NAMED ENTITY RECOGNITION FOR INDONESIAN TEXT USING CONDITIONAL RANDOM FIELDS

Narti Jaariyah1, Ednawati Rainarli, S.Si., M.Si.2

1.2 Information Engineering, Engineering and Computer Science, University Computer Indonesia
Jalan Dipatiukur No. 112-114, Coblong, Bandung, West Java 40132, Indonesia
E-mail: nartijaariyah@gmail.com1, ednawati.rainarli@email.unikom.ac.id2

ABSTRACT

Named entity recognition is parf of extraction information to classify entities such as the name of the person's name, location, organization, time, andquantity in the text. Named entity recognition in the Indonesian language text has been done using HMM. CRF also has many advantages over Hidden Markov Model methods and Maximum Entropy Markov Model. This was evident in the implementation of named entity recognition using the method of CRF in various languages that generate value for high accuracy. Simulator named entity recognition was made to test how well the CRF in recognizing named entities. Features that use is current word part-of-speech, current word part-of-speech and previous part-of-speech, and current word part-of-speech, previous part-of-speech, and next part-of-speech. Rate simulator using the f-measure by comparing the entity label simulator results with actual entity type. Tests using training data and test data the same result obtained the best accuracy of 90.53% with 63.09% recall and precision 31.55%. The test results of training data and test data of different shows best accuracy value is 90.06% with recall and precision was 68.38% and 41.35%.

Keywords: information extraction, named entity recognition, conditional random fields, class of words

1. INTRODUCTION

The introduction of named entities is widely used in natural language processing such as extraction of information, automatic answering machine, summarizing, information retrieval, machine translation, video annotations biofarmatik and so forth. Named entity recognition system for the Indonesian language text has been done using Hidden Markov Model [1]. However, the research questions that uses the phrase extraction recognition system entities still are failures resulting entity recognition systems are still not accurately identify the entity named in the text [2]. In addition to methods of HMM are other methods that can be used to perform named entity recognition that is conditional random fields. CRF method for cases of named entity recognition successfully implemented in various languages in the world and produces high accuracy values as in English with an accuracy of 92.29% and the Bengali language accuracy value reached 90.7% [4.5]. From research that has been conducted shows that the named entity recognition using CRF has a value of high accuracy. However, the study still has not been applied to the Indonesian while each language has rules of its own language writing. Therefore, it still needed a study to determine the accuracy of the method of CRF in recognizing named entities, especially in the Indonesian language text.

2. RESEARCH METHODS

The methodology used in this study is the collection of data, analysis methods, implementation methods, testing, and conclusion that can be seen in Figure 1.

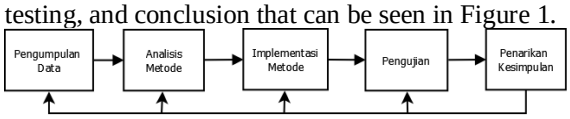


Figure 1 Phase Research

From Figure 1 stages of research methodology used is as follows.

1. Data Collection

Data collection methods used in this research is the study of literature. This phase is done by studying the literature such as books, journals, scientific articles, and websites related to the introduction of named entities and conditional random field. The purpose of this literature study is as basic theories of development simulator named entity recognition.

2. Analysis Method

In this stage, the analysis of the needs and processes of conditional random fields method to identify named entities. Once the needs and requirements simulator method collected then made the design of the simulator will be built.

3. Implementation Methods

The process is carried out in this stage is to build a simulator entity recognition based on the analysis method that has been done. Simulator is built using the Java programming language and text editor for saving all training data and testing.

4. Testing

Once the simulator is built next step is to test the results of the implementation of conditional random fields to identify the entity named in the Indonesian

language text. This testing is done by focusing on testing the accuracy of data.

5. Withdrawal conclusion

Conclusions obtained from the implementation of the method of conditional random fields to identify the entity named in the Indonesian language text based on the results of testing that was done in the previous stage.

2.1. Data Enter

Entries in the Indonesian language text that is in accordance with standard rules Indonesian. The text is divided into two, namely the text as training data and test data. Training data comprises 2,231 sentence consisting of 43 301 words to 1,756 words the person's name, the name says 1,546 locations, 1,417 words included in an organizational entity, and the remaining 38 582 words included in the entity other.¬ test data consists of 614 words with 747 words the person's name, location name 226, 444 name of the organization and not the entity as much as 9669 words.

