ABSTRACT

Calculating the similarity of Indonesian Sentences using Latent Semantic Indexing based on KBBI (Kamus Besar Bahasa Indonesia)

Muhammad Panji Muslim

Supervisor I: Dr. Arief Fatchul Huda, S.Si., M.Kom.

Calculating the semantic similarity between sentences is a long-discussed issue in the field of language processing. The field of semantic analysis has an important role in research related to text analysis. In this study, the data used is the definition of each word that has the same meaning by Indonesian dictionary. In this study also, the author presents two methodologies with semantic similarity using TF-IDF algorithm traditional and LSI method using TF-IDF with distribution terms of definition. All definitions are calculated by other definition, by analyzing the definition of which has a high similarity value along with accuracy. Each weight calculation and vectors generated by the TF-IDF algorithm each method, calculation steps to reduce the dimension of vector greatly affect the outcome of similarity values. Evident of this research method to add distribution terms, scores greater weight and influence the value of similarity. LSI method also strengthen the TF-IDF algorithm to determine the similarity distribution a better sentence, taking into account also the character of KBBI. The results obtained in this study, the value of an accuracy of 75.9% for the semantic similarity using traditional TF-IDF and 80% for LSI method and TF-IDF is a weighted term distribution.

Keywords: Sentences Similarity, TF-IDF, Semantic Similarity, Latent Semantic Indexing