International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 3, June 2019, pp. 2103~2111
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i3.pp2103-2111

Complete agglomerative hierarchy document's clustering based on fuzzy Luhn's gibbs latent dirichlet allocation

P. M. Prihatini[1], I. K. G. D. Putra[2], I. A. D. Giriantari[3], M. Sudarma[4]
[1]Study Program of Doctoral Engineering Science, Faculty of Engineering, Udayana University, Indonesia
[2]Study Program of Information Technology, Faculty of Engineering, Udayana University, Indonesia
[3,4]Study Program of Electrical Engineering, Faculty of Engineering, Udayana University, Indonesia

Article Info

Article history:

Received Apr 24, 2018
Revised Nov 18, 2018
Accepted Dec 11, 2018

Keywords:

Fuzzy sugeno
Gibbs sampling
Hierarchical clustering
Latent dirichl*et al*location
Luhn's idea

ABSTRACT

Agglomerative hierarchical is a bottom up clustering method, where the distances between documents can be retrieved by extracting feature values using a topic-based latent dirichl*et al*location method. To reduce the number of features, term selection can be done using Luhn's Idea. Those methods can be used to build the better clusters for document. But, there is less research discusses it. Therefore, in this research, the term weighting calculation uses Luhn's Idea to select the terms by defining upper and lower cut-off, and then extracts the feature of terms using gibbs sampling latent dirichl*et al*location combined with term frequency and fuzzy Sugeno method. The feature values used to be the distance between documents, and clustered with single, complete and average link algorithm. The evaluations show the feature extraction with and without lower cut-off have less difference. But, the topic determination of each term based on term frequency and fuzzy Sugeno method is better than Tsukamoto method in finding more relevant documents. The used of lower cut-off and fuzzy Sugeno gibbs latent dirichl*et al*location for complete agglomerative hierarchical clustering have consistent metric values. This clustering method suggested as a better method in clustering documents that is more relevant to its gold standard.

Corresponding Author:

P. M. Prihatini,
Study Program of Doctoral Engineering Science,
Faculty of Engineering,
Udayana University,
Kampus UNUD Bukit Jimbaran Kuta Selatan, Badung, Bali, Indonesia, 80361.
Email: manikprihatini@pnb.ac.id

1. INTRODUCTION

Clustering is one of the tasks in data mining to analyze large amounts of data and is able to generate hidden information that is very useful for decision making. Clustering is an unsupervised classification technique that grouping data with similarities into a cluster [1]. The techniques in clustering include partitioning, hierarchical, grid-based, and density-based method [2]. These methods have been used in various applications such as K-Means for SME Risk Analysis Documents [3], KPrototype for clustering big data based on MapReduce [4], K-Means for earthquake cluster analysis [5], Ward's linkage method for classifying the languages [6], grid and density-based for trajectory clustering [7], and DBSCAN for categorizing districts [8]. The hierarchical clustering method uses a tree concept, which is divided into agglomerative and divisive approaches [9]. Agglomerative is known as bottom up method, while divisive is known as top down method [10]. In general, agglomerative algorithm have three characteristics such as single link, complete link, and average link [11]. The difference between these algorithms is how to determine the distance between data that will be merged. The data that will be merged has diverse forms such

as text, images, sound or video. For text-shaped data, the text is processed first through several steps such as tokenization, filtering, lemmatization or stemming [12].

The results of text processing will be used to generate terms of indexing, which is a vocabulary extracted in the collection of texts, and determine a weight for each term [13]. The terms and its weights will be used to determine the distance between data to be merged in agglomerative algorithm. There are several methods of term weighting in text processing [14]. For vector space model, there is a commonly used Term Frequency-Inverse Document Frequency (TF-IDF) [15],[16]. To overcome the weakness of TF-IDF in addressing synonym and polysemy in natural language, Latent Semantic Indexing (LSI) was developed. Researches on text clustering with Agglomerative Hierarchical Clustering (AHC) algorithm has been done with those term weighting schemes. The AHC algorithm with TF-IDF has been used to cluster the web pages [17], construct taxonomies from a corpus of text documents [18], construct multi-keyword ranked search scheme [19], context aware document clustering [20], automatic taxonomy construction from keywords [21]. The AHC algorithm has also been developed with LSI for document clustering [22], clustering of news articles [23], information retrieval [24]. The weakness of LSI is overcome by developing a topic-based weighting term called Latent Dirichl*et al*location (LDA). LDA is a generative probabilistic model of a corpus, which documents are represented as random mixtures over latent topics, and each topic is characterized by a distribution over words [25]. A document in a corpus is not only identified as a single topic, but can be identified as several topics with their respective probabilities [26]-[28].

