Multidimensional Fuzzy Association Rules for Developing Decision Support System at Petra Christian University

Yulia[1], Siget Wibisono[2], Rolly Intan[1]

[1,2]Petra Christian University, Surabaya, Indonesia
[1]{yulia, rintan}@petra.ac.id, [2]m26411074@john.petra.ac.id

Abstract. Academic records of student candidates and students of Petra Christian University (PCU) which have been stored so far have not been used to generate information. PCU's top-level management needs a way to generate information from the records. The generated information is expected to support the decision-making process of top-level management.
Before starting the application development, analysis and design of the student academic records and the needs of top-level management are done. The design stage produces a number of modeling that will be used to create the application. The final result of the development is an application that can generate information using multidimensional fuzzy association rules.

Keywords. Application, Data Mining, Decision Support System, Multidimensional Fuzzy Association Rules

1 Introduction

During this time, PCU has stored academic records of student candidates who enroll in PCU, such as math and english grades at their schools. In addition, after entering the university, PCU will save GPA of all students.

Academic records of student candidates and students that have been kept, have not been used to produce valuable information. PCU's top-level management needs a way to generate information from the records. The generated information is expected to support the decision-making process of top-level management.

With academic records of student candidates and students, information can be generated in the form of relationship between students' data using multidimensional fuzzy association rules. The students' data that can be used are schools, math, and english grade in their schools, specialization (science, social, literature, etc.), GPA, faculty, majors, gender, religion, and batch. Therefore, PCU need a software that can generate information needed by top-level management related to academic records of student candidates and students.

2 Data Mining

Data mining is one of the most important steps of the knowledge discovery in databases process. It is considered as significant subfield in knowledge management. Research in data mining continues growing in business and in learning organization

over coming decades[8]. Data mining is a process of extraction of useful information and patterns from huge data. It is also known as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis[9].

The development of Information Technology has generated great amount of databases and huge data in various areas. The research in databases and information technology has resulted in approach to store and manipulate this precious data for further decision making. The important reason that attracted many attentions in information technology and the discovery of meaningful information from large collections of data industry towards field of "Data mining" is due to the perception of "we are data rich but information poor". There is huge volume of data but we hardly able to generate them in to meaningful information and knowledge for decision making process in business[10].

Data mining derives its name from the similarities between finding valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for valuable ore. Both processes require either sifting through a great amount of material, and intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business advantages and opportunities[10].

3 Multidimensional Association Rules

Association rule finds interesting association or correlation relationship among a large data set of items [1,2]. The discovery of interesting association rules can support decision making process.

Multidimensional association rules are association rules that involve two or more dimensions or predicates. Conceptually, a multidimensional association rule, $A \Rightarrow B$ consists of A and B as two datasets, called premise and conclusion, respectively.

Formally, A is a dataset consisting of several distinct data, where each data value in A is taken from a distinct domain attribute in D as given by

$$A = \{a_j \mid a_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_A \subseteq D$ is a set of domain attributes in which all data values of *A* come from.

Similarly,

$$B = \{b_j \mid b_j \in D_j, \text{ for some } j \in \mathrm{N}_n\},$$

where, $D_B \subseteq D$ is a set of domain attributes in which all data values of *B* come from.

For example, database of medical track record patients is analyzed for finding association (correlation) among diseases taken from the data of complicated several diseases suffered by patients in a certain time. Additional related information regarding the identity of patients, such as *age*, *occupation*, *sex*, *address*, *blood type*,

etc., may have a correlation to the illness of patients. Considering each data attribute as a predicate, it can therefore be interesting to mine association rules containing *multiple* predicates, such as:

Rule-1:

$Age(X, \text{"60"}) \wedge Smk(X, \text{"yes"}) \Rightarrow Dis(X, \text{"Lung Cancer"})$,

where there are three predicates, namely *Age*, *Smk* (*smoking)* and *Dis (disease)*. Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules.

From Rule-1, it can be found that A={60, *yes*}, B={Lung Cancer}, D_A={*age*, *smoking*} and D_B={*disease*}.

Considering $A \Rightarrow B$ is an interdimension association rule, it can be proved that $|D_A| \neq |A|$, $|D_B| \neq |B|$ and $D_A \cap D_B = \varnothing$.

