2nd International Conference on Intelligent Computing, Communication & Convergence

(ICCC-2016)

Srikanta Patnaik, Editor in Chief

Conference Organized by Interscience Institute of Management and Technology

Bhubaneswar, Odisha, India

Prediction of Online Lectures Popularity: A Text Mining Approach

Kavita S. Oza[a], Poornima G. Naik[b] *

[a]Shivaji University, Kolhapur, Maharshtra-416004 ,India

[b]CSIBER, Kolhapur, Maharshtra-416004 ,India

Abstract

Text mining is an emerging area of research with flavors of opinion mining, sentiment mining, document classification, content mining etc. Another flavor of text mining is text clustering. Proposed work is based on clustering the comments posted by users to online learning. The dataset is prepared using comments posted by users for text mining video lectures using R and Weka. In the proposed work learners comments for online text mining lectures have been clustered to observe the popularity of the lectures by analyzing the terms in each cluster.

Keywords: Text Mining, Clustering, R studio, Video lectures, online learning

1) INTRODUCTION

* Corresponding author. Tel.: +9850766660 ; fax: +0-000-000-0000 .

*E-mail address:*kso_csd@unishivaji.ac.in

Text mining is a specialized branch of Data mining. Data mining deals with mining hidden knowledge but in text mining information is plainly present in the text there is no concept of hidden information. Text mining main objective is to get text in computer understandable form directly so that it can be processed without human intervention. Data mining works with structured data like databases, data warehouse, online shopping data, mobile usage data etc. Text mining works with unstructured or semi-structured natural language data. Example of dataset for text mining is data generated by social media, which is natural language unstructured data. So biggest hurdle to text mining is natural language processing.

There are technologies coming up which will help computers to understand natural language to analyze and interpret. Some of the techniques used in text mining is information retrieval, summarization, clustering and classification. Information retrieval may be to mine some interesting patterns within the text. Human experts do exist who can summarize the text with fewer sentences and core concept. Attempts have been made to develop techniques which can help

computers in summarization of text. It may deal with summarization of one document or group of documents. Summarization is condensation of text into smaller version. Here information and meaning of text is maintained.

Clustering is another popular unsupervised technique of Data mining. Text data clustering has applications in customer categorization, document classification, pattern evaluation etc. In the proposed work learners comments for online text mining lectures have been clustered to observe the popularity of the lectures by analyzing the terms in each cluster. Paper is organized into four sections with section 1 dedicated to introduction of the text mining, section-2 deals with literature survey , section- 3 talks about data set creation followed by data analysis and paper is concluded with conclusion.

2) LITERATURE REVIEW

 New technologies are emerging to make text mining more comfortable like Data mining. Its not only text data which mined but also text stream mining is a new research area where new techniques are proposed for text stream classification and evolution analysis of the same[1],[2] . Clustering is widely studied data mining problem in the text domains with applications in number of domains. A detailed survey of the problem of text clustering has been carried out with text domain in focus [3]. A comparative study of document clustering using various techniques on twitter data has been carried out by[4]. Improvement in the clustering co-citation models by using full text along with bibliographic information has been proposed by[5]. Real world applications of text mining and its complexity in implementation has been studied by [6]. Lots of work has been carried out in clustering and analysis of complete biomedical article texts. An algorithm is introduced by [7] for Semi-supervised Affinity Propagation (SSAP) to improve