MULTI-DOCUMENT SUMMARIZATION BASED ON SENTENCE CLUSTERING IMPROVED USING TOPIC WORDS

Indra Lukmana[1], Daniel Swanjaya[2], Arrie Kurniawardhani[3], Agus Zainal Arifin[4], and Diana Purwitasari[5]

[1)] Department of Information System, Universitas Pesantren Darul Ulum
[2)] Department of Informatics Engineering, Universitas Nusantara PGRI Kediri
[3,4,5)] Department of Informatics Engineering, Institut Teknologi Sepuluh Nopember
e-mail: indra.lukmana.il@gmail.com[1)], swanjayadaniel@gmail.com[2)], emailku-wardhani@gmail.com[3)], agusza@cs.its.ac.id[4)], diana@if.its.ac.id[4)]

ABSTRAK

Informasi dalam bentuk teks berita telah menjadi salah satu komoditas yang paling penting dalam era informasi ini. Ada banyak berita yang dihasilkan sehari-hari, tetapi berita-berita ini sering memberikan konten kontekstual yang sama dengan narasi berbeda. Oleh karena itu, diperlukan metode untuk mengumpulkan informasi ini ke dalam ringkasan sederhana. Di antara sejumlah subtugas yang terlibat dalam peringkasan multi-dokumen termasuk ekstraksi kalimat, deteksi topik, ekstraksi kalimat representatif, dan kalimat rep-resentatif. Dalam tulisan ini, kami mengusulkan metode baru untuk merepresentasikan kalimat ber-dasarkan kata kunci dari topic teks menggunakan Latent Dirichlet Allocation (LDA). Metode ini terdiri dari tiga langkah dasar. Pertama, kami mengelompokkan kalimat di set dokumen menggunakan kesamaan histogram pengelompokan (SHC). Selanjutnya, peringkat cluster menggunakan klaster penting. Terakhir, kalimat perwakilan yang dipilih oleh topik diidentifikasi pada LDA. Metode yang diusulkan diuji pada dataset DUC2004. Hasil penelitian menunjukkan rata-rata 0,3419 dan 0,0766 untuk ROUGE-1 dan ROUGE-2, masing-masing. Selain itu, dari pembaca prespective, metode kami diusulkan menyajikan pengaturan yang koheren dan baik dalam memesan kalimat representatif, sehingga dapat mempermudah pemahaman bacaan dan mengurangi waktu yang dibutuhkan untuk membaca ringkasan.

***Kata Kunci**: Peringkasan Multidokumen, Latent Dirichlet Allocation, Clustering kalimat, similarity based histogram clustering, sentence information density.*

ABSTRACT

Information in the form of textual news has become one of the most important commodity in this information age. There are a lot of news produced every day, but often these news give same contextual content with different narratives. Therefore a method to collect these information into a simple summary is needed. Among a number of sub-tasks involved in multi-document summarization including sentence extraction, topic detection, representative sentences extraction, and sentence ordering. In this paper we propose a novel method for sentences ordering based on topic keyword using Latent Dirichlet Allocation (LDA). This method comprised of three basic steps. Firstly, we cluster the sentences in set of document using similarity histogram clustering (SHC). Next, we rank the cluster using cluster importance. Last, the representative sentences is selected by identified topics on LDA. The proposed method tested on DUC2004 dataset. The results show an average 0.3419 and 0.0766 for ROUGE-1 and ROUGE-2, respectively. Moreover, from reader perspective, our proposed method presents a good coherent arrangement in ordering representative sentences, so can improve their readability and reduce the time taken to read a summary.

***Keywords**: multi-document summarization, Latent Dirichlet Allocation, sentence clustering, similarity based histogram clustering, sentence information density.*

I. INTRODUCTION

TODAYS text based news are produced every day in large quantity. Many of these news have same contextual content with other news, and made information redundant. Because of this problem a system for collecting and summarizing these texts is needed.

