Proc. of 2nd International Conference on Advances in Computer, Control & Telecommunication Technologies (ACT 2010), pp.185-187

Mapping of Information and Communication Technology (ICT) Progress using Self Organizing Map (SOM)

Yuni Arti, Ninon Nurul Faiza, Sri Saraswati Wardhani, Anto Satriyo Nugroho, Dwi Handoko, Zain Saifullah, Wenwen Ruswendi, Edi Santoso
Center for Information & Communication Technology, Agency for the Assessment & Application of Technology
BPPT 2nd building 4F, Jalan M.H.Thamrin 8, Jakarta, Indonesia 10340
Email: yuni.arti@hotmail.com

***Abstract*—Information and Communication Technology (ICT) is one of the most inportant aspect in a country. Good progress in ICT will be a valuable factor to compete with other countries. The progress in DKI Jakarta province is considered as a measurement or representation, that the ICT in Indonesia has been developed well or not. This research will use data mining method with clustering, using Self Organizing Map (SOM) algorithm. The algorithm is used for clustering the villages in DKI Jakarta Province, based on the availability of telecommunication and internet facility. The clustering or grouping result, less or more, will describe the progress rate in the province. Last, the ICT progress in the country hoped will be represented in village clustering data of DKI Jakarta.**

***Keywords*—data mining, clustering, SOM, ICT**

I. INTRODUCTION

Information and communication technology (ICT) can support almost of all people life sector such as: economic, social, politic, and culture. Good progress in ICT means a country will have a bigger chance to win the competition with other countries.

ICT in Indonesia is measured in many indicators: telecommunication and internet, government ICT budged, ICT industy investation, ICT commodity export-import, university human resources, also ICT patent and copyright [1]. From all indicators, telecommunications and internet is the most important factor that can describe the ICT progress for the bigger scale because it is measured by identifying the availability of supporting facility in the whole of area.

In a country, the ICT progress can be represented in it's capital city. DKI Jakarta as Indonesian capital city can be considered as a ICT progress representation in this country. Identifying of ICT progress in DKI Jakarta can be positioned as first step to see the ICT progress in Indonesia in the bigger scale.

Mapping of ICT progress in DKI Jakarta needed to see the availability of telecommunication and internet supporting facility. Next, the availability of this facility is placed as one of indicator that determine ICT progress rate. Finally, the indicator can be a good input for legitimated people or organization to make a government decision.

This research uses data mining technics with clustering method, using Self Organizing Map (SOM) algorithm. The SOM can be used to visualize groups of similar items and how this groups are related to each other [2]. Many applications have been built using SOM. Two of the most popular application are WebSOM [3] and World Poverty Map [4].

This research use SOM for clustering village data in DKI Jakarta based on telecommunication and internet facility's ability. Clustering process will produce the village clusters with the topologies. The characteristic of each cluster is used to describe progress rate of the cluster. Last, the cluster topologies define a relationship closeness between one of cluster progress rate and other.

The structure of this paper is organized as follows: Sec.2 described the methodology of research from the preprocess step until the data mining step using SOM, Sec.3 reported the experimental results and last, the conclusion in Sec.4.

II. METHODOLOGY

The concept of similarity is important for many data mining related applications. Defining similarity can be very difficult if several aspects are involved. Self Organizing Map is a powerful tool to visualize how the data looks like from a certain perspective of similarity [2].

One of the main applications of the SOM is data analysis. Similarity relationships within a data set can be visualized on a graphical SOM display. Also other aspects of the data, distribution of the values of data variables, can be visualized on the same display [3].

This research uses the Village Potention Data for 2005 in DKI Jakarta province. The data is owned by Badan Pusat StatisICT (BPS) or Indonesian Statistic Agency. The research uses Data Discovery (KDD) process. Before doing the process, there are three pre-process steps: data integration, data selection, and data transformation

When data preprocessing is finished, the next step, the data will become an input in SOM algorithm [4]:

1 Initialize the centroids randomly then select the next object

2 Let $\vec{m}_1, \cdots, \vec{m}_M$ be the reference vector of M centroids. Determine the closest centroid $\vec{m}_c$ among the nodes to the object $\vec{x}$ using the following Euclidean distance equation:

$$dist(\vec{x}, \vec{m}_i) = \sqrt{\sum_{j=1}^{d} (x_j - m_{i,j})^2} \quad (1)$$
$$i = 1,2,\cdots,M$$

d is the dimensionality of the data.

