ISSN 2395-1621

Measuring and Rating System by using big data analysis

[#1]Balram Madam, [#2]Ratikesh Sajjanshette, [#3]Pritam Jain, [#4]Abhishek Jain, [#5]Prof. Mugdha Kirkire

[2]ratikeshrockz@gmail.com
[4]abhijain339@gmail.com

[#12345]Department of Computer

G. H. Raisoni College of Engineering and Management, Wagholi, Pune.

ABSTRACT

Now days social media and e-commerce mostly used. online social networking site which contains large amount of data that can be a structured, semi-structured and un-structured data on loud server. In this system we work, a method which performs classification of sentiment analysis in on hadoop framework is discussed. Hadoop is the latest technology for handling the large amount of data.To improve its scalability and efficiency, it is proposed to implement the work on Hadoop a widely distributed processing platform using the Map Reduce methodology of parallel processing paradigm. Finally, extensive experiments will be conducted on available data sets,we apply four algorithm in this proposed system to implement the system, semantic analysis and ranking algorithms 1. Map Reduce, 2. Apriori algorithm, 3. Hadoop word count, 4. Hadoop library, using this methodology we analysis the large amount data and calculate the analysis result.

ARTICLE INFO

Article History

Received: 2nd June 2017

Received in revised form :

2nd June 2017

Accepted: 4th June 2017

Published online :

4th June 2017

Keywords: Sentiment Analysis, Hadoop, Map reduce, HDFS, Apriori algorithm.

I. INTRODUCTION

We check live in a environment where the textual data on the Internet is growing and stored at a rapid pace, this is the major problem of the companies.Many companies are trying to use this deluge of data to extract people views towards their products. Online social network platforms, with their large-scale repositories of user-generated content, can provide unique opportunities to gain insights into the emotional "pulse of the nation", and indeed the global community. Manually it is not possible to judge the amount of data on social networks because it is an unstructured data. Today there are many social sites that make user possible to change, modify data also to express personal thoughts on a particular topic such as twitter. Twitter is a micro blogging site.. Micro blogging and more particularly,

Example: Twitter is used for the following reasons:
• Micro blogging site is a great platform for public opinion because people can express their own thoughts there on a particular issue.
• Twitter has great number of text posts and it's growing fast. Twitter audience can be a common man, a politician, an actor , a celebrity or a professional, students, or a job seeker
•Twitter audience represent from many countries sharing their thoughts, so it can be used for opinion mining..

Map reduce Technique:

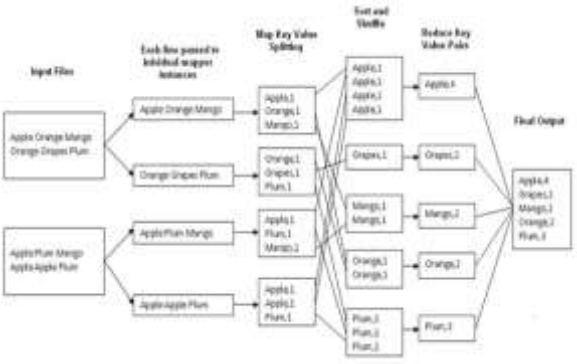


Fig 1. Map reduce technique on hadoop

II. LITERATURE SURVEY

2.1 Lin, Jimmy, and AlekKolcz. "Large-Scale Machine Learning at Twitter."In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pp. 793-804.ACM, 2012. [1] A case study of Twitter's integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform is presented in

this paper. Recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specifically on supervised classification is the base of this . In short, the stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data is identified by author The authors adopt a knowledge-poor, data driven approach. It provides a base-line for classification accuracy from content, given only large amounts of data.

2.2 Bian, Jiang, UmitTopaloglu, and Fan Yu. "Towards Large-Scale Twitter Mining for Drug-Related Adverse Events" In Proceedings of the 2012 international workshop on Smart health and wellbeing, pp. 25-32. ACM, 2012. [2] An approach to find drug users and potential adverse events by analyzing the content of twitter messages utilizing Natural Language Processing (NLP) and to build Support Vector Machine (SVM) classifiers is presented in this topic. The experiments were conducted on a High Performance Computing (HPC) platform using Map due to the size nature of dataset Reduce, which exhibits the trend of big data analytics. Daily-life social networking data could help early detection of important patient safety issues is the main suggestion here. A collection of over 2 billion Tweets collected from May 2009 to October 2010 is the dataset here, from which they try to identify potential adverse events caused by drugs of interest. The collected stream of Tweets was organized by a timeline. Twitter's user timeline API is used to crawl the raw twitter data that contains information about the specific Tweet and the user. The work is indexed only with the following four fields for each Tweet: 1) Tweet id that uniquely identifies each Tweet; 2) User identifier associated with each Tweet; 3) Timestamp of the Tweet; and 4) the Tweet text.

