Retraction

Retraction: Performance of news clustering using ant colony optimization (*J. Phys.: Conf. Ser.* 1566 012101)

Published 24 May 2022

This article has been retracted by the authors following an admission that this article is a duplicate publication [1].

IOP Publishing has investigated in line with the COPE guidelines, and agree that this article should be retracted.

The authors agree to this retraction.

[1] Simamora R. J, Girsang A.S, Jenson T 2019 Ant Algorithm Optimization for Clustering News Online, *2019 International Conference of Computer Science and Information Technology (ICoSNIKOM)*, pp 1-4

Retraction published: 24 May 2022
Performance of news clustering using ant colony optimization

T Jenson¹ and A S Girsang²

¹²Computer Science Department, BINUS Graduate Program-Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480,

Email : tomi.jenson@binus.ac.id , agirsang@binus.edu

Abstract. Humans are basically curious creatures. We always look for all the latest information through books or news. News usually contain tags or categories to make it easier to find similar news. It can be done manually by a human or with a machine to make it easier. Data mining techniques such as clustering and classification can be used to help the categorization process. Clustering is a technique that can be used to grouped news based on their similarity. There is some algorithm that can be used to clustering data. One of them is Ant Colony Optimization (ACO). ACO is a metaheuristic algorithm that mimics how ants bring food to the colony. ACO also can be used to solve the clustering problem. In this paper, we will use the ACO version for clustering to cluster news and find the performance of the said method. Then we compare the result with ground truth.

1. Introduction

News becomes an important part of society that can impact a lot of life. With the rapid development of information technology today, we can find news easily and quickly. Not only from original news sources, but we often get news from the results of posts that exist on various social media, forums, blogs, and chat applications. Especially on social media, the process of sharing news has become very easy and everyone is part of the distribution of the news to the global virtual community [1]. With this convenience, even the latest news is not difficult to get. The news we read usually categorized by editor based on the context of the news. The categorization process can be done manually by the editor or in this case some use machine learning to help the process [2]. Surely it would be better if there was human intervention in the categorization process because news must be seen semantically or in accordance with the meaning of the news. But how exactly is the result of using the machine to do that? Better or worse? Some data mining techniques such as clustering and classification can be used here. But what will be discussed in this research is clustering, because we want to eliminate label information so we can see the differences in machine and human performance.

Clustering is a data mining technique for grouping data in a dataset based on the characteristics and similarities (dissimilarity) of the data. Clustering is unsupervised learning which means there is no label on the data. The goal is to find out the structure of the distribution of the data to find out more about the data. One example of applying to a cluster is clustering on documents with the cosine similarity metric [3] and research by Luo et al. [4] with the cosine similarity metrics and neighbors and links. News is basically a document, which means clustering can also be used to classify news based on similarities (dissimilarity). The aim is to reduce information overload so that only relevant news will appear. But with a large amount of data, the clustering process will be longer and more complicated. A method is needed to help optimize the clustering process.

Ant Colony Optimization (ACO) [5] is a metaheuristic algorithm based on ant behavior used to find the best route between two nodes. Ant will find the best way between nest and food based on the
pheromone level dropped by its predecessor. Usually, ACO is used to solve the Travelling Salesman Problem (TSP) as that is the goal of the original algorithm. But as time goes by, there are a lot of ways found to use ACO with different goals in mind. One of them is to use ACO for clustering. Ant Colony Optimization for Clustering (ACOC) is an algorithm proposed by Shelokar [6] for clustering. The concept is that an ant is an agent which will find a solution in which each object in the solution associated with a certain cluster. The approach becomes one of the modern clustering techniques according to a survey conducted by Xu & Tian [7]. In this study, the concept of ACO clustering will be used with cosine similarity as a metric of the objective function. Then we compare the result with the ground truth.