2.2 Preprocessing

Preprocessing an early stage to perform named entity recognition on the text. Stages in the preprocessing consists of:

a. separation sentence

Separation sentence carried out using the rules in Indonesian like ending with a period (.), Exclamation mark (!) And question mark (?), And in it, there can be read that is a comma (,), semicolon (;), hyphens (-), dashes (-), sign the ellipsis (...), parentheses ((...)), brackets ([...]), quotation marks ("..."), single quotes ('...') the slash (/) and apostrophe or apostrophe [9]. In this process will be carried out using a regular expression and tagged <START> to the beginning of the sentence and <END> to the end of the sentence.

b. Tokenisasi

Tokenisasi is a separation process sentences into sets of words that correspond with the order. Punctuation in the text is considered as a word.

c. POS Tagging

After ditokenisasi, every word recognition process is carried out using a word-class POS Tagging to assist labeling kind words. POS Tagging is used IPOSTagger research results Alfan Wicaksono category Indonesian text using Hidden Markov Model.

a.2 Feature Extraction

On implementation, the features is a function that combines a feature with a corresponding label and the label adjacent. As examples of the features that

$$f_{1000}(y_{10}, x, 10) = \begin{cases} 1, & if\ x_{10} = \text{today} \\ and\ y_{10} = B - NP & \\ 0, & otherwise \end{cases}$$

Where, features in all 1000 will yield a value of 1 when the 10th word is "today" and the current label is B-NP and would be 0 for other labels.

a.3 Conditional Random Fields

Conditional Random Fields is a probabilistic model for segmentation and labeling of a sequence of data [6]. CRF shaped non-directional graph model that defines a single log-linear sequence to the sequence of the label certain observational data [3]. CRF is a probability model has many advantages compared Hidden Markov Model (HMM) and Maximum Entropy Markov Models (MEMM). CRF overcome the problem of high dependence on the assumption that HMM. This is because the CRF can determine how many features you want to build a model of CRF is different from the HMM that is local (restricted in transition binary and feature functionality of emissions, forcing each word depends only on the current label and any label depends only on label previous). Additionally CRF can have free weights while HMM must meet certain weight. CRF also overcome the problem of bias in MEMM label for CRF formulate a conditional distribution for the label as a whole compared to the data sequence formulating MEMM conditional distribution label for each data element [7]. One form of linear-chain CRF CRF is shown in Figure 2.

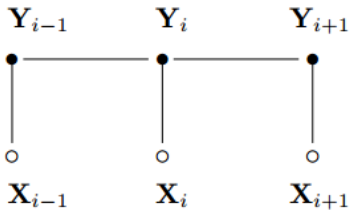


Figure 2 Linear Chain CRF

Given two random variables where x = (x_1, x_2, ..., x_n) is the set of observational data and y = (y_1, y_2, ..., y_n) is the set of labels that may be. The conditional probability of y with respect to x can be written in equation 2.1.

$$p(y|x) = \frac{1}{Z(x)} \prod_c \psi_c(y_c, x) \qquad (1)$$

Where ψc (yc ,, x) is a positive function hereinafter called the potential function and z (x) indicates the normalization of the probability distribution function conditionally label for all the data sequence x and denoted by

$$Z(x) = \sum_y \prod_c \psi_c(y_c, x) \qquad (2)$$

Potential function refers to the amount of data that bind label with features at the same time. Features usually referred to as a data pattern in several literature CRF. In the simplest case the potential is exponential function of a weighted sum of all the features.

In general CRF process is shown in Figure 3 below.

