TELKOMNIKA, Vol.14, No.1, March 2016, pp. 286~293
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i1.2330

Improving Multi-Document Summary Method Based on Sentence Distribution

Aminul Wahib*[1], Agus Zainal Arifin[2], Diana Purwitasari[3]
[1]Study Program of Informatics, Politeknik Kota Malang, Indonesia, 65132
[2,3]Department of Informatics, Faculty of Information Technology, ITS Surabaya, Indonesia, 60111
*Corresponding author, email: wahib@poltekom.ac.id[1], agusza@cs.its.ac.id[2], diana@if.its.ac.id[3]

Abstract

Automatic multi-document summaries had been developed by researchers. The method used to select sentences from the source document would determine the quality of the summary result. One of the most popular methods used in weighting sentences was by calculating the frequency of occurrence of words forming the sentences. However, choosing sentences with that method could lead to a chosen sentence which didn't represent the content of the source document optimally. This was because the weighting of sentences was only measured by using the number of occurrences of words. This study proposed a new strategy of weighting sentences based on sentences distribution to choose the most important sentences which paid much attention to the elements of sentences that were formed as a distribution of words. This method of sentence distribution enables the extraction of an important sentence in multi-document summarization which served as a strategy to improve the quality of sentence summaries. In that respect were three concepts used in this study: (1) clustering sentences with similarity based histogram clustering, (2) ordering cluster by cluster importance and (3) selection of important sentence by sentence distribution. Results of experiments showed that the proposed method had a better performance when compared with SIDeKiCK and LIGI methods. Results of ROUGE-1 showed the proposed method increasing 3% compared with the SIDeKiCK method and increasing 5.1% compared with LIGI method. Results of ROUGE-2 proposed method increase 13.7% compared with the SIDeKiCK and increase 14.4% compared with LIGI method.

***Keywords**: Multi-document summaries, Extracting important sentences, Sentence distribution*

1. Introduction

The number of digital documents has increased very rapidly, so that raises many new problems in digging and obtains information quickly and accurately. A growing number of documents led to the information diggers must spend extra time in searching and reading information. Another issue that arises is the magnitude of the potential loss of important information contained in the document. The researchers tried to resolve this problem by developing a method in document summaries.

Good document summary is a summary of which is capable of covering (coverage) as much as possible the important concepts (saliency) that exist in the source document [1] and saliency are a major problem in the document summary, the strategy to select a sentence is very important because it should be able to choose the main phrases and avoid redundancy so as to include many of the concepts [2]. Some studies [2-5] have developed a method of selecting an important sentence to address the issue of coverage and saliency.

One of good method is the method combination of sentence information density and keyword of sentence clusters (Sidekick) [2]. According [2] important sentence is a sentence that has the information density of the sentence and has many keywords of sentence clusters. Sentence information density can be extracted with an approach positional text graph and keyword of sentence clusters can be extracted using TF.IDF methods [2]. But the approach of the positional text graph on the condition there are some sentences with almost the same weight, it is difficult to determine the important sentences [6]. While the method of keyword of sentence clusters obtained with the TF.IDF concept not able to give maximum weight on cluster keywords [7].

Received July 14, 2015; Revised November 23, 2015; Accepted December 20, 2015

This study proposes a method of sentence weighting as a new strategy in selecting important sentences based on the sentence distribution method. The sentence distribution method will take into wherever the location of the elements forming of sentences, so as to give a higher weight to the sentences that should be the topic of the document. A selection of important sentence with this scheme is expected to select a representative sentence in multi-document summary and can improve quality of the summary result.

2. Research Method

The research method used in this study was adopted from the researcher [3] this framework was also employed by the researcher [2]. Figure 1 shows the phases which should be done to obtain final summaries. Those phases are text preprocessing, sentence clustering, cluster ordering, sentence extraction and the last is summarizing arrangements. Sentence extraction phase is the contribution of this research.

