into the set of final large itemsets if $s_x^{UB} = \dfrac{Count_x^{appearing}}{Match\_Trans}$ and $s_x^{UB} \geq s_q$.

Phase 3: Generation of association rules:
(a) Check whether each remaining candidate itemset x is large by scanning the underlying blocks of data for the matched tuples in which x does not appear.
(b) Generate the association rules satisfying $conf_q$ from the final set of found large itemsets.

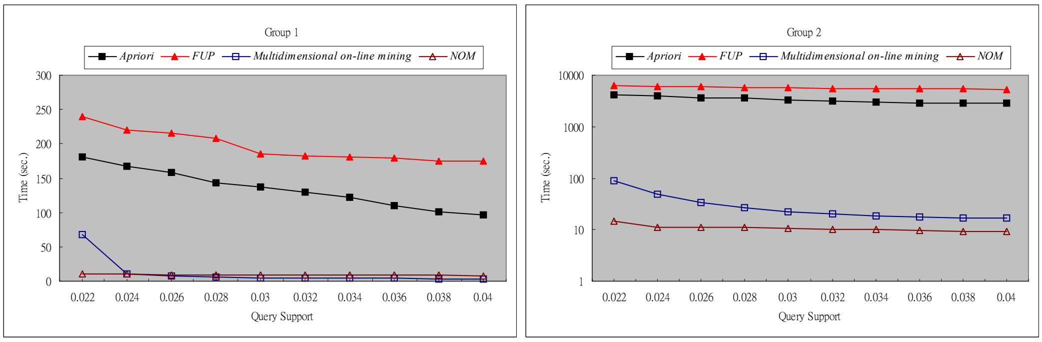
6. EXPERIMENTAL RESULTS

The experiments were implemented in Java on a workstation with dual XEON 2.8GHz processors and 2,048MB of main memory, running the RedHat 9.0 Linux operation system. The datasets were generated by a generator similar to that used in [3]. The generator first generated L maximal potentially large itemsets, each with an average size of I items. The items in a potentially large itemset were randomly chosen from N items. The generator then generated D transactions, each with an average size of T items. The items in a transaction were generated according to the L maximal potentially large itemsets in a probabilistic way. Two groups of datasets generated in the above way were used in our experiments and are listed in Table 1, where each dataset was treated as a block of data in the database. They could be treated as heterogeneous because of their varied N values.

In addition to our proposed NOM algorithm, the Apriori algorithm, the FUP algorithm, and the multidimensional online mining algorithm were run for these two groups along with different minimum supports in the mining requests. Among them, the Apriori algorithm, which is a batch-based mining algorithm, utilizes a level-wise candidate

Table 1. The two groups of datasets used in the experiments.

Group	*Size*	*Datasets*	*D*	*T*	*I*	*L*	*N*
1	10	$T10I8D10KN^1$ to $T10I8D10KN^{10}$	10,000	10	8	200	100 to 145
2	5	$T10I8D500KN^1$ to $T10I8D500KN^5$	500,000	10	8	400	200 to 360



(a) (b)

Fig. 2. The execution times needed by the four algorithms for the two groups.

generation approach to reduce its search space (such that only the large itemsets found in the previous level are treated as seeds for generating the candidate itemsets in the current level). The FUP algorithm treated each dataset of a group as a new addition of transactions. The execution times needed by the four algorithms are shown in Fig. 2.

It is easily seen that the execution time for the NOM algorithm was always much less than that for the Apriori algorithm or the FUP algorithm. The FUP algorithm can, in general, achieve better performance than the Apriori algorithm if the size of newly inserted transactions is smaller than the size of the original database. However, for the applications shown in this paper, each dataset in a group was treated as an increment. The performance of the FUP algorithm was worse than the Apriori algorithm.

It is also seen that the NOM algorithm performs better than the multidimensional online mining algorithm for Group 2. This is because most of the candidate itemsets appear in only one or a few tuples in the extended multidimensional pattern relation. The effect of negative border on finding tight upper bound supports thus becomes apparent. For Group 1, the NOM algorithm performs better for low support values and worse for high support values. The performance of the NOM algorithm is thus highly dependent on the pruning effects of candidate itemsets.

