ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 1, January 2018

The Patrons for Predicting Veracity of Rail Mishaps Using ID3 Algorithm

[1] M.Karthi, [2]R.Priscilla, [3]E.Benila
[1] Assistant professor, [2] Professor, [3] UG Scholar
[1][2][3] St.Joseph's Institute of technology, Shollinganallur, Chennai-600 119, Tamil Nadu, India.

Abstract- **Data Mining is the emerging technique for Knowledge Discovery in Databases. Data Mining gains Information by processing the raw data. The main advantage of Data mining is to predict accurate information from huge amount of data and thus paves way for decision making. In this project, the main advantage of data mining process is implemented to predict accurate information from the railway accident datasets provided by the railway department. Over 11 years' railway datasets provided by the railway department are analyzed to predict the cause for major rail accidents and thus contribution is made to the railway department for reducing the accident rates. Here Structured provided by the user as well as Un Structured data from railway datasets are mined using Text Mining technique. High level of veracity in Text Mining is obtained by using ID3 Algorithm. ID3 Algorithm analyzes the Unstructured data and predicts the cause for rail accident in a most predominant manner. Finally Report is generated from the mined information and the generated report is represented graphically and geographically by analyzing the latitude and longitude locations.**

Key words: **railway accident; decision making ;ID3 Algorithm**

I. INTRODUCTION

Railway accidents are very common these days. Intensity of rail accidents is on an upward curve towards highest accident record. Thousands of trains run over the rail track every day and some mishap sometimes cannot be ruled out. But the rate of Rail accidents can be reduced by predicting the major cause for rail accidents. By analyzing the past 11 years data ranging from 2001 to 2012, there were more than 45000 rail accidents and death rate is also tremendously increasing due to rail accidents. The Federal Railroad Administration (FRA) has collected data to understand and find ways to reduce the numbers and severity of these accidents. The FRA has set "an ultimate goal of zero tolerance for rail-related accidents, injuries, and fatalities" [15].A review is done by the FRA and it collected the data regarding a variety of accident types from derailments to Collisions. Most of the accidents are less severe. Because they cause little damage and no injuries. The prediction is made to understand the characteristics of these accidents that improve safety so that human lives can be saved. After each accident a report is completed and submitted to the FRA by the railroad companies. This report has a number of fields that include characteristics of the train, the Operators on the trains, track types involved, type of the train, the environmental impacts, operational conditions, and the primary cause of the accident. Additionally, the accident reports contain a narrative which provides a unstructured description of the accident. These unstructured data contain more description about the causes and contributors to the accidents and their circumstances. The FRA uses all of these data much as the Federal Aviation Administration uses reports on aviation accidents, namely, to "develop hazard elimination and risk reduction programs that focus on preventing railroad injuries and accidents" [15]. Till date, there is no analysis for rail accidents report on large scale to insist safety and policies. This paper does an investigation to predate the cause and effects of the rail accidents. By predicting these factors, intensity of rail accidents can be reduced in high level. Here the prediction is done by implementing Mining concepts. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. The ability of Data Mining is to uncover hidden patterns and relationships in data that can be used to make accurate predictions. In this project, structured data and unstructured data is analyzed to understand the accurate characteristics of rail accidents over years. Text mining technique can be used to predict the accurate cause for rail accidents and Railway Department can concentrate on highly predicted reason behind major rail accidents thus accident rates can be reduced. Text mining is the process of examining large collections of written resources in order to generate new information. It is a subset of the larger field of data mining. Text mining is referred as text analytics is used to make "unstructured" data usable by a computer. To implement Text Mining technique ID3 Algorithm is used find highly predicted words from the unstructured text document and by analyzing these datasets submitted by railway administrations. In [16], the work is concentrate on regression. Based on algorithm proposed the data classification are made in this work.

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 1, January 2018

II. RELATED WORK

This section explains the various researches related to the Accident data report.

In [6] Donald E. Brown describes the use of text mining with a combination of techniques to automatically discover accident characteristics that can inform a better understanding of the contributors to the accidents. The study evaluates the efficacy of text mining of accident narratives by assessing predictive performance for the costs of extreme accidents. The results show that predictive accuracy for accident costs significantly improves through the use of features found by text mining and predictive accuracy further improves through the use of modern ensemble methods. Importantly, this study also shows through case examples how the finding from text mining of the narratives can improve understanding of the contributors to rail accidents in ways not possible through only fixed field analysis of the accident reports. In this work RFA Algorithm is used to predict the cause for rail accidents which lacks in terms of accuracy and spatial temporal representation is not made in this study.