3 Update the winner centroid and the neighbor centroids. For time step t, let $p(t)$ be the current object (point), c be the closest centroid to $p(t)$ and $N_c(t)$ be the monotonically decreasing neighborhood function with respect to the winner neuron c. Then, for time $t+1$, the reference vector of the neurons $\vec{m}_i(t+1)$ are updated using the following equation.

$$\vec{m}_i(t+1) = \vec{m}_i(t) + \alpha(t)\left(\vec{p}_i(t) - \vec{m}_i(t)\right) \quad \text{for } i \in N_c(t) \quad (2)$$

$$\vec{m}_i(t+1) = \vec{m}_i(t) \quad \text{for } i \notin N_c(t) \quad (3)$$

Variable α(t) is a learning rate parameter, 0<α(t)<1, which decreases monotonically with time and controls the rate of convergence. The algorithm terminates when there are no significant changes of centroids position.

This research uses experiment as many as 144 times with parameter combination below:

1. radius: 1 (min) and 5, 10, 15, 20 (max)
2. learning rate: 0.1 (min) and 0.1 , 0.5 , 0.9 (max)
3. iteration: 100, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000,5500
4. grid: 10 x 10

III. EXPERIMENTS AND DISCUSSION

The data of village potention consists of two groups i.e. data that gives information about village potention in many aspects (69957 records and 25 attributes) and data about all province names, regencies, sub-ditricts, and villages in Indonesia (69957 records dan 9 attributes). Both of data are integrated in pre-process step, becoming a data with 69975 records and 34 attributes.

All of 34 attributes are eliminated to 18 attributes, which related with telecommunications and internet aspect. The selection doesn't works only in attributes, but also in record. The selected record is the village in DKI Jakarta province. Finally, there are 267 records and 18 attributes that is used in the research.

The transformation process that is applied to 267 records is a coding for attribute values. Code '0' informs that 'there is no supporting facility in an area', and code '1' means 'supporting facility in an area is available'.

Based on 144 experiments with many parameter combination, we selected the experiment that produce 11 cluster with combination: parameter radius = 5, rate max = 0.1, iteration = 2500 and Quantization Error (VQE value = 3.71. The experiment's selected because total of cluster and cluster member are consistently produced. The cluster details are listed in Table 1.

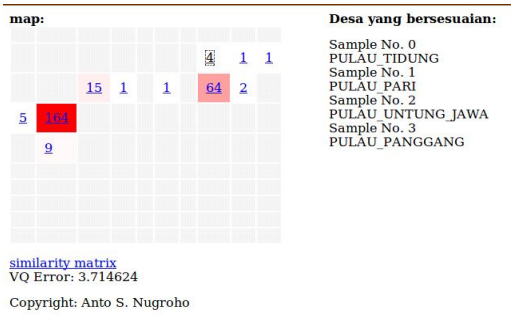

Figure 1. Data mapping in grid 10x10.

TABLE I. THE DETAILS OF EACH CLUSTER

Cluster	1	2	3	4	5	6
Member	4	1	1	164	2	64
Member (%)	1.5	0.38	0.38	61.42	0.75	24

Cluster	7	8	9	10	11
Member	15	1	9	1	5
Member (%)	5.6	0.38	3.37	0.38	1.87

A. *The cluster charecteristic*

The process of data clustering produces 11 clusters that have similarity each other. The charecteristics of each cluster are showed in Table 2. The value '0' and '1' in the that table reflect the availability of telecommunication and internet aspect supporting facility. The value '0' means 'unavailable' and '1' means 'available'. The availability of the facility in each cluster is the indicator that determine ICT progress a area.

The Cluster 7 has 15 members (Table 1) and it is categorized as a group of village with the complete facility. All of the 15 villages are distributed in some regency, i.e. South Jakarta, East Jakarta, Central Jakarta, and West Jakarta.

The largest cluster is cluster 4 with 164 members (Table 1). It is group of villages with sufficient facilities, because from all 18 facilities used as indicator for ICT progress, there is only 1 facility which is unavailable, i.e. overseas broadcast. All of them are distributed in 5 regency in DKI Jakarta, i.e. South Jakarta, East Jakarta, West Jakarta, and North Jakarta.

Cluster 2, 6, 8, 9, 10 dan 11 are the clusters with incomplete facilities, because from all 18 facilities used as indicators for ICT progress, there is 2 facilities that is unavailable. Facilities that is unavailable in cluster 2 are public telephone and internet kiosk. Facilities that is unavailable in cluster 6 are internet kiosk and overseas broadcast. Facilities that is unavailable in cluster 8 are RCTI broadcast and overseas broadcast.

TABLE II. CLUSTER CHARACTERISTIC

Cluster	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18
1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1
2	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
5	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1
6	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1
9	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
10	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1

x1 : electricity
x2 : public telephone
x3 : telephone kiosk
x4 : internet kiosk
x5 : TVRI broadcast
x6 : TRANS TV broadcast
x7 : TPI broadcast
x8 : RCTI broadcast
x9 : SCTV broadcast
x10 : INDOSIAR broadcast
x11 : TV7 broadcast
x12 : GLOBAL TV broadcast
x13 : ANTV broadcast
x14 : LATIVI broadcast
x15 : METRO TV broadcast
x16 : overseas broadcast
x17 : local broadcast
x18 : mobile phone signal

. Facilities that is unavailable in cluster 9 are public telephone and overseas broadcast. Facilities that is unavailable in cluster 10 is GLOBAL TV broadcast and overseas broadcast. Facilities that is unavailable in cluster 11 is overseas broadcast and local broadcast.

