ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse278952020.pdf
https://doi.org/10.30534/ijatcse/2020/278952020

A Synthesis of Optimal Unknown Number Clustering System and Categorical

Eka Arriyanti[1], Pitrasacha Adytia[2]
[1]Informatics Engineering, STMIK Widya Cipta Dharma, Indonesia, ekaarry@wicida.ac.id
[2] Information System, STMIK Widya Cipta Dharma, Indonesia, pitra@wicida.ac.id

ABSTRACT

Big data makes data dynamic and research moves from sample to population. These circumstances imply data mining, specifically clustering, should be able to produce the formed clusters of population as much as possible, thus leaving only a little "junk" of data. Unknown number clustering data mining is a term for set of clustering algorithms which cluster without defining number of formed clusters. Without intending to let others and their data types, this research proposes a synthesis of optimal unknown number clustering system and categorical, because the excellence of DBSCAN algorithm performance on spatial and non-spatial data, and the expectance of standard clustering system presence for population statistics. After reanalyzing its Spatial Coordinate Way and the radius parameter range, the synthesis is done with structural analysis based on general data mining steps for framing the complete clustering process. Then, the result of analysis is compared against the cluster analysis requirements to convince the conclusion on the system synthesis. The categorical data without weighting are also possibly converting into spatial form, but imply qualitative value. Therefore, the synthesis recommends that the standard clustering system should have special items on unweighted categorical data in addition to a numeric type.

Key words : Clustering, data, population, system_synthesis.

1. INTRODUCTION

One of the things that should influence high information management decisions in the 4.0 era is the result of data mining. Decisions are influenced by information obtained from the amount of data that has been used during the work period. The piles of mined data, produce information that provides expensive business value because the strategic level of management. In this position, data mining is like statistical analysis in research, but for processing big data. Regardless of sample-based research statistics, data mining will underlie population statistics [1], because it has collected information from almost all data of population.

Referring to the data mining tasks which consist of classification, estimation, prediction, affinity, clustering, description, and profile determination [2], the data mining optimization depends on optimization in each task (say type). After being studied, it turns out that each type of data mining has a concept of execution based on the original intent which then underlies the algorithm. In example, clustering. There are five main types of clustering methods [3], but in general, it is necessary to know how many clusters will be formed or how many members of the clusters are. In other words, one set algorithms based on the number of known clustering and another based on the unknown number of clustering. Due to these two execution concepts, there are some clustering algorithms.

Each theory or practice of clustering algorithms discussions mentions the specificity of one compared to the other. But basically in clustering, it is how to cluster as much data as possible. Collective clustered data have similarity with fellow members in the same cluster. If the mean values of this data are used to represent the data points in the population, then this will form the basis of statistics for population. However, statistics requires a standard of clustering. Therefore and also because an algorithm which represents the type of data mining can be chosen [4], DBSCAN is chosen as a representation of the unknown number clustering. Because, despite it is a spatial clustering based on Euclidean distance, the data type is numeric and it can detect clusters of arbitrary shape [5]. This is what has been attested by [6] with Spatial Coordinate Way; SCW, the non-spatial data converting way, where clustering of DBSCAN considers the data to be clustered as points (X, Y) in integers (not geospatial coordinates) and with Density Based Spatial Clustering Application with Noise-radius parameter range; DBSCAN-rpr. This is for finding the most formed clusters

This research is sponsored by Kemenristekdikti-BRIN (Ministry of Research, Technology, and Higher Education-National Research and Inovation Agency) Republic of Indonesia, 2020.

with the smallest noise, and all of formed clusters are proven good. It has confirmed [7] that DBSCAN performs excellently on any data (spatial or non-spatial).

The purpose of this research is to emphasize synthesis of the system resulted. The synthesis is important to initiate standard of clustering as the basis for population-based statistics. Therefore, the problem formulation of this research is an analysis to SCW and DBSCAN-rpr, a structural analysis on doing synthesis based on general data mining steps for framing the complete clustering process, and a confirmation on system synthesis with cluster analysis requirements for the convince system synthesis conclusion. The following is a figure of the problem formulation:

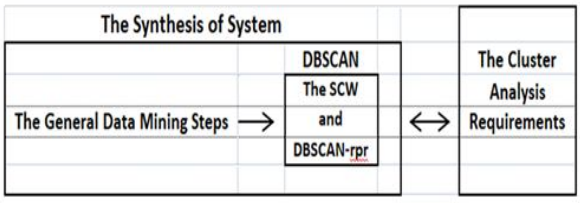

Figure 1: The Problem Formulation

2. OBJECTS OF RESEARCH REVIEW

2.1 Spatial Coordinate Way and DBSCAN-rpr

DBSCAN clustering considers the data to be clustered as points (X, Y) where for each point, it is used as a center to find its neighborhood friends according to the specified radius and minimum point number. By this way, each point is potentially clustered. Data is noise if it does not enter into any cluster.

