Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information). 9/1 (2016), 17-25
DOI: http://dx.doi.org/10.21609/jiki.v9i1.362

# WEB NEWS DOCUMENTS CLUSTERING IN INDONESIAN LANGUAGE USING SINGULAR VALUE DECOMPOSITION-PRINCIPAL COMPONENT ANALYSIS AND ANT ALGORITHMS

**Arif Fadllullah[1,2], Dasrit Debora Kamudi[1,3], Muhamad Nasir[1,4], Agus Zainal Arifin[1], and Diana Purwitasari[1]**

[1]Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Kampus ITS Sukolilo, Surabaya, 60111, Indonesia
[2]Universitas Borneo Tarakan, Jl. Amal Lama 1, Tarakan, 77115, Indonesia
[3]Politeknik Negeri Nusa Utara, Jl. Kesehatan 1 Tahuna, Sangihe, 95812, Indonesia
[4]Politeknik Negeri Bengkalis, Jl. Bathin Alam-Sungai Alam, Bengkalis, 28711, Indonesia

E-mail: arif14@mhs.if.its.ac.id

**Abstract**

Ant-based document clustering is a cluster method of measuring text documents similarity based on the shortest path between nodes (trial phase) and determines the optimal clusters of sequence document similarity (dividing phase). The processing time of trial phase Ant algorithms to make document vectors is very long because of high dimensional Document-Term Matrix (DTM). In this paper, we proposed a document clustering method for optimizing dimension reduction using Singular Value Decomposition-Principal Component Analysis (SVDPCA) and Ant algorithms. SVDPCA reduces size of the DTM dimensions by converting freq-term of conventional DTM to score-pc of Document-PC Matrix (DPCM). Ant algorithms creates documents clustering using the vector space model based on the dimension reduction result of DPCM. The experimental results on 506 news documents in Indonesian language demonstrated that the proposed method worked well to optimize dimension reduction up to 99.7%. We could speed up execution time efficiently of the trial phase and maintain the best F-measure achieved from experiments was 0.88 (88%).

**Keywords:** *web news documents clustering, principal component analysis, singular value decomposition, dimension reduction, ant algorithms*

**Abstrak**

Klasterisasi dokumen berbasis algoritma semut merupakan metode klaster yang mengukur kemiripan dokumen teks berdasarkan pencarian rute terpendek antar node (*trial phase*) dan menentukan sejumlah klaster yang optimal dari urutan kemiripan dokumen (*dividing phase*). Waktu proses *trial phase* algoritma semut dalam mengolah vektor dokumen tergolong lama sebagai akibat tingginya dimensi, karena adanya masalah *sparseness* pada matriks *Document-Term Matrix* (DTM). Oleh karena itu, penelitian ini mengusulkan sebuah metode klasterisasi dokumen yang mengoptimalkan reduksi dimensi menggunakan Singular Value Decomposition-Principal Component Analysis (SVDPCA) dan Algoritma Semut. SVDPCA mereduksi ukuran dimensi DTM dengan mengkonversi bentuk *freq-term* DTM konvensional ke dalam bentuk *score-pc Document-PC Matrix (*DPCM). Kemudian, Algoritma Semut melakukan klasterisasi dokumen menggunakan *vector space model* yang dibangun berdasarkan DPCM hasil reduksi dimensi. Hasil uji coba dari 506 dokumen berita berbahasa Indonesia membuk-tikan bahwa metode yang diusulkan bekerja dengan baik untuk mengoptimalkan reduksi dimensi hingga 99,7%, sehingga secara efisien mampu mempercepat waktu eksekusi *trial phase* algoritma se-mut namun tetap mempertahankan akurasi F-*measure* mencapai 0,88 (88%).

**Kata Kunci:** *klasterisasi dokumen web berita, principal component analysis, singular value decomposition, reduksi dimensi, algoritma semut*

## 1. Introduction

The emergence of internet has brought great changes in terms of the information presentation. Online sites become popular publications, for reliability in offering a variety of information quickly, hot, and plentiful with a wider range of readers. The fact that the number of articles is very large and disorganized would cause difficulty to the readers to find news relevant to the topic of the interest. For that, we need an automatic clustering for news web documents.

The problem of clustering in large datasets is a combinatorial optimization problem that is difficult to be solved with conventional techniques. Hierarchical and partition based clustering are the common clustering categories. 'K' means algorithm is the very popular partition based clustering algorithm. However, there are few limitations observed with this 'K' means algorithm from literature. The limitations are (a) It fails to scale with large datasets. (b) The quality of the algorithm results highly depends on the initial number of static clusters [1].