Support of A is then defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{r} \tag{1}$$

where r is the number of records or tuples (see Table 1, r=12).

Alternatively, r in (1) may be changed to $|Q(D_A)|$ by assuming that records or tuples, involved in the process of mining association rules are records in which data values of a certain set of domain attributes, D_A, are not null data. Hence, (1) can be also defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|}{|Q(D_A)|} \tag{2}$$

where $Q(D_A)$, simply called *qualified data* of D_A, is defined as a set of record numbers (t_i) in which all data values of domain attributes in D_A are not null data. Formally, $Q(D_A)$ is defined as follows.

$$Q(D_A) = \{t_i \mid d_{ij} \neq null, \forall D_j \in D_A\} \tag{3}$$

Similarly,

$$\text{supp}(B) = \frac{|\{t_i \mid d_{ij} = b_j, \forall b_j \in B\}|}{|Q(D_B)|} \tag{4}$$

Similarly, $\text{support}(A \Rightarrow B)$ is given by

$$\begin{aligned} \text{supp}(A \Rightarrow B) &= \text{supp}(A \cup B) \\ &= \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|Q(D_A \cup D_B)|} \end{aligned} \tag{5}$$

where $Q(D_A \cup D_B) = \{t_i \mid d_{ij} \neq null, \forall D_j \in D_A \cup D_B\}$ $\text{conf}(A \Rightarrow B)$ as a measure of certainty to assess the validity of $A \Rightarrow B$ is calculated by

$$\text{conf}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} = c_j, \forall c_j \in A \cup B\}|}{|\{t_i \mid d_{ij} = a_j, \forall a_j \in A\}|} \tag{6}$$

A and *B* in the previous discussion are datasets in which each element of *A* and *B* is an atomic crisp value. To provide a generalized multidimensional association rules, instead of an atomic crisp value, we may consider each element of the datasets to be a dataset of a certain domain attribute. Hence, *A* and *B* are sets of set of data values or sets of datasets. For example, the rule may be represented by

Rule-2:

$age(X, \text{"20...60"}) \wedge smoking(X, \text{"yes"}) \Rightarrow$

$disease(X, \text{"bronchitis, lung cancer"})$,

where *A*={{20…60}, {yes}} and B={{bronchitis, lung cancer}}.

Simply, let *A* be a generalized dataset. Formally, *A* is given by

$$A = \{A_j \mid A_j \subseteq D_j, \text{ for some } j \in \mathrm{N}_n\}.$$

Corresponding to (2), support of *A* is then defined by:

$$\text{supp}(A) = \frac{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|}{|Q(D_A)|} \tag{7}$$

Similar to (5),

$$\begin{aligned} \text{supp}(A \Rightarrow B) &= \text{supp}(A \cup B) \\ &= \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|Q(D_A \cup D_B)|} \end{aligned} \tag{8}$$

Finally, $\text{conf}(A \Rightarrow B)$ is defined by

$$\text{conf}(A \Rightarrow B) = \frac{|\{t_i \mid d_{ij} \subseteq C_j, \forall C_j \in A \cup B\}|}{|\{t_i \mid d_{ij} \subseteq A_j, \forall A_j \in A\}|} \tag{9}$$

To provide a more meaningful association rule, it is necessary to utilize *fuzzy sets* over a given database attribute called *fuzzy association rule* as discussed in [4,5]. Formally, given a crisp domain *D*, any arbitrary fuzzy set (say, fuzzy set *A*) is defined by a membership function of the form [2,3]:

$$A : D \rightarrow [0,1]. \tag{10}$$

To provide a more generalized multidimensional association rules, we may consider *A* and *B* as sets of fuzzy labels[6]. Simply, *A* and *B* are called fuzzy datasets.