The original spatial data, data defined by (x, y) coordinates that point out a location on the world, can be clustered immediately after the coordinates are converted to distances and the parameter values inputed. To cluster non-spatial data, data defined by attributes on the table (tabular data) [8] in tables, spreadsheets, or data cubes [9], DBSCAN needs conversion of tabular data being spatial. This conversion is called Spatial Coordinate Way, a way for preprocessing non-spatial data before clustering. Stages in this paper are rearranged after reanalysing the SCW. They are:

1. Converting dimensions;
   1) Define non-spatial datasets and the number of Dimensions; D
   2) Analyze the relationship between variables.
   3) Sort the coordinates formed based on the tendency of variables based on clustering objectives.
   4) Normalize the data in the previous step variable (3); the normalization scale must be Integer; I.
   5) Repeat steps 2-4 so that there are only 3 variables.
   6) Determine the representation of the variables in the coordinates formed to be the coordinates (X, Y).
   7) Normalize the data in the preceding step variable (5) ; the normalization scale must be integer.
2. Determining the radius parameter range, $\mathcal{E}$ range;

$\mathcal{E}_1, \ldots, \mathcal{E}_4 =$
$((((minPts - 1) - 1) - 1) - 1), \ldots, (minPts - 1)$ (1)
where
$minPts \geq D + 1$ (2)

D : number of Dimensions; number of tabular data attributes or variables of non-spatial data.
minPts : number of neighborhood density thresholds.

Data normalization is a formal process of decomposing relations with anomalies to produce smaller, well-structured, and stable relations. It is a general reduction of data to canonical form, where the canonical form of a positive integer in decimal representation is a finite sequence of digits that does not begin with zero. The positive integer; I = 1, 2, 3, … etc. [10]. In certain measurements, categorical data are given weight so that they are measured and together with numerical data are normalized, both directly and inversely proportional. In this case, numerical data are a reference for categorical weighting and needs direction from the independent variable.

DBSCAN-rpr is :

$\mathcal{E} = minPts - 1 \ni \mathcal{E} \leq minPts - 1 \; ; \mathcal{E}_1, \ldots, \mathcal{E}_4$ (3)

$$\mathcal{E}_1, \ldots, \mathcal{E}_4 \begin{cases} \varepsilon_1 = \left(\left(\left(\left(minPts - 1\right) - 1\right) - 1\right) - 1\right) = minPts - 4 \\ \varepsilon_2 = \left(\left(\left(minPts - 1\right) - 1\right) - 1\right) = minPts - 3 \\ \varepsilon_3 = \left(\left(minPts - 1\right) - 1\right) = minPts - 2 \\ \varepsilon_4 = \left(minPts - 1\right) = minPts - 1 \end{cases} \quad (4)$$

where

$$minPts \geq D + 1; \; D = \begin{cases} \text{Density; for spatial data} \\ \text{Dimension; for non} - \text{spatial data} \end{cases}$$

Density : amount of data per 1 unit square coordinates; taken the minimum for this algorithm.
Dimension : number of the tabular data attributes or variables of non-spatial data.

2.2 General Data Mining Steps

Reference [11] discusses a general five step experimental procedure that is adaptive to data mining problems. They are stating the problem and formulating the hypothesis, collecting the data, preprocessing the data, estimating the model, and interpreting the model and drawing conclusions. The five are adapted as the general data mining step whit adjustment:

1. Defining the data domain; purpose of data mining provides directions for collecting data that must exist. The research variables imply how the data domain can be found. Defining the data domain will frame the mining to search only the necessary or related data.
2. Collecting the data; random data collecting is assumed in most data mining applications. It is important to make sure

that, in clustering data mining, the data used for clustering and testing come from the same way.