LDA has been developed based on a hierarchy, known as hLDA [29], but this method is not able to capture the hierarchical relationship that is formed [30]. Therefore, research needs to be done to classify documents hierarchically by using hierarchical clustering method and LDA for weighting term. The research that integrates LDA into the hierarchical clustering method, especially agglomerative, has already been done. X. Li, H. Wang, G. Yin, T. Wang, C. Yang, Y. Yu, D. Tang [31] used LDA for inducing taxonomy from tags based on word clustering. AHC framework is used to determine how similar every two tags, and then LDA is used to capture thematic correlations among tags that resulted by AHC. D. Tu, L. Chen, G. Chen [32] used LDA to extract the most typical words in every latent topic and apply a multi-way hierarchical agglomerative clustering algorithm (AHC and WordNet) to cluster candidate concept words. The problem is those papers discussed about English text. Until now, the performance of using the LDA method and agglomerative hierarchical clustering in Indonesian text has never been published. If both of these methods are proven to have good performance in clustering Indonesian texts, then it can also be used on other text mining tasks, for example for document summarization.

To overcome this problem, in this research, AHC and LDA are used to cluster documents, where LDA is not used for clustering, but used to generate the weight of terms contained in document text. This research has differences with other related researches for Indonesian text. First, the term weighting calculation used Luhn's Idea to select the terms of text by defining upper cut-off and lower cut-off, and then extracts the feature of terms using Gibbs Sampling LDA combined with the term frequency values and fuzzy Sugeno logic. While, in other research, P.M. Prihatini, I.K.G.D. Putra, I.A.D. Giriantari, M. Sudarma [26] used only TF-IDF for term weighting calculation. Second, the calculation of the distance between documents for AHC is topic-based because it uses the value that resulted by Fuzzy Luhn's Gibbs LDA. Third, the document clustering with AHC uses three characteristics: single link, complete link and average link based on Fuzzy Luhn's Gibbs LDA, and then compares the best AHC characteristics with measurement metrics. While, in other researches, Yuhefizar, B. Santosa, I.K. Eddy, Y.K. Suprapto [33] used Euclidean for distance calculation and single linkage for document clustering. M.A.A. Riyadi, D.S. Pratiwi, A.R. Irawan, K. Fithriasari [34] used single link, complete link and Ward's link based on autocorrelation distance. The discussion in this research is divided as follows. Section 2 discusses about research method. Section 3 discusses about the results and its analysis. Section 4 discusses about the conclusion of this research.

2. RESEARCH METHOD

This research consists of several steps, such as document text processing, term weighting with Fuzzy Luhn's Gibbs LDA, documents clustering with Fuzzy Luhn's Gibbs LDA, and evaluation, as shown in Figure 1.

2.1. Document text processing

In this research, the documents used are news text files obtained from Indonesian news website. Each file is delimited into a collection of terms in the tokenization process. In the filtering process, each term is filtered using a stop words list resulting in a meaningful set of terms. The terms generated by the filtering process, some are already in the form of basic word, some are still have affixes. To make all terms in a uniform shape, all terms are parsing into basic words through the stemming process. In this research,

stemming uses the deletion of affixes method based on rules and basic dictionary. The stemming algorithm used is a modification of Nazief-Adriani [35].

