a Method of Discovering Interesting assocation rules from student admision dataset

by Arief Hermawan

A Method of Discovering Interesting Association Rules from Student Admission Dataset

Wiwik Novitasari[1], Arief Hermawan[2], Zailani Abdullah[3], Rahmat Widia Sembiring[4] and Tutut Herawan[5]

[1]Universitas Muhammadiyah Tapanuli Selatan, Sumatera Utara, Indonesia
[2]Universitas Teknologi Yogyakarta, Kampus Jombor, Yogyakarta, Indonesia
[3]Department of Computer Science, Universiti Malaysia Terengganu, Malaysia
[4]Politeknik Negeri Medan, Sumatera Utara, Indonesia
[5]AMCS Research Center, Yogyakarta, Indonesia
vitampd@yahoo.com, ariefdb@uty.ac.id, zailania@umt.edu.my,
rahmatws@polmed.ac.id, tutut@amcs.co

Abstract

For the past decades and until now, association rule mining is one of the most prominent research topics in data mining. However, the main challenge among public or private practitioners is to find the interesting rule from data repository. As a result, many efforts have been put forward to explore this rule by applying several methods and interesting measures. Therefore, in this paper, we introduced an enhanced association rule mining method namely Significant Least Pattern Growth (SLP-Growth), where the algorithm embeds with two interesting measures called Critical Relative Support (CRS) and Correlation (Corr). The experiment uses the dataset that contains the records of preferred programs being selected by post-matriculation or post-STPM students of Malaysia via Electronic Management of Admission System (e-MAS) for the year 2008/2009. The experimental results show that the SLP-Algorithm with the embedded measures can successfully in categorizing the association rules. In addition, this information can be used by educators and higher university authority personnel in the university to understand the programs' patterns being selected by the students. More importantly, it can assist them as a basis to offer more relevant programs to the potential students rather than by chance technique.

Keywords: Data mining, Association rule, significant least patterns, student dataset

1. Introduction

University opens a host of new opportunities and takes control to the future development. Anyone with a higher education qualification makes more attractive in the eyes of employers and increases the earning potential. It also shows the ability to learn at a higher level system with the preferred programs. In Malaysia, public universities are among the ultimate directions and choices for most of the post-matriculation or post-STPM students. After the students obtaining the actual result of the examination, they have to choose their preferred programs at Malaysian public universities via Electronic Management of Admission System (e-MAS) within certain period of time. E-MAS has been developed for Ministry of Education (MOE, previously known as Ministry of High Education) to facilitate the admission of students into their preferred programs. Prior to that, all new codes, programs, program requirements, etc. will be registered by the MOE authority personnel. However, the main concern is, for the average and the lower grades students, they might not be offered to their preferred programs. As a result, some of them might not perform well, less motivation and in an extreme case, drop-out from the

university. Based on this situation, many studies [1-3] have been intensively conducted to ensure prolong of that particular students at university.

From the literature, there is an increasing interest in data mining and educational systems, resulting educational data mining as a new growing research community [4]. One of the popular data mining methods is popularly known as Association Rules Mining (ARM). It has been widely studied by knowledge discovery community [5] at discovering the interesting correlations, frequent patterns, associations or casual structures among sets of items in the data repositories. The problem of ARM was first introduced by Agrawal for market-basket analysis [6-8]. After the introduction of Apriori [6], many studies [13-29] have been look forward. There are two main stages involved before producing the association rules (ARs). First, find all frequent items from transactional database. Second, generate the common association rule from the frequent items.

Generally, an item is said to be frequent if it appears more than a minimum support threshold. These frequent items are then used to form the ARs. Besides that, confidence is another measure that always used in pair with the minimum support threshold. By definition, least item is an itemset which rarely found in the database but it may produce interesting and useful ARs. These type of rules are very meaningful in discovering rarely occurring but significantly important, such as air pollution detection, critical fault detections, network intrusions, etc. and their possible causes. For the past developments, many series of ARs mining algorithms are using the minimum supports-confidence framework to avoid the overloaded of ARs. The challenge is, by increasing or decreasing the minimum support or confidence values, the interesting rules might be missing or untraceable. Since the complexity of study, difficulties in algorithms [9] and it may require excessive computational cost, there are very limited attentions have been paid to discover the highly correlated least ARs. In term of relationship, both of frequent and least ARs have a different degree of correlation.

Highly correlated least ARs are referred to the itemsets that its frequency does not satisfy a minimum support but are very highly correlated. ARs are classified as highly correlated if it is positive correlation and in the same time fulfils a minimum degree of predefined correlation. Until this moment, statistical correlation technique has been successfully applied in the transaction databases [10], which to find relationship among pairs of items whether they are highly positive or negative correlated. As a matter of fact, it is not absolute true that the frequent items have a positive correlation as compared to the least items. In previous papers, we address the problem of mining least ARs with the objectives of discovering significant least ARs but surprisingly highly correlated [11, 12]. A new algorithm named Significant Least Pattern Growth (SLP-Growth) to extract these ARs is proposed [11]. The proposed algorithm imposes interval support to extract all least itemsets family first before continuing to construct a significant least pattern tree (SLP-Tree). The Critical Relative Support (CRS) and Correlation (Corr) measures for finding relationship between itemset is also embedded to this algorithm [11]. In this paper, we employed SLP-Growth algorithm and the measures for capturing interesting rules in student admission dataset. The dataset was taken from Division of Academic, Universiti Malaysia Terengganu for 2008/2009 intake students in computer science program. The results of this research will provide useful information for educators or higher university personnel authority to offer more relevant programs to the potential students rather than by chance or unguided technique.

The contributions of this paper are as follows. First, we employed the enhanced SLP-Growth algorithm and SLP-Tree to mine the association rules. Second, we applied the Critical Relative Support (CRS) and Correlation (Corr) measures against extracted association rules in quantifying its interestingness. Third, we performed the experiment with the real dataset called Student Enrolment dataset which is the records of preferred programs being selected by post-matriculation or post-STPM students of Malaysia.

The reminder of this paper is organized as follows. Section 2 describes the related work. Section 3 describes the basic concepts and terminology of ARs mining. Section 4

describes the employed method, SLP-Growth algorithm. This is followed by performance analysis through student admission dataset in section 5 and the results are presented in Section 6. Finally, conclusions of this work are reported in section 7.

