International Journal of Computer Applications Technology and Research
Volume 9–Issue 06, 194-199, 2020, ISSN:-2319–8656

Use of Hybrid Data Mining in Identification of Crime Patterns and Trends in the Matatu Industry in Kenya

Duncan Nyale
Faculty of
Information Science
and Technology
Kisii University
Kisii, Kenya

Samuel Liyala
Faculty of
Information Science
and Technology
Kisii University
Kisii, Kenya

James Ogalo
Faculty of
Information Science
and Technology
Kisii University
Kisii, Kenya

Michael Kangethe
School of Computing
and Informatics
Gretsa Univesity
Thika, Kenya

Abstract: The aim of this study was to propose an automatable technological framework that identifies crime and misconduct patterns and trends in the matatu industry using data mining techniques for intelligence led policing in Kenya. The objectives of the study include to propose a framework for intelligent transport management system with patterns and trends identification capabilities, enhance formulation of policy developments, implementations and government regulations for the transport sector in Kenya, design model system for testing the framework to ascertain its practicability and effectiveness and identify challenges of the transport sector in Kenya. This was an application research which made use of dummy data. The study established that it is possible to use artificial intelligence to manage the transport sector by use of a system that will not only help identify the patterns and trends of matatus' on Kenyan roads but to answer the why's associated with the trends to help come up with meaningful applicable practical solutions to enhance security and integrity in the transport sector in general. The study also unearthed challenges in relation to the implementation of the above. Combination of classification and association rules based data mining approach was utilized for this study due to its effectiveness in bringing out patterns and trends that are interlinked and related to each other.

Keywords: WHO: - World Health Organization, OB: - Occurrence Book, GUI: - Graphical User Interface, SQL: - Structured Query Language, CCTV: - Closed-Circuit Television, CBD: - Central Business District

1. INTRODUCTION

The matatu industry in Kenya has been a regulation nightmare for all the stakeholders despite all the government regulations that have been put in place. This has been a long standing problem as current solutions and efforts to curb this menace have not yet provided effective and long lasting results to tackle this issue. Different criterions have exhibited different problems thus annulling the "one shoe fits all" existing solutions that have currently been proposed and implemented. Problems exhibited in the matatu industry range from the irritant to the diabolical. The issues range from overlapping, over speeding, careless driving, indiscipline, misconduct, theft robberies, accidents, corruption and even worse. This has made it difficult for the authorities to effectively regulate, monitor and maintain this industry as it has been influenced by several factors which include geographic locations, Sacco's, population, income levels (financial class), proximity to the city and matatu physical attributes.

2. BACKGROUND

The public transport industry is one of the most vital development industries in any country. It is the backbone of transportation where those who do not have private transportation can easily access public service vehicles for a fair price. Governments usually ensure that its population has adequate transport fertilities for its masses. In Kenya this industry is usually dominated by the private matatu industry as majority of the people in both urban and rural areas use this network of transport thus proper policies and regulations should be introduced that protect both the consumer and the operators at equal measure. (Nantulya and Muli, 2009).

Crime and criminality in the matatu industry has been a constant thorn in the flesh for both the citizens and government. This has been a constant undesirable feature of the matatu industry and the situation convoluted due to the blatant and constant corrupt activities involving the police and by a large margin bribes given to the police (WHO, 2012). This makes it difficult for the government to arrest offenders and control crime as the complexities involved in both prevention and detection measures are hampered by complicity by the police in the crime activities.

Misconduct and crime reporting can be done using several different channels such as Calling the police and reporting of incidents that occur. The reported incidents are usually recorded in the police occurrence book also known as the OB. Physical reporting directly at the police station where an individual or group physically makes the complaint directly to the police at the station or reporting to the traffic officers on duty. This form has proven less effective in having the offenders apprehended (Ogendi *et al*, 2013).

