International Journal of Pure and Applied Mathematics
Volume 119 No. 12 2018, 207-211
ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

CRIME ANALYSIS AND PREDICTION USING BIG DATA

Aarathi Srinivas Nadathur, Gayathri Narayanan, Indraja Ravichandran, Srividhya.S, Kayalvizhi.J
Department of Information Technology, SRM Institute of Science and Technology,
Kattankulathur, Kancheepuram, TamilNadu, India

Abstract—Cyber crime is the new age concern which has originated due to the increasing growth and advancements in computer systems and networks, which has given rise to a new environment for criminal activities. Regular occurrences of the crime incidents pose a serious threat to the safety and well-being of the society. This paper offers a extensive overview of crime occurrences and their relevance in literature by combining approaches. The paper initially reviews and identifies features of crime incidents proposing a combinatorial incident description schema. The proposed schema intends to raise a opportunity to systematically combine different elements, or crime characteristics. Additionally, a comprehensive list of crime-related offences is put forward. This enables a thorough understanding of the repeating and underlying criminal activities. This paper intends to be of use to specialists and law enforcement officers in discovering patterns and trends for making forecasts, finding relationship and possible explanations.

1 INTRODUCTION

Big data is simply a large, voluminous data collected from different sources which could be structured or unstructured. Ancient processing system[4] may not be suc-cessful in processing such voluminous data. Big data ana-lytics (BDA) uses extensive techniques and tools for ana-lyzing large,voluminous data sets and drawing meaningful conclusions from the same. The exponentially increasing population in our country leads to increase in crime and in turn generating huge chunk of data which could be analyzed for the government to make critical and essential decisions as to maintain law and order. With the increasing concern of the crime rate this becomes really necessary.

2 PROBLEM STATEMENT

Data mining is sensitive to quality of input data that may be inaccurate having missing information (noise, redundant data). Mapping real data to data mining attributes could be challenging in its own ways. Big data is extensively used to transform large unstructured or structured raw data into crucial and meaningful information which helps in forming a healthy decision support system for the judiciary and legislature to enforce law and order towards keeping crimes in check and making strategic decisions for safety and well being of the society. Increase in population leading to exponential increase in crime rate poses the biggest concern of handling, maintaining and analysing huge amount of data generated every year within a minimal time span.

3 EXISTING SYSTEM

In the existing system[1][2], K-means clustering technique has been implemented using RDBMS containing drawbacks such as data limitations, high processing time and data recovery problems.

4 PROPOSED SYSTEM

The proposed system deals with providing database with high throughput and low maintenance cost using Hadoop tools containing HDFS and map reduce programs. The project will be implemented using joins, partitions and bucketing techniques in Hadoop and be graphically pre-sented using R Tool.The proposed schema could be ex-tended with suggestions to recommend actions, correspond-ing measures and constructive policies according to the of-fence type and the subsequent crime incident. This matching will ensure better security,monitoring,handling and preven-tion of criminal act occurrences.

5 TECHNOLOGIES OR PLATFORMS USED

5.1 HADOOP

The Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models, scaling up from single servers to thousands of machines, each offering local computation and storage. The two major components of Apache Hadoop comprises of HDFS and YARN.

5.2 HDFS AND YARN

Hadoop File System [3] was developed using distributed file system design, holding very large amount of data and providing easier access. The files are stored in multiple machines ensuring redundancy leading to rescue of the system in case of sudden failures. YARN (Yet Another Resource Negotiator) is a cluster management technol-ogy.YARN is an operating system for Hadoop, which specif-ically manages the resources such as the RAM, CPU of all the nodes (machines) in the hadoop cluster. Applications like Hive,MapReduce,Spark requests YARN to allocate re-sources like processing power and memory to fulfill the jobs of the application.

6 IMPLEMENTATION

6.1 Data Preprocessing

The crime data set used has over 1 lakh rows and provides details regarding the crimes that have occurred in the chen-nai region between the years 2005-2015.The data set is first converted to a csv file. Using the My Sql Query Browser, it is then loaded into a structured database.

6.2 Data Migration Module with Sqoop

Sqoop is a tool for transferring data between relational databases and Hadoop(HDFS). Using sqoop we can perform lot of the functions such as, to fetch the particular column or fetch the dataset with a specific condition that will be supported by Sqoop Tool.