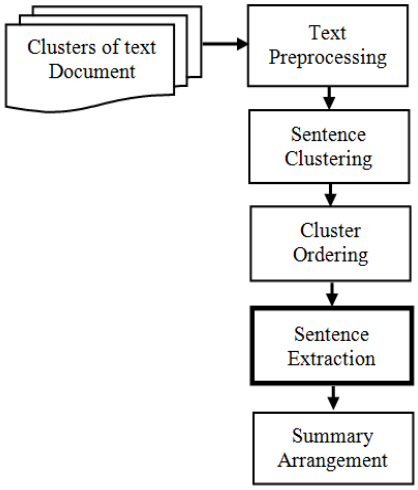

Figure 1. The framework of multi-document summary

2.1. Text Processing Phase

Text preprocessing phase includes the process of tokenizing, stopword and stemming. Tokenizing is a proses of beheading words so each word can stand alone. Stopword is the process of removing the key words which are not appropriate to be used, such as conjunctions, prepositions and pronouns. Stemming is the process of obtaining basic word of each word. In this study tokenizing is done by using Stanford of natural language processing, stopword removal is using stoplist dictionary, and stemming is using library English porter stemmer.

2.2. Sentence Clustering Phase with Similarity based Histogram Clustering (SHC)

Sentence clustering is an important part in a system of the automatic summary for each topic in the set of documents should be properly identified to find a similarity and dissimilarity in the document so as to ensure good coverage [3, 8].

The function of similarity which is used is uni-gram matching-based similarity based on the equation (1). Similarity between sentences is calculated based on corresponding words between the words *s* to-*i* and the words *s* to-*j* ($|s_i| \cap |s_j|$) is divided by the total length of the words *s* to-*i* and s to-*j* ($|s_i| + |s_j|$) :

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

The method of uni-gram matching-based similarity measure is a method used to measure similarity of each pair sentences in a cluster. If a cluster has *n* the number of a sentence, so the total number of the existing pair sentence is *m* where *m=n(n+1)/2* and *Sim={sim₁, sim₂, sim₃, …, simₘ}* is the collection of pair similarity between sentence with the total *m*. Similarity histogram from a cluster is noted with *H= {h₁, h₂, h₃, ..., hₙb}*. The function to calculate h_i is shown in the equation (2):

$$h_i = count\ (sim_j) \tag{2}$$

Amount similarity of pairs of each sentence in the bin to-*i* h_i in particular cluster is obtained by summing the similarity pairs of each sentence in the bin to-*i* (sim_j) in that cluster with a lower limit value of similarity in the bin to-*i* (sim_{li}) and the upper limit value of the similarity bin to-*i* (sim_{ui}). Histogram ratio (HR) from a cluster can be calculated by the equation of (3) and (4):

$$HR = \frac{\sum_{i=T}^{n_b} h_i}{\sum_{j=1}^{n_b} h_j} \tag{3}$$

$$T = \lfloor S_T * n_b \rfloor \tag{4}$$

Histogram ratio is a cluster which can be calculated by counting all of the number pairs of similarity in sentence (h_i) slipped away from threshold (S_T) is divided with the total number of pairs similarity sentences in all of the bin n_b. A sentence can be in a cluster if that sentence meets the criteria of the cluster. However, if the sentence does not meet the criteria in all of existing cluster, a new cluster will be formed. The SHC method for clustering sentences used in this study was adapted from study [3].

2.3. Cluster Ordering Phase

One of the weaknesses in the clustering sentence phase, is the similarity based histogram clustering (SHC) which are unknown total clusters that will be formed. Therefore, cluster ordering can be used as solution to determine the appropriate cluster to be a part of the process in making a summary. That is by testing every word which is available in the cluster based on threshold value *θ*. If the frequency of words *w* (*count* (*w*)) fulfills threshold *θ,* therefore those words are considered as frequent words.