Ant algorithm is a universal solution which is first designed to overcome Traveling Salesman Problem (TSP). This was proposed to find the shortest path, which the ants will choose the town (node) to be traversed by the probability function between the distance of nodes that must be taken and how much the level of trace pheromones. Ant algorithms is applied on clustering of text documents, by analogying documents as graph nodes. Idea of finding shortest path of those document nodes has triggered a clustering method which is known as Ant-based document clustering method [2-3]. This method consist of two phases, i.e. finding documents most alike (trial phase) and clusters making (dividing phase). The advantages of this algorithm are the ability to classify documents in unsupervised learning and flexibility to determine the initial number of clusters [3].

Before entering into a stage of Ant algorithms, VSM is required. VSM would build the relationship between the number of documents and terms of all documents in DTM (Document-Term Matrix). DTM is a structured data matrix obtained by forming a number of terms as column and documents as row. If all the terms entered into DTM, then the size of the document vectors will become larger. It will take a long time and increase the complexity of the trial phase calculation in Ant algorithms. For that reason, we need a strategy to optimize execution time of the trial phase Ant algorithms. One common approach is dimension reduction of the stop word and stemming specific to a particular language.

Stop word is used to remove the terms that they appear that often occur in for the entire document, but cannot be a significant feature of each document. While stemming is used to change affixes-word into basic-word, so that a number of affixes-word which has the same basic-word can be grouped into a single term. Most of the stemmer frame-work which has been standardized for English language documents, while the attention of the Indonesian domain is still relatively low, and continues to undergo development. Unlike English, Indonesian language has more complex category of affixes, which includes prefixes, suffixes, infixes (insertions), and confixes (combinations of prefixes and suffixes).

Several Indonesian language stemmer had been developed by Nazief and Adriani (1996), Vega, et al.(2001), and Arifin and Setiono (2002) [4]. In addition, there is also confix stripping (CS) stemmer made by Jelita Asian (2007) [5], and enhanced confix stripping (ECS) stemmer was an improvement of CS stemmer by Arifin, Mahendra and Ciptaningtyas (2009) [6].

Stemmer ECS produced document vectors with reduced terms size up to 32.66% of the 253 news documents in Indonesian language. It could create document clusters in dividing phase with F-measure of 0.86 [6]. Only the complexity of trial phase in document vectors processing is still comparatively high, although the term has been reduced by ECS stemmer. This happens because not only many documents and terms are processed, but also the sparse terms problem in DTM. Sparseness occurs because not all terms are in all documents, so that most conditions of the DTM term value is zero.

In this paper, we proposed a document clustering method for optimizing dimension reduction using Singular Value Decomposition Principal Component Analysis (SVDPCA) and Ant algorithms. SVDPCA will reduce the size of the DTM dimensions by converting freq-term of conventional DTM to score-pc of Document-PC Matrix (DPCM) without sparseness based on cumulative proportion of variance. Then, Ant algorithms will create document clustering using the vector space model based on the dimension reduction result of DPCM. The proposed method is expected to overcome the dimension reduction problem because a lot of affixes word and sparseness in DTM. Therefore it would speed up the execution time of the trial phase Ant algorithms and improve F-measure news documents clustering in Indonesian language.

**Enhanced Confix Stripping (ECS) Stemmer**

ECS stemmer is a stemmer framework as the improvement of the previous confix stripping stemmer. It does stemming in the news documents in Indonesia language. In this method affixes-word can be recorded into basic words, so that a number of affixes-word which has the same basic-word can be grouped into a single term [6]. The steps in the process of ECS stemmer recording exactly the same as the previous stemmer [5]. However, this stemmer corrects some examples of words that failures stemming by stripping stemmer confix previous methods as follows:

1. No prefix removal rule for words with construction of "mem+p...", for example, "mem-

TABLE 1
DTM FROM THE COLLECTION OF NEWS DOCUMENTS IN INDONESIAN LANGUAGE

| Document | Perang | tembak | … | emas | banjir | Bencana |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | … | 0 | 0 | 0 |
| 2 | 0 | 0 | … | 4 | 0 | 0 |
| … | … | … | … | … | … | … |
| 505 | 0 | 3 | … | 0 | 0 | 1 |
| 506 | 0 | 0 | … | 0 | 1 | 2 |

promosikan", "memproteksi", and "memprediksi".

2. No prefix removal rule for words with construction of "men+s...", for example, "mensyaratkan", and "mensyukuri".
3. No prefix removal rule for words with construction of "menge+...", for example, "mengerem".
4. No prefix removal rule for words with construction of "penge+...", for example, "pengeboman".
5. No prefix removal rule for words with construction of "peng+k...", for example, "pengkajian".
6. Suffix removal failures – sometimes the last fragment of a word resembles certain suffix. For examples, the words like "pelanggan" and "pelaku" failed to be stemmed, because of the "-an" and "-ku" on the last part of the word should not be removed.

