International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 3
ISSN: 2454-4248
383 – 386

Text Mining Applied to Rail Accidents

Nayana Kamerkar, Kamlesh Patil, Ankita Kale
Guide-Mrs. Nita Patil

Abstract: Indian Railway is the most popular way of transportation. So rail safety represents an important safety concern for transportation industry. In 6 years period between 2009-2010 and 2014-2015, there were a total of 803 accidents in Indian Railway killing 620 people and injuring 1855 people. 47% of these accidents were due to derailment of trains. To better understand the reason to these extreme accidents, Indian Railway has submitted reports that contain both fixed field and narratives that describe the characteristics of the accident. While a number of studies have looked at the fixed fields, none have done an extensive analysis of the narratives. It 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. It will let the user know about of type of accident which can occur. The analysis can also be used for the safety improvement of Indian Railway.

*****

I. INTRODUCTION:

The Indian Railway network (IRN) is fourth largest railway networks in the world [1], handling massive numbers of passengers and quantities of goods daily. Railways are the most popular means of long-distance transportation in India; hence the IRN is often described as the backbone of this nation's economy.

The present scenario in the transportation sector in India gives further motivation for a detailed analysis of the IRN - it is a commonly voiced opinion among economists that the current transportation network in India is too weak to meet the demands of the country's rapidly growing economy. For instance, factors such as the traffic between major cities exceeding the planned capacity and over-utilized railway tracks are resulting in trains having to travel at reduced speeds and carry lesser amounts of freight, thus increasing the cost and time of transportation.

The IR has long served as the backbone of this nation's economy by being the most popular means of long distance transportation in India. However, the IR is facing several grievous problems in the recent years. More alarmingly, there has been a spate of Railway Accidents in India in the year 2010, leading to loss of a significant number of human lives and frequent disruption of traffic over large regions of the country. Here we consider only those accidents that were caused due to collision among trains or derailment of trains and not due to terrorist activity or natural calamities like fire, floods. According to the Wikipedia page[9] enlisting the major rail-accidents in India, there have been 11 such accidents in 2010 alone as compared to 7 such accidents in the 5-year period of 2005-2009.

Thus, the accident patterns in the IRN have also been studied in this project to understand this repeated occurrence of accidents in a Indian Railway in recent times. Analyzing the current IR traffic as well as the increase in IR traffic over the last two decades, we find that traffic in the specific metropolitan region has increased exorbitantly and it is quite probable that the present amount of traffic has exceeded the allowable safety-limits considering the IR resources (e.g. railway-tracks, signaling systems) available in India.

II. LITERATURE REVIEW:

Moty Ben-Dov ,Wendy Wu ,Ronen Feldman ,Ramat Gan and Paul A. Cairns[2] stated that the availability of online text documents exposes the readers to a vast amount of potentially valuable information buried in those texts. The huge number of documents created the pressing need for automated methods of discovering relevant information without having the need to read it all. Information Extraction (IE) from documents is one of approaches in text mining which extracts the features (entities) from documents. They proposed to use link-analysis techniques over the extracted features for finding new knowledge. In order to use link analysis techniques they need to create links out of the features extracted by the IE process.

G. Nagamallika, A.Anuradh,K.Sridevi [3] analyzed the utilization of content mining with a blend of strategies to naturally find mischance qualities that can advise a superior comprehension of the supporters of the mishaps. The review assesses the viability of content mining of mischance stories by surveying prescient execution for the expenses of outrageous mishaps. The outcomes demonstrate that prescient precision for mishap costs altogether enhances using highlights found by content mining and prescient exactness additionally enhances using present day outfit strategies. Essentially, this review likewise appear through

 ISSN: 2454-4248
383 – 386

case illustrations how the discoveries from content mining of the accounts can enhance comprehension of the supporters of rail mischances in ways impractical through just settled field investigation of the mishap reports.

