IJASRM

International Journal of Advanced Scientific Research and Management, Vol. 2 Issue 5, May 2017.

www.ijasrm.com

ISSN 2455-6378

DESIGN & IMPLEMENTATION OF TRAFFIC ANALYSIS AND PREDICTION SYSTEM USING CDR DATA

Suja C Nair[1], Dr. Sudeep Elayidom M[2] and Dr. Sasi Gopalan[3]

[1] Division of Computer Science,Cochin University of Science and Technology,
Cochin, Kerala,682021, India

[2] Division of Computer Science,Cochin University of Science and Technology,
Cochin, Kerala,682021, India

[3] Division of Mathematics,Cochin University of Science and Technology,
Cochin, Kerala,682021, India

Abstract

Effective path prediction is a great challenge in many most sophisticated fields of research and also daily life from robotic motion till Ambulance motion .Actually there are a number of mechanisms exists nowadays ranging from normal infrared sensors to sophisticated sensors for the traffic analysis purpose. But the advantage of this system is that it doesn't need any extra hardware or even no worries of handling the huge data generated by sensors in even nano-seconds. We are planning to make a clear and precise traffic analysis and there by predicting the path by utilizing the most common device of this era the cell phones. Though the Call Detail Record is a collection of a bunch of details about the location and much more details we hope that will serve the data for an effective and cheaper analysis.

Keywords: *CDR, MSISDN, DUMP,GLM*

1. Introduction

Currently there are numerous set of sophisticated sensors are used for traffic analysis. Traffic analysis results are used for a range of application such as from yearly survey to robotic motion. Nowadays sensors are placed along both sides of road or even areal sensors are used. These sensors generate data per nano second and generates a huge data for analysis which the system finds pretty difficult to perform with. Because we need special systems like Hadoop or any distributed system to do the job. Even though the systems are commonly used now but computation wise it makes a limitation in number of devices made for analysis of one device is producing more results which are also not relevant sometimes. And the existing system setting requires extra hardware and thus its costly.

Proposed system is designed more focusing on the data collection part which is used for analysis. Though we focus on data collection part it makes the system completely different from what is existing. As we know all data requires different type of analysis to extract better and efficient results out of it. Here we make use of the CDR data as we said earlier. This requires no special hardware and also commonly all carry it all time which helps in a real time traffic analysis and tracking easier and cheaper. We plan to use R language for easy extraction and exhibition of results. Though R language is a complete solution for analysis purpose we expect this will solve the problem effectively.

The system aims to analyze the live and future road traffic by analyzing the Call details Record effectively and producing useful insights which can answer a series of questions. System is designed so as it can monitor the traffic and also predict by using insights from the analyzed data. The prediction can help in decreasing the loss of time for the users by opting the alternative path suggested by the system.System also helps the whole public not only th registered users by decreasing the accident rates and also by the proper planning of traffic police by using the results from the system.

System functions

• System provides traffic monitoring and predicting the future traffic pattern based on the analysis.

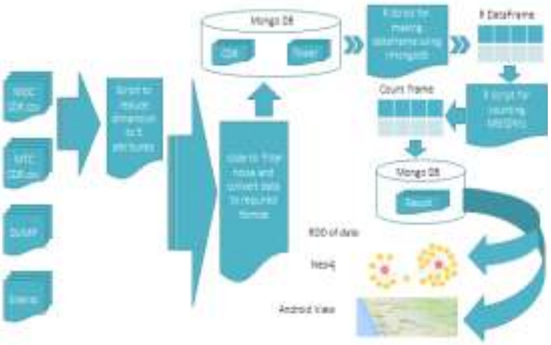
• The system analysis reports can be used for other research agencies for population surveys and also behavioral surveys.

• Help in proper and effective traffic navigation and controlling.

• A more sophisticated model can be used for robot navigation.

• Results can be used for making customized voice or data packs for customers by different telecom operators

2. Program Flow & System Structure



The Program Flow Structure defines the core working and application of the ”Traffic Analysis and Prediction” project .In other words ,it depicts a modular description of the project through its program flow .In terms of the Call Detail Record, we consider MOC(Mobile Originating Calls) CDR,MTC(Mobile Terminating Calls) CDR and DUMP data. Dump data is a form of CDR data which is generated when the cell phone is active but not necessarily calling. So we are considering the CDR data being generated when the cell phone is either calling or idle it is in the active state so that the computation is even more precise .A collection of these data is made in CDR. After the analysis of the CDR data is done it is necessary to preprocess it, preprocessing being a very important phase .A java code is implemented to reduce the number of attributes of the CDR data from 163 to 3 in number.