**Weighting *Document-Term Matrix* (DTM)**

DTM is a mathematical matrix that describes the frequency of term co-occurrence in documents. In DTM, a number of terms is defined as column and documents as row (see Table 1).

TF-IDF (combined with a term frequency-inverse document) method has shown better performance when compared with the binary method and frequency. In TF-IDF, $w_{ij}$ represents weight of term $i$ on document $j$ which is expressed in equation(1) [7]:

$$w_{ij} = tf_{ij}.\log_2\left(\frac{N}{df_i}\right) \quad (1)$$

where $tf_{ij}$ represents frequency of term $i$ on document $j$. $N$ represents number of processed documents and $df_j$ represents number of documents that actually have term $i$ at least one.

**Singular Value Decomposition-Principal Component Analysis (SVDPCA)**

Sparseness occurs because not all terms are in all documents, so that most conditions of DTM term value is zero (see Table 1). SVDPCA is a mechanism of DTM-dimensional transformation which sparse terms into a dimension of Document PC Matrix (DPCM) without sparse (not redundant and correlated) by utilizing the calculation of Principal Component Analysis (PCA), where eigenvalues and eigenvectors is not searched by covariance matrix, but utilizing the left eigen-vectors results of Singular Value Decomposition (SVD). Dimensional DPCM then reduced by selecting the number of score-pc when their cumulative proportion of variance is more than 50% (see Table 2). [8]

TABLE 2
DPCM WITH THREE PCS

| Document | PC1 | PC2 | PC3 |
|---|---|---|---|
| 1 | -7.05324 | -1.24343 | -3.675594 |
| 2 | 2.65687 | 4.434534 | 1.98538 |
| … | | | |
| 505 | -6.2431 | -1.23234 | 2.322449 |
| 506 | -2.4321321 | -3.129341 | 1.321393 |

PCA is an orthogonal linear transformation that maps given data onto the PC system, which consists of first PC, second PC, and so on, according to the order of their variance values [9]. PCA estimates the transformation matrix S as the equation(2).

$$X(N \times m) = S^T A(N \times p) \quad (2)$$

DTM $A$ is transformed to DPCM $X$ by $S$, which generates column dimensions $m$ is smaller than $p$ when using only conventional DTM. To implement classic PCA into $A$ with dimensions $N \times p$, $N$ has to be larger than $p$. If $N$ is smaller than $p$, then we cannot achieve convergence calculations for PCA. Therefore, we add SVD (Singular Value Decomposition) to overcome this problem. In the SVD transformation, the original matrix can be decomposed (similar to PCA eigen-decomposition) into three matrix components with the same size. If they was multiplied again, so they will produce the same value as the original matrix [10]. The decomposition of A is shown in equation(3):

$$A = P.D.Q^T \quad (3)$$

$P$ and $Q$ are $N \times N$ and $p \times p$ orthogonal matrices, respectively. D is a $(N \times p)$ matrix with singular value $d_{ij}$ shown in equation(4):

$$d_{ij} = \begin{cases} d_{ij} \geq 0, & i = j \; and \; i < r \\ 0, & i \neq j \; or \; i > r \end{cases} \quad (4)$$

The rank of $A$ is $r$, the columns of $P$ and $Q$ are the left and right eigenvectors of $XX^T$ and $X^TX$, respectively. The singular values on the dia-

gonal D is the square root of the eigenvalues of $AA^T$ and $A^TA$. The classic PCA depends on the decomposition of covariance (or correlation) matrix S by PCA eigenvalues decomposition. In fact, the covariance matrix S calculation of DTM will obviously take a long time, due to the large dimensions of DTM. While SVDPCA, we do not need to carry out covariance matrix S execution, because its value can be extracted from the matrix P which is left eigenvectors of the SVD results as space data records used to find PCA scores. As a SVDPCA result, we are not restricted by the constraint that N being larger than p [8].

**Ant algorithms**

Ant algorithms typically used to solve the Traveling Salesman Problem (TSP) algorithm that is inspired by the behaviour of ant colonies in search for food sources (documents) by leaving a trail pheromone [11].

In the case of Euclidean TSP, $d_{ij}$ is the euclidean distance between towns i to j, i.e. $\boldsymbol{d_{ij}} = \left[(x_i - x_j)^2 + (y_i - y_j)^2\right]^{1/2}$. Let *m* be the total number of ants. Each ant is a simple agent with the following characteristics: First, it chooses the town (node) that will be passed next, with a probability function between the distances that must be taken and how the level of existing pheromone trail. Second, it is given a memory of the towns that have been passed (by using tabu list) and the last after completing a full path, the path of the ants are given a number of pheromone [6].