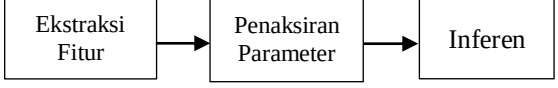


Figure 3 Stages of the CRF

a.3.1 Parameter Estimation

Parameter estimation is the process to get the optimal parameter values that features λ_k function. Λ_k value is calculated using maximum likelihood procedure. This method generally requires two calculations the log-likelihood and its first derivative on certain parameters [7]. Log-likelihood can be expressed mathematically in the following equation.

$$\mathcal{L} = \sum_{t\epsilon[1,T]}\sum_k \lambda_k f_k(y'_t, x, t) - \sum_{t\epsilon[1,T-1]}\sum_k \lambda_k f_k(y'_t, y'_{t+1} x, t) - \log Z(x) - \sum_k \frac{\lambda_k^2}{2\sigma^2} \quad (4)$$

From equation (2.6) are obtained derivative of the log-likelihood function to a feature-k is

$$G_k^x = \sum_{t\epsilon[1,T]}(f_k(y_t{}', x, t) - \sum_{y_t} P_t(y_t|x) f_k(y_t, x, t)) - \frac{\lambda_k}{\sigma^2} \quad (5)$$

$$G_{k'}^x = \sum_{t\epsilon[1,T-1]}(f_k(y'_t, y'_{t+1} x, t) - \sum_{y_t}\sum_{y_{t+1}} P_t(y_t, y_{t+1}|x) f_k(y_t, y_{t+1} x, t) - \frac{\lambda_{k\prime}}{\sigma^2} \quad (6)$$

Information :

G_k^x : the value of the gradient to-k to function node features

$G_{k'}^x$: value of gradient k-edge features to function

k : values 1 to the number of nodes feature

k' : values 1 to the number of features edge

T : the number of words in one sentence

$P_t(y_t|x)$: the probability of all the local node t

$P_t(y_t, y_{t+1}|x)$: Local probability edge all t

$f_k(y'_t, x, t)$: function node k-th feature of training data that have been labeled

$f_k(y_t, x, t)$: features in all k for all the possible label y_t

$f_k(y_t, y_{t+1} x, t)$: features in all k for all labels y_t and y_{t+1} are possible

$f_k(y'_t, y'_{t+1} x, t)$: the function of all k edge features from the training data that have been labeled

σ : standard deviation of Gaussian distribution of a range of values from 0.1 to 10 [8]

λ_k : the value of the function parameter node to the feature-k

$\lambda_{k'}$: the parameter value to the features edge-k

To get parameter values λk features optimal function, equation 5 and 6 used dynamic programming by utilizing the procedure forward-backward pass to explore the data sequences. Searches by the forward-backward pass intended for mendapatakan local probability values of all labels in the sequence data. Forward pass in each of the data xt mathematically to said 2nd through Nth word in one sentence can be written as follows.

For t = 1

$$\alpha_1[y_1] \quad (7)$$

For t = 2 to T

$$\alpha_t[y_t] = \kappa_t \sum_{y_{t-1}} \alpha_{t-1}[y_{t-1}]\, \phi_t(y_t, x)\, \psi_t(y_t, y_{t+1}, x) \quad (8)$$

Information :

S : the number of labels

κ_t : scaling factor so Σ_ (y_t) $\sum_{y_t} \alpha_t[y_t] = 1\rrbracket$, $\kappa_t > 0$

$\alpha_{t-1}[y_{t-1}]$: value of the forward pass to t-1

$\phi_t(y_t, x)$: the value of the node potential all t

$\psi_t(y_t, y_{t+1}, x)$: the value of the potential edge all t

The backward pass in each xt data is mathematically written as.

For t = T

$$\alpha_1[y_1] = 1/S \quad (9)$$

For t = 2 to T-1

$$\beta_t[y_t] = \gamma_t \sum_{y_{t+1}} \beta_{t+1}[y_{t+1}] \phi_{t+1}(y_{t+1}, x)\, \psi_t(y_t, y_{t+1}, x) \quad (10)$$

Information :

γ_t : scaling factor so $\sum_{x_t} \beta_t[x_t] = 1$, $\gamma_t > 0$

$\beta_{t+1}[y_{t+1}]$: value backward pass all t + 1

$\phi_{t+1}(y_t, x)$: value to a potential node t + 1

$\psi_t(y_t, y_{t+1}, x)$: value to a potential edge t

After all the forward pass and the backward pass to each of the data x already has a value, then the local probability calculation for any label that may be on the data xt can be written in mathematical equations 2:10 and 2:11 as follows.

$$P_t(y_t|x) = w_t \alpha_t[y_t] \phi_t(y_t, x) \beta_t[y_t] \quad (11)$$

$$P_t(y_t, y_{t+1}|x) = \text{Ш}_t \alpha_t[y_t] \phi_t(y_t, x) \psi_t(y_t, y_{t+1}, x) \phi_{t+1}(y_t, x) \beta_{t+1}[y_{t+1}] \quad (12)$$

Information :

$P_t(y_t|x)$: probability y_t local to each label that might

w_t : normalizing factor to ensure that $\sum_{y_t} P_t(y_t|x) = 1$.

