A NOVEL FRAMEWORK FOR PREDICTING CRIMINAL ACTIVITIES USING BIG DATA ANALYTICS

Dissertation proposal submitted in fulfillment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

RAHUL GARG

11305484

Supervisor

MR ARUN MALIK

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

December 2017

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal entitled "A NOVEL FRAMEWORK FOR PREDICTING CRIMINAL ACTIVITIES USING BIG DATA ANALYTICS" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Mr Arun Malik. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Rahul Garg

11305484

SUPERVISOR’S CERTIFICATE

This is to certify that the work reported in the M. Tech dissertation proposal entitled **“A NOVEL FRAMEWORK FOR PREDICTING CRIMINAL ACTIVITIES USING BIG DATA ANALYTICS”**, submitted by **Rahul Garg** at **Lovely Professional University, Phagwara, India** is a bona-fide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Arun Malik)

Date:

Counter Signed by:

1) **Concerned HOD:**

   HoD’s Signature: ________________

   HoD Name: ____________________

   Date: ___________________

2) **Neutral Examiners:**

   External Examiner

   Signature: _______________

   Name: __________________

   Affiliation: ______________

   Date: ___________________

   Internal Examiner

   Signature: _______________

   Name: __________________

   Date: ___________________

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the guidance to complete this dissertation proposal. I would like to express my special gratitude to my dissertation mentor, **Mr Arun Malik**, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my dissertation, especially in writing this proposal.
Without the support from Lovely Professional University, it would not have been possible to complete this proposal. The faculties have always been very responsive in providing necessary information, and without their generous support, I would have lacked in accurate information on current developments.

I perceive this opportunity as a big milestone in my career development. I will strive to use the gained skills and knowledge in the best possible way and will continue to work on their improvement, to attain desired career objective.

Hope to continue cooperation with all of you in the future.

Rahul Garg

TABLE OF CONTENTS

CONTENTS **PAGE NO.**

ABSTRACT

We know that in today’s technical world, Big Data is a name everyone hears every now and then from small businesses to big corporations. Big Data finds it’s applications in variety of fields ranging from health care services to business analytics, from disaster prevention to smart city designs.

Cities have been developing at a very fast pace. Technology is playing a crucial role in every sector of it. With the advent of Smart City idea, latest innovations have come into play and research is being carried out to identify problems and find solutions for the same.

Data Analytics is one such technology that helps to solve a number of problems. A lot of research has been done on it and new applications of Data Analytics are discovered. To site a few applications - an extensive research in the health sector, financial market and resource utilization is being carried out.

Crimes, however still remain to be a major problem in every part of the world. My research is about using data analytics and big data to deliver important insights which would help make significant decisions to reduce the crime rate.

This dissertation proposal is an attempt to study and yield a prediction model that could be used to help society take a step towards a safer city. This could be done by using data analytics and big data to deliver important insights which would help make significant decisions to reduce the crime rate.

CHAPTER 1
INTRODUCTION

The concept of the smart city is implemented using big data. The idea of smart city is for the betterment of quality of life. Smart city may include one or many of the following mentioned applications – Smart Healthcare facilities, Smart Energy Grid, Convenient Transportation, Green Environment, Public Safety, Smart Education, E-Governance are implemented across the globe in various smart cities. It is a domain open for lot of research and development. Innovations in big data technologies are going to help make smart cities more sustainable.

Another very significant and important domain of smart city is "Public Safety". Now public safety could be ensured via a number of factors – safety from disasters, accidents, healthcare, etc. But our focus is on the crimes that happen all around us and have a huge impact on public safety. We are going to study about the ways of how these crimes affect our society. We are going to examine if there are any patterns or any some sort of relationship between the crimes. Our goal is to find some sort of solution to control the crimes happening around us.

1.1 Big Data

Big Data is a technology used deal with the large amount of data that is generated at a very fast pace constituting different types of data. The collection, storage and processing of large amount of data that is in abundance in each and every field is handled using Big Data. Valuable Insights can be derived from this data using big data analytics which supports the new innovations and development in that field.

Simply put, Big Data is the making of the huge boom in the data. The last 2 decades have seen expansion of technology to the extents never imagined before. Internet, Cloud Computing, Internet of Things, etc. all these technologies have been crucial for the development of Big Data technology. Mobile devices, sensors, web, everything is continuously generating a large amount of data every second. Under this tremendous increase of global data, the term of big data is used to represent the tremendous piles of datasets gathered from a large number of varied sources. Compared to traditional datasets, Big Data comprises of "Big" unstructured data, generated at a very "Big" speed. This grants Big Data it's 3 V's. Volume, Velocity and Variety.

