Mining Rare Association Rules on Banpheo Hospital (Public Organization) via Apriori_MSG-P Algorithm

Taweechai Ouypornkochagorn[1], Non-member

ABSTRACT

Mining association rule is one of the important techniques in data mining to exploit hidden knowledge in large database. Many businesses in several areas need this technique for examine their enormous information, and public health is the one area that highly requires. Several hidden information conceal in daily operation data such as relation between visit time and symptom, relation between disease and patient age, etc. By the way, association rule discovery via traditional Apriori algorithm, the fundamental way to retrieve hidden rules, has to pay with tremendous resources and time. This research implements the modification of association rule mining technique, Apriori_MSG-P, in operational database of Banpheo hospital (Public Organization), Sumutsakon province, Thailand. The objectives target on epidemic information and patient behavior on hospital's services. The research's outcomes show that our implementation can evaluate a lot of valuable information that can be used by both of operation staffs and executive staffs. Moreover, the research's outcomes demonstrate that Apriori_MSG-P can be the proper one technique that can implement the real-world databases.

Keywords: Banpheo Hospital, Apriori_MSG-P, Multiple Supports, Association Rule, Data Mining Implementation

1. INTRODUCTION

Association rule mining is the one of the most popular techniques in data mining for discovering interesting patterns in database and this valuable information could be utilized and applied in several real-world applications. Apriori algorithm is the fundamental algorithm of the association rules discovery technique by using user minimum support threshold for selecting interesting patterns. Unfortunately, the process of Apriori method performs an exhaustive search for all possible itemsets that takes unacceptable time in real-world applications. Several researches proposed the ways to improve traditional Apriori algorithm. For example, by reducing the number of database scanning, Frequent-Pattern growth (FP Tree) [10] uses the tree concept which adopts a divide-and-conquer strategy (avoid to create candidate), PRICE [4] uses only once database scanning and using logical operations in the process. Furthermore, [1], [17], [21], [22] had been proposed the way to reduce database size with strategic planning and sampling. This leads to reducing the CPU computation time and disk access, but the qualification and the quality of sampling still are concerned. The other way, transaction reduction [14], [15] provide effective way in mining association rules by neglecting a transaction which does not contain any frequent k-itemsets in subsequent scans. Parallelization by partitioning approaches is the challenges for parallel data mining. FDM [5] and FSM [6] are parallelization of Apriori for mining association rules by its own partition which handled by separated machines. [3] was applied grid services for implementing high-performance distributed knowledge discovery systems that can use parallel and distributed mining system. Hashing technique is also the alternative effective way, Direct Hashing and Pruning (DHP) [11] uses a hashing technique to filter frequent 2-itemsets which are ineffective candidate and also avoids some database scanning. Hashing techniques have been applied for mining association rules in text database [9] and also over data streams [8].

In several applications, some interesting items appear very frequently in the data while others appear rarely. In the aspect of mining association rules with traditional Apriori with only one minimum support, when users set minimum support too high, those rules that involve with rare items will not be found. But when users set minimum support very low for retrieving rare itemsets, this may cause combinatorial explosion because those frequent items will be associated with one another in all possible ways. Users will receive the expected rare itemsets together with enormous number of unexpected itemsets. This dilemma is called the rare item problem. The interesting itemsets containing with some interesting item but those itemsets do not have support value enough to evaluate are called "rare itemset". Threshold improvements have been proposed. Multiple minimum supports approaches were proposed for evaluating rare itemsets in [2] (that called MSApriori), [7], [13], [16] but these still be the hard way to use in practical.

Manuscript received on September 1, 2011 ; revised on January 31, 2012.

[1] The author is with Faculty of Engineering at Si Racha Kasetsart University Si Racha Campus Chonbury, Thailand, E-mail: taweechai.ouy@gmail.com

In this research, we apply Apriori_MSG-P [18], which is the improvement way of multiple minimum supports approaches by only one specific user threshold for discovering rare itemsets, to real-world application. We choose hospital information system at Banpheo hospital (Public Organization), Sumutsakon province, Thailand as our database. Apriori_MSG-P proposed threshold bases on the statistic distribution that interpretable, easy to evaluate the desired outcome and able to find rare itemsets by examining the data characteristic of support values.

2. CHARACTERISTIC OF TRADITIONAL APRIORI'S RESULT

To show characteristic of traditional Apriori algorithm, let take a look on the sample database, Northwind database from http://www.microsoft.com. We take the attribute-oriented induction process following:

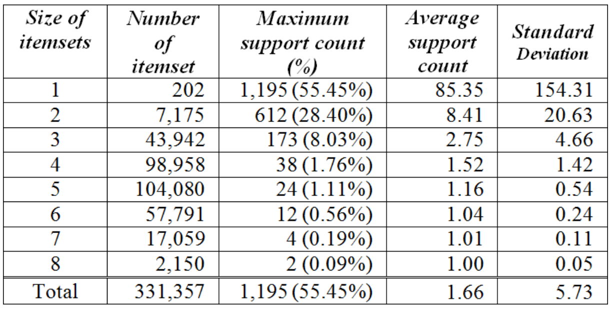
- Categorize the product's unit price into 3 ranges: high (> 50), medium ($20 - 50$) and low (< 20)
- Categorize the ordering quantity into 3 ranges: high (> 40), medium ($15 - 40$) and low (< 15)

Then, try to choose 8 attributes in the sale ordering subject on the trail: product, product's category, product's unit price, employee, ordering month, ordering quantity, customer's city and customer's country. The 202 items were found on 2,155 transactions. The results on running Apriori with the 1 transaction count minimum support show in table 1. Obviously, each size of itemsets has the different data characteristic. So, applying the single minimum support to all of itemsets may not suitable.

In case of choosing 1% minimum support (or 21.55 support count), this support will locate at percentile 33.96% of itemset size 1, 73.79% of itemset size 2 and 100% of the rest of upper sizes. This mean, the filtered itemsets with 1% minimum support will pass a lot of itemsets in itemset size 1 (passes 66.04%!), but it seems strictly screen in the upper sizes (only 26.21% in itemset size 2 and very few in the upper sizes).

In the other hand, when offering the 40% statistic percentile on each size, instead of using minimum support, it would be 2.15% minimum support in itemset size 1, 0.15% in size 2 and 0.07% in size 3 (omitted other sizes), that means 40% statistic percentile is able to generate multiple supports automatically depend on data characteristic. Additionally, with this statistic percentile, it can be used for expecting the number of result itemsets. The number of outcome itemsets of 40% percentile has the coverage about 100-40=60% in each size when the support distribution is symmetric curve. So, statistic percentile measurements should be used as threshold of interestingness justification instead of the single minimum support.

