International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 02
February 2017

A Review on Rail Accidents and Predictions Using Data Mining Techniques

S. L. Shalini
PG Scholar
Department of CSE,
DIET, ANAKAPALLE, Visakhapatnam

A. A. Narasimham
Associate Professor,
Department of CSE,
DIET, ANAKAPALLE, Visakhapatnam

Abstract- *The costs of fatalities and injuries from train accidents have a great impact on society. As part of our effort to understand the characteristics of past train accidents, this paper presents an analysis of significant train accidents occurring in around world from 2000 to 2015. Rough set theory and associated rules approaches are applied in analysing the collected data. The results show that although most derived rules are unique, some rules are worth noting. Collision accidents generally lead to more casualties than derailment accidents, and the most frequent cause of accidents is human error. Additionally, most*

Train accidents occur during summer. These findings can provide railway leaders with lessons and rules learned from past accidents, thus facilitating the establishment of a safer railway operation environment around world. Accident investigation and analysis are key to reinforcing and improving railway safety. Many railway accidents have been caused by degraded human performance and human error, and the tasks of train drivers and signallers have remained essentially the same. Although new technologies and equipment have gradually reduced railway operation accidents, no investigation has been conducted to investigate whether railway performance shaping factors (R-PSFs), attributed to degraded human performance, have changed or remained constant. 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 findings 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.

.

Keywords: Rail safety, safety engineering, latent Dirichlet Allocation, partial least squares, random forests.

1. INTRODUCTION

For a long time, ministry of railways in world is a relative monopoly agency. When an accident happened, the ministry of railways would form an investigation team to investigate; the team would release the investigation report on the accident ultimately. But the investigation report only could only be seen by the railway staff, there was no way of knowing for the outside world basically. After the reform of the ministry of railways in 2013, this situation has changed. In this year, world's railway system released annual railway traffic fatality statistics publicly at first time, 1336 people lost their life due to train accidents. In 2014, this figure reached 1232. On the whole, the number of deaths declined in 2014, but the outlook for railway safety in world is still grim.The railway system is a major component of the economy of most countries, daily transporting millions of passengers as well as millions of dollars' worth of goods from origin to destination (1).

Therefore, the relevant operational, regulatory, and governmental bodies of every country with a rail network aim for a safe, highly reliable, and excellent quality railway system (2). In the United Kingdom in particular, railways have played a substantial role in society's daily life and economy since the 1820s. Their crucial role is apparent in a recent Rail Safety and Standards Board Ltd. (RSSB) annual safety performance report (3), which states that for 2013 to 2014 about 1.59 billion passenger journeys, 60.1 billion passenger kilometers, and 48.5 million freight train kilometers were recorded.This paper describes an investigation to understand the possible predictors or contributors to accidents obtained from

"mining" the narrative text in rail accident reports. To do this the approach integrates a combination of analytical methods to first identify the accidents of interest and then look for relationships in the structured and unstructured data that may suggest contributors to accidents. This study evaluates the efficacy of the features found from text mining using models containing these features to predict the costs of extreme accidents. In performing this evaluation the study also considers the usefulness of modern ensemble approaches incorporating these text-mined features to predict accident costs. Finally, the study teases apart the text-mined features, whose importance is confirmed by predictive accuracy, for their insights into the contributors to rail accidents. The purpose of this final analysis is to understand the insights for rail safety that text mining can provide to the exclusion of fixed field reports.

These studies revealed some interesting results, however, they are unable to properly analyse the cognitive aspects of the causes of the crashes. They often opt to leave out significant qualitative and textual information from data sets as it is difficult to create meaningful observations. The consequence of textual ignorance results in a limited analysis whereby less substantial conclusions are made. Text mining methods attempt to bridge this gap. Text Mining is discovery of new, previously unknown information, by automatically extracting it from different written (text) resources. Text mining methods are able to extract important concepts and emerging themes from the collection of text sources. Used in a practical situation, the possibilities for knowledge discovery through the use of text mining is immense. To our knowledge, there is limited or no reputable studies that have utilised text mining in this data domain, however, earlier studies in the field indicate a real need for textual mining in order to better understand the contextual relationships of road crash data.

