Clustering Sports News in Indonesian Using Modified K-Medoid Method

Yoga Dwitya Pramudita[1], Sigit Susanto Putro[2], Eka Malasari Rochman[3], Ach. Yasir Rofiqi[4], Achmad Jauhari[5], Ika Oktavia Suzanti[6], Aeri Rachmad[7]
{yoga@trunojoyo.ac.id[1]}

Informatics Engineering Program, Faculty of Engineering, University of Trunojoyo Madura[1234567]

Abstract. Sports news is a topic with high internet access rankings in Indonesia. The number of documents that must be managed continues to increase. In this research, news documents are grouped using the k-Medoid method. Medoid amount is adjusted to the number of news topics. In the k-Medoid method the cluster initialization process greatly influences the cluster results. Medoid is taken randomly according to the number of clusters desired. If the randomization results in a high degree of similarity, the resulting cluster is not optimal. Modifications are made to the initialization process so that the cluster has a low level of similarity. The test results showed accuracy reached 0.584 with five clusters. Modification of the k-Medoid method by adding the cosine similarity method can increase the average accuracy value from 0.44 to 0.54.

Keyword: sport news speak, clustering, k-medoid, modified k-medoid.

1 Introduction

The news is a fact or opinion that makes many people feel interested to know [1]. Communication can be obtained from various media such as newspapers, television, internet, and others. At this time, the media most often used to get news is the internet. News published on the internet has a wide variety of topics. One of them is sports news. Sports news is one of the most highly rated news. This is evidenced on the site alexa.com where sports news sites such as sport.detik.com and bola.net entered into 25 top sections in Indonesia [2]. So sports news becomes favorite news for the citizens of Indonesia. In presentating the news, almost all sites have the same way. The story is divided into categories according to sports. On the site http://sport.detik.com, the types used are six sports that include: Basketball, Racquet, Formula 1, MotoGP, soccer and other sports. On the site http://sport.okezone.com, the categories used are five sports that involve F1, MotoGP, Netting, Basket and other Sport. Then on the site http://sports.sindonews.com, the groups used Racket, Motors sport, Boxing and all sports. Of the three sites above, it can be concluded that on average each site provides approximately five sports. In addition to these three sites, the data is taken from sport.idntimes.com

The process of grouping the news is still done manually according to predetermined categories. Problems arise if the news document to be grouped quite a lot. The process will take longer. Therefore it is necessary to grouping with other approaches. In this research approach used to grouping news documents using text mining approach.

ICCSET 2018, October 25-26, Kudus, Indonesia

DOI 10.4108/eai.24-10-2018.2280569

Text mining is an unsupervised process learning to group the similarities of a document with other documents so that it can be separated into several groups [1]. One example of methods in text mining is clustering. Clustering can be interpreted as one technique in text mining that is used for grouping documents based on similarity of news content without defining the previous category [2].

K-medoid is a grouping algorithm that is still associated with the k-means algorithm. Both algorithms are the partition by breaking the dataset into groups, and both try to minimize the squared error from a distance between the labeled points in the cluster and the point designated as the center of the cluster. Unlike the k-means algorithm, k-medoids choose data point as the center (medoid) [3]. K-medoid is a partition technique grouping the data set from n objects into k clusters with k that are known before. In this study, the number of k will be adjusted to the number of categories on the sports news to be grouped.

K-Medoid has flaws where at initial cluster initialization greatly affects the cluster results. This method directly takes randomly as many clusters as desired by the user. To overcome these problems, in this study added a similarity prevention process at initial cluster initialization. The number of clusters that users enter is also influential in testing the accuracy of the cluster results. So this research is expected to produce optimal cluster number to classify sports news in the Indonesian language using a modified k-Medoid method.

2 Methods

In this research clustering is used to group news documents in Indonesian language. The clustering process begins with collecting news, preprocessing, and choosing the right clustering method. Data collection can be done in two ways. The first way is by directly taking news from sports sites. While the second way is by utilizing the crawler.

After the news is obtained, then preprocessing is done by converting the document into the appropriate format in order to be processed in clustering. The result of preprocessing is a collection of words or terms that have certain weights. After the news is converted into a set of words, then proceed by choosing the clustering method.