3. Preprocessing the data; outlier detection and scaling-encoding-selecting features are at least two common tasks of preprocessing. To detect and eventually to remove outliers, and to develop robust modelling methods are two strategies for dealing with outliers. It is recommended to scale and bring data to the same features for further analysis.
4. Estimating the model; the implementation of this process is based on type of data mining. Selecting the best model is an additional task.
5. Interpreting the result; modern data mining methods are expected to yield highly accurate results using high-dimensional models. The problem of interpreting these models is considered a separated task with specific techniques to validate the results. A user does not want hundreds of pages of numeric results. He only needs to understand on the result from the right interpretation and for certain people, usually needs a printed version to hold on to.

2.3 Cluster Analysis Requirements

The cluster analysis requirements are used for comparing clustering methods [12]. The fulfillment of these requirements makes the clustering system ideal. The nine typical requirements are :

1. Scalability; work well on small or big data.
2. Ability to deal with different types of attributes; may require clustering any data (numerical, categorical, or a mix of both).
3. Discovery of clusters with arbitrary shape; can detect clusters of arbitrary shape.
4. Requirements for domain knowledge to determine input parameters; simple specification of domain knowledge and clear definition of parameter range.
5. Ability to deal with noisy data; robust to noise.
6. Incremental clustering and insensitivity to input order; insensitive to the input order is needed.
7. Capability of clustering high-dimensionality data; good in cluster low or high-dimensionality data.
8. Constraint-based clustering; good clustering behavior that satisfies specified constraints.
9. Interpretability and usability; clustering results are interpretable, comprehensible, and usable.

3. ANALYTICAL PROCESS

3.1 The Synthesis of System

There are 18 synthesis of the system as the result of a structural analysis from the general data mining steps until the complete clustering process. Some of them which are the order of the analysis that can be interpreted from the other, are not shown. Use Case of UML; Unified Modelling Language, is only a tool to ease drawing of the analysis.

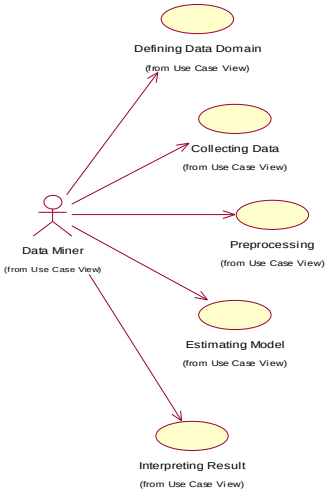

Figure 2: Synthesis 1

Figure 2 shows synthesis 1 that; system has main items of the five general data mining steps.

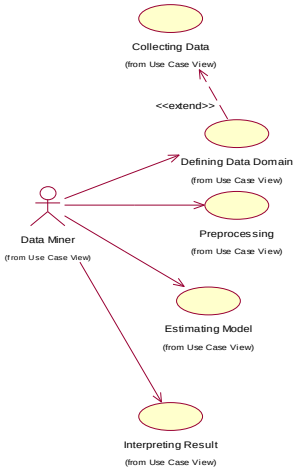

Figure 3: Synthesis 2

Figure 3 shows synthesis 2 that; system becomes four with the defining data domain item becomes requirement for the collecting data.

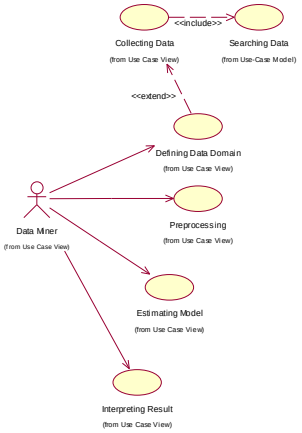

Figure 4: Synthesis 3

Figure 4 shows synthesis 3 that; the searching data is part item of the collecting data.

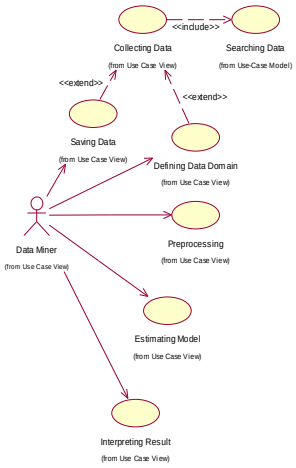

Figure 5: Synthesis 4

Figure 5 shows synthesis 4 that; system becomes five again with the saving data item becomes requirement for the collecting data.

