



THE PATRONS FOR ANTICIPATING VERACITY OF RAIL MISHAPS USING TEXT MINING and ID3 ALGORITHM

M.Karthi¹, R.Priscilla², E.Benila³

¹Assistant professor, ²Professor ³UG Scholar,

^{1,2,3}St.Joseph's Institute of technology, Shollinganallur, Chennai-600 119, Tamil Nadu, India.

¹mmuthukarthi@gmail.com, ²prisci.christa@gmail.com, ³benilaedwin@gmail.com

ABSTRACT

Data Mining is an emerging technique for Knowledge Discovery in Databases. Data Mining processes the raw data and it converts the raw data into a useful information. The Accuracy of the information mined is the major advantage of Data mining and the accurately mined information is used for decision-making purposes. 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 of major rail accidents and thus contribution is made to the railway department for reducing the accident rates. Here Structured data 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 of 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.

Keyword :railway accident; decision making ;ID3 Algorithm

I. INTRODUCTION

Railway accidents are very common these days. The intensity of rail accidents is increasing day by day and it is on an upward curve towards highest accident record. Each and every day thousands of trains run over the rail track and some mis-happenings cannot be completely stopped. But the rate of Rail accidents can be reduced by predicting the major cause of rail accidents. From 2001 to 2012, a wide range of analysis is done over railway datasets and it is observed that there were more than 44000 rail accidents and the death rate is also increasing tremendously due to these rail accidents. The Federal Railroad Administration (FRA) has collected data to analyze, understand the problem and to provide a better way to reduce the count and severity of these accidents. The FRA has set "an ultimate goal of zero tolerance for rail-related accidents, injuries, and fatalities" [15]. FRA conducted a review by collecting the variety of data regarding the types of rail accidents ranging from collisions to derailment. By analyzing, most of the accidents were found to be less severe as they did not cause extreme damage or injuries. To understand the characteristics of these accidents and to improve safety, a prediction is made to improve safety so that human lives can be saved. FRA collects a report from the railroad companies after each accident which includes a number of fields and characteristics of the train such as train operators, track types, train type, environmental constraints, operational constraints and the root cause of the accident. In addition to this, these reports contain a narrative section which includes the unstructured description of the mis-happening. This narrative section will contain information such as causes of the accident, contributors to the accident and circumstances of the accident. FRA exploits information from all these structured and unstructured data to "eliminate hazards, reduce risks" that prevents railroad injuries and mis-happenings [15]. Till date, rail accident reports are not

analyzed on large scale to improve safety measures. This paper does analysis and an investigation is done to predict the cause and effects of the rail mis-happenings. The intensity of rail accidents can be reduced at a high level by predicting these factors. Here the prediction is done by implementing Mining concepts. Data mining sorts large data sets for pattern analysis and it relates the datasets to make accurate decisions. 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 generates new information by examining a large collection of text resources. Text Mining is a text analytics technique in which unstructured data is used by the computer to mine information. Text Mining technique is implemented using ID3 Algorithm to find highly predicted words from the unstructured text document by analyzing the datasets submitted by railway administrations. In [16], the work is concentrated on regression. On the basis of the algorithm proposed, the data classifications are made in this work.

II. RELATED WORK

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

In [6] Donald E. Brown describes the applications of text mining by combining techniques to automatically discover characteristics of the mis-happenings that can provide a better understanding of the contributors to the mis-happenings. This paper evaluates the efficiency of text mining and costs of the extreme accidents are predicted with accuracy by accessing the accident narratives. The result shows the

predictive accuracy and it is observed that 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. This work shows an example where the process of text mining in the narrative section helps the contributors to better understand the characteristics of rail mis-happenings than that of fixed field analysis of the mis-happening reports. In this work, RFA Algorithm is used to predict the cause for rail accidents where accuracy is not specific and spatial temporal representation is not made in this study.