2.1 System Design

The system workflow consists of input in the form of test documents and produces an output in the form of the result of news grouping according to the resulting cluster. Figure 1 shows the process of grouping a sports news document performed by the system.

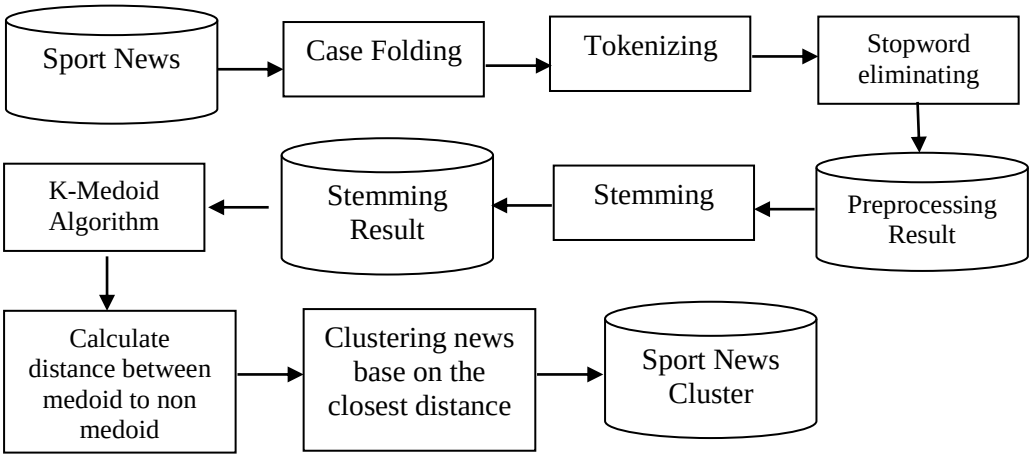

Fig. 1. System Architecture.

The documents are obtained using crawlers. The number of reports is limited to 25 sports news documents in Indonesian language. The preprocessing steps have an effect on success in extracting knowledge [4]. Pre-processing is done before clustering process. There are several sub-processes that must be done in pre-processing, among others:

1. Case folding is the stage of changing the capital letters in the news document into the same form. The form used is small capital.
2. Tokenizes is the stage of converting a material that originally shaped text into a collection of words called tokens. At this stage the deletion of symbols or punctuation marks such as (.), Comma (,), colon (:), and other symbols is also done. So that the original text-news is converted into a set of words with words produced without any logos or punctuation.
3. Stop word removal is a process to filter tokens generated by the tokenizes process so as not to stop word. This is done so that the words to be counted for resemblance are essential words. Stopword itself can be interpreted as a set of words that are considered not to be counted similarity with other words.
4. Stemming is the process of converting a token into a base form called a term. This is done so that tokens containing affixes can be returned to the primary word form so that when the process of calculating the similarity of documents, the results obtained are very optimal.
5. Weighting is the stage of calculating the frequency and weight of each term generated by the stemming phase.

2.2 ECSP (Enhanced Confix Stripping Porter) Stemming

After the document news is done preproccessing, then a collection of words from the news documents will proceed to the stemming process. The process of stemming is the conversion of words into basic words. In this study, the stemming used is ECSP for Indonesian as diagram in Figure 2 [5].

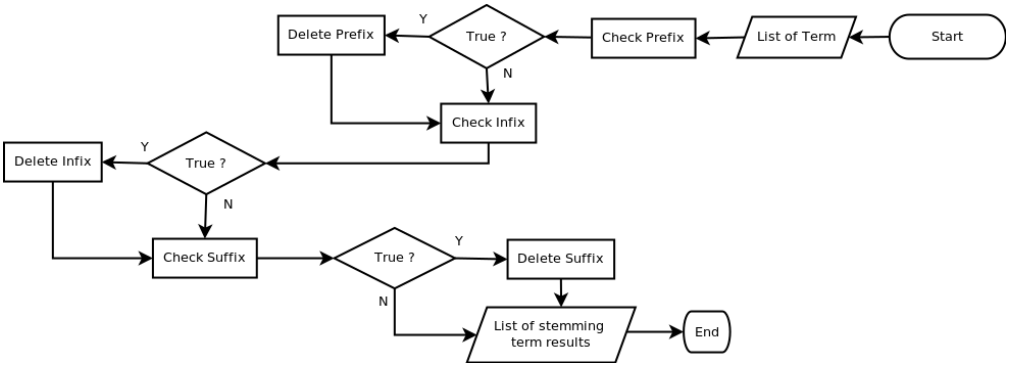

Fig. 2. ECSP Flowchart.

