TELKOMNIKA, Vol.16, No.2, April 2018, pp. 843~851
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v16i2.8431

Sentence Extraction Based on Sentence Distribution and Part of Speech Tagging for Multi-document Summarization

Agus Zainal Arifin*, Moch Zawaruddin Abdullah, Ahmad Wahyu Rosyadi, Desepta Isna Ulumi, Aminul Wahib, Rizka Wakhidatus Sholikah
Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia
[3]Sekolah Tinggi Teknik Qomaruddin, Gresik, 61152, Indonesia
*Corresponding author, e-mail: agusza@cs.its.ac.id

Abstract

Automatic multi-document summarization needs to find representative sentences not only by sentence distribution to select the most important sentence but also by how informative a term is in a sentence. Sentence distribution is suitable for obtaining important sentences by determining frequent and well-spread words in the corpus but ignores the grammatical information that indicates instructive content. The presence or absence of informative content in a sentence can be indicated by grammatical information which is carried by part of speech (POS) labels. In this paper, we propose a new sentence weighting method by incorporating sentence distribution and POS tagging for multi-document summarization. Similarity-based Histogram Clustering (SHC) is used to cluster sentences in the data set. Cluster ordering is based on cluster importance to determine the important clusters. Sentence extraction based on sentence distribution and POS tagging is introduced to extract the representative sentences from the ordered clusters. The results of the experiment on the Document Understanding Conferences (DUC) 2004 are compared with those of the Sentence Distribution Method. Our proposed method achieved better results with an increasing rate of 5.41% on ROUGE-1 and 0.62% on ROUGE-2.

Keywords*: multi-document summarization, sentence distribution, pos tagging*

1. Introduction

The number of documents is increasing rapidly daily. The information contained within multi-documents means that it is now a requirement to spend significant amounts of time finding a representative sentence from all documents. Such sentences are referred to as candidate important sentences. Sentences containing important words and which are well spread in the document are termed important sentences. Text summarization is a process for obtaining the shorter version in a text [1]. Multi-document summarization generates a representative summary or abstract of the entire document by reducing documents in size while retaining the main information of the original [2].

A sentence-weighting strategy needs to choose a representative sentence as a summary. Such a sentence should contain as much information as possible from the document [3] and arrange important words scattered in the document. If the word is spread throughout the document, it will have a higher value than words with low distribution. Thus, the method of weighting sentences must pay attention to the level of word distribution.

Various kinds of methods to solve the multi-document summarization problem have been proposed. One involves using hierarchical Latent Dirichlet Allocation (hLDA) in legal judgment clustering [4]. Another method uses the principle of vertex cover algorithms [5]. This method extracts a summary by selecting sentences that cover the input document's main concepts. It can generate a summary that covers the document's main concepts, required length, and minimum redundancy. Another method uses only frequent words in document summarization that will be used in document clustering [6]. Another approach involves optimizing diversity, coherence, and coverage among the summary sentences [7].

This approach uses a self-adaptive differential evolution (SaDE) algorithm to solve the optimization problem. Such an approach can produce a good summary with an appropriate level

Received October 5, 2017; Revised January 28, 2018; Accepted February 12, 2018

of readability. The Hypergraph-based Vertex-reinforced random walk framework is proposed for multi-document summarization by integrating multiple important factors for sentence ranking [8]. This method produces a summary that only considers the topic distribution relationship among sentences. Another method uses a statistical decision by calculating the weight of sentences based on word distribution (SDM) [9]. According to SDM, the important sentences that become candidates for summary sentences are obtained based on the distribution of important words. This method uses local and global sentence distribution for weighting sentences. Local sentence distribution can determine the importance of a sentence in the single cluster by assuming that the sentence that has more spread elements in a cluster is more important and has a higher position in that cluster. Global sentence distribution can determine the importance of a sentence in a set of clusters. The sentence that has more spread elements in its cluster but less scattered in another cluster is more important and has a higher global sentence distribution weight.