Table 1: Itemsets Charateristic with 1% min supp.



Size of itemsets	Number of itemset	Maximum support count (%)	Average support count	Standard Deviation
1	202	1,195 (55.45%)	85.35	154.31
2	7,175	612 (28.40%)	8.41	20.63
3	43,942	173 (8.03%)	2.75	4.66
4	98,958	38 (1.76%)	1.52	1.42
5	104,080	24 (1.11%)	1.16	0.54
6	57,791	12 (0.56%)	1.04	0.24
7	17,059	4 (0.19%)	1.01	0.11
8	2,150	2 (0.09%)	1.00	0.05
Total	331,357	1,195 (55.45%)	1.66	5.73

Turning to the aspect about the meaning of minimum support, minimum support seems to be unclear meaning for user to trail for the best outcomes if environment was changed. For example, firstly, you may need 3 from 10 transactions (30%) for minimum support, but when the transaction increases to 100 and supposes that the number of items increases in 2 times, how should the minimum support change? Remaining 3 or changing to 30 or another value? Other confusion occurs when it moves on another database. If the best result of database A set on 30% minimum support, it would unnecessarily set to 30% on database B even if they have same transaction size. So, minimum support cannot directly reflect to interestingness, especially when database is different or changed since minimum support doesn't analyze database characteristic first. Indeed, the meaning of interestingness should be measured from ranking or something else among their members.

3. HOSPITAL INFORMATION SYSTEM (HIS)

Hospital Information System (HIS) is the system designed for manage daily operation in hospital, both of front offices and back offices. The examples of important offices are shown as following:

- Medical record office
- Outpatient department (OPD)
- Inpatient department (IPD)
- Emergency Room (ER)
- Operating Room (OR)
- Labor Room (LR)
- Medicine (MED)
- Surgical (SUR)
- Orthopedic (ORTHO)
- Obstretic Gynecology (OB-GYN)
- Etc.

All information moves on system for multiple purposes such as curing or cashing purpose. Therefore, the enormous number of information is stored in hospital database. In addition, currently, HIS is implementing in standard. The well known standards, such as HL7 by American National Standards Institute (ANSI) or statXML by Thailand's National Information Center (NIC), are applied and are used for system integrating among hospitals or between hospital and government agency units.

Unfortunately, nowadays these information are used only in the small parts since they need a lot of times and human resources to exploit. A lot of knowledge still does not be used and some of knowledge is the thing that seems unclear but is accustomed and acceptable by long experience hospital's staffs which is the great obstacle for new staffs to understand it. These hidden information and unclear information are the high valuable knowledge for medical profession that may be found from symptom diagnosing results and patient's circumstance causes. So, a lot of knowledge still be laid and wait for someone to discover them.

Banpheo hospital (public organization) is the biggest local hospital in Sumutsakon province, Thailand which has 304,470 patients and 6,864,118 clinic visits in 2010. Even though many of executive information are prepared by HIS as executive or operative reports, the number of discovering information seems be in the small portion comparing with the number of data in database. In this research, we implement the one of our recent proposed technique, Apriori_MSG-P, to evaluate hidden knowledge from Banpheo hospital's operational database by develop an application for self-discovering and for interfacing with hospital staffs. The results can be found in section 5.

Fig.1: *Banpheo hospital (public organization), Sumutsakon province, Thailand*

4. MINING ASSOCIATION RULES WITH APRIORI_MSG-P

Apriori with Multiple Support Generating by statistic Percentile threshold (Apriori_MSG-P) was proposed in [18] and was shown in Fig. 2. This algorithm combined statistic concepts, Normal Distribution Curve, to traditional Apriori by analyzing all outcomes and defining the meaning of word "user's interestingness" with statistic percentile parameter. So, Apriori_MSG-P is able to provide the result meet users need and still preserves rare itemsets in finally.

Refer to Fig. 2, function *Apriori_MSG-P* that shown in line1 to 8 and 13 to 15 are the same as traditional Apriori - evaluating candidates and frequent itemsets. Apriori_MSG-P adds the calculation steps for support's statistic in line 9 to 11, itemset size by itemset size. User's statistic percentile value α is applied in line 11 by tracing Normal Distribution Curve for minimum support at its size (MSS). At line 12, the true minimum support at its size evaluates from the greatest value, between MSS and LS (shown in definition 1). The least support or LS is another user-specified threshold. Least support refers the lowest minimum support which should satisfy to become a frequent itemsets and prevents to create uninteresting itemsets when standard deviation of its size nears zero. Notice, *apriori_gen* and $has\_infrequent\_subset$ are also as same as traditional Apriori that do not be described here.

Definition 1: the minimum support of k-itemsets candidates: MS_k is defined as below where MSS_k is minimum support that calculates from user percentile value (α) and LS is user specific least support.

$$MS_k = \begin{cases} MSS_k & if \;\; MSS_k > LS; \\ LS & Otherwise; \end{cases} \tag{1}$$

Definition 2: the minimum support of k-itemsets: MSS_k is the generated support threshold which its value depends on statistic data of support values of all candidates in same itemset size, based on Normal Distribution Curve. MSS_k is the support value which located on percentile, calculated from (2), and that percentile have to equal with user specific percentile α (practically, Standard Curve Statistical Table is easier to use). $Z - score$ getting from (2) will be used in (3) to find out MSS_k. Remark that μ is Mean and σ is Standard Deviation.

$$Percentile(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-1/_2 z^2}\, dz \tag{2}$$

$$MSS = z \times \sigma + \mu \tag{3}$$

Notice that the single user-specified percentile value will be used to generate the multiple minimum supports, itemset size by itemset size. The infrequent itemsets, whose supports are less than MSS_k, will not be used for generating next size of candidates. So, this can reduce the number of uninterested itemsets and improves the creating time performance, especially on large database, and also preserves rare itemsets on high size of itemset.

The extraction of rare itemsets using Apriori_MSG-P illustrated in example 1.