Each term will be checked one by one to remove the affixes and returned to the basic word form. In the Indonesian language, there are 3 types of affixes, namely prefix prefix, inserts, and suffixes. After the term is changed into a base word, it will then go into the weighting process.

2.3 Distance Space

Distance space or known as distance measurement is a step to calculate the distance of each document with other documents [6]. Documents that are similar to other documents can be found using distance space. Suppose there are five documents, namely documents A, B, C, D, and E. If the distance space between document A to document B is smaller than the distance space of document A to document C, then it can be said that the distance between document A to document B closer than the distance between document A to document C. Distance space has several methods such as Euclidean Distance, Manhattan Distance, Canberra Distance, and others. In this study, the distance calculation used is Euclidean Distance Space. The formula of Euclidean Distance Space is:

$$d(o,m) = \sum_{i=1}^{n} |o_i - m_i|^2 \quad (1)$$

Information:
d = distance between documents
o = data from the first document
m = data from the second document
i = initialization index
n = the last index of data from the document

2.4 Cosine Similarity

Cosine similarity is a method of measuring the distance or similarity between object A and object B [7]. By adding cosine similarity method to new medoid retrieval system will make the system more accurate in preventing the similarity between medoid. So the medoid

initialization is more accurate for use in sports news groupings. Here is the cosine similarity method formula:

$$similarity(A_j, B) = \frac{\sum_{i=1}^{t}(W_{ij}.W_{iB})}{\sqrt{\sum_{i=1}^{t}(W_{ij}^2)}.\sqrt{\sum_{i=1}^{t}(W_{iB}^2)}} \quad (2)$$

Information:
A = previous medoid
B = temporary medoid
t = term
w ij = TF-IDF word to i from A to j
w iq = TF-IDF word to i from B

2.5 K Medoid

In this study applying a k-Medoid method to classify sports news documents so that news can be grouped following the similarity between news documents. Steps to classify objects using k-Medoid algorithm [8]:

1. Initialization: randomly select as many k objects from n objects as medoid.
2. Calculate the distance of the document to the medoid using the Euclidean Distance formula.
3. Replace the medoid with non-medoid data
4. Calculate the total length.
5. Repeat steps 2 through 4 until there is no change in the overall distance.

In the first step, the initial determination of the medoid is done randomly. So there is a possibility that the selected medoid has a high degree of similarity. Therefore, the researchers added a method for checking the similarities between the medoid when randomly selected as shown in Figure 3. The method used to check the similarity between medoid there cosine similarity method. By adding the cosine similarity method, medoid produced is medoid-medoid which has a meager degree of similarity. So that the result of news grouping will be more optimal if compared without adding cosine similarity method.

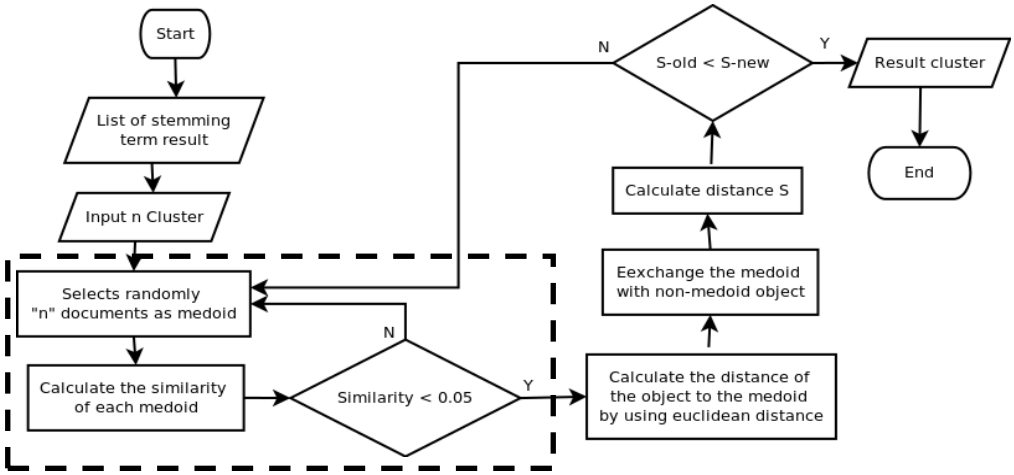

Fig. 3. Flow Chart of Modified K-medoid Method.