Let $\tau_{ij}(t)$ be the amount of pheromone trail on edge (*i,j*) at time *t*. Pheromone trail on each edge is updated on each tour, cycle, or iteration according to the equation(5):

$$\tau_{ij}(t+n) = \rho.\tau_{ij}(t) + \Delta\tau_{ij} \quad (5)$$

where ρ is a coefficient such that (1- ρ) represents the evaporation of pheromone trail between time *t* and *t+n*. The range of coefficient $\rho$ is (0.1). Δ is a segment that have been passed by ant *k* as part of the trajectory of the nest towards food sources. The decline in the amount of pheromone allows ants to explore different paths during the search process. It will also eliminate the possibility of choosing the bad path. In addition, it can also help limit the maximum value is achieved by a pheromone trajectory. The total amount of pheromone trail laid on edge (*i, j*) defined in equation(6).

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^k \quad (6)$$

where $\Delta\tau_{ij}^k$ is the quantity per unit length of pheromone trail laid on edge (*i,j*) by the *k-th* ant between time *t* and *t+n*; it is given by the equation (7).

$$\Delta\tau_{ij}^k \begin{cases} Q/L_k, & \\ if\ k-th\ ant\ use\ edge(i,j) & \\ in\ its\ tour & \\ 0, \quad otherwise & \end{cases} \quad (7)$$

*Q* is a constant and $L_k$ is the tour length of the *k-th* ant. Each ant equiped with a data structure called the tabu list $tabu_k$, that saves the towns already visited. After each complete cycle, the tabu list is used to compute the ant's current solution, to get the shortest path route and its length. The transition probability from town i to town j for the *k-th* ant is defined as the equation(8).

$$p_{ij}^k(t) = \frac{[\tau_{ij}(t)]^\alpha [\eta_{ij}]^\beta}{\sum[\tau_{ik}(t)]^\alpha [\eta_{ik}]^\beta}, k \in allowed_k \quad (8)$$

Where $\eta_{ij}$ is the visibility of $1/d_{ij}$, $allowed_k = \{N\text{-}tabu_k\}$, $\alpha$ and $\beta$ are parameters that control the relative importance of pheromone trail and visibility where $\alpha \geq 0$ and $\beta \geq 0$.

**Ant-Based Document Clustering**

Ant-Based Document Clustering is divided into two phase, which are finding the shortest path between the documents (trial phase) and separate a group of documents alike (dividing phase) based on previous trial phase result [1-3] and [6].
The finding shortest path adopts Ant algorithms trial phase, where the ant *k* probability ($p_{ij}^k$) to select a node (document) *j* from a node *i* position is defined as the equation(9).

$$p_{ij}^k(t) = \frac{[\tau_{ij}(t)]^\alpha [S_{ij}]^\beta}{\sum[\tau_{ik}(t)]^\alpha [S_{ik}]^\beta}, k \in Z_k \quad (9)$$

$Z_k$ represents set of nodes (documents) that is not visited yet by k-th ant, $\tau_{ij}(t)$ represents pheromone trail on edge (*i,j*), $\alpha$ represents parameter that control the importance of pheromone trail, and $\beta$ represents visibility parameter. $S_{ij}$ represents cosine distance between node (*i,j*) which is defined as the equation(10).

$$S_{ij} = \frac{\sum_k [w_{ki}] \cdot [w_{kj}]}{\|d_i\|^2 \cdot \|d_i\|^2} \quad (10)$$

where $w_{ki}$ and $w_{kj}$ are score-pc weighting of *k*-th pc on document *i* and *j* of DPCM. $\|d_i\|^2$ and $\|d_j\|^2$ are the length of document vector *i* and *j*. For exa-

mple, if $||d_i||^2 = (t_1^2 + t_2^2 + t_3^2 + ... + t_k^2)^{1/2}$, so $t_k$ is *k*-th pc of $d_i$ document vector.

After the search nodes by ants in one iteration is complete, the next step is to add the amount of pheromone on edge passed by each ant, and evaporated. In contrast to Lukasz Machnik [2] [3], Arifin, et al [6] research to modify the method of increasing the number of pheromone (see formula 12) on edge *i,j* ($\Delta\tau_{ij}$), which is defined by equation(11).

$$\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^{k} \tag{11}$$

where $\Delta\tau_{ij}^{k}$ is the amount of phremone trail laid on *edge* (*i,j*) by the k-th ant, which is defined as the equation(12).

$$\Delta\tau_{ij}^{k} \begin{cases} N \times L_k, \\ if\ k-th\ ant\ use\ edge(i,j) \\ in\ its\ tour \\ 0, \quad otherwise \end{cases} \tag{12}$$

where *N* is the total documents and $L_k$ is the total route length of k-th ant. The evaporation of pheromone trail is given by equation(13):

$$\tau_{ij}(t+n) = (1-\rho).\tau_{ij}(t) + \Delta\tau_{ij} \tag{13}$$

where $\rho$ is the pheromone evaporation coefficient.