1.2 Smart City

There is no fixed definition for the Smart City and the concept has varying connotations from people's perspective and as that of technological. The smart city phenomena prevail throughout the world. The idea of smart city is about enhancement of quality of life of citizens of the city. This is accomplished by bringing together various components of city - natural resources, infrastructures, power, transportation, education, healthcare, government, and public safety by utilizing the latest technological developments. In addition to that, smart city also emphasizes on the sustainability of resources and applications for the future generations. Big data plays an important role to realize the idea and concept of smart city.

However, we are going to talk about the least focused and talked about scenario of smart cities – crime control. It is very crucial to realize that increasing crime rates have a huge impact on the people and ultimately the city. We can't call a city smart when it's people are living under the fear of crimes happening all the time, ranging from petty thievery to murders and shootings. A city is not smart enough when it can't make use of the present technology to study the crimes and bring forth the solutions which could help to prevent or control these crimes. Crime Analysis – a term used for study of crimes in such a way so as to identify trends, patterns, statistical facts from the crime data which could help us to understand the crimes in a better way. Next section will discuss that in detail.

1.3 Crime Analysis

"The qualitative and quantitative study of crime and law enforcement information in combination with socio-demographic and spatial factors to apprehend criminals, prevent crime, reduce disorder, and evaluate organizational procedures".

The above definition is coined in a report submitted to the office of Community Oriented Policing Services (COPS, USA) in 2001. This gives us a very clear picture that crime analysis is nothing new. On the contrary, crimes have been studied since ages. But there has always been a big gap in terms of technology and availability of information to deduce crucial insights from crimes on a large scale. But with the advancement of technology to this huge extent, we are available with huge number of resources, which avail us with the large datasets using which we can identify patterns and trends in crimes. Crime analysis make use of both quantitative and qualitative data. The non-numerical data (categorical values) is referred as qualitative and analytical techniques are used for the analysis and examination of observations to discover significant insights and trends in the data. Quantitative data is normally present in numerical format. Quantitative analysis helps in the manipulation of relationships for the purpose of describing and explaining a particular phenomenon.

Particularly in Crime Analysis, my focus is on the *Tactical Crime Analysis,* where the recent crime incidents are analyzed based on the factors of location, time, medium, etc. where the data available with the authorities is analyzed to identify how the crimes are being scattered across a particular city and we could identify some patterns about how the crimes vary from location and across the time duration.

Crime analysis is field with vast number of problem to work upon. Crime prediction act as one of the most important problems that is worked upon. Using social networking to predict crimes is a latest problem being worked upon. However, it’s not an easy feat to achieve that. Working with the crime data offers a lot of challenges. The data usually is full of errors and is incomplete. And also, there is a lack of proper standards across the globe to record the crimes that happens. Infrastructure, which could accommodate the various aspects of crime analysis, is a major issue. Even if the data is available, it is not easy to accommodate that data with another framework to achieve more concrete predictions.

Another major issue is that the there are many law enforcement authorities working at different levels. It is quite a difficult task to design an infrastructure that can accommodate data for all of them.

And also, data privacy and security are big concerns. Which data should be accessible to which extent is also major issue that needs to be addressed?

To carry out this crime analysis, it is very important to understand the concept of data analysis. Data Analysis is simply not about just getting the data and start deriving insights from it. Data Analysis is a process which involves much more depth than just that. Data Analysis is a detailed process itself which is divided into multiple phases. Following are the phases of the process of Data Analysis –

- Identify the question that one needs to answer or the problem one needs to solve
- Data Wrangling is the next step in this process. It comprises of two parts of its own – data acquisition and data cleaning. Data acquisition means the collection of data from the relevant sources. Now the data collected is not always perfect. Most of the times it's dirty. It has some inconsistencies, redundancies, incorrect format, etc. present with which it's hard to work upon that data. That's where data cleaning comes into the play. This collected data is cleaned using various techniques and packages available to do so.
- Next step is the Data Exploration phase. Before deriving any insights from the data, one needs to be aware of the data they are working with. That's all is done here in this phase. Analysts plays with the data and try to get a feel for the data. Find the plus and minus about the data and how are they going to delve into it to identify the insights without being affected by the outliers.
- Then in this next step, conclusions are drawn from data after working with the data and carefully studying it.
- In the final step, these insights/findings are to be conveyed to the respective people who will know what to do with these insights. In our case for example, the insights are to be conveyed to the respective authorities which would know the decisions they should make based on these insights and control the crime rates.

