南　華　大　學

資訊管理學系

碩士論文

模糊關聯法則挖掘架構及其應用

Fuzzy Association Rules Mining Framework and Its Applications

研 究 生：湯鎰聰

指導教授：邱宏彬

中華民國　九十五　年　六　月　三　日

南 華 大 學

資訊管理研究所

碩士學位論文

Fuzzy Association Rules Mining Framework and Its Applications

研究生：湯鎰聰

經考試合格特此證明

口試委員：

謝[illegible]霖

李翔詣

指導教授：邱宏彬

系主任(所長)：資訊管理學系 系主任吳光閔

口試日期：中華民國 95 年 6 月 3 日

模糊關聯法則挖掘架構及其應用

學生：湯鎰聰　　　　　　　　　　指導教授：邱宏彬

南華大學　資訊管理學系碩士班

摘　　要

在本研究中，提出了一個整合式模糊關聯法則挖掘架構，並分為二大部分：第一部分為從交易資料庫中挖掘出模糊關聯法則，這是一項重要的資料挖掘議題，並且能挖掘出許多未知且重要的決策資訊供決策制定。而在許多模糊關聯法則文獻中，對於模糊關聯法則挖掘演算法之效率更是一項重要的研究議題。本研究提出改良式模糊關聯法則挖掘演算法來解決模糊挖掘過程中之效率問題。第二部分為從模糊大項目組與網路瀏覽資料萃取出二項資訊間之模糊關聯法則。

本研究提出以群聚為基礎之模糊關聯法則挖掘演算法，並以模糊群聚表概念來改善以往模糊關聯法則挖掘演算法之效率不佳問題。實驗結果證明，以群聚為基礎之模糊關聯法則挖掘演算法確實有效改善模糊大項目組之處理效率。本研究架構所挖掘出之整合式模糊關聯法則資訊將有助於決策者進行決策制定。

關鍵字：資料挖掘、模糊關聯法則、群聚

Fuzzy Association Rules Mining Framework and Its Applications

Student：Yi-Tsung Tang　　　　　　Advisor：Dr. Hung-Pin Chiu

Department of Information Management
The M.B.A. Program
Nan-Hua University

ABSTRACT

In this paper, two important issues of mining association rules are investigated. The first problem is the discovery of generalized fuzzy association rules in the transaction database. It’s an important data-mining task, because more general and qualitative knowledge can be uncovered for decision making. However, few algorithms have been proposed in the literature, moreover, the efficiency of these algorithms needs to be improved to handle real-world large datasets. The second problem is to discover association rules from the web usage data and the large itemsets identified in the transaction database. This kind of rules will be useful for marketing decision.

A cluster-based mining architecture is proposed to address the two problems. At first, an efficient fuzzy association rule miner, based on cluster-based fuzzy-sets tables, is presented to identify all the large fuzzy itemsets. This method requires less contrast to generate large itemsets. Next, a fuzzy rule discovery method is used to compute the confidence values for discovering the relationships between transaction database and browsing information database. An illustrated example is given to demonstrate the effectiveness of the proposed methods and experimental results show that CBFAR outperforms a known Apriori-based fuzzy association rules mining algorithm.

Keywords: Data Mining, Fuzzy Association Rules, Cluster

List of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Data mining is defined as knowledge discovery from databases, which can be further explained as the process of nontrivial extraction of implicit, previously unknown and potentially useful knowledge from large databases [12]. Different kinds of methods and techniques are needed to find different kinds of knowledge. Based on the kinds of knowledge, tasks in data mining can be classified into summarization, classification, clustering, association, and trend analysis [11].

Deriving association rules from transaction database is most commonly seen in data mining [2, 8, 14, 15, 21]. Association rules are used to discover the relationships, and potential associations, of items or attributes among huge data. These rules can be effective in uncovering unknown relationships, providing results that can be the basis of forecast and decision. Therefore, the application and development of association rules is a popular area of data mining research [7].

In the past, Agrawal and Srikant proposed the Apriori association rule algorithm [16]. It can discover meaningful itemsets and construct association rules within large databases, but a large number of the candidate itemsets are generated from single itemsets. This method also needs to perform contrasts against all of the transactions, level by level, in the process of creating association rules. The database is repeatedly scanned to contrast each candidate itemset, that performance is dramatically affected and shown in Figure 1-1[26]. In the literature, many approaches, including the DHP algorithm [9], the DIC [20], the Sampling algorithm [6], and the cluster-based association rule (CBAR) approach [26], have been proposed to improve the efficiency of the mining process.

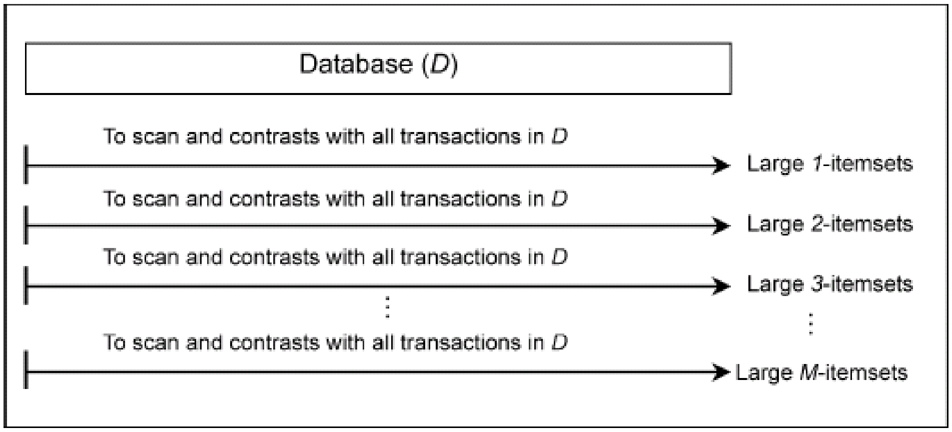

Database (*D*)
To scan and contrasts with all transactions in *D*
Large *1*-itemsets
To scan and contrasts with all transactions in *D*
Large 2-itemsets
To scan and contrasts with all transactions in *D*
Large 3-itemsets
To scan and contrasts with all transactions in *D*
Large *M*-itemsets

Figure 1-1: Illustrate database scans and contrasts of Apriori Algorithm[26]

Agrawal et al. also proposed a method for mining association rules from data sets using quantitative and categorical attributes[18]. Their proposed method first determines the number of partitions for each quantitative attribute, and then maps all possible values of each attribute onto a set of consecutive integers.

Recently, the fuzzy set theory [10] has been used more and more frequently in intelligent systems because of its simplicity and similarity to human reasoning [1]. Hong et al. also proposed a fuzzy mining algorithm for managing quantitative data [23]. The items considered in their approach had no hierarchical relationships. However, items in real-world applications are usually organized in some hierarchies. Mining multiple-concept-level fuzzy rules may lead to discover the generalized important knowledge from data [25].