$\alpha_t[y_t]$: value forward pass all t

$\phi_t(y_t, x)$: the value of the potential edge all t

$\psi_t(y_t, y_{t+1}, x)$: nilai edge potensial ke-t

$\beta_t[y_t]$: value backward pass all t

$P_t(y_t, y_{t+1}|x)$: Local probability for each label $y_t$and y_{t+1} are possible

Ш_t : normalizing factor ensuring $\sum_{y_t}\sum_{y_{t+1}} P_t(y_t, y_{t+1}|x) = 1$

$\phi_{t+1}(y_t, x)$: value to a potential node-t + 1

$\beta_{t+1}[y_{t+1}]$: value backward pass all t + 1

After the local probability value of each label is possible for each of data sequences x x_t is obtained, then the value of the first derivative of the log-likelihood be obtained by equation 6 and 5. Then the value of the first derivative of the log-likelihood

G_k^x will be used to renew the function parameters λk features at each iteration based on the equation

$$\lambda_k \leftarrow \lambda_k + \omega G_k^x \quad (13)$$

$$\lambda_k \leftarrow \lambda_k + \omega G_{k'}^x \quad (14)$$

Information :

ω : learning rate with ω ε [0.001,0.1] [8].

G_k^x : gradient value node

$G_{k'}^x$: the value of the gradient edge

3.3.2 Inferen

The next process after getting the parameter values of the features of the training process that implements the parameter values in the data pengujian.Himpunan feature function parameters obtained from the training process will be used to perform inferen against a set of test data that would predict all label by finding sequences label the most optimal whole. There are three stages to perform the decoding is to calculate the potential function, maximal forward pass, and backtracking.

In early iterations, α_1 ^ max [x_1] initialized with 1/S for each label with S is the number of labels that may be given to the sequence of test data. Furthermore, the forward pass maximal searches performed on sequence data is mathematically written:

For t = T

$$\alpha_1[y_1] = 1/S \quad (14)$$

For t = 2 to T-1

$$\alpha_t^{max}[y_t] = \kappa_t max_{yt-1}(\alpha_{t-1}[y_{t-1}^{max}]\phi_t(y_t,x)\psi_t(y_t,y_{t+1},x)) \quad (15)$$

Information :

S : the number of labels

κ_t : normalization factor

$\alpha_{t-1}[y_{t-1}^{max}]$: the maximum value of the forward pass t-1

$\phi_t(y_t,x)$: the value of the node potential all t

$\psi_t(y_t,y_{t+1},x)$: the value of the potential edge all t

In the next stage of backtracking, aims to trace back and review the optimal label on each data based on the maximum value of the forward pass were done previously. Backtracking mathematically equation can be written as:

$$y_t^* = arg\ max_{y_t}(a_t^{max}[y_t]\ \phi_t(y_t,z)) \quad (16)$$

so that the optimal sequence of labels is $\{y_1^*, y_2^*, \dots, y_T^*\}$ for the sequence number of the test data T.

Where :

$a_t^{max}[y_t]$ = the maximum value of the forward pass all t

$\phi_t(y_t,z)$ = value to a potential node-t

a.4 Examination

This study using the test method precision, recall, and F-measure. Calculations of precision, recall, and F-measure is as follows [10].

$$Precision = \frac{true\ positif}{\text{true positif+ false positif}} \quad (17)$$

$$Recall = \frac{\text{true positif}}{\text{true positif + false negatif}} \quad (18)$$

$$F - Measure = \frac{true\ positif + true\ negatif}{true\ positif + false\ positif + true\ negatif + false\ negatif} \quad (19)$$

For entities person, location, and oganization identified correctly counted as true positive, while the one counted as a false positive. As for the other entities identified correctly counted as true negative, while one counted as a false negative.