A significant number of tools and techniques are available in the market depending on the requirements. To carry out the analysis of this crime data, the most traditional and commonly used tool, MS Excel is used. And also, a very state of the art and feature rich visualization tool, Tableau is used. For statistical computing and visualization, we used RStudio, a very prominent tool for statistical purposes which uses R, the programming language which is very easy to use. In this paper, visualizations are created by using R and Tableau are discussed in detail. Some of the packages which were excessively used for exploratory analysis and visualizations are – Lubridate, ggplot2. Lubridate is used to carry out formatting of date and time whereas ggplot2 helps to create visualizations which can be used for exploratory analysis and as well as derive trends from the data.

These tools are adequate for the data we are working on can be easily handled by these tools. If we have a very huge amount of data which can’t be managed on a local system, there are advanced analytical tools available which make use of the Hadoop infrastructure to handle the data while the data processing is carried out in the same way. Various number of services are available on the cloud which provides the Hadoop environment to work upon and carry out the analysis necessary. However, that is not in the scope of our study.

CHAPTER 2

REVIEW OF LITERATURE

M Chen, et al. (2014) talks about the big data and related technologies - cloud, IoT, Hadoop and data centers. It discusses about the 4 phases of big data – data generation, acquisition, storage and analysis. Each phase is discussed in a little more detail, where the general background and technical challenges are discussed. Applications of big data is also discussed in this paper- medical, enterprises, social networks, smart grid, etc. And then some technological developments in Big Data are discussed which includes – real-time performance of big data, data transfer, data processing and data security. This paper has been very useful to understand the concepts of Big Data and how big data is making difference in today's world.

J Manyika, et al. (2011) talks about the significance of big data in various sectors. It emphasizes on how big data is going to transform economies and productivity. It discusses about the ways in which big data will create value in the corporate world. There would be transparency by making data available to the relevant stakeholders. Big Data enables experimentation, that can be performed on the high quality of digital data generated by the organizations. The most significant of all would be the role of big data in human decision making. In the corporate world, the whole process of decision making will rely on the insights generated by the big data, making process nearly automated. The scale and scope of the changes big data is bringing about are at an inflection point, set to expand greatly, with innovations and accelerating trends in technology. However, there are challenges to face to fully implement the big data.

S.R. Ahmed (2004) brings the light on the application of data mining techniques in the field of retail business. The paper also includes the detailed study of pros and cons of these techniques. This paper is the summary of all widely used data mining techniques.

G V Dhivyabharathi, & Prof. S. Kumaresan (2016) explains duplication as the presence of similar entities in a relation known as duplication. In this survey the architecture of duplication detection has been explained, the architecture states the whole process from selection of the data to the interpretation of the data. It also states the various adaptive techniques for reducing or decreasing the run time in the detection and progressive techniques helps to satisfy the progressive needs of the client. It also explains about the merits and demerits about the various duplication detection techniques. And also justifies how the adaptive and progressive algorithms are beneficial for the duplication cleaning.

J West, et al. (2006) has analyzed some of the relevant experimental issues of fraud detection with the focus on credit card fraud. According to the author, previous researchers have not made enough observations with respect to financial fraud detection. In general, many observations have been explored but not in context to financial fraud detection. Author has also investigated credit card fraud with the help of different performance metrics, controlled simulation, and concentration on detection algorithms.

D Jeffrey and G. Sanjay (2004) in this paper by google has been revolutionizing in the development of big data technologies. This paper plays a huge significance in the development of Hadoop. In this paper, the map-reduce technique used by google has been explained. MapReduce is programming model working on large clusters of commodity machines, parallelly executing the tasks. MapReduce is a highly scalable technology and computation processes TBs of data on thousands of machines. MapReduce is a programming model which process and generate key value pairs and then reduce function comes into play which merge all the value having same key. This paper is applicable for the large datasets, so that during querying or execution the processing time will be decrease up to an extent.

This paper also explains about the implementation of MapReduce in seven steps. It also states about the fault tolerance, that how the very large amount of data is processed by multiple machines, which minimize the chances of a node failure. This large size of cluster however does have a single point of failure- Name Node. But by replicating the functionalities of a name node, this issue can be resolved. This paper by Google laid the foundation for the development of the one of the most renowned distributed platform- Hadoop.