2. ACO For Clustering News

2.1. Algorithm

The algorithm used in this paper is based on [6]. The ACOC algorithm consists of 3 parts. The first is to generate a solution by each agent using the pheromone trail information table, second is to perform a local search on select L best solutions, and third is to update the pheromone table that considers the best L solutions. In this paper, we modified some of its steps. Accurately the step to calculate the objective function of the solution. In ACOC, the metric used to calculate objective function for a solution is to use Euclidean distance. Euclidean distance for each member of a cluster is calculated to the centroid, then summed. The total summed Euclidean distance for each cluster is the objective function of the solution. In this paper, we do not use Euclidean distance but using cosine similarity (cosine distance) as the replacement method. Cosine similarity is a method for comparing the similarity between two documents. By finding the Term Frequency (TF) of each word in the document, we can calculate the similarity between 2 documents using Eq. (1) and Eq. (2)

\[
\cos(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{1}
\]

\[
\text{cosine distance} = 1 - \cos(x, y) \tag{2}
\]

Where x and y are the TF of each word in a different document. If the value of cosine is close to 1, that means the two documents have a high level of similarity. If the value of cosine is close to 0, that means the two documents have almost had no similarity.

As explained earlier the first step in the algorithm is to generate a solution. Each ant R will be grouping N data into K clusters. An ant will make a solution S of N length, in which each element in S correspondent to one data. The value assigned to each element in S represents the cluster number the data belong to. For example, Figure 1 is a solution created by an ant. The value of the first element is 2, which means that the first data of a dataset belong to cluster number 2. As is the second element value is 1, which means data number 2 belongs to cluster number 1. To find a solution, each ant will decide which cluster each data belong to using the pheromone information. Each ant choosing cluster in this algorithm is by following 2 way, but first, we define \( q_0 \), \( 0 < q_0 < 1 \) which is a threshold coefficient:

1) \textit{Exploitation}: The cluster chosen based on the highest pheromone concentration in the associated cluster. But first, we must create a random number between 0 and 1. If said number \( \leq q_0 \), then exploitation is used to determine the cluster number.

2) \textit{Exploration}: Cluster chosen using a stochastic distribution with probability \( (1 - q_0) \), denoted as, \( p_{ij} \). If the random number that we generate before is \( \geq q_0 \), then exploration is used.

| 2 | 1 | 3 | 2 | 2 | 3 | 2 | 1 |

Figure 1. Solution example.

If exploration is used, then the cluster chosen based on Eq. (3).
\[ p_{ij} = \frac{\tau_{ij}}{\sum_{k=1}^{K} \tau_{ik}}, \quad j = 1, \ldots, K \quad (3) \]

Where \( p_{ij} \) is the normalized pheromone probability of element \( i \) belong to cluster \( j \) and \( \tau_{ij} \) is the pheromone concentration associated with \( j \) cluster. As an example, for exploration we have this normalized pheromone of element \( i \): 0.36, 0.34, and 0.3 each for clusters 1, 2 and 3. Let say the random number generated is 0.7. That means the random number is in the range of cluster 2 \((0.36 < \text{random} \leq 0.7)\). The chosen cluster for associated data now has been found which is cluster number 2.

The quality of the solution generated by ants then get calculated by using the cosine distance as the metric of the objective function (2). For each cluster in the solution, we will calculate the objective function using cosine distance. As there are more than one documents in the cluster, we need to define the centroid first to be used as the comparison document for each data in the cluster. As the data is documented, it is difficult to find the centroid of a cluster. The way I approach this is to aggregate all documents in a cluster to a corpus, which will be the centroid of the cluster. Then find the cosine distance of each data in cluster to the centroid. The sum of all cosine distance in a cluster is the objective function of the cluster. The objective function of the solution is the sum of the objective function of all clusters, as stated in Eq.(4).

\[ F(S) = \sum_{j=1}^{K} \sum_{i=1}^{n} (1 - \cos(v, m)) \quad (4) \]

Where \( F(S) \) is the objective function of the solution, \( v \) is the data in a cluster, and \( m \) is the centroid of the cluster.

The second step in the algorithm is to perform a local search. Local search is a way to improve the solutions by choosing some L best solution. After choosing L best solution, then we need to define threshold probability \( p_{ls} \) between 0 and 1. Then we generate another random number between 0 and 1 for each data in the solution. If the random number is \( \leq p_{ls} \), then we change the cluster number associated with the data to another cluster randomly with equal probability. If the newly formed solution has a smaller objective function than the solution before the local search performed, then the new solution is accepted. If else, the new solution will be abandoned.