2. Related Works

For the past decades, there are several efforts has been made to discover the scalable and efficient methods for mining frequent ARs. However, mining least ARs is still left behind. As a result, ARs that are rarely found in the database are pruned out by the minimum support-confidence threshold. As a matter of fact, the rarely ARs can also reveal the useful information for detecting the highly critical and exceptional situations. Zhou, *et al.,* [30] suggested a method to mine the ARs by considering only infrequent itemset. The drawback is, Matrix-based Scheme (MBS) and Hash-based scheme (HBS) algorithms are very expensive in term of hash collision. Ding [31] proposed Transactional Co-occurrence Matrix (TCOM for mining association rule among rare items. However, the implementation wise is quite complex and costly. Yun, *et al.,* [9] introduced the Relative Support Apriori Algorithm (RSAA) to generate rare itemsets. The challenge is, it takes similar time taken as performed by Apriori if the allowable minimum support is set to very low.

Koh, *et al.,* [32] suggested Apriori-Inverse algorithm to mine infrequent itemsets without generating any frequent rules. However, it suffers from candidate itemset generations and costly in generating the rare ARs. Liu, *et al.,* [33] proposed Multiple Support Apriori (MSApriori) algorithm to extract the rare ARs. In actual implementation, this algorithm is facing the "rare item problem". From the proposed approaches [9, 31–33], many of them are using the percentage-based approach to improve the performance as faces by the single minimum support based approaches. In term of measurements, Brin, *et al.,* [34] introduced objective measure called lift and chi-square as correlation measure for ARs. Lift compares the frequency of pattern against a baseline frequency computed under statistical independence assumption. Omiecinski [35] proposed two interesting measures based on downward closure property called all confidence and bond.

Lee, *et al.,* [36] suggested two algorithms for mining all confidence and bond correlation patterns by extending the pattern-growth methodology Han, *et al.,* [37]. In term of mining algorithms, Agrawal, *et al.,* [6, 7] proposed the first ARs mining algorithm called Apriori. The main bottleneck of Apriori is, it requires multiple scanning of transaction database and also generates huge number of candidate itemsets. Han, *et al.,* [38] suggested FP-Growth algorithm which amazingly can break the two limitations as faced by Apriori series algorithms. Currently, FP-Growth is one of the fastest approach and most benchmarked algorithms for frequent itemsets mining. It is derived based on a prefix tree representation of database transactions (called an FP-tree). Several domain applications have been adapted association rule mining technique [39-43] and one of them is in educational domain.

Recently, Educational Data Mining (EDM) has emerged as an important research area in order to resolve educational research issues [44]. Kumar *et al.* [45] enhanced the quality of students' performances at post graduation level via association rule mining. Garcial, *et al.,* [46] described a collaborative educational data mining tool based on association rule mining for the ongoing improvement of e-learning courses. Tair, *et al.,* [47] used association rule mining technique to analyze to improve graduate students' performance, and overcome the problem of low grades of graduate students. Chandra, *et al.,* [48] applied the association rule mining technique to identifies the students' failure patterns in order to improve the low capacity students' performances. Kularbphettong, *et al.,* [49] created a model of Student Motivation Behavior on e-learning based on association rule mining technique. Phankokkruad [50] captured optimal item-set from association rules to help the students improve their knowledge reach their full potential.

Radosav [51] employed WEKA as a tool to extract association rules to understand e-learning system and found this rules are very useful in the domain of education. Damaševičius [52] proposed a framework for mining educational data based on association rules for assessing student academic results and extracting recommendations for the improvement of course content. All this information provides a gold mine of educational data [53].

3. Essential Rudiments

3.1. Association Rules (ARs)

ARs were first introduced for market basket analysis to study customer purchasing patterns in retail stores [6, 7, 54]. Nowadays, ARs has been used in many applications or disciplines such as customer relationship management [38], image processing [55], webpage transactions [56] and mining air pollution data [57]. Typically, association rule mining is the process of discovering associations or correlation among itemsets in transaction databases, relational databases and data warehouses. There are two subtasks involved in ARs mining: generate frequent itemsets that satisfy the minimum support threshold and generate strong rules from the frequent itemsets.

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \leq |M| \leq |A|$ and each transaction can be identified by a distinct identifier TID.

Definition 1. *A set $X \subseteq I$ is called an itemset. An itemset with k-items is called a k-itemset.*

Definition 2. *The support of an itemset $X \subseteq I$, denoted $\mathrm{supp}(X)$ is defined as a number of transactions contain X.*

Definition 3. *Let $X, Y \subseteq I$ be itemset. An association rule between sets X and Y is an implication of the form $X \Rightarrow Y$, where $X \cap Y = \phi$. The sets X and Y are called antecedent and consequent, respectively.*

Definition 4. *The support for an association rule $X \Rightarrow Y$, denoted $\mathrm{supp}(X \Rightarrow Y)$, is defined as a number of transactions in D contain $X \cup Y$.*

Definition 5. *The confidence for an association rule $X \Rightarrow Y$, denoted $\mathrm{conf}(X \Rightarrow Y)$ is defined as a ratio of the numbers of transactions in D contain $X \cup Y$ to the number of transactions in D contain X. Thus*

$$conf(X \Rightarrow Y) = \frac{\sup p(X \Rightarrow Y)}{\sup p(X)}.$$

Definition 6. (Least Items). *An itemset X is called least item if $\alpha \leq \mathrm{supp}(X) \leq \beta$, where α and β is the lowest and highest support, respectively.*

The set of least item will be denoted as Least Items and

$$\text{Least Items} = \{X \subset I \mid \alpha \leq \mathrm{supp}(X) \leq \beta\}$$

Definition 7. (Frequent Items). *An itemset X is called frequent item if $\mathrm{supp}(X) > \beta$, where β is the highest support.*

The set of frequent item will be denoted as Frequent Items and

$$\text{Frequent Items} = \{X \subset I \mid \mathrm{supp}(X) > \beta\}$$

Definition 8. (Merge Least and Frequent Items). *An itemset X is called least frequent items if* $\mathrm{supp}(X) \geq \alpha$, *where* α *is the lowest support.*

The set of merging least and frequent item will be denoted as LeastFrequent Items and

$$\text{LeastFrequent Items} = \{X \subset I \mid \mathrm{supp}(X) \geq \alpha\}$$

LeastFrequent Items will be sorted in descending order and it is denoted as

$$\text{LeastFrequent Items}^{desc} = \left\{ \begin{array}{l} X_i \mid \mathrm{supp}(X_i) \geq \mathrm{supp}(X_j), \ 1 \leq i, j \leq k, \ i \neq j, \\ k = |\text{LeastFrequent Items}|, \ x_i, x_j \subset \text{LeastFrequent Items} \end{array} \right\}$$

Definition 9. (Ordered Items Transaction). *An ordered items transaction is a transaction which the items are sorted in descending order of its support and denoted as* t_i^{desc}, *where*

$$t_i^{desc} = \text{LeastFrequentItems}^{desc} \cap t_i, 1 \leq i \leq n, |t_i^{least}| > 0, |t_i^{frequent}| > 0.$$

An ordered items transaction will be used in constructing the proposed model, so-called LP-Tree.

