Multi-Document Text Summarization Based on Semantic Clustering and Selection of Representative Sentences Using Latent Dirichlet Allocation

Oktefvia Aruda LISJANA[1], Dian Palupi RINI[2*], and Novi YUSLIANI[3]

[1]*oktfvardl@gmail.com, Sriwijaya University, Indonesia*
[2]*dprini@unsri.ac.id, Sriwijaya University, Indonesia*
[3]*novi_yusliani@unsri.ac.id, Sriwijaya University, Indonesia*
**Corresponding author: dprini@unsri.ac.id*

ABSTRACT

Information in the form of online news texts has become one of the most important in this information age. There is a lot of online news that is produced every day, but this news often provides the same contextual content but with a different narrative. This makes it difficult for readers to get complete information. Therefore, we need a solution that can retrieve information in several online news texts to be more effective and efficient in the form of summarizing multi-document texts. In this research, extraction summarization is used, which is to arrange the sentences in the source document to a shorter form. The methods used are Latent Semantic Indexing (LSI) and Similarity-Based Histogram Clustering (SHC) to create semantic sentence clusters and the Latent Dirichlet Allocation (LDA) method to select representative sentences from the formed clusters. Recall-Oriented Understanding Testing for Gisting Evaluation (ROUGE) is used to measure test results. The proposed methods can reach a ROUGE-1 multi-F-measure value of 0.481.

Keywords: *multi-document, semantic clustering, text summarization, LDA*

INTRODUCTION

The amount of information that is presented online news makes it difficult for readers to get complete information. Readers need a lot of time to read a variety of different news sources, but with the same context. A solution is needed that can make searching information on several news texts more effective and efficient in the form of summarizing multi-document texts. Summarizing multi-document text automatically produces a more concise form of the document but still contains important information [1]. Summarizing the multi-document text is done by combining and integrating information in a collection of documents, synthesizing knowledge, and making it into a more concise form [2]. A good summary is a summary that has good coverage, covers as many important concepts as possible, and avoids the redundancies contained in the source document [3]. The clustering process is needed to achieve good coverage in a summary result. One clustering method that can guarantee cluster coherence is the SHC method [4]. To cover as many important concepts as possible and avoid redundancies, the process of selecting representative sentences is needed. In this study,
the LDA method is proposed to give weight to sentences by capturing topics as keywords.

METHODOLOGY

Data

The research data consisted of two, namely data from 70 Indonesian online news texts with ten news topics from seven different news sources and summary data from 70 news texts summarized by experts. Both data are copied into a file with a .txt extension and each file that has the same topic is stored in the same folder. Online news text data will go through text preprocessing stages which will convert text data into numerical data.

The Process of Summarizing Text

In this study, there are five stages of research that will be shown in Figure 1, namely preprocessing text, semantic clustering of sentences, sorting clusters using cluster importance, selection of repressive sentences and then getting summary results. The text preprocessing carried out at this stage are sentence segmentation, case folding, tokenizing, filtering, and stemming. After that, the weighting is done using the term frequency. The results of weighting words will be input into the process of semantic clustering of sentences consisting of LSI and SHC methods. LSI is done to obtain the degree of similarity between pairs of sentences based on relations semantics. Next, the grouping process is carried out the sentence using the SHC method, which is a process grouping sentences into each cluster coherent based on the histogram ratio. In sorting clusters, cluster importance sort according to the total weight value terms which are frequent words. LDA method is used to select sentences from formed clusters. Sentences with the highest LDA values will be used as summary sentence candidates in a cluster.

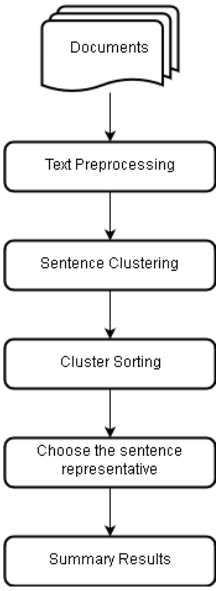


Figure 1. Proposed Multi-Document Text Summarization Framework

Sentence Semantic Clustering

In this research, a clustering process is needed to group the sentences resulting from word weighting. Each sentence in a document must be identified using an appropriate cluster to guarantee good coverage. The semantic process is used to measure the similarity between sentences. The Latent Semantic Indexing (LSI) method can determine the similarity between sentences in forming clusters [7]. In this study, a Similarity-Based Histogram Clustering (SHC) method is proposed which uses a cluster similarity histogram approach to guarantee the coherence of a cluster.