Example 1: Table 2 shown the dataset of 10 transactions with 12 items. This example applied thresholds with percentile 40% and least support 10%

```
$Apriori\_MSG-P(D,\alpha,LS)$
(1)  $L_1 = find\_frequent\_1-itemsets(D);$
(2)  $for(k=2; L_{k-1} \neq \phi; k++)\{$
(3)      $C_k = Candidate\_Gen(L_{k-1});$
(4)      $for\ each\ transaction\ t \in D\{$
(5)          $C_t = subset(C_k, t);$
(6)          $for\ each\ candidate\ c \in C_t$
(7)              $c.count++;$
(8)      $\}$
(9)      $\sigma_k = get\_stddev(C_k | c.count > 0, c \in C_k));$
(10)     $\mu_k = get\_mean(C_k | c.count > 0, c \in C_k));$
(11)     $MSS_k = get\_MSS(\alpha,\sigma,\mu);$
(12)     $MS_k = \begin{cases} MSS_k & If\ MSS_k > LS; \\ LS & Otherwise, \end{cases}$
(13)     $L_k = c \in C_k | c.count > MS_k;$
(14)     $return\ L = \cup_k L_k;$
(15) $\}$
```

Notation:
$D = Database$
$\alpha = User\ specific\ percentile$
$LS = Least\ support$
$L_k = Frequent\ itemset\ with\ k-items$
$C_k = Candidate\ itemset\ with\ k-items$
$MSS_k = Minimum\ support\ by\ significant\ of\ k-itemsets$
$MS_k = Minimum\ support\ of\ k-itemsets$

Fig.2: *Psuedo code of Apriori_MSG-P*

(or 1 support count) that the results shown in table 2-5.

Table 2: *Transformed Database*

TID	*Items*
1	A, D, H
2	B, G, I, L
3	B, F, J, L
4	A, E, H, K
5	C, D, H, L
6	C, F, J, L
7	A, D, H, K
8	C, F, J, L
9	C, F, J, L
10	J, L

Table 3: *Itemsets at Size 1*

Itemset	*Count*	*Is Frequent*
{A}	3	Yes
~~{B}~~	~~2~~	~~No~~
{C}	4	Yes
{D}	3	Yes
~~{E}~~	~~1~~	~~No~~
{F}	4	Yes
~~{G}~~	~~1~~	~~No~~
{H}	4	Yes
~~{I}~~	~~1~~	~~No~~
{J}	5	Yes
~~{K}~~	~~2~~	~~No~~
{L}	7	Yes

Mean = 3.083333
SD = 1.831955
MSS,MS = 2.619244
Coverage = 7/12 = 58.33%

From database shown in table 2, size 1 frequent itemsets were discovered and shown in table 3. Mean of support count of whole size 1 itemsets is 3.083333 and standard deviation is 1.831955. At user preference of 40 percentile and these statistic parameter values, MSS_1 is 2.619244 (reference from (3)). Whereas LS is 0.1, means least support count is 1 transaction, the MS_1 is the greater value between MSS1 and LS or equals 2.619244. The itemsets which have support count lower than 2.619244 will be neglected. We can found that the number of frequent itemsets is 7 from 12 candidate itemsets, or coverage is 58.33%. All of frequent itemsets in itemset size 1 will be used for generate candidate itemsets in size 2 and so on.

Table 4: *Itemsets at Size 2*

Itemset	*Count*	*Is Frequent*
~~{A,D}~~	~~2~~	~~No~~
{A,H}	3	Yes
~~{C,D}~~	~~1~~	~~No~~
{C,F}	3	Yes
~~{C,H}~~	~~1~~	~~No~~
{C,J}	3	Yes
{C,L}	4	Yes
{D,H}	3	Yes
~~{D,L}~~	~~1~~	~~No~~
{F,J}	4	Yes
{F,L}	4	Yes
~~{H,L}~~	~~1~~	~~No~~
{J,L}	5	Yes

Mean = 2.692307
SD = 1.377474
MSS,MS = 2.343352
Coverage =8/13 = 61.54%

Table 5: *Itemsets at Size 3*

Itemset	*Count*	*Is Frequent*
~~{C,F,J}~~	~~3~~	~~No~~
~~{C,F,L}~~	~~3~~	~~No~~
~~{C,J,L}~~	~~3~~	~~No~~
{F,J,L}	4	Yes

Mean = 3.25
SD = 0.50
MSS,MS = 3.123335
Coverage = 1/4 = 25.00%

We can notice from above example that the coverage approximately equals 100-percentile or 100-40=60%. This helps us to spot on top most of interesting itemsets in each size.

5. EXPERIMENTAL RESULTS

5.1 Experimental results on Northwind database

The experiments run on Northwind database which is performed the attribute-oriented induction process described above. We compare the results of Apriori_MSG-P and traditional Apriori at the stage that each of techniques provides the approximately same number of outcome itemsets by adjusting threshold. The results were shown in table 6-7 and Fig. 3-5. Obviously, the rare itemsets can be better found, especially at the high itemset size. At itemset size 5-6 in Fig. 5, Apriori_MSG-P can discover itemsets that Apriori scarcely found. Also at size 3-4 in Fig. 4, Apriori_MSG-P takes the better discovery performance with average 560.60% better. This means traditional Apriori cannot evaluate rare itemsets in upper size 3 effectively. In the contrary at size 1-2 in Fig. 3, traditional Apriori provide greater number of outcomes than Apriori_MSG-P but a lot of them seem uninteresting when comparing with support value among their members. Therefore Apriori_MSG-P is able to examine and choose the itemsets at all itemset sizes and those itemsets seem significant among their siblings with have the same itemset size.

Moreover in traditional Apriori, finding the best minimum support is very difficult to guessing at the first time. In contrast, Apriori_MSG-P is able to use easily because of its statistic meaning. Percentile is the statistic parameter that can estimate the number of outcomes (more accurately when the distribution looks like symmetry). For example, 40% of percentile value means about 60% of high support itemsets will be selected (left tail concept).

