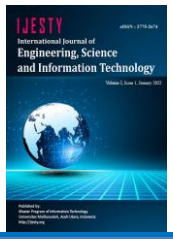
International Journal of Engineering, Science & InformationTechnology (IJESTY)
Volume 2, No. 1 (2022) pp. 70-75
eISSN: 2775-2674
Website: http://ijesty.org/index.php/ijesty
DOI: https://doi.org/10.52088/ijesty.v1i1.206
Research Paper, Short Communication, Review, Technical Paper



Online Newspaper Clustering in Aceh using the Agglomerative Hierarchical Clustering Method

Rizal Tjut Adek*, Rozzi Kesuma Dinata, Ananda Ditha

Department of Information Technology, Universitas Malikussaleh, Aceh, Indonesia
**Corresponding author E-mail:rizal@unimal.ac.id*

Manuscript received 29 October 2021; revised 10 Nov 2021; accepted 1 Jan 2022. Date of publication 10 Jan 2022

Abstract

The rapid progress in the field of information technology, especially the internet, has given birth to a lot of information. The ease of publishing an article on a website causes an explosion of news pages which will certainly confuse readers. The diversity and the increasing number of news articles make it increasingly difficult for internet users to find news and large piles of news data on online newspaper sites in Aceh. The grouping of text documents is needed to classify news in online newspapers in Aceh based on the content contained in news articles. In this study, the process of grouping online news in Aceh was tried using the Agglomerative Hierarchical Clustering method. News is grouped with a Bottom-Up design strategy that starts with placing each object as a cluster then combined into a larger cluster based on the similarity of keywords in each news, then the cluster results are compared and put into each news category. The research design was carried out in a structured manner using data flow diagrams in forming the research framework. The study was conducted by taking online news text data on 10 online news websites in Aceh from July 2016 to March 2017 with 1000 randomly generated documents. The process of crawling news data is done using a php script which will only take text files from the news on the website. News grouping is done based on religion, politics, law, sports, tourism, education, culture, economy and technology. The results of the grouping performance of the Agglomerative Hierarchical Clustering method in this study have an average accuracy of 89.84%.

Keywords: *Document Clustering, Online News, Agglomerative Hierarchical Clustering*

1. Introduction

The rapid progress in the field of information technology, especially the internet, has led to the development of great information [1]. For example, information in the form of news articles, many news article makers present their information online to the public so that the tendency for people to access information, especially news in online newspapers in Aceh is higher.
The diversity and increasing number of news articles published can make it more difficult for internet users to find the desired article [2] and the accumulation of news data in online newspapers in Aceh is very large [3] [4]. The accumulated text-form data will lose its useful value if it is not immediately handled, therefore it is necessary to group news in online newspapers in Aceh based on the content contained in news articles.
Clustering is a grouping of records or observations that form a class of similar objects [5]. Good clustering should classify or classify objects that have similarities in one group and separate objects that are not similar. Basically, there are two clustering methods [6], but in this final project a hierarchical clustering method will be used. One type of this method is agglomerative clustering, a method that is bottom-up clustering, which is the merging of several objects into a single cluster.
The agglomerative hierarchical method starts with individual objects [7] [8]. Initially the number of clusters is the same as the number of objects. First of all the objects that are most similar are grouped [9] [10], and these initial groups are combined according to their likeness. Finally, when the similarity decreases, all the subgroups are combined into one cluster.
In this study, the researcher tried to create a website grouping news for online newspapers in Aceh which will accommodate data in the form of news articles and will display several news articles based on various categories automatically [11].

2. Literature Review

Text mining has been defined by many research experts and practitioners. Text mining has the definition of mining data in the form of text [12] where the data source is usually obtained from documents[6], and the goal is to find words that can represent the contents of the document so that analysis of the relationship between documents can be carried out.
The text mining system consists of a text preprocessing component, feature selection [13], and a data mining component. The text preprocessing component functions to convert unstructured textual data [14] such as documents into structured data and stored in the database [15]. Feature selection will choose the right words and affect the classification process. The last component will run data mining techniques on the output of the previous component.
A minor element of text mining is the focus on document collections [16]. In simple terms, document collections can be in the form of grouping text-based documents [17]. In short, text mining aims to find patterns throughout a very large document collection [18]. A document collection can be either static, in that the initial complement to the document remains unchanged, or dynamic, which is the term used to document a collection marked by the entry of new documents or being updated from time to time. Very large document collections, as well as document collections at very high levels of document change, can pose a challenge to performance optimization for various components of a text mining system.
Algorithms used in text mining usually do not only do calculations on documents, but also on features (features). Four kinds of features are often used:

1. Character, which is an individual component, can be letters, numbers, special characters and spaces, is a building block at the highest level forming semantic features, such as words, terms and concepts.
2. Word (Words).
3. Term (Terms), is a single word (one word) and a multi word phrase (many words) which are selected directly from the document.
4. Concept, a feature that is generated from a document manually, rule-based, or other methodology.