Previous researches on mining association rules only focus on discovering the relationships among items in a transaction database. The relationships between different databases can find the different kinds of useful knowledge which are unknown and potentially. However, few researches have examined the important data-mining issue. Tsai et al.

proposed a mining algorithm named graph-based algorithm to discover the association rules between large items in the transaction database and attribute values in the customer database [13]. An example of such an association rule might be “80% of customers who specialties are data-mining buy itemset X.” These kinds of rules will be useful for marketing decision.

Recently, using the data mining techniques to extract information for knowledge discovery from web documents and services is an interesting and important data mining task for electronic commerce. Web mining refers to the use of data mining techniques to automatically retrieve, extract and evaluate, generalized, and analyze information for knowledge discovery from Web documents and services. Almost 90% of the data is useless, and often does not represent any relevant information that the user is looking for[3]. It can be broadly defined as the discovery and analysis of useful information from the World Wide Web [19].

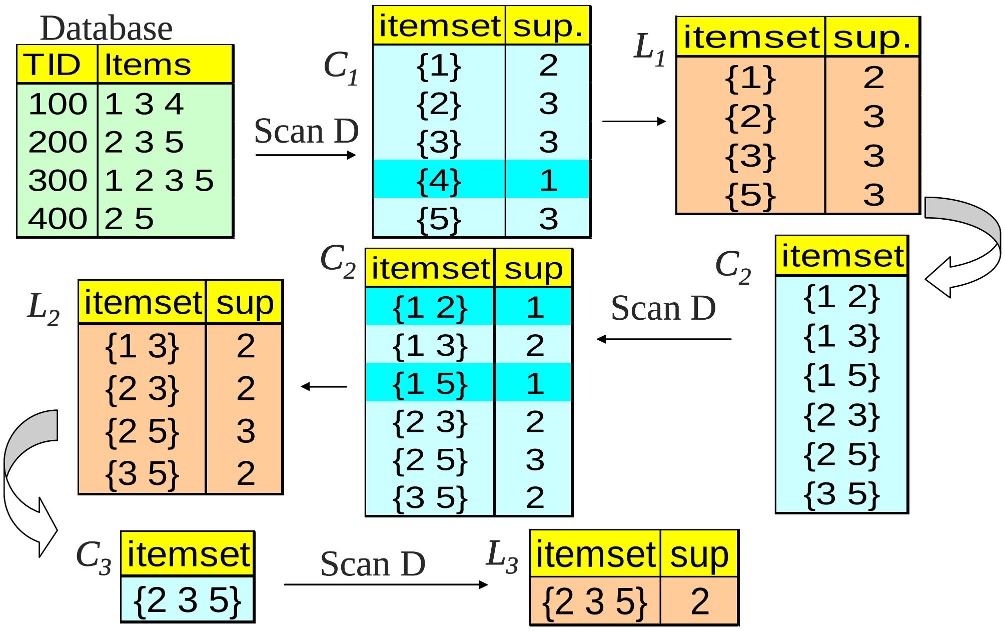
Web mining typically addresses semi-structured or unstructured data, like Web and log files with mixed knowledge involving multimedia, flow data, etc., often represented by imprecise or incomplete information. This implies that fuzzy set theory approaches are useful instruments in order to mine knowledge from such data[3]. It can be further divided into web content mining and web usage mining. Web usage mining can discover the unknown and potentially useful knowledge from web usage data. The growth of web-based B2B electronic commerce has shown the necessity of web usage mining.

In summary, two important issues of mining association rules are examined in the paper. The first issue is the discovery of generalized fuzzy association rules in the transaction database. The second issue is to discover association rules from the web usage data and the large itemsets identified in the transaction database. We propose a cluster-based mining architecture to address the two problems. The characteristics of the proposed architecture are discussed in detail in the paper. An example is given to demonstrate the effectiveness and feasibility of the proposed approach.

This paper is organized as follows. In chapter 2, we review relevant literature related to our study. In chapter 3, the proposed cluster-based mining architecture is explained. In chapter 4, illustrates the proposed approach with an example. Chapter 5 evaluated and discussed the performance with our mining method. Finally, conclusions and future works are summarized in chapter 6.

Chapter 2 Literature review

The early data mining method for association rules is the support-confidence framework. Agrawal and Srikant proposed a method to discover large itemsets and construct association rules named Apriori algorithm [16]. The Apriori algorithm requires to repeatedly scanning the database and generated the potentially large itemsets, called candidate itemsets by using the large itemsets found in the previous pass. Compute their supports during the pass over the data. For example, at the (k-1)th iteration, all large itemsets containing k-1 items, called large (k-1)-itemsets, are generated. In the next iteration, the candidate itemsets containing k items are generated by joining large (k-1)-itemsets. An example is shown in Figure 2-1. When all of the large itemsets generated, compute their confidence to construct the association rules.



Database

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

L_3

itemset	sup
{2 3 5}	2

Figure 2-1: An example of large itemsets generation of Apriori algorithm

After Agrawal et al. proposed the Apriori association rule algorithm [16]. Several researches focus on effectively reducing the number of database scans, however, it's still wasted scanning infrequent candidate itemsets. Park et al. proposed an effective algorithm DHP (direct hashing and pruning) for the initial candidate set generation [9]. This method efficiently controls the number of candidate 2-itemsets. Brin et al. proposed the dynamic itemset count (DIC) algorithm for finding large itemsets [20]. H. Toivonen proposed the sampling algorithm [6]. Tsay et al. have used cluster-based association rule (CBAR) approach [26]. This method used cluster-based table to reduce the number of database scan and requiring less contrast. These algorithms are to reduce the time of database scan and data contrast as shown in Figure 2-2.

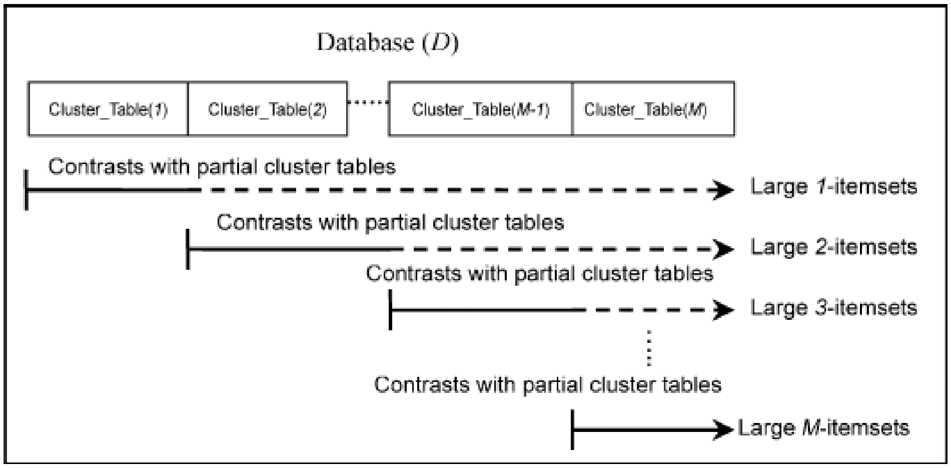

Database (D)
Cluster_Table(1)
Cluster_Table(2)
Cluster_Table(M-1)
Cluster_Table(M)
Contrasts with partial cluster tables
Large 1-itemsets
Contrasts with partial cluster tables
Large 2-itemsets
Contrasts with partial cluster tables
Large 3-itemsets
Contrasts with partial cluster tables
Large M-itemsets

Figure 2-2: Illustrate database scans and contrasts of CBAR algorithm[26]

Previous studies on data mining focused on finding association rules on the single-concept level hierarchical relationships. However, data taxonomies are usually predefined in real world applications and can be represented using hierarchy trees. Terminal nodes on the trees represent actual items appearing in transactions; internal nodes represent classes or concepts formed by lower-level nodes [17]. An example is given in

Figure 2-3.