Table 6: *The results from traditional Apriori*

Size of itemsets	*Number of itemsets at each min support*						
	1%	*2%*	*3%*	*4%*	*5%*	*6%*	*7%*
1	143	75	53	45	42	37	34
2	645	238	131	83	57	35	24
3	439	95	26	8	6	2	1
4	38	0	0	0	0	0	0
5	1	0	0	0	0	0	0
6-8	0	0	0	0	0	0	0
Total	1,266	408	210	136	105	74	59

Table 7: *The results from Apriori_MSG-P*

Size of itemsets	*Number of itemsets at each percentile*						
	50%	*55%*	*60%*	*65%*	*70%*	*75%*	*80%*
1	46	43	38	34	29	24	20
2	239	185	127	91	65	38	23
3	528	336	172	85	53	22	6
4	708	315	167	88	38	6	0
5	207	118	53	11	5	0	0
6	5	2	0	0	0	0	0
7-8	0	0	0	0	0	0	0
Total	1,733	999	557	309	190	90	49

a. Least Support sets to 0.1%

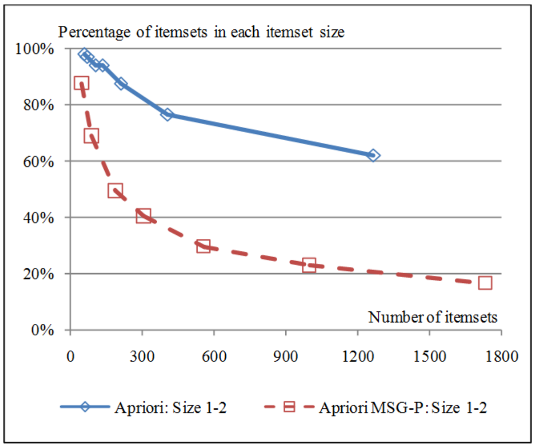


Fig.3: *Percentage of number of itemsets in itemset size 1-2 using Apriori and Apriori_MSG-P*

The experiments on creating time were shown in Fig. 6. The results obviously show that Apriori_MSG-P takes lower creating time, especially when the number of result itemsets is high. Note that the experiments run on the Pentium Dual Core 3.00GHz with 2GByte RAM.

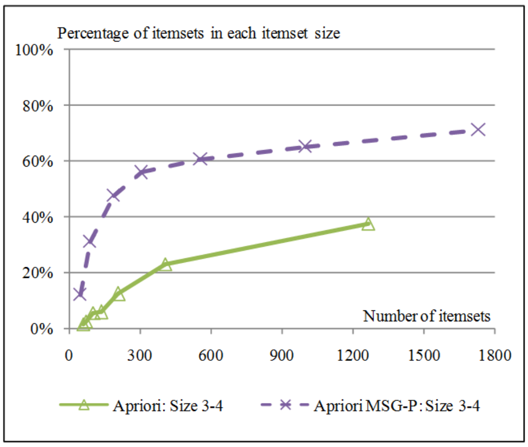


Fig.4: *Percentage of number of itemsets in itemset size 3-4 using Apriori and Apriori_MSG-P*

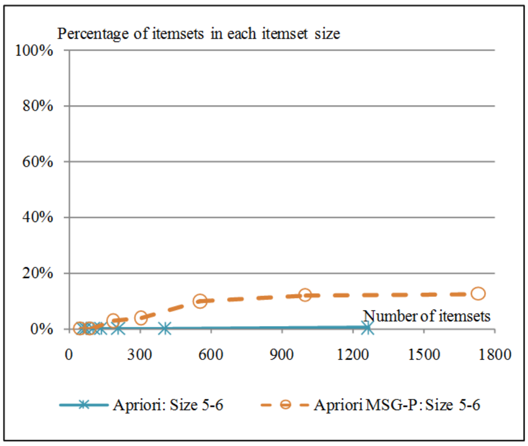


Fig.5: *Percentage of number of itemsets in itemset size 5-6 using Apriori and Apriori_MSG-P*

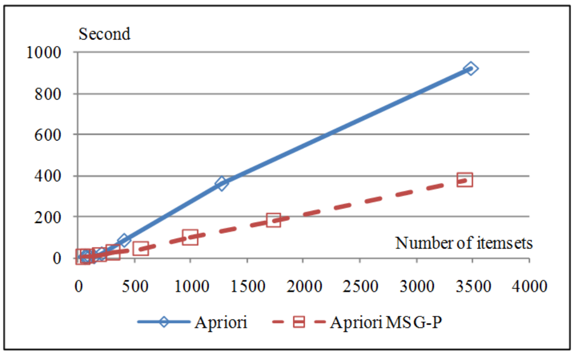


Fig.6: *Creating time using Apriori and Apriori_MSG-P*

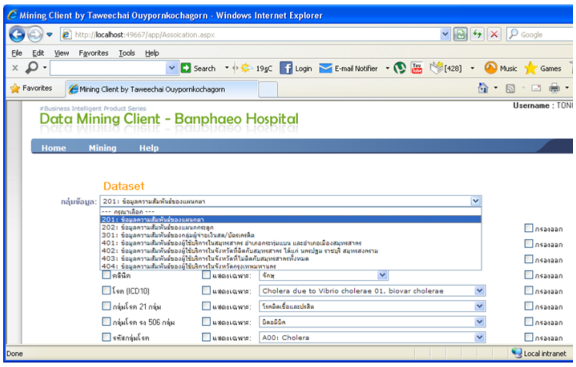
5.2 Experimental results on Banpheo hospital HIS

The experiments on Banpheo hospital HIS categorize into 7 categories. Each of categories target on discovering interesting behaviors or patterns between selected attributes that hidden on Banpheo hospital HIS. All of experimental results run on real operational database between February 2010 and February 2011 on 14 selected attributes as following:

- ■ Month of year such as January, February.
- ■ Day of week such as Monday, Tuesday.
- ■ Clinic such as Orthopedic, Obstretic Gynecology.
- ■ Disease (reference on standard ICD10 code) such as Tuberculosis of eye, Chancroid.
- ■ Disease group such as Erysipelas, Measles, Dermatophytosis.
- ■ Diagnostic 21 groups such as Neoplasms, Diseases of the nervous system.
- ■ Pestilence that have to report Bureau of Epidemiology, Thailand (Report 506) such as, ChoierA Chickenpox.
- ■ Patient age range consists of 0-5, 5-15, 15-25, 25-35, 35-45, 45-55, 55-65, 65-75, 65-75, 75+, and No data.
- ■ Patient pay right such as social security right, government officer right.
- ■ Marital status such as single, married, and divorced.
- ■ Sex consists of male, female, and not specific.
- ■ Education background such as Grade 4, 6, 9, 12, and bachelor.
- ■ Occupation such as Engineer, Farmer, and Reporter.
- ■ Region of present address consists of local area (Banpheo District), close neighbor area (other districts in Sumutsakon province), further neighbor area (provinces around Sumutsakon except Bangkok), Bangkok (Capital of Thailand), other provinces, and not specific.

Note that some of experiments neglect one or some attributes since they need to fix one or some attributes due to their research topic.