3 Result and Discussion

3.1 Implementation

This research will produce a comparative analysis accuracy value sports news classification system using modified k-Medoid method. The modification is to add cosine similarity method at the time of k-Medoid determination. Trials were conducted involving five users. The test is done by 4 test scenarios. Each scenario consists of some clusters 3, 4, 5, and 6. This trial involves five users to determine the matching of documents with the resulting cluster. Then calculate the average accuracy resulting from each of these trials.

The accuracy values generated by both approaches are compared with using four types of input clusters. To know the accuracy of the output then it can be calculated using the following formula of precision.

Accuracy Calculation:

$$accuracy = \frac{tp+tn}{tp+tn+fp+fn} \quad (3)$$

Where tp = number of relevant clusters being retrieved
fp = the number of irrelevant clusters being fetched
tn = the number of irrelevant clusters not captured.
fn = the number of relevant clusters not captured

3.2 Implementation

Trials were conducted to compare the value of accuracy generated in the sports news groupings using two approaches. The first approach using clustering k-Medoid method and

second approach using clustering modified k-Medoid method in which both approaches use four combinations of some clusters output.

3.3 Analysis

The first trial was performed using 3 clusters with the test results in Figure 2:

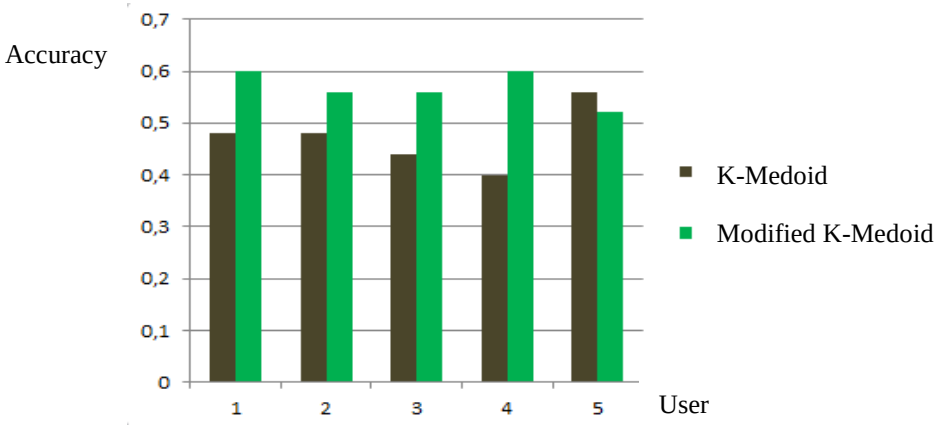

Fig. 4. Comparison of accuracy values by including 3 clusters.

The test result scenario first generates a comparison that the grouping system sports news using the modified k-Medoid by adding a cosine similarity method has a value greater accuracy. In the first approach produce average an accuracy of 0.472. While in the second approach to produce average an accuracy of 0.568.

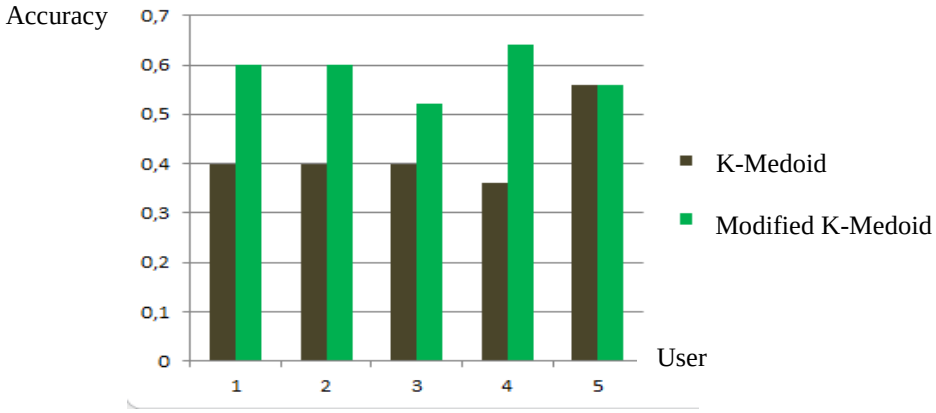

Fig. 5. Comparison of accuracy values by including 4 clusters.

