ACCIDENT DATA ANALYSIS USING HADOOP HIERARCHICAL CLUSTERING

Mr.K. Venkatesh
Asst. Professor,Department of computer science
MallaReddyEngineeringCollegeForWomen, Hyderabadh

Mr. M.Chandra Rao
Asst.professor,Department of computer science
Mallareddy Engineering college for women,Hyderabad

K.shiva Bhavani
Asst.professor,Department of computer science
MallaReddyEngineeringCollegeForWomen, Hyderabadh

Keywords:Bigdata,MapReduce,Hadoop,Trafficdataanalysis,Hierarchical Clustering

Abstract

Road accidents are more frequent today than ever before for a variety of reasons, including poor road conditions, poor weather, driver inexperience, etc. Due to these reasons, a sizable number of individuals in India have died or been hurt. Traffic accidents need to be analyzed to find issues and their influence on death rates and injury severity levels in order to increase traffic safety. In order to investigate the primary causes of accidents in all of India's states, we did cluster analysis using data sets from road accidents that were published by the Indian government. On the Hadoop platform, hierarchical clustering is used to evaluate the traffic data set for every state in India. The main agenda of Hierarchicalclusteranalysis is tofindingdatasetsthatbelongstogetherandat separating them from the other data resulting a cluster ofvariables.

1Introduction

Each and everydaythetrafficinIndiaisincreasinggradually.Itisshow that there is a positive correlation among growthofautomobiles, population of growth and change in income.[1]The main traffic causes twowheelers, fourwheelers and threewheelers, it createairpollution,andnoisepollution. In addition to the pollution the increase in traffic isalsoleadingtotrafficaccidentswhicharemajorcauseofdeath of people. Approximately 1.25 million people die everyyear and 20 to 50 million people suffer with injuries. If thepreventive actions are not taken it is predicted that road trafficcrashes may rise to become 7th place in theworld. [2]

The number of traffic accidents is rising for a number of reasons, including driver error, two-wheeler traffic, environmental factors, driver drug usage, etc. Traffic accidents must be examined to discover potential risk variables and their influence on injury severity levels in order to increase traffic safety. Planning for the improvement of road conditions and implementing the necessary corrective measures to stop traffic accidents requires conducting a traffic study. [3]

GovernmentofIndiapublishing alotofdatarelatedtovariousdepartmentslikehealthandfamilywelfare, Data over the last ten years on information and communications, agriculture, finance, rural areas, and traffic. With the research and analysis of this data, new business initiatives may be developed, precautionary measures may be taken, or some decision support systems may be supported.

2 LiteratureReview

After reading and examining 66 case studies about the effect of drugs on car accidents, Rune Elvik carried out a meta-analysis of them. His findings indicate that 264 estimations of the likelihood of accidents are associated with drug usage while driving.[4] nations. Based on local data and studies into the causes of accidents, safety rules should be put into practice.[6]

Ross After doing logistic regression analysis on various survey responses, Owen Phillips and Friduly Sagberg found that factors like driving off the road, poor road conditions, long commutes, and inexperienced drivers are linked to incidents involving tired drivers.[5]

Only 0.7% of studies published on road traffic injuries are from India, according to N.N. Borse and A.A. Hyder's 2009 study of 826 PubMed articles about injuries caused by traffic accidents. More research should be supported in developing nations in order to increase traffic safety and decrease injuries from accidents. Based on local data and studies into the causes of accidents, safety rules should be put into practice.[6]

3 Implementation

Hadoop is a free, open-source set of tools for scalable, distributed computing. For the analysis of huge datasets, MapReduce is used as the implementation. Hadoop employs the Hadoop Distributed File System, or HDFS, to manage storage resources throughout the cluster. A primary distributed storage file system called HDFS was created to allow dispersed jobs to communicate data across various hardware and software platforms.[7][16][17]

MapReduce: Based on Java, Map Reduce is a framework and programming model for handling huge data collections. Large datasets are processed using Map Reduce on a node cluster. Map Reduce leverages the proximity of data processing principle to minimize data transmission. The map function transforms a set of data into another set of data, where each element is divided into tuples of key/value pairs. Two crucial jobs, namely Map and Reduce, are part of the Map Reduce algorithm. using Map Reduce

theworkflowofInput→Map()→Copy()/Sort()→Reduce() →Output.Mapfunctionexecutesonasetofinputandoutputsasetofrecordsintheformofkey-valuepairs,andpassesthem to the Reduce function. The Reduce function, accepts theintermediatekeypairsandmergestogetherthesamekeyvaluestoformasmallersetofvaluesasfinaloutput.[8][9][3][10][15]

