Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Incremental News Mining Using Evolving Clustering with Functional Operators

Amalia Wirdatul Hidayah[1], Aliridho Barakbah[2], Iwan Syarif[3]
[1]wirdatulamalia@gmail.com, [2]ridho@pens.ac.id, [3]iwanarif@pens.ac.id
Politeknik Elektronika Negeri Surabaya

Article Information

Submitted : 29 Apr 2023
Reviewed: 30 Apr 2023
Accepted : 3 May 2023

Keywords

News, Online Media, Clustering, Evolving System, News Mining

Abstract

Online media publish journalistic products, one of which is news online (online news). This is in line with the findings of the Ministry of Communication and Informatics (Kemkominfo), that in 2018 there were 43,000 online media in Indonesia. On generally in getting actual news, humans tend to read the news on online media one by one. The activity is not effective because of the news that produced by online media have the same information with each other news. In this study, we propose an innovative solution to this issue by developing a news mining system that employs clustering based on an evolving system. This system has the potential to improve the effectiveness of news retrieval by grouping similar news together and identifying key information trends, ultimately enhancing the ability of individuals to obtain actual news. Based on research observations, the performance of growing news clustering with functional operators is quite good, as evidenced by an accuracy of 83%.

A. Introduction

Media is a tool or means used to convey messages from communicators to audiences, while the notion of mass media itself is a tool used in conveying messages from sources to audiences using communication tools such as newspapers, films, radio, and television [1]. Lately, new media have begun to emerge as the influence of the development of knowledge and technology. According to M.Romli [2] in the era of the third generation of mass media, various online platforms provided the latest news in a matter of minutes. Various media companies are competing to create online media to meet the human need for information. Online news makes the articles available for readers to access anytime, unlimited time. Online news writers can link recent news with old news, so there is no need to repeat old-related news. News writers only need to design a variety of new information into links, which contain the entire background of the report [3]. Online news also provides a space for the public to respond, interact, or even customize certain stories. Along with the convenience of today's technology, it has an impact on the proliferation of online media in Indonesia [4]. Meanwhile, each online media can publish as many as 1 to 2 stories every hour. This shows how massive the data produced by online media in Indonesia is. The amount of news circulating exceeds the information processing capacity of humans so it can cause confusion and psychological stress for humans [5]. It exceeds the information processing capacity of the human brain. This has an impact on human mental health, resulting in confusion and psychological distress [16]. So that news readers can not read the news effectively. In general, in getting actual news, humans tend to read news on online media one by one. This activity was not effective because Subašić's research [6] revealed that news produced by online media has the same information as one another. As we know that in today's modern era, the role of information technology in everyday life is of course very influential. This is inseparable from our activities, which are often supported by information technology itself, which can answer the demands of work that is faster, easier, cheaper, and saves time. Human limitations in processing information cause humans to be unable to process all information in the form of news. This causes only a little news that humans can read. Reading behavior that is not comprehensive can result in misinterpretation of events due to differences in perspectives [7].

Many studies implement the clustering method on news. Some of these studies have their respective objectives in clustering. They also have their approach to achieving high accuracy in grouping news. The increasing volume of Indonesian-language electronic news is a valuable source of information. Clustering text documents is one of the operations in text mining to group documents that have the same content. Clustering can be applied to find links between news stories. The experiment in this paper uses 4718 news from the www.kompas.com site taken from June 2005 to November 2005. The results of the experiment show that clustering can reveal links between news that were not seen before. Clustering in this experiment uses the K-means algorithm [8]. Based on research conducted by Joel [9] when an event occurs in the real world, many news reports describing this event start appearing on various news websites within minutes of the event occurring. This can generate large amounts of information for users, and it may need automated processes to help manage this information. In this paper, we describe a clustering system that can group news reports from different sources into event-centric groups — that is, groups of news reports that describe the same event. They applied the clustering method to news reports by representing them as Bag-of-Words with TF.IDF using a variation of the k-means algorithm that works in a single pass without cluster reorganization. When talking about clustering, feature extraction will not escape. This is the main foundation of the machine learning process. Research conducted by Piskorski [10] extracted the features of location names, people's names, and others as news features. This is the same as what was done by Florence [11] Florence tried to extract features from the news text in the form of person's name, organization name and geographical location. Features in the form of entities are unique in nature and discussion of events between one news and another can be easily distinguished by knowing the features of the entity. However, the resulting news groups are contextually indistinguishable. This will occur when discussing an event that is protracted and takes place over a long period of time. Meanwhile, research conducted by Laban [12] used the TF-IDF approach as a feature. News is processed to form bag-of-words, then processed into TF-IDF. Feature extraction using the TF-IDF approach is widely used and has proven to produce good results. However, this approach is not appropriate when used in the case of streaming data, where data comes continuously. When using TF-IDF, the system is charged with updating the TF-IDF value of all the words in the news every time a new news comes in. Therefore, we propose a new approach to create a news mining system with an evolving system based clustering. Evolving system-based clustering focuses on grouping stories into small groups to help organize news and identify patterns of news group development. Clustering based on the Evolving system is used to group news so that it can be used to organize news according to the right category and analyze the pattern of development of newsgroups from time to time. As well as helping users read the news in a more structured manner. This research attempts to uncover patterns of formation of an issue of discussion, which in the future can be utilized by interested parties in analyzing issues/discussions in the media.