2.1 Surveillance

Surveillance is the monitoring of the behavior, activities, or other changing information, usually of people for the purpose of influencing, managing, directing, or protecting. It most usually refers to observation of individuals or groups by government organizations; in this case, surveillance is monitoring and recording data about traffic offenses in various forms. This includes:

- Incidences reports
- Location route information
- Physical attributes of matatu's
- Sacco
- Matatu personnel information

2.2 Data Mining and Profiling

For the purpose of this research we focus on the application of a hybrid data mining model combining Classification based and Association based algorithms to discover patterns and trends within matatu industry data. Data profiling in this context is the process of assembling information about a particular individual or group in order to generate a profile - that is, a picture of their patterns and behavior based on the observable event features.

3. METHODOLOGY

3.1 Main Framework Model

The main framework was based on a hybrid model derived partially from the combination of the two case based reasoning techniques which is the Bayesian Belief Networks from the Bayesian Based learning class of algorithms, the main concept of the Bayesian Based Learning which in principle is as below.

Features of Bayesian learning methods:

i. Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct, unlike algorithms which completely eliminate a hypothesis if it is inconsistent with any single example

ii. Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. Prior probability is provided by asserting a prior probability for each candidate hypothesis and also a probability distribution over observed data for each possible hypothesis

iii. Bayesian learning can accommodate hypotheses which make probabilistic predictions e.g. this patient has a 93 % chance of recovery

3.2 Mathematical Model Design

This involves “mapping” the attributes to their occurrences and identification of their associations to the overall result outcomes. Each attribute will be analyzed separately and together with all others to identify the magnitude effect on the overall outcome they have. Once the attributes, their relationships and effects to the final outcome have been discretely identified, they will be combined with respect to their magnitude and effect to the overall outcome to provide the actual mathematical formula that will be mapped on to the framework to provide the complete theoretical blueprint of the system that can be implemented by a variety of different programming technologies and approaches that observe the algorithmic process. This process will involve several mathematical processes. The processes will focus on two main things; attribute incident association calculation and attribute effect calculation.

In the attribute incident association calculation each associated attribute impacts the overall outcome differently as some have a greater impact on the outcome than others. This will be achieved by computing the magnitude effect of the attribute individually and by combination with other attributes that result to the same outcome.

The general mathematical rule to be observed will be as below:

$$w_{c_a} = \sum_{1}^{n} a_{c_i}$$

Equation 1: Proposed Attribute incident Equation

Where:

i. w_{c_a} Is the overall weight of the attribute in terms of how it affects the outcome

ii. n Is the number of possible combinations of the attributes

iii. a_{c_i} Is the combination instance of the attribute combination

This initial model is subject to alteration and improvement during the research process. The result of this model will be the mathematical determination of the impact of specific attributes or conditions to an outcome.

In addition, the attribute effect calculation model will be used to identify the most dominant attribute that leads to an occurrence or incident. The reverse association rule will be as below:

$$a_{Dominant} = a_{iMAX}$$

Where:

i. $a_{Dominant}$ is the dominant attribute

ii. a_{iMAX} is the particular attribute instance *i* of all the available attributes

Equation 2: Attribute Effect Magnitude Equation

Attribute effect magnitude equation means the most dominant attribute that contributes to an occurrence is that attribute that has the highest number of correlation values in the said instance of computation.

This first main stage will involve the most research effort and time of implementation as the whole concept and purpose of the research is based on this stage.

4. PROTOTYPE IMPLEMENTATION

This model will be used to compute and not only identify the attribute that mostly affects the outcome in a particular way but the value that is most prevalent in an event of such a specific outcome.

The attributes that have been considered as necessary variables for the purpose of the framework model include:

i. **Time** in which the incident has been reported by the observer

ii. **Location** where the incident occurred. This also helps in providing a more accurate picture of what happened by where it happened.

iii. **Registration number** which would have initially been entered into the system for accountability purposes.

iv. **Incident type** of the specific crime or misconduct that the vehicle personnel will have been reported to have committed at any one particular instance.

v. **Route plying** of the specific route that the Vehicle and subsequently the crew will have been registered.

vi. **Economic levels** of the specific route the registered vehicle is plying.

vii. **Security levels** in the area the matatu is plying.

viii. **Proximity to any major city** the matatu route is within.