2. Literature review

Railway accidents cause numerous casualties every year, thereby attracting the interest of many researchers and analysts. Existing research on train accidents is empirically and methodologically diverse. From an empirical standpoint, most studies have attempted to identify the risk factors that influence the severity of train accident casualties. Many of these studies report that train accidents often result from a chain or sequence of events as opposed to a single cause.3,5,6 For example, Reinach and Viale7 analyzed six train accidents and developed a human error analysis and classification model with 36 probable contributing factors. The results demonstrate that each accident was associated with multiple contributing factors. Related studies in other countries have reported similar results. Although these models can be used to assign blame for accidents, they are usually ineffective in preventing future ones. Additionally, numerous studies have focused on the factors influencing the severity of accidents. Ilkjær and Lind8 analyzed a railway accident that occurred in Denmark in 1994 and concluded that carriage interior (the layout of the seat, is there an armrest on the seat, do the luggage place safe or not and so on) has a major influence on personal injuries. Mirabadi and Sharifian3 investigated Iranian Railways accident data from 1996 to 2005 and concluded that human error, wagon and track contribute most often to increased accident severity. Evans9 analyzed fatal train accidents on Europe's main line railways from 1980 to 2009. The results show that the most common immediate causes of serious accidents at railroad crossings are errors or violations by railroad users. Other studies have concentrated on developing computer-controlled systems to prevent the train accidents.1,10,11 From a methodological standpoint, researchers have applied a variety of statistical approaches to analyze the train accident rates and trends in various countries. Evans9 estimated the overall trend in the number of fatal train collisions and derailments per train-kilometer to be 6.3% per year from 1990 to 2009 in Europe, with a 95% confidence interval. However, there are statistically significant differences among different European countries in the rate of fatal train accidents. Evans10 analyzed data on almost all fatal railway accidents in the United Kingdom from 1967 to 2003, found downward trends in all the main classes of accidents per train kilometre in the 27 years leading up to 1993. In addition, nonparametric statistical methods have been widely used in analyzing train accidents. Chong et al.12 applied artificial neural networks and decision trees to model the severity of injury resulting from train accidents. In all cases, the decision trees outperformed the neural networks. Yan et al.13 applied hierarchical tree-based regression to explore the train-vehicle crash prediction and analysis at passive highway-rail grade crossings and concluded that installing stop signs can effectively improve safety at these crossings. Liu et al.14 developed two regression models provided a better understanding of train derailment severity distribution. Knowledge discovery and data mining techniques have also been applied in many recent explorations of this topic. Wong and Chung15 explored accident occurrence in Taiwan using rough set theory, a technique that has been used to make decisions in the presence of uncertainty and vagueness.16,17 Mirabadi and

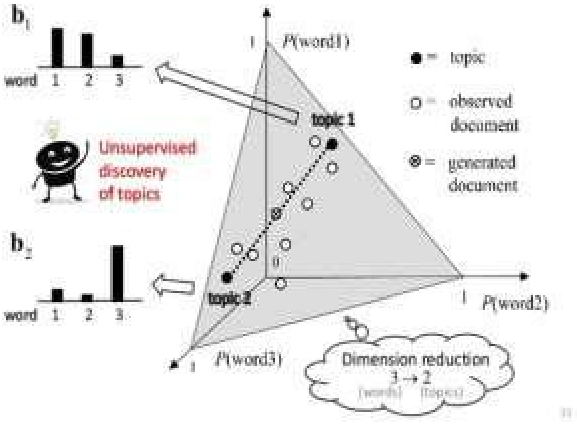
Sharifian3 applied associated rules techniques to analyze the data from accidents on Iranian railways in order to reveal previously unknown relationships among the data. The patterns extracted through these methods can be utilized to develop regulations and rules to help prevent similar accidents from occurring in the future.