2.5 Scenario Testing

Testing is done by finding the optimal parameter values with toggle combinations of parameters are included. Test plan is carried out as follows:

a. Examining the effect of standard deviation and the learning rate using training data and test data similar to the feature class of the word now (feature 1), feature word classes now and grade the previous word (feature 2), and features a class of words now, the class said before, and after the word class (feature 3)
b. Examining the effect of standard deviation and the learning rate using training data and test data differently to feature 1, 2, and 3.

3.2 Evaluation Testing

Tests using the standard deviation in the value range of 0.1 to 10 and learning rate is in the range 0.001 to the value of 0.1 [8]. Standard deviation of 0.1 Gauss tested value with changes in value multiple of 3. As for the learning rate parameter testing is given a value of 0.1 prior learning rate. After that, it was tested with multiple 10-1 to see the value of learning rate that will produce the most good value. The test results using training data and test data similar to the first feature, features 2 and 3 features can be seen in the Table 1,2, and 3.

Table 1 Results of Testing Features 1

σ	ω	P	R	F
0.1	0.0001	0	0	87.04
	0.001	0.35	19.23	87.07
	0.01	0	0	84.92
	0.1	0	0	0
3	0.0001	42.7	61.86	89.31
	0.001	58.86	21.21	66.8
	0.01	24.28	10	62.39
	0.1	12.7	40	86.41
6	0.0001	42.7	61.86	89.31
	0.001	58.86	21.21	66.8
	0.01	58.86	21.32	66.99
	0.1	6.77	26.23	85.65
9	0.0001	42.7	61.86	89.31
	0.001	58.86	21.21	66.8
	0.01	58.86	21.32	66.99
	0.1	6.77	26.23	85.65

σ	ω	P	R	F
12	0.0001	42.7	61.86	89.31
	0.001	58.86	21.21	66.8
	0.01	58.86	21.32	66.99
	0.1	41.35	68.38	90.06

Table 5 The Test Features Features 2

σ	ω	P	R	F
0.1	0.0001	0.35	18.52	87.06
	0.001	0.35	14.71	87
	0.01	0	0	82.71
	0.1	0	0	0
3	0.0001	4.59	30.66	86.48
	0.001	32.46	26.65	79.95
	0.01	32.46	26.65	79.95
	0.1	21.74	19.36	78.42
6	0.0001	4.59	30.66	86.48
	0.001	32.46	26.65	79.95
	0.01	43.9	32.55	81.2
	0.1	32.46	26.54	79.88
9	0.0001	4.59	30.66	86.48
	0.001	32.46	26.65	79.95
	0.01	43.9	32.55	81.2
	0.1	32.46	26.45	79.83
12	0.0001	4.59	30.66	86.48
	0.001	32.46	26.65	79.95
	0.01	43.9	32.55	81.2
	0.1	43.9	32.33	81.08

Table 6 The Test Results Features 3

σ	ω	P	R	F
0.1	0.0001	1.55	30.56	86.97
	0.001	0.21	1.47	85.43
	0.01	0	0	82.55
	0.1	0	0	0
3	0.0001	4.52	29.91	86.44
	0.001	4.59	29.95	86.43
	0.01	4.59	29.55	86.41
	0.1	7.55	24.94	85.28
6	0.0001	4.52	29.91	86.44
	0.001	4.59	29.95	86.43
	0.01	12.42	47.44	87.05
	0.1	3.46	17.56	85.59
9	0.0001	4.52	29.91	86.44
	0.001	4.52	29.63	86.42
	0.01	12.42	47.44	87.05
	0.1	4.59	19.82	85.43
12	0.0001	4.52	29.91	86.44
	0.001	4.52	29.63	86.42
	0.01	12.42	47.44	87.05
	0.1	8.26	29.32	85.73

Implementation method of conditional random fields in recognizing the entity named in the Indonesian language text showed the best accuracy using training data and test data similar to the features 1 with a deviation of 12 and a learning rate of 0.1 was 90.53%, which can recognize the entity names of people, locations, and organizations as much as 1,489 words of the 4,719 words, to feature two entities that can be identified as many as 137 words of the 4,582 words the entity with an accuracy of 89.71% with a deviation of 0.1 and a learning rate 0.0001, and features 3 accuracy obtained to reach 89.13% by an entity that can be identified as many as 103 words of 4,616 words learning rate deviation of 0.1 and 0.0001. From these results can be seen in the feature 2 and 3 has a value greater accuracy although only recognize a few words of entities than feature 1. This is because the feature 2 and 3 recognize words instead of one entity is predicted respectively 91 108 words and words with deviation election and learning the same rate that is 0.1 and 0.0001.