The weight of the word *w* is calculated based on the frequencies of all of the words in the input document. Calculation of the cluster weight refers to all of the frequencies of the word *w* which is owned by particular cluster. Cluster ordering phase method used in this study was adopted cluster importance method which was suggested by in the study [3]. Cluster ordering based on the weight of cluster importance can be calculated by equation (5):

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

The weight of a cluster to-*j* (*Weight* (c_j)) can be calculated by summing all of the frequencies of the word *w* from the document input which is found in the cluster to-*j*.

2.4. Sentence Extraction phase

Sentence extraction is a phase selection of important sentence for forming summaries. To select important sentence, this study proposes a new strategy of weighting sentences using a distribution of local and global sentence methods. This strategy is called a sentence distribution method.

2.4.1. Sentence Distribution Method

Sentence distribution method is formed from the distribution of local and global sentence which can be seen in equation (6):

$$Weight_{(s_{ik})} = W_{ls}(s_{ik}) \times W_{gs}(s_{ik}), \quad (6)$$

Weights sentence ($Weight_{(s_{ik})}$) is obtained from multiplicative the weight distribution of local sentence ($W_{ls}(s_{ik})$) with the weight distribution of global sentence ($W_{gs}(s_{ik})$).

Distribution of local sentence is used to determine the position of each sentence in a cluster, by assuming that the sentence which has spreader elements in a cluster will have a higher position in that cluster. This method is expected to select the most representative sentences which are able to represent the cluster.

Distribution of global sentence is used to give a position in each sentence in a cluster by assuming that the sentences which have spreader elements will have a higher position. This method is expected to determine the level of interest or position of each sentence globally in the cluster.

The weight distribution of local and global sentences will be multiplied with each other so that the weight of both the local and global sentence can be mutually reinforcing. If the local position of a sentence has a great weight, on the other hand the global position of a sentence has a little weight, so by using this multiplication it will decrease the weight of the sentence. As apposite, if the weight of local sentence and global sentence are both higher, so that sentence deserves to represent a cluster than another sentence which has a lower value.

2.4.1.1. Local Sentence Distribution Method

Local sentence distribution, is a distribution of important words which is forming sentences in a cluster. Local sentence method is formed through the process of: (1) calculating probability of distribution, (2) calculating the total distribution, (3) calculating the expansion of distribution, (3) calculating weight of sentence components, and (4) calculating weight based on local sentence method.

Examples of clusters which contains *i* sentences and *j* words, are then S_{ik} represented as a sentence to-*i* in cluster to-*k*. Of which every word to-*j* in the set of sentences, are part of the sentence if those words has an equal distribution of words, thus the chances of a sentence to-*i* is calculated by using the theory of K. Pearson as equation (7):

$$r_{ij} = \frac{|\, s_{ik} \,|_{dt}}{|\, c_k \,|} \quad (7)$$

Distribution opportunities (r_{ij}) are obtained from difference of total forming sentences *s* to-*i* in the cluster to-*k* ($|\, s_{ik} \,|_{dt}$) which is divided by the number ($|\, s_{ik} \,|_{dt}$) in the cluster to-*k* ($|\, c_k \,|$). The amount of the difference between frequencies of words with frequencies of words distribution to-*j* in a sentence to-*i* can be calculated by using chi-square test statistics. Thus, the distribution of words to-*j* in a cluster to-*k* are the same with equation (8):

$$\chi^2_{jk} = \sum_{j=1}^{|c_k|_{dt}} \frac{(v_{ij} - n_{jk} r_{ij})^2}{n_{jk} r_{ij}} \quad (8)$$

Distribution of sentence component (χ^2_{jk}) is derived from total of different quadrate between the frequencies of the sentence components (v_{ij}) with frequencies of the distribution sentence component to-*j* in cluster to-*k* ($n_{jk} r_{ij}$) is divided by the frequencies of the distribution of sentence component to-*j* in the cluster to-*k*. Variabel n_{jk} is the frequency of sentence component to-*i* in cluster to-*k* and $|c_k|_{dt}$ is the number of different words in cluster to-*k*.