2.1 Current Limitations

The current design The analysis in this study is limited to mining the causal text relating to 'Groundings', 'Collisions', 'Machinery Failures' and 'Fire' related accidents. The scope of the study has also been limited by focusing only on pattern classification and connectives methods for extracting the causal relations to keep the study to a reasonable size. There are quite a few challenges when dealing with accident investigation reports. The reports are written in the natural language with no standard template. Misspellings and abbreviations are often found. Detection of compound words such as "safety culture", "spirit status", etc are difficult as order of importance is unknown. The contextual meaning of the words "safety" and "culture" differs significantly but the word "safety culture" has a different meaning altogether. Therefore, context and semantics play an important role in text mining. To date, they have not reported large scale analysis of the narratives for information that could inform safety policies and design. They focused on retrieval not prediction.

3. Related Work

This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data and text mining techniques. This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data and text mining techniques. Based on the collected data, this paper proceeds to establish a rail accident decision table. To the author's knowledge, this paper is the first such attempt to analyse Chinese train accidents. Our approach consists of two stages. In the first stage, we use rough set theory as a tool for data pre-processing, to remove redundant knowledge from established information systems and to provide the means to deal with missing data. In the second stage, we adopt associated rules analysis to propose rules based on the past accidents' data. These rules can uncover unknown relationships that can be the basis of forecast and decision. The number of published reports was not constant over time. With continuous technological improvements such as installation of the Train Protection and Warning System, the number of accidents and significant accidents in recent years has dropped significantly and so fewer cases were extracted for the period of 1997 to 2012. To achieve a more accurate distribution of the number of the R-PSFs for the whole period of interest, the proportion of operator errors in the 467 reports was intended to be similar to the number of errors of the final set of 237 reports. Text mining is concerned with finding patterns in unstructured text. This field has become increasingly important because of the large amounts of data available in documents, news articles, research papers, and accident reports. In many cases text databases are semi structured because in addition to the free text they also contain structured fields that have the titles, authors, dates, and other meta data. The accident reports used in this paper are semi structured. One of the key goals of text mining is to characterize the contents of the documents through pattern discovery. These patterns may then be used for improved information retrieval or, as in this paper, for input into predictive models. Regardless of the ultimate goal, most text mining begins with vector space models where documents are represented by term-document matrices. These matrices have terms as headers for the rows and documents as headers for the columns. The values in the cells give the count or frequencies of a term (row) in a document (column).



4.1 Generate Accident Report

This paper integrates methods for safety analysis with accident report data and text mining to uncover contributors to rail accidents. This section describes related work in rail and, more generally, transportation safety and also introduces the relevant data and text mining techniques.

4.2 Characteristics of Accident Report

This report has a number of fields that include characteristics of the train or trains, the personnel on the trains operational conditions (e.g., speed at the time of accident, highest speed before the accident, number of cars, and weight), and the primary cause of the accident.

This field has become increasingly important because of the large amounts of data available in documents, news articles, research papers, and accident reports.

4.3 Stored In Databases

Text databases are semi structured because in addition to the free text they also contain structured fields that have the titles, authors, dates, and other Meta data. The accident reports used in this paper are semi structured.

5 CONCLUSION

In this Paper, show that the combination of text analysis with ensemble methods can 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 narratives can provide a much richer amount of information than is possible in the fixed fields. Finally, as described in the work here used standard methods to clean the narratives. However, train accident narratives use jargon common to the rail transport industry and classical stemming and stop word removal do not necessarily do a good job of characterizing the words used in this industry. 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.A second of fundamental research need is to characterize the variation and uncertainty inherent in text mining techniques. In this study the use of both LDA and PLS did not give consistent results with different training and test set selections.

These differences need to be formally characterized and, ideally, described with a probabilistic model that further enhances understanding of the contributors to accidents. Finally, as described in Section V the work here used standard methods to clean the narratives. However, train accident narratives use jargon common to the rail transport industry and classical stemming and stop word removal do not necessarily do a good job of characterizing the words used in this industry. 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.

References

[1] "Railroad safety statistics—2009 Annual report—Final," Federal Railroad Admin., Washington, DC, USA, Apr. 2011. [Online]. Available: http://safetydata.fra.dot.gov/OfficeofSafety/publicsite/Publications.aspx

[2] "Office of safety analysis," Federal Railroad Administration, Washington, DC, USA, Oct. 2009. [Online]. Available: http://safetydata.fra.dot.gov/officeofsafety/

[3] G. Cirovic 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.