Another study that discusses clustering, namely research by Sigita [13], uses a news clustering approach that develops over time. This online grouping uses the principle vector quantization (VQ) algorithm dynamically and obtains an accuracy of 70.9%. Research conducted by Bakr [14] used an incremental density-based algorithm. The experimental results show that this algorithm is significant increased program execution time at runtime. it affects performance of the clustering algorithm. Part of Incremental Density based algorithm, there are several algorithm related to news grouping.

B. Research Method

We divide this research into 4 main sections, namely data input & data preprocessing, data preprocessing & Keyword Extraction, evolving clustering system, and output. We will explain each part's stages as follows.

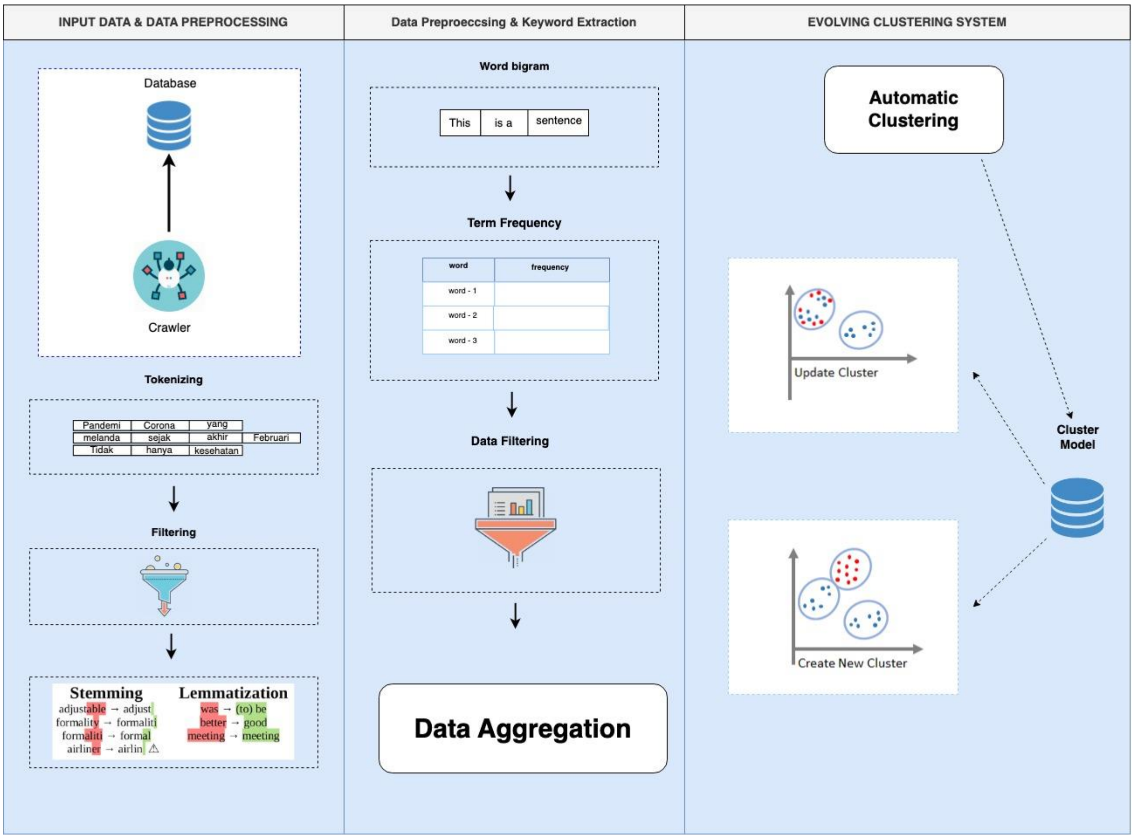

Figure1. System Design

a. Data Acquisition

Data acquisition is a process to get data to be processed, in this process it runs a crawler. At this stage, news data is obtained from RSS news. This research takes news data from 60 RSS links that have been done by previous research [15]. Table 1 shows an example of the RSS link used. It tasked the crawler with retrieving data from these sources on a regular basis.

Table 1. RSS Link Example

Number	Name	Link
1	kompas	http://rss.kompas.com/get/all
2	okezone	http://rss.okezone.com/get/all

b. Tokenizing

Data preprocessing and keyword extraction is one of the important stages for data in the mining process. The data used in the mining process is not always in ideal conditions for processing. Sometimes in the data there are various problems that can interfere with the results of the mining process itself, such as missing values, redundant data, outliers, or data formats that are not in accordance with the system. Therefore, to overcome these problems, the Preprocessing stage is needed. Preprocessing is one of the stages of eliminating problems that can interfere with the results of data processing. In the case of document classification using text-type data, there are several kinds of processes that are generally carried out including tokenizing, filtering (remove punctuation), stopword removal, stemming, and so on.

We will execute tokenizing as the first stage in the data preprocessing section. At the tokenizing stage, the news text data is cleaned of characters that contain other than letters of the alphabet (A-Z) and then split into an array of words. At this stage the text is broken down into an array of words based on spaces. In addition, this stage also removes text in the form of numbers and punctuation.