It should be noted that the best shortest path is route that has the largest total similarities. It is because the length of path is calculated using cosine distance similarity between documents.

In the next phase, dividing phase is the formation of clusters based on the order of the documents obtained from the previous trial phase. The first document of that order, is regarded as the centroid $\mu$ of the first cluster. The clustering process begins from $\mu$ and the next document in sequence (called a comparative document D) is defined as equation(14).

$$\delta < \cos(\mu, D) \tag{14}$$

where $\delta$ is the attachment coefficient that has value range at (0,1), and *cos(μ,D)* is cosine distance of document $\mu$ and *D*.

If the condition is true, then the document *D* into a group in centroid $\mu$, and the next document in sequence into the next document *D*. If the condition is false, then the current document *D* becomes new centroid $\mu$ for the new cluster. The testing process is repeated to the next document until finished when the whole sequence of documents is done.

## 2. Methods

### Documents clustering using SVDPCA and Ant Algorithms

Figure 1 shows order of the proposed method that consisting of some stages, where the first stage begins with the collection of corpus data as input is derived from web news documents in Indonesia language. Corpus data was consisted of 506 news

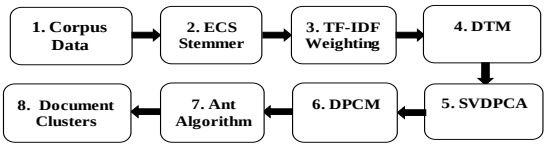


Figure 1. Design of the Proposed Method

TABLE 3
THE CORPUS CLASS WITH MANUALLY IDENTIFIED EVENTS

| ID | Event | Medoid | Count file |
|---|---|---|---|
| 1 | Kasus sodomi Anwar Ibrahim | Tunangan Saiful Bicara Tentang Kasus Sodomi Anwar Ibrahim | 52 |
| 2 | Kasus suap Artalyta | Jadi "Sinterklas", Artalyta Bagi-bagi Makanan di Pengadilan | 54 |
| 3 | Bencana gempa bumi di China | Hampir 900 Terperangkap dan 107 Tewas di Gempa China | 38 |
| 4 | FPI dalam insiden Monas | Kasus Kekerasan oleh FPI | 70 |
| 5 | Konflik Israel-Palestina | Menteri Israel Tolak Gencatan Senjata di Gaza | 50 |
| 6 | Dua tahun tragedi lumpur Lapindo | Pengungsi Lapindo Tuntut Jatah Makan Dilanjutkan | 42 |
| 7 | Badai nargis di Myanmar | Yangon Rusak Parah, 351 Orang Tewas | 48 |
| 8 | Pengunduran jadwal Pemilu | Jadwal Pemilu Tidak Berubah Cuma Digeser | 20 |
| 9 | Perkembangan harga emas | Berkilaunya Investasi Emas | 36 |
| 10 | Kiprah Khofifah dalam Pilkada Jatim | Khofifah, Perempuan Lembut Setangguh Lelaki | 30 |
| 11 | Kematian mantan presiden Soeharto | Kematian Soeharto Tragedi Bagi Korbannya | 56 |
| 12 | Festival film Cannes | The Class Raih Penghargaan Tertinggi Cannes | 10 |

from www.kompas.com, www.detik.com, www. wartakota.com and www.jawapos.com .Then, the corpus data processed by stop list and stemming using ECS stemmer. Stop word is used to remove the terms that they appear often for the entire document, but cannot be a significant feature of each document. While stemming is used to change affixes word into basic word, so that a number of affixes word which has the same basic word can be grouped into a single term. We also use stop list with 792 stop words and Indonesian language dictionary with 29337 basic-words of the proposed ECS stemmer.

After all terms of basic-word has been formed, then the term value processed by TF-IDF weighting on each document to make DTM. Stages of the fifth, sixth and seventh is proposed contribution of our research, which serves to overcome the sparseness problem when reducing the DTM dimensions, so that the executing time of cluster formation during process of finding the shortest path in the trial phase Ant algorithms can be optimized. SVD is used to find the set of orthogonal that divided into two spaces, right eigenvectors (Q) for a range of dimensions space and left eigenvectors (P) for a range of data records space. P values in the SVD is used as a data matrix S in the calculation of PCA. Then PCA transform $S^T$ which maps original data $X$ normalized (each term minus the average term) into new dimensions of data, by changing the TF-IDF weighting of term into a score-pc and reduce the dimensions $S^TA$ based on significant score-pc. This selection is based on the largest proportion of variance of score-pc, so obtained a new DPCM without sparseness and dimensions have been reduced. In our research, the threshold limit of cumulative proportion of variance modifies the cumulative proportion limit of previous research [8], when it over 40%.