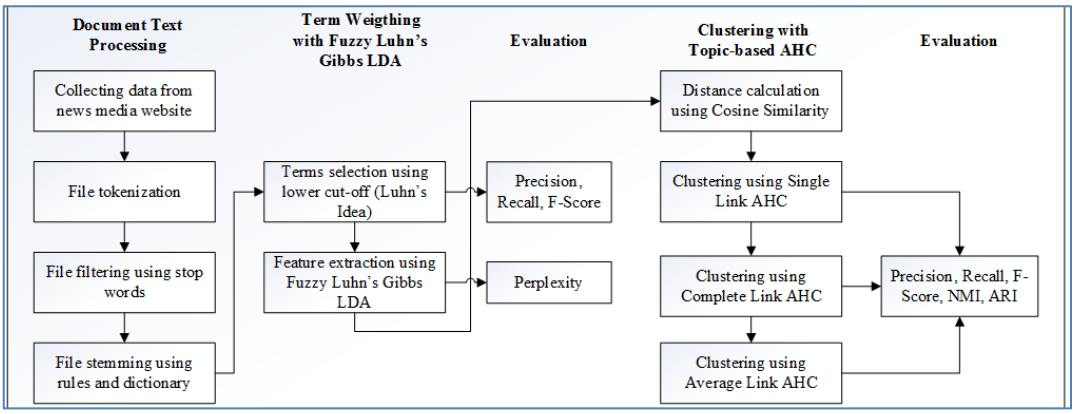

Figure 1. The research method design

2.2. Term weigthing with Fuzzy Luhn's Gibbs LDA

In this research, term weighting is done through term selection and feature extraction. The term selection is based on the concept of Luhn's Idea, where each term is calculated based on its relative frequency against all terms in the document text [36]. Luhn describes the relationship between the occurrence frequencies of a term (term frequency) with the importance of a term in the document. The term with medium-frequency is more important than high or low frequency terms. Low frequency terms are included in the lower cut-off, while high frequency terms are included in the upper cut-off. Medium-frequency terms can be obtained by cutting the upper and lower cut-off. To eliminate terms in the upper cut-off can be done by filtering terms based on stop words list. However, to eliminate terms in the lower cut-off, so far there has been no research that can determine effective ways to determine the lower cut-off limits.

In this research, the elimination of terms in the upper cut-off limit is done twice. First, it done by filtering of terms based on stop word list. Second, the filtering results are filtered again through stemming proccess. For the elimination of terms in the lower cut-off limit is based on the stemming result, with different percentage removal values for each text document, as in (1). This is based on the idea that each document has different text lengths, so that no single constant value can be taken for all documents. Variable lco_d (lower cut-off document) refers to the lower-cut-off constant value for document d (in the form of a positive integer). Variable fs_d (false-stemming document) refers to the number of unsuccessful term stemming in document d. Variable fr_d (filtering result document) refers to the number of terms in document d used for the stemming process. Variable ts_d (true-stemming document) refers to the number of successful term stemming in document d.

$$lco_d = \left(\frac{fs_d}{fr_d}\right)|ts_d| \quad (1)$$

The term selection result is a collection of selected terms of each document that have important meanings to be processed at the feature extraction. In this research, feature extraction is done by topic-based LDA method. LDA has some reasoning algorithms, one of which is Gibbs Sampling that have proven effective in conducting the topic sampling process [28]. In general, in the initialization process, Gibbs Sampling assigns the topic of each term randomly using a multinomial random function. However, the use of this function cannot represent the existence of each term in the topic. Therefore, in this research, the determination of topic for each term in the initialization process is done based on the highest occurrence frequency (*tf*) of the term in all topics, as in (2). Variable $z_{t,k}$ similar with k refers to the topic. Variable $tf_{t,k}$ refers to *tf* value of term t on topic k. To calculate the probability of each term in the sampling process used the formula as in (3). Variable $p_{t,k}$ refers to probability value of sampling for term t on topic k. Variable nkw_{-1} refers to the value of the topic-term matrix by ignoring the current term value. Variable V is the unique number of terms in all documents. Variable ndk_{-1} refers to value of the document-topic matrix by ignoring the current term value. Variable β determine the mixing proportion of documents on the topics, while α determines the mixture components of words on the topics [37]. Variable K is the number of topic.

$$z_{t,k} = k \sim \max(tf_{t,k}) \quad (2)$$

$$p_{t,k} = \frac{nkw_{-1}\ \beta}{(\sum nkw_{-1}-1)+(V\ \beta)}\ \frac{ndk_{-1} \propto}{(\sum ndk_{-1}-1)+(K \propto)} \quad (3)$$

In general, Gibbs Sampling in the LDA requires several times iterations for the sampling process until it reaches convergent conditions. This takes time and high complexity. The addition of fuzzy Tsukamoto logic into the sampling process can accelerate the achievement of convergent conditions with good measurement values [26]. The fuzzy logic concept that used in that research will be improved through this research by using Sugeno method to increase the accuracy value, considering the output of fuzzy logic which needed for sampling is a constant value. In this research, the upper and lower limits for the fuzzy curve are determined based on the *tf* value of each term. Fuzzification uses a triangular curve with the probability value of the sampling result for each term p, as in (4). Variable *u[t]* refers to the degree of membership for term *t*. Variable *a* refers to the lower bound of the curve. Variable *b* refers to the peak of the curve. Variable *c* refers to the upper bound of the curve. The implication function used is OR because fuzzy logic here is used to determine the probability value of term for one topic, while all topic will be determined in sampling process. The rule composition generates the α_p value based on the maximum value of all *u[t]* as in (5), and the value of z_o is based on the term probability value across the topic whose value is not equal to zero, as in (6). Variable *t* refers to the term probability of sampling result. Variable *zo* refers to the composition output. Variable *n* refers to the number of topics whose term probability is not equal to zero. For the defuzzification, the final output of fuzzy *z* is obtained by calculating the mean value, as in (7). The value of *z* is used as the probability value of term *p* for topic *k* and will be used for the next sampling process until it reaches convergent conditions. After convergence, the final value of *z* will be the feature value for each term and ready for clustering.

