Abstract

KNN is one of the accepted classification tool, it used all training samples in the classification which cause to a high level of computation complexity. To resolve this problem, it is necessary to combine traditional KNN algorithm and K-Means cluster algorithm that is proposed in this paper. After completing the preprocessing step, the first thing to do is weighting the word (term) by using Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF weighted the words calculating the number of words that appear in a document. Second, grouping all the training samples of each category of K-means algorithm, and take all the cluster centers as the new training sample. Third, the modified training samples are used for classification with KNN algorithm. Finally, calculate the accuracy of the evaluation using precision, recall and f-measure. The simulation results show that the combination of the proposed algorithm in this study has a percentage accuracy reached 87%, an average value of f-measure evaluation = 0.8029 with the best k-values = 5 and the computation takes 55 second for one document.
- Zhou Yong, LiYouwen and Xia Shixiong. 2009. An Improved KNN Text Classification Algorithm Based on Clustering. Journal of Computers, vol. 4, no. 3
- Keno Buss. Literature Review on Preprocessing for Text Mining. STRL, De Montfort University.
- Yang Lihua, Dai Qi, GuoYanjun, Study on KNN Text Categorization Algorithm, Micro Computer Information, No. 21, 2006, pp. 269-271
Applications Vol 25 No. 1 July 2011
- Xinhao Wang, Dingsheng Luo, Xihong Wu, Huisheng Chi, Improving Chinese Text Categorization by Outlier Learning, Proceeding of NLP-KEapos;05, pp. 602-607

**Index Terms**

- Computer Science
- Artificial Intelligence

**Keywords**

- Text Classification
- KNN classification Algorithm
- K-means Cluster Algorithm
- TF-IDF Method