Latent Semantic Indexing (LSI)

LSI was chosen because it can get semantic relations that are not visible in the case of text clustering. For example, the word “die” and “passed away” has the same meaning but with a different form of the word. If the two words are measured using other methods such as Uni Gram Matching Based Similarity, it will produce a small similarity value, even though the two words have a high semantic relationship. So the semantic relationship between words is needed to calculate sentence similarity. One of the main features of LSI is its expertise in creating relationships between terms that appear in the same context.

LSI requires a Singular Value Decomposition (SVD) process for the matrix obtained with the word rows and sentence columns [4]. SVD is done as a matrix composition process with the aim of the matrix being a clean matrix from the noise dimension. The value of similarity is measured by cosine similarity. Cosine similarity is a calculation of similarity based on the cosine angle between two vectors shown in figure 2, the first sentence vector denoted by qi and the second sentence vector notated by s_j [4]. Based on the cosine of the angle between the two vectors, the similarity value has a range between 0-1. If the similarity value is close to 1, the pair of sentences will become more similar. The Cosine Similarity formula can be written as follows:

$$\cos\left(\overrightarrow{q_\iota}, \overrightarrow{s_J}\right) = \frac{\overrightarrow{q_\iota}.\overrightarrow{s_J}}{|\overrightarrow{q_\iota}||\overrightarrow{s_J}|} \qquad (1)$$

$\overrightarrow{q_\iota}$ is i-th sentence vector obtained from the centroid of the product of the term vector to the singular value matrix in SVD. $\overrightarrow{s_J}$ is j-th sentence vector obtained from the product of the singular value matrix of the sentence vector in SVD. While $\overrightarrow{|q_\iota|}$ and $\overrightarrow{|s_J|}$ are sentence length vector from q_i and $s_{j.}$

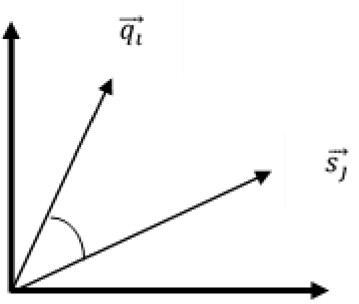


Figure 2. Vector in Cosine Similarity

Similarity based Histogram Clustering

This research uses the Similarity-Based Histogram Clustering (SHC) method for sentence clusters. SHC can maintain the coherence of formed clusters [4]. At SHC there is the concept of cluster similarity histogram. The concept is a statistical representation of a distribution of similarity pairs between members in a cluster [4]. The number of a bin in the histogram shows the interval of a certain similarity value [5].

The high degree of coherence in a cluster can be achieved by maintaining a high degree of similarity between members remains high [5]. In the concept of similarity histogram, this can be interpreted by maintaining the distribution of similarity so that it tends to the right. Illustration of the SHC concept can be seen in Figure 3. The method used by the SHC to get a good level of coherence in a cluster is to maintain a similarity level [5]. Every new sentence that will be included in the new cluster will be tested first by doing calculations before and after the addition of the sentence. If the sentence reduces the quality of coherence, then the sentence is not added.

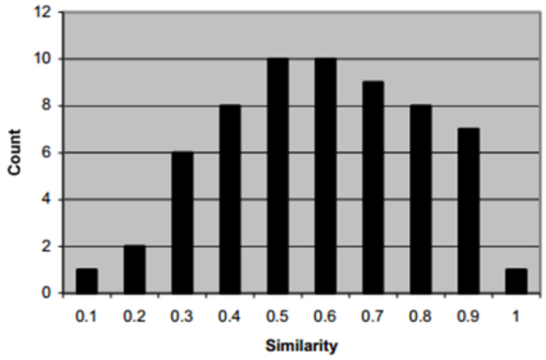


Figure 3. Similarity Histograms in Clusters

The quality level of a similarity histogram is determined by calculating the ratio of similarity that is above the threshold. If the number of sentences in a cluster is denoted by n, then the sum of the pair of the sentences is n (n + 1) / 2, denoted by m. Sim = {sim_1, sim_2, ..., sim_m} is a collection of pairs of similarities between sentences. Similarity histogram in a cluster is denoted by H = {h_1, h_2, h_3, ..., h_{nb}}. Calculation of the h-i value can be seen as follows:

$$h_j = count(sim_j) \qquad (2)$$

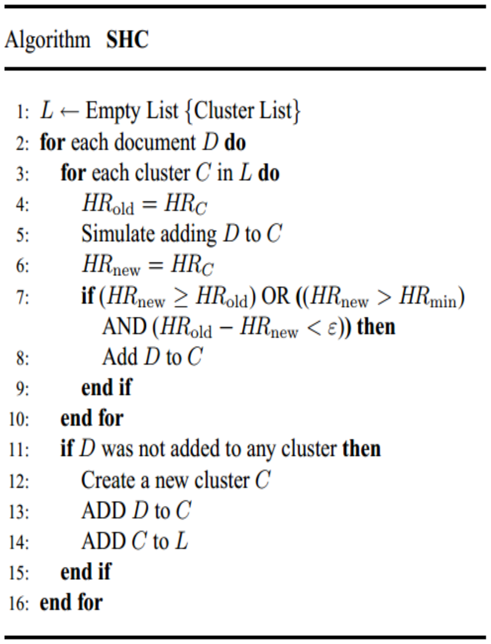
for $sim_{li} \leq sim_j \leq sim_{ui}$

h_j is the amount of similarity in j. sim_{li} is lower limit of similarity in bin i while sim_{ui} is upper limit of similarity in bin i. Calculation of Histogram Ratio (HR) in a cluster is written in the following equation:

$$HR = \frac{\sum_{i=T}^{nb} h_i}{\sum_{j=1}^{nb} h_j} \qquad (3)$$

$$T = [Similiairy\ threshold * n_b] \qquad (4)$$

n_b is number of bin.



```
Algorithm SHC
L ← Empty List {Cluster List}
for each document D do
    for each cluster C in L do
        $HR_{old} = HR_C$
        Simulate adding D to C
        $HR_{new} = HR_C$
        if ($HR_{new} \geq HR_{old}$) OR (($HR_{new} > HR_{min}$)
          AND ($HR_{old} - HR_{new} < \varepsilon$)) then
            Add D to C
        end if
    end for
    if D was not added to any cluster then
        Create a new cluster C
        ADD D to C
        ADD C to L
    end if
end for
```

Figure 4. Pseudocode Similarity-based Histogram Clustering

Cluster Importance

Cluster importance is a method of sorting clusters according to the number of weights of words in a cluster that often appears. A threshold (θ) is used to specify a word that appears frequently or not to all input documents. If the frequency of a word meets the threshold (θ) then the word has weight. In the SHC algorithm, there is no knowledge of the number of clusters that will be formed so cluster sequencing is needed. The determination of the selected clusters that will be the final summary candidates is very important to know [5]. The process of making clusters will continue to follow the algorithm in Figure 4. This study does not use parameters to determine the ideal number of clusters. All clusters formed will be sorted according to their weight and some of the best clusters will be selected. Cluster importance is calculated according to the following equation:

$$Weight(c_j) = \sum_{w \in c_j} \log(1 + count(w)) \qquad (5)$$

cj is the j-th cluster while count(w) is the number of word w in the whole documents and should be greater than threshold. This weight represents on how much information that a cluster has[6].

Selection of Representative Sentences

LDA method has been widely applied in various multi-document text summary studies. [6] and [8] have proposed LDA method for summarization which focused on selecting representative sentences.