$$\mu[t] = \begin{cases} 0, & t \le a, t \ge c \\ \frac{t-a}{b-a}, & a \le t \le b \\ \frac{b-t}{c-b}, & b \le t \le c \end{cases} \quad (4)$$

$$\propto_p = \max(\mu[t_1], \mu[t_2], \dots, \mu[t_k]) \quad (5)$$

$$z_o = \frac{\sum(t \neq 0)}{n} \quad (6)$$

$$z = \frac{\propto_p z_o}{\propto_p} \quad (7)$$

2.3. Documents clustering with Fuzzy Luhn's Gibbs LDA

The feature values obtained at the feature extraction are used to calculate the distance between documents to be used in the clustering process. In this research, distance calculations using the Cosine Similarity, as in (8). Variable |*di-dj*| refers to the distance between documents *i* and *j*. Variable d_i refers to document *i*, while d_j refers to document *j*.

$$|d_i - d_j| = \frac{d_i\ x\ d_j}{\sqrt{d_i^2\ x\ d_j^2}} \quad (8)$$

The distance between documents is used to cluster document using three types of AHC algorithms. In the Single Link AHC algorithm, clusters are based on the smallest distance between pairs of two documents, as in (9). In the Complete Link AHC algorithm, clusters are based on the largest distance between pairs of two documents, as in (10). In the Average Link AHC algorithm, clusters are based on the average distance between pairs of two documents, as in (11). Variable d_{ij} refers to the selected pair of documents *i* and *j*.

$$d_{ij} = \min(|d_i - d_1|, |d_i - d_2|, \dots, |d_i - d_j|) \quad (9)$$

$$d_{ij} = \max\left(|d_i - d_1|, |d_i - d_2|, \dots, \left|d_i - d_j\right|\right) \quad (10)$$

$$d_{ij} = \text{avg}(|d_i - d_1|, |d_i - d_2|, \dots, |d_i - d_j|) \quad (11)$$

2.4. Metrics evaluation

In this research, evaluation is done in two steps: evaluation of the feature extraction results and evaluation of the clustering results. The text document used in this research has been classified into five categories by Indonesian news media websites, so it can be used as gold standard for the evaluation process.

Evaluation of feature extraction results is done by comparing results with lower cut-off and without lower cut-off. An evaluation was also performed to compare the feature extraction results between the Fuzzy Gibbs LDA method [26] and Fuzzy Luhn's Gibbs LDA that used in this research. The evaluation was performed using two measurement metrics. First, the perplexity is used to measure the ability of the Fuzzy Luhn's Gibbs LDA feature extraction method to generalize the hidden data, as in (12) and (13) [25]. The smaller value of the perplexity indicates the better performance of the method. Variable $P(\widetilde{W}|M)$ refers to perplexity value. Variable *M* refers to the number of documents. Variable *V* is the unique number of terms in all documents. Variable N_m^t refers to the number of occurrences of the word *t* in document *m*. Variable *K* is the number of topic. Variable $\varphi_{k,t}$ refers to the number of documents for each topic. Variable $\vartheta_{m,k}$ refers to the number of words for each topic.

$$P(\widetilde{W}|M) = \exp -\frac{\sum_{m=1}^{M} \log p(\vec{w}_m|M)}{\sum_{m=1}^{M} N_m} \quad (12)$$

$$\log p(\vec{w}_m|M) = \sum_{t=1}^{V} N_m^t \log(\sum_{k=1}^{K} \varphi_{k,t} \ . \vartheta_{m,k}) \quad (13)$$

Second, Precision (P), Recall (R) and F-Score (F) metrics are used to measure the ability of the Fuzzy Luhn's Gibbs LDA feature extraction method in finding relevant documents according to the gold standard, as in (14)-(16) [14]. Variable *TP* is true positive, refers to the number of relevant items retrieved. Variable *FP* is false positive, refers to the number of non-relevant items retrieved. Variable *FN* is false negative, refers to the number of relevant items that cannot retrieved. The greater value of P, R and F indicates the better performance of the methods.