E Rahm, & H Hai Do (2000) explains the concept of data cleaning. Data cleaning is all about to extract some valuable information from the dataset, this can be done only when we have good tools to extract data from the data warehouse because the data collected from heterogeneous places always create mesh, so to handle that mesh and filtered and extract the correct data is a challenge. The data warehouse collects the info of so many data coming from different resources and to get valuable data as result from that warehouse which is explain in this paper by the ETL process that is extraction transformation and Loading of the data to the database. So, this paper explains about the ETL process and the problems that researcher face during the data cleaning. It also shows that how we can improve the data quality of those data at schema and instance level which comes from single source as well as from multiple source database system.

The prime focus of this paper is to integrate the data from different sources and addressed togather with structure transformations. And also states the various phases that are involved in data cleaning which are also shown below:

- Data analysis
- Definition of transformation workflow and mapping rules
- Verification
- Transformation
- Backflow of cleaned data

These transformation required large amount of metadata at instance level as well as schema level. Thsese transformation defination help to handle the conflicts and resolve them.

Dr. V. Deepali, et al. (2017) in this paper proposed a framework in order to achive the data quality. In this framework the whole process is divided it into five phase in which first phase is for collecting data from hetrogenous sources and do profiling on the data. The next phase is for cleaning the data in which they detect the missing values and after detection of missing value the fill those missing values with nominal values like if the field contains the string values then they filled those values by NA or they remove the field from the dataset. In same phase if they detact that the missing value is a number then they fill that missing filled with the mean of that column so that instead of removing the value they have something that represent that field for the visualization. Then the next phase is for detecting the duplicacy in the dataset. This can be done manually or by running a python script for the detection. When they are done with finding duplicacy they try to remove it from the database. This is because if they didn’t remove it then it will decrease the effectiveness and correctness of the data. So, with this phase they remove the duplicacy and missing values completely from the dataset. So in next phase they have check on grammer or spellings. So for this they will detect the misspelling and corresponding to that they show some suggestion to the user for correcting the spelling.

P. Raghavendra and S. Lokesh (2011) in this paper have tried to detect fraudulent transaction with the help of two famous data mining approaches- the neural network along with the genetic algorithm. According to author even though artificial neural network can work as a human brain when trained properly but still it is impossible for the artificial neural network to completely imitate the human brain. The author has made use of the Genetic algorithm for the better decision making about the features of a neural network like network topology, number of hidden layers, number of nodes etc. Author has used Supervised learning feed forward backpropagation algorithm to train the neural network. The author also describes the working of the neural network as well as problems faced during the training of neural networks. Author has made use of both the algorithms considering the fact that a talented man can do exclusively well when trained properly. With this thought in mind, the author has combined both the algorithms to find better results.

T. Ubon (2016) in this paper reviews the literature on various data mining applications applied to solve the crimes. In addition to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining. The author has beautifully described various data mining techniques along with its pros and cons. The author describes the application areas associated with these data mining techniques. The paper also states the research gaps and challenges present in the data mining with respect to crime.

A M Emad, et al. (2014) in this paper reviews the MapReduce programming framework and its existing applications and its implementation on Hadoop platform for the purpose of big data generated by clinical sources and related medical informatics fields. Various genetic algorithms for example have been implemented using the MapReduce framework in top of Hadoop Distributed File System. Applications of MapReduce can be extended to all kinds of data in clinical field - Publicly available clinical datasets, Biometrics datasets, Bioinformatics datasets, Biomedical signal datasets. Along with these applications, there are quite a few challenges - Technology straggling, Data dispersion, Security concerns and privacy issues, Standards and regulation. Despite these challenges, Big Data has the potential to revolutionize the Healthcare sector.

K. Zhou, et al. (2016) in this paper focuses on the application of big data in smart energy management. It explains the working of smart grid and implications of big data analytics on that. Four major aspects have been discussed – management at the power generation end, microgrid management and renewable energy management, asset management and collaborative operation, as well as demand side management (DSM).

A N Eiman, et al. (2015) in this paper explains the concept of smart city. There are various definitions of smart city across the globe, but the idea behind it is all the same – to improve the quality of life of citizens. This paper talks about the applications of Big Data, how it helps to implement the concept of smart city. This paper further targets to understand the benefits that a big data can incorporate in the smart city and various challenges that it might face.