2.3 Liu, Bingwei, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier" In Big Data, 2013 IEEE International Conference on, pp. 99-104.IEEE, 2013. [3] Because of their ability to "learn" from the training dataset to predict or support decision making with relatively high accuracy the machine learning technologies are widely used in sentiment classification. But, some algorithms might not scale up well when the dataset is large .Here the author evaluates Naive Bayes classifier. A simple and complete system for sentiment mining on large datasets using a Naive Bayes classifier with the Hadoop framework is presented. They implemented NBC to achieve finegrain control of the analysis procedure for a Hadoop implementation instead of using Mahout Library, demonstrated that NBC is able to scale up to analyze the sentiment of millions movie reviews with increasing throughput.

2.4Skuza, Michal, and AndrzejRomanowski. "Sentiment analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction" In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, pp. 1349-1354. IEEE, 2015. [4] A possibility of making prediction of stock market basing on classification of data coming from Twitter micro blogging platform is discussed in this paper. Twitter Streaming API is used to retrieve ral time twitter streaming. From 2nd January 2013 to 31st March 2013 the tweets were collected over 3 month's period. Reposted messages are deleted. Each message was saved as bag of words model a standard technique of simplified information representation used in information retrieval after pre-processing. Following is the system design: Retrieving Twitter data, pre-processing and saving to database (1), stock data retrieval (2), model building (3) and predicting future stock prices (4). Considered large volumes of data resulted also in decision to apply a map reduce version of Naïve Bayes algorithm.

2.5 Tare, Mohit, IndrajitGohokar, Jayant Sable, DevendraParatwar, and RakhiWajgi. "Multi-Class Tweet Categorization Using Map Reduce Paradigm" In International Journal of Computer Trends and Technology. pp 78 - 81 (2014) [5] The strategy that uses Apache Hadoop framework, an open source java framework, which relies on Map – Reduce paradigm and a Hadoop Distributed File System (HDFS) to process data this is the topic presented by the author. Using Naïve Bayes classifier relies on two Map-Reduce passes the proposed Map – Reduce strategy for classification of tweets. Twitter4j library to gather tweets which internally uses twitter REST API is used which requires OAuth support to access the API. OAuth to provide authorized access to its API is used by Twitter. After preprocessing of tweets the final step is the labeling of tweets which is based on categories politics, sports and technology.

III. PROPOSED SYSTEM

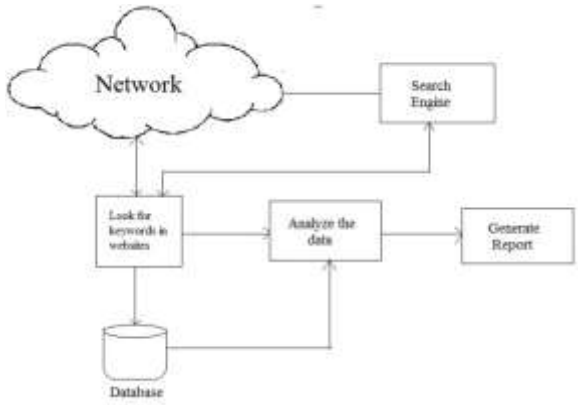


Fig 2. System architecture

We can develop a useful rating system that will help more student's to realize the dream of getting degree from reputed college that unleashes their potential and opens doors to a better life. College rating system will provide the analysis rating done on the basis of reviews, feedback and comments given by students and parents in various websites and social media.

Module:
Search collage name
In this system we can search the college name for submit the any query or feedback.

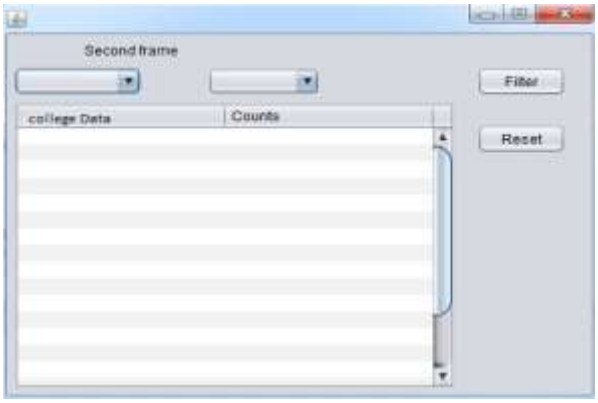
Data collection
All upcoming user/student data we can collect and analysis that data.