Definition 10. (Significant Least Data). *Significant least data is one which its occurrence less than the standard minimum support but appears together in high proportion with the certain data.*

3.2. Critical Relative Support

Critical Relative Support (CRS) is introduced by Abdullah *et al.* [11] in an attempt to discover critical least association rules from the given dataset. This measure has been employed in various domain applications such as in medical, educational, environmental sciences, etc. To ensure that only significant rules are selected, a predefined minimum CRS (min-CRS) must be set up. Any rule that appears equal or more than min-CRS is classified as critical least association rules. Domain expert is employed to confirm the criticality and usefulness of these types of rules.

Definition 6. (Critical Relative Support). *A Critical Relative Support (CRS) is a formulation of maximizing relative frequency between itemset and their Jaccard similarity coefficient.*

The value of Critical Relative Support denoted as CRS and

$$\text{CRS}(I) = \max\left(\left(\frac{\mathrm{supp}(X)}{\mathrm{supp}(Y)} \right), \left(\frac{\mathrm{supp}(Y)}{\mathrm{supp}(X)} \right) \right) \times \left(\frac{\mathrm{supp}(X \Rightarrow Y)}{\mathrm{supp}(X) + \mathrm{supp}(Y) - \mathrm{supp}(X \Rightarrow Y)} \right).$$

CRS value is in a range of 0 and 1, and is determined by multiplying the highest value either supports of antecedent divide by consequence or in another way around with their Jaccard similarity coefficient. It is a measurement to show the level of CRS between combination of the both Least Items and Frequent Items either as antecedent or consequence, respectively.

3.3 Correlation Analysis

After the introduction of ARs, many researches including Brin, *et al.,* [17] had realized the limitation of the confidence-support framework. Utilizing this framework alone is quite impossible to discover the interesting ARs. Therefore, the correlation measure can be used as complimentary measure together with this framework. The correlation rule is a measure based on the minimum support, minimum confidence and correlation between itemsets *A* and *B*. There are many correlation measures applicable for ARs. One of the simplest correlation measures is Lift. The occurrence of itemset *A* is independence of the occurrence of itemset *B* if $P(X \cup Y) = P(X)P(Y)$; otherwise itemset *A* and *B* are dependence and correlated.

Definition 7. *The lift for an association rule* $X \Rightarrow Y$ *, denoted* $\text{lift}(X \Rightarrow Y)$ *is defined as a ratio of the confidence for an association rule* $X \Rightarrow Y$ *,* $\text{conf}(X \Rightarrow Y)$ *to the number of transactions in D contain X,* $\text{supp}(X)$ *. Thus*

$$\text{lift}(X,Y) = \frac{\text{conf}(X \Rightarrow Y)}{\text{supp}(Y)}$$

or

$$\text{lift}(X,B)Y = \frac{P(X \cap Y)}{P(X)P(Y)}$$

or

$$\text{lift}(X,Y) = \frac{P(Y \mid X)}{P(Y)}$$

The strength of correlation is measure from the lift value. If $\text{lift}(A,B) = 1$ or $P(B \mid A) = P(B)$ $\left(\text{or } P(A \mid B) = P(B)\right)$ then B and A are independent and there is no correlation between them. If $\text{lift}(A,B) > 1$ or $P(B \mid A) > P(B)$ $\left(\text{or } P(A \mid B) > P(B)\right)$, then A and B are positively correlated, meaning the occurrence of one implies the occurrence of the other. If $\text{lift}(A,B) < 1$ or $P(B \mid A) < P(B)$ $\left(\text{or } P(A \mid B) < P(B)\right)$, then A and B are negatively correlated, meaning the occurrence of one discourage the occurrence of the other. Since lift measure is not down-ward closed, it definitely will not suffer from the least item problem. Thus, least itemsets with low counts which per chance occur a few times (or only once) together can produce enormous lift values.

3.4 FP-Growth

Candidate set generation and tests are two major drawbacks in Apriori-like algorithms. Therefore, to deal with this problem, a new data structure called frequent pattern tree (FP-Tree) was introduced. FP-Growth was then developed based on this data structure and currently is a benchmarked and fastest algorithm in mining frequent itemset [36]. The advantages of FP-Growth are, it requires two times of scanning the transaction database. Firstly, it scans the database to compute a list of frequent items sorted by descending order and eliminates rare items. Secondly, it scans to compress the database into a FP-Tree structure and mines the FP-Tree recursively to build its conditional FP-Tree.

A simulation data [56] is shown in Table 1. Firstly, the algorithm sorts the items in transaction database with infrequent items are removed. Let say a minimum support is set to 3, therefore alphabets f, c, a, b, m, p are only kept. The algorithm scans the entire transactions start from T1 until T5. In T1, it prunes from {f, a, c, d, g, i, m, p} to {f, c, a, m, p, g}. Then, the algorithm compresses this transaction into prefix tree which f becomes the root. Each path on the tree represents a set of transaction with the same prefix. This process will execute recursively until the end of transaction. Once the complete tree has been built, then the next pattern mining can be easily performed.

Table 1. A Simulation Data

TID	Items
T1	a c m f p
T3	b f h j o
T4	b c k s p
T5	a f c e l p m n

4. Methodology

4.1. Algorithm Development

Determine Interval Support for least Itemset

Let I is a non-empty set such that $I = \{i_1, i_2; \cdots, i_n\}$, and D is a database of transactions where each T is a set of items such that $T \subset I$. An item is a set of items. A k-itemset is an itemset that contains k items. An itemset is said to be least if the support count satisfies in a range of threshold values called Interval Support (ISupp). The Interval Support is a form of ISupp (ISMin, ISMax) where ISMin is a minimum and ISMax is a maximum values respectively, such that $\text{ISMin} \geq \phi$, $\text{ISMax} > \phi$ and $\text{ISMin} \leq \text{ISMax}$. The set is denoted as L_k. Itemsets are said to be significant least if they satisfy two conditions. First, support counts for all items in the itemset must greater ISMin. Second, those itemset must consist at least one of the least items. In brevity, the significant least itemset is a union between least items and frequent items, and the existence of intersection between them.