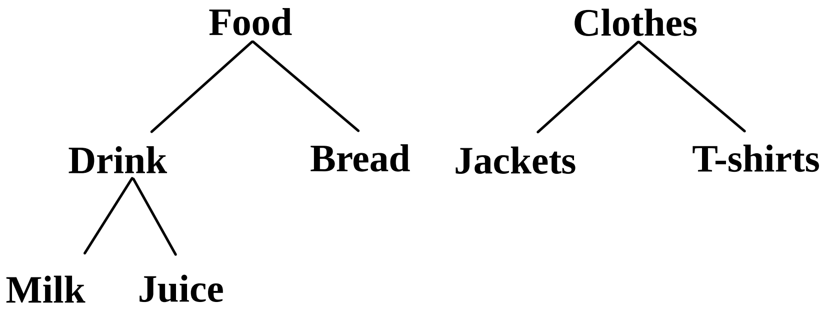


Figure 2-3: An example of taxonomic structures

In this example, the food falls into two classes: drink and bread. Drink can be further classified into milk and juice. Similarly, assume clothes are divided into jackets and T-shirts. In the transaction database only the terminal items (milk, juice, bread, jacket and T-shirt) can appear in transactions.

Agrawal and Srikant proposed a method for finding generalized association rules at multiple levels [17]. Their mining process can be divided into four main phases. In the first phase, ancestors of items in the given transactions are added according to the predefined taxonomy. In the second phase, candidate itemsets are generated and counted by scanning the expanded transaction data, then using the minimum support to discover all large itemsets. This process is repeated until all large itemsets has been discovered. In the third phase, all possible association rules are found from the large itemsets, and calculated confidence values. In the fourth phase, extract all the generalized association rules by using the minimum confidence.

Previous researches on mining association rules focus on discovering the relationship among items in the transaction database. Another important issue of the association rules mining is to find the relationships from different databases. Tsai et al. have proposed a method for mining association rules from the customer database and the transaction database. This method can be decomposed into two sub-problems. First, all large itemsets are discovered from the transaction database. Second, a relationship graph and graph-based algorithm are used to discover the association rules from the customer database and the transaction database.

Fuzzy set theory was first proposed by Zadeh and Goguen in 1965[10]. Fuzzy set theory is primarily concerned with quantifying and reasoning using natural language in which words can have ambiguous meanings. This can be thought of as an extension of traditional crisp sets, in which each element must either be in or not in a set[24].

Formally, the process by which individuals from a universal set X are determined to be either members or non-members of a crisp set can be defined by a characteristic or discrimination function[10]. For a given crisp set A, this function assigns a value $\mu_A(\mathrm{x})$ to every $\mathrm{x} \in X$ such that

$$\mu_A(\mathrm{x}) = \begin{cases} 1 & \text{if and only if } \mathrm{x} \in A \\ 0 & \text{if and only if } \mathrm{x} \notin X \end{cases}$$

Thus, the function maps elements of the universal set to the set containing 0 and 1. This kind of function can be generalized such that the values assigned to the elements of the universal set fall within specified ranges, referred to as the membership grades of these elements in the set. Larger values denote higher degrees of set membership. Such a function is called the membership function, $\mu_A(\mathrm{x})$, by which a fuzzy set A is usually defined. This function is represented by $\mu_A : \mathrm{X} \rightarrow [0, 1]$ where [0, 1] denotes the interval of real numbers from 0 to 1 and function can also be generalized to any real interval instead of [0, 1]. The triangle fuzzy membership function is used frequently and efficiently. An example of triangle fuzzy membership is given in Figure 2-4.

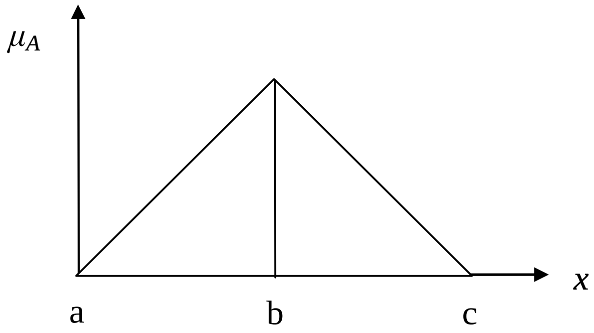


Figure 2-4: An example of triangle fuzzy membership function

Assume that x_1 to x_n are the elements in fuzzy set *A*, and μ_1 to μ_n are, respectively, their grades of membership in *A*. *A* is then usually represented as follows:

$$A = \mu_1 / x_1 + \mu_2 / x_2 + \ldots + \mu_n / x_{n.}$$

An α-cut of a fuzzy set A is a crisp set A_α that contains all elements in the universal set *X* with membership grades in *A* greater than or equal to a specified value of α. This definition can be written as

$$A_\alpha = \{ x \in X \mid \mu_A(x) \geq \alpha \}.$$

The scalar cardinality of a fuzzy set *A* defined on a finite universal set *X* is the summation of the membership grades of all the elements of *X* in A. Thus,

$$|A| = \sum_{x \in X} \mu_A(x). \qquad (2\text{-}1)$$

Among operations on fuzzy sets are the basic and commonly used union and intersection, as proposed by Zadeh.

(1) The union of fuzzy sets A and B is denoted by A∪B, and the membership function of A∪B is given by

$\mu_{A\cup B}(\mathrm{x}) = \max\{\mu_A(\mathrm{x}), \mu_B(\mathrm{x})\}, \quad \forall_{\mathrm{x}} \in X.$ (2-2)

(2) The intersection of two fuzzy sets A and B is denoted by A∩B, and the membership function of A∩B is given by

$\mu_{A\cap B}(\mathrm{x}) = \min\{\mu_A(\mathrm{x}), \mu_B(\mathrm{x})\}, \quad \forall_{\mathrm{x}} \in X.$ (2-3)

The applications of fuzzy theory, “Fuzzy if-then rules” is important concept of fuzzy partition. Ishibuchi proposed the simple fuzzy partition and multiple fuzzy partition concepts[4, 5]. In this paper, we focus on simple fuzzy partition issue and as shown in Figure 2-5. Its defined as follow.

$$m_i^D(x) = \begin{cases} 0, x < a_i \\ (x-a_i)/(b_i-a_i), a_i \le x \le b_i \\ \\ (c_i - x)/(c_i - b_i), b_i \le x \le c_i \\ 0, x > c_i \end{cases} \qquad 1 \le i \le K$$