In this study, the LDA method is proposed to give weight to sentences by capturing topics as keywords. LDA is used to increase representative sentence extraction and reduce incorrect sentences in multi-document summarization [6]. To calculate the weight of each sentence in each cluster using the following equation.

$$F_{LDA}(S_{kj}) = \sum_{Wi\ \varepsilon\ Sr} \frac{p(Wi\,|\,Tk) * p(Tk\,|\,Dm)}{length(Sr)} \qquad (6)$$

$F_{LDA}(S_{kj})$ is the value of the sentences based on the total weight of all words that make up k-th sentence s in j-th cluster. p(Wi | Tk) is the probability of the i-th word in the k-th topic while p(Tk | Dm) is the probability of the k-topic in the m-th document. length(Sr) is number of words in of *k*-th sentence *s* in *j*-th cluster.

RESULTS AND DISCUSSION

Based on the results of the research conducted, obtained an accurate ROUGE-1 score from 10 different topics and each topic has 7 different news sources that are displayed in Table 1.

Table 1. Detailed ROUGE-1 value for each test

	Recall	Precision	F-measure
Minimal Value	0.275	0.405	0.374
Avarage Value	0.396	0.629	0.481
Maximum Value	0.581	0.709	0.630

From the results of the accuracy in Table 1, the average ROUGE-1 multi-F produced is 48.1%. Summarizing using semantic clustering produces clusters that are good enough to form sentence clusters in multi-document summarizing [4]. The advantage of semantic clustering is being able to find semantic relations in creating sentence clusters. That is because of the use of the LSI method before measuring the sentence similarity value.

The accuracy calculation of this system is not optimal because in the process of selecting sentences using the LDA method it does not consider information that belongs to a sentence but only considers the keywords of the sentence. This is evidenced by the sentences in the expert summary, some clusters are formed, but only a few sentences are not selected as representative sentences from the cluster. Also, there is no correct or ideal summary, either the system summary or expert summary.

SUMMARY

Semantic clustering and selection of representative cluster sentences can be implemented in multi-document text summarizing. The formation of sentence clusters using semantic clustering is done by two methods, namely LSI and SHC. The advantage of Semantic Clustering is being able to find semantic relations in creating sentence clusters. That is because the use of the Latent Semantic Indexing method before measuring sentence similarity values.

From 70 test data consisting of 10 different topics and each topic containing 7 news sources, the summary results have an average value of ROUGE-1multi F-measure which is 0.481 or 48.1%. The calculation of the accuracy of this system is not optimal because in the process of selecting sentences using the Latent Dirichlet Allocation method does not consider information that belongs to a sentence but only considers the keywords of the sentence. This is evidenced by the sentences in the manual summary, there are several clusters that are formed, but only a few sentences are not selected as representative sentences from the cluster.

In the future study, we expect that the proposed method can choose representative sentences better and can obtain accuracy value higher.

REFERENCES

[1] Gupta, V. and G. S. Lehal, "A survey of text summarization extractive techniques." Journal of emerging technologies in web intelligence 2(3): 258-268, 2010.

[2] Alguliev, R. M., et al, "Multiple documents summarization based on evolutionary optimization algorithm." Expert Systems with Applications 40(5): 1675-1689, 2013.

[3] Ouyang, Y., et al, "Applying regression models to query-focused multi-document summarization." Information Processing & Management 47(2): 227-237. [4] "Optimizing K-Means by Fixing Initial Cluster Centers," vol. 4, no. 3, pp. 2101–2107, 2014, 2011.

[4] Santika, P. P. and G. N. Syaifuddin, "Semantic Clustering Dan Pemilihan Kalimat Representatif Untuk Peringkasan Multi Dokumen." Jurnal Teknologi Informasi dan Ilmu Komputer 1(2): 91-97, 2014.

[5] Sarkar, K, "Sentence clustering-based summarization of multiple text documents". TECHNIA–International Journal of Computing Science and Communication Technologies, 2(1), 325-335, 2009.

[6] Lukmana, I., et al, "Multi-Document Summarization Based On Sentence Clustering Improved Using Topic Words." JUTI: Jurnal Ilmiah Teknologi Informasi 12(2): 1-8, 2014.

[7] Christopher, G., & Yusliani, N. Rancang Bangun Sistem Peringkasan Teks Multi-Dokumen. Paper presented at the Annual Research Seminar (ARS), 2017.

[8] Hennig, l., et al, "Identifying Sentence-Level Semantic Content Units with Topic Models". In 2010 Workshops on Database and Expert Systems Applications, pages 59–63. IEEE, 2010.