This is done as a part of Dimensionality reduction so that not much of the time is wasted in processing those attributes that are of no use to our project. After dimension a laity reduction is done, the data is preprocessed by applying a java code to the CDR data collection which filters the noise and converts the data to the required format. Filtering involves converting hex values to decimal or desirable format ,removing extra spaces and indentation, removing null entries etc.. So after preprocessing is done we have the data in its required format. The CDR data is now imported to MongoDB along with the Tower data for storage on a distributed platform. Furthermore, the interaction of MongoDB with various tools technologies such as R , Android(via Json) etc. results in its high usage across various platforms. Also being a NoSQL database ,MongoDB helps in the storage of semi-structured as well as structured data ,being of great help to our project.

The data is finally imported to a MongoDB collection. Then from that collection we can make a data frame in R. The data frames can be of several types containing various information’s such as a data frame specifically for the served MSISDNs, Location Area Codes ,answer Date and answer Time for the purpose of calculation of Density, another data frame containing the Location Area Code,CellId and Area for the purpose of retrieving information about the location and specifications of the tower. For making a data frame from MongoDB in R we use the rmongo db package of R. The point to be noted here is that the collection of MOC,MTC and DUMP data is finally accumulated into a single data frame,used for the purpose of density calculation. At this stage we have the required data frames in R necessary for our project.

The density calculation is done using the first data frame mentioned above. The frequency output of this data frame represents the number of ISDNs present around a particular Location Area Code.The pattern of cell distribution in a Location Area Code is nearly uniform. he various proposed models of this Representation say that different cells can be in form of one of the geometrical shapes such as hexagon, circular ,square etc. .These individual cells have different cell Ids and each of these cells contain a single tower representing a unique Longitude and Latitude. Every location Area Code has many no.of Cells each containing a single tower represented by a tower Id. So different Cell Ids can be a part of the same Location Area Code. After density calculation is done, the GLM model is imparted to the data frames for density prediction purpose. In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. In GLM model we have considered four parameters namely answerDate, weather,Day and Density. For this system we take dataframe of each location by taking its all history values with date, weather,day and density. The weather value is represented as 0 if not rainy and 1 if rainy this weather value can be updated while in a batch process. The 7 days in a week is coded to a numeric value as Sunday is represented by 1 Monday by 2 and Saturday by 7 etc. .

For the prediction first we have to make our dataset as two one is the train dataset and other the test dataset. The train dataset is used to train the model built using the specified formula. The predictors in a model can be set by using a + symbol. The actual output from the prediction is put forward to the android app for mapping to google Maps. Frequent Pattern Mining Algorithm is applied to the CDR collection to find out the Frequent pattern of travel for a particular served MSISDN so that we can suggest him a different path ,if his path is densily populated for that part of the day.

We use arulesNBMiner for this purpose. arulesNBMiner presents an algorithm, which is implemented at R-package and uses a simple stochastic model (Negative Binomal model or NB-model) to estimate a minimum support utilizing knowledge of the process which generates transaction data and allows for highly skewed frequency distributions. The name
of package in R program is arulesNBMiner that is the Java implementation of a depth first search algorithm to mine NB-frequent itemsets of NB-precise rules .Beside the algorithm utilize the information contained in own data structure to estimate the minimum support, it uses a precision limit to estimate min support and for each k-item set plus 1 extension it calculates a different minimum support. The output of this part is a set of transactions representing the frequent patterns of travel for high precision data out of which a search for a particular served MSISDN can be made.

The view is made into three components namely :Android View Neo4j Shiny Android view is an app in android which has Google Maps API implemented. The view contains of two features : Density depiction and prediction.The initial view contains a window which has the various ISDNs plotted.

