The CSI Journal on Computing Science and Information Technology
Vol. 16, No. 2, 2019

Intelligent Selection of Initial Centers in K-means Clustering Algorithm to Improve the Performance in Topic Detection

Sepehr Arvin **Ali Vardasbi** **Hesham Faili** **Azadeh Shakery**

School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran

ABSTRACT

Topic detection is one of Natural Language Processing (NLP) tasks that recently has been taken into consideration. The main goal of this problem is to cluster documents based on their topics. Several solutions have been introduced for this task, among which the use of clustering algorithms such as K-means clustering is quite popular. In addition to clustering algorithms, topic modeling has also been used with promising results. In this paper, first, we show practically the sensitivity of K-means algorithm to the initial centers. Then we propose novel methods for intelligent selection of initial centers in K-means algorithm and improve the performance of K-means algorithm. In the proposed methods, we have used Latent Dirichlet Allocations (LDA) for intelligent selection of initial centers and have used K-means algorithm for document clustering. We use LDA's topic distribution calculate the distance between documents. Experiments show that the proposed method performs considerably better in two datasets out of three standard datasets, compared to the LDA methods. In addition, the proposed method always performs better in two datasets and in the other dataset 70% of the time in selecting initial centers.

Keywords: Topic Detection, LDA (Latent Dirichlet Allocation), Clustering, Selecting Initial Centers, K-Means, Distance Metric, Silhouette.