Multi-document summarization has drawn much attention due to its applicability in real world applications. Multi-document summarization provide concise information from multiple documents in a shorter version while preserving its information content. In order to obtain the desired information.

Methods to arrange summary can be divided into extractive summarization and abstractive summarization [1] [2]. An extractive summarization chooses a subset of the sentences in the set of documents without any modification to the sentence and simply combining sentences together to form a summary. Whereas an abstractive summarization can be described as reading and understanding the text to recognize its content, then arranged in a concise text that employs words not appearing in the original document. An abstractive summarization can produce summaries that are more like what a human might generate but it requires deep natural language processing techniques [1]. Because of simple but robust method for text summarization, most of multi-document summarization focus on extractive summarization.

Extractive summarization methods can be divided into two categories e.g., query focused and generic multi-document summarization. Query focused methods, give summary that answers the given queries. Lin Zhao, et al [3] presented a study about multi-document summarization that used extractive summarization method based on query. They proposed a novel query expansion algorithm used in the graph-based ranking approach. You Ouyang, et al [1], presented a study about query-focused multi-document summarization using regression models for sentence ranking. The regression models that they used are Support Vector Regression (SVR). Ercan Canhasi, et al [4] study about query-focused multi-document summarization using a graph representation based on weighted archetypal analysis (wAA). wAA estimate the important sentences from documents set.

Generic methods contain overall information of the document's content. It can be either sentence based approaches or keyword-based approaches [5]. Sentence-based approaches focuses on partitioning documents in sentences and generating a summary that consists of the subset of most informative sentences. Whereas keyword-based approaches focus on detecting salient document keywords or topics. Rasim M. Alguliev, et al [6][7] present a study about generic document summarization using an optimization-based model. The model builds a summary by extracting salient sentences from document set and reduce redundancy in the summary. Elena Baralis, et al [5] study about multi-document summarization based on the Yago ontology. Yago based Summarizer relies on an ontology based evaluation to select the most representative sentences from document set. Suputra, et al [8] study about multi-document summarization using sentence information density and keyword of sentence cluster to select the most representative sentences from document set.

Topic identification has been one of identification method for describing the contextual content of a text documents. Latent topics can be used as correlation measures between documents [9]. One of the most commonly topic models used is Latent Dirichlet Allocation (LDA) [10].

LDA has been applied on many research in text processing including multi-document summarization. Some of multi-document summarization that use keyword-based approaches utilize a topic model to extract a summary. LDA has been applied successfully in multi-document summarization [7] [11] [12] [9]. In [7], the LDA used as a base for source code classification, [12] used as knowledge structure in searching and recommendation for enterprise social software, [12] used LDA as text segmenter, which construct a retrieval method for searching desired topics. [11] Used LDA as documents extraction. [9] has proposed a method for summarization which focused on selecting representative sentences. [13] Shown some redundancy of sentences when summarizing textual documents.

Inspired by the success of LDA, in order to enhance the extraction of representative sentences and

reduce improperly ordering sentences in multi-document summarization. In this paper we propose a novel method for sentences ordering based on topic keyword using LDA.

II. SUMMARIZATION USING SENTENCE BASED CLUSTERING

Suputra [8], has proposed a new feature called keywords of sentence cluster which combined with the sentence information density feature as a new strategy for selecting salient sentence in multi-document summarization based on a sentence clustering method.

First step, the documents from dataset were preprocessed using Porter Stemmer. Second step, each sentence in the set of document was clustered using similarity based histogram clustering. Third step, clusters that have been built were ordered using cluster importance. Fourth step, a representative sentence was selected from each cluster using sentence information density and keyword of sentence cluster. Fifth step, the summaries were arranged from those representative sentences obtained from each cluster. Those representative sentences were ordered according to the ordering of cluster in third step.