Smaller value χ^2_{jk} in equation (8) shows that component sentence to-*j* is closer to the maximum of the distribution where the value is contradicted with the weighting and distributing word which has positive correlation with non linear [7]. Therefore, equation (9) can be obtained from:

$$U_{jk} = \frac{1}{1+\chi^2_{jk}} \quad (9)$$

Weighting of component sentence to-*j* in a cluster to-*k* is spread (U_{jk}) and capsized with distribution of sentence component (χ^2_{jk}). In order to obtain calculation the weighting of sentence distribution to-*j* in cluster to-*k* optimally carried out expansion of the calculation in order to obtain equation (10):

$$St_{jk} = \log_2(1+\frac{p_{jk}}{P_k}) \quad (10)$$

The expansion of distributing component sentence to-*k* in cluster to-*k* (St_{jk}) is obtained from a number of sentences which contains the word to-*j* in cluster to-*k* (p_{jk}) with the number of sentences from the entire sentences in cluster to-*k* (P_k). So, the weighting of the component sentences to-*j* in a cluster to-*k* can be calculated by an equation (11):

$$Wt_{l,jk} = \log_2(1+U_{jk} * St_{jk}) \quad (11)$$

The weighting of local sentence components to-*j* in cluster to-*k* ($Wt_{l,jk}$) will form the weighting of local sentence to-*i* by summing all of the components forming sentences *s* to-*i* in cluster to-*k* ($W_{ls}(s_{ik})$) which is divided by the number of components forming sentences *s* to-*i* in a cluster to-*k* ($|s_{ik}|$), shown in equation (12):

$$W_{ls}(s_{ik}) = \frac{1}{|s_{ik}|}\sum_{Wt_{l,jk}\in s_{ik}} Wt_{l,jk} \quad (12)$$

Equations (7) to (11) are adopted from [7] which is originally used to calculate the distribution of words in a paragraph of the document, in this study are developed for weighting sentence on sentence cluster.

2.4.1.2. Global Sentence Distribution Method

Global sentence distribution is a distribution of important components forming sentences in sets of clusters. Global sentence method is formed in a similar manner to that of local sentence method, that is: (1) calculating probability of distribution, (2) calculating the total distribution, (3) calculating the expansion of distribution, (4) calculating the weight of component sentences, and (5) calculating the weight of local sentence.

Example the set of cluster which contains *m* cluster and sequences of a cluster is in *k* cluster, where k= (1,2,3..m). Hence the total of difference of words in a cluster to-*k* is given by ($|c_k|_{dt}'$) and total difference of words in the sets of clusters is given by ($|c|_{dt}'$). Therefore, the chances of component sentence to-*j* in a cluster to-*k* has a chance to spread as in the equation (13):

$$r_{jk}' = \frac{|c_k|_{dt}'}{|c|_{dt}'} \quad (13)$$

Variabel n_j is the frequency of component sentence to-*i* in a collection cluster. The total differences of quadrate between the frequencies of words ($v_{jk}{}'$), with the frequencies of distribution words to-*j* in a cluster to-*k* ($n_j{}' r_{jk}{}'$) is divided by the frequencies of word distribution to-*j* in that cluster to-*k* is used to calculate the distribution of component sentence to-*j* in the set of cluster so that it can be obtained the equation (14):

$$\chi_j^{2'} = \sum_{j=1}^{|c|_{dt}{}'} \frac{(v_{jk}{}' - n_j{}' r_{jk}{}')^2}{n_j{}' r_{jk}{}'} \quad (14)$$

