International Journal of Research in Engineering, Science and Management
Volume 4, Issue 4, April 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Social Media Platform Using K-Mean Clustering

Akshay Gupta[1], Atul Tiwari[2], Kelvin[3*], Chirag Sanwal[4]
[1,2,3,4]*Department of Computer Science and Engineering, Abdul Kalam Technical University, Lucknow, India*

***Abstract*: In today's scenario there exists piles of information on any social media platform which raises the issue of finding and producing the most relevant piece of information which are data chunks and serves as feeds to be consumed by users that lies in the domain of interest and online behavioural activity pattern of the same users and help them find alike personalities with same interest. The current structure of social media platform is to have the most engagement by the user on their platform irrelevant of the content of information which may just serves as noise just another chunk of information for more user engagement instead of knowledge which turns out to be productive for the user. In this paper we present an acknowledgement to this information serving as chunks of noise to device a structure of a platform that will enable implementation of a social media platform that will help its users to develop and find skillset enabling them to be more productive and grow their skill tree of their knowledge. Our platform structure makes use of a machine learning technique called K-Mean clustering algorithm used in data mining over very large amount of data, an iterative approach to partition data sets into distinctive non-overlapping clusters which can be used for optimizing the content delivering engine to any specific individual or groups of users.**

***Keywords*: K-mean Algorithm, Clustering, Machine Learning, Data Mining.**

1. Introduction

In recent years there has been a tremendous growth in the volume of data. To draw meaningful insights from this mountain of data we need algorithms which can perform analysis on this data. Basically, as of now social media is really an important and crucial part of a person's life. Another very important use of social media is that new issues and discussions are initiated and shared on it. And to complete this task and to enjoy this process a bit more we need a good number of participants to initiate the discussions. The discussion we all know basically comprises of positive as well as negative comments and reviews by different type of people all over the social media globally. So, in this manner we can say that people can be categorized into various categories. Now what is this category and how to put people into such groups is based on their reviews and comments. In this way we actually proposed a system to group different kind of people into different categories. E.g., politician, film industry etc. So, now basically what we do is we parse the social media data such as the user messages and identify the network relations and then data mining techniques are applied to it and for this purpose, we basically use K-MEANS CLUSTER Algorithm. Clustering is the process of grouping of data or dividing large data set into smaller data sets of some similarity so that the objects in the same cluster are more similar to each other and more different from the objects in the other group. Clustering is important analysis techniques that is employed to large datasets and finds its application in the fields like search engines, recommendation systems, data mining, knowledge discovery, bioinformatics and documentation. Nowadays, the data being generated is not only huge in volume, but is also stored across various machines all around the world. We need to process this data in parallel to reduce the cost of processing.

2. Methodology

The technology stack and techniques used are as given below.

A. Mern

MERN stack is a JavaScript stack that is used for easier and quicker deployment of full-stack web based applications.

MERN stack is simply a group of four technologies:

- MongoDB,
- Express,
- React and
- Node.js.

It is designed to make the development process smoother and better.

B. K-Mean's algorithm

K-Means Algorithm is a clustering algorithm to partition the number of observations into clusters in which each observation belongs to the cluster with the nearest mean.

K-means takes two variables as inputs:

- The first variable is the observations that we want to cluster.
- The second variable is the size of the cluster.

C. K-Means Implementation in Node.js Using node-k means

We will build an API REST server using Node.js and create a function to cluster our observations using K-Means Algorithm. We can use a library named node-k means.

In our case, the target user that we had to arrange contained

*Corresponding author: userkelvin99@gmail.com

their interest information. We would consider this information as the first variable. Second target user who has same interest as of first would be our second variable.

K-Means will cluster the users into three groups. Each group may not contain the same number of users, but we can be sure that the interests of the users in the same group are close to each other.

D. Node.js API server

This API server will contain functions to implement K-Means Algorithm.

E. React web application

This will be our user interface where one user will greet with some another user of same interests.

F. Working of K-mean clustering

K-mean is the simplest unsupervised learning algorithm that solve the well-known clustering problem. It's a method of vector quantization, originally from signal processing, that is popular for clustering analysis in data mining. The procedure follows a simple way to classify a given data set through a certain number of clusters fixed initially. The algorithm aims at minimizing an objective function known as squared function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{ci} (\|xi - vj\|)^2$$

Where,
||xi – vj || = Euclidian distance between xi and vj
Ci = number of data points in ith cluster
C = number of cluster centers

Step 1: Randomly select 'c' cluster centers.

Step 2: Calculate the distance between each data point and cluster centers.

Step 3: Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster center.

Step 4: Recalculate the new cluster center using:

$$v_i = 1/c_i \sum_{j=1}^{c_i} x_i$$

Where, c_i represents the number of data points in ith cluster.

Step 5: Recalculate the distance between each data point and new obtained cluster centers.

Step 6: If no data point was reassigned then stop, otherwise repeat from step 3.

3. Conclusion

Clustering is technique by which large datasets are dividing into small data collections that are called clusters. There are number of algorithms that work well for clustering the data that can divide a dataset into clusters. Survey on K-Means clustering algorithm proposes different advantage and disadvantage in different K-Means application algorithm.

Acknowledgment

The authors of this paper would like to thank the AWS documentation on K-mean clustering.

References

[1] Social Media Analysis using Optimized K-Means Clustering by Department of Computer Science Bowie State University
[2] 180_Social_Media_Analysis_Using_Optimized_K-Means_Clustering20190521-27362-ht1qoe