Rule-2 is an example of such rules, where *A*={*young*, *yes*} and B={*bronchitis*}. Here *young*, *yes* and *bronchitis* are considered as fuzzy lables. A fuzzy dataset is a set of fuzzy lables/ data consisting of several distinct fuzzy labels, where each fuzzy label is represented by a fuzzy set on a certain domain attribute. Let *A* be a fuzzy dataset. Formally, *A* is given by

$$A = \{A_j \mid A_j \in \mathrm{F}(D_j),\ \text{for some}\ \ j \in \mathrm{N}_n\},$$

where $\mathrm{F}(D_j)$ is a fuzzy power set of D_j, or in other words, A_j is a fuzzy set on D_j.

Corresponding to (7), support of *A* is then defined by:

$$\mathrm{supp}(A) = \frac{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}}{|Q(D_A)|} \tag{11}$$

Similar to (5),

$$\begin{aligned} \mathrm{supp}(A \Rightarrow B) &= \mathrm{supp}(A \cup B) \\ &= \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{|Q(D_A \cup D_B)|} \end{aligned} \tag{12}$$

$\mathrm{conf}(A \Rightarrow B)$ is defined by

$$\mathrm{conf}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\}} \tag{13}$$

The correlation betwen two fuzzy datasets can be defined by the following definition.

$$\mathrm{corr}(A \Rightarrow B) = \frac{\sum_{i=1}^{r} \inf_{C_j \in A \cup B} \{\mu_{C_j}(d_{ij})\}}{\sum_{i=1}^{r} \inf_{A_j \in A} \{\mu_{A_j}(d_{ij})\} \times \inf_{B_k \in B} \{\mu_{B_k}(d_{ik})\}} \tag{14}$$

4 Research Methodology

4.1 Problems Analysis

There are several problems faced by PCU, such as:

1. PCU's top-level management takes decisions for the promotion or cooperation purpose based solely on estimates and habits, has not taken advantage of the existing academic records.
2. PCU's Faculties/Majors Promotion Team has not equipped with information or facts about the academic condition of PCU's students while promoting faculties/majors to high schools.
3. There is no feature in the current academic information system that can show the relationship between students' data.

4.2 Requirements Analysis

From the problems listed above, it can be concluded that the PCU's top-level management requires a computer-based system to assist in generating PCU's students academic records, that is a data mining-based information systems that can produce association rules of students' attributes. This system obtains data from the ETL process and has a multidimensional concept that shows the relationships between students' attributes. The dimensions used are schools, math and english grade in their schools, specialization (science, social, literature, etc.), GPA, faculty, majors, gender, religion, and batch.

4.3 Extract, Transform, and Load

Extract, Transform, and Load (ETL) is a function that integrates data and involves extracting data from sources, transforming it to be more valid, and loading it into a data warehouse[7]. This process begins by importing the data from the database. The imported data is religions, majors, schools, specializations, student candidates, students, and student admissions. Next, the imported data is transformed into more valid data and loaded into data warehouse.

4.4 Determination of Fuzzy Values

Determination of fuzzy values is done by establishing a group fuzzy set. First, user must input the name and choose the attribute, such as religions, majors, schools, GPA, math grade, etc. Next, user can make as many fuzzy sets as he/she wants inside the group fuzzy set made. User need to fill the name and the description of the fuzzy set. There are two types of fuzzy set based on the attribute of the group fuzzy set, numerical and non-numerical. For numerical, user can input as many points as he/she wants to form fuzzy membership function. A point includes crisp value and the membership degree of the crisp value to the fuzzy set. For non-numerical, user must input membership degree for every members of the attribute. Flowchart for determination of fuzzy values can be seen on Figure 1.