The data generated includes the complaints sent by the passengers and public against specific matatu/personnel. This has proved important as it acts as a factor for determining the level of association between the matatus/routes of interest and the incidence trends.

$$R_i = \sum_{1}^{n} W_j(A_i)$$

Which is achieved by summing up the collective weight and attribute scores as below:

$$R_i = [W(A_1)+W(A_2)+W(A_3)+,,,+W(A_{n-1})+W(A_n)]$$

where:

R_i is the overall Incidence Score Rank ***e.g. Over speeding =123***

W_j is the attribute weight which represents the attribute impact to the incidence ***e.g. Age has a greater influence to offenses such as over speeding than location.***

This will be achieved by computing it as below.

$$\frac{\sum_{1}^{n} C_{a_i}}{\sum_{1}^{n} \theta C_{a_i}}$$

Where $\sum_{1}^{n} C_{a_i}$ is the total Number of times The Attribute *I* will be considered based on a unique case type ***e.g. the value will be 15 as there are 15 reported cases of over speeding thus the total sum of times the attribute*** a_i ***appears in over speeding.***

$$\sum_{1}^{n} \theta C_{a_i}$$

Is the total sum of all the unique non-repeating Attribute values e.g. the sum will be 4 since there are seven different attribute (Route) values associated to over speeding {44, 19, 9, 237} this might be because the over speeding offense has only been reported on those routes.

The above associations ensure that the less the number of deviations or rather distributions the greater the overall impact weight of the attribute to the offense. This ensures that whatever attributes combinations will be done. Also the weights significance of the attributes will be taken into consideration. This will eliminate guess estimation as the final ranking will be purely based on the data and frequency and distributions of the values and attributes.

After computing the weights of the different attributes to each output case the final ranking will be achieved by combining and ordering the different Case scenarios in terms of the frequencies of reported cases and in combination of different multiple attribute search to identify the most probable offense that might be committed based on the attributes and with that the system will produce the most prevalent value of offense based on a specific case or attribute combination.

The final output is a list of numbers in descending order where the Indecent or offense with the combined most likely association value to a particular event is displayed as the first item on the list of entities and the Indecent or offense with the overall least possible likelihood to be associated to the incident of interest is at the bottom of the list.

The validity of this framework is based on the observation of its ranking order of the Offenses that have been given to it based on all attributes individually and combined. This will measure the validity of the output based on all cases and their recorded deviations. There will be dummy records combined specifically for the system testing phase that will contain specific attributes and the system will have to either reach the same conclusion or fail with a certain degree of error margin.

4.1 Framework Validity Computation

Percentage of viability is measured as the accuracy of the system producing the same results and in the same order to the generated case scenarios validation data as described below.

i. By Individual attribute: this will be achieved by computing first the deviations of the attributes and average them to get the mid-point. This will then be used to compute the deviation of the midpoint of the computed data on the existing test records to measure weather it falls within a certain range from the midpoint using the Euclidean distance.

ii. The Second measurements will be by testing the overall system performance by summing all the test results accuracy/Deviation values and averaging them all to give the cumulative framework validity value that will in effect be used as the overall system measure of “correctness and reliability”

The main achievement of this project framework is the separation of the stages of development into two subsections.

i. The development of an association model that is used to compute the value of the weights of the attributes, based on each attribute value of the unique indecent type.

ii. The development of an offense ranking model that will compute the commonalities in order of frequency of appearance to any Offenses.

5. PROOF OF CONCEPT

This is an intelligent transport management system intended to mine data and bring out certain crime patterns and trends influencing road etiquette on Kenya roads. The sytem has the following features and structure:

i. Saves and retrieves data on Matatus, Matatu personnel, Matatu owners, Matatu sacco details, Route details and Incidence reports.

ii. Mines and displays dominant values for each of the above attributes for any selected crime.

iii. Predicts likelihood of a certain offence occurring using the attributes above.

iv. It is web based system

v. Uses a GUI (graphical user interface) user access mode

vi. Needs SQL server e.g. Apache to run and connect to its database

Analysis Module

This is accessible via the Analysis menu and allows one to choose a particular offence and analyze it as shown below.