$$P = \frac{TP}{TP+FP} \quad (14)$$

$$R = \frac{TP}{TP+FN} \quad (15)$$

$$F = \frac{2 \ . \ P \ . \ R}{P+R} \quad (16)$$

Evaluation of clustering results is done by comparing the clustering results between Single Link, Complete Link and Average Link AHC using feature extraction results. In this research, the evaluation was performed using five measurement metrics. Precision, Recall, and F-Score (PRF) are used to measure the ability of methods to cluster relevant documents according to the gold standard, as in (14)-(16). The fourth is Normalized Mutual Information (NMI), as in (17)-(20) [38]. Variable $I(\Omega, C)$ refers to the value of mutual information for class (gold standard) and cluster. Variable $H(\Omega)$ refers to the entropy of class. Variable $H(C)$ refers to the entropy of cluster. Variable w_k refers to the number of document belongs to class *k*. Variable c_j refers to the number of document belongs to cluster *j*.

$$NMI = \frac{2\, I(\Omega,C)}{H(\Omega)+H(C)} \quad (17)$$

$$I(\Omega, C) = \sum_k \sum_j \frac{|w_k \cap c_j|}{N} \log \frac{N|w_k \cap c_j|}{|w_k||c_j|} \quad (18)$$

$$H(\Omega) = -\sum_k \frac{|w_k|}{N} \log \frac{|w_k|}{N} \quad (19)$$

$$H(C) = -\sum_j \frac{|c_j|}{N} \log \frac{|c_j|}{N} \quad (20)$$

The fifth is Adjusted Rand Index (ARI), as in (21)-(24) [39]. Variable m_{ij} refers to the number of document belongs to class *i* and cluster *j*. Variable C_i refers to the number of document belongs to class *i*. Variable C'_j refers to the number of document belongs to cluster *j*. The greater value of P, R, F, NMI and ARI indicates the better performance of the methods.

$$ARI = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2) - t_3} \quad (21)$$

$$t_1 = \sum_{i=1}^{k}\binom{|C_i|}{2} \quad (22)$$

$$t_2 = \sum_{j=1}^{l}\binom{|C'_j|}{2} \quad (23)$$

$$t_3 = \frac{2t_1t_2}{n(n-1)} \quad (24)$$

3. RESULTS AND ANALYSIS

3.1. Fuzzy Luhn's Gibbs LDA

The evaluation results of the Fuzzy Luhn's Gibbs LDA feature extraction method can be seen in Table 1. The value in the table shows the comparison between Fuzzy Luhn's Gibbs LDA feature extraction method and Fuzzy Gibbs LDA that have published by P.M. Prihatini, I.K.G.D. Putra, I.A.D. Giriantari, M. Sudarma [26]. The Fuzzy Luhn's Gibbs LDA feature extraction method uses lower cut-off and without lower cut-off for the selection feature method, while Fuzzy Gibbs LDA did not use Luhn's concept.

Table 1. Comparison of Fuzzy Luhn's Gibbs LDA and Fuzzy Gibbs LDA

Metrics evaluations	Fuzzy Luhn's Gibbs LDA		The difference	Fuzzy Gibbs LDA	The difference	
	Lower cut-off (1)	Without lower cut-off (2)	(1) & (2)	(3)	(1) & (3)	(2) & (3)
Perplexity	0.0375	0.0339	0.0036	0.0376	0.0001	0.0037
Precision	0.9435	0.9515	0.0080	0.8975	0.0460	0.0540
Recall	0.9280	0.9360	0.0080	0.8486	0.0794	0.0874
F-Score	0.9296	0.9387	0.0091	0.8420	0.0876	0.0967

The evaluation results in Table 1 shows that the feature extraction with lower cut-off using equation (1) gives the evaluation value not much different than without the lower cut-off. The difference of metric measurement values between the two methods is very small with the range from 0.0036 to 0.0091. This insignificant difference occurs because the feature selection in this research has been done through two step of the upper cut-off, which at filtering step with stop word list and then at stemming step. These two steps have filtered the term with frequencies that appear frequently and rarely appear, so it results a list of meaningful terms for the feature extraction. The lower cut-off process with the value adjusted to the length of the document only removes a small portion of the meaningful term in the feature selection so it does not significantly affect the feature extraction results.

