Soft and Hard Clustering for Abstract Scientific Paper in Indonesian

1st Johannes Petrus
Teknik Informatika Universitas Sriwijaya Doctoral program student (lecturer at STMIK GI MDP)
Palembang, Indonesia
johannes@mdp.ac.id

2nd Ermatita*
Computer Science Faculty Universitas Sriwijaya
Palembang, Indonesia
ermatitaz@yahoo.com

3rd Sukemi
Computer Sciemce Faculty Universitas Sriwijaya
Palembang, Indonesia
sukemi@ilkom.unsri.ac.id

***Abstract*—For ease in grouping research papers is by doing clustering. Clustering is a method to classify the objects into subsets with similar attributes. Clustering method divided into two categories ie hard and soft clustering. Hard clustering is method to grouping the data items such that each item is only assigned to one cluster, K-Means is one of them. While Soft clustering is method to grouping the data items such that an item can exist in multiple clusters, Fuzzy C-Means (FCM) is an example. Most research papers are documented into groups that are associated with the area of expertise of the researcher, even though there is also research whose contents relate to other fields outside the area of expertise of the researcher, so it should also be documented in the group of other fields so that its contribution in other fields can be known. Here in this paper we analyse the abstract of papers written in Indonesian as data set. Data samples were taken from 3 fields, namely information technology, health and economics. Clustering process using k-means and FCM to find out whether scientific paper's abstracts from different fields of research can be in the same group / cluster as a whole, not whole or different groups. As unstructured data, abstracts must be processed through a text mining procedure first to become vector data.**

Keywords—Hard and soft clustering, k-means, fuzzy c-means (FCM), Abstract.

I. INTRODUCTION

In the world of research, abstracts are an important part that must be present in research reports. Abstracts are usually displayed at the beginning of a research report as general and brief information about the contents of the overall research report. In scientific papers, abstracts are the most important part for readers to identify the basic contents of documents quickly and accurately. By using abstracts, the reader can determine the relevance of the document to their interests, and henceforth decide to read the document or not as a whole set. Thus, each scientific paper must be accompanied by an abstract[1]. Generally abstracts are written in English but in Indonesia it is also written in Indonesian known as Bahasa. The use of these two languages is intended so that research reports can be used both by people from the country where the research report was made, and by people from countries outside the research report. Many scientific works are written by authors in certain fields but their research is related to other fields of science. The grouping of scientific papers makes it possible to find out the contribution of research to certain fields of study by researchers who do not have that background. Clustering is the organization of unlabeled data into similarity groups called clusters. A cluster is a collection of data items which are "similar" between them, and "dissimilar" to data items in other clusters. Hard clustering is about grouping the data items such that each item is only assigned to one cluster. K-Means is a famous hard clustering algorithm whereby the data items are clustered into K clusters such that each item only blogs to one cluster. Soft clustering is about grouping the data items such that an item can exist in multiple clusters. Fuzzy C-means is a famous soft clustering algorithm. It is based on the fuzzy logic and is often referred to as the FCM algorithm.

II. LITERATURE REVIEW

There is an increasing amount of textual data available every day. Textual data is unstructured, unclear and manipulation is difficult[2]. Text mining is an increasing field that attempts to gather meaningful information from natural language text. Text mining is to handle textual data. Text mining is the process of extracting the high quality of information from such large text corpus. There are various Text mining techniques are available for handling these large amount of documents such as information extraction, clustering, classification, categorization, summarization, concept linkage etc[3]. Although Text Mining is similar to Data Mining, yet they are quite different. Data Mining are design to handle structured data from database, but Text Mining can also work with unstructured or semi-structured datasets such as emails, text documents and HTML files, etc[4]. Text Mining is also process of turning text into numeric data[5], so that it can be used in an analysis or predictive modelling. Some of the application of text mining as text categorization, document clustering and information retrieval. Clustering techniques are mostly unsupervised data mining process is widely used to organize data into groups - called clusters - based on similarities among the individual data items. Objects within the cluster tend to be similar while objects belonging to different cluster are dissimilar[6]. There are 2 types of clustering namely hard and soft clustering. K-means is classified as hard clustering while Fuzzy C-Means is classified as soft clustering. Text clustering is "unsupervised" learning that do not require training data, the algorithms themselves are generally far more computation-intensive then supervised scheme. Text clustering is a special problem in data clustering, where the objects are in form of texts like paper's abstract. As in other countries, in Indonesia its own language is used for both written and oral communication, known as Bahasa Indonesia or Bahasa. In 1928 a youth congress was held. The congress comes with an important decision of a national language, i.e. Bahasa Indonesia. It is declared as the unifying language in the new nation of Indonesia and should be used instead of Dutch for formal and nation-wide communications[7], included in writing scientific papers.