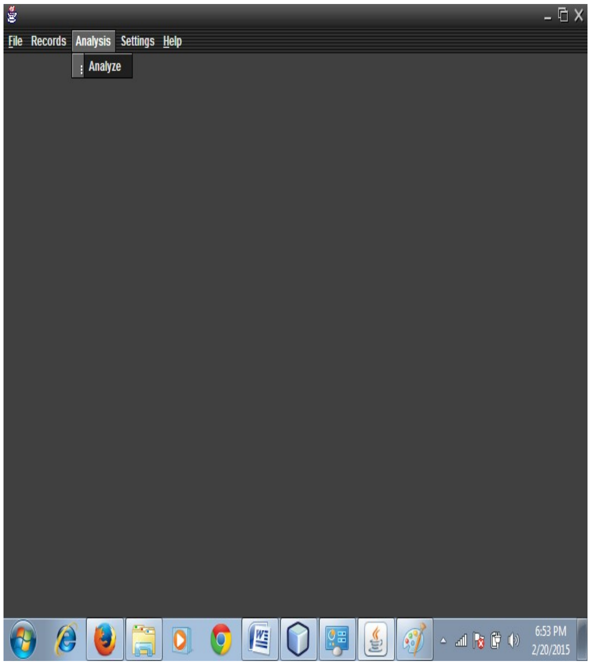


Figure 1. Dashboard Screen

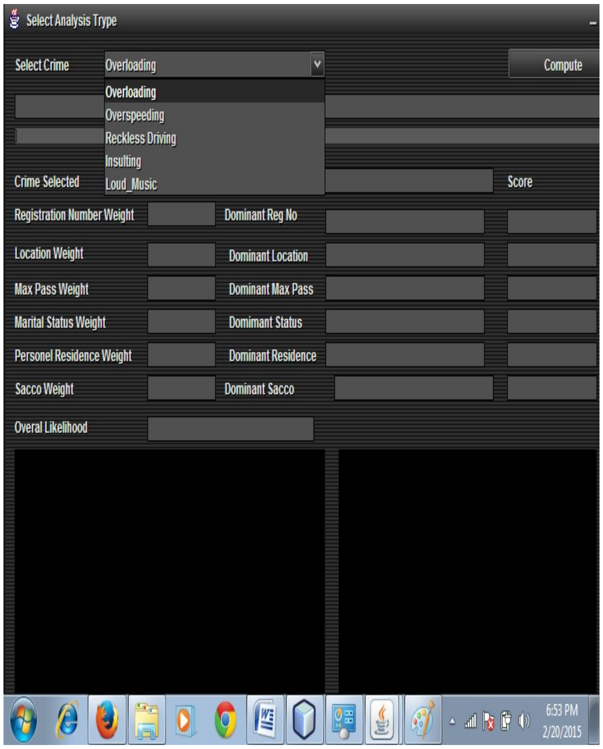
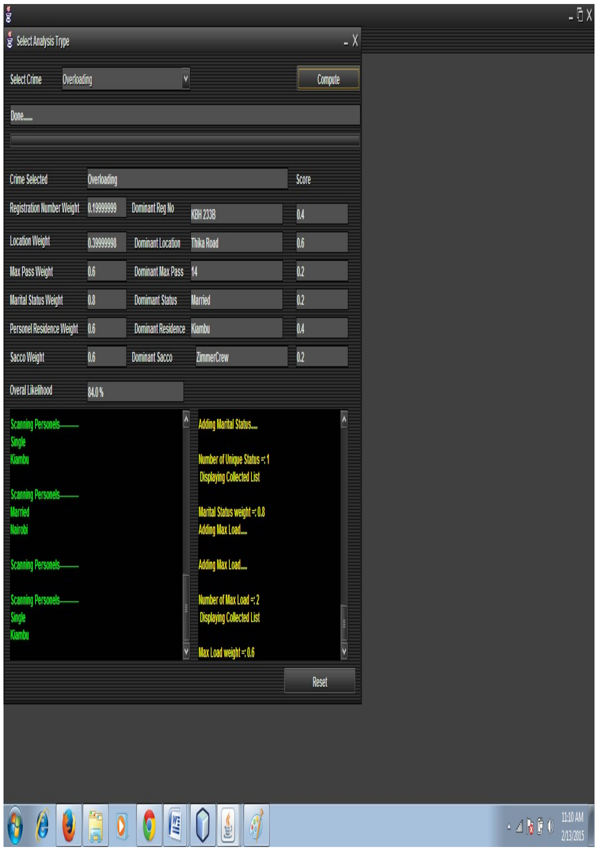