The third step of the algorithm is to update the pheromone information. The updating process of pheromone is using Eq.(5) and Eq.(6)

\[ \tau_{ij}(t+1) = (1-\rho) \cdot \tau_{ij}(t) + \sum_{l=1}^{L} \Delta \tau_{ij}^{l} \quad (5) \]

\[ \Delta \tau_{ij}^{k} = \frac{\tau_{ij}(t)}{F_k} \quad (6) \]

Where \( \rho \) is the evaporate coefficient of pheromone and \( \Delta \tau_{ij}^{k} \) is the addition of new pheromone to pheromone information. \( F_k \) is the objective function of a solution created by ant \( k \). In equation (5) we have \( \Delta \tau_{ij}^{l} \), which means that the addition of new pheromone to the pheromone information only done by select L best solutions. If there are 2 select best solutions, then only the sum of all data associated with certain cluster get the additional new pheromone. The other pheromone information will only get smaller based on the evaporation coefficient.

The algorithm repeatedly carries out these three steps for a maximum number of given iterations, and a solution having the lowest objective function value represents the optimal partitioning of data of a dataset into several clusters.

2.2. Preparing data and performance evaluation criteria

The data used in this paper are crawled from https://kumparan.com. The data crawled are from 5 different topics with each topic contains 1000 news, a total of 5000 news. After getting the dataset, we begin by preprocessing the data. First, we perform stemming on the data. Stemming in this paper will use Sastrawi [8], a Bahasa Indonesia stemmer. After stemming the data then we normalized the data by removing punctuation and perform lower cases for all words. The last step is to tokenize the news into word and perform stopword removal. Stopwords are a collection of common words that often appear and are considered to have no meaning. The list of stop words for each language is different. Stopwords used in this paper are a Bahasa Indonesia stopwords.
For the evaluation, we use 4 external clustering evaluations to see the performance. The evaluation method used is purity, normalized mutual index (NMI), rand index, and F-measure [9]. Purity is a simple and transparent test method. Purity will see to what extent a cluster is related to its class. For each cluster, it will be seen which class is dominant in the cluster, then the size of the dominant class data in each cluster will be totaled divided by the total amount of data as shown Eq. (7).

\[
purity = \frac{1}{N} \sum \max |m \cap d|
\] (7)

Where \(m\) is the set of clusters, \(d\) is the set of classes and \(N\) is the total amount of data. A bad cluster has a purity value close to 0 and a perfect cluster has a purity value of 1. High purity values can be obtained with the increasing number of clusters. Therefore, we cannot just use purity to evaluate the results of clustering. Mutual information (MI) measures how much information can be obtained about a variable from other variables. Normalized mutual information (NMI) is the normalized value of mutual information by dividing it by the entropy value to eliminate problems such as purity where the more clusters, the higher the MI value. Next is the formulation of NMI as shown Eq.(8), Eq.(9) and Eq.(10)

\[
NMI(K, C) = \frac{2 I(K; C)}{H(K) + H(C)}
\] (8)

\[
I(K; C) = \sum_{k,j} P(m_k \cap d_j) \log \frac{P(m_k \cap d_j)}{P(m_k)P(d_j)}
\] (9)

\[
H(K) = -\sum_k P(m_k) \log P(m_k)
\] (10)

Where \(I(K; C)\) is the MI of the K cluster and the C class, \(H(K)\) is the entropy value of the cluster \(K\). A value of 0 on NMI means that both variables are independent of each other, a value of 1 means that both variables are dependent on each other. The Rand index (RI) measures the similarity of the cluster to the data benchmark. RI can also be interpreted as a percentage of true decisions made by an algorithm or it can be said as accuracy. The evaluation test formula on the rand index is as shown Eq. (11).

\[
RI = \frac{TP+TN}{TP+FP+FN+TN}
\] (11)

Where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative. The problem with using RI is that RI places the same weight on FP and FN. To overcome this, we can use F-measure. The F-measure is obtained from the calculation of precision and recall values and can be formulated as shown Eq.(12) and Eq.(13)

\[
F_\beta = \left( \frac{\beta^2 + 1}{\beta^2} \right) \frac{P \cdot R}{\beta^2 P + R}
\] (12)

\[
P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN}
\] (13)

Where \(P\) is precision and \(R\) is the recall. The value of \(\beta\) affects how much the weight of the recall of the F-measure.