In Figure 3 shows the results of the second trial scenario, the highest average accuracy value is generated by the approach using modified k-Medoid method. In the first approach yields average an accuracy of 0.392. While in the second approach to produce average an accuracy of 0.512.

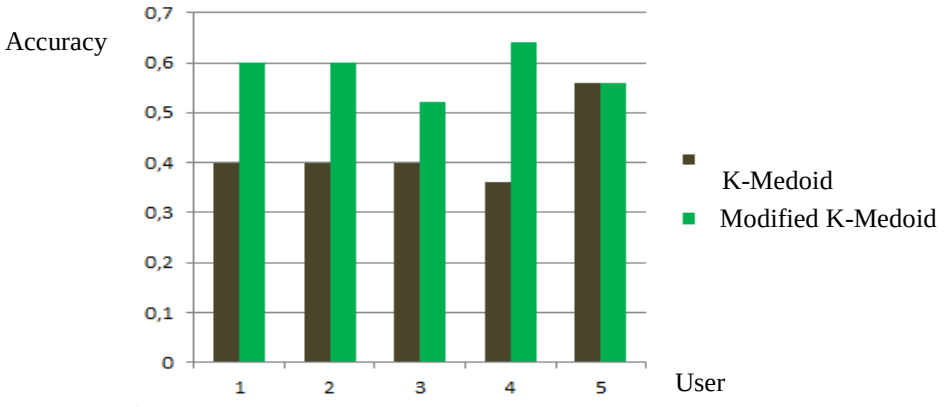

Fig. 6. Comparison of accuracy values by including 5 clusters.

Figure 4 shows the results of third trials scenarios, where the average value of the highest accuracy generated by the system using the modified k-Medoid. In the first approach yields an accuracy of 0.424. While in the second approach to produce average an accuracy of 0.584.

Figure 5 shows the results of the fourth trial scenario, where the highest average accuracy value is generated by the system using the modified k-Medoid method. In the first system yields average an accuracy value of 0.488. While on the second system yields average an accuracy value of 0.512.

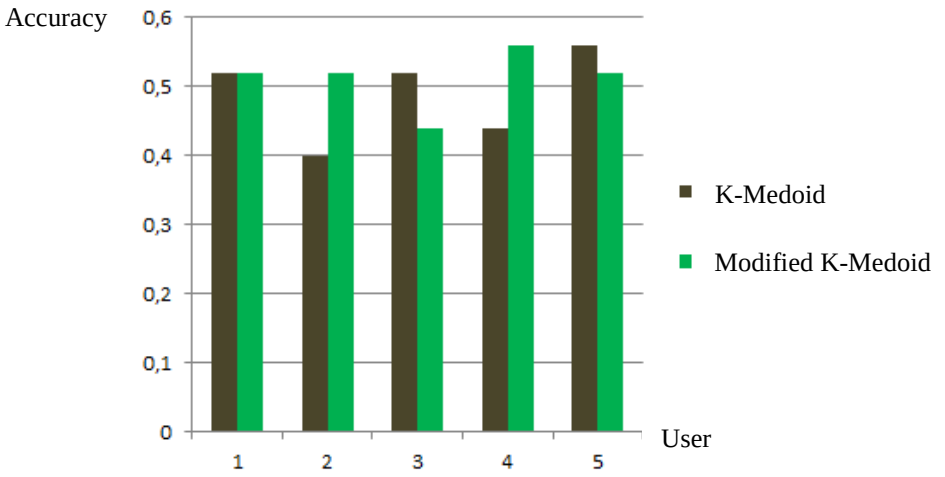

Fig. 7. compares the accuracy value by entering 6 clusters.