The experiments performed by implementing Apriori_MSG-P on each categories of researching topics, on the Pentium Dual Core 3.00GHz with 2GByte RAM. As the results, Apriori_MSG-P gave the interesting association rules in each item sizes which were stored in database. This database recommended to be used together with rules discovery application, "Data Mining Client v1.0", which was developed for users of Banpheo hospital. This application will permit users to select researching topic outcomes, to specific interesting/ uninteresting attributes or interesting/ uninteresting item of each attributes, and to specific itemset by the way of forming between itemsets which are contained only with specific attributes and itemsets which are contained at least with specific attributes. The examples of "Data Mining Client v1.0" application were shown in Fig. 7-8.


Mining Client by Taweechai Ouypornkochagorn - Windows Internet Explorer
http://localhost:49667/app/Assoication.aspx
Google
File Edit View Favorites Tools Help
Search
19°C
Login
E-mail Notifier
[428]
Music
Games
Favorites
Mining Client by Taweechai Ouypornkochagorn
#Business Intelligent Product Series
Data Mining Client - Banphaeo Hospital
Username : TONI
Home
Mining
Help
Done
Local intranet

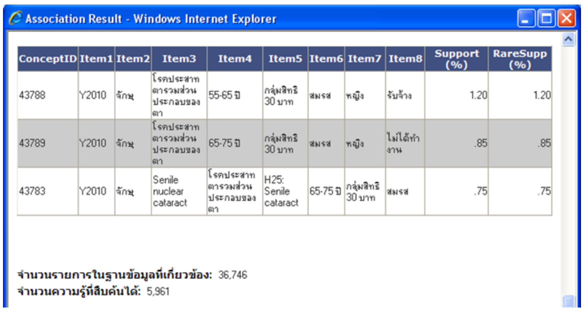
Dataset

กลุ่มข้อมูล: 201: ข้อมูลความสัมพันธ์ของแผนกตา

--- กรุณาเลือก ---
201: ข้อมูลความสัมพันธ์ของแผนกตา
202: ข้อมูลความสัมพันธ์ของแผนกกระดูก
301: ข้อมูลความสัมพันธ์ของกลุ่มผู้จ่ายเงินสด/บัตรเครดิต
401: ข้อมูลความสัมพันธ์ของผู้ใช้บริการในสมุทรสาคร อำเภอกระทุ่มแบน และอำเภอเมืองสมุทรสาคร
402: ข้อมูลความสัมพันธ์ของผู้ใช้บริการในจังหวัดที่ติดกับสมุทรสาคร ได้แก่ นครปฐม ราชบุรี สมุทรสงคราม
403: ข้อมูลความสัมพันธ์ของผู้ใช้บริการในจังหวัดที่ไม่ติดกับสมุทรสาครทั้งหมด
404: ข้อมูลความสัมพันธ์ของผู้ใช้บริการในจังหวัดกรุงเทพมหานคร

☐ กรองออก
☐ กรองออก
☐ กรองออก
☐ คลินิก ☐ แสดงเฉพาะ: จักษุ ☐ กรองออก
☐ โรค (ICD10) ☐ แสดงเฉพาะ: Cholera due to Vibrio cholerae 01, biovar cholerae ☐ กรองออก
☐ กลุ่มโรค 21 กลุ่ม ☐ แสดงเฉพาะ: โรคติดเชื้อและปรสิต ☐ กรองออก
☐ กลุ่มโรค รง 506 กลุ่ม ☐ แสดงเฉพาะ: บิดมีตัว ☐ กรองออก
☐ รหัสกลุ่มโรค ☐ แสดงเฉพาะ: A00: Cholera ☐ กรองออก

Fig.7: *"Data Mining Client v1.0" application*



ConceptID	Item1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Support (%)	RareSupp (%)
43788	Y2010	จักษุ	โรคประสาทตารวมส่วนประกอบของตา	55-65 ปี	กลุ่มสิทธิ 30 บาท	สมรส	หญิง	รับจ้าง	1.20	1.20
43789	Y2010	จักษุ	โรคประสาทตารวมส่วนประกอบของตา	65-75 ปี	กลุ่มสิทธิ 30 บาท	สมรส	หญิง	ไม่ได้ทำงาน	.85	.85
43783	Y2010	จักษุ	Senile nuclear cataract	โรคประสาทตารวมส่วนประกอบของตา	H25: Senile cataract	65-75 ปี	กลุ่มสิทธิ 30 บาท	สมรส	.75	.75

จำนวนรายการในฐานข้อมูลที่เกี่ยวข้อง: 36,746
จำนวนความรู้ที่สืบค้นได้: 5,961

Fig.8: *Discovering association rules which are exploited by Apriori_MSG-P algorithm*

5.2.1 Patient behavior of Eye clinic

This research topic aims to study patient behavior of Eye clinic from 36,746 clinic visits between February 25, 2010 and February 22, 2011. This clinic has the different customer behavior from most of clinics in Banpheo hospital and the hospital executive staffs need to exploit some facts that may hide in HIS. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 8. The period of time that was used to process is 39 minutes.

Table 8: *Examples of interesting behavior on Eye clinic*

Rules	*Support*
{Z09: Follow-up examination after treatment for conditions other than malignant neoplasms, Age range 65-75, Present address: local area}	2.10%
{ H40: Glaucoma, Age range 65-75, Present address: local area}	2.06%
{ H25: Senile cataract, Age range 65-75, Present address: local area}	2.01%

a. Show only itemset size 3 which have highest support.
b. Some of rules are prohibited to uncover.

5.2.2 Patient behavior of Orthopedic clinic

This research topic aims to study patient behavior of Orthopedic clinic from 33,675 clinic visits between February 25, 2010 and February 22, 2011. This clinic has the different customer behavior from most of clinics in Banpheo hospital (like Eye clinic) and the hospital executive staffs need to exploit some facts that may hide in HIS. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 9. The period of time that was used to process is 32 minutes.

Table 9: *Examples of interesting behavior on Orthopedic clinic*

Rules	***Support***
{ Gonarthrosis [arthrosis of knee], Age range 55-65, Present address: local area}	3.99%
{ Gonarthrosis [arthrosis of knee], Age range 65-75, Present address: local area}	3.10%
{ Gonarthrosis [arthrosis of knee], Age range 55-65, Present address: local area}	2.33%
{ Other spondylopathies, Age range 55-65, Present address: local area}	1.53%

a. Show only itemset size 3 which have highest support

b. Some of rules are prohibited to uncover.