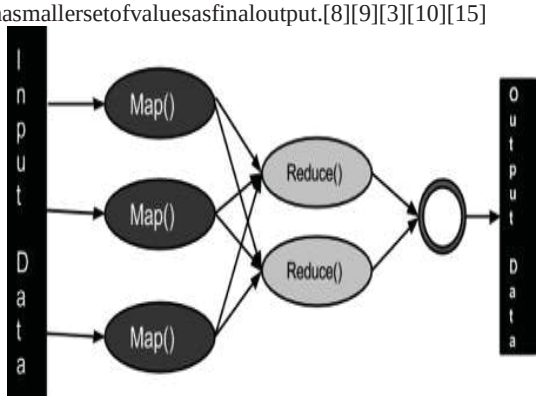

Fig.1:MapReduceframework

4 Methodology

A cluster analysis technique called hierarchical clustering aims to create a hierarchy of groups. It is a commonly used tool for data analysis. The comparable data are grouped together and separated from the other data using hierarchical cluster analysis. The clusters will be made up of uniform variables. Each data point is treated as its own cluster when using hierarchical clustering, and the distance between clusters is determined. The clusters that are closest to one another are combined. Based on the correlation coefficients between the variables, clusters will be created. Depending on the Euclidian distance between the variables, clusters can form. [11][12]

Theroadaccidentdatasetsarecollectedfromthedata.gov.in. and performed the vertical processing of dataandclusteringanalysisisdoneusinghierarchicalalgorithm.

MapReduce

Stage of the map: The job of the map or mappers is to process the supplied data. The input data is often stored in the Hadoop file system as text files. The mapper function receives the input file line by line. The mapper breaks up the data into multiple little pieces after processing it.

Reduce stage: This step combines the shuffle stage with the previous stage. Processing the data that arrives from the mapper is the Reducer's responsibility. It produces a fresh set of output after processing.[13]

[12]

Steps

1. Theinputdatasetisloadedandapplypreprocessing onthat data
2. Partitionthedatabymapping
3. Takingthemapperoutputasaninput,reducethedata
4. Afterreducing,clusteringthedatabyreadingjarfiles.

The input dataset is loaded and apply preprocessing onthatdata.

The initial action taken when data enters the process state is parsing. Parsing is the process of separating or identifying data from a text file. The Indian government publishes information about road accidents that is gathered. Accidents involving two-wheelers, those caused by driver error, and those caused by vehicle defects are all included in the data collection. Preprocessing is carried out based on the data that are available. In the absence of data, zero is used in its place.

2. Partitionthedatabymapping

Theinputdatasetisdistributedtothreemappers.Eachmapperwillseparatelygivetheresultasthepartitionsofdata. The inbuilt class for partition is MyPartitioner which isimplemented from the class Partitioner. The inbult method isgetPartitionbytakingtheparametersastextkey,textvalue. Partitionnumber

publicintgetPartition(Textkey,Textvalue,intpartionerNo).[13]

Taking the fields such as cause of accidents, years and totalnumber of accidents, data is partitioned into three sets anddistributed to three mappers. Each contains the data related tonumber of accidents took place in a particular year in all thestates.

3. Taking the mapper output as an input, reduce the data.The mapper output is taken as an input to reducer and collectdatafromeachmapper.TheinbuiltclassisMyReducerwhich is implemented from Reducer and inbuilt method isreduce. As per the year, the total number of accidents will bedisplayed.

Thepartitioneddataisjoinedintoasinglefilebyusingreducefunction.

4. Clusteringandvisualizingtheclusters

The data is grouped according to the type of accident that occurred in a particular year using hierarchical clustering. Barcharts are used to display the output data. The data in that graph shows the annual total number of accidents caused by three factors: two-wheeler use, driver error, and vehicle problem.

5 ResultsAndDiscussions

The data published by governmentof India is collected forthe analysis.[14] The data is collected for theyears from2006 to 2011 for all the states inIndia. Thesampledatausedfortheanalysisisshownbelowin Fig.No.2.

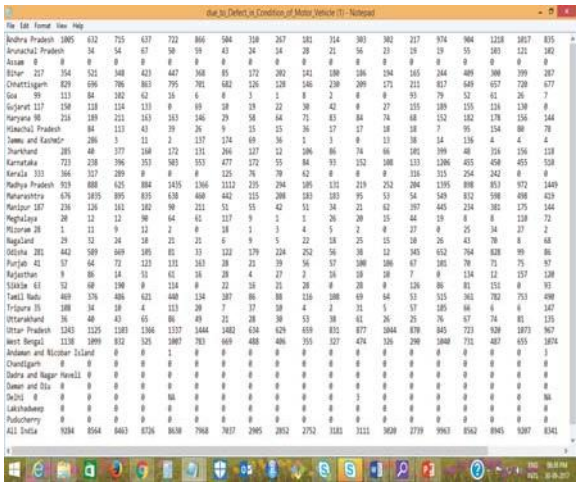

Fig.2:Sampledataset

Weconsideredthreedimensioneddatawithyear,stateandreason foraccident.AfterthepreprocessingstepThe data is partitioned based on the year. and the samplepartitioneddata isshowin thebelowfig.no.3

