ABSTRACT

The development of the data in text form is growing very rapidly from a variety of sources such as social media, blogs, and news. And there is a need of search methods on a large amounts of textual data. Various solutions have been developed to perform document searching. In general, these techniques are based on the document bag-of-words and projecting them into the vector space, from here the search is done by comparing the vectors of texts to determine the most similar document. Disadvantages of these techniques are they still weak in dealing with the concept of synonyms and polysemy, also they stil not taking into account the implicit context of text.

This research develop a search method by semantic analysis based on the topics contained in a text using Latent Dirichlet Allocation, and also a research to improve the method by analyzing metadata. By integrating the search techniques based on the text metadata and the topic a search technique based on weighted importance can be done. The proposed searching technique uses extraction Gibbs sampling of topics based on LDA, this topic becomes the representation of text searching. The search results then combined with the search results based on metadata. This combination is done by adding the search results based on LDA and metadata based on a weight.

The proposed technique is tested using precision, recall, and F-measure. Searching method for comparison has been carried out for searches based on LDA with the F-measure harmonic highest and lowest obtained 0.78 and 0.71 respectively, searching based on metadata with the F-measure harmonic highest and lowest, 0.49 and 0.70 respectively. For search results with the proposed technique to representation 20 topics obtained F-measure harmonic highest and lowest, 0.81 and 0.71 respectively.

Keywords: Topic Models, Latent Dirichlet Allocation, Metadata, Documents Searching, Indonesian Language.