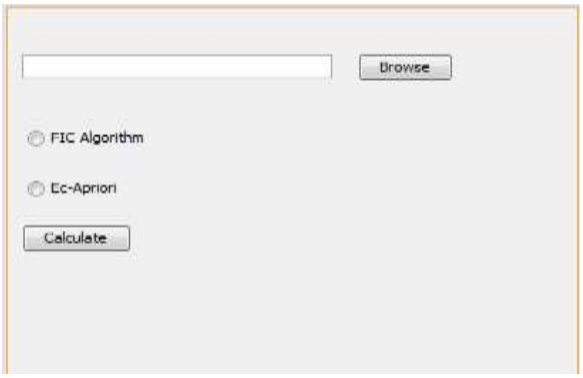
Data clustering (Map reduce)
Data clustering we apply on hadoop server here we use the Map Reduce technique for clustering the upcoming data.

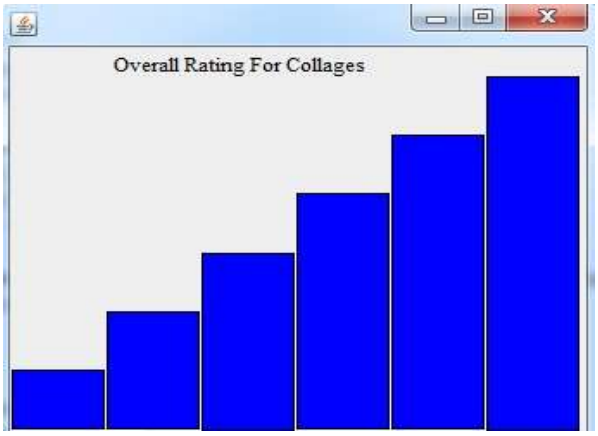
Data classification
Apriori algorithm we can use the data classification for upcoming data through MapReduce technique.

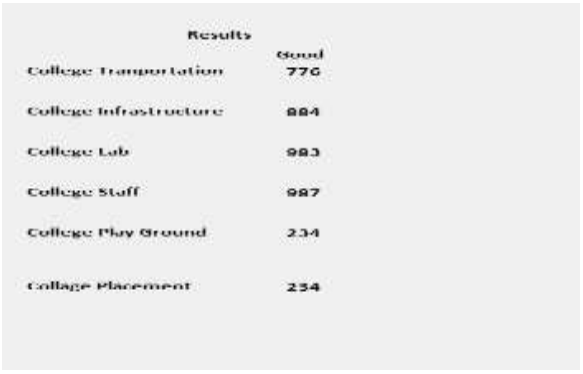
Suggestions and feedback
Here we also user can submit the suggestion and feedback on any activity for future planning.

IV. SYSTEM RESULT









V. CONCLUSION

The researchers in learning analytics, educational data mining, and learning technologies will have a great help from our article. The workflow for analyzing website data for educational purposes which will overcome limitations of both manual qualitative analysis and large scale computational analysis of user-generated textual content. The educational administrators, practitioners and other relevant decision makers will be informed by our article to gain further understanding of engineering students' college experiences. We propose many possible directions as an initial attempt to instrument the uncontrolled website space, for future work for researchers who are interested in this area. Good education and services to them. In the future, it will analyse the student's learning experiences by giving solutions to their problems. To attain the privacy of student and for improving security by a novel secure algorithm the suggested solution is forwarded to the student's individual email-ids. Finally get the feedback from the students about solution provided to generate comparison graph.

REFRENCES

[1] L. Jimmy, and A. Kolcz, "Large-scale machine learning at twitter", In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, ACM, (2012), pp. 793-804

[2] B. Jiang, U. Topaloglu and F. Yu, "Towards large-scale twitter mining for drug-related adverse events", In Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM, (2012), pp. 25-32.

[3] L. Bingwei, E. Blasch, Y. Chen, D. Shen and G. Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naive Bayes Classifier", In Big Data, 2013 IEEE International Conference on, IEEE, (2013), pp. 99-104.

[4] S. Michal and A. Romanowski, "Sentiment analysis of Twitter data within big data distributed environment for stock prediction", In Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, IEEE, (2015), pp. 1349-1354.

[5] T. Mohit, I. Gohokar, J. Sable, D. Paratwar and R. Wajgi, "Multi-Class Tweet Categorization Using Map Reduce Paradigm", In International Journal of Computer Trends and Technology. (2014), pp. 78- 81