III. TEXT PROCESSING

A. Text Mining

In general, text mining process starts with collecting text documents. Because collected document texts are unstructured which cannot directly use for processing, it require several preprocessing activities.

1) Text Cleanup : Text Cleanup means removing of any unnecessary or unwanted information such as normalize text converted from binary formats, deal with tables, figures and formulas.

2) Case Folding : Activity to convert the entire text in a document into a standard form i.e lowercase letters.

3) Tokenizing : Activity to breaking up a sequence of strings into pieces of words called tokens.

4) Stop words removal : Stop words are common words that considered to have no meaning. This filtering activity to reduce the number of words that need to be processed.

5) Stemming : Activity stemming is the process of reducing words to their word stems or it is the process of removing prefixes and suffixes.

B. Term document binary matrix

A document-term matrix or term-document matrix is a mathematical matrix describes the frequency of terms or words that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. Term document matrix is tracking the term frequency for each exist term by each document and represent as a numeric structure. Every word that appears in a document is stated with numbers 1 and 0 if the word does not appear in another document. To get more meaningful words, filtering can be done by selecting the frequency of occurrence of words above a certain value

	t_1	t_2	t_3	…	t_n
d_1					
d_2					
d_3					

Fig. 1. Example of Term Document Binary Matrix format.

C. Term Weighting

The most successful and widely used scheme for giving term weights is the TF-IDF weighting scheme. TF (*Term Frequency)* represents the number of times of a word appearing in a document[8]. The importance of word ti in a document can be expressed as:

$$TF(word) = \frac{Count(word)}{\sum_{i=0}^{n} Count(word_i)} \quad (1)$$

In formula (1), *Count(word)* presents the number of occurrences of the word in the document. The denominator is the sum of the number of occurrences about all the words in the documents.

IDF (*Inverse Document Frequency*) is a calculation of how the terms are widely distributed in the collection of documents concerned.

$$IDF(word) = log\left(\frac{Count(docs)}{Count(word, docs)}\right) \quad (2)$$

In formula (2), *Count(docs)* is the total number of documents. The denominator is the total number of documents that contain the word.

Formula (1)'s result is multiplied by the result of Formula (2) to obtain the result of Formula (3), which represents the weight of words.

$$Weight(word) = TF(word) * IDF(word) \quad (3)$$

IV. CLUSTERING

Clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar to each other than to other groups or clusters. Clustering can be distinguished into hard and soft clustering. Hard clustering means an object either belongs to a cluster or it does not. K-means is one of the hard clustering method. Soft clustering (fuzzy clustering) means that an object is possible to be in several clusters. Fuzzy C-Means classified as soft clustering method.

A. K-Means

After the construction of the document vector, the process of clustering is carried out. As categorized as hard clustering, K-means algorithm is one of the algorithms reviewed in this paper. The K-Means is one of the famous hard clustering algorithm[9]. K-means is the simplest and most widely used algorithm in many areas. This algorithm similarly measure is based on Euclidean distance and works only for datasets that consist of numerical attributes.

The basic algorithm of K-means is as following :
Input : k: the number of clusters
Method :
Step 1: Choose k numbers of clusters to be determined.
Step 2: Choose C_k centroids randomly as the initial centers of the clusters.
Step 3 : Repeat
  3.1: Assign each object to their closest cluster center using Euclidean distance
  3.2: Compute new cluster center by calculating mean points.
Step 4: Until
  4.1: No change in cluster center OR No object changes its clusters.

