ComTech: Computer, Mathematics and Engineering Applications, 14(2), December 2023, 119−127 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v14i2.9732 E-ISSN: 2476-907X

Fuzzy C-Means in Content-Based Document Clustering for Grouping General Websites Based on Their Main Page Contents

Sri Probo Aditiyo[1]*, Eni Sumarminingsih[2], and Rahma Fitriani[3]

[1-3]Department of Statistics, Mathematics and Natural Science Faculty, Brawijaya University
Jln. Veteran Malang, Kota Malang 65145, Indonesia
[1]dutyprobo@gmail.com; [2]eni_stat@ub.ac.id; [3]rahmafitriani@ub.ac.id

Received: 13th March 2023/ **Revised:** 24th May 2023/ **Accepted:** 25th May 2023

How to Cite: Aditiyo, S. P., Sumarminingsih, E., & Fitriani, R. (2023). Fuzzy C-Means in Content-Based Document Clustering for Grouping General Websites Based on Their Main Page Contents. *ComTech: Computer, Mathematics and Engineering Applications, 14*(2), 119−127. https://doi.org/10.21512/comtech.v14i2.9732

Abstract **-** The research aimed to use Fuzzy C-Means clustering in content-based document clustering to classify general websites based on their content. The data used were a table ranking of the most visited websites for Indonesia, taken from https://dataforseo.com/top-1000-websites/ on September 24th, 2022. The research was conducted with two different cases using Fuzzy C-Means clustering, which had two different iteration parameter values, namely 100 and 200 in maximum iteration. The research results on Fuzzy C-Means clustering in content-based document clustering are based on the two cases. These different maximum iteration parameters result in a different amount of website name data in the cluster. They are formed in the first and second clusters only. However, in the other clusters, the numbers are all the same. The results of the cluster research are validated using the silhouette coefficient, with case no. 1 and no. 2 values being 0,977783879 and 0,977788457. The use of Fuzzy C-Means clustering in content-based document clustering has an excellent performance when this method is applied to group general websites based on their content. With that result, content-based clustering can be also applied in other cases. Hence, the results can be considered to be applied to other cases for content-based clustering in the future.

Keywords: Fuzzy C-Means, content-based document clustering, general websites

I. INTRODUCTION

Website, commonly referred to as the web, is a service obtained by computer users connected to the Internet. A website is a collection of pages that are used to display information in the form of words, pictures, sounds, videos, or all forms of that information. A collection of these pages forms a series of interrelated buildings connected by a link (Andriyan et al., 2020).

Today, many people use websites in their daily lives because websites generally function to provide visitors with the information they want quickly. With the information they get from the website, they can use this information for various things, such as getting new news in various places, gaining new knowledge, being a work tool, getting entertainment, talking to each other, and getting other benefits that can be used. Hence, nowadays, the Internet cannot be separated from people not using the website (Nurrahman et al., 2021).

Even though the website has many advantages, it has problems. Websites, in general, can be made by governments, organizations, companies, or even ordinary people. Many people or groups use websites with malicious intent by providing false information because websites are easy to create. In addition, information from the website can be changed or updated in large quantities every day. Hence, a common problem that usually occurs when opening a website is that the visitors do not get the desired information and get false information. Because of this situation, the research is conducted to group existing general websites into several clusters based on the similarities of words and sentences that have been chosen before. Hence, when people search for websites based on topics, there are choices of websites that can be used from the clusters formed.

There is previous research related to content-based clustering. For example, Ting (2004) discussed

*Corresponding Author

grouping articles in digital libraries based on article content and usage. The results showed that users' short-term information needs were generally not limited to one subject category, and the content-based approach resulted in fewer or no unclusterable articles. Next, Wang (2005) conducted research using the content-based document clustering method with category preferences to support knowledge maps and effective document management. It showed that the use of the content-based document clustering method produced knowledge maps from different preference perspectives because it took into account the user's categorization preferences.

The research applies the Fuzzy C-Means clustering method as the basis of content-based clustering for websites in general because the use of data for clustering uses website content. Hence, a content representation is needed for each website to conduct research (Wang, 2005). The grouping is based on 11 website topic variables, with each variable having predetermined keywords that represent the website topic or variable. Because each website can have more than one topic, the cluster method used in content-based clustering is Fuzzy C-Means clustering. So, the Fuzzy C-Means clustering method is used because each object from the data center can be grouped into more than one cluster (Maimon & Rokach, 2010).