5.2.3 Patient paying behavior on cash, credit card or equivalent

This research topic aims to study patient behavior on paying behavior on cash, credit card or equivalent from 88,542 visits between February 25, 2010 and February 22, 2011. Since Banpheo hospital is bureaucratic organization that accepts all of social security rights, the hospital customers who choose to pay by cash, credit card, or equivalent are in highly interesting of executive staffs. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 10. The period of time that was used to process is 4 hours 57 minutes.

Table 10: *Examples of interesting patient paying behavior on cash, credit card, or equivalent*

Rules	***Support***
{ Paediatrics clinic, Age range 0-5, Present address: local area}	3.88%
{ Medicine clinic, Age range 75+, Present address: local area}	3.81%
{ Medicine clinic, Age range 65-75, Present address: local area}	3.22%
{ Medicine clinic, Age range 55-65, Present address: local area}	3.01%
{ Medicine clinic, Age range 45-55, Present address: local area}	1.95%
{ Medicine clinic, Age range 65-75, Present address: further neighbor area }	1.53%

a. Show only itemset size 3 which have highest support

b. Some of rules are prohibited to uncover.

5.2.4 Patient behavior of close neighbor area customers (other districts in Sumutsakon province)

This research topic aims to study patient behavior of close neighbor area customers (other districts in Sumutsakon province: Kratumban and Mueng district) from 31,610 visits between February 25, 2010 and February 22, 2011. This topic wants to study why customers on other areas outside Banpheo district prefer to take the hospital services at Banpheo hospital and what are their preferred clinics. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 11. The period of time that was used to process is 55 minutes.

Table 11: *Examples of interesting patient behavior of close neighbor area customers (other districts in Sumutsakon province)*

Rules	***Support***
{ Medicine clinic, Age range 55-65 }	8.21%
{ Medicine clinic, Age range 45-55}	7.37%
{ Medicine clinic, Age range 65-75 }	6.22%
{ Medicine clinic, Age range 75+}	6.11%
{ Paediatrics clinic, Age range 0-5 }	4.50%
{ Medicine clinic, Age range 35-45}	4.31%
{ Medicine clinic, Age range 25-35}	4.19%
{ Eye clinic, Age range 65-75}	3.32%

a. Show only itemset size 2 which have highest support

b. Some of rules are prohibited to uncover.

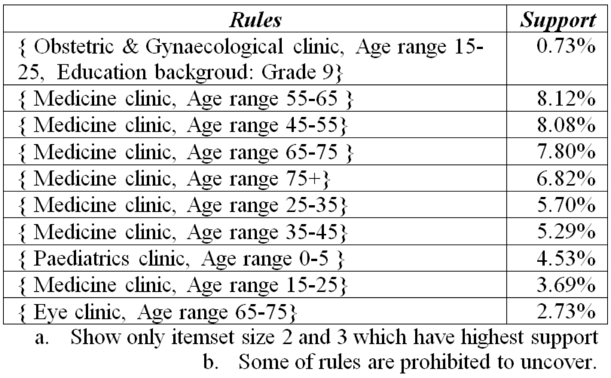
5.2.5 Patient behavior of further neighbor area customers (provinces around Sumutsakon except Bangkok)

This research topic aims to study patient behavior of further neighbor area customers (provinces around Sumutsakon except Bangkok, Nakornpathom, Rachabury and Sumutsongkram province) from 59,726 visits between February 25, 2010 and February 22, 2011. This topic wants to study why customers on other provinces around Sumutsakon province except Bangkok prefer to take the hospital services at Banpheo hospital and what are their preferred clinics. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 12. The period of time that was used to process is 1 hour 50 minutes.

5.2.6 Patient behavior of Bangkok

This research topic aims to study patient behavior of Bangkok province customers (Capital of Thailand) from 12,927 visits between February 25, 2010 and February 22, 2011. This topic wants to study why customers at Bangkok (which place that has highest medical resources of Thailand) prefer to take the hospital services at Banpheo hospital and what are their preferred clinics. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 13. The period of time that was used to process is 23 minutes.

Table 12: *Examples of interesting patient behavior of further neighbor area customers (provinces around Sumutsakon except Bangkok)*

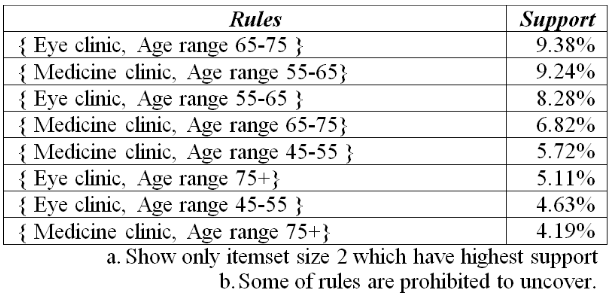


Rules	Support
{ Obstetric & Gynaecological clinic, Age range 15-25, Education backgroud: Grade 9}	0.73%
{ Medicine clinic, Age range 55-65 }	8.12%
{ Medicine clinic, Age range 45-55}	8.08%
{ Medicine clinic, Age range 65-75 }	7.80%
{ Medicine clinic, Age range 75+}	6.82%
{ Medicine clinic, Age range 25-35}	5.70%
{ Medicine clinic, Age range 35-45}	5.29%
{ Paediatrics clinic, Age range 0-5 }	4.53%
{ Medicine clinic, Age range 15-25}	3.69%
{ Eye clinic, Age range 65-75}	2.73%

a. Show only itemset size 2 and 3 which have highest support

b. Some of rules are prohibited to uncover.

Table 13: *Examples of interesting patient behavior of Bangkok, the capital of Thailand*



Rules	Support
{ Eye clinic, Age range 65-75 }	9.38%
{ Medicine clinic, Age range 55-65}	9.24%
{ Eye clinic, Age range 55-65 }	8.28%
{ Medicine clinic, Age range 65-75}	6.82%
{ Medicine clinic, Age range 45-55 }	5.72%
{ Eye clinic, Age range 75+}	5.11%
{ Eye clinic, Age range 45-55 }	4.63%
{ Medicine clinic, Age range 75+}	4.19%

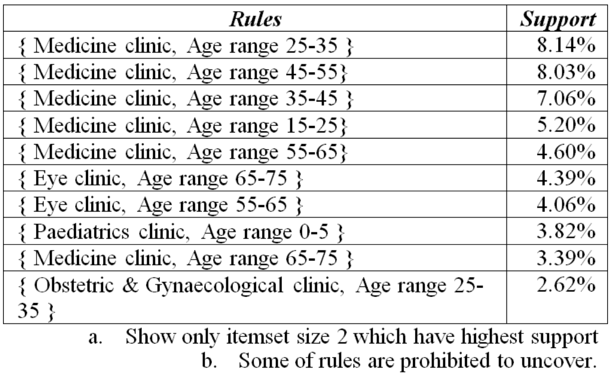
a. Show only itemset size 2 which have highest support

b. Some of rules are prohibited to uncover.