B. Fuzzy C-Means

Fuzzy C-Means (FCM) is a data clustering technique where the existence of each data point in a cluster is determined by the degree of its membership. This technique first time introduced by Jim Bezdek in 1981. The basic concept of FCM, First time is to determine the center of the cluster, which will mark average location for each cluster. In the initial condition, the cluster center is still inaccurate. Each data has a degree of membership for each cluster. By improving the cluster center and the membership value of each data repeatedly, it will be seen that the cluster center will move to the right location[10]. The main advantage of fuzzy C–means clustering is that it allow each data can belong to more than one cluster.

Fuzzy c-means algorithm is as following:

Step 1: Determine

1.1: The X matrix is n x m in size, with n = the amount of data to be clustered and m = the number of criteria (variables).

1.2: Number of clusters to be formed (c>=2)

1.3: rank (weighting w>1)

1.4: maximum iteration

Step 2: Initialize randomly the value of the matrix.

Step 3: Repeat

3.1: Calculate the centroids G_k by taking into account the cluster membership

3.2: Recalculate the cluster membership for each individual.

Step 4: Until

4.1: Convergence (the centroids don't change).

V. EXPERIMENT

A. Data Set

The data used in this paper are abstract documents of scientific research paper in Indonesian. There are 15 papers chosen.. Selected papers from 3 fields of science are information technology (IT) related to health (HE) or economics (EC) and vice versa, each of the 5 abstracts.

TABLE I. SAMPLE DATA SET

n	Field relation	d 1	d 2	d 3	d 4	d 5	d 6	d 7	d 8	d 9	d 10	d 11	d 12	d 13	d 14	d 15
1	IT-HE	√	√	√	√	√										
2	IT-EC						√	√	√	√	√					
3	HE-EC											√	√	√	√	√

Fig. 2. Sample data set

The initial group of sample data taken is as follows: documents 1 to 5 are abstract papers in the field of information technology relating to health, documents 6 to 10 are abstract papers in the field of information technology relating to economics and documents 11 to 15 are abstract papers on health related to economics. Before to the preprocessing text mining stage, the total number of terms or words was 1570 and after the preprocessing phase and filtering is done, 159 unique words are generated.

B. Experiment

The motivation is to find out whether scientific paper's abstracts from different fields of research can be in the same group / cluster as a whole, not whole or different groups.

1) *Experimental Environment :* Experiments used Python for pre-processing phase and Matlab software for the clustering phase. Scientific paper's abstract in Indonesian language used as the experimental data. The corpus contained three journal categories include information technology, health and economics. Vector space model was introduced in the experiment with TF-IDF algorithm.

2) *Experimental Flow* : there are three stages in the experiment: preprocessing, calculating term-weighting to build vector and clustering process using Matlab software for both K-Means and Fuzzy C-Means methods.

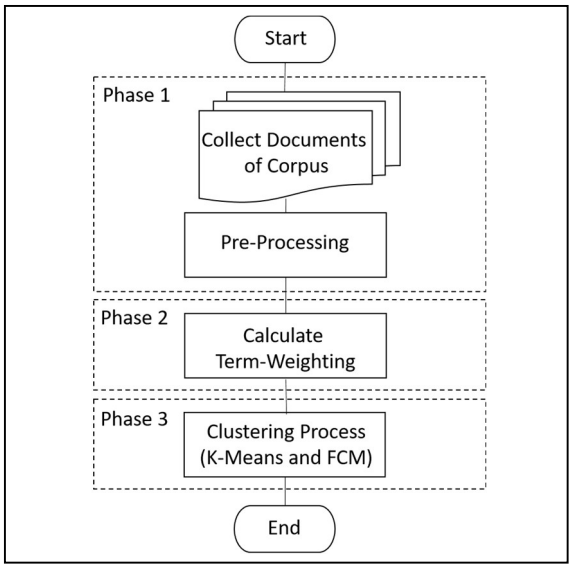

Fig. 3. Experimental Flow

3) *Experimental Basic Data* : after the 1st phase, the following basic data were created :