Similarity Based Histogram Clustering (SHC) is one of clustering method that can monitor and control the coherence of clusters. More coherent a cluster, higher the similarity of their component. SHC keep the coherence of each cluster in good cover age, in order to reduce redundant sentences. SHC monitors the coherence through the number of component that have high similarity value. Similarity measurement that used to measure the similarity between two sentences is uni-gram matching-based similarity measure using equation 1.

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

where s_i and s_j are *i*-th and *j*-th sentences, respectively. $|s_i + s_j|$ is number of words that match between *i*-th and *j*-th sentences. $|s_i|$ is the length of *i*-th sentences, namely the number of words arranging that sentence.

Moreover, similarity histogram that is represented cluster coherence is built. The distribution of histogram must be kept in the right side or similarity value must be equal to one. To examine the distribution of histogram, the similarity ratio was calculated. The similarity ratio of a cluster should be above the threshold. If n is number of sentences in a cluster, then number of sentence pairs in a cluster is $n(n + 1)=2$. $sim = sim_1; sim_2; sim_3; ::::; sim_m$ are set of similarity between two sentences, where $m = n(n+1)=2$. So, similarity histogram for each cluster is $h = h_1; h_2; h_3;$ ::::; hnb . nb is number of bin in histogram, and hi is number of sentence similarity. h_i is calculated using equation 2.

$$h = count(sim_j) \; \boldsymbol{for} \; sim_{li} \leq sim\_j \leq sim_{ui} \quad (2)$$

where sim_{li} is similarity lower limit in *i*-th bin, whereas sim_{ui} is similarity upper limit in *i*-th bin. Histogram Ratio (HR) for a cluster is calculated using 3.

$$HR = \frac{\sum_{i=T}^{n_b} h_i}{\sum_{j=1}^{n_b} h_j}, \quad (3)$$

with

$$T = |S_T \times n_b|, \quad (4)$$

where S_T is similarity threshold. Equation 4 is interpreted number of bin that appropriate with similarity threshold that is denoted by T.

Cluster importance is used to order the clusters formed using SHC. Cluster importance orders the cluster according to the sum of weight of frequent word in a cluster. A threshold is determined to

limit when the word can become a frequent word. The more important a cluster, the greater number of frequent word in a cluster. If there are *N* cluster namely $c = c_1; c_2; c_3; :::; c_N$. The ordering is done by calculating the weight of each cluster,

$$Weight(c_j) = \sum_{w \in c\_j} \log(1 + count(w)), \quad (5)$$

where c_j is the *j*-th cluster, *count(w)* is the number of word *w* in the whole documents and *count(w)* should be greater than threshold θ . This weight represents on how much information that a cluster has.

Sentence Information Density (SID) is built according to positional text graph approach. SID can represent how much information that a sentence has, thus it can represent any other sentences in a cluster. The similarity value of each sentence in a cluster that gets from equation 1 is arranged into a similarity matrix. Further, the similarity matrix is used to form a graph of sentences that represents position information. Graph is represented as *P = (V;E)*, where *P* represent graph, $V = s_1; s_2; :::; s_n$ is vertex in graph that represents sentences in a cluster, and $E = (s_i; s_j)$ is a set of an edge in graph where edge weight is calculated using equation 1. Graph *P* is formed according to sentences in a cluster. Firstly, the graph is empty, afterward all sentences in a cluster assign as vertex. Second step, similarity value is calculated for each pairs of sentences in *P*. If the similarity value of pairs of sentences satisfies the value of threshold then edge is formed, and its similarity value become to be its weight value. Third step, graph have been formed, SID is calculated by equation 6.

$$F_{sid}(S_{kj}) = \frac{W_{S_{kj}}}{\max l \in n\, W_{sij}} \quad (6)$$

where *n* is number of sentences in *j*-th cluster, W_{Skj} is the sum of all weight from all edge in *k*-th sentence *s* in *j*-th cluster, whereas W_{sij} is the sum of maximum edge weight among all similarity value in *j*-th cluster.