7. CONCLUSION AND FUTURE WORK

In this paper, the concept of *negative border* has been used to enlarge the mining information in the multidimensional pattern relation [21] to help get tight upper-bound supports of candidate itemsets and reduce the number of candidate itemsets to be considered. Based on the extended multidimensional pattern relation, a corresponding online mining approach called *Negative-Border Online Mining* (NOM) has been proposed to utilize the information of *negative itemsets* in the negative border. From the experimental results, the NOM algorithm can get good performance for heterogeneous datasets. In the future, we will attempt to use other techniques to further improve the performance of the proposed methodology. For example, we can use materialized views [6] for the proposed extended multidimensional pattern relation to provide more efficient online association rule generation and more powerful mining services.

REFERENCES

1. C. C. Aggarwal and P. S. Yu, "A new approach to online generation of association rules," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 13, 2001, pp. 527-540.
2. R. Agrawal, T. Imielinksi, and A. Swami, "Database mining: a performance perspective," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 5, 1993, pp. 914-925.
3. R. Agrawal and R. Srikant, "Fast algorithm for mining association rules," in *Proceedings of the ACM International Conference on Very Large Data Bases*, 1994, pp. 487-499.
4. W. G. Aref, M. G. Elfeky, and A. K. Elmagarmid, "Incremental, online, and merge mining of partial periodic patterns in time-series databases," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, 2004, pp. 332-342.
5. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *Proceedings of the ACM SIGMOD Conference*, 1997, pp. 255-264.
6. S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD Record*, Vol. 26, 1997, pp. 65-74.
7. M. S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from database perspective," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, 1996, pp. 866-883.
8. D. W. Cheung, J. Han, V. T. Ng, and C. Y. Wong, "Maintenance of discovered association rules in large databases: an incremental updating approach," in *Proceedings of IEEE International Conference on Data Engineering*, 1996, pp. 106-114.
9. D. W. Cheung, S. D. Lee, and B. Kao, "A general incremental technique for maintaining discovered association rules," in *Proceedings of Database Systems for Advanced Applications*, 1997, pp. 185-194.
10. G. Grahne, L. V. S. Lakshmanan, X. Wang, and M. H. Xie, "On dual mining: from patterns to circumstances, and back," in *Proceedings of IEEE International Conference on Data Engineering*, 2001, pp. 195-204.
11. J. Han, L. V. S. Lakshmanan, and R. Ng, "Constraint-based, multidimensional data mining," *IEEE Computer Magazine*, 1999, pp. 2-6.
12. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

13. C. Hidber, "Online association rule mining," in *Proceedings of the ACM SIGMOD Conference*, 1999, pp. 145-156.
14. T. P. Hong, C. Y. Wang, and Y. H. Tao, "A new incremental data mining algorithm using pre-large itemsets," *International Journal on Intelligent Data Analysis*, Vol. 5, 2001, pp. 111-129.
15. W. H. Immon, *Building the Data Warehouse*, Wiley Computer Publishing, 1996.
16. H. Mannila, H. Toivonen, and A. I. Verkamo, "Efficient algorithm for discovering association rules," in *Proceedings of the AAAI Workshop on Knowledge Discovery in Databases*, 1994, pp. 181-192.
17. H. Mannila and H. Toivonen, "On an algorithm for finding all interesting sentences," in *Proceedings of the European Meeting on Cybernetics and Systems Research*, Vol. II, 1996, pp. 973-978.
18. J. S. Park, M. S. Chen, and P. S. Yu, "Using a hash-based method with transaction trimming for mining association rules," *IEEE Transactions on Knowledge and Data* Engineering, Vol. 9, 1997, pp. 812-825.
19. A. Savasere, E. Omiecinski, and S. Navathe, "An efficient algorithm for mining association rules in large databases," in *Proceedings of the ACM International Conference on Very Large Data Bases*, 1995, pp. 432-444.
20. S. Thomas, S. Bodagala, K. Alsabti, and S. Ranka, "An efficient algorithm for the incremental update of association rules in large databases," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, 1997, pp. 263-266.
21. C. Y. Wang, T. P. Hong, and S. S. Tseng, "Multidimensional online mining," in *Proceedings of the Workshop on Foundation and New Directions in Data Mining in the Third IEEE International Conference on Data Mining*, 2003, pp. 196-202.