6.3 Data Analytic Module with Hive

Hive is one of the data ware house software facility for Hadoop.In this module, we have analysed the dataset using Hive Query Language (HQL). Using hive we perform Tables creations, joins, Partition, Bucketing concept.

6.4 Partitioning

Partitioning is a more effective way for querying. It basically divides the table into parts that are related to one another making querying process easy. Partitioning could be applied on one or more column, imposing multi-dimensional struc-ture on directory storage.

6.5 Bucketing

Tables are divided into manageable parts known as Buckets or Clusters based on a function. The Bucketing concept is based on Hash function, which depends on the type of the bucketing column.It is done to evenly distribute the data in files or buckets.Columns having the same hash value will belong to the same bucket.

6.4 Data Analytic module with MapReduce

MapReduce[6] is a programming method to analyze large data sets.The mapreduce program is written in java. The data is processed in form of clusters using various parallel , distributed algorithms. The MapReduce algorithm contains a mapper,reducer and a driver[5].

Map Stage: The mapper's job is to process the input which is stored in HDFS. The mapper function processes the input file line by line and creates several small chunks of the data.

Reduce stage : The Reducer processes the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS. Map Reduce Process: (i) Splitting: MapReduce job is divided into fixed-size pieces called input splits that is consumed by a single map.

(ii) Mapping: data in each split is passed to a mapping function to produce output values. For example, to count number of occurrences of each crime from input splits and prepare a list in the form of (crime,total) .

(iii) Shuffling: output of Mapping phase. Its task is to consolidate the relevant records from Mapping phase output.

(iv) Reducing: This phase combines values from Shuf-fling phase and returns a single output value. It summarizes the complete dataset.

Map Reduce Processes the data set and analyses effi-ciently and fastly than other hadoop ecosystems such as Hive,Pig,etc.

6.5 R

R is a programming language used for statistical analysis of data.R can perform data manipulation,data analysis and data visualization. In our project we have used R-tool which is used to visualize the analyzed data graphically.

6.6 Correlation matrix

A correlation matrix is a symmetric matrix showing the relation between various attributes. One can know which pairs have the highest correlation. There are various types of correlation matrices but in our project we are using heat maps to analyse the dependencies among the attributes. Heat maps depict the relationship using different colour shades. Darker the shade, more dependency between the two attributes. One can easily study the dependencies using heat maps.

7 ANALYSIS PERFORMED

7.1 Pre-processing and sqoop

(i)First the crime data which is in excel format is converted to .csv file.

(ii)Then the .csv file is converted into a table using MySql Query Browser.

(iii)Using the sqoop tool we export the table to HDFS.

7.2 Hive

i. Using sqoop we create a table in hive.
ii. Then using HQL language we analyse different crimes and their count in a particular area.
iii. Total crime for a particular period of time.
iv. Partitioning technique in hive to partition the type of crime column and analyse.
v. Bucketing in hive: by clustering the dataset based on "area" having a bucket size of 20.

7.3 Map Reduce

(i) For a MapReduce Pogram we need 3 classes.The Driver class,The mapper class and the reducer class.

(ii) Using mapreduce we have analysed the average and total crime in each area of the chennai city.

(iii) Analysed the area where maximum number of crimes occur.

(iv) Analysed number of crimes and the year.

7.4 R-Tool

(i)Bar Graph-Rate of different crimes in an year. (ii)Ggplot-Different crimes in Royapuram (iii)Ggplot-Crimes in entire chennai area wise. (iv)Ggplot-Yearly distribution of all crimes in chennai. (v)Ggplot-Hourly distribution of all crimes in chennai. (vi)Analysis using correlation matrix. Fig 1. shows that the "All other offence" crime category had the maximum count of crime. Fig 2. graph shows that majority of the crime happens between 18-23 which could be regarded as late evening to night. Fig. 3. matrix shows that user generated code which is crime specific and hour of the crime are strongly correlated while year and Distance from the area of crime to nearest police station are slightly less correlated.