Keyword of Sentence Cluster (KCK) uses approach according to frequency of a word in a cluster and the distribution of that word in another cluster. KCK is used to assign a weight in a word. The higher the weight value of a word, the more representative that word as a keyword in a cluster. Weight of each word in a cluster is calculated by

$$tf.iscf_{w_{ij}} = tf_{w_i Cluster_j} \times iscf_{wij} \quad (7)$$

with

$$iscf_{wij} = \log \frac{N}{scf_{wij}} \quad (8)$$

where tf_{wi} Cluster_j is number of all *i*-th word in jth cluster, scf_{wij} is number of cluster that contain *i*-th word *w* in *j*-th cluster, and *N* is the number of all cluster.

The KCK itself is calculated by

$$F_{kck}(s_{kj}) = \frac{1}{len(s_{kj})} \sum_{w_{ij} \in S_{kj}} tf.iscf_{w_{ij}} \quad (9)$$

where $F_{kck}(s_{kj})$ is the value of sentences according to total weight of all word that arrange *k*-th sentence *s* in *j*-th cluster, *len*(*skj*) is a length of *k*-th sentence *s* in *j*-th cluster. The length of *k*-th

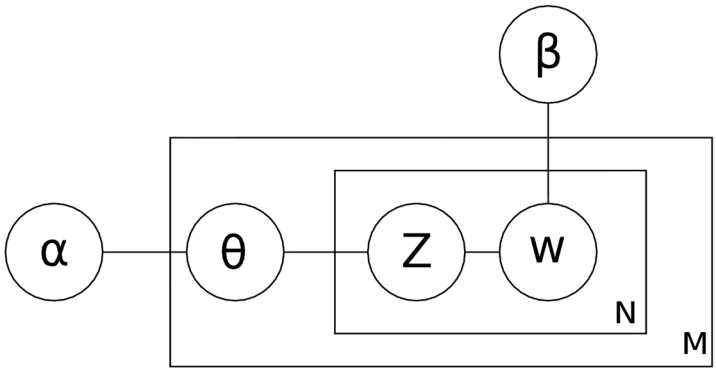

Fig. 1. LDA Diagram

sentences is the number of all word arranging that sentences.

After selected from each cluster, the representative sentences are arranged to build summaries. Summaries is arranged from the representative sentence that obtained from the most important cluster to the least important cluster. The representative sentences selecting is repeated until the expected length of summaries is satisfied.

III. Latent Dirichlet Allocation

LDA which firstly proposed by Blei, et al [10] is generative probabilistic topic modeling. The basic concept of LDA assumes that documents are viewed as a distribution over latent topics while each topic is a distribution over words. Fig. 1. represented the probabilistic graphical model of LDA. LDA modeling consist of three-level. The first level, second level, and third level consist of corpus-level parameter, document-level variable, and word-level variable respectively.

The corpus-level parameter includes parameters α and β. Parameter α is Dirichlet distribution that represented topic distribution in one document, while parameter β is Dirichlet distribution that represented word distribution in one topic. Those parameters are sampled once in the process of generating corpus. The document-level variable includes variable θ. Variable θ is represented of distribution of each topic in one document. That variable is sampled once per-document. Finally, the word-level variable includes variable z and w. Variable z is represented various topics in corpus, while variable w is represented various word in corpus. Those variables are sampled once for each word in each document.

IV. Summarization using topic words

In this study we proposed another method for keyword selection in [8] Using LDA, we weight the sentences by capturing the topic as keyword the proposed method as shown in Fig. 2.