Construct Significant Least Pattern Tree

A Significant Least Pattern Tree (SLP-Tree) is a compressed representation of significant least itemsets. This trie data structure is constructed by scanning the dataset of single transaction at a time and then mapping onto path in the SLP-Tree. In the SLP-Tree construction, the algorithm constructs a SLP-Tree from the database. The SLP-Tree is built only with the items that satisfy the ISupp. In the first step, the algorithm scans all transactions to determine a list of least items, LItems and frequent items, FItems (least frequent item, LFItems). In the second step, all transactions are sorted in descending order and mapping against the LFItems. It is a must in the transactions to consist at least one of the least items. Otherwise, the transactions are disregard. In the final step, a transaction is transformed into a new path or mapped into the existing path. This final step is continuing until end of the transactions. The problem of existing FP-Tree are it may not fit into the memory and expensive to build. FP-Tree must be built completely from the entire transactions before calculating the support of each item. Therefore, SLP-Tree is an alternative and more practical to overcome these limitations.

Generate Least Pattern Growth (LP-Growth)

SLP-Growth is an algorithm that generates significant least itemsets from the SLP-Tree by exploring the tree based on a bottom-up strategy. 'Divide and conquer' method is used to decompose task into a smaller unit for mining desired patterns in conditional databases, which can optimize the searching space. The algorithm will extract the prefix path sub-trees ending with any least item. In each of prefix path sub-tree, the algorithm will recursively execute to extract all frequent itemsets and finally built a conditional SLP-Tree. A list of least itemsets is then produced based on the suffix sequence and also sequence in which they are found. The pruning processes in SLP-Growth are faster than FP-Growth since most of the unwanted patterns are already cutting-off during constructing the SLP-Tree data structure. The complete SLP-Growth algorithm is shown in Figure 1.

```
1:    Read dataset, D
2:    Set Interval Support (ISMin, ISMax)
3:    for items, I in transaction, T do
4:        Determine support count, ItemSupp
5:    end for loop
6:    Sort ItemSupp in descending order, ItemSuppDesc
7:    for ItemSuppDesc do
8:        Generate List of frequent items, FItems > ISMax
9:    end for loop
10:   for ItemSuppDesc do
11:       Generate List of least items,  ISMin <= LItems < ISMax
```

```
12:  end for loop
13:  Construct Frequent and Least Items, FLItems = FItems U LItems
14:  for all transactions,T do
15:     if (LItems ∩ I in T > 0) then
16:        if (Items in T = FLItems) then
17:           Construct items in transaction in descending order,
TItemsDesc
18:        end if
19:     end if
20:  end for loop
21:  for TItemsDesc do
22:     Construct SLP-Tree
23:  end for loop
24:  for all prefix SLP-Tree do
25:     Construct Conditional Items, CondItems
26:  end for loop
27:  for all CondItems do
28:     Construct Conditional SLP-Tree
29:  end for loop
30:  for all Conditional SLP-Tree do
31:     Construct Association Rules, AR
32:  end for loop
33:  for all AR do
34:     Calculate Support and Confidence
35:     Apply Critical Relative Support (CRS)
36:     Apply Correlation (Corr)
37:  end for loop
```

Figure 1. SLP-Growth Algorithm

4.2. Weight Assignment

Apply Critical Relative Support (CRS) and Correlation (Corr)

The weighted ARs (ARs value) based on CRS and Corr are derived from the Definition 6 and Definition 7, respectively. The processes of generating weighted ARs are taken place after all patterns and ARs are completely produced.

Discovery Highly Correlated Least ARs

From the list of weighted ARs, the algorithm begins to scan all of them. ARs will be then categorized and counted as positive correlation, no correlation, and negative correlation and significant based on their meeting of the predefined thresholds.

5. Scenario on Capturing Rules

5.1. Dataset

The data was obtained from Division of Academic, Universiti Malaysia Terengganu in a text file and Microsoft excel format. There were 160 students involved and their identities were removed due to the confidentiality agreement. In the original set of data, it consists of 35 attributes and the detail information were explained in 10 tables which provided in Microsoft excel format. From the original flat file, the 8 chosen programs by the students are extracted according to the fix location. The actual location for each programs are based on the fix column range as presented in Table 2. There were in total of 822 bachelors programs offered in Malaysian public universities for July 2008/2009 students' intake. From this figure, 342 bachelor programs were selected by our 160 students and it can be generalized into 47 unique general fields. For simplicity, only 5 bachelor programs were extracted as illustrated in Table 3. In addition, SLP-Growth algorithm with lift measurement to determine the degree of correlation of association rules was employed. The total of 4,177 association rules was successfully extracted.

Table 2. Mapping of Students Chosen Programs' and its Column Range

Chosen Programs Description	Column range
First choice	2 – 6
Second choice	8 – 12
Third choice	14 – 18
Fourth choice	20 – 24
Fifth choice	26 – 30
Sixth choice	32 – 36
Seventh choice	38 – 42
Eighth choice	44 – 48

Table 3. Mapping a Part of Bachelor Programs' Offered, General Field and its Code

Program Description	Field	Code
Bachelor of Computer Science (Software Engineering)	Computer	14
Bachelor of Counselling	Counselling	15
Bachelor of Dental Surgery	Dentist	16
Bachelor of Science Education	Education	17
Bachelor of Engineering (Electrical)	Engineering	18

The existing dataset does not have information about field of the program. Thus, the information provided by the office website of Malaysian Qualification Agency, Ministry of High Learner Education Malaysia (http://www.mqa.gov.my) is used as a guideline to match between the programs offered by the respective university and its specific field (domain). For example, Bachelor in Computer Science (Software Engineering) is in a general field of Computer. Here, a flat file called "FieldData" is produced. This file contains program's field and program's code. They are split by a blank space. Table 4 presents the detail characteristics pertinent to Student Admission dataset and Figure 2 shows some portion from the flat file of the dataset.