2.3. Result and discussion

For this research, the equipment used is a personal computer with specifications showed in the Table 1. For the algorithm, the parameter used as shown in Table 2. The heuristic coefficient is set to zero because the algorithm depends on the pheromone information to find the solution. Exploitation probability is set to 0.98 which means most of the cluster chosen based on the highest pheromone level. Local search probability value is set to 0.01 which means that the probability a data change cluster in the local search is only 1 %. Stagnant parameter is used to reduced time by stopping the iteration when the same best solution is produced for the last 20 iteration.

Table 1. Computer specification.
Table 2. Parameter used on the clustering program.

| Parameter  | Value | Description                      |
|------------|-------|----------------------------------|
| n          | 10    | Number of Artificial Ant          |
| α          | 1     | Pheromone Coefficient             |
| β          | 0     | Heuristic Coefficient             |
| ρ          | 0.01  | Evaporation Coefficient           |
| τ          | 0.1   | Pheromone Starting Value          |
| q₀         | 0.98  | Exploitation Probability          |
| p₁₀        | 0.01  | Local Search Probability          |
| iteration  | 100   | Iteration Number                  |
| stagnant   | 20    | Max Iteration number for stagnant |
| trial      | 10    | Number of trials                  |

Clustering will be done using a program created in python. Python used is version 3.7.3 by using several libraries to facilitate and simplify the program. The initial parameters used in the algorithm can be seen in Table 2.2, some initiation of variables is also carried out to simplify the process of clustering but not presented in the table or algorithm that will be discussed. The results of clustering have several different performances. To test the algorithm there are several datasets that are used with different data.

Table 3. Dataset used for evaluation purposes.

| Dataset     | Number of Data | Cluster Number | Description |
|-------------|----------------|----------------|-------------|
| Kalimat10   | 10             | 3              | Short Sentences |
| Berita10    | 10             | 3              | 10 News     |
| Berita50    | 50             | 5              | 50 News     |

For Berita10 dataset, there are 10 news data taken from 3 different categories. Of the 10 trials conducted, the best objective function value obtained was 4,087. For the evaluation results, in aside from RI value, the three evaluation methods have the highest values when compared to other datasets. Purity is at 0.7 which means that on average around 70% of cluster members are dominated by certain news categories. The NMI value is 0.442, is at the middle threshold, the data for each other is not completely independent. RI value at 0.688 means 68% similarity between the results of clustering with test data, and finally the F-measure value (β = 5) at 0.416.

Table 4. Evaluation result using 4 different method for 3 different datasets.

| Evaluation Result | Purity | NMI  | RI    | F5   |
|-------------------|--------|------|-------|------|
| Min. Value        | 0      | 0    | 0     | 0    |
| Max. Value        | 1      | 1    | 1     | 1    |
| Berita10          | 0.7    | 0.442| 0.688 | 0.416|
| Berita50          | 0.46   | 0.194| 0.72  | 0.246|
| Berita500         | 0.282  | 0.026| 0.685 | 0.209|

Whereas for the Berita50 dataset, the performance results decreased considerably except for the RI value which is the highest value of all datasets in the number of 0.72. Berita50 itself consists of 50 news taken from 5 different news categories. With more news in dataset, the calculations carried out by the algorithm became more complex. The purity value obtained by Berita50 is 0.46 which means that almost half of the cluster members are dominated by certain news categories. The NMI and F-measure scores
are half compared with the News10 results. This means that most of the news are independent of each other and the FN value generated is quite large. The News50 objective function value is 30,003.