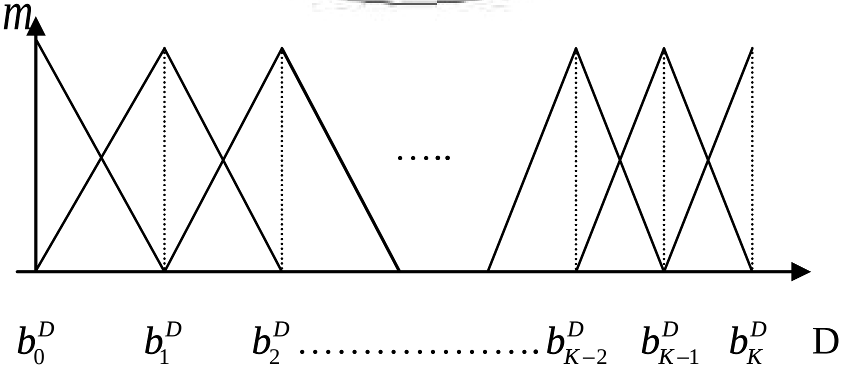


Figure 2-5: $\mathbf{i_{th}}$ fuzzy sets of fuzzy membership function (K : The number of fuzzy sets of dimension D)

We can compute the range distance of (a, b, c), that $x_{\max}^{D}$ is the maximum values and $x_{\min}^{D}$ is the minimum values of dimension D[22].

$$S = \frac{(x_{\max}^{D} - x_{\min}^{D})}{(K-1)} \qquad \begin{aligned} b_i &= x_{\max}^{D} + S \cdot (i-1) \\ a_i &= b_i - S \\ c_i &= b_i + S \end{aligned} \tag{2-4}$$

Fuzzy logic has employed fuzzy sets to represent knowledge that is vague or imprecise [10]. Recently, fuzzy logic used more and more frequently. The fuzzy association rules can extract more generalized and useful knowledge. Hong et al. proposed the fuzzy data mining method for mining interesting generalized association rules [25].

The Web mining can be broadly categorized as Web Content Mining, Web Structure Mining, and Web Usage Mining[3] and shown in Figure 2-6.

(1)Web Content Mining of multimedia documents, involving text, hypertext, images, audio and video information deals with the extraction of concept hierarchies/relations from the Web, and their automatic categorization.

(2)Web Structure Mining of inter-document links, provided as a graph of links in a site or between sites.

(3)Web Usage Mining of the data generated by the user’s interactions with the Web, typically represented as Web server access logs, user profiles, user queries and mouse-clicks. This includes trend analysis and Web access association/sequential pattern analysis.

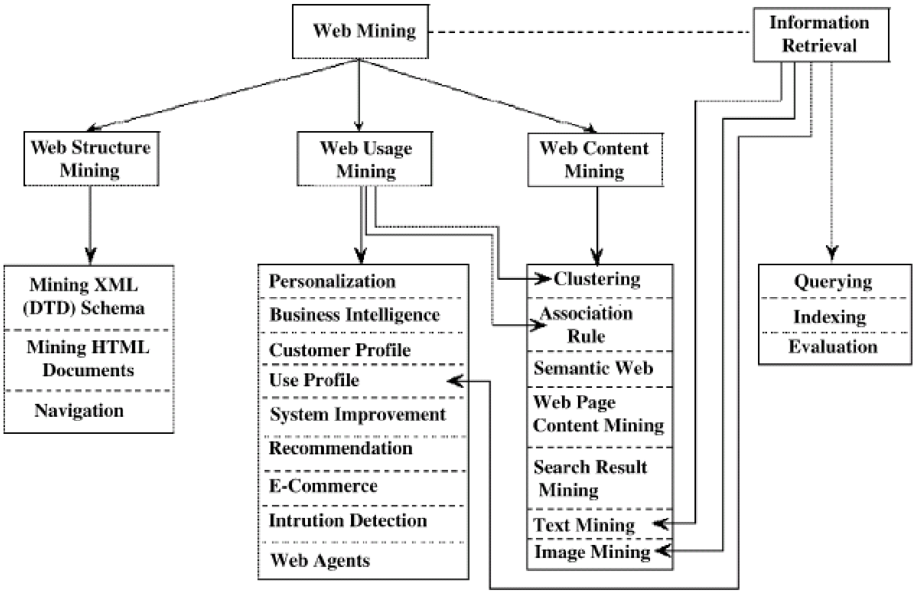


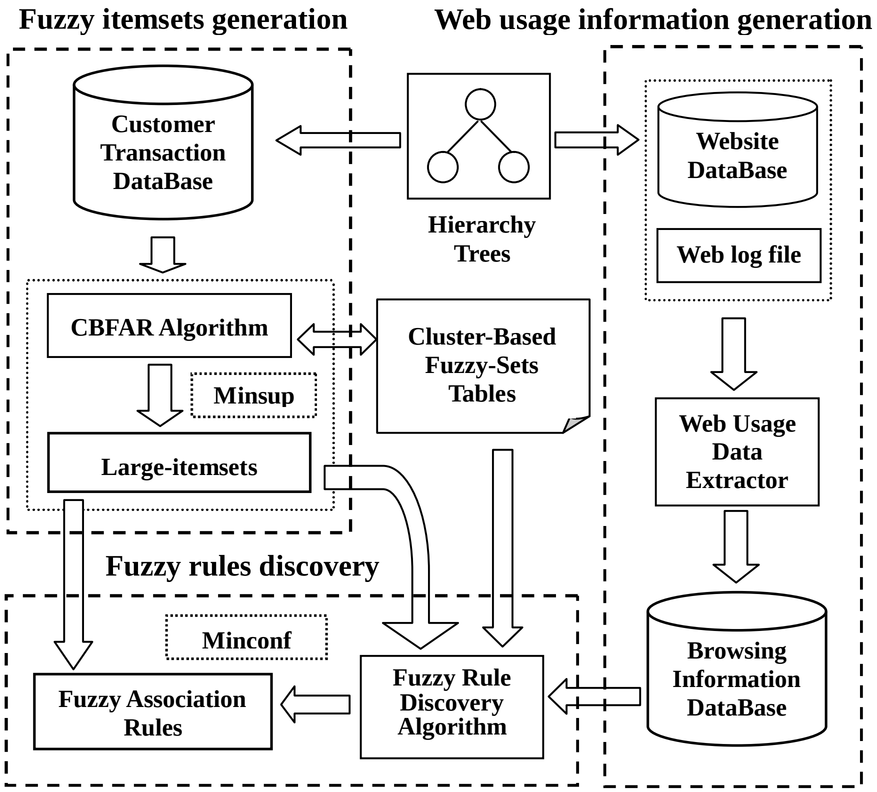
Figure 2-6: A Web mining taxonomy[3]

All the researches have indicated that the times of database scan and data contrast are the important factors to the efficiency of the mining process. Using the multi-concept level hierarchical relationships to discover the generalized fuzzy association rules is also an important data-mining task. Furthermore, in the e-commerce era, discovering the potential and unknown knowledge from the Web usage databases and transaction databases has recently brought to light. All the issues indicated above will be addressed in this paper.