In [1], Intersection crashes due to driver, vehicle, roadway surface characteristics are predicted using neural network model. It explains the advantages of neural networks to model intersection crashes and its characteristics, such as lighting, surface materials, etc. This paper shows the use of data mining to understand the factors influencing and improving safety at rail level crossings. This study is aimed at uncovering important determinants of accident frequency. While relating the mis-happening types with climatic conditions, roadway, driver and vehicle characteristics, the research reports give a perception that probable measures to counter the worse effects of road conditions, road drivers and vehicle types on highway sections with proper geometrics and roadway elements can provide secure transportation. The modeling results are obtained from a single year data belonging to the year 1995 and 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 rail transportation systems in the urban area. Here three interdependent aspects are elaborated: Points, Lines, and Networks. Points represent the urban rail stations modeling, the microscopic characteristics of urban rail connections between designated stations are represented using lines, and the macroscopic properties of all the urban rail connections are established using networks. An artificial system is constructed to test and develop effective emergency control and management strategies for real rail transport systems are constructed. Therefore, this work enhanced reliability, security, the robustness of urban rail transport systems in case of an emergency.

In this paper, Wang, X [13] Proposed a new modelling process where two recently developed approaches are combined in order to develop a model for criminal accidents. The first approach of the process involves the Spatio-Temporal Generalized Additive Model (STGAM) in which probability of criminal action at a particular location and the corresponding time is predicted using feature-based approach. Text analysis is done in the second approach. Here twitter datasets are analyzed and its posts are automatically analyzed periodically to gain rich, event-based context from criminal incidents. Additionally, important features are identified using a new feature selection method. Here event's spatial and temporal content are ignored from textual information by the SRL-LDA model. But this information is

used to map tweets to a specific spatio-temporal grid location. This process is related to spatio-temporal criminal happening patterns and it improves model's ability to identify textual information.

In this paper [8], H. Gonzalez proposed a framework for multidimensional mining. This framework identifies a brief set of deviations from the traffic monitoring data and then it monitors the data, contrasts and explores deviations in a multi-dimensional space. In this work, two novel methods are developed. First is an efficient anomaly mining stemming from the discovery of the atypical fragment and the second is a multidimensional anomaly overlay model that enables the clustering of multiple atypical fragments according to multiple criteria. Brief, global view of the anomaly situation is provided by the atypical fragment and power of online analytical processing for discovering patterns from different deviation types, anomaly navigation at multilevel abstractions are provided by the anomaly overlay framework.

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 an overall view of the proposed system; here the user will submit the report in the form of 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) algorithm is used to generate a decision tree from the provided railway dataset. For modelling the classification process, a tree is constructed using the decision tree technique. When a tree is constructed, modifications are applied to each row in the railway database and it results in the classification for that particular row. Based on [16], various algorithms based on classification is discussed, analyzed and finally, the ID3 algorithm is chosen for this work. The following issues are faced by most of the decision tree algorithms:

- Choosing splitting attributes
- Order of splitting attributes
- Number of splits to be taken
- Balancing of tree structure and pruning
- The stopping criteria

Entropy is the major concept in this decision tree algorithm as it maps all the examples with various categories on the basis of different values of the conditional attribute set. The Main objective of entropy is to identify the best attribute set from various classifications. Here Information Gain is used as the attribute selection criteria. An attribute with highest Information Gain is selected as the splitting attribute of the current node. Based on the different values of the attributes, Branches are established and it is processed recursively on each branch to create further nodes until all the samples in a branch are categorized to a similar category. Concepts of Entropy, Information Gain, Gain Ratio is used for selecting the splitting attributes.