SDM focuses on statistical decisions to develop the weight of words, which means that they need to find important sentences by determining the frequency and spread of words in the corpus. Further, SDM ignores the grammatical information that indicates instructive content. Grammatical information carried by part of speech (POS) label can indicate to an extent the presence or absence of informative content in sentence and increase the quality of translation [10]-[11]. SDM will produce a good result if the frequent and spread word has informative content in the sentence. In contrast, it may inferior if the frequent and spread word is fewer content-bearing words. Sentence distribution and grammatical information carried by POS can be a great combination to find important sentence because it arranges the summary from sentences that have many frequent, spread, and most content-bearing words.

Summarization based on frequency and spread of words is not enough. Such as SDM which cannot distinguish the different function or meaning of a word. Any word that has various functions or meanings but the root form is same will be considered as a single word. It can be solved by grammatical information that is carried by a POS label. In fact, the content-bearing level of the word that can be obtained from such a POS label is also required. Both of these correspond to sentence distribution and grammatical information. Therefore, a weighting strategy that can integrate sentence distribution and grammatical information is needed.

In this paper, we propose a new sentence weighting method by incorporating sentence distribution and POS tagging for multi-document summarization. The proposed weighting method will integrate the power of frequency and content-bearing words for selecting important sentences. The proposed method can improve the quality of summary-containing sentences which have many frequent, well spread, and the most content-bearing words.

2. Research Method

In this study, the research method from Sarkar [12] was adopted and the framework by Lukmana [13] was also employed. There are 5 steps that are conducted to obtain final summaries: 1) text preprocessing, 2) sentence clustering for clustering sentences in the dataset, 3) cluster ordering to order the cluster by descending order, 4) sentence extraction for extracting the representative sentence from each cluster as shown in Figure 1 and Figure 5 arranging summaries from representative sentences.

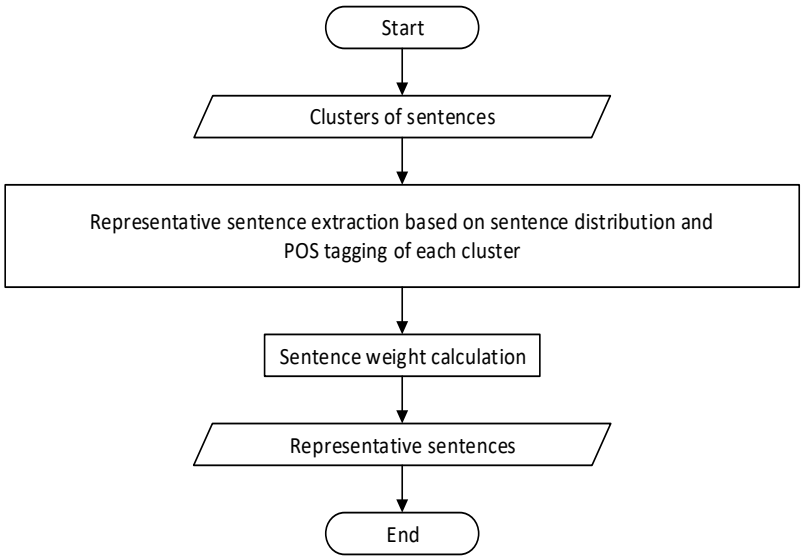

Figure 1. The steps involved in sentence extraction.

2.1. Text Preprocessing Phase

Text preprocessing involves ensuring that the text is more structured and compatible with the system. In this study, we use four steps in text preprocessing: 1) tokenizing is conducted to split the sentence into words so that each word can stand alone, 2) then, POS tagging is used to label each word with its POS label of a sentence, 3) after that, stopword removal is processed to remove the inappropriate keywords, such as prepositions, pronouns, and conjunctions, and then finally 4) in order to obtain the basic word of each word, a stemming process is conducted. In this study, Stanford of natural language processing is used in the tokenizing and POS tagging processes. A stoplist dictionary is used for stopword removal and an English porter stemmer library is used for stemming.