Furthermore, Ant algorithms divides the cluster process into two steps. The first step is to find the shortest path in the trial phase, where the results of the DPCM are converted to vector space model. Then search distances based on some rule of trial phase (see section Ant-Based Document Clustering) and the best shortest path that produces the greatest of total similarities. Due to the large number of parameters on how the real total similarities value or actual largest path, the parameters were tested regarding Marco Dorigo experiment in shortest path searching to the TSP using Ant algorithms [11]. The parameter values are: α= 2, β= 5, ρ= 0.5, ants=30, and the number of iterations or maximum rotation of 100. The second step of the dividing phase Ant algorithm that makes clusters based on the order of the documents obtained from the previous phase. The first document is regarded as the centroid $\mu$ of the first cluster. The clustering process begins from $\mu$ and the next document in sequence based on the condition $\delta$ (see formula 14) with parameter value is 0.008, until generated news documents clustering.

To verify whether documents clustering obtained by the proposed method is successful, so we require evaluation techniques. Evaluations of information retrieval are recall, precision and F-measure. Each cluster is generated by the proposed method is considered as the retrieval result. The manual-predefined class of documents as the ideal cluster that should be retrieved [4][6] with 12 manually identified news events from previous research [6] (see Table 3).

Specifically, for each predefined class *i* and cluster *j*, the recall and precision are defined as equation(15) and equation(16).

$$Rec(i,j) = R_{ij} = n_{ij}/n_i \tag{15}$$

$$Prec(i,j) = P_{ij} = n_{ij}/n_j \tag{16}$$

where $n_{ij}$ is the number of documents of class *i* in cluster *j*, $n_i$ is the number of documents of class *i*, and $n_j$ is the number of documents of cluster *j*. The F-measure of class *i* on all cluster *j* is defined as the equation(17).

$$F_{ij} = (2 \times R_{ij} \times P_{ij})/(R_{ij} + P_{ij}) \tag{17}$$

Then, the overall F-measure is calculated by the following the equation(18).

$$F = \sum_i \frac{n_i}{n} \max\{F_{ij}\} \tag{18}$$

where *n* is the total number of documents, $\boldsymbol{n_j}$is the number of documents in class *i*, and $\boldsymbol{max\{F_{ij}\}}$ is the maximum value $\boldsymbol{F_{ij}}$ of class *i* on all cluster *j*. Because we will see increased performance of the proposed method, so the F-measure value and execution time of trial phase in our research will be compared with the previous method (Ant Algorithms without SVDPCA) [6].

## 3. Results and Analysis

Table 4 shows that ability ECS stemmer in reduce insignificant term. ECS stemmer could reduce the number of term up to 4758 terms, from a total of 7327 different terms in 506 corpus data. With the use of ECS stemmer, term had been reduced to 35%. Table 5 shows the standard deviation, proportion of variance (variance of each score-pc divided by the total variance), and the cumulative proportion of variance for each score pc results from SVDPCA. Score-pc sorted by descending,

TABLE 4
COMPARISON OF TOTAL TERM WITH AND WITHOUT ECS STEMMER

| Stop list Removal | Stemmer | Total Term |
|---|---|---|
| No | Without *Stemmer* | 7327 |
| Yes | ECS *Stemmer* | 4758 |

TABLE 5
VARIANCES OF THE TOP- FIFTEEN PCS

| PCs | Standard Deviation | Proportion of Variance | Cumulative Proportion |
|---|---|---|---|
| PC1 | 19.703 | 0.050 | 0.050 |
| PC2 | 19.394 | 0.048 | 0.098 |
| PC3 | 18.944 | 0.046 | 0.144 |
| PC4 | 15.635 | 0.031 | 0.175 |
| PC5 | 15.134 | 0.029 | 0.205 |
| PC6 | 14.545 | 0.027 | 0.232 |
| PC7 | 13.826 | 0.025 | 0.257 |
| PC8 | 13.379 | 0.023 | 0.279 |
| PC9 | 13.088 | 0.022 | 0.301 |
| PC10 | 12.465 | 0.020 | 0.321 |
| PC11 | 11.556 | 0.017 | 0.339 |
| PC12 | 11.295 | 0.016 | 0.355 |
| PC13 | 10.758 | 0.015 | 0.370 |
| PC14 | 10.503 | 0.014 | 0.384 |
| **PC15** | 10.258 | 0.014 | **0.397** |
| PC16 | 9.671 | 0.012 | 0.409 |
| PC… | … | … | … |

so that PC1 has the highest standard deviation or variance from total of variance data.