A neural network (NN) model to predict intersection crashes based upon driver, vehicle and roadway surface characteristics are discussed in [1]. it describes the use of neural networks to model intersection crashes and intersection characteristics, such as, lighting, surface materials, etc. Taken together these papers show the use data mining to better understand the factors that can influence and improve safety at rail crossings. These studies aimed at uncovering important determinants of accident frequency. By studying the relationship of accident types with weather, roadway, driver and vehicle characteristics, the research offers insight into potential measures to counter the adverse effects of road conditions, road users and vehicle types on highway sections with proper geometrics and roadway elements which can provide safer transportation. The modeling results are obtained using only a single year data belonging to year 1995and future research can be done by expanding the datasets and by setting up some random variables in the design parameters, it may be possible to predict types of vehicle crashes based on various crash properties for a year in the future works.

In this paper Dong.H[7] introduces a basic framework of parallel control and management (PCM) for emergency response of urban rail transportation systems. This work elaborates three interdependent aspects: Points, Lines, and Networks. Points represent the modeling of urban rail stations, Lines describe the microscopic characteristics of urban rail connections between designated stations, and Networks present the macroscopic properties of all the urban rail connections. Furthermore, the constructed artificial system can be used to test and develop effective emergency control and management strategies for real rail transport systems. Therefore, work enhanced the reliability, security, robustness of urban rail transport systems in case of an emergency.

In this paper, Wang.X[13]presents a new modeling process that combines two of recently developed approaches for model in criminal incidents. The first component to the process is the spatio-temporal generalized additive model (STGAM), which predicts the probability of criminal activity at a given location and time using a feature-based approach. The second component involves textual analysis. In this work, Twitter posts are automatically analyzed, which provide a rich, event-based context for criminal incidents. In addition, a new feature selection method to identify important features is described. In this SRL-LDA model, textual information describing an event's spatial and temporal location are ignored. This information could be used to map tweets to particular spatio-temporal grid locations. This would improve the model's ability to identify textual information that correlates with spatio-temporal criminal incident patterns.

In H.Gonzalez[8] proposed a multidimensional mining framework. It can be used to identify a concise set of anomalies from massive traffic monitoring data and then overlay, contrast, and explore such anomalies in multidimensional space. This work is based on the development of two novel methods: (1) efficient anomaly mining stemming from the discovery of the atypical fragment and (2) a multidimensional anomaly overlay model that enables the clustering of multiple atypical fragments according to different criteria. The atypical fragment provides a concise, global view of the traffic anomaly situation, whereas the framework for anomaly overlay provides the power of online analytical processing to facilitate the discovery of patterns associated with different anomaly types and the navigation of anomalies at multilevel abstraction.

III. PROBLEM STATEMENT

The Fig1 (Architecture Diagram) describes the overall structure of the system and complete overview is obtained from its view.

Architecture diagram describes overall view of the proposed system; here the user will submit the report in the form of

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 1, January 2018

structured or unstructured data. This report has a number of fields that include characteristics of the train or trains, the personnel on the trains, the environmental conditions (temperature and precipitation), operational conditions (speed at the time of accident, highest speed before the accident, number of cars, type of track, type of train and weight), and the primary cause of the accident. Unstructured data will be in the form of narratives which includes the information that is provided by the Federal Rail Administration. Once the data is submitted successfully the submitted reports are stored in the database. Then those datasets are mined and retrieved using text mining and the retrieved information are consolidated in the form of reports and the results are represented graphically using Radar Charts, Bar charts, Line charts and Pie chart for better enhancements. Furthermore, the results of this work are geographically represented using Google Maps. By analyzing the latitude and longitude position in the accident reports, the exact location where most of the accidents take place is represented for better understanding.

A.ID3 ALGORITHM

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from the dataset. To model the classification process, a tree is constructed using the decision tree technique. Once a tree is built, it is applied to each tuple in the database and these results in classification for that tuple. Based on [16], the various algorithms are for classification is discussed and finally the ID3 is chosen for this work. The following issues are faced by most decision tree algorithms:

- To choose splitting attributes
- Order of splitting attributes
- Number of splits to be taken
- Balance of tree structure and pruning
- The stopping criteria

The decision tree algorithm is based on Entropy, its main idea is to map all examples to different categories based upon different values of the condition attribute set; its core is to determine the best classification attribute from condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of the current node. Branches can be established based on different values of the attributes and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain and Gain Ratio are used.