2.2. Sentence Clustering Phase

Each sub-topic or sentence in the document set should be identified properly by using similarity-based histogram clustering (SHC) [12]. This is used to cluster each sentence based on its similarity to another sentence. Uni-gram matching-based similarity is used to obtain the similarity between a candidate of cluster member sentence and each sentence in a cluster [9]. The similarity between two sentences is calculated based on the number of similar words between two sentences ($|s_i| \cap |s_j|$), which is divided by the number of words in the first sentence s_i and the second sentence s_j ($|s_i| + |s_j|$), as shown in Equation (1).

$$sim(s_i, s_j) = \frac{(2 * |s_i| \cap |s_j|)}{|s_i| + |s_j|} \quad (1)$$

The similarity value of each pair of sentences in a cluster is saved in a list of similarity values of a cluster $Sim=\{sim_1, sim_2, sim_3, \ldots, sim_m\}$. The length of *Sim* depends on the number of existing pairs of sentences m. The value of m is obtained by Equation (2).

$$m = n(n+1)/2, \quad (2)$$

where n is the number of sentences in a cluster.

A similarity histogram of a cluster can be noted as $H= \{h_1, h_2, h_3, \ldots, h_{nb}\}$. The value of h_i denotes the number of similar pairs of each sentence in a cluster with a similarity lower limit value in the bin sim_{li} and similarity upper limit value in the bin sim_{ui}, as shown in Equation (3). The histogram ratio of a cluster *HR* can be obtained by using Equations (4) and (5).

$$h_i = count(sim_j) \tag{3}$$

$$HR = \frac{\sum_{i=T}^{n_b} h_i}{\sum_{j=1}^{n_b} h_j} \tag{4}$$

$$T = \lfloor S_T * n_b \rfloor, \tag{5}$$

where S_T denotes similarity threshold, T is the number of bins corresponding to the similarity threshold, and n_b denotes the number of bins. A sentence will be a member of a cluster if it satisfies the characteristics of the cluster. However, a new cluster will be made from a sentence if the sentence does not satisfy the characteristics of the cluster [9].

2.3. Cluster Ordering Phase

The set of clusters that have been obtained from the previous phase has to be ordered, because the number of clusters is unknown. It is conducted to obtain the convenient cluster that will be used in making the summary process. The more important cluster will be in the top of ordered cluster list. The importance of a cluster can be determined based on the number of important or frequent words in it [14]. Each word in the cluster is tested based on the value of threshold *θ*. The word *w* is considered as a frequent word if its frequency *count (w)* meets the threshold *θ.*

The weight of the word *w* can be obtained by calculating the frequencies of all words in the input document. The weight of a cluster is calculated with respect to the number of frequent words in the cluster. The cluster importance method used in this study is based on [12]. The cluster importance weight of each cluster *weight (cj)* can be calculated using Equation (6). Cluster ordering is conducted by sorting the cluster based on the cluster importance weight with descending order.

$$weight(c_j) = \sum_{w \in c_j} \log(1 + count(w)) \tag{6}$$

2.4. Sentence Extraction Phase

Sentence extraction is used to determine the most important sentences from each ordered cluster. Those sentences will be the representative sentences from each cluster and will be used to form summaries. In this study, we use a new method to select the most important sentences using sentence distribution and POS tagging.

2.4.1. Part of Speech (POS)

POS has a function for natural language processing that can provide some information about a word (noun, verb) and the words around it (possessive pronoun, personal pronoun) [11] The presence of informative content can be indicated by POS's grammatical information [7]. Based on Jespersen's Rank Theory, POS can be ranked into four degrees: 1) nouns, because they have the most content-bearing labels, 2) adjectives, verbs and participles, 3) adverbs, and finally 4) all remaining POS [7]. In this study, the POS label of each word in a sentence will be combined with the word to be a term. The weight of local and global sentence distribution will be generated for each term.

2.4.2. Sentence Distribution Method

Weight of local $W_{ls}(s_{ik})$ and global sentence distribution $W_{gs}(s_{ik})$ are used to determine the weight of a sentence $Weight(s_{ik})$, as shown in Equation (7).

$$Weight(s_{ik}) = W_{ls}(s_{ik}) * W_{gs}(s_{ik}). \tag{7}$$

In this study, we combine both local and global sentence distribution with the weight of POS label. Both local and global sentence distribution of each term are calculated based on the POS label of the term. It is different from research [9] which only calculated both local and global sentence distribution of each term, without paying attention to the POS label.