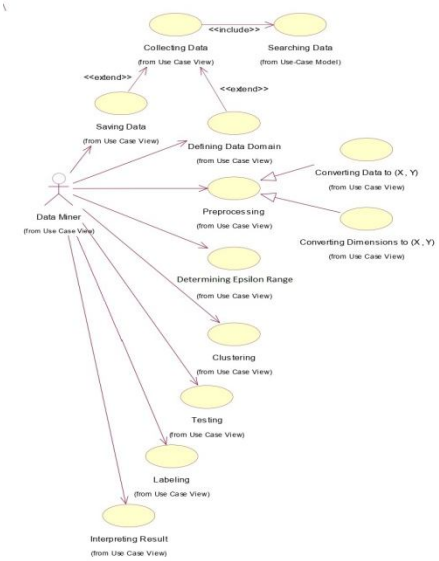

Figure 6: Synthesis 5

Figure 6 shows synthesis 5 that; system becomes eight with the estimating model item changed by the clustering (the model) and followed by the determining epsilon range, the testing, and the labeling. The preprocessing has two sub items, the converting data to (X, Y) and the converting dimensions to (X, Y).

...

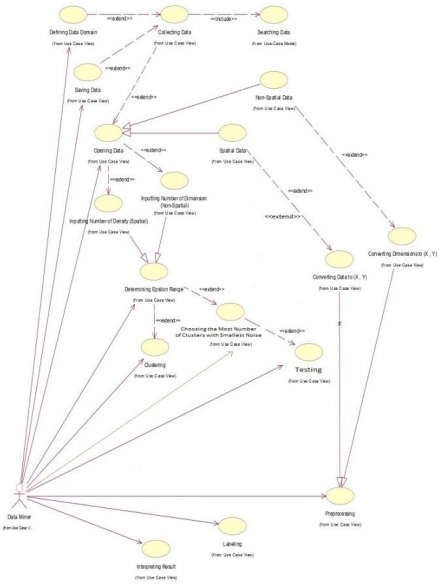

Figure 7: Synthesis 14

Figure 7 shows synthesis 14 that; system becomes ten with adding two items, the opening data and the choosing the most clusters with smallest noise. The collecting data is requirement for the opening data which is requirement for the inputing number of density (spatial) and the inputing number of dimension (non-spatial), where both are sub items for the determining epsilon range. The determining epsilon range is requirement for the clustering and the testing, where the choosing the most clusters with smallest noise is requirement for the testing. The opening data also has two sub items, the spatial and the non-spatial, where respectively, is the requirement for the converting data to (X, Y) and the converting dimensions to (X, Y).

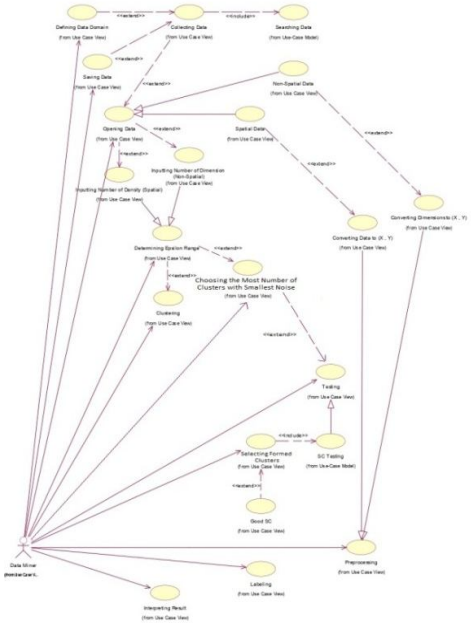

Figure 8: Synthesis 15

Figure 8 shows synthesis 15 that; system becomes eleven with adding the selecting formed clusters items. The SC testing is part item of the selecting formed clusters and the good SC is its requirement.

...

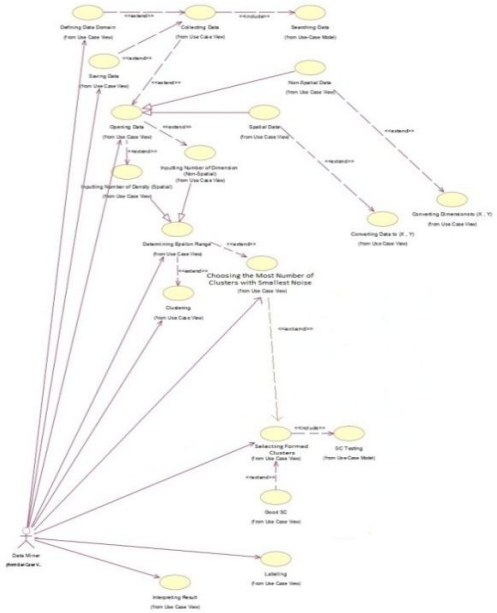

Figure 9: Synthesis 17

Figure 9 shows synthesis 17 that; system becomes nine with deleting the testing and the preprocessing, and the most clusters with smallest noise becomes requirement for the selecting formed clusters which replaces the testing. The testing is deleted because it is super item for the SC testing, which is part item of the selecting formed cluster. The preprocessing is deleted because it is sub item of the opening data.