C. Sergio, et al. (2017) in this paper presents a prototype based on the case of Catania with the aim of sharing data and information, which can be reused as reference practices in other cases with similar requirements. This paper discusses the syntactic and semantic interoperability and how it is important while transforming heterogenous sources into Linked Data. The application of Semantic Web technologies on smart cities data has been discussed. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data technologies like RDF and OWL are used.

Y. Sun, et al. (2016) in this paper evolves around the concept of smart and connected communities SCC, which is another concept of smart cities focusing on the lifestyle of the people. In this paper, it is explained how SCC is a step ahead to improve life style, standards, aspirations, and the productivity of a community. The SCC aims to achieve a goal to intricate the past, present and future and bring together a lifestyle which would be very ideal for a community. This paper talks about how IOT and big data analytics can help us achieve that. A case study of TreSight, is discussed which implements IOT and big data analytics for smart tourism and sustainable cultural heritage in the city of Trento, Italy.

C. Xu, et al. (2016) in this paper enlightens the basic of data cleaning. This paper focus on quantitative approaches like as outlier detection. With this they also discuss the various approaches and challenges incipient in data cleaning. This show the Qualitative error detection need to face some question shown below:

- What to detect: This describes the pattern of the instance. Or in other words it can be classified as that which type of error need to capture. For this they purposed the ICs.
- How to detect: This states that detection of fields automatically, which occurs more than once in the table. As in other approaches human helps for the detection of so.
- Where to detect: This states that where we need to detect he error so that the time in finding the error will be decreased. So, in this step they make sure where they have need to detect the error.

Apart from this this paper also tells the error repairing techniques which are shown below:

- What to repair: This states that what type of errors need to be targeted. As the majority of techniques repair the data with one respect or one type of error only and some techniques which interact among more than one type of error are provide with holistic repair.
- How to repair: this stated that how to the errors without manual help, their prime focus is on the automation. So that the machine automatically cleans the errors.
- Where to repair: For this phase they state to create a repair model. And the queries initiated by the user are handled and answered by that query model only.

The various challenges arise during this are scalability, user engagements, un-structured and semi-structured data, new applications for streaming data, growing privacy and security concern. So, it become very important to have a look on it carefully so that the results and output get after analysis aren't inappropriate for the data analysis.

L. Hong, et al. (2016) in this paper discuss the uniform framework that that they provide. It's main feature is that based on the context and usage patterns only, it unifies the ID and connection. This paper makes the data complete by filling the null values present in the dataset.

This paper also states the framework for data cleaning. Which states that the data coming from different sources in different form like media, text or in the form of image. Which is stored in one database and create the metadata of it then with the help of machine and expert the ranking of the data is done and then they make a report on the data.

The main and core part of this paper is data association and repairing process, there also a framework for this in which the metadata assists in identification of items and their uniqueness for all type of dataset. The metadata generator stores both structural metadata as well as descriptive metadata at whatever time new dataset is stored in the Hadoop. This paper also shows the repairing if the dataset like when the header field is unavailable for the dataset then the metadata generator generated the relevant name for it. With this if we delete any field form the Hadoop cluster then the corresponding metadata of that filed is also get delete. So this paper enlighten us about the fixing and removing of the damaged and inaccurate data with their purposed algorithm.

S. Iaca, et al. (2014) in this is a white paper published by the International Association of Crime Analysts. This paper discusses about the significance of crime analysis in police force throughout the globe. But the various terms used in the study of crime don't necessarily convey a standard meaning to everyone. These terms are used ambiguously in different circles and settings. However, this paper focusses on the standardization of these terms of the study of criminology and also discuss which techniques are involved in various types of crime analysis.

This paper makes use of the widely known definitions - tactical, strategic, and administrative crime analysis with slight modifications. Intelligence analysis has been recognized as a form of crime analysis when practiced at local level authorities, where intelligence refers to the people connected to the crimes, chiefly used to identify suspects or the perpetrator.

H. Wickham (2014) says that in Data Science, 80% of the time and cost involved is spend on the cleaning of the large amount of data sets, which could then be analyzed. Data Cleaning incurs a huge amount of effort to get the data ready for analysis purpose. This paper covers the gap of the very less research done regarding the process of data cleaning and how to make it easier and effective. This paper discusses the concept of data tidying, a small part of data cleaning and yet has a very huge significance.