There are several previous studies related to Fuzzy C-Means clustering applied to websites. For example, Lalang and Lanmay (2022) discussed grouping text data from a social network, namely Twitter, using the fuzzy clustering method to identify trending topics about preventing the spread of COVID-19. They showed that the words appearing most often on the topic of preventing the spread of COVID-19 were vaccine, COVID, and vaccination. Then, Wang and Jiang (2022) used characteristic data and records of e-commerce users to obtain information that could increase sales in e-commerce. The analysis showed that it was necessary to improve the cluster analysis method in the case of e-commerce.

Compared with previous studies, the research has a big difference. The difference comes from the data used for research. In the previous study, the data used are from company documents and Twitter chat. Meanwhile, in the research, content data on the main page of a general website are used. Hence, it has a very different case from previous studies because the range of data used is very large.

The research purpose is to classify general websites using clustering based on the main page content of the website. When the general website name data have formed several clusters, it can be easier for website visitors to find information on the website based on the grouping done. Besides that, when website visitors do not get the information needed on a visited website, they can find new information from new websites based on website topic clusters originating from previously visited websites.

II. METHODS

The mindset concept in the research uses content-based clustering for websites based on their content. The clustering method used for g website content data is carried out using Fuzzy C-Means clustering. Each website can be categorized as more than one, so it requires Fuzzy C-Means clustering. The data taken are the words on the main page of the website. Before website content data are used for research, they need to be processed first into structured data using text preprocessing.

Text preprocessing is the first step that is usually done when getting raw text data. Text preprocessing is the process of cleaning raw data and turning it into structured text data because the data retrieved are still unstructured text data (Petrović & Stanković, 2019). Unstructured text has a lot of redundant, unnecessary, and useless information, such as repeated words, numbers, punctuation marks, HyperText Markup Language (HTML) tags, Uniform Resouce Locators (URLs), and other unnecessary words. If an analysis is carried out using unstructured text data, the research results are not appropriate or deviate from the original results, thereby wasting analysis time and costs (Işik & Dağ, 2020). The text preprocessing stage consists of several steps: case folding, tokenization, filtering, and stemming (Sigit et al., 2019).

The explanation of the text preprocessing process based on Sigit et al. (2019) is as follows. The first step of text preprocessing is case folding. The process in this text preprocessing is done by standardizing the characters in the data. The case folding process is the process of converting all letters to lowercase. The second step is tokenization. The process cuts the input string based on each constituent word. This process also removes numbers, punctuation, and other characters besides alphabet letters. The third step is filtering. The process in this text preprocessing is to choose important words from the results of tokenization and discard non-descriptive or unimportant words. The last step is stemming. The process changes the form of words into basic words or the stage of finding the root words of each filtered word.

Content-based document clustering is one of the grouping methods used in grouping data. It is used to group documents into several groups based on the text content in the documents, and aims to determine the keywords that represent the contents of the document. Hence, after determining the keywords, it can be used in the clustering process (Ting, 2004).

The variables used in the research are website variables based on the type of website topic. The number of variables formed based on the website topic is 11. They come from an explanation of the types of website topics based on https://www.mysch.id/blog/detail/61/jenis-jenis-website-dan-penjelasan-lengkapnya. An explanation of the website topic variables formed is shown in Table 1.

Table 1 Variable Name in Website Topic

Variable	Explanation
x_1	The number of words that match the data with the keyword for the e-commerce website
x_2	The number of words that match the data with the keyword for the business/ company website
x_3	The number of words that match the data with the keyword for the entertainment website
x_4	The number of words that match the data with the keyword for the search engine website
x_5	The number of words that match the data with the keyword for the personal website
x_6	The number of words that match the data with the keyword for the mass media website
x_7	The number of words that match the data with the keyword for the portfolio website
x_8	The number of words that match the data with the keyword for the education website
x_9	The number of words that match the data with the keyword for the social media website
x_{10}	The number of words that match the data with the keyword for the website forum
x_{11}	The number of words that match the data with the keyword for the agency/ organization's website

The results of the text preprocessing data are converted into numeric variable data shown in Table 1 to continue the clustering analysis. Before starting the clustering analysis, the data must be checked to determine whether the data are suitable for use. If it is not suitable for use, the results of the clustering model cannot be used. In examining the data, two methods are used in the research: the Kaiser Meyer Olkin (KMO) and the multicollinearity tests.