The first step of this method is assigning each word in a set of document to be a part of a topic randomly. Then, the former topic was calculated again to obtain the better topic by

$$P(D) = \frac{D_T}{W_D} \tag{10}$$

$$P(T) = \frac{W_T}{N_T} \tag{11}$$

where $P(D)$ is the probability of a document, D_T is the number of document which have topic T, $W(D)$ is the number of words in document; with $P(T)$ is the probability of topics, W_T is the number of words in topic T, and N_T is the number of topic. Then for the probability of a word is the identified

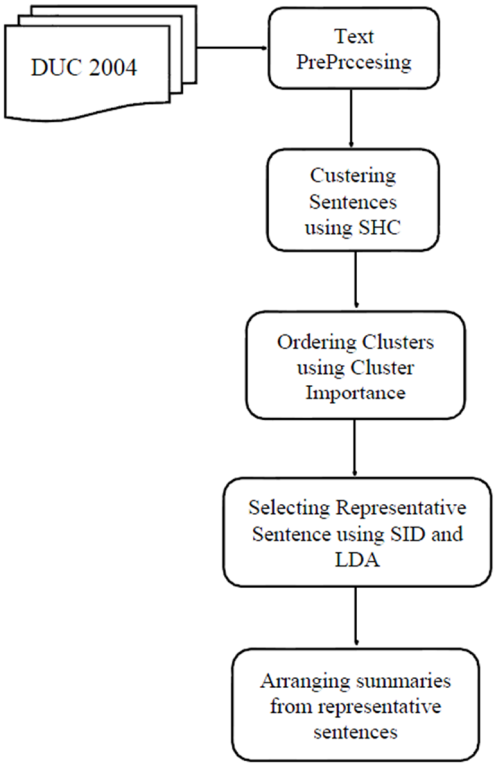

Fig. 2. Summarization using Topic Words

topic is computed by

$$P_w = P(D) \times P(T) \tag{12}$$

Each word in a topic is reset according to *P(T)*.

Those methods is repeated until reaches a steady state or until a certain threshold. Weight sentences using LDA is calculated by

$$F_{LDA}(s_{kj}) = \frac{1}{len(s_{kj})} \sum_{w_{ij} \in S_{kj}} P_{wij} \tag{13}$$

where $F_{LDA}(s_{kj})$ is the value of sentences according to the total weight of all word that arrange *k*-th sentence *s* in *j*-th cluster, *len(skj)* is a length of *k*-th sentence *s* in *j*-th cluster. The length of *k*-th sentences is the number of all weight of word composing that sentences. P_{wij} is probability value of *i*-th word in *j*-th cluster.

V. RESULTS AND DISCUSSION

We use the Document Understanding Conference (DUC) 2004 data sets to test the proposed method empirically. DUC2004 is open benchmark data sets for automatic summarization evaluation. It is English-written news articles which range over different subjects. Data set consists of 50 clusters according to their subject. Each cluster contains approximately 10 documents. For each cluster, there are four human summaries provided as references for evaluation. The length of human summaries approximately 250 words.

TABLE I
SUMMARY EVALUATION RESULTS

Run	R-1.Min	R-1.Max	R-1.Avg	R-2.Min	R-2.Max	R-2.Avg
KCK	0.2933	0.3792	0.3362	0.0434	0.1094	0.0753
LDA	0.2955	0.3888	0.3419	0.0443	0.1115	0.0766

TABLE II
SUMMARY COMPARISON

KCK	LDA
(1) Cambodian leader Hun Sen on Friday rejected opposition parties' demands (2) Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen's party to form a new government failed. (3) Cambodia's two-party opposition asked the Asian Development Bank Monday. (4) Ranariddh and Sam Rainsy renewed their international lobbying campaign against the old government Monday in a letter to ADB President Mitsuo Sato calling for the bank to stop lending money to it. "We respectfully advise the Asian Development Bank not to provide any new loans to the current regime in Cambodia," the two party leaders wrote.	(1) Cambodian leader Hun Sen on Friday rejected opposition parties' demands for talks outside the country, accusing them of trying to "internationalize" the political crisis. (2) Government and opposition parties have asked King Norodom Sihanouk to host a summit meeting after a series of post-election negotiations between the two opposition groups and Hun Sen's party to form a new government failed. (3) Opposition leaders Prince Norodom Ranariddh and Sam Rainsy, citing Hun Sen's threats to arrest opposition figures after two alleged attempts on his life, said they could not negotiate freely in Cambodia and called for talks at Sihanouk's residence in Beijing. (4) In it, the king called on the three parties to make compromises to end the stalemate: "Papa would like to ask all three parties to take responsibility before the nation and the people.