c. Filtering

At the filtering stage, stopwords have to be removed because the words contained in stopwords are words that commonly appear in Indonesian. This causes that if the stopword is contained in most of the news, it will cause habits in the clustering process. Examples of words that are included in the stopword part and are omitted are the words “oleh”, “dan”, “yang”, and so on are omitted from the text.

d. Stemming & Lemmatization

The next process in data preprocessing is stemming and lemmatization, which is removing affixes and then turning them into basic words.

e. Word Bigram

At the word bigram stage, we assemble the group of words into a phrase of 2 words. This is done without changing the previous set of words.

f. Term Frequency

This stage is to form a Term Frequency (TF) table which contains all words or phrases (w_i) in a news document (d) which is then transformed into a word frequency matrix (F_d) shown in equation (1).

$$F_d = \left[f_{w_i d} \cdots f_{w_n d}\right], w \in d. \tag{1}$$

g. Data Filtering

The next stage is data filtering. At this stage filtering of keywords that can represent news is carried out. The determination of keywords is obtained from words that have a frequency that exceeds the threshold. Threshold (th_d) is obtained from the highest frequency value divided by two which is shown in equation (2). Then, in equation (3) we apply word filtering using the threshold on the term frequency (F_d) that has been obtained to filter out the words that will be transformed into keywords.

$$th_d = \frac{1}{2} max\{F_d\} \tag{2}$$

$$F'_d = \left\{f_{w_i d} \mid f_{w_i d} \geq th_d\right\}, w \in d \tag{3}$$

h. Data Aggregation

At this stage, we form a large matrix (M) which contains the keywords from each document represented vertically as shown in equation (4). But before combining all the frequency terms for each document, we collect the words from all documents in equation (5). If there are No. occurrences of the word, it will be given a value of zero. The following is a matrix table form for data aggregation exemplified in table 2.

$$F'_{d_j} = \left[f_{w_i d_j} \cdots f_{w_n d_j} \right], w \in \cup_{j=0}^{m} \quad d_j \tag{4}$$

$$M = \begin{bmatrix} F'_{d_j} \\ . \\ . \\ F'_{d_m} \end{bmatrix}, m > 0 \tag{5}$$