Text that will be carried out by the text mining process generally has several characteristics including high dimensions, noise in the data, and bad text structure [1]. The method used in studying text data is by first determining the features that represent each word for each feature in the document.
Stemming is an integral part of Information Retrieval (IR). There are not many algorithms specifically for Indonesian stemming with various limitations in it. Porter's algorithm is one of them, this algorithm takes a shorter time than stemming using the Nazief & Adriani Algorithm [19], but the stemming process using Porter's Algorithm has a smaller percentage of accuracy (precision) compared to stemming using the Nazief & Adriani Algorithm. The Nazief & Adriani algorithm as a stemming algorithm for Indonesian text which has the ability to have a percentage of accuracy (precision) better than other algorithms. This algorithm is indispensable and decisive in the IR process in Indonesian documents.
The process of stemming in Indonesian text is more complicated / complex because there are variations of affixes that must be removed to get the root word (root word) of a word [20]. In general, basic Indonesian words consist of a combination:

$$\text{Prefix 1} + \text{Prefix 2} + \text{Root word} + \text{Suffix 3} + \text{Suffix 2} + \text{Suffix 1} \quad (1)$$

Clustering Hierarchy builds a cluster hierarchy or in other words a cluster tree, which is also known as a dendrogram. Each cluster node contains a child cluster; sister clusters that divide the poit covered by their parent. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down). Agglomerative clustering begins with a single point (singleton) cluster and repeatedly combines two or more of the most appropriate clusters. The divisive cluster begins with one cluster of all data points and iteratively divides the cluster that is most appropriate. The process continues until the termination criteria (often, the required number of k from the cluster) are met.
The Agglomerative Hierarchical Clustering method is a method that uses a Bottom-Up design strategy which starts by putting each object as a separate cluster (atomic cluster) and then combining atomic clusters - atomic clusters into clusters that are bigger and bigger until finally all the objects converge. in a cluster or process can also stop if it has reached certain conditions[21].
According to [22], the steps in the Agglomerative Hierarchical Clustering algorithm to group N objects (items / variables) are:

1. Starting with N clusters, each cluster contains a single entity and a symmetric matrix of the distance (similarities) D = {dik} with the matrix type is NxN.
2. Find the distance matrix for the closest (most similar) cluster pairs, namely by looking for the greatest similarity. Suppose that the distance between the U and V clusters that are the most similar is the duv.
3. Merge clusters U and V. Label the new clusters with (UV). Update the entries in the distance matrix to represent the closeness between the new group and the remaining groups by:
   a. Delete rows and columns corresponding to clusters U and V
   b. Add rows and columns providing distances between the clusters (UV) and the remaining clusters.
4. Repeat steps 2 and 3 (N-1) times. The process will continue until finally a cluster consisting of all objects is formed.
5. Done.

There are 3 (three) hierarchical cluster methods, namely the single linkage method [23], the complete linkage method, the average linkage method. Single linkage [24] gives the result when groups are joined according to the distance between the members who are closest, complete linkage occurs when groups are joined according to the distance between the members who are furthest. For average linkage, it is combined according to the average distance between pairs of members of each member on the set [25].

In this study, the single linkage method was used [26] to form document clusters. The input for the single linkage algorithm is the distance or similarity between pairs of objects. Groups are formed from a single entity by combining the shortest distance or the greatest similarity (similarity). Initially, we have to find the shortest distance in D = {dik} and combine the corresponding objects, for example, U and V, to get the cluster (UV). For the next step of the algorithm, the distances between (UV) and other W clusters are calculated using the formula:

$$d(uv)w = \min\{ duw , dvw \} \quad (2)$$

Information:

duw, dvw = smallest distance between groups (u, v) and w

Here the quantities duw and dvw are the shortest distance between clusters U and W as well as clusters V and W.