Figure 2. Incidences Analysis Screen


Select Analysis Trype

Select Crime: Overloading — Compute

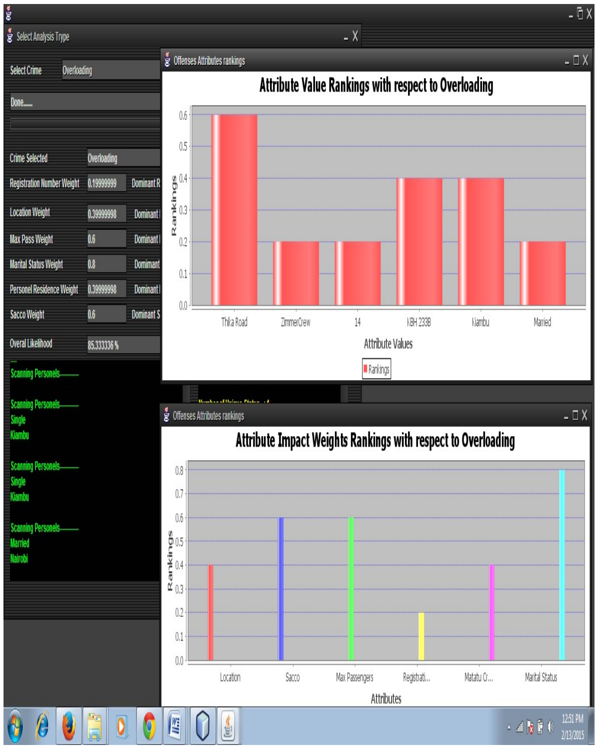
Done.......

Crime Selected	Overloading			Score
Registration Number Weight	0.19999999	Dominant Reg No	KBH 233B	0.4
Location Weight	0.39999998	Dominant Location	Thika Road	0.6
Max Pass Weight	0.6	Dominant Max Pass	14	0.2
Marital Status Weight	0.8	Domimant Status	Married	0.2
Personel Residence Weight	0.6	Dominant Residence	Kiambu	0.4
Sacco Weight	0.6	Dominant Sacco	ZimmerCrew	0.2
Overal Likelihood	84.0 %			

```
Scanning Personels-----------
Single
Kiambu

Scanning Personels-----------
Married
Nairobi

Scanning Personels-----------

Scanning Personels-----------
Single
Kiambu
```

```
Adding Marital Status.....

Number of Unique Status =: 1
 Displaying Collected List

Marital Status weight =: 0.8
Adding Max Load.....

Adding Max Load.....

Number of Max Load =: 2
 Displaying Collected List

Max Load weight =: 0.6
```

Figure 3. Analysis Output Screen 1


Select Analysis Trype

Select Crime: Overloading

Done.......

Crime Selected	Overloading	
Registration Number Weight	0.19999999	Dominant R
Location Weight	0.39999998	Dominant I
Max Pass Weight	0.6	Dominant I
Marital Status Weight	0.8	Domimant
Personel Residence Weight	0.39999998	Dominant I
Sacco Weight	0.6	Dominant S
Overal Likelihood	85.333336 %	

```
Scanning Personels-----------

Scanning Personels-----------
Single
Kiambu

Scanning Personels-----------
Single
Kiambu

Scanning Personels-----------
Married
Nairobi
```

Figure 4. Analysis Output Screen 2

6. KEY FINDINGS

Some key issues have become apparent during the research based on surveillance and below are the noting's.