For evaluating the summary produced by the method, we used ROUGE measurements. ROUGE is the most commonly used metric of content selection quality, and the scores which it produces are highly correlated with manual evaluation scores for generic multi-document summarization of text. ROUGE compute the n-gram overlap between a generated summary and a set of models (expert made summaries). It is a recall-oriented measurements, which is suitable for summarization evaluation with variations in human content selections.

The results show that our proposed method does not fall back on result with the method proposed by [8]. Because of news text documents natures, the first sentence of paragraphs usually contain the information which represent the contextual content of the news, and [8] has done a fine result. This implicated in the evaluation results of our study, without our improvements the average ROUGE-1 shows 0.3362 and average ROUGE-2 shows 0.0753, when with our improvements the average ROUGE-1 shows 0.3419 and average ROUGE-2 shows 0.0766.

The average results show overall improvements albeit small as shown in Table I. Beside quantitative results there is another view point which we discovered, that is the summary results of our proposed method have a better sense of information as shown in comparison on Table II.

VI. CONCLUSION

By modifying the summaries with identified topic from LDA the results show a slight improvement evaluation metrics. We hypothesis this is because of the nature of news documents. For future study there are possibilities on using other methods for identifying important sentence for clustering, hence we recommend so.

REFERENCES

[1] You Ouyang,Wenjie Li, Sujian Li, and Qin Lu. Applying regression models to query-focused multi-document summarization. Information Processing & Management, 47(2):227– 237, 2011.

[2] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Nijat R. Isazade. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications, 40(5):1675 – 1689, 2013.

[3] Lin Zhao, Lide Wu, and Xuanjing Huang. Using query expansion in graph-based approach for query-focused multi-document summarization. Information Processing & Management, 45(1):35–41, January 2009.

[4] Ercan Canhasi and Igor Kononenko. Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Systems with Applications, 41(2):535–543, 2014.

[5] Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. Multidocument summarization based on the Yago ontology. Expert Systems with Applications, 40(17):6976–6984, December 2013.

[6] R.M. Alguliev and R.M. Aliguliyev. Effective Summarization Method of Text Documents. In The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05), pages 264–271. IEEE.

[7] Rachit Arora and Balaraman Ravindran. Latent Dirichlet Allocation and Singular Value Decomposition based Multi-Document Summarization, November 2008.

[8] I Putu Gede Hendra Suputra, Agus Zainal Arifin, and Anny Yuniarti. Pendekatan positional text graph untuk pemilihan kalimat representative cluster pada peringkasan multi dokumen.

[9] Leonhard Hennig, Thomas Strecker, Sascha Narr, Ernesto William De Luca, and Sahin Albayrak. Identifying Sentence-Level Semantic Content Units with Topic Models. In 2010 Workshops on Database and Expert Systems Applications, pages 59–63. IEEE, August 2010.

[10] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. Jour nal of Machine Learning Research, 3(4-5):993–1022, 2003.

[11] Hongling Wang and Guodong Zhou. Topic- Driven Multi-document Summarization. In 2010 International Conference on Asian Language Processing, pages 195–198. IEEE, December 2010.

[12] Dongmei Zhang, Jun Ma, Xiaofei Niu, Shuai Gao, and Ling Song. Multi-document summarization of product reviews. In 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, pages 1309–1314. IEEE, May 2012.

[13] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive LDA model selection. Neurocomputing, 72(7-9):1775–1781, March 2009.