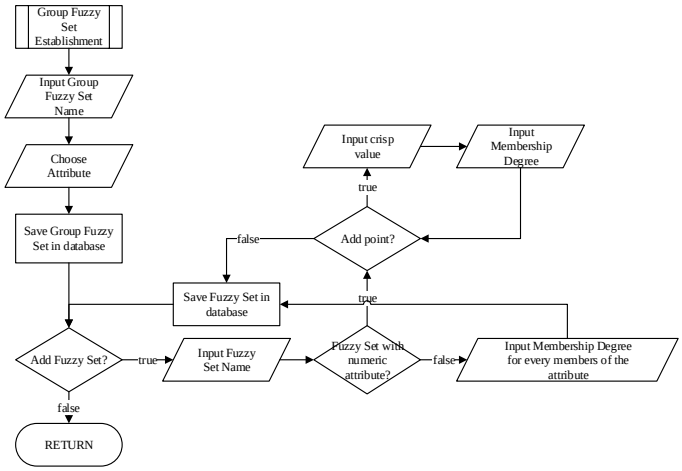

Fig. 1. Flowchart for Determination of Fuzzy Values

4.5 Customization of Fuzzy Association Rules

Customization of fuzzy association rules is done to generate fuzzy association rules report to support the decision-making process of top-level management. First, user must input the name and choose the attributes that will be used to generate the rules. After choosing the attributes, user must choose the group fuzzy set(s) of the attributes. Next, the application will generate the rules and save them in database. The user can see the whole report and filter the rules based on the support, confidence, and correlation value of the rules. Flowchart for customization of fuzzy association rules can be seen on Figure 2.

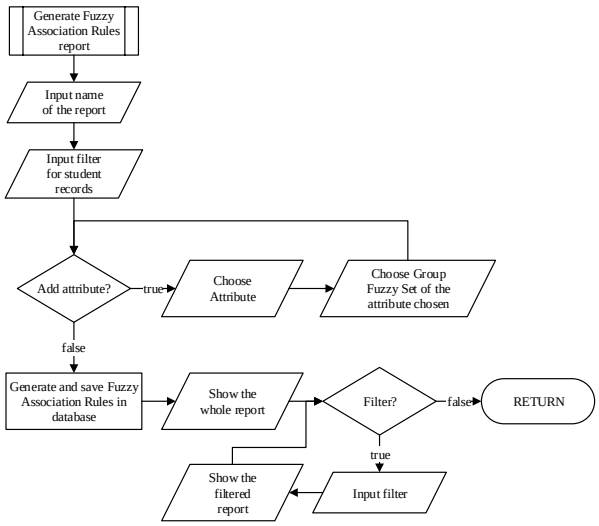

Fig. 2. Flowchart for Fuzzy Association Rules

5 Results

A test is conducted to prove the accuracy of the developed application to calculate support, confidence, and correlation of the multidimensional fuzzy association rules generated. The test is started from a given simple academic records of students with three attributes, such as major, math grade, Grade Point Average (GPA) as shown in Table 1.

Table 1. Academic Records of Students

Student	Major	Math grade	GPA
1	English Literature	74	3.34
2	Civil Engineering	75	3.41
3	Civil Engineering	90	3.9
4	Interior Design	86	3.75
5	Interior Design	78	3.45
6	Business Management	76	3.23
7	Business Management	68	3.35
8	Business Management	89	3.56
9	Informatics Engineering	91	3.84
10	Informatics Engineering	71	3.01
11	Science Communication	79	2.71
12	Science Communication	76	3.03

The test is conducted using three attributes, such as major, math grade, and GPA. First, we must determine how to convert each crisp value into fuzzy value for every attributes. Major is a non-numerical attribute, so we must determine the fuzzy value for each major. For example, we make a group fuzzy set for major named 2014 which has a fuzzy set named Engineering. Inside this fuzzy set, we determine business management has a membership degree of 0.2, civil engineering has a membership degree of 1, and so on as shown on Figure 3.

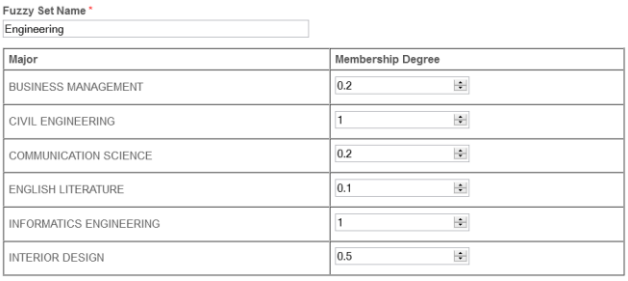

Fig. 3. Input for Major Fuzzy Values

Math grade is a numerical attribute, so that the fuzzy value of math grade will be calculated through a fuzzy membership function which is formed from the points stored in the fuzzy set. For example, we make a group fuzzy set for math grade named 2014 which has a fuzzy set named High. Inside this fuzzy set, we determine math grade of 0 has a membership degree of 0, math grade of 75 has a membership degree of 0, math grade of 95 has a membership degree of 1, and math grade of 100 has a membership degree of 1 as shown on Figure 4.