Table 4. Characteristic of Student Admission Dataset

Dataset	Size	#Transactions	Average Transaction	# Items
Student Admission	3KB	160	6	132

```
FPM01  PG32  PG11  PM02  KS54  MG38  GG27  SS02  AF101B02SS880502086536SITI NORZUNAINAH BIN
FEE02 *UE02 *SE02 *ME02 *PE02 *YE02 *UE17 *GE02 *         SS790820025297RASHIDI BIN HAMZAH
FBK07 *RK07 *MK07 *KK07 *KS11 *PP43 *PG07 *GY06 *         SS870506085240NORAZLINDA BINTI MAT
FCK47  CK40  CK59  GP18 *CK12  CK13  BK81  KK08           SS860518386375NUR ANWAR FATIMI BIN
EGP04  MP04  DI13 *QA19 *HA19  HA20  KA14  PP02  JF101    SS860429235436NORHAYATI BINTI MOHD
FGG10  KS05  GG27  GS28  ES70 *HS27  ES08 *ES60 *JC101J01SS870820235630NORAZLINA NORDIN
EGP04  KA14  KB06  EA39 *QA19 *         BH102B02SS860911565206KHAIRUL NISA BT MOHD
FPK05 *PG11  DS55 *ES57 *PG27  PG09 *PG07 *GS04 *MA101J02TS870203145023MOHD ZAHID BIN AMIR
FTT26 *EC01 *MC01  UC01 *BC01  GC10  CC10 *TC10  PB102P01SS860719355428NOOR AQILAH BINTI HA
EGE02  UP09 *UP01 *UE02 *UE07  DE02 *UE17 *BP25          SS860404235632SITI ZURAIDAH BINTI
FGC10 *BC01 *CC25 *CC10 *CC14 *JC10  PC10 *PK17 *BH101B01SS870308295140IZZAN RUSMAWATI BINT
FGC10 *CC25 *CC10 *CC14 *BC01  PC04 *PK17  PC10 *DG101D01SS871222295478PAH SALMIZA BT SAYID
FTK01  MK01  SK01 *PK01  GS40 *HK01  KM11 *PK15 *BD302B01SS870617105646SURAYA BT. HADI DARS
FGS40  KS01  GG10  HS40 *GG15  PG27  WA05 *UP08          SS870402055397YEE WOEI SHYAN
FCC10 *JC10 *BC01 *PC10 *CC14  CC13 *GC10  KC03 *JB101J01SS871115235218NURULLIZA BINTI PANG
FGG10  GG27  SG02  PS02 *KS05  PG07  SS02  PG13          SS860418236628TAY SIN YEN
PGE02  SE02  KE00  ME00  ME02  EE02  EE23  PE02  JB201J01SS890628135064SITI KHADIJAH BINTI
FGG10  GS40  GG27  GG15  PG27  PP05  PG07  SG02  CJ101C01SS870427065364NURRUL ALIAA BINTI A
FES40 *GG37 *SK15 *KK88 *YK15 *CK40 *TK40 *YK48 *         SS870303085697NOOR AMIN BIN ABDUL
FPG09  GS04  DY01 *PG07  QG07 *KM06  PG17  KS04          SS870103075064SURIANI BINTI AHMAD
FRK07 *BK07 *ES63 *MK07 *GY06 *HS11 *KS11 *WS49 *BG101B02SS871013145596SHAHIRA BT SHAARI
FGG10  GS40  GG27  GS68 *GG37 *GS02  GG15  GS03  DH101D01SS860314296271MOHD FIRDAUS BIN AHM
EGP04  QA19 *HA19  HA20  YA16 *DA14 *DI13 *YA14 *NA101B01SS860725595778SALBIAH BINTI MAMAT
FRK07 *BK07 *MK07 *KS11 *GY06 *GS44 *SY06 *HS11 *RA101K01SS870722405026SITI NOOR BAYA BINTI
EKT35 *MT35 *GP18 *TP16 *BP23 *TP22 *TP09 *GP22 *CJ101C01SS870817065474NORHANIM BINTI SAID
FGS40  GG15  KS01  HS27  DB00 *UP08  EP13 *WS49          SS860115065120NURUL ASHIKIN ISMAIL
EGE02  UT08  UE02 *UP15  SE02  YE02 *WE07  UP15  PB101P01SS860127355846HASMIDAH BT MD KASA
```

Figure 2. A Portion of Original Student Admission Dataset

5.2. Design

The design for capturing interesting rules on in student admission dataset is described in the Figure 3.
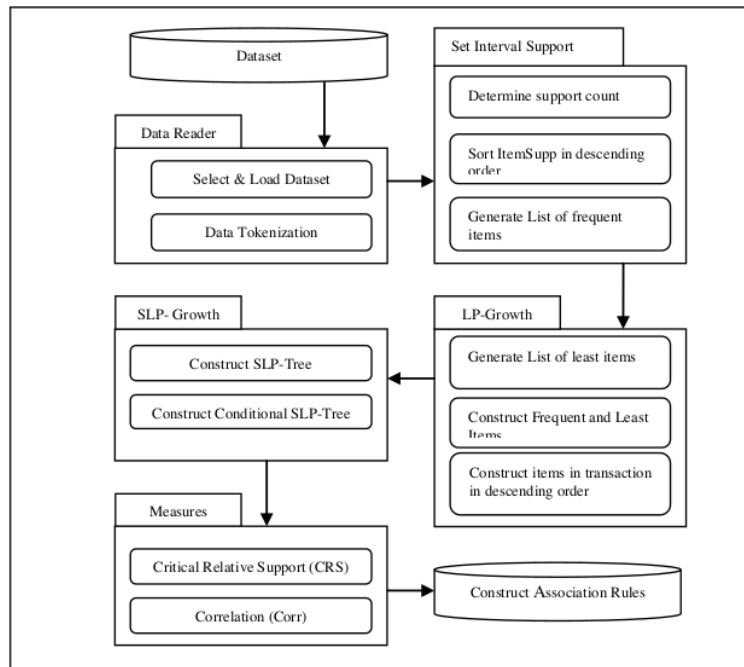

Figure 3. The Procedure of Mining Critical Least Association Rules

In order to capture the interesting rules and make a decision, the experiment using SLP-Growth method will be conducted on Intel® Core™ 2 Quad CPU at 2.33GHz speed with 4GB main memory, running on Microsoft Windows Vista. The algorithm has been developed using C# as a programming language. The mathematics anxiety dataset used and SLP-Growth produced in this model are in a format of flat file.