Figure 6 shows the highest accuracy results in the fourth trial scenario for the k-Medoid approach. The number of clusters that can be used to group sports news with optimum results is an experiment on a fourth trial scenario conducted using six clusters with an accuracy value of 0.488. When adding cosine similarity method at the time of the declaration of medoid will

prevent the presence of medoid initializations that have a high degree of similarity. The trial results prove that the addition of cosine similarity method is helpful for optimizing the cluster results. Where the accuracy of the modified k-medoid is higher than k-Medoid method. The average value of accuracy of the modified k-Medoid method obtains average a value of 0.54 with the number of clusters yielding the most optimal efficiency is the experiment conducted using five groups with an accuracy of 0.584.

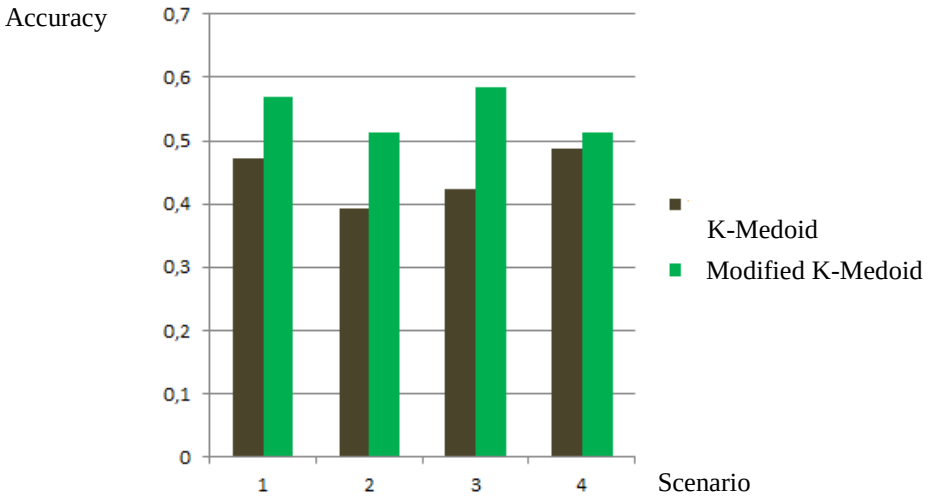

Fig. 8. Comparison of the accuracy values of the four scenarios.

4 Conclusion

From the test of two approach, the difference of average accuracy value was 0.44 when using a k-Medoid method and 0.54 when using modified k-Medoid method. So, it can be concluded that the improvement of a k-Medoid method by adding cosine similarity method at the time of the declaration of medoid very helpful to increase accuracy value.

References

[1] M. V Charnley, *Reporting*. New York: Holt, 1965.
[2] "https://www.alexa.com/topsites/countries/ID." .
[3] S. S. and M. Singh, "Comparison of a Time Efficient Modified K-mean Algorithm with K-Mean and K-Medoid Algorithm," *Int. Conf. Commun. Syst. Netw. Technol.*, pp. 435–437, 2012.
[4] Y. N. C. and N. H. A. A. I. Kadhim, "Text Document Preprocessing and Dimension Reduction Techniques for Text Document Clustering," *4th Int. Conf. Artif. Intell. with Appl. Eng. Technol.*, pp. 69–73, 2014.
[5] H. Alif, M., Solihin, F., "Perbandingan Metode Enhanced Confix Stripping Dan Porter Stemmer Untuk Stemming Konten Bahasa Indonesia," *J. Buana Inform.*, vol. 4, 2013.
[6] M. A. G. and T. C. R. S. Canuto, D. X. Sousa, "A Thorough Evaluation of Distance-Based Meta-Features for Automated Text Classification," *IEEE Trans. Knowl. Data Eng.*
[7] X. X. and C. M. X. Wang, Z. Xu, "Computing User Similarity by Combining SimRank++ and

Cosine Similarities to Improve Collaborative Filtering," *14th Web Inf. Syst. Appl. Conf.*, pp. 205–210, 2017.
[8] & T. G. P. Sengottuvelan, "Efficient Web Usage Mining Based on K-Medoids Clustering Technique," 2015.