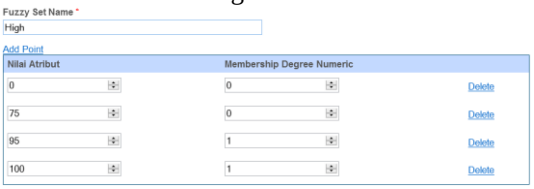

Fig. 4. Input for Math Grade Fuzzy Membership Function

These four points will form fuzzy membership function as shown on Figure 5.

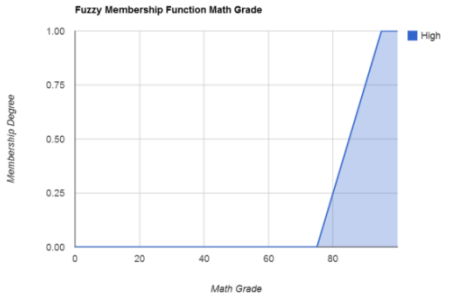

Fig. 5. Visualization of Math Grade Fuzzy Membership Function

For example, if a student has math grade of 90, then the application will look for its membership degree through the equation of the line that formed the point (75, 0) and (95, 1), as given by $y = \{ x \mid 0.05x - 3.75, \text{for } 75 \leq x \leq 95 \}$. Thus, membership degree of 90 is 0.05 * 90 - 3.75 = 0.75.

GPA is a numerical attribute like the math grade, so that the fuzzy value of GPA will also be calculated through a fuzzy membership function. For example, we make a group fuzzy set for GPA named 2014 which has a fuzzy set named High. Inside this fuzzy set, we determine points, such as (0, 0), (3.2, 0), (3.4, 0.6), (3.7, 1), and (4, 1) as shown on Figure 6.

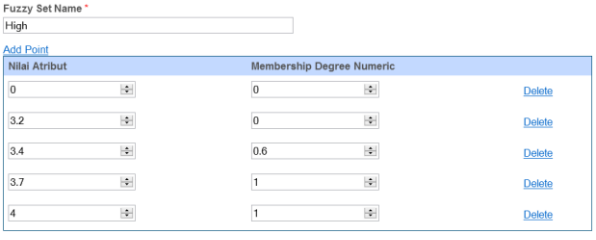

Fig. 6. Input for GPA Fuzzy Membership Function

This example of engineering fuzzy set for major attribute, high math grade fuzzy membership function, and high GPA fuzzy membership function are determined by interviewing one of PCU's structural officers. Next, we choose the attributes that are used during this test and each attribute's group fuzzy set that we just made before as shown on Figure 7.

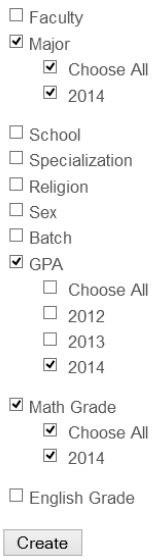

Fig. 7. Input Attributes for Fuzzy Association Rules

This test will generate all combinations of fuzzy association rules using every fuzzy sets of the attributes chosen. For example, one of the rules may be represented by:

Rule-3:

$$Major(X, \text{"Engineering"}) \wedge Math(X, \text{"high"}) \Rightarrow GPA(X, \text{"high"})$$

Rule-3 is a fuzzy rule, where A={Engineering, high} and B={high}. Next, each academic records of students shown in Table 1 will be converted using the fuzzy sets to fuzzy values as shown in Table 2.

