ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

Study on Feature Selection Methods for Text Mining

Divya P[1], G.S. Nanda Kumar[2]

M.E, Department of CSE, Kumaraguru College of Technology, Coimbatore, India[1]

Associate Professor, Department of CSE, Kumaraguru College of Technology, Coimbatore, India[2]

Abstract: Text mining has been employed in a wide range of applications such as text summarisation, text categorization, named entity extraction, and opinion and sentimental analysis. Text classification is the task of assigning predefined categories to free-text documents. That is, it is a supervised learning technique. While in text clustering (sometimes called document clustering) the possible categories are unknown and need to be identified by grouping the texts. Clustering of documents is used to group documents into relevant topics. Each of such group is known as clusters. It is an unsupervised learning technique. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. The high dimensionality of data is a great challenge for effective text categorization. Each document in a document corpus contains much irrelevant and noisy information which eventually reduces the efficiency of text categorization. Most text categorization techniques reduce this large number of features by eliminating stopwords, or stemming. This is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. This paper gives a literature survey on the feature selection methods. The survey mainly emphasizes on major feature selection approaches for text classification and clustering.

Keywords: Text mining, Text classification, Text clustering, and Feature selection.