Partitioneddata

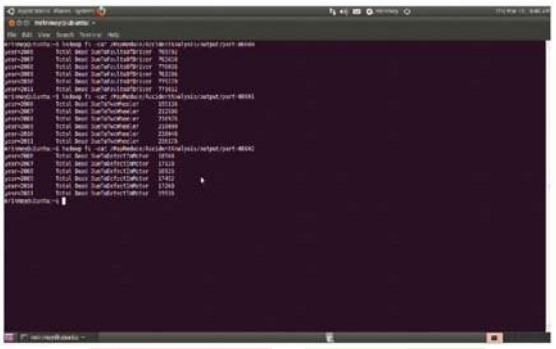

Fig.3:Partitioneddata

In heirarchial clustering the data is clustered based on theproximityfromoneclustertoanother.Theproximityconsidered here is the type of reason for the accident. TheresultantclustersarevisualizedusingbarchartsinFig.4.

Clusteredoutput

Fig.4:Barchartshowingtheclusters

6 ConclusionAndFutureScope:

As the population and earning capability of the people is increasing the rate of purchase of motor vehicles is also increasing. But, the infrastructure facilities and safety measures are not provided on par with the available vehicles. Every year a huge number of people are killed or injured in road accidents. The reasons for the road accidents are many like due to fault of driver, due to bad road conditions, due to consumption of drugs by drivers, due to two wheelers, due to defect in motor etc.. The need for analyzing road accidents due to heavy traffic is most important. The government can suggest appropriate safety measures once the cause of the event has been determined. In order to determine the primary reason for accidents, we examined the traffic statistics in this study. Hierarchical clustering is used to examine the data using the MapReduce architecture. We can infer from the facts that the majority of accidents occur as a result of driver error. The government can lower the number of fatalities and injuries caused by traffic accidents by taking the required action. In the future, we can examine which states have more accidents based on the weather and the state of the roads.

References

[1] RameshwarDASHARMA&SandeepJAIN&KewalSINGH ”Growth rateofMotorVehicles in India -Impactof Demographic and Economic Development”” Journal ofEconomic and Social Studies, Volume 1 ,Number 2, July2011,P137-p153

[2] http://www.who.int/mediacentre/factsheets/fs358/en/(accessedinJanuary,2016)

[3] Vadivel.MandRaghunath.V“EnhancingMap-ReduceFrameworkforBigdatawithHierarchicalClustering”.,InternationalJournalofInnovativeResearchinComputerandCommunicationEngineering,March2014.

[4] Rune Elvik “Risk of road accident associated with the use ofdrugs: A systematic review and meta-analysis of evidencefromepidemiologicalstudies”AccidentAnalysis&

Prevention ,Volume 60, November 2013, Pages 254–267doi:10.1016/j.aap.2012.06.017

[5] RossOwenPhillips· , FridulvSagberg,"Roadaccidentscaused by sleepy drivers: Update of a Norwegian survey"Accident Analysis & Prevention Volume 50, January 2013,Pages138–146.

[6] N.N. Borse & A.A. Hyder, "Call for more research on injuryfromthedevelopingworld:Resultsofabibliometricanalysis"Indian JMedRes129,March2009,pp321-326.

[7] JeffreyShafer;ScottRixner;"The[::Hadoop::]distributedfilesystem: Balancing portability and performance", ;AlanL.CoxPublicationYear:2010,Page(s):122-133.

[8] Ronald C Taylor,"AnoverviewoftheHadoop/MapReduce/HBaseframeworkanditscurrentapplicationsin bioinformatics", BMCBioinformatics2010**11(Suppl12)**: S1**DOI:** 10.1186/1471-2105-11-S12-S1

[9] Jeffrey Dean and Sanjay Ghemawat "MapReduce: SimpliedDataProcessingonLarge Clusters"October2013.

[10] Dr. Siddaraju and Sowmya "Efficient Analysis of Big DataUsingMapReduceFramework" International JournalofRecentDevelopmentinEngineeringandTechnology.

[11] ShenWangandHaimontiDutta,PARABLE:APArallelRAndom-partitionBasedHierarchicaLClustEringAlgorithmfortheMapReduceFramework"https://www.researchgate.net/publication/266422725

[12] Georg Foerster; "Traffic State estimation using hierarchicalclustering:apraticalapproach"YoungResearchers Seminar2007.CD-ROM:27to30May2007,Brno,CzechRepublicBrno: CDV,2007,12pp.

[13] PrasadkumarKale1andArtiMohanpurkar2"BigDataAnalysis usingPartitionTechnique"

[14] www.data.gov.in

[15] HadoopMapReducehttp://hadooptutorial.wikispaces.com/MapReduce

[16] www.youtube.com/hadoopinstallation

[17] http://mvnrepository.com/artifact/org.apache.hadoop/hadoop-common/2.2.0