SISTEM PERINGKASAN MULTI DOKUMEN BAHASA INDONESIA BERDASARKAN LDA-*SIGNIFICANCE SENTENCES*

ABSTRAK

Sistem peringkasan multi dokumen pada Bahasa Indonesia dapat membantu masyarakat dalam mendapatkan informasi berita *online* yang lebih komprehensif. Algoritma klustering pada domain data teks yang banyak dikembangkan pada satu dasawarsa adalah *Latent Dirichlet Allocation* (LDA). Metode LDA memberikan kontribusi yang cukup baik dalam bidang klasifikasi teks dan temu balik informasi. Salah satu penggunaan LDA adalah peringkasan dokumen, dikarenakan LDA mampu mendapatkan kerangka pemikiran dari suatu dokumen. Penelitian peringkasan multi dokumen Bahasa Indonesia dengan menggunakan pendekatan *unsupervised learning* khususnya LDA masih terbatas. Metode LDA dan *Significance Sentence* mempuyai keunggulan untuk memilih kalimat representatif dari dokumen sumber. Pengujian model dilakukan dengan menggunakan kombinasi parameter alfa 0,1 dan 0,001 serta beta 0,001 dan 0,1 , dengan dikombinasikan bobot level peringkasan 10%, 30% dan 50% pada proses perangkingan kalimat di masing-masing dokumen. Hasil pengujian berdasarkan nilai c*osine similarity* menunjukkan hasil uji terbaik diperoleh pada kombinasi parameter alfa 0,01 dan beta 0,1 serta level ringkasan 50% dengan nilai *cosine similarity* sebesar 0,931.

Kata kunci: *peringkasan; multi dokumen; LDA; Significance Sentence*

MULTI DOCUMENT SUMMARIZATION FOR THE INDONESIA LANGUAGE BASED ON LATENT DIRICHLET ALLOCATION AND SIGNIFICANCE SENTENCE

ABSTRACT

Automatic Multi-document summarization in Bahasa Indonesia can help people get more comprehensive online news information. The clustering algorithm which is widely developed over a decade in the text data domains is Latent Dirichlet Allocation (LDA). The LDA method contributes quite well in the field of text classification and information retrieval. One of LDA's usages is a document summarization method, since LDA is able to get the framework in a document. The multi-document summarization in Indonesian language using unsupervised learning approach, especially LDA, is still limited. The LDA and Significance Sentence methods have the advantage of choosing representative sentences from source documents. The testing model was performed using a combination of alpha parameters 0.1 and 0.001 as well as beta 0.001 and 0.1, which is combined with compression rate at 10%, 30% and 50% in the sentence ranking process of each document. Testing results show that the best result was obtained under parameters combination as follows: alpha value is 0.01, beta value is 0.1, compression rate is 50% and cosine similarity value is 0.931.

Keywords: summarization; *multi document; LDA; Significance Sentence*