Ching-Yao Wang (王慶堯) received a B.S. degree in Information Management from Chang Jung University in 1998 and received a M.S. degree in Information Engineering from I-Shou University in 2000. He is currently a Ph.D. student in Institute of Computer and Information Science at National Chiao Tung University. His current research interests include data mining, data warehousing, expert system, artificial intelligence and soft computing, and Internet application.

Shian-Shyong Tseng (曾憲雄) received the Ph.D. degree in Computer Engineering from the National Chiao Tung University in 1984. Since August 1983, he has been on the faculty of the Department of Computer and Information Science at National Chiao Tung University, and is currently a Professor there. From 1988 to 1991, he was the Director of the Computer Center at National Chiao Tung University. From 1991 to 1992 and 1996 to 1998, he acted as the Chairman of Department of Computer and Information Science. From 1992 to 1996, he was the Director of the Computer Center at Ministry of Education and the Chairman of Taiwan Academic Network (TANet) management committee. Since 1999, he has participated the National Telecommunication Project and acted as the Chairman of the Network Planning Committee, National Broadband Experimental Network (NBEN). In Dec. 1999, he founded Taiwan Network Information Center (TWNIC)

and is now the Chairman of the board of directors of TWNIC. He is also the Director of the e-learning and application research center at National Chiao Tung University. His current research interests include expert systems, data mining, computer-assisted learning, and Internet-based applications. Dr. Tseng is an associate editor of Information and Education, and a member of IEEE and Phi Tau Phi Societies. He was named an Outstanding Talent of Information Science of the Republic of China in 1989. He obtained the 1992, 1994, and 1995 Outstanding Research Awards of the National Science Council of the Republic of China. He was the winner of the 1990, 1991, 1998 and 2000 Acer Long Term Awards for outstanding M.S. Thesis Supervision and the winner of 1992 and 1996 Acer Long Term Awards for outstanding Ph.D. Dissertation Supervision. He was also awarded Outstanding Youth Honor of R.O.C. in 1992.

Tzung-Pei Hong (洪宗貝) received his B.S. degree in Chemical Engineering from National Taiwan University in 1985, and his Ph.D. degree in Computer Science and Information Engineering from National Chiao Tung University in 1992. From 1987 to 1994, he was with the Laboratory of Knowledge Engineering, National Chiao Tung University, where he was involved in applying techniques of parallel processing to artificial intelligence. He was an Associate Professor at the Department of Computer Science in Chung Hua Polytechnic Institute from 1992 to 1994, and at the Department of Information Management in I-Shou University (originally Kaohsiung Polytechnic Institute) from 1994 to 1999. He was a professor in I-Shou University from 1999 to 2001. He was in charge of the whole computerization and library planning for National University of Kaohsiung in Preparation from 1997 to 2000 and served as the first director of the library and computer center in National University of Kaohsiung from 2000 to 2001. He is currently a professor at the Department of Electrical Engineering and the Dean of Academic Affairs in National University of Kaohsiung. His current research interests include parallel processing, machine learning, data mining, soft computing, management information systems, and www applications. Dr. Hong is a member of the Association for Computing Machinery, the IEEE, the Chinese Fuzzy Systems Association, the Taiwanese Association for Artificial Intelligence, and the Institute of Information and Computing Machinery.

Yian-Shu Chu (曲衍旭) received a B.S. degree in Department of Computer and Information Science from National Chiao Tung University in 2000 and received a MS degree in Institute of Computer and Information Science from National Chiao Tung University in 2002. He is currently a Ph.D. student in Institute of Computer and Information Science at National Chiao Tung University. His current research interests include Data Mining, Expert System and Artificial Intelligence.