PC2 and PC3 have the highest second and third standard deviation. And so on, thus obtained PCs limit which have a cumulative proportion of variance which reached 0.4 (40%), that was PC15. Selection score-pc could convert DTM which contains 506 documents x 4758 term become DPCM which contains 506 documents x 15 score-pc. This shows that the SVDPCA had been successfully optimize term into score-pc until 99.7%. DPCM form are converted into vector space model and then processed by Ant algorithms until obtain some cluster. In the proposed method, vector space model for each document did not use 4758 term but only 15 score-pc.

Table 6 shows the performance of the proposed method (SVDPCA and Ant algorithms) with the parameter values were set $\alpha = 2$, $\beta = 5$, $\rho = 0.5$, 30 ants, the number of iteration is 100, and $\delta = 0.008$, where the proposed method successfully created 14 clusters. It was more than two clusters from total of 12 classes defined manually. Because of the cluster label of the proposed method does not represent the same label with the class, so we need to identify clusters by comparing the members of each class to each cluster. From the comparison, a total of 6 classes (1, 3, 8, 9, 10, 12) defined manually has been identified in 6 different clusters (13, 3, 2, 11, 7, 14) of the proposed method. Even 4 from 6 clusters (13, 2, 7, 14) have exactly the same or similar all document members

TABLE 6
OVERALL F-MEASURE OF THE PROPOSED METHOD

| Class | Cluster | Rec | Prec | F-measure | F-max | F-max Weighting |
|---|---|---|---|---|---|---|
| 1 | 13 | 1 | 1 | 1 | 1 | 0.10 |
| 2 | 9 | 0.13 | 0.14 | 0.13 | 0.90 | 0.10 |
| | 10 | 0.81 | 1 | 0.90 | | |
| 3 | 3 | 1 | 0.92 | 0.96 | 0.96 | 0.07 |
| 4 | 8 | 0.47 | 1 | 0.64 | 0.64 | 0.09 |
| | 9 | 0.52 | 0.61 | 0.56 | | |
| 5 | 1 | 0.93 | 1 | 0.96 | 0.96 | 0.09 |
| | 9 | 0.04 | 0.04 | 0.04 | | |
| 6 | 11 | 0.10 | 0.1 | 0.10 | 0.95 | 0.08 |
| | 12 | 0.91 | 1 | 0.95 | | |
| 7 | 3 | 0.04 | 0.05 | 0.05 | 0.96 | 0.09 |
| | 4 | 0.92 | 1 | 0.96 | | |
| 8 | 2 | 1 | 1 | 1 | 1 | 0.04 |
| 9 | 11 | 1 | 0.82 | 0.90 | 0.90 | 0.06 |
| 10 | 7 | 1 | 1 | 1 | 1 | 0.06 |
| 11 | 5 | 0.25 | 1 | 0.4 | | |
| | 6 | 0.54 | 1 | 0.70 | 0.70 | 0.08 |
| | 9 | 0.11 | 0.02 | 0.03 | | |
| 12 | 14 | 1 | 1 | 1 | 1 | 0.02 |
| | | | | Overall | F- | 0.88 |

TABLE 7
COMPARISON PERFORMANCE OF METHODS

| Description | SVD-PCA and Ant Algorithms | Ant algorithms (Arifin, etc, 2009) |
|---|---|---|
| SVDPCA | 6 sec | - |
| Graph (Network) | 5 sec | 1 minutes 13 sec |
| The Search Shortest Path | 8 minutes 56 sec | 8 minutes 35 sec |
| Total Time of Trial Phase | 9 minutes 7 sec | 9 minutes 48 sec |
| Cluster Obtained | 14 cluster | 14 cluster |
| F-measure | 0.88 | 0.78 |

with members of the class defined manually (1, 8, 10, 14), because each that clusters took the best value of recall and precision is 1. Furthermore, a total of 6 classes (2, 4, 5, 6, 7, 11) have the set document members spread to more than one cluster. If it happens, so we choose one cluster that has highest F-measure to became F-max for each class. The overall F-measure obtained by the sum of F-max weighting from all class where the result of this calculating performance of the proposed method was 0.88.

Table 7 shows comparison performance between proposed method (SVDPCA and Ant algorithms) and control method (only Ant algorithms) with the same parameter value. It shows the total time trial phase or finding the best path for 506 documents with the proposed method is faster

than the control method. This is due to the control method, terms that used to construct the vector graph is larger than the proposed method. It took 4758 terms from DTM for each document, so it needed 1 minutes 13 sec to calculate cosine distance between documents of that terms. While the proposed method only took 15 score-pc from DPCM for each document, so that the time required to calculate the cosine distance between documents just 5 sec, 1 minute 8 sec was faster than the control method.