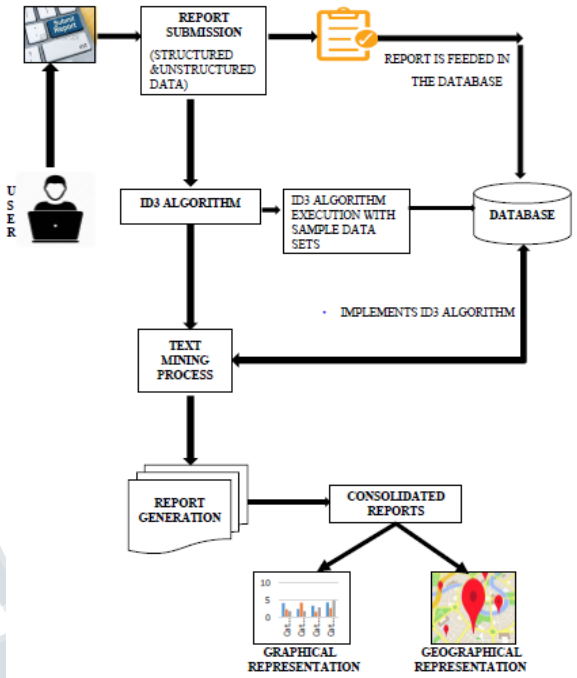

Fig. 1 Process of Finding Rail Mishaps using ID3 algorithm

In this Fig.1 ID3 Algorithm is implemented by using a Sample dataset which consist of Weather Report. The Reason that weather report as a dataset in this work is that weather related factor of rail accidents are mined. To Generate the report for the rail accident the ID3 algorithm validate the following factors involved in rail accident such as Structural Defects, Signal Issues, Miscommunication, Miscellaneous, Operational Factors, Human Factor, Over Speed, Lack of Safety etc. The sample dataset were extracted from the UCI repository for structured data process) and user defined dataset for unstructured data process.

IV EXPERIMENTS AND RESULTS

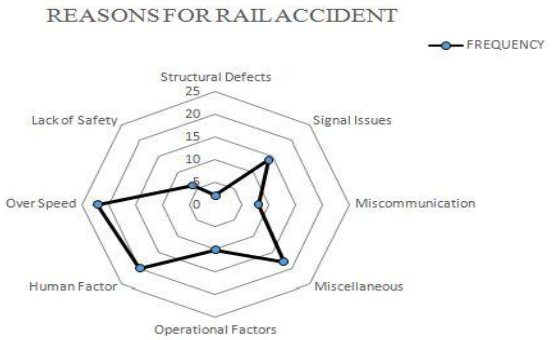

Fig. 2 Reasons for Rail Accident

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 1, January 2018

Fig. 2 shows the reason for rail accident using RFA algorithms approach, and the determinants of accident frequency is shown.

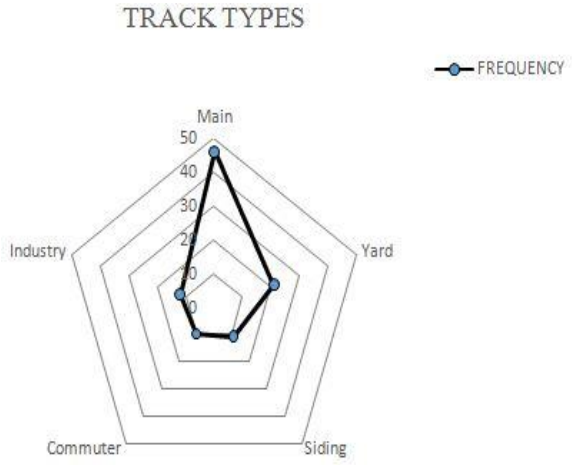

Fig.3 Track Types

Fig. 3 shows the different track types involved in the primary cause of the accident.

Fig.4 Mishap Types

Fig.4 shows the several mishap types involved to predict the cause of the rail accidents.

Table.1 Determinants of accident frequency

CODE	CAUSE	FREQUENCY (%)
1	Structural Defects	2
2	Signal Issues	14
3	Miscommunication	8
4	Miscellaneous	18
5	Operational Factors	10
6	Human Factor	20
7	Over Speed	22
8	Lack of Safety	6

Table.1 shows the different causes of rail accidents. It has been observed from the above results the main factor of the accident is caused by over speed.