The second platform for representation is used to give the possible alternate paths or the probable paths.The output of Frequent Pattern miner is supplied as an input to Neo4j for depicting the frequent path of travel for a particular served MSISDN .Neo4j is a graph database management system developed by Neo , Inc. Described by its developers as an ACID compliant transactional database with native graph storage and processing Neo4j is the most popular graph database according to db-engines.com Neo4j is available in a GPL3-licensed open-source ”community edition”, with online backup and high availability extensions licensed under the terms of the Affero General Public License Neo also licenses Neo4j with these extensions under closed-source commercial terms.

2.1 Preprocessing

Data preprocessing is an important and critical step in the data mining process, and it has a huge impact on the success of a data mining project. The purpose of data preprocessing is to cleanse the dirty/noise data, extract and merge the data from different sources, and then transform and convert the data into a proper format. A typical data set in data mining application tends to be high dimensional (hundreds even thousands of feature variables) with both numerical and symbolic type and has millions of tuples.Many actual applications, such as telephone billing, text categorization, and supermarket transactions, may collect hundreds to thousands of feature variables. Nonetheless, not all of the feature variables inherent in these applications are useful for sophisticated data analysis, for example, for data mining. One reason for this phenomenon is because most of the time, the data are collected without ”mining” in mind. In addition, the existence of numeric data and the primitive symbolic values of symbolic attributes create a huge data space determined by the numeric data and primitive symbolic values. In order to mine the knowledge pattern from the data efficiently, it is essential to reduce the data set before the mining algorithm can be mined. Data in the real world is dirty and incomplete lacking attribute values, lacking certain attributes of interest or containing only aggregate data. It is noisy contains errors or outliers and it is inconsistent.

Data goes through a series of steps during preprocessing:

• Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

• Data Integration: Data with different representations are put together and conflicts within the data are resolved.

• Data Transformation: Data is normalized, aggregated and generalized. • Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.

• Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

Java Programm is used to clean the data from the unwanted noise in the cdr file. There were a total of 163 attribute that are available in the cdr file. It is then reduced to 3 attributes which we think that are the only usable attributes for the current senario. Out of 163 attribute ,for the processing we took only the

- MSISDN
- LOCATION AREA CODE
- DATE AND TIME

3. Density Calculation

This section deals with the calculation of density at each Location. For this to be done we should first obtain the count of total cell phones traveled or appeared in that location on that particular day. The count is divided with the geo-location area of that location to find the density. The data frame used here is the same cdr-nb of a specified date. Though the system works in a batch wise manner every days density is calculated at end of the day.

3.1 Counting users at Locations

As per the availability of the data we choose 3 locations of Trivandrum city. The major reason to restrict the study to trivandrum is because the maximum customers of the telecom operator is there. Now we will have a number of around 1 lac records of CDR in a particular day. So counting on this large set of data is not feasible by a normal programming language and also the database to be used for the creation of this dataset is also challenging

In this system we use R's data frame which is very fast in these kind of data storage and analysis. So by using R language and its data frame the challenge of high volume is solved. But now a logical error while counting can happen. The normal way to count a user at a location is to check the appearance of that particular location in the whole cdr transaction list of that day. Here comes the logical error, as we know a cdr record is generated while a user makes or terminates a call so if a user has made multiple calls and also received a call after making one or its all possible combinations can make the count high by counting the same user at multiple

To solve this issue we iterate through the record and set of combinations are checked and also new a temporary data frame named red is made. The logic is that we will count first the occurrence of each location in cdr by neglecting redundant users. Now the red data frame is made from cdr-nb by counting the occurrence of a user's MSISDN and that particular location. So whenever the iterating code over red data frame finds a value more than one it will decrement a value of(frequency-(frequency-1)) from the first count. The resultant value at count data frame will be the value after removing the redundant users.

Though study has performed on trivandrum city we need to consider the geo location area of trivandrum only got density calculation.There are only 3 locations out of 15 found on cdr can we used for mapping purpose because only those 3 lies within the telecom operator which supplies data to us.So now the available count dataframe of R can be accessed and a simple R script can find the density of the location by dividing the count of users at that location by the area of that location. Actually Trivandrum is a district of total area 2192Kmsq .As we have already said only 3 locations can be usable we assume that the three locations of that telecom serves the entire region of trivandrum so as we can divide 2192 by 3 and get corresponding location area's. Now each location's area is about 730.66 Kmsq.