The KMO test is used to evaluate how strong the correlation is between variables. If the KMO value is bigger than 0,5, the variable can be used in the next analysis step (Xia et al., 2020). Besides that, the correlation relationship between variables is getting stronger if the results of the KMO value are close to 1 (Maiolo & Pantusa, 2021). The KMO method measures overall sampling adequacy. It measures sampling adequacy for each indicator or item in the questionnaire. So, using the KMO method, the researchers can find out whether the sample data used in the research is sufficient. If it is not sufficient, it only needs to add sample data. The KMO formula is shown in Equation (1) (Aithal et al., 2019). It has r_{ij} as a simple correlation coefficient between the *i*-th and *j*-th variables, a_{ij} as a partial correlation coefficient between the *i*-th and *j*-th variables, *i* as 1,2, … *p, j* as 1,2, … *p*, and *p* as number of objects.

$$KMO = \frac{\sum\sum_{i \neq j} r_{ij}^2}{\sum\sum_{i \neq j} r_{ij}^2 + \sum\sum_{i \neq j} a_{ij}^2} \quad (1)$$

The multicollinearity test is used to reveal the sources of multicollinearity in the correlation matrix of interpretive properties and their exclusion. Multicollinearity tests have been used to determine whether there is multicollinearity among the traits measured (Al-Ashkar et al., 2021). One way to check whether there is multicollinearity is by using Variance Inflation Factor (VIF) formula. The VIF value indicates how much variance of the estimated coefficient is inflated by multicollinearity. If any VIF values exceed 10, the associated regression coefficient is poorly estimated due to multicollinearity (Oke et al., 2019). The VIF formula is shown in Equation (2) (Teerenstra et al., 2019). It has VIF_i as the VIF value of predictor variable *i* and R_i^2 as the coefficient of determination between the *i*-th predictor variable and other predictor variables.

$$VIF_i = \frac{1}{1-R_i^2} \quad (2)$$

Fuzzy C-Means is a data clustering technique to group each object. It is determined by its membership degree. The use of data in Fuzzy C-Means means that it can belong to any class or cluster that is formed with membership varying between 0 and 1. The determination to place objects in a cluster is determined by the degree of membership. The advantage of using the Fuzzy C-Means clustering method is its ability to detect high-level clusters very well and reveal relationships between different cluster models (Rohmah & Saputro, 2020).

Fuzzy C-Means is a soft clustering technique in which each clustered object can become more than one cluster, depending on its membership level (Alam et al., 2019). Fuzzy C-Means clustering uses the degree of membership to determine which objects will be included in a cluster. It also determines whether the object can belong to more than one cluster. The criterion for Fuzzy C-Means clustering is finding the membership matrix and clustering center to minimize the objective function (Khang et al., 2020). The objection function of Fuzzy C-Means clustering is in Equation (3) (Arora et al., 2023). It consists of *n* as the total number of the pattern in data, *c* as the number of the clusters, μ_{ik} as the membership function, $d(x_k, v_i)$ as the Euclidean distance between x_k and v_i, x_k as the data point in -*k*, v_i as the center of the cluster in -*i*, and *m* as the degree of the fuzziness.

$$J_{FCm} = \sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{ik}^{m}d(x_k, v_i)^2\ 1 < m < \infty \quad (3)$$

Fuzzy C-Means clustering is the most frequently used method in fuzzy clustering cases. Fuzzy logic is used so that each object is not only associated with one cluster but also clusters with a certain degree of similar membership. The ambiguity of the solution becomes closer to 1 as *m* approaches infinity (Rajkumar et al., 2019).

The results of Fuzzy C-Means clustering calculations are needed to validate first to find out whether the model is suitable. One of the methods used is the silhouette coefficient method. The silhouette coefficient is the separation between each data point. The cluster centroid is allocated to the nearest centroid that has a place with other clusters. It is a good criterion for validating cluster quality (Yuan & Yang, 2019). This comparison is achieved with values from silhouette, which are in the range of −1 and 1. If the silhouette value has a value close to 1, the result indicates that objects and clusters have a close relationship. Hence, the closer the silhouette value is to 1, the more feasible and acceptable the cluster results will be (Jujjuri & Rao, 2019). The silhouette coefficient formula is shown in Equation (4) (Crnogorac et al., 2021). It has S_i as silhouette coefficient values, a_i as the average distance from object *i* to all objects that are in the same cluster, and b_i as the smallest value of the average distance of object *i* from other objects in different clusters.