5.2.7 Patient behavior of other provinces of Thailand

This research topic aims to study patient behavior of other provinces customers from 35,687 visits between February 25, 2010 and February 22, 2011. This topic wants to study why customers on other provinces of Thailand prefer to take the hospital services at Banpheo hospital and what are their preferred clinics. The experiment sets the user percentile of Apriori_MSG-P at 50%. The results were shown in table 14. The period of time that was used to process is 1 hour 16 minutes.

Table 14: *Examples of interesting patient behavior of other provinces of Thailand*



Rules	Support
{ Medicine clinic, Age range 25-35 }	8.14%
{ Medicine clinic, Age range 45-55}	8.03%
{ Medicine clinic, Age range 35-45 }	7.06%
{ Medicine clinic, Age range 15-25}	5.20%
{ Medicine clinic, Age range 55-65}	4.60%
{ Eye clinic, Age range 65-75 }	4.39%
{ Eye clinic, Age range 55-65 }	4.06%
{ Paediatrics clinic, Age range 0-5 }	3.82%
{ Medicine clinic, Age range 65-75 }	3.39%
{ Obstetric & Gynaecological clinic, Age range 25-35 }	2.62%

a. Show only itemset size 2 which have highest support

b. Some of rules are prohibited to uncover.

5.3 Apriori_MSG-P's result analysis comparing with traditional Apriori's results

The experiments repeatedly process on patient behavior of Eye clinic of Banpheo hospital, on 36,746 clinic visits between February 25, 2010 and February 22, 2011, by evaluating very interesting itemsets in users' view. After running Apriori_MSG-P with 85 percentile and 0.1 least support, which is the setting that seems to represent emphasis on high interesting itemsets, it products 187 itemsets with lowest 8.09% support value. This number of results is only about 0.70% when comparing with the number of results of 1% minimum support of traditional Apriori. We try to study how Apriori_MSG-P is better than traditional Apriori. Our analysis separates into 3 subjects.

5.3.1 Interesting itemsets that found in Apriori_MSG-P but don't found in traditional Apriori

This experiment tried to change minimum support threshold of traditional Apriori to stage that gives the number of itemsets equal with Apriori_MSG-P at 85 percentile (187 itemsets outcome) and then we found at 15.64% minimum support: 147 itemsets are same with Apriori_MSG-P and 40 itemsets are different. Among the different itemsets of Apriori_MSG-P outcomes, 23 itemsets are item size 1 and 17 itemsets are size 2. To prove the itemsets that were found by Apriori_MSG-P are interesting, we pick up a different itemset that has lowest support value in each item size for analyzing.

At item size 1, we found item February. This itemset has 2,974 support count and 8.09% support value. We investigate its interesting by comparing among its siblings and found that among size 1 itemsets they have 461 support count average and 2,377.24 support count standard deviation. That means {February} locates on 85.47 percentile among its siblings.

By the same proving way, 30 Baht health security right, Age range: 65-75 is the lowest support itemset in item size 2 with 5,102 support count and 13.88% support value. Statistic parameters on 2 item size are 2,075 support count average and 2,836.52 support count standard deviation. That means {30 Baht health security right, Age range: 65-75} locates on 85.70 percentile among its siblings.

Therefore, both of picked up itemsets are interesting due to their percentile, we can conclude that Apriori_MSG-P can retrieve interesting itemsets more effectively than traditional Apriori since its different outcomes locate in high percentile among their siblings.

5.3.2 Itemsets that found in traditional Apriori but Apriori_MSG-P judges to be uninteresting itemsets

This experiment tried next from previous experiment by staring on the 40 different outcomes that traditional Apriori retrieved. 20 itemsets are item size 3, 16 itemsets are size 4 and 4 itemsets are size 5. We pick up a highest support value itemsets in each itemset size to determine why Apriori_MSG-P denied these itemsets.

At item size 3, {Year: 2010, Eye clinic, General pay right} has 6,597 support count and 17.95% support value. Among its siblings, statistic parameters are 8,902 support count average and 3,174.14 support count standard deviation. This item locates at 23.39 percentile.

At item size 4, {Eye clinic, ICD10 group: G00 - G99 Diseases of the nervous system, Married, Female} has 7,604 support count and 20.69% support value. Among its siblings, statistic parameters are 7,820 support count average and 1,731.30 support count standard deviation. This item locates at 45.03 percentile.

At item size 5, {Year 2010, Eye clinic, 30 Baht health security right, Married, Present Address: Bangkok-Thailand} has 6,532 support count and 17.78% support value. Among its siblings, statistic parameters are 6,615 support count average and 797.68 support count standard deviation. This item locates at 45.86 percentile.

Therefore, from the low percentile value of picked up itemsets that shown above, we can conclude that even though these itemsets have a high support value but they seem uninteresting comparing with their siblings. So, traditional Apriori with minimum support may not the good way to evaluate "interesting itemset".

5.3.3 Interesting itemsets evaluating performance

Due to last 2 discussed subjects, we can conclude that traditional Apriori with minimum support is the threshold technique that does not try to interpret "interestingness" among their sibling itemsets. Therefore traditional Apriori may accidentally evaluate uninteresting itemsets and this seems very difficult to get rid of this problem. From 187 itemsets of Apriori_MSG-P with lowest 8.09% support, when we try to evaluate itemsets with 8.09% minimum support on traditional Apriori, it products 729 itemsets or 3.90 time comparing with the number of outcome from Apriori_MSG-P. This analysis can conclude that traditional Apriori gives the lower evaluating performance. When users want to evaluate some of interesting itemsets, traditional Apriori produces a lot of uninteresting itemsets (comparing with their sibling itemsets) in the same time.