The equation (14) shows the smaller value of the sentence component to-*j* that is spread ($\chi_j^{2'}$). Hence the sentence component to-*j* is closer to the maximum of distributions. This value is in line with the relations of weighting and distributing words which have correlation negative non linear [7] so the equation (15) can be derived:

$$U_j{}' = 1 + \chi_j^{2'} \quad (15)$$

The equation (15) shows the weighting of the components of sentence to-*j* is spread ($U_j{}'$) straightly compared with the distributing from components of a sentence (χ_{jk}^2). To get calculation of the weighting of components of sentence to-*j* which is spread optimally done by the expansion of calculation to get the equation (16):

$$St_j{}' = \log_2 (1 + \frac{P'}{p'_j}) \quad (16)$$

The expansion of distributing sentence component to-*j* (St_j) is obtained from the number of sentences which contains word to-*j* (p'_j) with total sentences in set of cluster (P'). So that weighting from component of sentence to-*j* ($Wt_{g,j}$) in the set of cluster can be calculated with the equation (17):

$$Wt_{g,j} = \log_2 (1 + U_j{}' * St_j{}') \quad (17)$$

Equation (18) shows the weighting global sentence ($W_{gs}(s_{ik})$) is derived by summing all of the weighting of global word of forming sentences to-*i* in a cluster to-*k* by considering the length of the sentence or the number of component forming sentences *s* to-*i* in a cluster to-*k*. This can be done by avoiding sentences which have the biggest number of the sentences components which will always appear as an important sentence without considering meaning that contains in the sentence.

$$W_{gs}(s_{ik}) = \frac{1}{|s_{ik}|} \sum_{Wt_{g,j} \in s_{ik}} Wt_{g,j} \quad (18)$$

Equation (13) to (17) are adopted of the study [7] which are originally used to calculate the distribution of words in a paragraph of the document. They are therefore used in this study to develop the weighting of sentence in sentence cluster. Equation (6), (12) and (18) are our original contribution in this study.

2.5. Summary Arrangement Phase

The summary arrangement phase is a step of summary arrangement which is obtained from the extracting important sentence phase. The clusters which are in cluster ordering phase will be a reference in summary arrangement. A sentence that has the higher weight in each cluster will be the main point of the expected summary. The number and order of summary sentence equal to the number and sequence of cluster sentence.

2.6. Evaluation of Summary Results

The evaluation of the summary result employed in this study is called ROUGE. ROUGE measure the quality of the result from summary by calculating overlap units such as N-gram, ordering words, pairs of words between the candidate of summary and summary as reference. ROUGE is effectively employed to evaluate the summary of the document [9].

ROUGE used in this study is the ROUGE-1 and ROUGE-2. ROUGE-1 is a measurement by using anagram matching concept where this measurement is calculated based on the total number of each word (unigram) which is appropriate between the result of the summary system with a summary of references made manually by the experts. ROUGE-2 is calculated based on the total of each two pairs (bigram) which is appropriate between the result of the summary system with the reference summary made by the experts, where the best of the maximum score in the ROUGE 1 and ROUGE 2 in better condition is 1.

3. Results and Analysis

Experiment in this study was done by using three of extracting important sentences methods, there are sentence distribution method, SIDeKiCK method [2] and the last is LIGI method [3]. The data used in this study is DUC (document understanding conferences) 2004 task 2 which consists of 50 groups of documents. The result of the evaluations used in this study was ROUGE-1 and ROUGE-2 where the higher value of ROUGE shows the better quality of the result summary obtained.

3.1. Testing of Sentence Distribution Method

Testing was used to know the result of proposed method compared to LIGI (local importance and global importance) method and SIDeKiCK (*sentence information density and keyword the cluster of sentence)* method.

Parameter used in the testing process is a parameter with a combination score of HR_{min}=0.7, ε=0,3 , similarity threshold (S_T)=0,4 and parameter θ=10. Parameter β the LIGI method which has been established in the study [3] was 0,5. The score α and λ used in SIDeKiCK method was α=0,4 and λ=0,2 [2] where those scores consider as the optimal of recommended score.

The result of testing sentence distribution by using LIGI method and SIDeKiCK method can be seen in Table 1. Table 1 show that the sentence distribution method has a higher average score of ROUGE compared to LIGI method and SIDeKiCK method both in the testing ROUGE-1 and ROUGE-2.