However, the addition time of SVDPCA was not so affected because making vector graph (network) from matrix is very fast, so the total time trial phase of proposed method was 41 sec faster than the control method. This happens because the SVDPCA calculation can decrease the processing complexity for dimension reduction of DTM. If the classic PCA, we convert the DTM into DPCM in three stages. Firstly, we depend the decomposition of covariance (or correlation) matrix of DTM. Secondly, we have to calculate the eigenvalues using covariance matrix results, and the last we make eigenvectors based on eigenvalues that is used to get the DPCM matrix. While the SVDPCA, we just require one stage but gained three matrix at once, and choosing left eigenvector matrix to get the DPCM matrix, so that SVDPCA can support the efficiency of trial phase Ant algorithm.

In addition, with same parameters, both methods could make 14 clusters. However, the F measure performance of proposed method was 0.88 better than control method that only 0.78 (see table VII). The proposed method has proved that dimension reduction optimally by selecting the number of score-pc when their cumulative proportion of variance is more than 0.4. It has no effect to decrease in information quality of DPCM for making clustering documents. In fact, it could maintain F-measure is even better than control method with 30 ants and 100 iterations, although the limit cumulative proportion of variance suggested in previous study is 0.5 [8]. This is due to differences in the type and number of documents used previous study [8]. Although the control method did not reduce dimension of DTM, with parameters are 30 ants and 100 iterations, it could only reach the F-measure was 0.78. If the control method want to have the percentage of F-measure up to 80%, so it is still needed changes parameter values with more number of ants and iterations that would in turn be increase processing time when compared it with the proposed method.

The fact that the number of configuration parameters used in ant algorithm make difficult to determine the best configuration in order to make optimal cluster with the highest F-measure. Therefore, we plan to classifying news of documents with cluster algorithm that will develop classification method of documents based on Ant algorithms with supervised classification.

## 4. Conclusion

In this paper, we have proved that in order to speed up execution time of the trial phase Ant algorithms can be done by reducing sparse terms. SVDPCA could solve that sparse terms problem in DTM based acquisition term of ECS stemmer (taking 65% of all term) into a score-pc DPCM up to 99.7%, thus speeding up execution time of the trial phase Ant algorithms for finding the shortest path between documents. Then Ant algorithms could make document clusters with the best F-measure value achieved was 0.88 (88%) for the number of ants and iterations are not too large. The experimental results have shown that proposed method is more efficient and accurate.

## References

[1] Vaijayanthi, P., Natarajan, A.M., & Murugadoss, R., *"Ants for Document Clustering,"* International Journal of Computer Science (IJCSI), vol. 9, no. 2, pp. 493-499. 2012.

[2] L. Machnik, *"Ants in Text Document Clustering"* In Proceeding of Advances in Systems, Computing Sciences and Software Engineering, pp. 209-212, 2006.

[3] L. Machnik, *"Documents Clustering Method based on Ant algorithms"* In Proceeding of the Internetional Multiconference on Computer Science and Information Technology, 2006.

[4] A. Z. Arifin and A. N. Setiono, "Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algortima Single Pass Clustering" In Prosiding Seminar on Intelligent Technology and Its Application (SITIA), 2002.

[5] J. Asian, *"Effective Techniques for Indonesian Text Retrieval,"* PhD thesis School of Computer Science and Information Technology, RMIT, University Australia, 2007.

[6] A. Z. Arifin, I. P. A. K. Mahendra and H. T. Ciptaningtyas, *"Enhanced Confix Stripping Stemmer and Ant algorithms for Classifying News Document in Indonesia Language"* In Proceeding International Conference on Information and Communication Technology and Systems (ICTS), vol. 5, pp. 149-158, 2009.

[7] G. Salton, Automatic Text Processing, Addison Wesley, 1989.

[8] S. Jun, S.-S. Park and D.-S. Jang, *"Document clustering method using dimension reduction and support vector clustering to overco-*

*me sparseness,"* Expert Systems with Applications, vol. 41, pp. 3204–3212. 2014.

[9] V. Cherkassky and F. Mulier, *"Learning from data: concepts, theory, and Methods"*, John Wiley & Sons, 2007.

[10] R. V. Ramirez-Velarde, M. Roderus, C. Barba-Jimenez and R. Perez-Cazares, *"A Parallel Implementation of Singular Value Decomposition for Video-on-Demand Services Design Using Principal Component Analysis"* In Proceeding of International Conference on Computational Science (ICCS), vol 29, pp. 1876-1887, 2014.

[11] M. Dorigo, V. Maniezzo and A. Colorni, *"The Ant System: Optimization by a Colony of Cooperating Agents,"* IEEE Transactions on Systems, Man, and Cybernetics-Part B, vol. 26, no. 1, pp. 29-41. 1996.