$$S_i = \frac{(b_i - a_i)}{max(b_i, a_i)} \quad (4)$$

The parameters used in the Fuzzy C-Means clustering analysis are the number of clusters, rank, maximum iteration, the smallest expected error, initial objective function, and initial iteration. In the research, two cases are carried out with different parameters, as listed in Table 2 which shows the various parameters used in fuzzy C-means clustering with each value. The research uses two different cases in the iteration section with a number of iterations of 100 in the first study and 200 in the second study.

Table 2 Parameter for Research Cases

Parameter	Value	
	Case No. 1	Case No. 2
Number of Clusters	11	11
Rank	2	2
Maximum Iteration	100	200
The Smallest Expected Error	0,01	0,01
Number of Clusters	11	11

In the research, the website data used for each website content are primary data derived from the ranking table of Indonesia's most frequently visited websites. The data come from https://dataforseo.com/top-1000-websites/. The number of website data used are 250. Then, the content website data taken on each website is the main page of the website. The ranking table for the most visited websites in Indonesia was taken on September 24th, 2022.

The reason for determining the number of clusters formed (11 clusters) is because it follows the number of research variables used. The number of research variables is determined based on the number of types of website topics on https://www.mysch.id/blog/detail/61/jenis-jenis-website-dan-penjelasan-lengkapnya. The researchers want to group general websites based on website topics, so 11 clusters are determined. After the cluster is formed, the names of the collected websites can be seen from the contents of the cluster so that they can determine which website the cluster belongs to. The analysis steps in the research are shown in Figure 1.

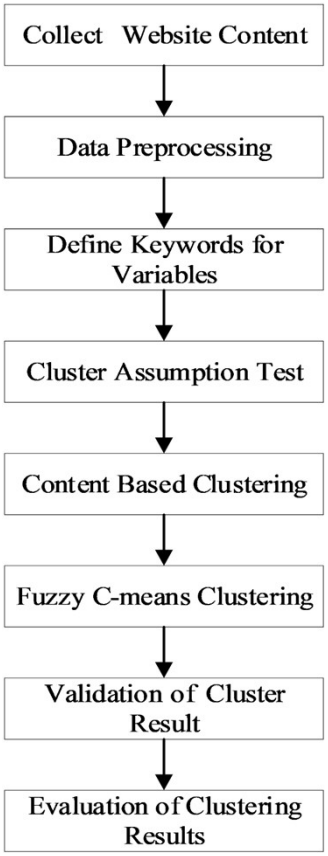

Figure 1 The Steps of Analysis in the Research

The steps of the research are described as follows. The first step is to collect website content from the most viewed websites in Indonesia from https://dataforseo.com/top-1000-websites/ as many as 250 websites. The second step is to perform data preprocessing to convert website content data into structured text data. The third step is to determine the keywords that represent the website content with the previously determined variables. The fourth step is to change the results of structured text preprocessing

data into numeric variable data using predetermined variables. The fifth step is to test the cluster analysis assumptions on the variables used to see whether they meet the requirements for performing Fuzzy C-Means cluster calculations, namely the KMO and VIF tests. If it does not meet the requirements, the sample needs to be added to meet the requirements for Fuzzy C-Means clustering. The sixth step is to start calculating Fuzzy C-Means grouping with predetermined parameter values. The seventh step is validating the cluster results to determine whether the Fuzzy C-Means clustering results are feasible by calculating the efficient silhouette coefficient. The last step is to make conclusions from the clusters that are formed.

III. RESULTS AND DISCUSSIONS

The content data containing words are taken from the ranking table of most frequently visited websites for Indonesia from https://dataforseo.com/top-1000-websites. The data are in the form of unstructured text on each website ranking. Because the research data are in the form of unstructured text data, it is necessary to change the data. Hence, it can facilitate the further analysis process. Examples of original content website data that have not been processed are listed in Table 3.