3. Methods

This study uses data taken from online newspaper news text documents on 10 online news websites in Aceh from July 2016 to March 2017 at random. The data is in the form of a text file format of 1000 news document files. To validate the website system created, the data collection is grouped into 8 categories, namely religion, culture, economy, law, sports, education, politics, technology, and tourism.
This analysis method stage is the stage of analyzing the system to be built. After the analysis is obtained, the next step is to make an analysis result. The results of the analysis will become a reference for the design of the system being built. The requirements needed to build the desired system are divided into 2 parts, namely input requirements and output requirements, namely as follows:

1. Input Requirements
   In a system that is built, it has a need for data to be entered, including the admin managing the system such as adding and deleting data, then users can enter keywords or select categories.
2. Output Requirements
   The output produced by this system is that it can display news articles that are similar to the entered keywords and can display news articles that have been clustered based on their respective categories.

System testing was carried out using text documents from one of the online newspaper news websites in Aceh with a total of 10 documents for each category. After being able to carry out the preprocessing process, the frequency of each term (word) is calculated and weighted using the TF-IDF method and then the terms are stored in the database and the clustering process with the agglomerative hierarchical method can be done to find out the category results of the test news. After the clustering process is complete, the user or user performs a keyword input test (query) or selects news categories to get the desired news or information.

4. Results And Discussion

The context diagram that shows a system process on the online newspaper news clustering website in Aceh is as follows:

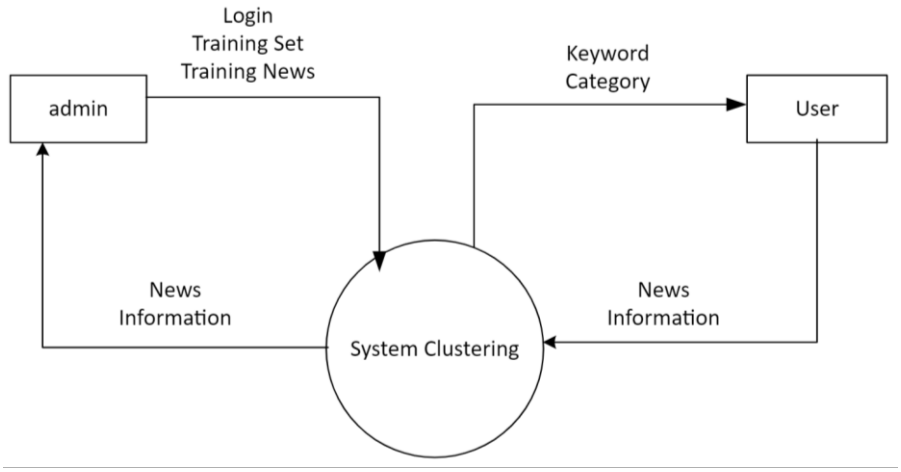


Fig 1. Context Diagram

In the picture above, the admin must log into the system first, then the admin can add new news to be tested by entering the title, content of the news, and the url of the news to be tested. After the news data is entered, the system will process the news data by doing preprocessing. The results of preprocessing will be clustered by the system, then the results of news clusters in the form of categories of test news are stored in the database. Admin can also manage news data in the database by deleting news data contained in the database. After that the user can enter a query (keyword) or select a category on the main page of the website to search for the news desired by the user.
The process carried out by the method used in this study is by weighting, then the similarity of the query is seen by looking for similarities in the keywords in the test news and training news. Then calculated the number of clusters and the distance matrix to get the value between matrices. After that, the distance matrix for the closest (minimum) cluster pair is searched and combined. Matrix rows that have been merged will be deleted and form a new matrix row. Then the system will repeat the process until the distance matrix remains one. The system will take the distance value from the test news clustering process and perform calculations with the category distance value, so that the largest distance value is obtained as a result of the clustering process. The results of the cluster will be updated into the database and the admin can also see whether the results of the clustering of news groups are appropriate.
The Agglomerative Hierarchical Clustering algorithm in this online newspaper news grouping system is used on the added news page during the news testing cluster process to be able to group each new test news processed into its respective categories. This Agglomerative Hierarchical Clustering process is to create a similarity matrix that contains the level of similarity between the grouped data. The algorithm of the Agglomerative Hierarchical Clustering process is as follows:

Given:

A set X of objects $\{x_1, \ldots, x_n\}$

A distance function dist(c_1, c_2)

for i = 1 to n

$c_1 = \{x_i\}$

end for
C = $\{c_1, \ldots, c_n\}$
L = n+1
while C.size > 1 do
(c_{min1},c_{min2}) = minimum dist(c_i,c_j) for all c_i,c_j in C
Remove c_{min1} and c_{min2} from C
Add {c_{min1},c_{min2}} to C
l = l + 1
end while