Local sentence distribution can determine the importance of a sentence in a cluster. The sentence that has more spread elements in a cluster is more important and has a higher position in that cluster [9]. This method will give a high local sentence distribution weight to a sentence that has more widely spread elements.

Global sentence distribution can determine the importance of a sentence in a set of clusters. The sentence that has more spread elements in its cluster but less scattered in another cluster is more important and has a higher global sentence distribution weight [9]. The weight of local distribution will be multiplied by the weight of global distribution to obtain the weight of a sentence. This process will determine high sentence weight if both local and global distribution weight is high. However, the weight of a sentence will be low if one of them is low.

2.4.2.1. Local Sentence Distribution with POS Label

The weight of local sentence distribution is calculated for each term *j* in sentence *i.* The term *j* is a combination of the word with its POS label in sentence *i.* For example, if in a sentence there are three similar words with two different POS labels, the term will be two: the word with the first POS label and the word with the second POS label. There are 5 steps of calculation in local sentence distribution method: distribution opportunities, total distribution, distribution expansion, sentence component or term weight based on POS label, and local sentence distribution weight of a sentence.

Distribution opportunities r_{ij} of a term *j* is obtained by dividing the number of different term in sentence s_i of cluster *k* $|s_{ik}|_{dt}$ with the number of $|s_{ik}|_{dt}$ in cluster *k* $|c_k|$, as shown in Equation (8). Distribution of term *j* χ^2_{jk} in cluster *k* is obtained by calculating chi-square test statistics using Equation (9). To obtain the value of χ^2_{jk}, total of quadrate different between the frequency of term *j* in sentence *i* v_{ij} with the distribution frequency of term *j* in cluster *k* n_{jk} multiplied by r_{ij} divided by $n_{jk}r_{ij}$, where $|c_k|_{dt}$ is the amount of different terms in *k*.

$$r_{ij} = \frac{|s_{ik}|_{dt}}{|c_k|} \tag{8}$$

$$\chi^2_{jk} = \sum_{j=1}^{|c_k|_{dt}} \frac{(v_{ij} - n_{jk}\, r_{ij})^2}{n_{jk}\, r_{ij}} \tag{9}$$

A smaller value χ^2_{jk} of term *j* in sentence *i* is closer to the distribution maximum [15]. The spread rate of term *j* in *k* U_{jk} can be obtained from Equation (10).

$$U_{jk} = \frac{1}{1 + \chi^2_{jk}} \tag{10}$$

Then, expansion is optimally conducted to calculate the spread of term *j* in the cluster, as we can see in Equation (11).

$$St_{jk} = \log_2 (1 + \frac{p_{jk}}{P_k}), \tag{11}$$

where p_{jk} is number of sentences containing term *j* in cluster *k*, and P_k is the total number of sentences in cluster *k*. In this study, we consider the POS label of each term in a sentence. To do this, we make a list of POS label weight *Wp*. Equation (12) shows that there are four weight values in *Wp* because POS labels can be ranked into four degrees based on

Jespersen's Rank Theory [10]. The values in *Wp* list are determined by our experiment. The value of *Wp* that will be used in the calculation of weight depends on the POS label of term *j*. The POS label weight of term *j* has to be determined before the weight of term *j* in a cluster $Wt_{l,jk}$ is calculated. POS label weight to term *j* can be determined based on the POS label of term *j*. In this study, we propose a new method to combine sentence distribution weight and POS label weight. To calculate the weight of term *j* that is a combination of local sentence distribution and POS label weight in a cluster, we use Equation (13) inspired by [9].

$$Wp = \{0.72,0.59,0.41,0.25\} \quad (12)$$

$$Wt_{l,jk} = \log_2(1 + U_{jk} * St_{jk}) * Wp_{jk} \quad (13)$$

The local sentence distribution weight of sentence $W_{ls}(s_{ik})$ is obtained by summing all of local sentence distribution weight of sentence's term $Wt_{l,jk}$ and dividing it by the number of terms forming the sentence s_i $|s_{ik}|$, as shown in Equation (14).