Tidy datasets are very easy to work with. They have a specific structure and are easy to visualize and manipulate. Everything is sorted properly in the form of tables. This makes it easy to develop a framework which will tidy the messy datasets using small set of tools to deal with the large number of un-tidy datasets.

Ngai, et al. (2010) in this paper presents a review of classification structure of data mining techniques for the detection of financial fraud. The author also provides a systematic academic literature review of different data mining techniques that are used for financial fraud detection.

Articles published with reference to financial fraud detection in between 1997and 2008 are deeply studied by the author and beautifully summarized in the paper. This paper also discourses the difference between the needs of the commerce to inspire additional research and FFD and suggests for further Financial Fraud Detection research.

H. Benjamin and A. Suruliandi (2017) talks about Data Mining and according to them data mining refers to the process of working with the large number of datasets and looking for patterns and trends in the data which would be very significant for an organization to get the value out of its databases. Prediction for extracting new information is performed based on the present trends in the data. Although a lot of work has been done in the field of the data mining, and many new approaches have been identified, the work performed specifically in the field of crimes is very negligible. The government authorities are though equipped with the hardware capabilities to store the large amount of data in government organizations including police headquarters, etc. But a very few efforts have been made to make use of these facilities and perform work on them so as to deduce important information. By performing analysis on that data, some very crucial insights can be derived. This paper focusses on the application of supervised and un-supervised learning techniques for criminal prediction.

G. Vikas, et al. (2007) in this paper talks about the existing techniques that are used for crime prediction. Though these techniques existed for some time now, only a limited success is achieved after refining these techniques. The techniques used these days focusses on the areas which are most affected by the crimes using location info, the ages of the perpetrators or the victims might have some relation, the crimes which are repeated based on either the offender, victim, or the location, etc. Machine learning is being used to identify different trends and patterns to make predictions. But these techniques are very generic. There is a need to carry out these analysis at a very local level so as to integrate more information with the identified stats which are very local to that area. In this way, a practical way of policing can be achieved, and crime analysis can play some real-life role in the control or even the prevention of these crimes.

J. Harihara, et al. (2017) in this paper has enlightened us by explain that the problems and their solutions in the process of data preprocessing, with this also explains about the problems and issues related to data cleaning, this paper also explains about the existing methods and for the removal of inconsistency's mentions below:

- Filter-based Method
- Imputation Method
- Hybrid method
- Wrapper method
- Global imputation of stray and non-missing attribute
- Embedded Method

This paper also states the methods for handling the noisy data as it states that the other paper prime focus is also on removal of noise but it is on the low level which results in imperfect data collection and those imperfect collection interprets the incorrect results for the organization. This also states the improving in stability of numerical values by centering process, which directly indirectly effect the skewness in the data

A Eldawy, et.al (2015) in this paper shows the Demonstration of Shahed which is a Map Reducer based system which is implemented on satellite data collected from NASA from last 15 years. This paper also explains about the four modules of Shahed system that is:

- Uncertainty module
- Indexing module
- Querying module
- Visualization module

This model removes the missing values using 2D interpolation then do indexing which replies all kind of queries (Spatio-temporal and aggregate query) and in Querying module it takes the queries from user in real time by engaging some pruning techniques and using partial aggregate available in index. In last module that is Visualization module allows users to visualize the output which is the result of querying module as heat map. So that the user or observer can easily visualize the changes. The user can see the result of real time or the user can deliver the result through the email if he wants to analyze the visualizations later.

The literature survey states the existing technology and work done by the various researchers in the field of data cleaning in-order to gain a valuable insight from it and try to come up with new and innovative unique solutions for the benefit of company as well as society.

S. Kotsiantis, et al. (2006) has explored the effectiveness of machine learning algorithms and different techniques that can be used for detection of fraud. Author has also tried to identify factors associated with financial fraud. To sum up, the author indicates that the examination of financial evidence can be used in the documentation of FFS and emphasize the importance of financial ratios. In this paper different machine learning algorithms have been compared that can be used in the detection of fraud. Author has a clear emphasis on different financial variables and their effect in determining fraud detection.

CHAPTER 3
PROBLEM DEFINITION

We know that in today’s technical world, Big Data is a name everyone hears every now and then from small businesses to big corporations. Big Data finds it’s applications in variety of fields ranging from health care services to business analytics, from disaster prevention to smart city designs.

