International Journal of Web-Based Learning and Teaching Technologies
Volume 15 • Issue 2 • April-June 2020

Topic Sensitive User Clustering Using Sentiment Score and Similarity Measures:
Big Data and Social Network

Bharat Tidke, SVNIT, Surat, India
Rupa Mehta, SVNIT, Surat, India
Dipti Rana, SVNIT, Surat, India
Hullash Jangir, SVNIT, Surat, India

ABSTRACT

Social media data (SMD) is driven by statistical and analytical technologies to obtain information for various decisions. SMD is vast and evolutionary in nature which makes traditional data warehouses ill suited. The research aims to propose and implement novel framework that analyze tweets data from online social networking site (OSN; i.e., Twitter). The authors fetch streaming tweets from Twitter API using Apache Flume to detect clusters of users having similar sentiment. Proposed approach utilizes scalable and fault tolerant system (i.e., Hadoop) that typically harness HDFS for data storage and map-reduce paradigm for data processing. Apache Hive is used to work on top of Hadoop for querying data. The experiments are performed to test the scalability of proposed framework by examining various sizes of data. The authors' goal is to handle big social data effectively using cost-effective tools for fetching as well as querying unstructured data and algorithms for analysing scalable, uninterrupted data streams with finite memory and resources.

KEYWORDS

Big Data, Big Social Data, Sentiment Analysis, Stream Mining

INTRODUCTION

In recent scenario the modern social media, mobile or web strategy involved in communication has been technology concentric that makes data to grow rapidly, ultimately creates large noisy and unstructured data. This gave sudden rise (Felt, 2016) to concept of Bigdata. (Kitchin, 2014) describe Bigdata by 3 V's: large volume, uninterruptable velocity, different data structure as variety which can be extensive in opportunity, interactive in nature, and springy in quality. Many theories in social science like correlation has been proven to be pertinent to social media. As per social correlation theory (Tang, Tan, & Liu, 2014), contiguous users in a social media have similar behaviors or attributes. These phenomena clarify user's inclination to connect or follow with others having certain similarity or sharing the same surroundings. The quantity of information available for harnessing in social media is massive and growing every second. Increasing volumes of data (Tang et al., 2014), has been a major challenge for the data oriented companies like Google, Yahoo, LinkedIn, Twitter, and

DOI: 10.4018/IJWLTT.2020040103

Facebook for which different solutions are proposed and implemented. Managing the voluminous and evolutionary real-world data demands the scalable data management system. Emerging distributed storage system like Hadoop, NoSQL and Cloud based infrastructure aids to reduce the cost for data storage. Researchers and practitioners (Beyer & Laney, 2012; Chen & Zhang, 2014; Wang, Kung, & Byrd, 2018) work dedicated to such huge volume of data shows constant growing interest. Similarly Open source software communities like Apache comes with opinion that due to massive increase in size of dataset it has been quiet difficult to acquire, store and analysed such large volume of data.

Social media (Zafarani, Abbasi, & Liu, 2014) is indeed a way to communicate virtually, in terms of opinions and sentiments of people that can be used by businesses and governments organisations to act accordingly. The process of collecting, integrating, storing and processing of Big social media data to gain information is highly tedious task which yet to be solved fully. In addition, such data needs pre-processing as it contains outliers and noisy data. Similarly, post, opinions or replies (Tang, Tan, & Liu, 2009) from various user on same or different topics have sentiments attached to it. Twitter is a microblogging sites which become one of the main platforms for capturing data to do further analysis. These analysis can be useful for finding out polarity in terms of positive or negative (Tang et al., 2009), detecting trends (Alsaedi, Burnap, & Rana, 2017; Lambrecht, Tucker, & Wiertz, 2018), community detections (Wen et al., 2017), recommendation of product and services (Abbas, Zhang, & Khan, 2015). This paper mainly focuses on stream data management and data analytics for finding similarity and sentiment analysis of user using tweets from Twitter data. Also, investigates and implement various imminent technologies for acquiring, storing and analysing Big social media data. The contributions of proposed work are highlighted below:

- Configured a highly reliable system i.e. Hadoop to store very large files in distributed environment. To ingest data as stream we configure Apache Flume to fetch Twitter data;
- To store Twitter data in structure format, we need to pre-process by storing data in Hive tables. On Hive table we implement a process to calculate sentiments of tweets using HiveQL based on AFFINE Dictionary;
- Proposed architecture for extracting information from large number of tweets to cluster similar user. For calculating Similarity between Tweets, we designed a MapReduce process for efficiency and fault-tolerance which uses text mining approach such as TF-IDF and cosine similarity measure to calculate values for similar tweets and users;
- Finally, results generates output clusters based on sentiments and similarity score.

RELATED WORK

The fast growth of social media networks (Tang et al., 2009) permit various users to relate, which helps to form a group of people who are eager to interact, share, and collaborate using social media platforms. Analysing social media is a tedious task and the existing approaches as well as methods needs to adapt and integrate them to emerging Bigdata models (Chen & Zhang, 2014) for enormous storage as well as processing. Various paradigm like Apache Hadoop and Spark comes into existence that makes possible to have scalable and distributed application of ML algorithms in diverse fields. These Bigdata paradigms consist of numerous in built libraries to improve performance of existing techniques and algorithms (Beyer & Laney, 2012; Chen & Zhang, 2014).

Social Media Mining

Social media mining has been divided into three categories [Figure 1], i.e. user based, link based and content based (Etter, Colleoni, Illia, Meggiorin, & D'Eugenio, 2018; Dridi & Recupero, 2017).

User based techniques explore behaviour modelling and build feature patterns from particular social usernames, idiom and linguistics. This information can be leverage for user classification,