Cluster 1, 3, and 5 are the clusters with insufficient facilities. In cluster 1 and 5, from all 18 facilities used as indicators for ICT progress, there are 3 facilities that is unavailable in those clusters. Whereas in cluster 3, from 18 facilities used as indicators for ICT progress, there are 4 facilities that is unavailable in that cluster. Facilities that is unavailable in cluster 1 are public telephone, internet kiosk, and overseas broadcast. Facilities that is unavailable in cluster 3 are public telephone, telephone kiosk, internet kiosk, and overseas broadcast. Facilities that is unavailable in cluster 5 are , internet kiosk, and overseas broadcast.

B. The topology inter clusters

Based on Figure 1, the progress rate of ICT in village of cluster 1 is closer with cluster 2, 5 and 6. The difference facility between cluster 1 and cluster 2 is only in overseas broadcast. The difference facility between cluster 1 and cluster 5 is in public telephone and internet kiosk. The difference facility between cluster 1 and cluster 6 is in public telephone and local broadcast.

The progress rate of ICT in villages in cluster 2 is closer with cluster 1, 3, 5, and 6. The difference facility in cluster 2 and cluster 3 is in telephone kiosk and overseas broadcast. The difference facility between cluster 2 and cluster 5 is in public telephone, overseas broadcast, and local broadcast. The difference facility between cluster 2 and cluster 6 is in public telephone and overseas broadcast. The progress rate of ICT in village in cluster 3 is closer with cluster 2. The progress rate of ICT in village in cluster 4 is closer with cluster 7, 9, and 11. The difference between cluster 4 and cluster 7 is only in overseas broadcast. The difference between cluster 4 and cluster 9 is only in public telephone. The difference between cluster 4 and cluster 11 is only in local broadcast. The progress rate of ICT in village in cluster 5 is closer with cluster 1, 2, and 6. The difference between cluster 5 and cluster 6 is only in local broadcast. The progress rate of ICT in village in cluster 6 is closer with cluster 1, 2 and 5. The progress rate of ICT in village in cluster 7 is closer with cluster 4 and 8. The difference between cluster 7 and cluster 8 is in RCTI broadcast and overseas broadcast. The progress rate of ICT in village in cluster 8 is closer with cluster 7. The progress rate of ICT in village in cluster 9 is closer with cluster 4 and 11. The difference between cluster 9 and cluster 11 is in public telephone and local broadcast. The progress rate of ICT in village in cluster 10 is not closer with other clusters. The progress rate of ICT in village in cluster 11 is closer with cluster 4 and 9.

IV. CONCLUSIONS

From grouping based on the availability of facilities in some villages in DKI Jakarta region, there are 11 clusters with different characteristic. This fact can be used to describe the ICT progress in each area of the cluster's members.

There is 61.42% area in DKI Jakarta which is included in cluster 4. Cluster 4 is a group of villages with sufficient availability of facilities, because from all of 18 facilities used as indicator for the ICT progress, there is only 1 facility that is unavailable, i.e overseas broadcast. Based on its topology, cluster 4 is closer with cluster 7, 9, and 11. Thus, majority villages in DKI Jakarta region had some progresses in ICT sector.

In Table 2, seems the availability of telecommunication and internet aspect supporting facility in DKI Jakarta haven't been distributed well. Thus, it is necessary for government to concern about this condition so that the growth and development of ICT sector in Indonesia can be better in the future.

REFERENCES

[1] Pusat Teknologi Informasi dan Komunikasi, Badan Pengkajian dan Penerapan Teknologi, *Indikator Teknologi Informasi dan Komunikasi* edisi 2007, PTIK BPPT Jakarta, 2007.

[2] E. Pampalk, "Aligned self-organizing maps," Proceedings of the Workshop on Self-Organizing Maps, pp. 185–190, Ktakyushu, Japan, September 2003.

[3] K.Lagus, S.Kaski, T.Kohonen. "Mining massive document collections by the WEBSOM method," Information Sciences: an International Journal, vol.163, pp. 135-156, 2004.

[4] Example of application of SOM for world poverty map http://www.cis.hut.fi/research/som-research/worldmap.html (last accessed: 5 July 2010)

[5] S.Kaski, "SOM-Based Exploratory Analysis of Gene Expression Data," Advances in Self-Organizing Maps, Springer, pp. 124-131, 2001.

[6] P.N.Tan, M.Steinbach, V.Kumar, Introduction to Datamining, Addison Wesley, 2006.