Table 2. Data Aggregation Example

	Word 1	**Word 2**	**Word n**
News-1	2	3	0
News-2	1	4	1
News-3	0	2	3
News-4	0	0	4
News-n	0	0	0

i. Predefined Cluster

Predefined Cluster is a cluster that is generated for the first time in the clustering process. This stage is only executed once to create a cluster as the initial model. At the Predefined Cluster stage, we use the Automatic Clustering algorithm to create clusters automatically, without defining the number of target clusters [3]. Automatic Clustering developed by Barakbah [3] runs using the Single Linkage Hierarchical Algorithm with the Agglomerative method. Determination of the number of clusters is done automatically by finding the global optimum from the data provided using the Valley Tracing method. The following is a picture of the automatic clustering algorithm.

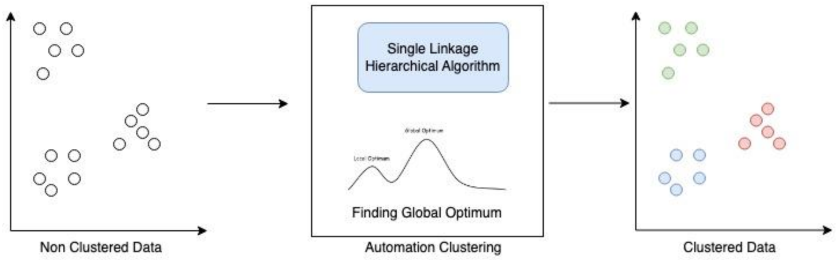

Figure2. Illustration of Automation Clustering Algorithm

The automatic clustering stage will produce the cluster model shown in table 3. Which attributes of the clustering results include Cluster, Radius, Member, and Centroid. Radius is the distance from the centroid to the outermost point of the cluster, which is obtained from the Euclidean distance calculation. And the centroid is the center point in a cluster. The centroid (Ck) is obtained from calculating the average frequency term of the cluster members shown in equation (6).

Table 3. Cluster Model Example

Cluster	**Radius**	**Member**	**Centroid**
1	20.543	4	(9,1.367), (2,1.854),

j. Evolving Clustering

Evolving Evolving System is a model system that can evolve in an environment that is continuously changing (dynamic) and developing. The model system can organize itself in order to adapt to a dynamic environment. This allows the system to fully adapt automatically both in terms of structure and parameters. This research uses a clustering approach in creating an evolving model system, so it is called the Evolving Clustering System. An illustration of the evolving clustering system process is shown in Figure 3.

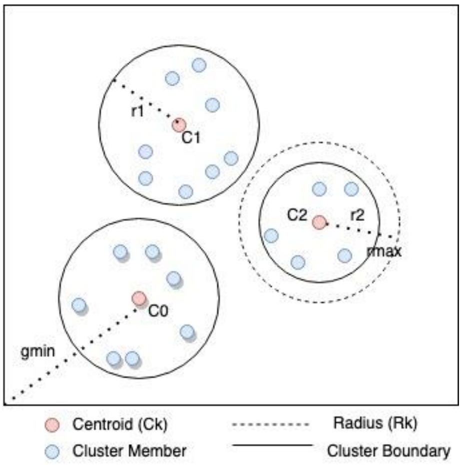

Figure3. Illustration of Evolving Clustering Algorithm

In the early stages, we will define unknown clusters, which are clusters that have no meaning and are indicated as the shortest distance from the gmin from the centroid point (ck) to the zero point. Then the radius of the cluster (rk) is obtained by calculating the centroid distance to the outermost point in the cluster. We will get the maximum distance from the centroid (ck) to all cluster members (Fdj) indicated by equation (7).

$$\left(C_k, F_{d_j}\right) = \sqrt{\sum_{i=0}^{n} \left(c_{kw_i} - f_{w_i d_j}\right)^2} \tag{7}$$

The cluster radius will be an identity for each cluster. We will automatically create a general radius for all clusters which is obtained from the maximum radius (rmax) of all cluster radii (rk) except for the unknown clusters shown in equation (8).

$$\begin{aligned} r_k &= max_{k\in[o,m)}\left\{d\left(C_k, F_{d_j}\right)\right\} \\ r_{max} &= max_{k\in[0,n)}\{r_k\} \end{aligned} \tag{8}$$

Evolving clustering system has two possible conditions which we call operators, namely Create New Cluster and Update Cluster. The first operator will work if there is no similarity between the new news topic and the pre-existing news topic, so that the new news will form a separate cluster. An illustration of this operator is depicted in figure (4) below. The red dot is an illustration of new incoming data. When the news topics do not have anything in common, the new data forms a new cluster with the name cluster 3.