Table 2. Calculation of Fuzzy Values

	α	β	ɣ	X	Y	X*Y	Z
1	0.1	0	0.42	0	0.42	0	0
2	1	0	0.613	0	0.613	0	0
3	1	0.75	1	0.75	1	0.75	0.75
4	0.5	0.55	1	0.5	1	0.5	0.5
5	0.5	0.15	0.667	0.15	0.667	0.10005	0.15
6	0.2	0.05	0.09	0.05	0.09	0.0045	0.05
7	0.2	0	0.45	0	0.45	0	0
8	0.2	0.7	0.813	0.2	0.813	0.1626	0.2
9	1	0.8	1	0.8	1	0.8	0.8
10	1	0	0	0	0	0	0
11	0.2	0.2	0	0.2	0	0	0
12	0.2	0.05	0	0.05	0	0	0
Σ	6.1	3.25	6.053	2.7	6.053	2.31715	2.45

Note:
$\alpha = \mu_{engineering}$ (major)
$\beta = \mu_{high}$ (math)
$\gamma = \mu_{high}$ (GPA)
$X = \min(\alpha,\beta)$
$Y = \min(\gamma)$
$Z = \min(\alpha,\beta,\gamma)$

Therefore, support of Rule-3 can be calculated by (12),
supp(Rule-3) = 2.45 / 12 = 0.20417

On the other hand, confidence of Rule-3 can be calculated by (13),
conf(Rule-3) = 2.45 / 2.7 = 0.90741

On the other hand, correlation of Rule-3 can be calculated by (14),
corr(Rule-3) = 2.45 / 2.31715 = 1.05733

The manually calculated support, confidence, and correlation of Rule-3 are match with the output of the fuzzy association rules generated by this test as shown on Figure 8.

Fuzzy Association Rules with 3 Attributes				Support	Confidence	Correlation
Major = Engineering (2014)	Math Grade = High (2014)	==>	GPA = High (2014)	0.20417	0.90741	1.05733

Fig. 8. Example Output of Fuzzy Association Rules

To evaluate this application, research on the use of this application is conducted. Samples of this research is five structural officers of PCU. To collect the data, distributed a questionnaire containing indicators to evaluate the use of the application. The indicators include display of application, determination of fuzzy values, customization of fuzzy association rules, ease of use, the ability to address the needs of users, and overall. From the data collected, the calculation of the percentage of user satisfaction in using this application is done.
Assessment of the feasibility of the application:

1. Display of application is 100% good
2. Determination of fuzzy values is 80% good
3. Customization of fuzzy association rules is 80% good
4. Ease of use is 100% good
5. The ability to address the needs of users is 60% good
6. Overall is 100% good

6 Conclusion

The generated fuzzy association rules have been tested and matched with the Multidimensional Fuzzy Association Rules algorithm and the reality of academic situation of PCU's students. From the assessment, obtained that overall application is 100% good. This suggests that the application developed has benefits for PCU and can be continued for the purpose of decision-making process by top-level management.

References

1. Han, J., Kamber, M., Pei, J.: Data mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann, San Fransisco (2012)
2. Klir, G. J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, New Jersey (1995)
3. Zadeh, L. A.: Fuzzy Sets and Systems. In: International Journal of General Systems, Vol. 17, pp. 129-138 (1990)
4. Intan, R.: A Proposal of Fuzzy Multidimensional Association Rules, Jurnal Informatika Vol. 7 No. 2 (2006)
5. Intan, R.: A Proposal of an Algorithm for Generating Fuzzy Association Rule Mining in Market Basket Analysis. In: Proceeding of CIRAS (IEEE). Singapore (2005)
6. Intan, R.: Hybrid-Multidimensional Fuzzy Association Rules from a Normalized Database. In: Proceedings of International Conference on Convergence and Hybrid Information Technology. IEEE Computer Society, Daejeon, Korea (2008)
7. Gour, V., Sarangdevot, S. S., Tanwar, G. S., Sharma, A.: Improve Performance of Extract, Transform and Load (ETL) in Data Warehouse. In: International Journal on Computer Science and Engineering, 2(3), 786-789 (2010)
8. Silwattananusarn, T., Tuamsuk, K.: Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012. In: International Journal of Data Mining & Knowledge Management Process (IJDKP), 2(5), 13-24 (2012)
9. Ramageri, B. M.: Data Mining Techniques and Applications. In: Indian Journal of Computer Science and Engineering, 1(4), 301-305 (2010)
10. Padhy, N., Mishra, P., Panigrahi R.: The Survey of Data Mining Applications and Feature Scope. In: International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), 2(3), 43-58 (2012)