$$W_{ls}(s_{ik}) = \frac{1}{|s_{ik}|} \sum_{Wt_{l,jk} \in s_{ik}} Wt_{l,jk} \quad (14)$$

2.4.2.2. Global Sentence Distribution with POS Label

The distribution of terms in sets of clusters can be defined as global sentence distribution. There are five calculation steps in global sentence distribution that are similar to the steps of local sentence distribution: distribution opportunities, total distribution, distribution expansion, sentence component or term weight based on POS label, and global sentence distribution weight of a sentence.

Similar to local sentence distribution, the weight of global sentence distribution is calculated to each term *j* in sentence *i.* The term *j* is a combination of a word with its POS label in sentence *i.* In global sentence distribution, the number of clusters *m* is needed because term *j* will be processed in the entire cluster. Distribution opportunities r_{jk} of a term *j* in cluster *k* is obtained by dividing the number of different terms in cluster *k* $|c_k|_{dt}$ with the number of different terms in the set of clusters $|c|_{dt}$. It is defined as

$$r_{jk} = \frac{|c_k|_{dt}}{|c|_{dt}}. \quad (15)$$

Distribution of term *j* χ_j^2 in the set of cluster *m,* is obtained by a total of quadrate different between the frequency of term *j* in *k* v_{jk} with the distribution frequency of term *j* in set of clusters n_j multiplied by r_{jk} divided by $n_j r_{jk}$, as shown in Equation (16).

$$\chi_j^2 = \sum_{j=1}^{|c|_{dt}} \frac{(v_{jk} - n_j r_{jk})^2}{n_j r_{jk}} \quad (16)$$

The spread rate of term *j* in the set of cluster U_j can be determined by using Equation (17).

$$U_j = 1 + \chi_j^2 \quad (17)$$

The spread of term *j* in the set of clusters optimally can be calculated using expansion, as can be seen in Equation (18).

$$St_j = \log_2\left(1 + \frac{P}{p_j}\right), \quad (18)$$

where p_j is the number of clusters containing term *j* and *P* is the number of clusters.

Similar to local sentence distribution, we also consider the POS label of each term in a sentence. To consider it, *Wp* is also used for calculating the weight of global sentence distribution. The POS label weight of term *j* has to be determined before the weight of term *j* in the set of clusters $W_{g,j}$ is calculated. To calculate the weight of term *j* that is a combination of global sentence distribution and POS label weight in the set of clusters, we use Equation (19) inspired by [9]. So, the weight of term *j* can be obtained by using Equation (19).

$$Wt_{g,j} = \log_2(1 + U_j * St_j) * Wp_j \quad (19)$$

The global sentence distribution weight of sentence $W_{gs}(s_{ik})$ is obtained by summing all of global sentence distribution weight of sentence's term $Wt_{g,j}$ and dividing it by the number of terms forming the sentence s_i $|s_{ik}|$, as shown in Equation (20).

$$W_{gs}(s_{ik}) = \frac{1}{|s_{ik}|} \sum_{Wt_{g,j} \in s_{ik}} Wt_{g,j} \quad (20)$$

2.5. Summary Arrangement Phase

The summary arrangement phase is conducted after the important sentences are obtained from the sentence extraction phase. The ordered cluster from the previous phase will be the reference for this phase. Sentences with higher sentence weight from the ordered cluster will be selected as cluster representative sentences. These representative sentences will be ordered based on the sequence of the ordered clusters. The selection of sentences is continuously performed until the length of the summary is fulfilled.

3. Results and Analysis

The experiments in this study were conducted using two important sentence methods for comparison, the sentence distribution method [9], and our proposed method. The data used in this study is the document understanding conference's (DUC) 2004 task 2, comprising 50 groups of documents. The evaluation of summary results uses ROUGE-1 and ROUGE-2, where the higher the value, the better the quality, as shown in [16].