The evaluation results in Table 1 also shows **Fuzzy Gibbs LDA** method resulted perplexity of **0.0376**, while **Fuzzy Luhn's Gibbs LDA** in this research gives the value of perplexity **0.0375** for **lower cut-off** and **0.0339 without lower cut-off**. This indicates that the Fuzzy Luhn's Gibbs LDA algorithm performs as well as Fuzzy Gibbs LDA in generating hidden data. But, the results of the P, R, and F metric indicate that the **Fuzzy Luhn's Gibbs LDA** algorithm performed gives better results ranging from **0.9280 to 0.9515** than **Fuzzy Gibbs LDA** algorithm ranging from **0.8420 to 0.8975**. The increasing value of PRF metric shows that the topic determination of each term for initial sampling that performed based on the highest occurrence frequency (*tf*) of term to all topics by using Luhn's Idea and the use of the Fuzzy Sugeno method is better able to find documents relevant to the gold standard. This indicates that Fuzzy Luhn's Gibbs LDA algorithm is a better choice in performing feature extraction for clustering.

2.2. AHC with Fuzzy Luhn's Gibbs LDA

The evaluation results of the AHC algorithms performed based on the Fuzzy Luhn's Gibbs LDA feature extraction can be seen in Table 2. The evaluation results in Table 2 shows that the feature selections with lower cut-off or without lower cut-off do not affect the performance of the AHC algorithms in the clustering process. It can be seen from the measurement metric values that both feature selection methods produce Complete Link AHC algorithm as the AHC clustering algorithm with the best metric value. The differences for the Complete Link AHC algorithm with both feature selection methods ranges from 0.0003 to 0.0263. This shows that both feature selection methods can be used as a good choice in clustering process with AHC. But, in terms of the consistency of the value generated by the five metric measurements, **Complete Link AHC and Fuzzy Luhn's Gibbs LDA with lower cut-off** have consistent metric values,

ranging from **0.9196 to 0.9989**, with differences ranging from 0.0213 to 0.0793; while **Complete Link AHC and Fuzzy Luhn's Gibbs LDA without lower cut-off** have values ranging from **0.8933 to 0.9974**, with differences ranging from 0.0213 to 0.0793, and decreased the NMI metric value 0.0263 compared to lower cut-off. The results of AHC with Fuzzy Luhn's Gibbs LDA compared with the results of AHC with autocorrelation distance that have published by M.A.A. Riyadi, D.S. Pratiwi, A.R. Irawan, K. Fithriasari [34]. In their research, **Complete Link AHC with Autocorrelation distance** resulted accuracy value of **0.8235**. Therefore, the use of Complete Link AHC and Fuzzy Luhn's Gibbs LDA with lower cut-off is more relevant as a better clustering method in clustering documents especially Indonesian text news.

Table 2. Evaluation Results of AHC with Fuzzy Luhn's Gibbs LDA

Metrics Evaluations	Fuzzy Luhn's Gibbs LDA with Lower cut-off			Fuzzy Luhn's Gibbs LDA without Lower cut-off			The difference
	Single Link AHC	Complete Link AHC (1)	Average Link AHC	Single Link AHC	Complete Link AHC (2)	Average Link AHC	(1)(2)
Precision	0.8255	0.9549	0.9169	0.8642	0.9583	0.9340	0.0034
Recall	0.6021	0.9273	0.8179	0.6714	0.9247	0.8201	0.0026
F-Score	0.6963	0.9409	0.8646	0.7557	0.9412	0.8733	0.0003
NMI	0.5827	0.9196	0.7208	0.6075	0.8933	0.6474	0.0263
ARI	0.9523	0.9989	0.9128	0.9534	0.9974	0.9017	0.0015

4. CONCLUSION

Complete Link AHC and Fuzzy Luhn's Gibbs LDA with lower cut-off algorithm that has built in this research can improve the quality of clusters generation for document clustering especially for Indonesian text news. This is shown by the value of evaluation metrics, which are Precision, Recall, F-Score, Perplexity, Normalized Mutual Information, and Adjusted Rand Index. The values of Precision, Recall and F-Score for lower cut-off have less difference than without the lower cut-off, which means both methods can be used in term selection process. The values of Perplexity, Precision, Recall and F-Score for Fuzzy Luhn's Gibbs LDA algorithm was increased, which means it performed better than Fuzzy Gibbs LDA. The value of Precision, Recall, F-Score, Perplexity, Normalized Mutual Information, and Adjusted Rand Index showed that the Complete Link AHC and Fuzzy Luhn's Gibbs LDA algorithm as the best AHC clustering algorithm, with or without lower cut-off. But, the Complete Link AHC algorithm and Fuzzy Luhn's Gibbs LDA with lower cut-off produce more consistency value for the five metric measurements, which means is more relevant to its gold standard.