The degree of similarity can be calculated in various ways such as by frequency comparison. Starting from the similarity of this matrix, it is possible to use which type of linkage will be used to group the analyzed data.
The data used to test this system is by using a text document from an online newspaper news website in Aceh. Between a category and another category has a total of 10 documents with a specification of the number of documents for each category in the test data can be seen in table 1 as below:

Table 1. Data testing

Category	Number of Document
Religion	10
Culture	10
Economic	10
Law	10
Sport	10
Education	10
Political	10
Technology	10
Traveling	10
Total	90

In addition to the data used for system testing, there is data that is used as training data and has the same character as the test data, only in the database creation, the data has been labeled a category according to the category provided by the news site. The specification of the amount of training data can be seen in the table 2.

Table 2. Data set for data training

Category	Number of Document
Religion	100
Culture	110
Economic	110
Law	130
Sport	110
Education	110
Political	110
Technology	110
Traveling	110
Total	1000

From the results of tests carried out on test data based on training data taken from online newspaper news text documents on 10 online news websites in Aceh with a total of 1000 documents, the system accuracy values obtained in the table 3.

Table 3. Accuracy Test Results

Category	*Accuracy*
Religion	90,47%
Culture	89,02%
Economic	90,24%
Law	90,24%
Sport	89,28%
Education	89,77%
Political	88,76%
Technology	90,47%
Traveling	90,36%
Average	89,84%

5. Conclusion

This online newspaper reporting grouping system in Aceh was built using the PHP programming language, MySQL as a database, and system design using DFD (Data Flow Diagram) modeling. Grouping online newspaper news in Aceh by conducting a preprocessing process consisting of case folding, tokenization, filtering, and stemming processes. Then the grouping is done by applying the Agglomerative Hierarchical Clustering method. The system can display news documents that have close similarity to the entered keywords (queries) so that users (users) can easily find information from different articles based on the searched keywords or categories. The results of system testing carried out on test data based on data taken from online news text documents on 10 online news websites in Aceh with a total of 1000 documents, showed the average accuracy of program work was 89.84%.

References

[1] R. T. Adek, M. Fikry, and A. Helmina, "OPINION MINING ABOUT PARFUM ON E-COMMERCE BUKALAPAK.COM USING THE NAÏVE BAYES ALGORITHM," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 6, no. 1, pp. 107–114, 2020, doi: 10.33480/jitk.v6i1.1448.

[2] M. D. Devika, C. Sunitha, and A. Ganesh, "Sentiment Analysis: A Comparative Study on Different Approaches," *Procedia Comput. Sci.*, vol. 87, pp. 44–49, 2016, doi: 10.1016/j.procs.2016.05.124.

[3] K. Kim, O. joung Park, S. Yun, and H. Yun, "What makes tourists feel negatively about tourism destinations? Application of hybrid text mining methodology to smart destination management," *Technol. Forecast. Soc. Change*, vol. 123, pp. 362–369, 2017, doi: 10.1016/j.techfore.2017.01.001.

[4] D. Riyan Rizaldi, A. Doyan, Z. Fatimah, M. Zaenudin, and M. Zaini, "Strategies to Improve Teacher Ability in Using The Madrasah E-Learning Application During the COVID-19 Pandemic," *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 2, 2021, doi: 10.52088/ijesty.v1i2.47.

[5] L. Oliveira and Á. Figueira, "Benchmarking Analysis of Social Media Strategies in the Higher Education Sector," *Procedia Comput. Sci.*, vol. 64, pp. 779–786, 2015, doi: 10.1016/j.procs.2015.08.628.

[6] R. T. Adek and M. Ula, "A Survey on The Accuracy of Machine Learning Techniques for Intrusion and Anomaly Detection on Public Data Sets," in *2020 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA)*, 2020, pp. 19–27, doi: 10.1109/DATABIA50434.2020.9190436.

[7] C. Shen and C.-J. Kuo, "Learning in massive open online courses: Evidence from social media mining," *Comput. Human Behav.*, vol. 51, pp. 568–577, 2015, doi: 10.1016/j.chb.2015.02.066.

[8] S. Ali Rafsanjani, F. E. Rooslan Santosa, and R. Durrotun Nasihien, "Analysis of Planning for Clean Water Needs at Grand Sagara West Surabaya Hotel With the Green Buillding Concept," *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 2, 2021, doi: 10.52088/ijesty.v1i2.55.