Chapter 3 Cluster-Based Mining Architecture

In this paper, we proposed a cluster-based architecture to generate fuzzy association rules from web usage database and transaction database. The proposed framework is shown in Figure 3-1.

The cluster-based mining framework is decomposed into three subproblems: (1) Using cluster-based fuzzy association rule algorithm (CBFAR) to identify all the large fuzzy itemsets from the transaction database, (2) Using browsing information database (BIDB) to store web usage data from World Wide Web, and (3) Finally, given predefined minimum confidences, the fuzzy association rules among items are generated from large fuzzy itemsets, and relationships from large fuzzy itemsets and web usage data are discovered with a fuzzy rule discovery method.



Fuzzy itemsets generation
Web usage information generation
Customer Transaction DataBase
Hierarchy Trees
Website DataBase
Web log file
CBFAR Algorithm
Cluster-Based Fuzzy-Sets Tables
Minsup
Large-itemsets
Web Usage Data Extractor
Fuzzy rules discovery
Minconf
Fuzzy Association Rules
Fuzzy Rule Discovery Algorithm
Browsing Information DataBase

Figure 3-1: Cluster-Based Fuzzy Rules Mining Framework

3.1 Large fuzzy itemsets generation

We propose a CBFAR mining algorithm for discovering generalized fuzzy association rules from transaction database based on the hierarchical relationships and cluster-based fuzzy-sets tables. The proposed algorithm, as shown in Figure 3-2 is divided into three phases as described below.

(1) In the first phase, we scan the database to create cluster-based fuzzy-sets tables. At first, the ancestors of items in each given transaction are added according to the predefined taxonomy. Then, transform the quantitative value v_{ij} of each transaction data D_i (i=1 to n), for each expanded item name I_j appearing into a fuzzy set f_{ij}. The f_{ij} are represented as ($f_{ij1}/R_{j1} + f_{ij2}/R_{j2} + \mathbf{L} + f_{ij1}/R_{j1}$) using the given membership functions, where h is the number of fuzzy regions for I_j . R_{jl} is the lth fuzzy region of I_j, 1≤l≤h, and f_{ijl} is v_{ij}'s fuzzy membership value in region R_{jl}. Calculate the summation of each fuzzy region R_{jl} in the transaction data. Finally, if the length of transaction record is k, the transaction record and the fuzzy region value of items in this transaction will be stored in the table, named cluster-based fuzzy-sets table (k), 1≤k≤M, where M is the length of the longest transaction record in database.

(2) In the second phase, the set of candidate itemsets C_n is generated by the self-join of L_{n-1}. When the length of candidate itemset is k, the support is calculated with reference to the cluster-based fuzzy-set table(k). If the fuzzy region value of c_n is greater than or equal to the predefined minimum support value a , the candidate itemsets becomes the large itemsets, put c_n in the large

itemsets L_n. Otherwise, it is contrasted with the cluster-based fuzzy-sets table(k+1). The contrast process terminates when the calculated support is greater than or equal to the predefined minimum support or the end of the cluster-based fuzzy-sets table(M) has been reached.

(3) In the third phase, we use the predefined minimum confidence value to discover fuzzy association rules. If the candidate fuzzy association rule is larger than or equal to the predefined confidence value, put it in the rule base.

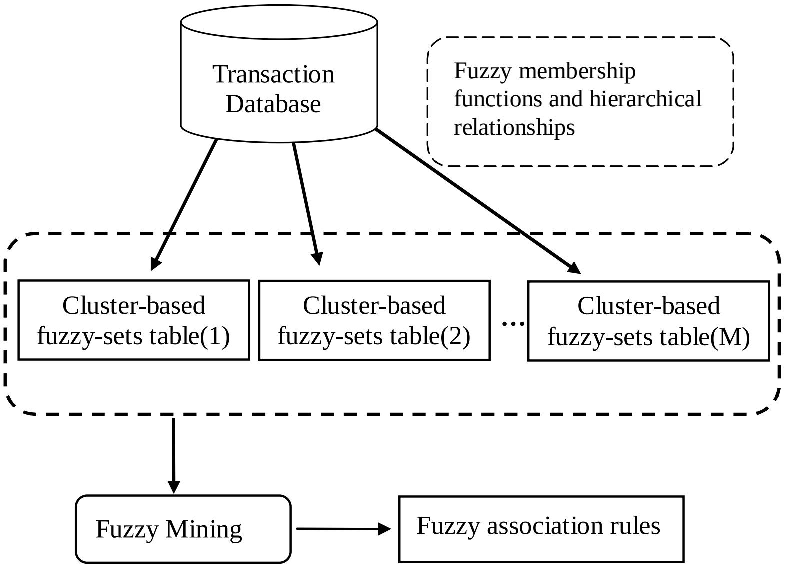

Transaction Database
Fuzzy membership functions and hierarchical relationships
Cluster-based fuzzy-sets table(1)
Cluster-based fuzzy-sets table(2)
...
Cluster-based fuzzy-sets table(M)
Fuzzy Mining
Fuzzy association rules

Figure 3-2: CBFAR Mining Approach

3.2 Web usage information generation

There is significantly potential information that is worth to be explored in the e-commerce environment. When users click the web pages or browsing the content in the web pages, their interest behaviors have been revealed. Therefore, we regard the clicks information of user navigations as the explored web usage data in this study. Moreover, fuzzy membership functions are employed to qualitatively represent the clicks frequency.

We assume that the web page for each item is organized and stored in the website. The information about the users who have visited a web site is recorded in the related web server log files. The web server can also record every click that users make on the web site. The web usage information generation method is described as follow.

(1) Predefined taxonomy trees and fuzzy membership functions are given first.

(2) The web log data is preprocessed and transformed into the form suitable for our work.

(3) Each item ancestor in the taxonomy tree is regarded as a class. All the users clicking the same item webpage are extracted from the transformed data and grouped together as a set. At the same time, the total clicking number (frequency) for each class is counted.

(4) Using the given fuzzy membership function, the membership grade of the frequency of each class belonging to every fuzzy set is computed, and the class is extended as fuzzy classes with fuzzy terms.

(5) All the information generated above are stored in the browsing information database (BIDB), by which the fuzzy rules between fuzzy classes and large fuzzy itemsets can be extracted easily.

3.3 Fuzzy rule discovery

The fuzzy rule discovery mining framework is decomposed into two ways: (1) using the large fuzzy itemsets which are generated with CBFAR method and predefined minimum confidence to discover the generalized fuzzy association rules from transaction database, and (2) discovering the relationships from large fuzzy itemsets and large fuzzy frequency classes by using a fuzzy rule discovery method, then transforming the relationships into association rules.

The two kinds of fuzzy association rules can provide much more flexible and effective information, by which the managers can make correct business decisions for marketing strategies. The fuzzy-based rule discovery method is divided into four steps.

(1) Given a minimum threshold for frequency membership grade, all large fuzzy frequency classes are filtered out from the BIDB. The set of customers for a large fuzzy frequency class Ui is represented as Ui-CIDsets.

(2) Given a minimum threshold, we scan the cluster-based fuzzy-sets table to filter out the set of customers for each large fuzzy itemsets Lj, which is defined as Lj-CIDsets.