TABLE II. BASIC DATA PREPROCESSING

No	Term	d 1	d 2	d 3	d 4	d 5	d 6	d 7	…	d1 4	d1 5	df
1	Teliti	1	1	0	1	1	1	0	…	1	5	30
2	Hasil	4	3	0	1	0	4	0	…	1	2	21
3	sistem	7	7	2	2	1	2	0	…	0	3	40
..	…	…	…	…	…	…	…	…	…	…	…	…
159	vital	1	1	0	0	0	0	0	…	0	0	2

Fig. 4. Basic data preprocessing

4) *Term-Weighting* : by using formulas 2) and 3) the calculation results are obtained below:

TABLE III. TERM WEIGHTING

No	idf	Tf-idf					
		d1	d2	d3	…	d14	d15
1	0.0621	0.0621	0.0621	0.0000	…	0.0621	0.3107
2	0.1347	0.5388	0.4041	0.0000	…	0.1347	0.2694
3	0.1761	1.2326	1.2326	0.3522	…	0.0000	0.5283
..	…	…	…	…	…	…	…
159	0.8751	0.8751	0.8751	0.0000	…	0.0000	0.0000

Fig. 5. Term Weighting

C. Clustering Process

As stated earlier that the clustering process uses the MatLab application.

1) *Hard Clustering*

Hard clustering is the process of grouping objects where one object can only be in one group.

In this paper, the K-Means clustering proceed using Szy-Young's code that applied in Matlab by filling

in the name of the data file to be processed and set the desired number of clusters to 3. The Code as follow.

Fig. 6. K-Means code in Matlab

```
clc;
clear;
my_data = load('data-proses.txt');
data_size = size(my_data);
num = data_size(1);
data = my_data(:, 2:160);
label = my_data(:, 1);

% normalize data
epsilon = 0.01;
data_mean = mean(data);
data_mean = repmat(data_mean, [num,1]);
data_var = var(data);
data_var = repmat(data_var, [num,1]);
data = (data - data_mean)./sqrt(data_var + epsilon);

% K-Means algorithm
clusters = 3;
[cluster_label, step] = fungsi_k_means(data, clusters, num);

eval = zeros(3, clusters);
for i = 1:3
   for j = 1:clusters
      for k = 1:num
         if ((label(k)==i) && (cluster_label(k)==j))
            eval(i, j) = eval(i, j) + 1;
         end
      end
   end
end

function [label, step] = k_means(data, clusters, num)
  % initialize
  index = randperm(num, clusters);
  dis = zeros(num, clusters);
  label = zeros(num, 1);
  center = data(index, :);
  step = 0;

  while(1)
     % save the centers for each clutters of last iteration
     pre_center = center;
     % calculate distance between data points and clutter
centers
     for i = 1:num
        for j = 1:clusters
           dis(i, j) = norm(data(i,:) - center(j, :));
        end
     end
     % construct new clutters
     for i = 1:num
        label(i) = find(dis(i,:)==min(dis(i,:)));
     end
     % attain new centers
     for i = 1:clusters
        one_clutter = data(find(label==i), :);
        center(i, :) = mean(one_clutter);
     end
     % test the terminating condition
     if (center == pre_center)
        break;
     end
     step = step + 1;
  end
end
```

2) Soft Clustering

Soft clustering is the process of grouping objects where 1 object can be in only a few groups based on the degree of membership. Fuzzy C-Means (FCM) Clustering method is used for data execution. In the Matlab code, specify the name of the data file to be processed and also specified the number of clusters to 3. The code as below.

```
fcmdata = load('data-proses.txt');
[centers,U] = fcm(fcmdata,3);
maxU = max(U);
index1 = find(U(1,:) == maxU);
index2 = find(U(2,:) == maxU);
index3 = find(U(3,:) == maxU);
```