Table 3 Raw Data of YouTube

Website Name	Website Content Words
youtube.com	Home, Explore, Shorts, Subscriptions, Library, History, Sign in to like videos, comment, and subscribe., SIGN IN, EXPLORE, Music, Sports, Gaming, Live, 360° Video, Browse channels, MORE FROM YOUTUBE, YouTube Premium, YouTube Music, YouTube Kids, YouTube TV, Settings, Report history, Help, Send feedback, About, Press, Copyright, Contact usCreators, Advertise, Developers, Terms, Privacy, Policy & Safety, How YouTube works, Test new features, © 2022 Google LLC

Raw data that contains website content in Table 3 are processed using text preprocessing to convert unstructured data into structured data. The text preprocessing process is divided into four steps: case folding, tokenization, filtering, and stemming. The first step in text preprocessing is case folding. This process changes the words of the website content that have been taken. All capital letters in the website content are changed to lowercase. An example of the results of the case folding process that has been carried out using data in Table 3 is shown in Table 4.

Table 4 Result of Case Folding on YouTube

Website Name	Website Content Words
youtube.com	home, explore, shorts, subscriptions, library, history, sign in to like videos, comment,and subscribe., sign in, explore, music, sports, gaming, live, 360° video, browse channels, more from youtube, youtube premium, youtube music, youtube kids, youtube tv, settings, report history, help, send feedback, about, press, copyright, contact us, creators, advertise, developers, terms, privacy, policy & safety, how youtube works. test new features, © 2022 google llc

The second step is tokenization. This process separates every word that makes up a website. This process also removes numbers, punctuation, and other characters besides alphabet letters. An example of the results of the tokenization process that has been carried out using data in Table 4 is shown in Table 5.

Table 5 Result of Tokenization on YouTube

Website Name	Website Content Words
youtube.com	home, explore, shorts, subscriptions, library, history, sign in to like videos, comment,and subscribe., sign in, explore, music, sports, gaming, live, video, browse channels, more from youtube, youtube premium, youtube music, youtube kids, youtube tv, settings, report history, help, send feedback, about, press, copyright, contact us, creators, advertise, developers, terms, privacy, policy & safety, how youtube works. test new features, google llc

The third step is filtering. This process removes words from website content that are not descriptive or important. This process removes common words or unimportant words leaving only text with words that represent the text after the filtering process. An example of the results of the filtering process that has been carried out using data in Table 5 is shown in Table 6.

The final step in text preprocessing is stemming. This process changes the website content words into basic words or initial words. This process changes the words in the text into base words to make it easier to process in the research. An example of the results of the stemming process that has been carried out using data in Table 6 is shown in Table 7.

Table 6 Result of Filtering on YouTube

Website Name	Website Content Words
youtube.com	home, explore, shorts, subscriptions, library, history, sign in to like videos, comment and subscribe., sign in, explore, music, sports, gaming, live, video, browse channels, premium, music, kids, tv, settings, report history, help, send feedback, about, creators, advertise, developers, terms, privacy, policy & safety, how works, test new features

Table 7 Result of Stemming on YouTube

Website Name	Website Content Words
youtube.com	home, explore, shorts, subscriptions, library, history, sign in to like videos, comment and subscribe, sign in, explore, music, sports, gaming, live, video, browse channels, premium, music, kids, tv, settings, report history, help, send feedback, about, creators, advertise, developers, terms, privacy, policy & safety, how works. test new features

Based on Table 7, the results of preprocessing text data are converted into numeric data for clustering the calculation of variables. The results of the variable values are formed based on the similarities of the words in the data that have been text-preprocessed with the keywords in the variables. An example of the results of changing text preprocessing data into numeric data on YouTube is shown in Table 8. Table 8 present the result of the number of similarities between YouTube's text data that has been text preprocessed and the keywords in each variable that represent the types of websites. Hence, the value that appears for each variable is the number of matching similarities between the text data and each variable.

The results of the website content in numerical variable data are checked first to determine whether they are feasible to use in clustering calculations with an assumption test. The assumption test used is the sample adequacy test with the KMO test and the multicollinearity test with the VIF test. The results of the KMO calculation are shown in Tabel 9. The results of the KMO test are 0,515. Because the result of the KMO value on the website data is 0,515, it means 0,5 < KMO value < 1. The website data truly represents the population and can be used in cluster analysis. Hence, the sample data taken for research are sufficient for cluster analysis.