N.M.Deepika , K. Rashmi ,V Divyavani [4] analyzed that Transportation plays an important role in humans life and now in major cities rail transportation is widely being used by various people most of the studies are not been focused on predicting accidents in rail transportation . A survey shows that since 11 years from 2001 to 2012, the U.S. had more than 40,000 rail accidents which cost more than $45 million. Most of the accidents are very expensive during this period up to $141 500. The Federal Railroad Administration has required the railroads involved in accidents to submit reports that contain both fixed field entries and narratives that describe the characteristics of the accident .So their paper proposes a clear idea regarding safety design and policies in railroad transportation and based on earlier estimation on accident reports the cost of repair can be made accurately by using the text mining with combination techniques to automatically discover accident characteristics that can inform better understanding of the contributors to the accidents. The findings from text mining of the narratives can recover understanding of the contributors to railroad accidents in ways not convenient through detached fixed field analysis of the accident reports.

IV Elder and John[5] studied that as the Internet expands and our natural capacity to process the unstructured text that it contains diminishes, the value of text mining for information retrieval and search will increase dramatically. This comprehensive professional reference brings together all the information, tools and methods a professional will need to efficiently use text mining applications and statistical analysis. The Handbook of Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications presents a
comprehensive how- to reference that shows the user how to conduct text mining and statistically analyze results. In addition to providing an in-depth examination of core text mining and link detection tools, methods and operations, the book examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection using real world example tutorials in such varied fields as corporate, finance, business intelligence, genomics research, and counterterrorism activities.

G. Cirovic and D. Pamucar [6] studied that every year, more than 400 people are killed in over 1.200 accidents at road-rail level crossings in the European Union (European Railway Agency, 2011). Together with tunnels and specific road black spots, level crossings have been identified as being a particular weak point in road infrastructure, seriously jeopardizing road safety. In the case of railway transport, level crossings can represent as much as 29% of all fatalities caused by railway operations. In Serbia there are approximately 2.350 public railway level crossings (RLC) across the country, protected either passively (64%) or by active systems (25%). Passive crossings provide only a stationary sign warning of the possibility of trains crossing. Active systems, by contrast, activate automatic warning devices (i.e., flashing lights, bells, barriers, etc.) as a train approaches. Securing a level crossing (whether it has an active or passive system of protection) is a material expenditure, and having in mind that Serbian Railways is a public company directly financed from the budget of the Republic of Serbia, it cannot be expected that all unsecured level crossings be part of a programme of securing them. The most common choice of which level crossings to secure is based on possible consequences of a rise in the number of traffic accidents at the level crossings. The process of selecting a level crossing where safety equipment will be installed is accompanied by a greater or lesser degree of uncertainty of the essential criteria for making a relevant decision. In order to exploit these uncertainties and ambiguities, fuzzy logic is used. Here also, modeling of the Adaptive Neuro Fuzzy Inference System (ANFIS) is presented, which supports the process of selecting which level crossings should receive an investment of safety equipment. The ANFIS model is a trained set of data which is obtained using a method of fuzzy multi-criteria decision making and fuzzy clustering techniques. 20 experts in road and rail traffic safety at railway level crossings took part in the study.

H. Gonzalez, J. Han, Y. Ouyang, and S. Seith [7] outline the identification and characterization of traffic anomalies on massive road networks is a vital component of traffic monitoring and control. Anomaly identification can be used to reduce congestion, increase safety, and provide transportation engineers with better information for traffic forecasting and road network design. However, because of the size, complexity, and dynamics of transportation networks, automating such a process is challenging. A multidimensional mining framework is proposed; 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. The framework is based on the development of two novel methods: (1) efficient anomaly mining stemming from the discovery of the atypical fragment (a compact representation of a set of abnormal traffic patterns happening across a sequence of connected road segments, possibly spanning multiple roads, and occurring at overlapping time intervals) and (2) a multidimensional anomaly overlay model that

 ISSN: 2454-4248
383 – 386

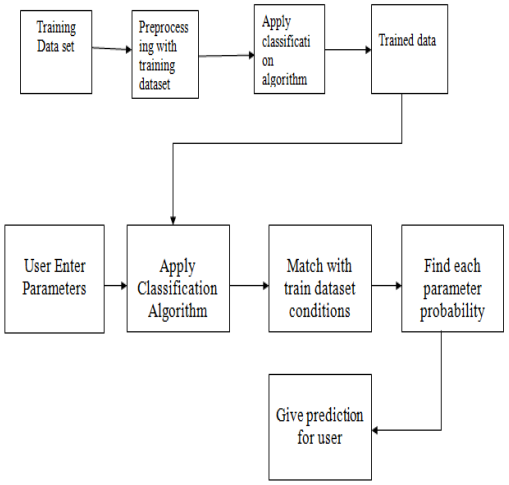
enables the clustering of multiple atypical fragments according to different criteria (e.g., severity, topology, or spatiotemporal characteristics).