REFERENCES

[1] Bansal A., *et al*., "Improved K-mean clustering algorithm for prediction analysis using classification technique in data mining," *International Journal of Computer Applications*, vol. 157, pp. 35-40, 2017.

[2] Arora P., *et al*., "Analysis of K-Means and K-Medoids algorithm for big data," *Procedia Computer Science*, vol. 78, pp. 507-12, 2016.

[3] Wahyudin I., *et al.*, "Cluster analysis for SME risk analysis documents based on Pillar K-Means," *TELKOMNIKA Telecommunication Computing Elctronics and Control*, vol. 14, pp. 674-83, 2016.

[4] Bathla G., *et al*., "A novel approach for clustering big data based on MapReduce," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, pp. 1711, 2018.

[5] Kamat R. K., *et al*., "Earthquake cluster analysis: K-Means approach," *Journal of Chemical and Pharmaceutical Sciences*, vol. 10, pp. 250-3, 2017.

[6] Strauss T., *et al*., "Generalising ward's method for use with manhattan distances," *PloS one*, vol. 12, 2017.

[7] Mao Y., *et al*., "An adaptive trajectory clustering method based on grid and density in mobile pattern analysis," *Sensors (Basel)*, vol. 17, 2017.

[8] Majumdar J., *et al*., "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big Data*, vol. 4, 2017.

[9] Balcan M. F., *et al.*, "Robust hierarchical clustering," *Journal of Machine Learning Research*, vol. 15, pp. 4011-51, 2014.

[10] Goulas C., *et al.*, "HCuRMD: Hierarchical clustering using relative minimal distances," in Chbeir R. M. Y., *et al.*, "Artificial Intelligence Applications and Innovations," IFIP Advances in Information and Communication Technology, Springer, pp. 440-7, 2015.

[11] Marathe M., *et al*., "A survey of clustering algorithms for similarity search," *International Journal of Pure and Applied Mathematics*, vol. 114, pp. 343-51, 2017.

[12] Allahyari M., *et al*., "A brief survey of text mining: classification, clustering and extraction techniques," KDD Bigdas, 2017.

[13] Ribeiro M. N., *et al*., "Local feature selection in text clustering," in Köppen M. K. N. and Coghill G., "Advances in Neuro-Information Processing ICONIP 2008," Lecture Notes in Computer Science. Berlin, Heidelberg, Springer, 2009.
[14] Manning C. D., *et al*., "An Introduction to Information Retrieval," England, Cambridge University Press, 2008.
[15] Islam M. R., *et al*., "Technical approach in text mining for stock market prediction: a systematic review," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, pp. 770-7, 2018.
[16] Amoli P. V., *et al*., "Scientific douments clustering based on text summarization," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 5, pp. 782-7, 2015.
[17] Ramage D., *et al*., "Clustering the tagged web," *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.
[18] Knijff J., *et al*., "Domain taxonomy learning from text: the subsumption method versus hierarchical clustering," *Data & Knowledge Engineering*, vol. 83, pp. 54-69, 2013.
[19] Indhuja A., *et al*., "A multi-keyword ranked search scheme over encrypted based on hierarchical clustering index," *International Journal On Smart Sensing And Intelligent Systems*, vol. 10, pp. 539-59, 2017.
[20] Venkateshkumar P., *et al*., "Using data fusion for a context aware document clustering," *International Journal of Computer Applications*, vol. 72, pp. 17-20, 2013.
[21] Song Y., *et al*., "Automatic taxonomy construction from keywords via scalable bayesian rose trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, pp. 1861-74, 2015.
[22] Kuta M., *et al*., "Comparison of latent semantic analysis and probabilistic latent semantic analysis for documents clustering," *Computing and Informatics*, vol. 33, pp. 652–66, 2014.
[23] Rott M., *et al*., "Investigation of latent semantic analysis for clustering of Czech news articles," *25th International Workshop on Database and Expert Systems Applications (DEXA)*, 2014.
[24] Park H., *et al*., "Agglomerative hierarchical clustering for information retrieval using latent semantic index," *IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 426-31, 2015.
[25] Blei D. M., *et al*., "Latent dirichl*et al*location," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.
[26] Prihatini P. M., *et al*., "Fuzzy-gibbs latent dirichl*et al*location model for feature extraction on Indonesian documents," *Contemporary Engineering Sciences*, vol. 10, pp. 403-21, 2017.
[27] Prihatini P. M., *et al*., "Feature extraction for document text using Latent Dirichl*et al*location," *Journal of Physics: Conference Series*, vol. 953, pp. 012047, 2018.
[28] Prihatini P. M., *et al*., "Indonesian text feature extraction using gibbs sampling and mean variational inference latent dirichl*et al*location," *Quality of Research (QIR): International Symposium on Electrical and Computer Engineering, 2017 15th International Conference on*, 2017.
[29] Blei D. M., *et al*., "Hierarchical topic models and the nested chinese restaurant process," *NIPS'03 Proceedings of the 16th International Conference on Neural Information Processing Systems*, pp. 17-24, 2003.
[30] Yerebakan H. Z., *et al*., "Hierarchical latent word clustering," *Bayesian Nonparametrics: The Next Generation NIPS 2015 Workshop*, 2015.
[31] Li X., *et al*., "Inducing taxonomy from tags : an agglomerative hierarchical clustering framework," *International Conference on Advanced Data Mining and Applications ADMA 2012: Advanced Data Mining and Applications*, pp. 64-77, 2012.
[32] Tu D., *et al*., "WordNet based multi-way concept hierarchy construction from text corpus," *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1647-8, 2013.
[33] Yuhefizar Y., *et al*., "Combination of Cluster Method for Segmentation of Web Visitors," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 11, pp. 207, 2013.
[34] Riyadi M. A. A., *et al*., "Clustering stationary and non-stationary time series based on autocorrelation distance of hierarchical and k-means algorithms," *International Journal of Advances in Intelligent Informatics*, vol. 3, pp. 154-60, 2017.
[35] Asian J., *et al*., "Stemming Indonesian," *Proceedings of the Twenty-eighth Australasian conference on Computer Science*, 2004.
[36] Kocabas I., *et al*., "Investigation of Luhn's claim on information retrieval," *Turk J Elec Eng & Comp Sci*, vol. 19, pp. 993-1004, 2011.
[37] Heinrich G., "Parameter estimation for text analysis," University of Leipzig, Germany, 2008.
[38] Fred A. L. N., *et al*., "Robust Data Clustering," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 3, pp. 128-36, 2003.
[39] Kuncheva L. I., *et al*., "Using diversity in cluster ensembles," vol. 2, pp. 1214-9, 2004.