i. Framework for Intelligent Transport Management System

The first objective was to propose a framework for intelligent transport management system with patterns and trends identification capabilities. The findings were that it is possible to use artificial intelligence to manage the transport sector. This was achieved as illustrated by the and screen shots of the live version of the framework shown earlier.

ii. Enhancement of Policy Formulation and Development

The second objective was to enhance formulation of policy developments, implementations and government regulations for the transport sector in Kenya Events are as a result of active human activity over a period of time. This was shown possible by the capacity of the system to generate and mine vast data generated by the framework which can prove useful decision making.

iii. Testing Model for the Intelligent Prototype

The third objective was to design model system for testing the framework to ascertain its practicability and effectiveness. This was achieved through generation of appropriate dummy data conforming to the information requirements of the model which was then used as a proof of concept of the model proposed by the research.

iv. Identify Challenges of Transport Sector in Kenya

The fourth objective was to identify challenges of the transport sector in Kenya. This was achieved during data generation and extraction which mainly focused on secondary data sources since a lot of research has been conducted previously in this area. Also, through the patterns and trends clearly brought out by the framework

7. KEY ACHIEVEMENTS

The research has resulted to some noted achievements which include:

i. The development of a quantifiable mathematical Model that can be used with desirable confidence to compute the possible relationships between specific attributes both individually and collectively to certain specific crime/Misconduct.

ii. An Automatable framework has been developed that can externalize the human computation process in using the mathematical Model to compute the Associations and identify trends.

8. KEY CHALLENGES

This research like any other has not been without its own challenges, both technical and resource wise as they include.

i. Unavailability of actual real world data to compute real world events thus limiting the models accuracy.

ii. Limited Research time as this research has proved to be wide and has a lot of factors that still need both interpretation and analysis.

9. ASSUMPTIONS AND LIMITATIONS OF SCOPE

The completion of this research and development of this framework and its prototype has come with its own challenges that have been overcome and some ignored due to the near impossibility to tackle natures which are described below.

i. The unavailability of the actual crime data from the police due to bureaucracy and confidential nature of the data.

ii. Also it is assumed that the data collected has been generated from the previous collection of records from offenders.

iii. The above provide limitations to the accuracy of the framework model suggested results as not the actual data has been used but a generated and less accurate data have been used.

iv. The nature of this research is not without its controversial aspect of invasion of privacy to which it is illegal by some government laws. Thus the adoption of this framework means policies within the law will have to be altered to allow the successful implementation of this system.

10. RECOMMENDATIONS

i. Possible Use of Computation

Further Research on the possible use of computations should be considered as it will provide better and more accurate results of the rankings.

ii. Possible Surveillance-Intelligent System Linkage

Actual integration methods that use this framework to various other forms of surveillance such as CCTV and online activity patterns should be considered.

iii. Possible Increase of Attributes Utilized By the System

The use of more attributes should be factored in during the computing process due to the fact that some individuals might exhibit more traits than are currently captured.

11. SUGGESTED AREAS FOR FURTHER RESEARCH

i. Possible Use of Enhanced Security

A study should be conducted on the most effective security mechanisms that can be employed to secure the critical data the system holds and uses.

ii. Attribute - Effect Comparative Study

A comparative study on the effects of each attribute on overall outcome should be conducted to ascertain ranking of attribute – effect relationship.

iii. Broad Study on Challenges of Transport Sector in Kenya

A study on the challenges faced in policing the whole road transport sector in Kenya should be conducted.

iv. Study Expansion to Areas Outside Nairobi CBD

The same research should be carried out in other areas outside Nairobi CBD

12. REFERENCES

[1] Aduwo, I. G. (1990), The Role, Efficiency and Quality of Service of the Matatu Mode of Public Transport in Nairobi: A Geographical Analysis, M.A. Thesis, University of Nairobi.

[2] Asingo, P.O. (2004) The institutional and organizational structure of public road transport in Kenya. IPAR Discussion Paper No.50

[3] Asingo, P. and W. Mitullah (2007). Implementing Road Transport Safety Measures in

[4] Kenya. Working Paper 545. IDS: University of Nairobi

[5] Bruno Agard (2007) Mining Public transport user behavior from smart card data Bruno Agard, Catherine Morency, Martin Trépanier.