Fig. 7. Fuzzy C-Means code in Matlab

Execution in FCM until a convergence is achieved occurs 44 times iteration

D. Experiment Result

Data execution through Matlab both K-Means and FCM shows the following results

TABLE IV. CLUSTERING RESULTS

Doc	K-Means	FCM			
	Cluster	*Membership degree*			*Cluster*
		c1	*c2*	*c3*	
d1	2	0.32376	0.50499	0.17125	2
d2	3	0.30073	0.57451	0.12475	2
d3	2	0.31044	0.54915	0.14040	2
d4	2	0.36468	0.42356	0.21175	2
d5	1	0.43178	0.45283	0.11539	2
d6	2	0.42553	0.36369	0.21078	1
d7	2	0.38859	0.33509	0.27632	1
d8	1	0.42906	0.28448	0.28647	1
d9	3	0.47706	0.20294	0.32000	1
d10	2	0.38090	0.21854	0.40056	3
d11	2	0.34359	0.20354	0.45286	3
d12	1	0.30057	0.19589	0.50354	3
d13	1	0.21971	0.13005	0.65024	3
d14	1	0.14313	0.08196	0.77491	3
d15	1	0.16613	0.10004	0.73383	3

CONCLUSION

K-Means and FCM can easily create clusters from text data in the form of abstract scientific paper that has previously been through the mining process. With K-Means, the new cluster created shows that there has been a change in the cluster of experimental data from the initial grouping. The results of K-Means clustering are diverse, allowing members not to be the same as the researcher's field of science. While FCM shows that many groups of experimental data are still in the same cluster as the original category. Only a few group changes occur, but in terms of the degree of membership, it can be assumed that there has been changes in the focus of the

group. FCM cluster results tend to be in line with the researcher's field of science.

REFERENCES

[1] Mahyuddin K. M. Nasution, Abstract - A Scientific Work, University of Sumatera Utara, 2017.

Mahyuddin K. M. Nasution, Abstrak - Suatu Karya Ilmiah, Universitas Sumatera Utara, 2017.

[2] Yogapreethi.N, Maheswari.S, A Review on Text Mining in Data Mining, International Journal on Soft Computing (IJSC) Vol.7, No. 2/3, 2016

[3] Neha Garg, Dr. R. K. Gupta, Clustering Technique on Text Mining : A Review, International Journal of Engineering Research Volume No.5, Issue No.4, pp : 241-243.

[4] Lokesh Kumar, Parul Kalra Bhatia, Text Mining : Concepts, Process and Applications, Journal of Global Research in Computer Science, vol. 4 no. 3, 2013.

[5] Deepti Gupta, Er. Jitendra Dangra, a new approach for clustering of text data based on fuzzy logic, International Journal of Advanced Research in Computer Engineering & Technology (*IJARCET*), Vol 4 Issue 8, 2015.

[6] Min Ren, Peiyu Liu, Zhihao Wang, and Jing Yi, A Self-Adaptive Fuzzy c -Means Algorithm for Determining the Optimal Number of Clusters, vol. 2016.

[7] Risa R. Simanjuntak, Bahasa Indonesia: Policy, Implementation, and Planning, English Department, Faculty of Language and Culture, Bina Nusantara University,

[8] Jie Chen, Cai Chen and Yi Liang, Optimized TF-IDF Algorithm with the Adaptive Weight of Position of Word, Advances in Intelligent Systems Research, vol. 133, 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016),

[9] Dibya Jyoti Bora, Dr. Anil Kumar Gupta, A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm, International Journal of Computer Trends and Technology (IJCTT) – vol. 10 number 2 – 2014

[10] Robbie Shugara, Ernawati, Desi Andreswari, Implementasi Algoritma Fuzzy C-Means Clustering dan Simple Additive Weighting dalam pemberian bantuan program peningkatan kualitas kawasan pemukiman (Studi Kasus : Kelurahan/RT se-Kota Bengkulu), Jurnal Pseudocode, Vol. III Nomor 2, September 2016.