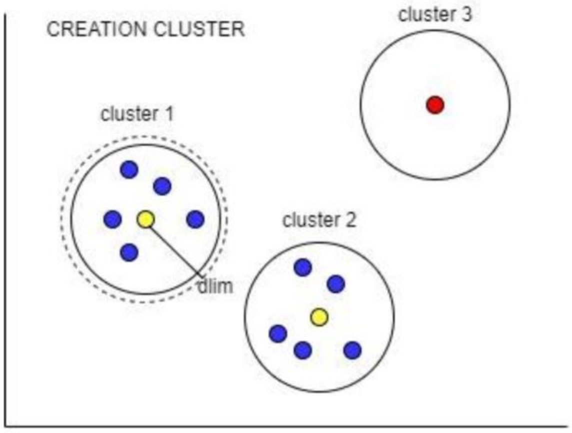

Figure3. Create New Cluster Illustration

The second operator is Update Cluster. This will be executed if there is a similar topic with one of the previously formed news clusters, and the new news will join one of the clusters that have the same topic. So this operator will be executed if this new data point enters one of the cluster limit circles. An illustration of this operator is shown in Figure (5). The red dot is an illustration of new data. When clustering is run, it turns out that the new data has the same topic as one of the clusters, namely cluster 2. So the new data enters or joins cluster 2.

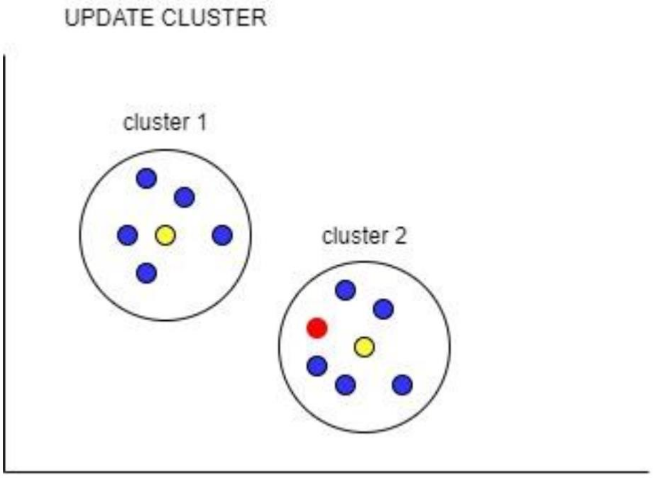

Figure4. Update Cluster Illustration

C. Result and Discussion

In this experiment, there are 4 discussions, namely regarding keyword feature extraction and data aggregation, automatic clustering, Evolving Clustering, and Evaluation.

1. Keyword Feature Extraction and Data Aggregation

In this topic, there are several stages, namely tokenizing, filtering, stemming & lemmatization, word bigram, term frequency, and data aggregation. Testing at the Data Aggregation stage consists of two stages, namely taking keywords from each article then transforming them into words and the frequency of these words then forming a large matrix using words and news id. The first experiment is to take keywords from each news, so that the keywords from each news are taken and used as columns of the large matrix shown in table 4. The last experiment is to make a large matrix that contains keyword data from each news. Making this matrix requires careful consideration, because the matrix that is created must be able to increase according to the number of words and incoming news. Then, as the data growth is two-dimensional, it can grow vertically and horizontally.

Table 4. Data Aggregation Example

Cluster	Bola	Covid	Presiden
News-1	2	0	0
News-2	0	1	0

2. Predefined Cluster

In this study, we obtained data from several online media, as many as 500 news, which will be used as predefined clusters. The attribute used for automatic clustering is the TF (Term Frequency) value of the keywords that represent each news. There are 7430 keywords from all the news that have been processed. A total of 500 news was processed using the Automatic Clustering algorithm and 53 clusters were obtained. The composition of each cluster varies. Table 5 below is an example of the results of the automatic clustering process.