[4] L.-S. Tey, G. Wallis, S. Cloete, and L. Ferreira, "Modelling driver behaviour towards innovative warning devices at railway level crossings," Neural Comput. Appl., vol. 51, pp. 104–111, Mar. 2013.

[5] D. Akin 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.

[6] H. Gonzalez, 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.

[7] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from Twitter stream analysis," IEEE Trans. Intell. Transp. Syst., vol. 16, no. 4, pp. 2269–2283, Mar. 2015.

[8] F. Oliveira-Neto, L. Han, and M. K. Jeong, "An online self-learning algorithm for license plate matching," IEEE Trans. Intell. Transp. Syst., vol. 14, no. 4, pp. 1806–1816, Dec. 2013.

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

[10] J. Burgoon 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.

[11] Y. Zhao, T. H. Xu, and W. Hai-feng, “Text mining based fault diagnosis of vehicle on-board equipment for high speed railway,” in Proc. IEEE 17th Int. Conf. ITSC, Oct. 2014, pp. 900–905.

[12] T. Hofmann, “Probabilistic latent s emantic indexing,” in Proc. 22nd Annu.Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1999, pp. 50–57.

[13] R. Nayak, N. Piyatrapoomi, J. W. R. Nayak, N. Piyatrapoomi, and J. Weligamage, “Application of text mining in analysing road crashes for road asset management,” in Proc. 4th World Congr. Eng. Asset Manage.,

Athens, Greece, Sep. 2009, pp. 49–58.

[14] “Leximancer Pty Ltd.” [Online]. Available: http://info.leximancer.com/ academic

[15] A. E. Smith and M. S. Humphreys, “Evaluation of unsupervised semantic mapping of natural language with Leximancer concept mapping,” Behav. Res. Methods, vol. 38, no. 2, pp. 262–279, 2006.

[16] U.S. Grant, The Personal Memoirs of U.S. Grant., 1885. [Online]. Available: http://www.gutenberg.org/files/4367/4367-pdf/4367-pdf.pdf

[17] W. Jin, R. K. Srihari, H. H. Ho, and X. Wu, “Improving knowledge discovery in document collections through combining text retrieval and link analysis techniques,” in Proc. 7th IEEE Int. Conf. Data Mining, Omaha, NE, USA, Oct. 2007, pp. 193–202.

[18] D. Delen et al., Practical Text Mining and Statistical Analysis for Non- StructuredText DataApplications. Waltham,MA, USA:Academic, 2012.

[19] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA, USA: Wadsworth, 1984.

[20] T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer-Verlag, 2009.

[21] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.

[22] H.Wold, “Estimation of principal components and related models by iterative least squares,” in Multivariate Anal., P.Krishnaiaah, Ed. NewYork, NY, USA: Academic, 1966, pp. 391–420.

[23] L. Li, R. D. Cook, and C. Tsai, “Partial inverse regression,” Biometrika, vol. 94, no. 3, pp. 615–625, Aug. 2007.

[24] M. Taddy, “Multinomial inverse regression for text analysis,” J. Amer. Statist. Assoc., vol. 108, no. 503, 2012. [Online]. Available: http://dx.doi.org/10.1080/01621459.2012.734168

[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J.Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.

[26] M. Steyvers and T. Griffiths, “Probabilistic topic models,” in Handbook of Latent Semantic Analysis, vol. 427. Hillsdale, NJ, USA: Erlbaum, 2007.

[27] D. Blei, L. Carin, and D. Dunson, “Probabilistic Topic Models,” IEEE Signal Process. Mag., vol. 27, no. 6, pp. 55–65, Nov. 2010.

AUTHORS

S.L. Shalini
PG Scholar
Department of CSE,
DIET, ANAKAPALLE,
Visakhapatnam

A.A.Narasimham
Associate Professor,
Department of CSE,
DIET, ANAKAPALLE,
Visakhapatnam