6. Results and Discussion

We evaluate the proposed algorithm to Student Enrolment dataset as in Table 2. To this, we have a dataset comprises the number of transactions (student) is 160 and the maximum number of items (attributes) is 8. By embedding FP-Growth algorithm, 3,768 ARs are produced. ARs are formed by applying the relationship of an item or many items to an item (cardinality: many-to-one). Figure 4 depicts the correlation's classification of interesting ARs. For this dataset, the rule is categorized as significant and interesting if it has positive correlation, confidence is 100% and CRS value should be equal to 1.0
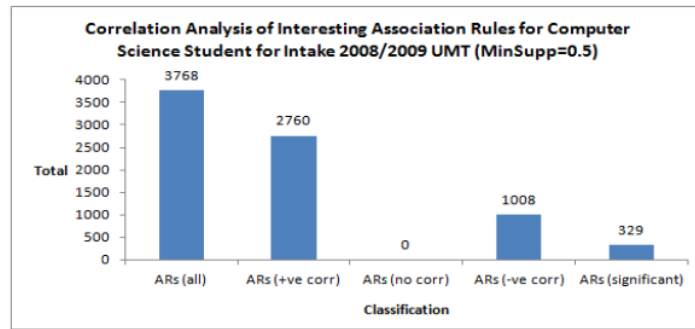
Figure 4. Classification of ARs using Correlation Analysis, only 8.73% from the Total of 3,768 ARs are Classified as Interesting ARs

Table 5 shows top 20 interesting ARs with numerous types of measurements. The highest correlation value from the selected ARs is 80.00 (No. 1 to 6). From these ARs, there are only one dominant of consequence items, item 4 (Agro-technology). In fact, item 4 only appears 1.25% from the entire dataset. Besides item 4, others consequent item that occur in to 20 interesting ARs is item 26. Item 26 is stand for "Human Resource". For item 26, it occurrences in the dataset is 2.50%. Table 1 also indicates that all interesting ARs have a value of CRS is equal to 1. Therefore, further analysis and study can be used to find out others interesting relationships such as academic performance, personality, attitude, *etc*.

Figure 5 depicts the correlation of interesting ARs based on several interval supports. The result indicates that CRS successfully in producing the less number of ARs as compared to the others measures. The typical support or confidence measure alone is not a suitable measure to be employed to discover the interesting ARs. Although, correlation measure can be used to capture the interesting ARs, it ratio is still nearly 12 times larger than CRS measure. Therefore, CRS is proven to be more efficient and outperformed the benchmarked measures for discovering the interesting ARs from the dataset. Generally, the total numbers of ARs are kept decreased when the predefined Interval Supports thresholds are increased.

Table 5. Top 20 of Highest Correlation of Interesting Association Rules Sorted in Descending Order of Correlation

No.	Association Rules	Supp	Conf	Corr	CRS
1	25 45 32 11 --> 2	0.62	100.00	80.00	1.00
2	45 32 11 --> 2	0.62	100.00	80.00	1.00
3	17 39 8 --> 4	0.62	100.00	80.00	1.00
4	28 17 39 8 --> 4	0.62	100.00	80.00	1.00
5	21 16 17 39 --> 4	0.62	100.00	80.00	1.00
6	25 17 39 8 --> 4	0.62	100.00	80.00	1.00
7	45 21 39 38 --> 26	0.62	100.00	40.00	1.00
8	33 38 41 --> 26	0.62	100.00	40.00	1.00
9	45 41 15 --> 26	0.62	100.00	40.00	1.00
10	25 45 38 15 --> 26	0.62	100.00	40.00	1.00
11	21 33 41 --> 26	0.62	100.00	40.00	1.00
12	16 34 15 --> 26	0.62	100.00	40.00	1.00
13	21 33 38 15 --> 26	0.62	100.00	40.00	1.00
14	36 34 15 --> 26	0.62	100.00	40.00	1.00
15	25 34 15 --> 26	0.62	100.00	40.00	1.00

16	45 21 33 38 --> 26	0.62	100.00	40.00	1.00
17	25 45 33 41 --> 26	0.62	100.00	40.00	1.00
18	25 45 21 41 --> 26	0.62	100.00	40.00	1.00
19	25 45 33 38 --> 26	0.62	100.00	40.00	1.00
20	25 33 15 --> 26	0.62	100.00	40.00	1.00

Interesting Association Rules for Computer Science Student for Intake 2008/2009 (UMT)

	< 0.7	0.7 - 2.0	2.1 - 5.0	5.1 - 7.0	> 7.0
ARs	2469	806	346	45	102
ARs (+ve corr)	1938	557	195	18	52
ARs (significant)	426	39	5	1	1
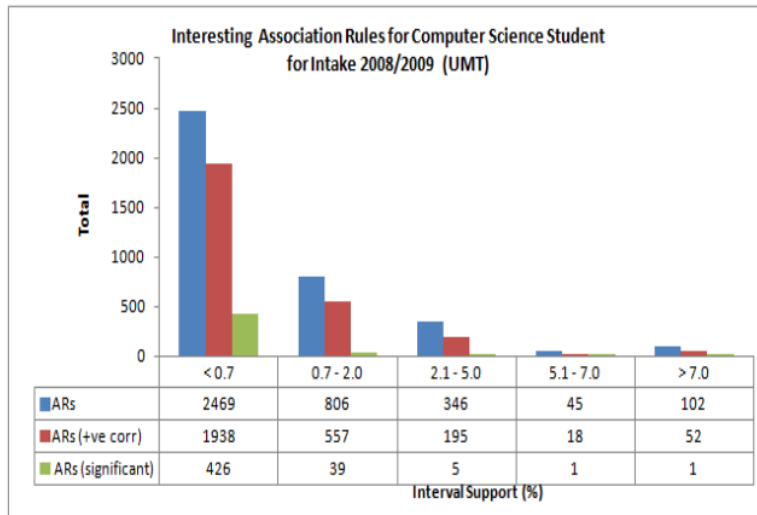
Figure 5. Correlation Analysis of Interesting ARs using Variety Interval Supports

7. Conclusion

Data mining is an influential new technology with great potential to reveal a new knowledge from data repositories. It has been applied in various domain applications. For the less than a decade, there is an increasing interest in data mining and educational systems, making educational data mining as a new growing research community [4]. In data mining, one of the most popular techniques is Association Rules Mining (ARM). ARM is a technique of discovering dependency or relationship between seemingly unrelated data from data repository. In order to quantify this affiliation, there are several interesting measures have been put forward. Therefore, in this paper we had successfully applied an enhanced association rules mining method, so called SLP-Growth (Significant Least Pattern Growth) proposed by [11] for capturing interesting rules in student enrollment admission dataset. There are two interesting measures embedded in the algorithm known as Critical Relative Support (CRS) and Correlation (Corr). The dataset for the experiment was taken from Division of Academic, Universiti Malaysia Terengganu (UMT) that related to the intake student of the year 2008/2009. It is found that SLP-Growth method is suitable to mine the interesting rules which provide more meaningful information. Based on the results, educators or the university' higher authority personnel can obtain guideline from the rules captured. Moreover, the results of this research will provide useful information to make a decision in offering appropriate programs for students. It also can be helpful to reduce the risk of drop-out students among computer science student at UMT.

Acknowledgment

References

[1] E. Yukselture and F. A. Inan, "Examining the Factors Affecting Student Dropout in an Online Certificate Program", Turkish Online Journal of Distance Education, vol. 7, no. 3, (2006) July, pp. 76-88.