Average scores gained in testing ROUGE-1 using sentence distribution was 0,4042, in SIDeKiCK method gained the average score 0,3924 and LIGI method gained average scores 0,3845. It means that the sentence distribution method is better, or there was 3% of improvement compared to SIDeKiCK method and there was an increasing 5,1% compared to LIGI method in the testing schema with ROUGE-1. Testing in sentence distribution using ROUGE-2 gained average scores 0,1209, SIDeKiCK method gained 0,1063 and LIGI method gained 0,1057. It means that sentence distribution method was better or increased 13,7% compared to SIDeKiCK method and increased 14,4% compared to LIGI method.

Table 1. Testing of Extracting Important Sentence Method

Summary Method	ROUGE-1	ROUGE-2
Sentence Clustering (SHC) + Cluster Ordering + Sentence Distribution	0,404	0,1209
Clustering kalimat (SHC) + Cluster Ordering + SIDeKiCK	0,392	0,1063
Clustering kalimat (SHC) + Cluster Ordering + LIGI	0,384	0,1057

4. Conclusion

To improve quality of summaries result in multi-document summarization can be done by extracting important sentence using the sentence distribution method. Sentence distribution method has proven better compared to SIDeKiCK and LIGI methods with the average achieved in ROUGE-1 is 0,404 and ROUGE-2 is 0,121.

Selecting important sentences using sentence distribution method parameter optimally are *harming*=0.7, *epsilon* (ε)=0,3, similarity threshold (S_T)=0,4 and cluster ordering threshold (θ)=10 where ROUGE-1 was gained from distribution of sentence method increased 3% compared to the SIDeKiCK method (0,392) and increased 5,1% compared to LIGI method (0,385). The result of ROUGE-2 using sentence distribution method increased 13,7% compared to the SIDeKiCK method (0,106) and increased 14,4% compared to LIGI method (0,105).

References

[1] Ouyang Y, Li W, Zhang R, Li S, Lu Q. A Progressive Sentence Selection Strategy for Document Summarization. *Journal of information Precessing and Management*. 2013; 49(1): 213-221.
[2] Suputra HGI, Arifin ZA, Yuniarti A. Strategi Pemilihan Kalimat pada Peringkasan Multi-Dokumen Berdasarkan Metode Clustering Kalimat. Master Thesis. Surabaya: Postgraduate ITS; 2013.
[3] Sarkar K. Sentence Clustering-based Summarization of Multiple Text Documents. *International Journal of Computing Science and Communication Technologies*. 2009; 2(1): 325-335.
[4] He T, Li F, Shao W, Chen J, Ma L. *A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi-document Summarization*. Proceeding of International Conference Advance Language Processing and Web Information Technology. Eds: Ock C. et al., University of Normal, Wuhan. China. 2008: 81-86.
[5] V Kumar R, Raghuveer K. Legal Documents Clustering and Summarization using Hierarchical Latent Dirichlet Allocation. *IAES International Journal of Artificial Intelligence (IJ-AI)*. 2014; 2(1): 27-35.
[6] Kruengkrai C, Jaruskulchai C. *Generic Text Summarization Using Local and Global Properties of Sentences*. Proceedings of the IEEE/WIC International Conference on Web Intelligence (WI'03), IEEE Computer Society Washington DC, Halifax, Canada. 2003: 201-206.
[7] Tian X, Chai Y. An Improvement to TF-IDF: Term Distribution based Term Weight Algorithm. *Journal of Software*. 2011; 6(3): 413-420.
[8] Amoli VP, Sh Sojoodi O. Scientific Documents Clustering Based on Text Summarization. *IAES International Journal of Electrical and Computer Engineering* (IJECE). 2015; 5(4): 782-787.
[9] Lin CY. *ROUGE: A Package for Automatic Evaluation of Summaries*. In Proceedings of Workshop on Text Summarization Branches Out. Eds: Moens, M. F. and Szpakowicz S. Association for Computational Linguistics. Barcelona. 2004: 74-81.