1 features the best accuracy results using training data and testing the different acquired 90.06% with a deviation of 12 and a learning rate of 0.1 that can be recognized entity of 586 words 1,417 words entity. For features 2 dik

Implementation method of conditional random fields in recognizing the entity named in the Indonesian language text showed the best accuracy using training data and test data similar to the features 1 with a deviation of 12 and a learning rate of 0.1 was 90.53%, which can recognize the entity names of people, locations, and organizations as much as 1,489 words of the 4,719 words, to feature two entities that can be identified as many as 137 words of the 4,582 words the entity with an accuracy of 89.71% with a deviation of 0.1 and a learning rate 0.0001, and features 3 accuracy obtained to reach 89.13% by an entity that can be identified as many as 103 words of 4,616 words learning rate deviation of 0.1 and 0.0001. From these results can be seen in the feature 2 and 3 has a value greater accuracy although only recognize a few words of entities than feature 1. This is because the feature 2 and 3 recognize words instead of one entity is predicted respectively 91 108 words and words with deviation election and learning the same rate that is 0.1 and 0.0001.

1 features the best accuracy results using training data and testing the different acquired 90.06% with a deviation of 12 and a learning rate of 0.1 that can be recognized entity of 586 words 1,417 words entity. For two recognizable features five said entities and 9,647 words is the entity that obtained the best accuracy is 87.06% with the deviation and the selected learning rate is 0.1 and 0.0001. Features 3 has the best accuracy value when a deviation 6, 9, 12 and 12:01 learing rate of 87.05% with an entity recognized word for word 116 of 1,241 entities.

3. CONCLUSION

From the research that has been done can be concluded simulator named entity recognition in the Indonesian language text using conditional random fields method can be performed using the word class features. Rated the best accuracy for every difitur

using training data and test data are the same or different, namely:

a. Values greater accuracy using training data and test data are the same for class features the word now is 90.53%, said the current class features and previous word class is 89.17%, and now word-class features, class and grade of the previous word after word was 89.13% ,
b. Values greater accuracy using training data and test data are different for class features the word now is 90.06%, said the current class features and previous word class is 87.06%, and now word-class features, the previous word classes and word-class 87.05% thereafter.

Suggestions for subsequent research are as follows.

a. Subsequent research can use the shape of words (word shape), n-grams, morphology, nonmorfologi, or use other features to enhance the value of accuracy.
b. Subsequent research can use Baum-Weltch algorithm to adjust the model obtained from the training process becomes maximum.

BIBLIOGRAPHY

[1] A. F. Wicaksono and A. Purwarianti, "HMM Based Part-of-Speech Tagger for Bahasa Indonesia," in *Proceeding of the 4th-International MALINDO Workshop (MALINDO2010)*, 2010.

[2] Y. Syaifudin, "Identifikasi Kalimat Kutipan Dari Teks Berita Online Berbahasa Indonesia Dengan Metode Berbasis Aturan," Skripsi, Ilmu Komputer, UGM, Yogyakarta, 2016.

[4] J. R. Finkel, T. Grenager and C. D. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[5] A. Ekbal, R. Haque and S. Bandyopadhyay, "Named Entity Recognition in Bengali: A Conditional Random Field Approach," in *IJCNLP*, 2008.

[6] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields," *Machine Learning*, vol. 4, no. 4, pp. 267-373, 2011.

[7] H. M. Wallach, "Conditional Random Fields: An Introduction," Technical Reports (CIS), 2004.

[8] T. T. Truyen and P. Dinh, "A Practitioner Guide to Conditional Random Fields for Sequential Labelling," Curtion University of Technology, 2008.

[9] "Kemendikbud, n.d., Pedoman Umum Ejaan Bahasa Indonesia," [Online] Available : http://badanbahasa.kemdikbud.go.id/lamanbahasa/sites/default/files/pedoman_umum-ejaan_yang_disempurnakan.pdf, [Accesed 3 Januari 2017].

[10] E. Prasetyo, Mengolah Data Menjadi Informasi Menggunakan Matlab, Yogyakarta: ANDI, 2014.