While researching the smart city domain and the innovations being carried out in it’s different sectors, I wanted to identify a problem in our society that’s really significant and somehow this could be solved using the modern technology. One such problem that our society faces is the problem of crimes. Every nook and corner of our society is filled with crimes, ranging from petty thievery to the murders. And there is a serious need to control these crimes. And our innovating technology can play a significant role in controlling these crimes and make our society secure for everyone.

We need to identify patterns, recognize chief crimes happening and identify the areas where most of the crimes occur to make decisions and take actions which could help us to deal with the increasing crimes rates. Though crime analysis is being carried out for centuries, in different types – crime mapping, criminal analysis, police planning, operation analysis, etc. the availability of data was never the same. And with the changing ways of crime, which now involve more risk and more damage to the life and property of the civilians, a different approach is needed to study these crimes and find some information which could help us to take initiatives to control these crimes and ensure the safety of people.

CHAPTER 4
SCOPE OF STUDY

Crime Analysis is a very important field where the growing technologies such as big data and data analytics can contribute significantly. Though crime analysis has been carried out since ages, there have not been enough developments which could provide deep understanding of the patterns and relationships between the crimes happening across the city, so that concrete decisions could be taken to control these crimes.

Making use of the historical data available with the city police department, the scope of this study to analyze crimes across a city. And then identify some types of patterns in the crimes which could help the authorities to derive some crucial insights which could help make some important decisions which in turn can control these crimes in the city.

However, suggesting these decisions (control measures for the crimes) is not the scope of this study. This study only focusses on the analysis part, making use of the data to identify trends and patterns in the crimes.

The scope of this study is limited to the development of a framework, using which a detailed analysis of the crimes happening across the city is performed and making use of machine learning techniques, a prediction model is developed which would enable authorities to take more concrete decisions to control the increasing rates.

CHAPTER 5
OBJECTIVES OF STUDY

Following are the objectives of my study –

I. To perform analysis of crime data in a city based on historical data available with the police department.

II. To perform real time analysis of crimes happening across a city based on live feeds.

III. To develop a framework, for predicting criminal activities using big data analytics and machine learning techniques

CHAPTER 6
PROPOSED RESEARCH METHODOLOGY

6.1 Sources of Data

The dataset for crime analysis is an open source data made available by the police department of Baltimore city. This dataset consists of nearly 275k entries which includes the crimes ranging from year 2012 to September 2017.

6.2 Techniques used

To carry out this study of crime analysis, I have made use of the standard data analysis techniques which could be implemented either in Python or in R. The process of the data analysis has clearly defined phases which starts with the pre-processing of the data in which the data is treated through standard checks to remove any unwanted entries and make sure that the data is correct and useful

Python and R, both are open source technologies. Both are rich with packages that allows to work on the data and help in the analysis process.

6.3 Tools Used

We will make use of Anaconda environment, which is a feature rich python environment and allows us to experiment a great deal with it. We will be making use of jupyter notebooks, a very user friendly and easy to share tool which can record the whole process carried out in Python.

For R analysis, we will be using R Studio, another open source tool, which is great to use while working with R (a programming language developed for statisticians).

We will be using one more tool, tableau, a very advanced and feature rich tool for visualization purposes. Tableau is very easy and straight forward to use tool which generates amazing visualizations within clicks to analyze the complex data very easily.

CHAPTER 7
EXPECTED OUTCOME

By the end of this study, we would have developed a framework which could make some crucial crime predictions based on the availability of the crime data. It would be deployed in the form of a prediction model, designed using some data analytics and machine learning techniques.

Throughout this study, the main focus has been the careful analysis of the crimes, using the large crime datasets, made available by the authorities containing basic information regarding the crimes – location, date, time, weapon, type, neighborhood, etc. Using this information, the final aim is to deduce some crucial insights from this information which could help the authorities to take some significant actions which could help control the increasing crime rates.

So, with this study, the final outcome would be a model, which would provide us very significant predictions based on crime trends which would help the authorities to take the necessary actions to control the crimes.

CHAPTER 8
SUMMARY AND CONCLUSION

Cities have been developing at a very fast pace. Technology is playing a crucial role in every sector of it. With the advent of Smart City idea, latest innovations have come into play and research is being carried out to identify problems and find solutions for the same.

Data Analytics is one such technology that helps to solve a number of problems. A lot of research has been done on it and new applications of Data Analytics are discovered. To site a few applications - an extensive research in the health sector, financial market and resource utilization is being carried out.

Crimes, however still remain to be a major problem in every part of the world. My research is about using data analytics and big data to deliver important insights which would help make significant decisions to reduce the crime rate. My research intends to yield a prediction model that could be used to help society take a step towards a safer city.