BIOGRAPHIES OF AUTHORS

Putu Manik Prihatini was born in Bali (Indonesia) on March 17, 1980. She is earned a bachelor's degree of informatics engineering at Sekolah Tinggi Teknologi Telkom Bandung (Indonesia) in 2002, and her master's degree at Universitas Udayana (Indonesia) in 2012. She is a lecturer at Politeknik Negeri Bali (Indonesia) since 2002 until now. Currently, she is being a doctoral student of Ilmu Teknik at Universitas Udayana (Indonesia). Her research and interest is Text Mining, Information Retrieval System, and Soft Computing. Putu Manik Prihatini, ST, MT. Email: manikprihatini@pnb.ac.id

I Ketut Gede Darma Putra was born in Bali (Indonesia) on April 24, 1974. He is earned a bachelor's degree of informatics at Institut Teknologi Sepuluh November (ITS-Indonesia); masters and doctoral degrees at Universitas Gadjah Mada (UGM-Indonesia). He is a lecturer at Universitas Udayana (Indonesia) since 1999 until now. His research and interest is Data Mining and Image Processing. Prof. Dr. I Ketut Gede Darma Putra, S.Kom., MT. Email: ikgdarmaputra@unud.ac.id

Ida Ayu Dwi Giriantari was born in Bali (Indonesia) on December 13, 1965. She is earned a bachelor's degree of electrical engineering at Universitas Udayana (Indonesia); masters and doctoral degrees at The University of New South Wales (Australia). She is a lecturer at Universitas Udayana Indonesia since 1991 until now. Her research and interest is electric power system, renewable energy technology and application, smart grid and control. Prof. Ir. Ida Ayu Dwi Giriantari, M.Eng,Sc., PhD. Email: dayu.giriantari @unud.ac.id

Made Sudarma was born in Bali (Indonesia) on December 31, 1965. He is earned a bachelor's degree of informatics at Institut Teknologi Sepuluh Nopember (ITS-Indonesia); master's degree at School of Information Technology and Engineering, Ottawa University (Canada) and doctoral degrees at Universitas Udayana (Indonesia). He is a lecturer at Universitas Udayana Indonesia since 1993 until now. His research and interest is Data Mining and Image Processing. Dr. Ir. Made Sudarma, M.A.Sc. Email: msudarma@unud.ac.id