(3) A relation matrix R is created with all pairs of Ui and Lj. Each entry R(Ui, Lj) in the relation matrix stores the confidence value for Ui and Lj. If the (Ui-CIDsets) – (Lj-CIDsets) =*f*, then the confidence = 1; otherwise, the confidence = n(Lj-CIDsets) / n(Ui-CIDsets), where n(s) is the cardinality of a set s.

(4) If the confidence of a pair (Ui, Lj) is greater than or equal to the user predefined minimum confidence, the relationship is transformed into a fuzzy rule in the form of “IF users browse Ui THEN buy Lj frequently.”

Chapter 4 An Example

4.1 Large fuzzy itemsets generation

In this section, an example is given to illustrate the large fuzzy itemsets generation process. This is a simple example to show how the proposed method can be used to discover large fuzzy itemsets from transaction data. There are six transactions and five items in a transaction database: A, B, C, D and E. An example transaction database is shown in table 4-1. The taxonomy tree is shown in Figure 4-1. The ancestors of appearing items are added to transactions according to the predefined taxonomy tree.

Table 4-1: Six transactions in this example

CID	TID	Items
1	T1	(A,3) (C,4) (E,2)
2	T2	(B,3) (C,7) (D,7)
3	T3	(A,4) (B,2) (E,5)
4	T4	(C,9) (E,10)
5	T5	(B,3)
6	T6	(B,8)(D,4)

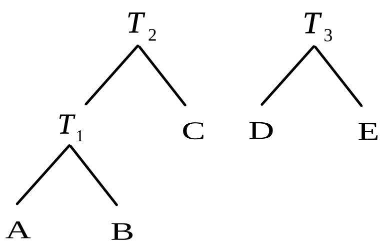


Figure 4-1: Taxonomy tree in this example

In this example, assume that the fuzzy membership functions are the same for all the items and are as shown in Figure 4-2. The fuzzy membership function is represented by three regions: ***Low(L), Middle(M) and High(H),*** and three fuzzy membership values are produced for each item according to the predefined membership function.

The length of the longest expanded transaction record in this database is six, and creates six cluster-based fuzzy-sets tables as shown in table 4-2. The fuzzy region value of items in this transaction will be stored in the cluster-based fuzzy-sets tables.

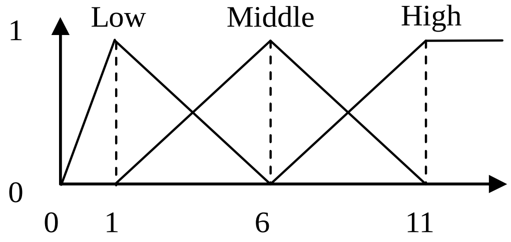


Figure 4-2: The membership function in this example

Assume the minimum support value is 2.0. We can discover the Large-1 itemsets (L_1) which is large than or equal to the predefined minimum support value according to the cluster-based fuzzy-sets tables. The itemsets of L_1 are {(B.Low) = 2.0, (T_1.Middle) = 2.8, (T_3.Middle) = 2.6}.

Next, we try to discover the large 2-itemsets L_2. Combining the items of L_1 in order to generate candidate 2-itemsets C_2. The procedure is similar to the candidate generation of Apriori algorithm [16]. The itemsets of C_2 are {(B.Low, T_1 .Middle), (B.Low, T_3 .Middle), (T_1.Middle, T_3.Middle)}. In order to generate L_2, it is necessary to compute the fuzzy region values of each candidate itemset, c, from cluster-based fuzzy-sets table(2) to cluster-based fuzzy-sets table(6). If the value is larger than or equal to the predefined minimum support value, put c in the L_2. Otherwise, compute the fuzzy region values in the next cluster table (cluster-based fuzzy-sets table(3)). The contrast process continues until the calculated support is greater than

or equal to the predefined minimum support or the end of the cluster-based fuzzy-sets table(M) has been reached. For example, the fuzzy support value of (T_1.Middle, T_3.Middle) = $\sum_2^6$ (T_1 .Middle $\cap$ T_3 .Middle) = 0+0+0+0.6+0.8+0.2+0.4 = 2.0. The itemsets of L_2 is {(T_1.Middle, T_3.Middle) = 2.0}. The other large

itemsets L_n are discovered in the similar way. Based on the process described above, the large itemsets in this example are (B.Low), (T_1.Middle), (T_3.Middle), and (T_1.Middle, T_3.Middle).

Table 4-2: cluster-based fuzzy-sets tables

TID	A	B	C	D	E	T_1	T_2	T_3

Cluster-based fuzzy-sets table(1)

NULL

Cluster-based fuzzy-sets table (2)

T5	0	L,0.6 M,0.4	0	0	0	L,0.6 M,0.4	0	0

Cluster-based fuzzy-sets table (3)

NULL

Cluster-based fuzzy-sets table (4)

T4	0	0	M,0.4 H,0.6	0	M,0.2 H,0.8	0	M,0.4 H,0.6	M,0.2 H,0.8
T6	0	M,0.6 H,0.4	0	L,0.4 M,0.6	0	M,0.6 H,0.4	0	L,0.4 M,0.6

Cluster-based fuzzy-sets table (5)

T3	L,0.4 M,0.6	L,0.8 M,0.2	0	0	L,0.2 M,0.8	M,1	0	L,0.2 M,0.8

Cluster-based fuzzy-sets table (6)

T1	L,0.6 M,0.4	0	L,0.4 M,0.6	0	L,0.8 M,0.2	L,0.6 M,0.4	L,0.4 M,0.6	L,0.8 M,0.2
T2	0	L,0.6 M,0.4	M,0.8 H,0.2	M,0.8 H,0.2	0	L,0.6 M,0.4	M,0.8 H,0.2	M,0.8 H,0.2

4.2. Web usage information generation

In this section, an example is given to illustrate the web usage information generation process. This is a simple example to show how the proposed method can be used to reserve web usage information from the World Wide Web. There are five items in items database: A, B, C, D and E, as shown in table 4-3. The taxonomy tree is shown in Fig. 4. The corresponding ancestor classes are shown in table 4-4.

Table 4-3: Items used in this example

No.	Item
1	A
2	B
3	C
4	D
5	E

Table 4-4: Ancestor classes

Classifications	Items
T_1	A, B
T_2	A, B, C
T_3	D, E

Ancestor classes with frequencies and the sets of customer IDs are given in table 4-5. We use the fuzzy membership function as shown in Figure 4-3 to calculate the membership grades of frequencies belonging to respective fuzzy set. The final Browsing information database is shown in table 4-6.