3.1. POS Tagging weight

The weight of POS Tagging label was determined by a manual experiment using 568 sentences comprising 14,488 words. The experiment involves selecting every important word in each sentence, which led to 6,860 important words being chosen. The weight of POS Tagging can be calculated by using the total POS labels that are selected as important words IW_{pos} divided by the total POS labels that appear in experiment T_{pos}. This number can be defined as

$$W_{pos} = IW_{pos}/T_{pos}. \quad (21)$$

The detailed results of POS label weight can be seen in Table 1. This table shows that the POS label noun is the label with the highest weight.

Table 1. POS label weight.

Label	Important (IW_{pos})	Total Appearances (T_{pos})	Weight
Noun	3223	4461	0.72
Verb	1868	3148	0.59
Adverb	175	432	0.41
Other	1594	6447	0.25

3.2. Testing of Sentence Distribution and Part of Speech Tagging

Testing is used to determine the result of the proposed method compared with the Sentence Distribution Method (SDM). The parameters used in this testing process are based on those parameters which exist in the SDM. An example of a summary generated from the

proposed method can be seen in Table 2. A summary result from the proposed method contains sentences with more content-bearing words than SDM. This happens because of the addition of POS weight on the weighting process. The result of testing for the proposed method and SDM can be seen in Table 3.

Table 2. Summary Results Comparison for a Sample Document Topic

The Proposed Method	Sentence Distribution Method
The radical group Islamic Jihad claimed responsibility.	The radical group Islamic Jihad claimed responsibility.
Netanyahu's Cabinet delayed action on the new peace accord following Friday's suicide bombing at a Jerusalem market, and his remarks about building on Har Homa may be seen as a provocation by the Palestinians at a politically sensitive moment.	Israel's Cabinet announced within hours of a market bombing Friday that it will put off a vote indefinitely on whether to ratify the Wye River accord until Palestinians crack down further on terrorism.
Ramadan Abdallah Shallah, the Damascus-based leader of Islamic Holy War, said martyrs from his movement had carried out the Jerusalem attack in response to Israel's settlement policy and Judaization of the West Bank.	A Palestinian security official said several Islamic Holy War members were arrested in the West Bank on Friday night.
	'We have no knowledge in the movement about the operation that occurred in Jerusalem,'' said Nafez Azzam, a senior leader of the Islamic Holy War in Gaza.
Israel radio said the 18-member Cabinet debate on the Wye River accord would resume only after Yasser Arafat's Palestinian Authority fulfilled all of its commitments under the agreement, including arresting Islamic militants.	Palestinian security sources and the families of the dead bombers had already identified them as Islamic Jihad activists.
	Although Hamas initially claimed responsibility through anonymous phone calls to the police, all sides now have agreed that it was Islamic Jihad that carried it out.

Table 3. Summary Evaluation Results

Summary Method	ROUGE-1	ROUGE-2
Sentence Distribution Method	0.3899	0.1187
The Proposed Method	0.4110	0.1194

Table 4. Example of Terms With Multiple POS Label

Term	POS label	Example
Arrest	Verb	the spread of the disease can be **arrested**
	Noun	I have a warrant for your **arrest**
Call	Verb	She heard Alleria **calling** her
	Noun	She made a phone **call** to the office
Ceremony	Noun	the new Queen was proclaimed with due **ceremony**
	Adjective	**ceremonial** robes
	Adverb	**ceremonially**

The proposed method does not treat every same word equally because the same word does not necessarily have the same function, as shown in Table 4. The function of every word can be known by determining the POS label. Consequently, it treats every term based on the word and its POS label in a sentence. The POS label will affect the weight of every term. The weight of any terms that have labels other than noun will be reduced based on the level of POS label. However, the weight of any terms that have labels with respect to nouns will remain. By doing this, the proposed method can improve the quality of the summary result and describe the content from a multi-document well because it arranges the summary from sentences that are frequent, well spread, and have the most content-bearing words. In contrast, our proposed method cannot handle homonyms, which can lead to missed places regarding POS tagging labels for words.