Table 5. Automatic Clustering Result Example

Cluster	News Title	Keyword
1	Janji Prabowo di Depan Ulama se-Jawa: Swasembada Pangan Hingga Turunkan Tarif Listrik	kpu, lapor, prabowo, kait, debat, pemilu, evaluasi, jokowi, serang, laku
2	Dubes RI Sebut Tak Ada WNI Korban Ledakan Bom di Kairo	kairo, bom, ledak, orang, korban, tewas, al, mesir, laku, polisi
7	Ambisi Rachmat Irianto Bawa Timnas Indonesia U-22 Kalahkan Malaysia	timnas, u, timnas indonesia, timnas indonesia u, malaysia, indonesia, indonesia u

3. Evolving Clustering

In evolving clustering, we experiment by using news data incrementally. In March 2022, for a while, the news data we got was 500 news. Previously, we had run automatic clustering of 500 news and obtained 53 clusters, then we processed the additional news using incremental clustering and obtained an additional cluster of 47 so that the current total of clusters is 100 clusters. New news was successfully grouped into clusters according to keywords, one of which was cluster 1, which discussed specific politics regarding elections. New news data made it into the cluster. Table 6 shows the updated sample clusters.

Table 6. Updated Cluster Sample

Cluster	Status	News Title	Keyword
1	Existing	Janji Prabowo di Depan Ulama se-Jawa: Swasembada Pangan Hingga Turunkan Tarif Listrik	kpu, lapor, prabowo, kait, debat, milu, evaluasi, jokowi, serang, laku
	Existing	KPU Gelar Rapat Evaluasi Debat Kedua	
	New	Pengamat: Dukungan Gubernur Riau Tak Signifikan Akan Tambah Elektabilitas Jokowi-Amin	
7	Existing	Indra Sjafri Takkan Lakukan Rotasi Pemain saat Hadapi Malaysia	timnas, u, timnas indonesia, timnas indonesia u, malaysia, indonesia, indonesia u
	Existing	Ambisi Rachmat Irianto Bawa Timnas Indonesia U-22 Kalahkan Malaysia	
	New	Minim Peluang, Akhiri Babak Pertama Timnas U-22 Indonesia Vs Malaysia	

And evolving clustering also produces new clusters of 47 clusters. In the existing model cluster, there is no news cluster that discusses the military, so a new cluster that discusses the military is formed with cluster number 90. Samples from the new cluster are attached in table 7.

Table 7. New Cluster Sample

Cluster	Status	News Title	Keyword
90	New	0	0

4. Evaluation

We have made observations to evaluate the cluster results obtained. In April 2022, we evaluated the clusters that have been produced, we used 10 random cluster samples. Half of the clustering results update the previously formed clusters, and the other half form new clusters. Table 8 shows a sample from the evaluation of clustering results. The observation results in this study showed that there were 17 wrong clusters out of 100 clusters, so that an accuracy of 83% was obtained.

Table 8. Clustering Evaluation Sample

Cluster	News Title	Keyword	Evaluation
1	Menteri Rini Teteskan Air Mata Saat Peresmian Menara Astra	astra	False
7	Winger Lincah Timnas Indonesia U-22 Siap Turun Kontra Malaysia	witan, timnas, u, timnas, u, timnas indonesia, timnas indonesia u, malaysia, indonesia, indonesia u	True
2	Sandiaga Usul Debat Tanpa Panelis, KPU Bilang 'Yang Buat Materi Soal Siapa'	kpu, lapor, prabowo, kait, debat, milu, evaluasi, jokowi, serang, laku	True

D. Conclusion

In conclusion, this study has successfully developed a news clustering system using Evolving Clustering that consists of five stages: Data Acquisition, Keyword Feature Extraction, Data Aggregation, Predefined Clusters, and Evolving Clustering. Through this system, we were able to cluster 1,000 news articles, resulting in a total of 96 clusters formed that were grouped according to their uniform content. Manual observations of the clusters showed that the algorithm performed well, with a cluster accuracy value of 83%. Although some clusters were not quite accurate, the overall performance of the Evolving Clustering algorithm was promising in effectively classifying news articles.