Table.2 Accident frequency on Track Types

CODE	TRACK TYPES	FREQUENCY (%)
1	Main	46
2	Yard	21
3	Siding	11
4	Commuter	10
5	Industry	12

Table.2 shows the accident frequency depend on the different Track types. It has been observed from the above results the accident frequency is higher in the main track type.

Table.3 Accident Types with Frequencies

CODE	ACCIDENT TYPES	FREQUENCY (%)
1	Derailment	38
2	Front End Collision	7
3	Back End Collision	8
4	Parallel Collision	5
5	Racking Clash	3
6	Damaged Train Crash	2
7	Expressway Rail Collision	4
8	Outburst Eruption	18
9	Variant Impacts	15

Table.3 is used to analyze frequency of the accident types. The frequency percentage is higher in the Derailment accident type due to over speed.

Table.4 Accident occurring time

CODE	TRACK TYPES	FREQUENCY
1	Day	74
2	Night	26

Table.4 shows the time of occurrence of accidents. The frequency percentage is higher in the day time which proves the accidents are taken more in the day time.

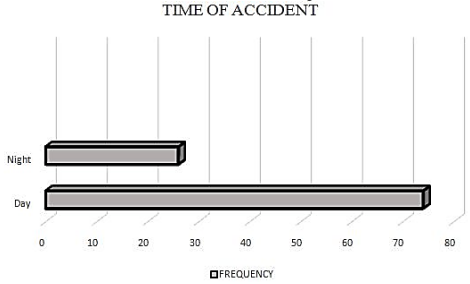

Fig.5 Time of Accident

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)
Vol 3, Issue 1, January 2018

Fig.5 depicts the more number of accidents occurring in the day time.

Table.5 Accident frequency on Train Types

CODE	TRAIN TYPES	FREQUENCY
1	Metro	8
2	Loader	6
3	Passenger	27
4	Freight	12
5	Single Car	3
6	Industry	7
7	Yard	1
8	Locomotive	22
9	Maintenance	4
10	Flat Cars	2
11	Commuter	3
12	Container	5

Table.5 is used to analyze frequency of the accident captivating in the train types. The higher range of frequency in the passenger type proves that more number of accidents involved in the passenger trains.

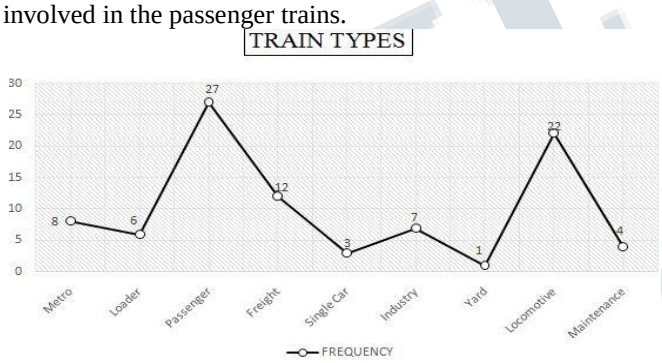

Fig.6 Train Types

Fig.6 depicts the more number of accidents occurring in the passenger train types.

Table.6 Accident related to weather condition

CODE	Weather Condition	FREQUENCY
1	Clear & Sunny	40
2	Windy	12
3	Cloudy	32
4	Rainy	7
5	Snowy	4
6	Fog	5

Table.6 Rail accidents are influenced by the seasonal effects of weather. The frequency percentage is higher in the clear & sunny weather condition which proves the accidents are taken more in the clear & sunny seasonal effects.

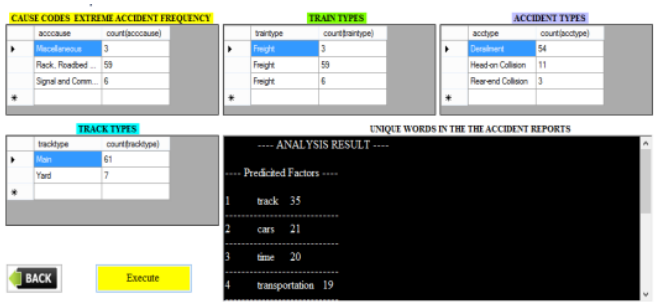

Fig.7 Analysis Report

In Fig.7 ID3 Analysis table is represented in which highly predicted cause for Rail accidents are analyzed by comparing the trained data and analyzed data are represented as unique words in result console.