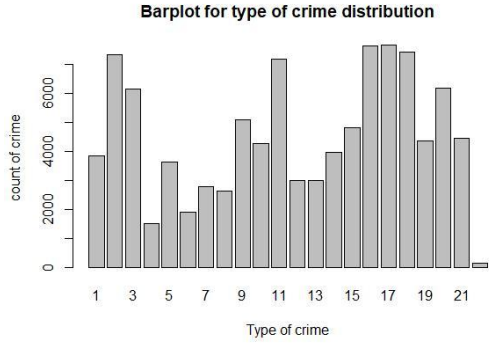

Fig. 1. **Crime distribution**

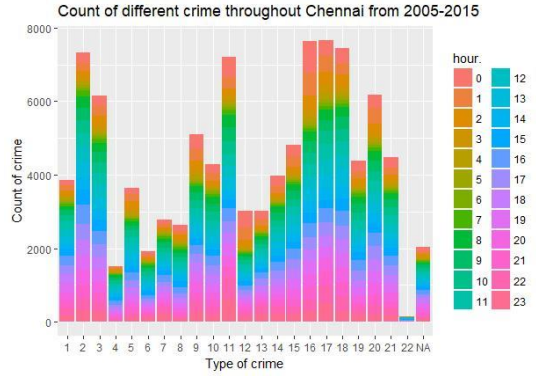

Fig. 2. **Hourly analysis**

8 CONCLUSION

Big Data analytics plays a key role in transforming raw data into important decision support system for the legislature and judiciary to take steps to handle the day-to-day crimes and keep a check. With the increasing population and shoot-ing crime rates, it is of at most necessity to acknowledge the large crime data sets as to use them to identify trends.

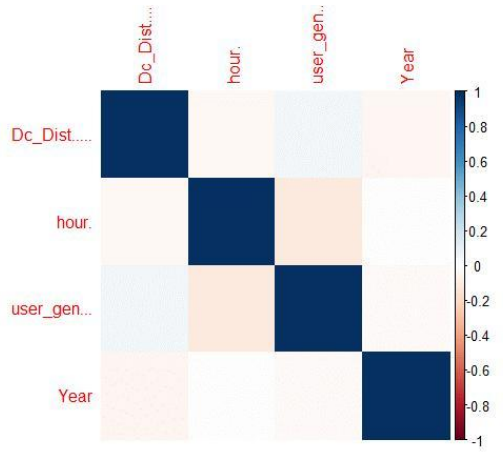

Fig. 3. **Correlation Matrix**

It not only take well informed decision but also helps to maintain decorum in the society by abiding law and order. Precautious measures and extra security could be given area specific to concentrate on the crime around that portion of the locality to ensure safety and well being. For example considering a particular area with a particular crime be-ing predominantly present, security could be enforced at the prime spots to ensure strict vigilance by dispatching required police force depending upon the intensity of the crime. This could help build a sense of safety in the minds of the citizens of a residing country.

ACKNOWLEDGMENTS

We owe our profound gratitude to our Head of the department Dr.G.Vadivu for giving us the opportunity to carry out the research work. We are thankful for all support and guidance from all the Teaching staff and non-teaching staff of the Department of Information Technology, SRM University who helped us in successfully completing our project work.

REFERENCES

[1] S. Sathyadevan, M. Devan, and S. Surya Gangadharan, "Crime analysisand prediction using data mining," in Networks Soft Computing(ICNSC), 2014 First International Conference on, Aug 2014, pp. 406–412.

[2] T. Pang-Ning, S. Michael, and K. Vipin, Introduction to Data Mining,1st ed. Pearson, 5 2005.

[3] S. Kaza, Y. Wang, and H. Chen, "Suspect vehicle identification forborder safety with modified mutual information," in Proceedings ofthe 4th IEEE International Conference onIntelligence and SecurityInformatics, ser. ISI'06 Berlin, Heidelberg: Springer-Verlag, 2006,pp. 308–318.

[4] V. Vaithiyanathan, K. Rajeswari, R. Phalnikar, and S. Tonge, "Improvedapriori algorithm based on selection criterion," in ComputationalIntelligence Computing Research (ICCIC), 2012 IEEE InternationalConference on, Dec 2012, pp. 1–4.

[5] Yun-cheng,"An improvement apriori arithmetic based on rough set theory," inCircuits, Communications and System (PACCS), 2011 Third Pacific-Asia Conference on, July 2011, pp. 1–3.

[6] S. Kaza, T.Wang, H. Gowda, and H. Chen, "Target vehicle identificationfor border safety using mutual information," in Intelligent TransportationSystems, 2005. Proceedings. 2005 IEEE, Sept 2005, pp. 1141–1146.