In the Berita500 dataset, the data used is 10 times that of Berita50. With the addition of very large data, the evaluation results appear to be decreasing again. Purity is at 0.282 and the F-measure value is 0.209. Only the RI value is still at 0.681. For the NMI value is the lowest value, which is 0.026 7 times smaller than the NMI News50 value. This means that each other's news are independent and their information cannot be retrieved to find out other news information. With the increasing amount of news data that must be clustered, it will make the calculations on the algorithm more complicated. The amount of data makes the centroid cluster bigger so that the time needed to find a centroid and calculate the objective function is longer.

To compare the performance of clustering algorithms with manual categorization, the measure to be used is the objective function value using the cosine distance metric. The objective function values in the clustering results have been explained in the previous explanation. While the objective function values for manual data are obtained by entering manual data into the objective function calculation functions that have been created previously. For more information, see the table below.

**Table 5. Objective function value for clustering result and ground truth.**

| Objective Function | Prediction Data | Actual Data |
|--------------------|-----------------|-------------|
| Berita10           | 4.087           | 4.175       |
| Berita50           | 30.003          | 26.849      |
| Berita500          | 376.755         | 334.445     |

The table shows that the objective function value of the clustering is greater than the actual data, except for News10. This means that the algorithm does not work as expected, which is getting the results of the cluster with the smallest objective function value. Manual process of categorization news by the editor is much better when viewed from the semantic meaning of the news. This means that news clustering using cosine distance method has not been able to replace or resemble manual results from humans. It can be seen from the RI value that is not high enough to be said this algorithm produces good clustering.

3. Conclusion
This research was conducted to see how to do news clustering using the ACO algorithm and how its performance is compared with manual categorization. Clustering news using ACO has a slightly different application than it does clustering using other type of data. Because the data is in the form of news the method used is cosine distance which is a method of measuring 2-sentence similarity. The ACO algorithm used is the version of Shelokar et al. (2004) with a slight change in the objective function calculation method. The application used is to calculate the cosine distance between news in one cluster with the centroid cluster. But the obstacle appeared in how to determine the centroid cluster. What we do is to combine the news in one cluster into a corpus that will become the centroid of the cluster. The next obstacle that we found is the diversity of news content. Even though they have one category or topic, sometimes the news is very close to one another in terms of word similarity. This is because TF (Term Frequency) is a unique word that can differ greatly. Similarly, sometimes different categories of news can have a smaller cosine distance. For example, news at a computer exhibition can be a technology category if it is highlighted in the contents of the exhibition and can be entertainment if what is reported is the entertainment filler. But we still conduct research by conducting tests according to the problem that is to find out its performance.

The results of the evaluation of the clustering algorithm are good for small and short data, whereas for big data, the evaluation results decrease with the increasing number of data and clusters. The evaluation results also concluded that the algorithm did not approach the categorization performance manually and needed to be optimized because for large and large news data, the time needed to process the data was far more than the author's estimate.
References

[1] Lee CS, Ma L. News sharing in social media: The effect of gratifications and prior experience. *Comput Human Behav* 2012; 28: 331–339.

[2] Suleymanov U, Rustamov S. Automated News Categorization using Machine Learning methods. *IOP Conf Ser Mater Sci Eng* 2018; 459: 12006.

[3] Muflikhah L, Baharudin B. Document clustering using concept space and cosine similarity measurement. In: *ICCTD 2009 - 2009 International Conference on Computer Technology and Development*. IEEE, 2009, pp. 58–62.

[4] Luo C, Li Y, Chung SM. Text document clustering based on neighbors. *Data Knowl Eng* 2009; 68: 1271–1288.

[5] Dorigo M, Maniezzo V, Colorni A. Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man, Cybern Part B* 1996; 26: 29–41.

[6] Shelokar PS, Jayaraman VK, Kulkarni BD. An ant colony approach for clustering. *Anal Chim Acta* 2004; 509: 187–195.

[7] Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. *Ann Data Sci* 2015; 2: 165–193.

[8] Librian A, Kukuh R. Sastrawi. *GitHub*, https://github.com/sastrawi.

[9] Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.