[2] S. Mohammad and A.Z. El-Masri, "Factors Affecting Dropouts Students in Arab Open University - Bahrain Branch", International Journal of Science and Technology, vol. 2, no. 7, (2012) July, pp. 435-442.

[3] E. Stearns and E. J. Glennie, "When and Why Dropouts Leave High School", Youth Society, SAGE Journal, vol. 38, no. 1, (2006) September, pp. 29-57.

[4] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, vol. 33, (2007), pp. 135–146.

[5] A. Ceglar and J. F. Roddick, "Association mining", ACM Computing Surveys, vol. 38, no. 2, (2006), pp. 1–42.

[6] R. Agrawal, T. Imielinski and A. Swami, "Database Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 5, no. 6, (1993), pp. 914–925.

[7] R. Agrawal, T. Imielinski and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of the ACM SIGMOD '93 International Conference on the Management of Data, (1993), pp. 207–216.

[8] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), (1994), pp. 487–499.

[9] H. Yun, D. Ha, B. Hwang and K. H. Ryu, "Mining Association Rules on Significant Rare Data Using Relative Support", The Journal of Systems and Software, vol. 67, no. 3, (2003), pp. 181-19.

[10] H. Xiong, S. Shekhar, P. N. Tan and V. Kumar, "Exploiting A Support-Based Upper Bond Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs", In The Proceeding of ACM SIGKDD, (2004).

[11] Z. Abdullah, T. Herawan and M. M. Deris, "Mining Significant Least Association Rules using Fast SLP-Growth Algorithm", In T. H. Kim and H. Adeli (Eds.): AST/UCMA/ISA/ACN 2010, Lecture Notes in Computer Science, Springer-Verlag, vol. 6059, (2010), pp. 324–336.

[12] Z. Abdullah, T. Herawan and M. M. Deris, "Scalable Model for Mining Critical Least Association Rules", In Rongbo Zhu, et al., ICICA, Lecture Notes in Computer Science, Springer-Verlag, vol. 6377, (2010), pp. 509-516.

[13] Z. Abdullah, T. Herawan, A. Noraziah and M. M. Deris, "Extracting Highly Positive Association Rules from Students' Enrollment Data", Procedia Social and Behavioral Sciences, vol. 28, (2011), pp. 107–111.

[14] Z. Abdullah, T. Herawan, A. Noraziah and M. M. Deris, "Mining Significant Association Rules from Educational Data using Critical Relative Support Approach", Procedia Social and Behavioral Sciences, vol. 28, (2011), pp. 97–101.

[15] Z. Abdullah, T. Herawan and M. M. Deris, "An Alternative Measure for Mining Weighted Least Association Rule and Its Framework", In J. M. Zain, et al., (Eds.): ICSECS, Communication of Computer and Information Sciences, Springer-Verlag, vol. 188, no. II, (2011), pp. 475–485.

[16] Z. Abdullah, T. Herawan and M. M. Deris, "Visualizing the Construction of Incremental Disorder Trie Itemset Data Structure (DOSTrieIT) for Frequent Pattern Tree (FP-Tree)", In H. B. Zaman, et al., (Eds.): IVIC, Lecture Notes in Computer Science, Springer-Verlag, vol. 7066, (2011), pp. 183–195.

[17] T. Herawan, I. T. R. Yanto and M. M. Deris, "Soft set approach for maximal association rules mining", In D. Ślęzak, et al., (Eds.): DTA, Communication of Computer and Information Sciences, Springer-Verlag, 64, (2009), pp. 163–170.

[18] T. Herawan, I. T. R. Yanto and M. M. Deris, "SMARViz: Soft maximal association rules visualization", In H. Badioze Zaman, et al., (Eds.): IVIC, Lecture Notes in Computer Science, Springer-Verlag, vol. 5857, (2009), pp. 664–674.

[19] T. Herawan and M. M. Deris, "A soft set approach for association rules mining", Knowledge Based Systems, vol. 24, no. 1, (2011), pp. 186–195.

[20] T. Herawan, P. Vitasari and Z. Abdullah, "Mining interesting association rules of student suffering mathematics anxiety", In J. M. Zain, et al., (Eds.): ICSECS, Communication of Computer and Information Sciences, Springer-Verlag, vol. 188, no. II, (2011), pp. 495–508.

[21] Z. Abdullah, T. Herawan and M. M. Deris, "Efficient and Scalable Model for Mining Critical Least Association Rules", In a special issue from AST/UCMA/ISA/ACN 2010, Journal of The Chinese Institute of Engineer, Taylor and Francis, vol. 35, no. 4, (2012) June 27, pp. 547–554.

[22] T. Herawan, Z. Abdullah, A. Noraziah, M. M. Deris and J. H. Abawajy, "IPMA: Indirect Patterns

Mining Algorithm", In N. T. Nguyen, *et al.*, (Eds.): ICCCI, Advanced Methods for Computational Collective Intelligence Studies in Computational Intelligence, Springer-Verlag, vol. 457, **(2012)**, pp. 187–196.

[23] T. Herawan, Z. Abdullah, A. Noraziah, M. M. Deris and J. H. Abawajy, "EFP-M2: Efficient Model for Mining Frequent Patterns in Transactional Database", In N. T. Nguyen, *et al.*, (Eds.): ICCCI, Lecture Notes in Computer Science, Springer-Verlag, vol. 7654, **(2012)**, pp. 29–38.

[24] N. Ahmad, Z. Abdullah, T. Herawan and M. M. Deris, "Scalable technique to discover items support from trie data structure", In B. Liu, *et al.,* (Eds.): ICICA, Lecture Notes in Computer Science, Springer-Verlag, vol. 7473, **(2012)**, pp. 500–507.

[25] N. Ahmad, Z. Abdullah, T. Herawan and M. M. Deris, "WLAR-Viz: Weighted Least Association Rules Visualization", In B. Liu, *et al.,* (Eds.): ICICA, Lecture Notes in Computer Science, Springer-Verlag, vol. 7473, **(2012)**, pp. 592–600.

[26] T. Herawan and Z. Abdullah, "CNAR-M: A Model for Mining Critical Negative Association Rules", In Zhihua Cai, *et al.,* (Eds): ISICA, Communications in Computer and Information Science, Springer-Verlag, vol. 316, **(2012)**, pp. 170–179.