III. Existing System

A review of the data collected by the IR shows a variety of accident types from derailments to truncheon bar entanglements. Most of the accidents are not serious; since, they cause little damage and no injuries. However, there are some that cause over 7 crores in damages, deaths of crew and passengers, and many injuries. Every station or location has its contributor which manages and maintains all the information and report of the accidents that occurred in that area. These information is only accessed by authorized people. The railway drivers do not have right to use to this information. Thus drivers are not aware of what type of accidents can occur.

IV. Proposed System

The propose system that 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 how can we avoid the accidents and give security. In performing this evaluation the study also considers the usefulness of modern ensemble approaches incorporating these text-mined features to predict for avoiding accident. These will help the railway driver to a large extent



1. ID3 Algorithm

Iterative Dichotomiser 3 or ID3 algorithm which is used to generate decision tree. It classifies the data using the attributes. Tree consists of decision nodes and decision nodes.

Training Phase –

Building the decision tree:

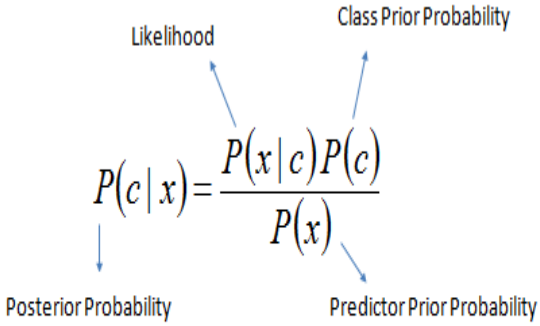
i. In the ID3 algorithm, we begin with the original set of attributes as the root node.
ii. On each iteration of the algorithm, we iterate through every unused attribute of the remaining set and calculates the entropy (or information gain) of that attribute.
iii. Then, we select the attribute which has the smallest entropy (or largest information gain) value.
iv. The set of remaining attributes is then split by the selected attribute to produce subsets of the data.
v. The algorithm continues to recurse on each subset, considering only attributes never selected before

2. Naive Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Algorithm

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

- P(c|x) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.

 ISSN: 2454-4248
383 – 386

- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

3.Agglomerative hierarchical clustering (AHC) Algorithm:

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The
classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.

The hierarchy within the final cluster has the following properties:

Clusters generated in early stages are nested in those generated in later stages.

Clusters with different sizes in the tree can be valuable for discovery.

V. Conclusion:

The output will show the analysis of the giving parameter using classification algorithm for giving which type of accident may occur. Classification algorithm analyses the probability of each parameter for giving result and 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 not available from only the fixed field entries. The improvements provided by text and ensemble modeling are dramatic even without working to optimize the performance of the ensemble methods for these data using AHC algorithm.

Reference:

[1]. https://en.wikipedia.org/wiki/List_of_countries_by_rail_transport_network_size

[2]. Moty Ben-Dov ,Wendy Wu ,Ronen Feldman ,Ramat Gan and Paul A. Cairns "Improving Knowledge Discovery By Combining Text-Mining And Link-Analysis Techniques "DOC: 28-31 Oct. 2007

[3]. Nagamallika, A.Anuradh,K.Sridevi " To Characterize The Contents Of The Documents Through Pattern Discovery In Text Mining " Vol 6, Issue-7,July-2017

[4]. N.M.Deepika , K. Rashmi ,V Divyavani " Text Mining Analysis And Predicting Techniques Of Accidents " vol 6,Issue-March 2017

[5]. IV Elder and John “Text Mining and Statistical Analysis for Non-structured Text Data Applications” DOP-2012-01-15

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

[7]. 5.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.