6. CONCLUSION

In this paper, we implement the improvement of traditional Apriori, Apriori_MSG-P, with the real-world application, Banpheo's hospital information system (HIS). Apriori_MSG-P processing is based on multiple minimum supports by a single statistic percentile value. It can mine the rare itemsets effectively better than traditional Apriori because it examines and considers about the different characteristic of support value in each item size of itemsets during evaluation processes, and then automatically generates multiple minimum support value from supports characteristic of its size. In additional, the statistic percentile value is more understandable and readily choices, especially when trying to choose or to adjust for the best one at first time. Experimental results show that rare itemsets evaluating performance by Apriori_MSG-P are more effective in evaluating performance and in creating time performance.

Turning to consider about the experimental results of Banpheo's database implementing, Apriori_MSG-P can be used for discovering hidden information in aspect of association rules with acceptable running time. The outcome rules, especially on the high itemset sizes, give a lot of valuable patient behaviors and medical notices that both operation staffs and executive staffs can enrich their services or improve their arrangement more appropriately.

References

[1] B. A. Mahafzah, A. F. Al-Badarneh and M. Z. Zakaria, "A New Sampling Technique for Association Rule Mining," in *Journal Of Information Science*, vol.35, 2009, pp.358–376.

[2] B. Liu, W. Hsu, and Y. Ma, "Mining Association Rules with Multiple Minimum Supports," *SIGKDD Explorations*, 1999.

[3] C. Mitica and A. Cristian, "Association Rules Discovery using Grid Services," In *3rd International Conference: Sciences of Electronic, Technologies of Information and Telecommunications*, 2005.

[4] C. Wang and C. Tjortjis, "Prices: An Efficient Algorithm for Mining Association Rules," *Lecture Notes in Computer Science*, vol. 2447, 2002, pp.77–83.

[5] D. Cheung, J. Han, V. Ng, A. Fu, and Y. Fu, (1996), "A Fast Distributed Algorithm for Mining Association Rules," In *Proc of 1996 Int'l Conference on Parallel and Distributed Information Systems'*, 1996, pp.31–44.

[6] D. Cheung and Y. Xaio, "Effect of Data Skewness in Parallel Mining of Association Rules," *Lecture Notes in Computer Science*, vol. 1394, 1998, pp.48-60.

[7] D. Xiangjun, Z. Zhiyun, N. Zhendong and J. Qiuting, "Mining Infrequent Itemsets Based on Multiple Level Minimum Supports," *Innovative*

Computing, Information and Control (ICICIC '07), 2007, pp.528.

[8] En Tzu Wang and L. P. Arbee Chen, "A Novel Hash-Based Approach for Mining Frequent Itemsets over Data Streams Requiring Less Memory Space," *Data Mining and Knowledge Discovery*, vol. 19, no. 1, 2009, pp 132-172.

[9] J. D. Holt and S.M. Chung, "Mining of Association Rules in Text Databases Using Inverted Hashing and Pruning," *Lecture Notes in Computer Science*, Volume 1874/2000, 2000, pp.290–300.

[10] J. Han and J. Pei, "Mining Frequent Patterns by Pattern-Growth: methodology and implications," *ACM SIGKDD Explorations Newsletter2*, 2000, pp.14–20.

[11] J. Soo, M. S. Chen, and P. S. Yu, "Using a Hash-Based Method with Transaction Trimming and Database Scan Reduction for Mining Association Rules," *IEEE Transactions On Knowledge and Data Engineering*, vol., no.5., 1997, pp.813-825.

[12] K. Waiyamai and L. Lakhal, "Knowledge Discovery from Very Large Databases Using Frequent Concept Lattices," In *Proc. ECML 2000*, 2000, pp.437–445.

[13] O. Weimin and H. Qinhua, "Mining Direct and Indirect Association Patterns with Multiple Minimum Supports," *Computational Intelligence and Software Engineering (CiSE)*, 2010, pp.1–4.

[14] R. Agarwal and R. Srikant, "Fast Algorithms for Mining Association Rules," In *Proc.20th Int. Conf. Very Large Data Bases*, 1994, pp. 487–499.

[15] R. E. Thevar, and R. Krishnamoorthy, "A New Approach of Modified Transaction Reduction Algorithm for Mining Frequent Itemset," In *Proc. 11th conference on Computer and Information Technology ICCIT 2008*, 2008.

[16] R. Uday Kiran and P. Krishna Re, "An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules," *Computational Intelligence and Data Mining (CIDM '09), IEEE Symposium*, 2009, pp.340–347.

[17] S. Parthasarathy, "Efficient Progressive Sampling for Association Rules," *IEEE International Conference on Data Mining*, 2002, pp.354–361.

[18] T. Ouypornkochagorn and K. Waiyamai, ""Apriori_MSG-P: A Statistic-Based Multiple Minimum Support Approach to Mine Rare Association Rules," *The 3rd International Conference on Knowledge and Smart Technologies (KST-2011)*, Chonburi, Thailand, 2011.

[19] [T. Ouypornkochagorn and K. Waiyamai, "Domain Ontology Development and Maintenance using Pertinent Concept Lattice," *GESTS International Transactions on Computer Science*, vol. 4, no. 1, 2005.

[20] T. Ouypornkochagorn and K. Waiyamai, "Formal Concept Mining: A Statistic-based Approach for Pertinent Concept Lattice Construction," In *Proc. 9th Asian Computing Science Conference ASIAN2004 Publish in Lecture Notes in Computer Science*, 2004, pp.195–204.

[21] V. Umarani and M. Punithavalli, "Developing a Novel and Effective Approach for Association Rule Mining Using Progressive Sampling," In *the proc of 2nd Int'l Conference on Computer and Electrical Engineering (ICCEE 2009)*, vol. 1, 2009, pp.610–614.

[22] V. Umarani and M. Punithavalli, "On Developing an Effectual Progressive Sampling Based Approach for Association Rule Discovery," In *the proc of 2nd IEEE Int'l Conference on Information and data Engineering (2nd IEEE ICIME 2010)*, 2010.

Taweechai Ouypornkochagorn is doing a Ph.D. on University of Manchester, UK in Biomedical Engineering, financial supporting by Office of The Civil Service Commission, Thailand. He achieved Master degree (M.Eng) in computer engineering at Kasetsart University in 2003. He was a lecturer at Kasetsart University, Si Racha campus. His researches involve medical imaging technologies such as EIT, MRI, data mining on real medical databases, knowledge base system, and embedded system applications. He received awards "Best 50 Ideas for Thailand", by The Prime Minister's Office, Thailand in 2010 and "Best 60 Thailand Embedded Product Award", by Thai Embedded Systems Association in 2010 and 2009.