LIST OF REFERENCES

[1] Min Chen, Shiwen Mao, & Yunhao Liu. Big Data: A Survey. Springer (2014)

[2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, & Angela Hung Byers. Big Data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute (2011)

[3] Ahmed, S.R. (2004). Applications of data mining in the retail business, International Conference on Information Technology: Coding and Computing 2 (2) (2004) 455 – 459

[4] Dhivyabharathi G V, & Prof. S. Kumaresan. "A Survey on Duplicate Record Detection in Real World Data". 3rd International Conference on Advanced Computing and Communication Systems (2016).

[5] Emad A Mohammed, Behrouz H Far, & Christopher Naugler. "Applications of the MapReduce programming framework to clinical big data analysis: current landscape and future trends". BioData Mining, 2014

[6] J. West, M. Bhattacharya, "Some Experimental Issues in Financial Fraud Mining", Proceedings of ICCS, 2016.

[7] Jeffrey Dean, & Sanjay Ghemawa. "MapReduce: Simplified Data Processing on Large Clusters". Google, Inc. (2004).

[8] Erhard Rahm, Hong Hai Do. "Data Cleaning: Problems and Current Approaches" 2000

[9] Dr. Deepali Virmani, Priti Arora, Ektha Sethi, & Neha Sharma. "Variegated Data Swabbing: An Improved purge approach for data cleaning" 7th International Conference on Cloud Computing, Data Science & Engineering – Confluence (2017)

[10] Kaile Zhou, Chao Fu, & Shanlin Yang. "Big data driven smart energy management: From big data to big insights". ELSEVIER, 2016

[11] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, & Jameela Al- Jaroodi. "Applications of big data to smart cities". Journal of Internet Services and Applications, 2015

[12] Sergio Consoli, Valentina Presutti, Diego Reforgiato Recupero, Andrea G. Nuzzolese, Silvio Peroni, Misael Mongiovi, & Aldo Gangemi. "Producing Linked Data for Smart Cities: The Case of Catania", ELSEVIER, 2017

[13] Yunchuan Sun, Houbing Song, Antonio J. Jara, & Rongfang Bie. "Internet of Things and Big Data Analytics for Smart and Connected Communities", IEEE ACCESS, 2016

[14] Hsinchun Chan, Roger H. L. Chiang, & Veday C. Storey. “Business Intelligence and Analytics: From Big Data to Big Impact”. MIS Quarterly (2012)

[15] Navjot Kaur. “Data Mining Techniques used in Crime Analysis: - A review”. IRJET (2016)

[16] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan and Jiannan Wang. “Data Cleaning: Overview and Emerging Challenges”. Proceedings of the 2016 on SIGMOD'16 PhD Symposium San Francisco, CA, USA (2016).

[17] Hong Liu, Ashwin Kumar TK, Johnson P Thomas and Xiaofei Hou. “Cleaning Framework for Big Data – An interactive approach for data cleaning “IEEE Second International Conference on Big Data Computing Service and Applications (2016).

[18] Kostas Kolomvatsos, Christos Anagnostopoulos, & Stathes Hadjiefthymiades. “An Efficient Time Optimized Scheme for Progressive Analytics in Big Data”. Big Data Research (2015)

[19] Mohammad Naimur Rahman, Amir Esmailpour. “A Hybrid Data Center Architecture for Big Data”. Big Data Research

[20] IACA S, Elder J, IACA S, Bruce CW, Santos RB, Rodriguez E, Los Angeles County CA, Steiner F, Police AF, Wyckoff L. Definition and Types of Crime Analysis. 2014

[21] Wickham H. “Tidy data”. Journal of Statistical Software. 2014

[22] Wickham H, Chang W. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 2.2.0.

[23] H. Benjamin Fredrick David and A. Suruliandi. “Survey on Crime Analysis and Prediction Using Data Mining Techniques”, ICTACT Journal on Soft Computing, 2017

[24] Vikas Grover, Richard Adderley, & Max Bramer. “Review of Current Crime Prediction Techniques”, 2007

[25] Ahmed Eldawy, Saif Alharthi, Abdulhadi Alzaidy, Anas Daghistani Sohaib Ghani, Saleh Basalamah and Mohamed F. Mokbel. “A Demonstration of Shahed: A MapReduce-based System for Querying and Visualizing Satellite Data”. International Conference on Data Engineering (2015).