Table 8 Results of Variable Value on YouTube

Variable	Variable Explanation	Values
x_2	Keyword from e-commerce website	1
x_3	Keyword from business/company website	0
x_4	Keyword from entertainment website	4
x_5	Keyword from search engine website	1
x_6	Keyword from personal website	0
x_7	Keyword from mass media website	1
x_8	Keyword from portfolio website	1
x_9	Keyword from education website	1
x_9	Keyword from social media website	4
x_{10}	Keyword from website forum	0
x_{11}	Keyword from agency/ organization website	4

Table 9 KMO Calculation

Formula Calculation	Values Results
Kaiser Meyer Olkin (KMO) Test	0,515

For the multicollinearity test, it uses VIF calculations. It has been calculated with each existing variable with other variables. One of the VIF calculation results that has been carried out is shown in Table 10. It calculates the VIF value on the X_1 variable. The results of the VIF value for the X_1 variable in Table 10 have shown that all VIF values for the existing variables for X_1 have values below 10, so they do not have multicollinearity. In the case of calculating other VIF values for the remaining variables, they also do not have a VIF value above 10. So, all variables do not have multicollinearity. All variables can be used in cluster analysis.

Table 10 Variance Inflation Factor (VIF) Value for X_1 Variable

Variable	Variance Inflation Factor (VIF) Values
x_2	1,019
x_3	1,205
x_4	1,228
x_5	1,113
x_6	1,265
x_7	1,184
x_8	1,116
x_9	1,040
x_{10}	1,080
x_{11}	1,375

After testing the assumptions on website content data variables, the next step is content-based clustering analysis by applying Fuzzy C-Means clustering. The research uses a self-created program through a program framework. The results of the Fuzzy C-Means clustering research for two different cases are shown in Table 11. There is a difference in the number of cluster values in C_1 and C_2 compared to the results of cluster calculations in case no. 1 and case no. 2. The difference in the number of website names in the cluster between case no. 1 and case no. 2 occurs because the research cases are differentiates based on the maximum iteration value, namely 100 and 200. The results of clustering values of C_1 and C_2 in case no. 1 are 22 and 19. Meanwhile, the results of clustering values of C_1 and C_2 in case no. 2 are 23 and 18. For the cluster, the remaining values in both cases have the same cluster values when compared.

Based on the results of the cluster analysis, the clustering result still needs to be examined to determine whether the results are optimal or not. In this analysis, the method used for cluster validation is using silhouette coefficient validation to measure how close the relationship between one object is to another object in a cluster. The silhouette coefficient results are listed in Table 12. The results of the calculations are 0,977783879 and 0,977788457. The values of the silhouette coefficient in case no. 1 and case no. 2 are close to 1. It means that the cluster results from both cases have good values that can be used. The results of the silhouette coefficient calculations from the two research cases show that the names of the websites that have been placed in the cluster have strong ties. Hence, the cluster results do not need to be corrected.

Table 11 Cluster Calculation Results

Cluster	Number of Websites	
	Case No. 1	Case No. 2
C_1	22	23
C_2	19	18
C_3	17	17
C_4	28	28
C_5	53	53
C_6	26	26
C_7	29	29
C_8	20	20
C_9	2	2
C_{10}	7	7
C_{11}	27	27

Table 12 Cluster Validation Results

Cluster	Cluster Explanation	Silhouette Coefficient (SC)	
		Case No.1	Case No.2
C_1	Cluster results formed from research with the 1st cluster	0,986702386	0,986713093
C_2	Cluster results formed from research with the 2nd cluster	0,986706416	0,986716773
C_3	Cluster results formed from research with the 3rd cluster	0,984093677	0,984061559
C_4	Cluster results formed from research with the 4th cluster	0,969103239	0,969154748
C_5	Cluster results formed from research with the 5th cluster	0,966516355	0,966735675
C_6	Cluster results formed from research with the 6th cluster	0,985062593	0,985059070
C_7	Cluster results formed from research with the 7th cluster	0,978024482	0,977770562
C_8	Cluster results formed from research with the 8th cluster	0,986066911	0,985906946
C_9	Cluster results formed from research with the 9th cluster	0,999928352	0,999928365
C_{10}	Cluster results formed from research with the 10th cluster	0,991487327	0,991487575
C_{11}	Cluster results formed from research with the 11th cluster	0,972788952	0,972746168
SC Global		0,977783879	0,977788457

The cluster research results in the first and second cases have strong cluster ties. It happens because the silhouette coefficient value in both cases is close to one. The silhouette coefficient is used to check whether the objects that have been placed in a cluster have ties or not. Hence, the cluster does not need to be corrected and is directly used to obtain conclusions because the value of both cases is close to one.