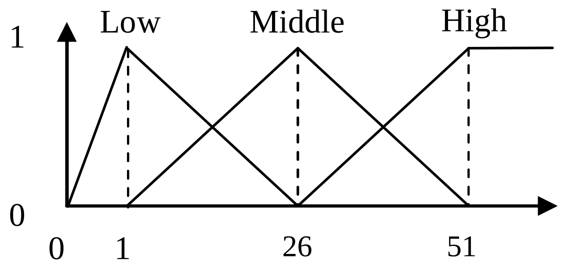


Figure 4-3: The membership function in this example

Table 4-5: Ancestor classes with frequencies and CIDsets

Classes	Frequency	Ui-CIDsets
T_1	25	{1, 2, 4, 5}
T_2	31	{3, 5}
T_3	16	{1, 4, 7, 8}

Table 4-6: Browsing information database

Fuzzy frequency classes	Fuzzy membership grades for frequencies	Ui-CIDsets
T_1.Low	0.04	{1, 2, 4, 5}
T_1.Middle	0.96	{1, 2, 4, 5}
T_2.Middle	0.8	{3, 5}
T_2.High	0.2	{3, 5}
T_3.Low	0.4	{1, 4, 7, 8}
T_3.Middle	0.6	{1, 4, 7, 8}

4.3 Fuzzy Rule discovery

At first, we use large fuzzy itemsets, which are generated with cluster-based fuzzy association rule mining method, and predefined minimum support to discover the generalized fuzzy association rules. The large fuzzy itemsets are {B.Low}, {T_1.Middle}, {T_3.Middle}, {T_1.Middle, T_3.Middle} which are generated in section 4.1. Assume the minimum confidence value is 2.0. The discovered rule is “IF users buy item T_1 with middle degree THEN buy item T_3 with middle degree frequently”.

Next, we can discovering the relationships from large fuzzy itemsets and large fuzzy frequency classes with fuzzy rule discovery method, then transform the relationships into fuzzy rules. Assume the minimum support value is 0.6. The large fuzzy frequency class are {(T_1.Middle) =

0.96, (T_2.Middle) = 0.8, (T_3.Middle) = 0.6}, as shown in table 4-7. The large fuzzy itemsets are {B.Low}, {T_1.Middle}, {T_3.Middle}, {T_1.Middle, T_3.Middle} which are generated in section 4.1. Let the minimum threshold be 0.6, we scan the cluster-based fuzzy-sets table to generate the CIDsets of large fuzzy itemsets as shown in table 4-8. Then, we construct the relation matrix as shown in table 4-9.

Table 4-7: Large fuzzy frequency classes

Fuzzy frequency classes	Fuzzy membership grades for frequencies	Ui-CIDsets
T_1.Middle	0.96	{1, 2, 4, 5}
T_2.Middle	0.8	{3, 5}
T_3.Middle	0.6	{1, 4, 7, 8}

Assume the minimum confidence value is 0.5. Then, {Large fuzzy Frequency Class T_1.Middle, Large Fuzzy Itemset B.Low}, {Large fuzzy Frequency Class T_2.Middle, Large Fuzzy Itemset B.Low}, {Large fuzzy Frequency Class T_2.Middle, Large Fuzzy Itemset T_1.Middle}, {Large fuzzy Frequency Class T_2.Middle, Large Fuzzy Itemset T_3.Middle}, and {Large fuzzy Frequency Class T_2.Middle, Large Fuzzy Itemset (T_1.Middle, T_3.Middle)} are the significant relations. The discovered rules are represented as “IF users browse class T_1.Middle THEN buy item B.Low frequently”, “IF users browse class T_2.Middle THEN buy item B.Low frequently”, “IF users browse class T_2.Middle THEN buy item T_1.Middle frequently”, “IF users browse class T_2.Middle THEN buy item T_3.Middle frequently”, and “IF users browse class T_2.Middle THEN buy item (T_1.Middle, T_3.Middle) frequently”

Table 4-8: The CIDsets of large fuzzy itemsets

Large fuzzy itemsets	Li-CIDsets	Counts
B.Low	{2, 3, 5}	3
T_1.Middle	{3, 6}	2
T_3.Middle	{6, 3, 2}	3
(T_1.Middle , T_3.Middle)	{3, 6}	2

Table 4-9: Relation matrix with confidences

(LFI: large fuzzy itemsets, LFFC: large fuzzy frequency classes)

LFI / LFFC	B.Low	T_1.Middle	T_3.Middle	(T_1.Middle , T_3.Middle)
T_1.Middle	$\frac{2}{4} = 0.5$	$\frac{0}{4} = 0$	$\frac{1}{4} = 0.25$	$\frac{0}{4} = 0$
T_2.Middle	$\frac{2}{2} = 1$	$\frac{1}{2} = 0.5$	$\frac{1}{2} = 0.5$	$\frac{1}{2} = 0.5$
T_3.Middle	$\frac{0}{4} = 0$	$\frac{0}{4} = 0$	$\frac{0}{4} = 0$	$\frac{0}{4} = 0$

Chapter 5 Experimental Results

To evaluate the efficiency of the proposed method, we have implemented the CBFAR, along with a known Apriori-based fuzzy association rule algorithm (ABFAR)[12], using Microsoft Visual Basic 6.0 on a Pentium IV 1600 MHz PC with 1GB of available physical memory. Experiments have been carried out on testing the spending time of large itemsets generation using the FoodMart transaction database provided with Microsoft SQL Server.

The performance of CBFAR algorithm is compared to the ABFAR algorithm under various user specified minimum supports, such that 0.35, 0.4, 0.45, 0.5, 0.55%. The results are shown in Figure 5-1. The experimental results show that the CBFAR algorithm has better performances than ABFAR. When there is an decrease in the values of minimum supports, the performance gap between the algorithms is shown in greater evidence.

2,000, 4,000, 6,000, 8,000, 10,000 transaction records of experimental data are randomly sampled from FoodMart transaction database. The performance of CBFAR algorithm is compared to the ABFAR algorithm where minimum support is set at 0.45%. The results are shown in Figure 5-2. From the figure, we can see that the CBFAR outperforms the ABFAR. It is noted that the superiority becomes much more obvious along with increase in the number of transaction records.