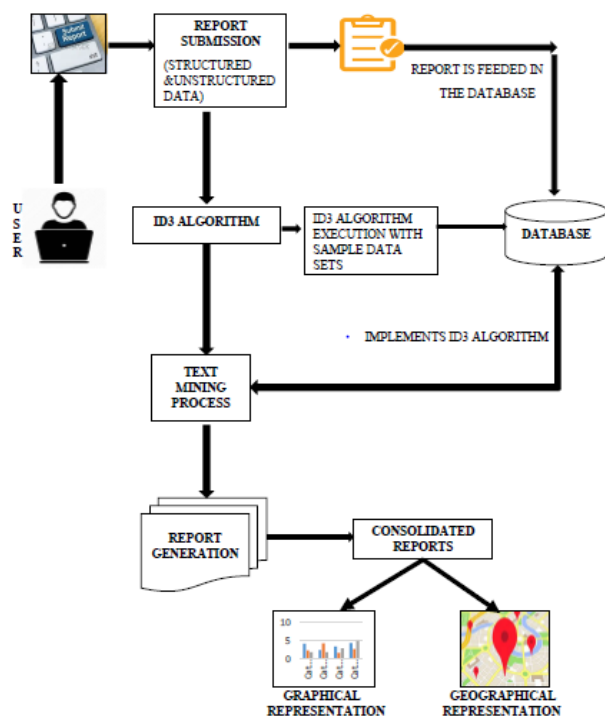


Fig.1 Process of Finding Rail Mishaps using ID3 algorithm

In this Fig.1 ID3 Algorithm is implemented by using a Sample dataset which consists of Weather Report. The Reason that weather report as a dataset in this work is that weather related factor of rail accidents are mined.

IV EXPERIMENTS AND RESULTS

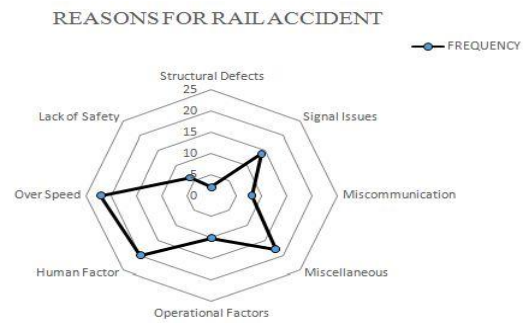


Fig. 2 Reasons for Rail Accident

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

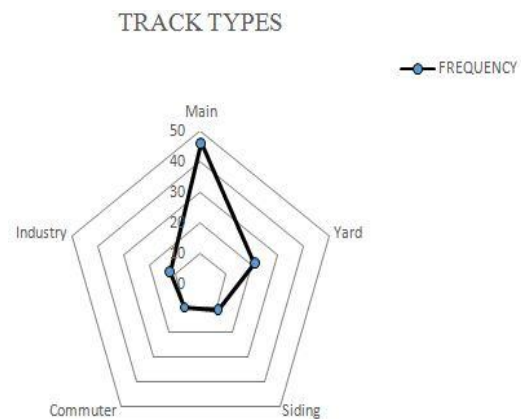


Fig.3 Track Types

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

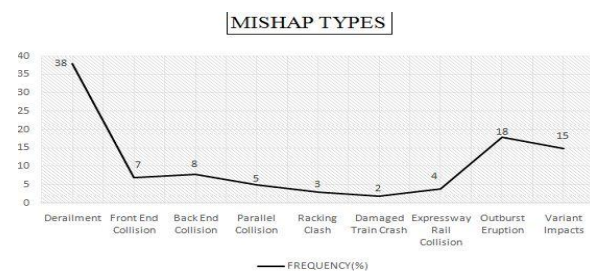


Fig.4 Mishap Types

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.

TIME OF ACCIDENT

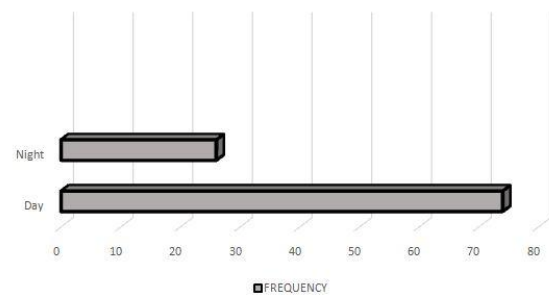


Fig.5 Time of Accident

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 capturing in the train types. The higher range of frequency in the passenger type proves that more number of accidents involved in the passenger trains.