The results of the general website cluster formed based on two different maximum iteration cases have a difference in the number of the first cluster and the second cluster by a difference of one. For the first cluster, the number of website names in the first and second cases is 22 and 23. For the second cluster, the number of website names in the first and second cases is 19 and 18. Meanwhile, the remaining clusters have the same number. The cluster with the lowest number of website names occurs in the ninth cluster, with as many as 2. Meanwhile, the cluster with the highest number of website names occurs in the fifth cluster, with as many as 53.

Based on the research that has been carried out, Fuzzy C-Means clustering analysis in content-based document clusters has good performance in grouping websites. Websites included in the cluster formed can be used as alternative websites if the website initially accessed cannot be used or does not provide enough information. The alternative website used is based on the initial website, including the cluster formed, so people check the names of the websites from that cluster.

IV. CONCLUSIONS

Based on the results, Fuzzy C-Means is successfully applied to content-based clustering with good cluster validation results. These results indicate that Fuzzy C-Means clustering can be applied to content-based clusters in everyday research. It has been shown that other cluster methods can be applied in content-based clustering depending on the case studied. From the calculation of content-based document clustering by applying Fuzzy C-Means in the two research cases, the cluster with the most website name data is in cluster of C_9. Meanwhile, the cluster with the least website name data is in cluster of C_5. There are only different cluster values in the two cases in clusters of C_1 and C_2 with each cluster having a difference of 1.

Next, from the results of the content-based document cluster analysis, the names of the websites that are included in the cluster can be seen. It can draw conclusions from the names of the websites that have been collected. Moreover, the keywords chosen for the variables and the number of keywords can affect the grouping with the content-based clustering method.

Nevertheless, the researchers only use the Fuzzy C-Means clustering method, which is applied in content-based clustering. Hence, the application of the content-based clustering method is only applied to one cluster method. It is suggested to use other methods for content-based clustering in one case study to see the results of the clusters formed and how good the cluster validation is based on other clustering methods applied in content-based clustering. In addition, the research can be developed further by determining the number of unique keywords to be used in symbolizing research variables so that changes in clustering results can be seen in research cases that use text data.

REFERENCES

Aithal, P. K., Dinesh, A. U., & Geetha, M. (2019). Identifying significant macroeconomic indicators for Indian stock markets. *IEEE Access*, *7*, 143829–143840. https://doi.org/10.1109/ACCESS.2019.2945603

Alam, M. S., Rahman, M. M., Hossain, M. A., Islam, M. K., Ahmed, K. M., Ahmed, K. T., ... & Miah, M. S. (2019). Automatic human brain tumor detection in MRI image using template-based K Means and improved Fuzzy C Means clustering algorithm. *Big Data and Cognitive Computing, 3*(2), 1–18. https://doi.org/10.3390/bdcc3020027

Al-Ashkar, I., Al-Suhaibani, N., Abdella, K., Sallam, M., Alotaibi, M., & Seleiman, M. F. (2021). Combining genetic and multidimensional analyses to identify interpretive traits related to water shortage tolerance as an indirect selection tool for detecting genotypes of drought tolerance in wheat breeding. *Plants, 10*(5), 1–23. https://doi.org/10.3390/plants10050931

Andriyan, W., Septiawan, S., & Aulya, A. (2020). Perancangan website sebagai media informasi dan peningkatan citra pada SMK Dewi Sartika Tangerang. *Jurnal Teknologi Terpadu, 6*(2), 79–88. https://doi.org/10.54914/jtt.v6i2.289

Arora, J., Tushir, M., & Dadhwal, S. K. (2023). A new suppression-based possibilistic Fuzzy C-Means clustering algorithm. *EAI Endorsed Transactions on Scalable Information Systems, 10*(3), 1–14. https://doi.org/10.4108/eetsis.v10i3.2057

Crnogorac, V., Grbić, M., Đukanović, M., & Matić, D. (2021). Clustering of European countries and territories based on cumulative relative number of COVID 19 patients in 2020. In *2021 20th International Symposium INFOTEH-JAHORINA (INFOTEH)* (pp. 1–6). IEEE. https://doi.org/10.1109/INFOTEH51037.2021.9400670

Işik, M., & Dağ, H. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences, 28*(3), 1405–1421. https://doi.org/10.3906/elk-1907-46

Jujjuri, R. D., & Rao, M. V. (2019). Evaluation of enhanced subspace clustering validity using silhouette coefficient internal measure. *Journal of Advanced Research in Dynamical and Control Systems, 11*(1), 321–328.