[6] Chitere, P. 2006. Public Service Vehicles in Kenya: Their characteristics and Compliance

[7] with Traffic Regulations and Prospects for the Future DP no. 081. Nairobi: IPAR.

[8] Chitere, P., and T. Kibua, 2004. Efforts to Improve Road Safety in Kenya: Achievements

[9] and Limitations of Reforms in the Matatu Industry. Nairobi: IPAR.

[10] Chitere P. O. (2004), Matatu Industry in Kenya. A Study of the performance of their Owners, Workers and theirs Associations and Potentials for Improvement .Institute of Policy Analysis and Research Discussion Paper Series. Discussion Paper No. 055/2004.

[11] Chitere P.O and Kibua T. N. (2001). Efforts to improve road safety in Kenya: Achievements and limitations of reforms in the Matatu industry. Institute of Policy Analysis and Research (IPAR).

[12] Daily Nation Newspapers See Issues of 7th November, 2003 and 3rdFebruary, 4th, 6th, 17th, 19th and 20th, 28th and 30th August, 11th 12th and 14th Sept., 2004

[13] Dale Dzemydiene, (2010),Pattern recognition based on statistics and structural equation models in multi-dimensional data warehouses of social behavioral data Dale Dzemydiene,Raimundas Vaitkevicius, Ignas Dzemyda pg 4 -10

[14] East African Standard Newspaper See Issue of 14th Sept, 2004

[15] East African Newspaper Issue of November 24-30, 2003

[16] Information on transport action plan from Central Connecticut Regional planning agency retrieved from http://www.ccrpa.org/transportation/LRTP_action_plan/ ACTION%20PLAN.pdf

[17] Khayesi, M (1999), The struggle for regulatory and economic sphere of influence in the matatu means of transport in Kenya – a stakeholder analysis at Sixth International Conference on Competition and Ownership in Land Passenger Transport

[18] Odero, W., Khayesi, M. and Heda, P.M. (2003) Road traffic injuries in Kenya: Magnitude, causes and status. In: Injury Control and Safety Promotion, Vol.10 (1-2)

[19] Ogonda, R. T. (1992), Post-Independence Trends in Development of Transport and Communications, in Ochieng' W. R. and R. M. Maxon, eds., An Economic History of Kenya, Nairobi: East African Publishers, pp. 313-326.

[20] Republic of Kenya (2003 and 2004) Legal Notices Nos. 161, 83 and 97

[21] Republic of Kenya (2004) Transformation of Road Transport Report, MOTC, Nairobi

[22] Republic of Kenya (2010). Kenya Police Report 2010.

[23] Republic of Kenya (2010). The Constitution of Kenya (27th August (2010). Nairobi: Government Printers

[24] Republic of Kenya (2012). Kenya National Highway Authority (KeNHA): Quality Highways, Better Connections. Feasibility studies and detailed engineering design of the multinational Arusha – Holili/Taveta – Voi road. RESETTLEMENT ACTION PLAN (RAP) FINAL REPORT, October 2012.

[25] Tom M. Mitchell (March 1, 1997), Machine Learning

[26] The World Bank (2007a), India: World Bank Supports National Highway Systems Improvements in Uttar Pradesh and Bihar, Press Release No: 2005/251/SAR Washington, D.C., December 21, 2004, available from: http://www.worldbank.org/ in website publication

[27] The World Bank (2007b), Uganda Signs US$ 107 million Agreement for Road Sector Development, Press Release No: 2005/354/AFR Kampala dated February 23, 2005 available at http://web.worldbank.org/afr/ug

[28] Van der Veer (2009), Data mining for intelligence led policing, van der Veer, H.T. Roos, A. van der Zanden

[29] Vikas.Grover (2009), Review of Current Crime Prediction Techniques, Vikas Grover, Richard Adderley, Max Bramer.

[30] Noemi Dreyer Galvão; Heimar de Fátima Marin (2009), Data mining: a literature review, Técnica de mineración de datos: una revisión de la literature