3.3 Future Work

In the future, we propose to improve our proposed method in order to handle the homonyms in weighting sentences. Homonyms are words that have the same pronunciation and spelling but have different meanings. For example, the word 'rose' (a type of flower) and 'rose' (past tense of rise) are homonyms. Other examples, such as 'river bank', 'savings bank', and 'bank of switches' share a common spelling and pronunciation regarding word bank, but differ in meaning. Thus, they will produce different weights and different POS labels.

4. Conclusion

In this work, we explored a sentence weighting method for multi-document summarization which selects important sentences with sentence distribution and part of speech tagging. The proposed method was successful and shows a better summary than the sentence distribution method.

Sentence distribution with the POS tagging method gained an average score of 0.4110 for ROUGE-1 and 0.1194 for ROUGE-2, which is better than results obtained using the sentence distribution method. There is an increased value of 5.41% on ROUGE-1 and 0.62% on ROUGE-2. The increasing number is due to the calculation of weighting of words that have important labels of the sentence and the extent of their distribution. The results show that the proposed method can improve the weighting for summarization in multi-documents which use part of speech tagging.

References

[1] M Suman, T Maddu, M Mohan. An Integrated Approach for Compendium Generator using Customized Algorithms. 2015; 4(1): 6–12.
[2] J Atkinson, R Munoz. Rhetorics-based multi-document summarization. *Expert Syst. Appl.*, 2013; 40(11): 4346–4352.
[3] T He, F Li, W Shao, J Chen, L Ma. *A new feature-fusion sentence selecting strategy for query-focused multi-document summarization*. Proc. - ALPIT 2008, 7th Int. Conf. Adv. Lang. Process. Web Inf. Technol., 2008: 81–86.
[4] K Raghuveer. Legal Documents Clustering using Latent Dirichlet Allocation. *IAES International Journal of Artificial Intelligence (IJ-AI),* 2012; 2(1): 34–37.
[5] A John, M Wilscy. Vertex Cover Algorithm Based Multi-Document Summarization Using Information Content of Sentences. *Procedia - Procedia Comput. Sci.*, 2015; 46(Icict 2014): 285–291.
[6] PV Amoli, OS Sh. Scientific Documents ClusteringBased on Text Summarizatio. *International Journal of Electrical and Computer Engineering (IJECE).* 2015; 5(4): 782-787.
[7] K Umam, FW Putro, G Qorik, O Pratamasunu. Coverage, Diversity, and Coherence Optimization For Multi-Document Summarization. *J. Ilmu Komput. dan Inf. (Journal Comput. Sci. Information)*. 2015; 1(8): 1–10.
[8] S Xiong, D Ji. Query-focused multi-document summarization using hypergraph-based ranking. *Inf. Process. Manag.*, 2016; 52(4): 670–681.
[9] A Wahib, AZ Arifin, D Purwitasari. Improving Multi-Document Summary Method Based on Sentence Distribution. *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2016; 14(1): 286.
[10] C Lioma, R Blanco. Part of Speech Based Term Weighting for Information Retrieval. 2009: 412–423.
[11] H Sujaini, K Kuspriyanto, A Akhmad Arman, A Purwarianti. A Novel Part-of-Speech Set Developing Method for Statistical Machine Translation,' *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, 2014; 12(3): 581.
[12] K Sarkar. Sentence Clustering-based Summarization of Multiple Text Documents. *Tech. – Int. J. Comput. Sci. Commun. Technol.*, 2009; 2(1): 974–3375.
[13] I Lukmana, D Swanjaya, A Kurniawardhani, AZ Arifin, D Purwitasari. Sentence Clustering Improved Using Topic Words. 1–8.
[14] P Bhole, AJ Agrawal. Extractive Based Single Document Text Summarization Using Clustering Approach. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 2014; 3(2): 73–78.
[15] T Xia, Y Chai. An improvement to TF-IDF: Term distribution based term weight algorithm. *J. Softw.*, 2011; 6(3): 413–420.
[16] CY Lin. Rouge: A package for automatic evaluation of summaries,' *Proc. Work. text Summ. branches out (WAS 2004)*, 2004; (1): 25–26.