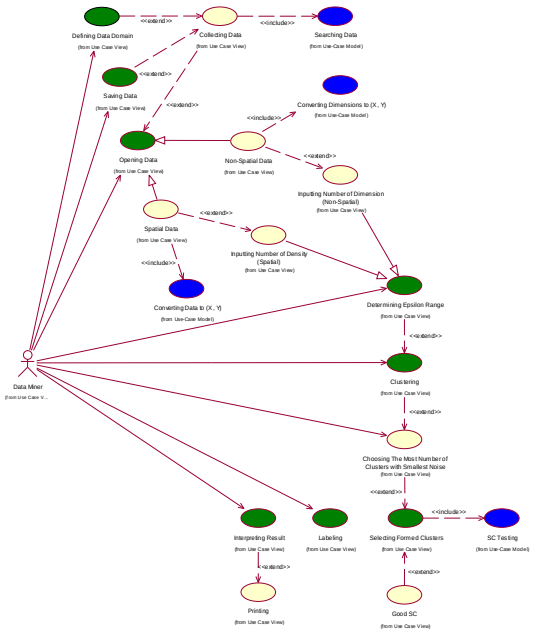

Figure 10: Synthesis 18

Figure 10 shows synthesis 18 that; system becomes eight again because the formed clusters is ending of a series of requirements beginning from the opening data. Because the determining epsilon range is requirement for the clustering and the choosing the most clusters with smallest noise, then the clustering becomes requirement for the choosing the most clusters with smallest noise. Because the spatial and the non-spatial are sub items of the opening data, they respectively become requirements for the converting data to (X, Y) and the converting dimensions to (X, Y). The interpreting result is requirement for the printing.

3.2 The Synthesis for Categorical Data

For instance, there are non-spatial data in a data cube, where the categorical data are unweigthed, that are all variables have the same meaning; colored or uncolored:

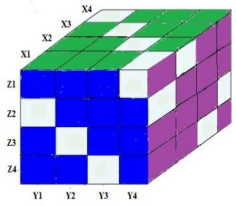

Figure 11: X(X1, X2, X3, X4) = Green, Y(Y1, Y2, Y3, Y4) = Blue, and Z(Z1, Z2, Z3, Z4) = Purple

Due to figure 11, the tabular data for Z(Y(X)) from Z(Y(X1) to Z(Y(X4) are in table 1, 2, and 3:

Table 1: Z(Y(X1)) = (Y1, Y2, Y3, Y4, (X1,Y1))

Z	D					(x, y)
	Y				X	
	Y1	Y2	Y3	Y4	(X1,Y1)	(D,y)
Z1	■	■	■		■	5,1
Z2		■	■	■	■	5,3
Z3	■		■	■	■	5,4
Z4	■	■		■	■	5,5
Number of Variables	4				1	

...

Table 2: Z(Y(X4)) = (Y1, Y2, Y3, Y4, (X4,Y1))

Z	D					(x, y)
	Y				X	
	Y1	Y2	Y3	Y4	(X4,Y1)	(D,y)
Z1	■	■	■			5,3
Z2		■	■	■		5,3
Z3	■		■	■		5,3
Z4	■	■		■		5,3
Number of Variables	4				1	

...

Table 3: Z(Y(X4)) = (Y1, Y2, Y3, Y4, (X4,Y4))

Z	D					(x, y)
	Y				X	
	Y1	Y2	Y3	Y4	(X4,Y4)	(D,y)
Z1	■	■	■		■	5,1
Z2		■	■	■	■	5,3
Z3	■		■	■	■	5,4
Z4	■	■		■	■	5,5
Number of Variables	4				1	

Table 1, 2, and 3 show that;
There are five attributes of Z(Y(X)). They are Y1, Y2, Y3, Y4, and (X,Y). This is the Dimensions; D, where D is 4 plus 1. For Z1…4, once Z1(Y(X)) is colored then Z2…4(Y(X)) is colored, and once Z1(Y(X)) is uncolored then Z2…4(Y(X)) is uncolored.