V. CONCLUSION

This paper shows that the combination of text analysis with ensemble methods to improve the accuracy of models for predicting accident severity and that text analysis can provide insights into accident characteristics. Modern text analysis methods make the narratives in the accident reports almost as accessible for detailed analysis as the fixed fields in the reports. More importantly as the examples illustrated, text mining of the semi-structured data in the form of narratives can provide a much richer amount of information than is possible in the structured data as fixed fields. Finally, as described in the paper standard methods to clean the narratives and the required information is extracted from the submitted unstructured data. For better understanding purpose of the text mining concept, the report is generated graphically and geographically. For train safety analysis, text mining could benefit from a careful look at ways to extract features from text that takes advantage of language characteristics particular to the rail transport industry for safety analysis.

VI. FUTURE ENHANCEMENT

The work described in this paper only focused on incidents with extreme accident damage and not less severe accidents. Hence future work is needed for accidents with extreme numbers of casualties to determine their contributors and the similarities and differences of these contributors to those of accidents with extreme costs. In this paper contribution is done only for railways and not for other modes of transport, so these methodologies can also be implemented for other

modes of transport too. There are also several areas of future work that will provide more fundamental advances in the use of text mining for train safety engineering.

.

REFERENCES

[1] Akin.D and B.Akbas, "A neural network (NN) model topredict intersection crashes based upon driver, vehicle and roadway surface characteristics," Sci. Res. Essays, vol. 5, pp. 2837–2847, 2010.

[2] Burgoon.J et al., "Detecting concealment of intent in transportation screening: A proof of concept," IEEE Trans. Intell. Transp. Syst., vol. 10, no. 1, pp. 103–112, Mar. 2009.

[3] Cao.J et al., "Web-based traffic sentiment analysis: Methods and applications," IEEE Trans. Intell. Transp. Syst., vol. 15, no. 2, pp. 844–853, Apr. 2014.

[4] Cirovic.G and D. Pamucar, "Decision support model for prioritizing railway level crossings for safety improvements: Application of the adaptive neuro-fuzzy system," Expert Syst. Appl., vol. 40, pp. 2208–2223, 2013.

[5] D'Andrea.E, P. Ducange, B. Lazzerini, and F. Marcelloni, "," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 4, pp. 2269–2283, Mar. 2015.

[6] Donald E. Brown, Fellow, "text mining the contributors to rail accident's" IEEE Trans. Intell. Transp. Syst., vol. 17, no. 2, february 2016.

[7] Dong.H et al., "Emergency management of urban rail transportation based on parallel systems," IEEETrans.Intell.Transp.Syst.,vol.14, no.2, pp. 627–636, Jun. 2012.

[8] Gonzalez.H, J. Han, Y. Ouyang, and S. Seith, "Multidimensional data mining of traffic anomalies on large-scale road networks," Transp. Res. Rec., vol. 2215, pp. 75–84, 2011.

[9] Meyers.T, A. Stambouli, K. McClure, and D. Brod, "Risk assessment of positive train control by using simulation of rare events," Transp. Res. Rec., vol. 2289, pp. 34–41, 2012.

[10] Nayak et al.,"Application of text mining in analysing road crashes for road asset management," in Proc. 4th World Confr. Eng. Asset Manage., Athens, Greece, pp. 49–58, Sep. 2009.

[11] Tey.S, G. Wallis, S. Cloete, and L. Ferreira, "Modelling driver behaviour towards innovative warning devices at Real-time detection of traffic from Twitter stream analysis railway level crossings," Neural Comput. Appl., vol. 51, pp. 104–111, Mar. 2013.

[12] Taddy.M, "Multinomial inverse regression for text analysis," J. Amer. Statist. Assoc., vol. 108, no. 503, 2012.

[13] .X and D. E. Brown, "The spatio-temporal modeling for criminal incidents," Security Inf., vol. 1, no. 2, pp. 1–17, Feb. 2012.

[14] "Positive train control(PTC)," Federal Railroad Admin(RFA)., Washington, DC, USA, 2012.

[15] "Railroad safety statistics—2009 Annual report—Final," Federal Railroad Admin., Washington, DC, USA, Apr. 2011.

[16] Glandence, L. Mary, M. Karthi, and V. Maria Anu. "A Statistical Comparison of Logistic Regression and different Bayes Classification methods for machine Learning." ARPN Journal of Engineering and Applied Sciences 10.14 (2015): 5947-5953.