The developed system can potentially contribute to the field of news mining by improving the effectiveness of news retrieval through clustering based on an evolving system. Future studies can further improve the accuracy of the clustering algorithm by addressing the limitations and challenges identified in this study. Overall, this research provides a useful contribution to the field of news clustering and can be a valuable reference for researchers and practitioners interested in developing similar systems.

E. Acknowledgment

I am very grateful to Politeknik Elektronika Negeri Surabaya, which has provided me with a masters level scholarship. We also thank all the researchers involved in this research, and Politeknik Elektronika Negeri Surabaya who have provided facilities for this research. I also want to thank my family and friends who have supported me to complete this research.

F. References

[1] Hafied Cangara, Pengantar Ilmu Komunikasi, Jakarta: Rajawali Pers, 2010.

[2] A. S. M. Romli, Jurnalistik Online : Panduan Mengelola Media Online, Bandung: Nuansa Cendekia, 2012.

[3] D. Z. E. Puspitasari, A. R. Barakbah & I. Winarno, “Automatic Representative News Generation using Automatic Clustering,” Industrial Electronics Seminar (IES) 2011, Surabaya, 2012.

[4] T. M. R. P. Pasya, “Kasus Jesica dalam Bingkai Berita Online : Analisis Framing Peradilan Kasus Jessica pada Portal Berita Detik.com dan Liputan6.com,” Skripsi Sarjana Fakultas Psikologi dan Ilmu Sosial Budaya, Universitas Islam Indonesia, Yogyakarta, 2017.

[5] M. Choirun, “Analisis Framing Berita Vonis Gayus Tambunan Pada Harian Tempo Edisi 24 - 30 Januari 2011,” Skripsi Sarjana UIN Sunan Ampel, Surabaya, 2011.

[6] A. Amalia, “Karakteristik Bahasa Jurnalistik Pada Berita Running Text Di Metro Tv Edisi Oktober 2012,” Skripsi Sarjana Universitas Muhammadiyah Surakarta, Surakarta, 2013.

[7] S. S. K, Jurnalisme Kontemporer, Jakarta: Yayasan Pustaka Obor Indonesia, 2017.

[8] Yudi Wibisono, “Clustering Berita Berbahasa Indonesia” , 2005.

[9] Joel Azzopardi, “Incremental Clustering of News Report”, Algorithms – Open Access Journal, vol.5, no.3, pp. 364 – 378, Dec. 2012.

[10] J. Piskorsi, H. Tanev, M. Atkinson and E. V. D. Goot, “Cluster-Centric Approach to News Event Extraction,” dalam Proceedings of the 2008 conference on New Trends in Multimedia and Network Information Systems, 2008.

[11] R. Florence, B. Nogueira and R. Marcacini, “Constrained Hierarchical Clustering of News Events,” Proceedings of the 21st International Database Engineering & Applications Symposium (IDEAS) 2017, Bristol, 2017.

[12] P. Laban and M. Hearst, “newsLens: building and visualizing long-ranging news stories,” Proceedings of the Events and Stories in the News Workshop, Vancouver, 2017.

[13] M. Sigita, A.R Barakbah, E.M. Kusumaningtyas, I.W., “Automatic Representative News Generation using On-Line Clustering”, EMITTER International Journal of Engineering Technology, vol.1. no. 1, pp.107-113, Dec.2013.

[14] A. M. Bakr, N. M. Ghanem and M. A. Ismail, “Efficient Incremental Density-based algorithm for clustering large datasets,” Alexandria Engineering Journal, vol. 54, no. 4, pp. 1147-1154, Dec. 2015.

[15] I. Shabirin, “Cluster Based News Representative Generation with Automatic Incremental Clustering”, M.Eng. thesis, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia, Jul. 2017.

[16] J. B. Schmitt, C. A. Debbelt, F. M. Schneider, “Too much information? Predictors of information overload in the context of online news exposure”, Information Communication and Society, vol. 21, no. 8, pp. 1151-1167, Apr. 2017.