Khang, T. D., Vuong, N. D., Tran, M. K., & Fowler, M. (2020). Fuzzy C-Means clustering algorithm with multiple fuzzification coefficients. *Algorithms, 13*(7), 1–11. https://doi.org/10.3390/A13070158

Lalang, D., & Lanmay, M. (2022). Aplikasi metode Fuzzy Clustering Means untuk data trending kasus vaksin Corona pada jejaring sosial Twitter. *EduMatSains: Jurnal Pendidikan, Matematika dan Sains*, *6*(2), 431–442. https://doi.org/10.33541/edumatsains.v6i2.3447

Maiolo, M., & Pantusa, D. (2021). Multivariate analysis of water quality data for drinking water supply systems. *Water, 13*(13), 1–14. https://doi.org/10.3390/w13131766

Maimon, O., & Rokach, L. (Eds.) (2010). *Data mining and knowledge discovery handbook*. Springer. https://doi.org/10.1007/978-0-387-09823-4

Nurrahman, A., Dimas, M., Ma'sum, M. F., & Ino, M. F. (2021). Pemanfaatan website sebagai bentuk digitalisasi pelayanan publik di Kabupaten Garut. *Jurnal Teknologi dan Komunikasi Pemerintahan, 3*(1), 78–95. https://doi.org/10.33701/jtkp.v3i1.2126

Oke, J. A., Akinkunmi, W. B., & Etebefia, S. O. (2019). Use of correlation, tolerance and variance inflation factor for multicollinearity test. *GSJ*, *7*(5), 652–659.

Petrović, Đ., & Stanković, M. (2019). The influence of text preprocessing methods and tools on calculating text similarity. *Facta Universitatis: Series Mathematics and Informatics, 34*(5), 973–994. https://doi.org/10.22190/fumi1905973d

Rajkumar, K. V., Yesubabu, A., & Subrahmanyam, K. (2019). Fuzzy clustering and Fuzzy C-Means partition cluster analysis and validation studies on a subset of CiteScore dataset. *International Journal of Electrical and Computer Engineering (IJECE), 9*(4), 2760–2770. https://doi.org/10.11591/ijece.v9i4.pp2760-2770

Rohmah, D. S., & Saputro, D. R. S. (2020). Clustering data dengan algoritme Fuzzy C-Means berbasis Indeks Validitas Partition Coefficient and Exponential Separation (PCAES). In *PRISMA, Prosiding Seminar Nasional Matematika* (pp. 58–63).

Sigit, K., Dewi, A. P., Windu, G., Nurmalasari, Muhamad, T., & Kadinar, N. (2019). Comparison of classification methods on sentiment analysis of political figure electability based on public comments on online news media sites. In *IOP Conference Series: Materials Science and Engineering* (Vol. 662, No. 4). IOP Publishing. https://doi.org/10.1088/1757-899X/662/4/042003

Teerenstra, S., Taljaard, M., Haenen, A., Huis, A., Atsma, F., Rodwell, L., & Hulscher, M. (2019). Sample size calculation for stepped-wedge cluster-randomized trials with more than two levels of clustering. *Clinical Trials, 16*(3), 225–236. https://doi.org/10.1177/1740774519829053

Ting, K. D. (2004). *Clustering articles in a literature digital library based on content and usage* (Master dissertation). National Sun Yat-sen University.

Wang, L., & Jiang, Y. (2022). Collocating cecommendation method for e-commerce based on Fuzzy C-Means clustering algorithm. *Journal of Mathematics*, *2022*, 1–11. https://doi.org/10.1155/2022/7414419

Wang, S. (2005). *Preference-anchored document clustering technique for supporting effective knowledge and document management* (Master dissertation). National Sun Yat-sen University.

Xia, S., Cai, J., Chen, J., Lin, X., Chen, S., Gao, B., & Li, C. (2020). Factor and cluster analysis for TCM syndromes of real-world metabolic syndrome at different age stage. *Evidence-Based Complementary and Alternative Medicine*, *2020*, 1–10. https://doi.org/10.1155/2020/7854325

Yuan, C., & Yang, H. (2019). Research on K-Value selection method of K-Means clustering algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016