[9] T. Hachaj and M. R. Ogiela, "Clustering of trending topics in microblogging posts: A graph-based approach," *Futur. Gener. Comput. Syst.*, vol. 67, pp. 297–304, 2017, doi: 10.1016/j.future.2016.04.009.

[10] J. S Pasaribu, "Development of a Web Based Inventory Information System," *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 2, 2021, doi: 10.52088/ijesty.v1i2.51.

[11] R. Rinaldy and M. Ikhsan, "Determinant Analysis Of Conflict On Project Results In Aceh Province," *Int. J. Eng. Sci. Inf. Technol.*, vol. 1, no. 1, 2021, doi: 10.52088/ijesty.v1i1.37.

[12] G. Vinodhini and R. M. Chandrasekaran, "A comparative performance evaluation of neural network based approach for sentiment classification of online reviews," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 28, no. 1, pp. 2–12, 2016, doi: 10.1016/j.jksuci.2014.03.024.

[13] E. V. Kotelnikov and M. V. Pletneva, "Text sentiment classification based on a genetic algorithm and word and document co-clustering," *J. Comput. Syst. Sci. Int.*, vol. 55, no. 1, pp. 106–114, 2016, doi: 10.1134/S1064230715060106.

[14] D. Tang, B. Qin, F. Wei, L. Dong, T. Liu, and M. Zhou, "A Joint Segmentation and Classification Framework for Sentence Level Sentiment Classification," *Audio, Speech, Lang. Process. IEEE/ACM Trans.*, vol. 23, no. 11, pp. 1750–1761, 2015, doi: 10.1109/TASLP.2015.2449071.

[15] R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt, "Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events," *Comput. Human Behav.*, vol. 56, pp. 179–191, 2016, doi: 10.1016/j.chb.2015.11.040.

[16] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic Sarcasm Detection: A Survey," *ACM Comput. Surv.*, vol. 50, no. 5, 2016, doi: 10.1145/3124420.

[17] A. Babour and J. I. Khan, "Tweet sentiment analytics with context sensitive tone-word lexicon," *Proc. - 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT 2014*, vol. 1, pp. 26–34, 2014, doi: 10.1109/WI-IAT.2014.61.

[18] M. Hu and B. Liu, "Mining and summarizing customer reviews," *Proc. 2004 ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '04*, p. 168, 2004, doi: 10.1145/1014052.1014073.

[19] R. T. Adek, Bustami, and M. Ula, "Systematics Review on the Application of Social Media Analytics for Detecting Radical and Extremist Group," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1071, no. 1, p. 012029, Feb. 2021, doi: 10.1088/1757-899X/1071/1/012029.

[20] K. Liu, L. Xu, and J. Zhao, "Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking," *Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. (Volume 1 Long Pap.*, pp. 314–324, 2014, doi: 10.1109/TKDE.2014.2339850.

[21] A.-M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 339–346, doi: 10.3115/1220575.1220618.

[22] N. Nurdin, R. T. Adek, and R. Rizwan, "PENDETEKSIAN DOKUMEN PLAGIARISME DENGAN MENGGUNAKAN METODE WEIGHT TREE," *J. Telemat.*, vol. 1, no. 1, pp. 31–45, 2019, doi: http://dx.doi.org/10.35671/telematika.v12i1.775.

[23] X. Lv and N. El-Gohary, "Text Analytics for Supporting Stakeholder Opinion Mining for Large-scale Highway Projects," *Procedia Eng.*, vol. 145, pp. 518–524, 2016, doi: 10.1016/j.proeng.2016.04.039.

[24] C. S. Rao and S. Viswanadha Raju, "Concurrent Information Retrieval System (IRS) for Large Volume of Data with Multiple Pattern Multiple ($$2^\\mathrm{N}$$ 2 N) Shaft Parallel String Matching," *Ann. Data Sci.*, vol. 3, no. 2, pp. 175–203, 2016, doi:

10.1007/s40745-016-0080-1.
[25] F. L. Cruz, J. A. Troyano, B. Pontes, and F. J. Ortega, “Building layered, multilingual sentiment lexicons at synset and lemma levels,” *Expert Syst. Appl.*, vol. 41, no. 13, pp. 5984–5994, 2014, doi: 10.1016/j.eswa.2014.04.005.
[26] W. Hochwarter, “On the merits of student-recruited sampling: Opinions a decade in the making.,” *J. Occup. Organ. Psychol.*, vol. 87, no. 1, pp. 27–33, Mar. 2014.