Now divide the count by the area of each location and find the density. The density ranges from a value of 0.7 to 3.5 because the max observed value of count is 2500 and minimum is 500.Hence we get each days density of these 3 locations and these will be maintained as a 3 separate data frames which will be used as a history data from prediction models. Now along with the density value two new predictors are also collected to make the prediction of prediction models perfect. Those are the weather and day values. That will be covered in the next chapter the prediction model using GLM.

4. Conclusions

Aims at plotting the availability of people contributing to the traffic congestion in a particular area. This information can be used by the government, Traffic Police ,Crime Branch etc. for their various purposes. The caller data records are taken from the service provider.This data record is then analyzed using R platform.In this project we plotted the ISDN in the map.Also various prediction algorithms were used such as arulesnbMinor Rules as well as the Generalized Linear Model also. Thus our project avails real time density as well as density prediction in the app and using neo4j we can depict various possible paths . This particular project can be simply implemented in SparkR also with very minor changes .Future scope of this project can include building up a fully automated system which is cloud based so that only we need to update the CSV file on the cloud and then it will automatically map the same. This project can be easily implemented in sparkR with only minor changes. This system can also be used in accordance with the climate .This can be done with a completely synced database of the weather forecasting department Manuscript should have relevant brief conclusion (limit of 200) and should reflect the importance and future scope.

.

References

[1] Chi-Hua Chen, Hsu-Chia Chang and Chun-Yun Su “Traffic speed estimation based on normal location updates and call arrivals from cellular networks”, Simulation Modelling Practice and Theory, pp26–33, 2013.

[2] Sihai Zhang “Computing on Base Station Behavior using Erlang Measurement and Call Detail Record”, Journal Of Latex Class Files, vol. 6, no. 1, 2014.

[3] Paola Pellegrini, Gregory Marli and Joaquin Rodriguez “A detailed analysis of the actual impact of real-time railway traffic management optimization”, Journal of Rail Transport Planning & Management, 2016.

[4] Andreas Janecek, Danilo Valerio and Karin Anna Hummel"The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring", IEEE Transactions On Intelligent Transportation Systems, 2015.

[5] Seungwoo Jeon and Bonghee Hong "Monte Carlo simulation-based traffic speed forecasting using historical big data", journal homepage: www.elsevier.com/locate/fgcs, Future Generation Computer Systems, 2015.

[6] Mahmood A. Khan, Syed Yasir Imtiaz and Mustafa Shakir “Automatic Monitoring & Detection System (AMDS) for Grey Traffic”, Proceedings of the World Congress on Engineering and Computer Science, vol II, 2015.

[7] Rainer Kujala, Talayeh Aledavood and Jari Saramaki “Estimation and monitoring of city-to-city travel times using call detail records”, Springer Open journal, 2016.

[8] Michele Berlingerio, Francesco Calabrese, Giusy Di Lorenzo, Rahul Nair, Fabio Pinelli, and Marco Luca Sbodio, "AllAboard: a System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data",2014

[9] R.C. Sidle, D. Taylor and X.X. Lu “Interactions of natural hazards and society in Austral-Asia: evidence in past and recent records”, Quaternary International, pp118–119, 2004.

[10] Saptarsi Goswami, Sanjay Chakraborty and Sanhita Ghosh “A review on application of data mining techniques to combat natural disasters”, Ain Shams Engineering Journal, 2016.

[11] Jonathan Cinnamon, Sarah K. Jones and W. Neil Adger, “ Evidence and future potential of mobile phone data for disease disaster management” Journal of Science direct, vol. 75, pp.253–264, 2016.

[12] Vanessa Frias-Martinez, Enrique Frias-Martinez and Nuria Oliver “A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records”, Association for the Advancement of Artificial Intelligence, 2009.

[13] William Hsu, Ricky K. Taira, and Suzie El-Saden “Context-Based Electronic Health Record: Toward Patient Specific Healthcare”, IEEE Transactions On Information Technology In Biomedicine, vol. 16, no. 2, 2012

[14] Corey W. Arnold , William Hsu and Ricky K. Taira “A Neuro-Oncology Workstation for Structuring, Modeling, and Visualizing Patient Records”, ACM, 2010.

[15] Gennaro Boggia, Pietro Camarda and Alessandro D’Alconzo “Modeling of Call Dropping in Well-Established Cellular Networks”, EURASIP Journal on Wireless Communications and Networking, 2007.