TRAIN TYPES

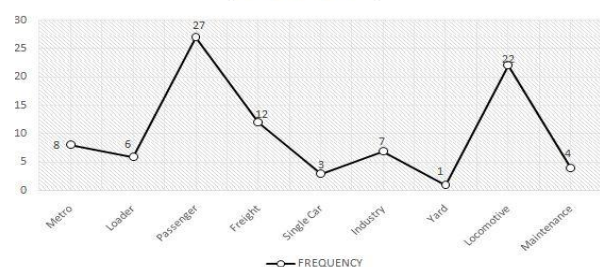


Fig.6 Train Types

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

Table.7 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.

V. CONCLUSION

This paper shows that the combination of text analysis with ensemble method improves the accuracy of models which helps in predicting the severity of the accident and the text analysis provides an intuition into accident characteristics. Modern text analysis methods mines information from the narrative section of the accident report for detailed analysis of the fixed fields in the reports. From the illustrated example, it is inferred that the information gained by mining the narrative semi-structured section is more accurate than mining information from structured fixed field analysis. For Railway safety analysis, Text mining contributes its part to effectively predict the accurate cause for rail mis-happenings and it benefits by extracting important features from the text related to railway transport industry for ensuring safety.

VI. FUTURE ENHANCEMENT

The work described in this paper only focused on incidents of 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, a 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.

To Generate the report for the rail accident the ID3 algorithm validates 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 dataset was extracted from the UCI repository.

REFERENCES

- [1] Akin.D and B.Akbas, "A neural network (NN) model to predict 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," *IEEE Trans. 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] Gladence, L. Mary, M. Karthi, and T. Ravi. "A Novel Technique for Multi-Class Ordinal Regression-APDC." *Indian Journal of Science and Technology* 9.10 (2016).
- [17] Glandence, L. Mary, M. Karthi, and V. Maria Anu. "A Statistical Comparison of Logistic Regression and different Bayes Classification methods for machine Learning." *ARN Journal of Engineering and Applied Sciences* 10.14 (2015): 5947–5953.
- [18] R.Joseph Manoj, M.D.Anto Praveena, K.Vijayakumar, "An ACO-ANN based feature selection algorithm for big data", *Cluster Computing The Journal of Networks, Software Tools and Applications*, ISSN: 1386-7857 (Print), 1573-7543 (Online) DOI: 10.1007/s10586-018-2550-z, 2018.
- [19] K. Vijayakumar,C.Arun,Automated risk identification using NLP in cloud based development environments,*J Ambient Intell Human Computing*,DOI 10.1007/s12652-017-0503-7, Springer May 2017
- [20] K.Vijayakumar-C.Arun,Continuous security assessment of cloud based applications using distributed hashing algorithm in SDLC,Cluster Computing DOI 10.1007/s10586-017-1176-x,Sept 2017
- [21] K. Vijayakumar and C. Arun, "Continuous Security Assessment of Applications in Cloud Environment", *International Journal of Control Theory and Applications*, ISSN: 0974-5645 volume No. 9(36), Sep 2016, Page No. 533-541.
- [22] K. Vijayakumar, N. Divya Sri, M. Vijayashree, "An Effective User Revocation and Anti Collusion System for Dynamic Groups in Cloud", *International Journal for Research in Applied Science & Engineering Technology*, ISSN: 2321-9653, Volume 4 Issue V, May 2016
- [23] Vijayakumar .K, Aravind S, Ullas Kumar S "Preventing data loss and increasing data security using file instance sharing", *International Journal Of Advanced Research in Biology, Engineering, Science and Technology (JARBEST)*, volume 2, special issue 10, May 2016, Page No. 144-151