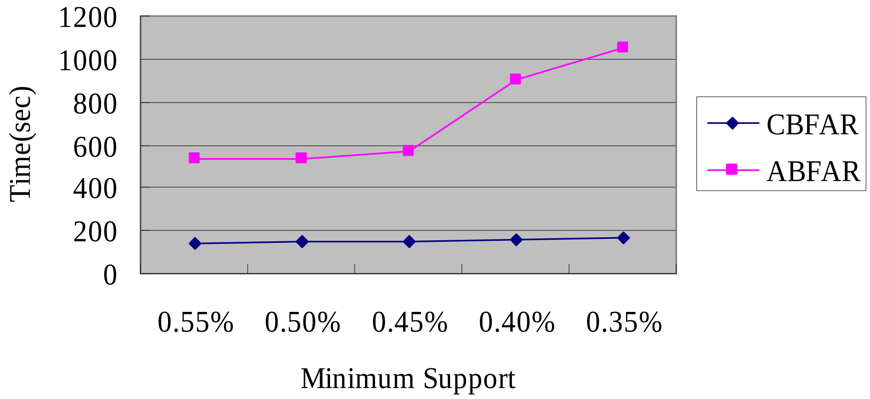


Figure 5-1: Performance of CBFAR and ABFAR on 4,000 transaction records

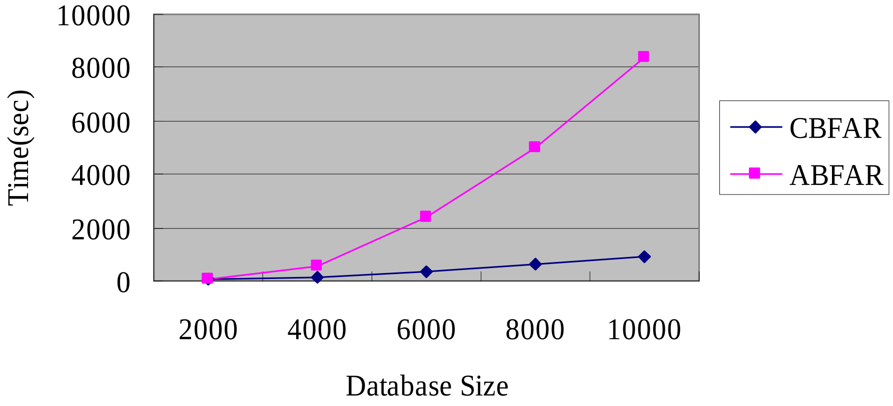


Figure 5-2: Performance of CBFAR and ABFAR at minimum support 0.45%

Chapter 6 Conclusions and Future Works

The relationships of the transaction database and browsing information database are very important. There is a lot of potential information in the relationships. The managers can make correct business decisions for marketing strategies. In this paper, we have proposed a cluster-based association rules mining architecture for extracting generalized fuzzy association rules from transaction database and browsing information database.

At first, we employ the cluster-based fuzzy association rule (CBFAR) method to discover the fuzzy association rules from the transaction database. CBFAR only requires a single scan of the transaction database to create cluster-based fuzzy-sets tables in advance. Contrasts are then performed only against the partial cluster tables to generate all the large fuzzy itemsets. The mining performance is effectively improved due to the method needs only one database scan and also requires less contrast. Next, the fuzzy-based rule discovery method is employed to discover the relationships between the identified large fuzzy itemsets and the interesting web usage data.

Experiments show that the efficiency of CBFAR is greater than the existing ABFAR algorithm. When there is an increase in the number of transaction records and decrease in the minimum supports, the performance between the algorithms becomes more evident.

In the future, we will consider the extension of the CBFAR algorithm to incremental mining of generalized fuzzy association rules from real-time databases. Besides, it is unrealistic that the most approximate fuzzy sets can always be provided in advance. We will attempt to develop an automatic method to find the most suitable fuzzy sets that cover the domains of quantitative data for fuzzy association rules mining. Furthermore, the practical applications of the proposed architecture will be also explored, for example, the hidden CRM knowledge between the RFM data and the large itemsets is worth to be discovered for making right marketing strategies.

References

[1] A. Kandel, Fuzzy Expert Systems, CRC Press, Boca Raton, FL, 1992 pp.8-19.

[2] C.C.Chang, Y.C.Li, and J.S.Lee, ”An Efficient Algorithm for Incremental Mining of Association Rules”, *Proceedings of 2005 RIDE-SDMA IEEE International Workshop on Research Issues in Data Engineering: Stream Data Mining and Application*, 2005 pp.3-10.

[3] D. Arotaritei and S Mitra, “Web mining: a survey in the fuzzy framework”, *Fuzzy sets and systems*, 2004 pp.5-19.

[4] H. Ishibuchi, K. Nozaki, and H. Tanaka, “Distributed representation of fuzzy rules and its application to pattern classification”, *Journal of Fuzzy Sets and Systems*, Vol.52(1), 1992 pp.21-32.

[5] H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka, “Selecting fuzzy if-then rules for classification problems using genetic algorithms”, *IEEE Transaction on Fuzzy Systems*, Vol.3(3), 1995 pp.260-270.

[6] H. Toivonen, “Sampling Large Database for association Rule”, *VLDB*, 1996 pp.134-145.

[7] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, Los Altos, 2000.

[8] J. Han, Y. Fu, “Discovery of multiple-level association rules from large database”, *The International Conference on Very Large Databases*, 1995 pp.420-431.

[9] J.S. Park, M. S. Chen, and P. S. Yu, “An Effective Hash Based Algorithm for Mining Association Rules”, *Proceedings of ACM SIGMOD*, 1995 pp.175-186.

[10] L.A. Zadeh, “Fuzzy sets”, *Information and Control*, 1965 pp.338-353.

[11] M.S. Chen, J. Han, P. S. Yu, “Data mining: an overview from a database perspective”, *IEEE Transactions on Knowledge and Data Engineering*, 1996 pp.866-883.

[12] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, Discovering data mining from concept to implementation, Prentice-Hall, Englewood Cliffs, 1997.

[13]P.S.M. Tsai, C.M. Chen, “Mining interesting association rules from customer databases and transaction databases”, *Information Systems,* 2004 pp.685-696.

[14]R.Agrawal, T. Imielinksi, A. Swami, “Mining association rules between sets of items in large database” *The 1993 ACM SIGMOD Conference*, Washington, DC, USA, 1993 pp.1-10.

[15]R.Agrawal, T. Imielinksi, A. Swami, “Database mining: a performance perspective”, *IEEE Transactions on Knowledge Data Engineering,* 1993 pp.914-925.

[16]R.Agrawal, R. Srikant, “Fast algorithm for mining association rules in large databases”, *Proceedings of 1994 International Conference on VLDB*, 1994 pp.487-499.

[17]R.Agrawal, R. Srikant, “Mining generalized association rules”, *The International Conference on Very Large Databases,* 1995 pp.407-419.

[18]R.Agrawal, R. Srikant, “Mining quantitative association rules in large relational tables”, *The 1996 ACM SIGMOD International Conference on Management of Data*, Monreal, Canada, June 1996 pp.1-12.

[19]R.Cooley, B. Mobasher and J. Srivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web”, *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICYAI’97)*, November, 1997 pp.558-567.

[20]S.Brin,R. Motwani, J.D. Ullman, and S. Tsur, “Dynamic Itemset Counting and Implication Rules for Marketing Basket Data”, *1997 ACM SIGMOD Conference on Management of Data*, 1997, pp.255-264

[21]S.L.Wang, and A Jafari, “Hiding Sensitive Predictive Association Rules”, *2005 IEEE International Conference on Systems, Man and Cybernetics*, 2005 pp.164-169.

[22]T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, “Mining optimized association rules for numeric attributes”, *The ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems,* 1996 pp.182-191.

[23]T.P. Hong, C.S. Kuo, S.C. Chi, “A data mining algorithm for transaction data with quantitative values”, *Intelligence Data Analysis*, 1999 pp.363-376.

[24]T.P. Hong, C.S. Kuo, S.C. Chi, “Mining association rules from quantitative data”, *Intelligence Data Analysis*, 1999 pp.363-376

[25]T.P. Hong, K.Y. Lin, S.L. Wang, “Fuzzy data mining for interesting generalized association rules”, *Fuzzy sets and systems*, 2003 pp.255-269.

[26]Y-J Tsay, J-Y Chiang, “CBAR: an efficient method for mining association rules”, *Knowledge-Based Systems*, 2005 pp.99-105