[27] Z. Abdullah, T. Herawan, N. Ahmad and M. M. Deris, "DFP-Growth: An efficient algorithm for mining pattern in dynamic database", In B. Liu, *et al.,* (Eds.): ICICA, Lecture Notes in Computer Science, Springer-Verlag, vol. 7473, **(2012)**, pp. 51–59.

[28] Z. Abdullah, T. Herawan and M. M. Deris, "Detecting Critical Least Association Rules in Medical Databasess", International Journal of Modern Physics: Conference Series, World Scientific, vol. 9, **(2012)**, pp. 464–479.

[29] T. Herawan, Z. Abdullah, A. Noraziah, M. M. Deris and J. H. Abawajy, "EFP-M2: Efficient Model for Mining Frequent Patterns in Transactional Database", In N. T. Nguyen, *et al.,* (Eds.): ICCCI, Lecture Notes in Computer Science, Springer-Verlag, vol. 7654, **(2012)**, pp. 29–38.

[30] L. Zhou and S. Yau, "Assocation Rule and Quantative Association Rule Mining Among Infrequent Items", The Proceeding of ACM SIGKDD'07, **(2007)**.

[31] J. Ding, "Efficient Association Rule Mining among Infrequent Items", Ph. D Thesis, University of Illinois at Chicago, **(2005)**.

[32] Y. S. Koh and N. Rountree, "Finding Sporadic Rules using Apriori-Inverse", Lecture Notes in Computer Science, Springer, vol. 3518, **(2005)**, pp. 97–106.

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

[34] S. Brin, R. Motwani and C. Silverstein, "Beyond Market Basket: Generalizing ARs to Correlations", Special Interest Group on Management of Data (SIGMOD), **(1997)**, pp. 265–276.

[35] E. Omniecinski, "Alternative Interest Measures for Mining Associations", IEEE Trans. Knowledge and Data Engineering, vol. 15, **(2003)**, pp. 57–69.

[36] Y.-K. Lee, W.-Y. Kim, Y. D. Cai and J. Han, "CoMine: Efficient Mining of Correlated Patterns", The Proceeding of ICDM'03, **(2003)**.

[37] J. Han, H. Pei and Y. Yin, "Mining Frequent Patterns without Candidate Generation", The Proceeding of SIGMOD'00, ACM Press, **(2000)**.

[38] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd ed., **(2006)**.

[39] Z. Zhang and C. Guo, "Association Rules Evaluation by a Hybrid Multiple Criteria Decision Method", International Journal of Knowledge and Systems Sciences, vol. 2, no. 3, **(2011)**, 14-25.

[40] K. Chihara and Y. Nakamori, "Clarification of Abilities and Qualities of Knowledge Coordinators: The Case of Regional Revitalization Projects", International Journal of Knowledge and Systems Sciences, vol. 2, no. 1, **(2011)**, pp. 1-16.

[41] T. Herawan, P. Vitasari and Z. Abdullah, "Mining Interesting Association Rules of Students Suffering Study Anxieties Using SLP-Growth Algorithm", International Journal of Knowledge and Systems Sciences, vol. 3, no. 2, **(2012)**, pp. 24-41.

[42] Y. Zhang, L. Zhang, Y. Lliu and Y. Shi, "Mining intelligent knowledge from a two-phase association rules mining", Int. J. Data Mining, Modelling and Management, vol. 2, no. 4, **(2010)**, pp. 403-419.

[43] G. Yang, K. Shimada, S. Mabu, K. Hirasawa and J. Hu, "Mining Generalized Association Rules with Ontology Using Genetic Network Programming", International Journal of Knowledge and Systems Sciences, vol. 3, no. 4, **(2006)**, pp. 43-50.

[44] R. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions", Journal of Educational Data Mining, vol. 1, no. 1, **(2010)**, pp. 3–17.

[45] V. Kumar and A. Chadha, "Mining Association Rules in Student's Assessment Data", IJCSI, International Journal of Computer Science Issues, vol. 9, no. 5, **(2012)** September 3, pp. 211-216.

[46] E. Garcia, C. Romero, S. Ventura and C. Castro, "An architecture for making recommendations to courseware authors using association rule mining and collaborative filtering", User Modeling and User-Adapted Interaction: The Journal of Personalization Research, vol. 19, **(2009)**, pp. 99–132.

[47] M. M. A Tair and A. M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, vol. 2, no. 2, **(2012)**, pp. 140-146.

[48] E. Chandra and K. Nandhini, "Knowledge Mining from Student Data", European Journal of Scientific Research, vol. 47, no. 1, **(2010)**, pp. 156-163.

[49] K. Kularbphettong, P. Waraporn and C. Tongsiri, "Analysis of Student Motivation Behavior on e-Learning Based on Association Rule Mining", World Academy of Science, Engineering and Technology, vol. 66, (2012), pp. 502-505.

[50] M. Phankokkruad, "Association rules for data mining in item classification algorithm: Web service approach", Second International Conference of Digital Information and Communication Technology and it's Applications (DICTAP), (2012), pp. 463 – 468.

[51] D. Radosav, E. Brtka and V. Brtka, "Mining Association Rules from Empirical Data in the Domain of Education", Int. J Comput Commun, ISSN 1841-9836, vol. 7, no. 5, (2010), pp. 933-944.

[52] R, Damaševičius, "Analysis of Academic Results for Informatics Course Improvement Using Association Rule Mining Information Systems Development", Springer Science, (2010), pp. 357-363.

[53] J. Mostow, J. Beck, H. Cen, A. Cuneo, E. Gouvea and C. Heiner, "An educational data mining tool to browse tutor–student interactions: Time Will Tell!", In Proceedings of the workshop on educational data mining, (2005), pp. 15–22.

[54] C. C. Aggrawal and P. S. Yu, "A New Framework For Item Set Generation", In Proceedings of the ACMPODS Symposium on Principles of Database Systems, Seattle, Washington, (1998).

[55] W. H. Au and K. C. C. Chan, "Mining Fuzzy ARs In A Bank-Account Database", IEEE Transactions on Fuzzy Systems, vol. 11, no. 2, (2003), pp. 238–248.

[56] H. Li, Y. Wang, D. Zhang, M. Zhang and E. Y. Chang, "Pfp: Parallel Fp-Growth for Query Recommendation", In Proceedings of RecSys, (2008), pp. 107–114.

[57] M. D. Mustafa, N. F. Nabila, D. J. Evans, M. Y. Saman and A. Mamat, "Association rules on Significant Rare Data using Second Support", International Journal of Computer Mathematics, vol. 83, no. 1, (2006), pp. 69–80.

a Method of Discovering Interesting assocation rules from student admision dataset

a Method of Discovering Interesting assocation rules from student admision dataset