For instance, the attribute (X1,Y1) is uncolored:

Table 3: (X1,Y1) is uncolored

Z	D					(x, y)
	Y				X	
	Y1	Y2	Y3	Y4	(X1,Y1)	(D,y)
Z1	■					5,1
Z2	■	■				5,2
Z3	■	■	■			5,3
Z4	■	■	■	■		5,4
Number of Variables	4				1	

and the attributes Y1, Y2, Y3, and Y4 are uncolored:

Table 4: Z1(Y) = Y1, Y2, Y3, and Y4 are uncolored

Z	D					(x, y)
	Y				X	
	Y1	Y2	Y3	Y4	(X1,Y1)	(D,y)
Z1					■	5,1
Z2	■	■			■	5,3
Z3	■	■	■		■	5,4
Z4	■	■	■	■	■	5,5
Number of Variables	4				1	

Due to table 3 and 4 ;
It is clear that y is number of attributes colored in D. Therefore, (x, y) = (D, y) is a spatial form of a categorical data, if D is the number of attributes and y is the number of colored attributes. If the colored (X,Y) is 1 and the uncolored (X,Y) is 0, then the 1 and 0 represents the qualitative value of Z.

4. CONFIRMATION AND CONCLUSION

The result of the analysis (synthesis 18) with the synthesis for categorical data, has been confirmed by point 1, 2, 3, 4, 5, 6, 7, and 9 of the cluster analysis requirements with the items shown. Point 2 requires the opening data item to have a branch for categorical, because it needs a separate step. The Fulfillment of point 8 depends on the application of the case. Hence, in a clustering system, categorical clustering is a separate sub-system.

5. FUTURE WORK

The research following this research are on categorical clustering based on spatial form, and on defining population-based statistics.

REFERENCES

1. A.D. Ciaccio and G.M. Giorgi, "**Statistics in the big data era**", Rivista Italiana di Economia Demografia e Statistica, vol. LXX (4), Oct.-Dec. 2016.
2. "**Types of clustering methods**", datanovia.com, accessed Oct., 2020.
3. S.A.D. Budiman, D. Safitri, and D. Ispriyanti, "**Comparison of the K-Means method and the DBSCAN method in the student boarding house grouping in Tembalang village, Semarang [Perbandingan metode K-Means dan metode DBSCAN pada pengelompokan rumah kos mahasiswa di kelurahan Tembalang, Semarang]**", Gaussian Journal, vol. 5(4), pp. 757-762, 2016.
4. C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse, and A. Folleco, "**An empirical study of the classification performance of learners on imbalanced and noisy software quality data"**, Elsevier Journal On Information Sciences, vol. 259, pp. 571-595, 2014.
5. P. Perner, "**Data mining-concepts and techniques**", researchgate.net, article 220633275, 2015.
6. E. Arriyanti, I. Arfyanti and P. Adytia, "**Spatial coordinate trial : converting non-spatial data dimension for DBSCAN**," in *Proc. 6th International Conf. on Electrical Engineering, Computer Science and Informatics* (EECSI), Bandung, 2019, pp. 223-228.
7. E. Schubert, J. Sander, M. Ester, H.P. Kriegel, and X. Xu, "**DBSCAN revisited, revisited : why and how you should (still) use DBSCAN**", ACM Transactions on Database Systems, vol. 42(3), article 19, 21 pages, Jul. 2017.
8. A.K. Saraf, "**Non-spatial data (attributes) and their typhes**," *Lecture-06*, Departmentof Earth Sciences, Indian Institute of Technology, Roorkee, 2016.
9. J. Han, M.Kamber, and J. Pei, "**Data mining : concepts and techniques**", 2nd ed., pp. 110-112, the Morgan Kaufmann series, Elsevier, 2006.
10. A. Shlapentokh, "**Defining integers**", Bulletin of Symbolic Logic 17(2), Aug. 2010.
11. "**Data mining**", Article ID 047122854.pdf, accesed Sep., 2020.
12. J. Han, M. Kamber, and J. Pei, "**Data mining : concepts and techniques**", 3rd ed., pp. 445-447, the Morgan Kaufmann series, Elsevier, 2012.