I J C T A, 9(40) 2016, pp. 213-219

Opportunities and Advances of Data mining and Data Analytics in Railways

Rashmi Thakur[1] and M. V. Deshpande[2]

ABSTRACT

Transportation industry is not able to handle the situation of analyzing bulk amount of data which is collected from various sources which in is not necessarily a structured data. The techniques used in data mining and tools are not sufficient to handle those data and analyze for better Planning and management. This paper gives the overview of various challenges in the railways sector. It also briefs about the various mining and analytical developments done in rail sector in countries apart from India.

Keywords: Transportation, Data mining, Big Data, Data analytics, Rail Sector

I. INTRODUCTION

Data mining is the method of extraction of information or the patterns from the existing data. Data mining tools are able to detect behavioral patterns in future which makes possible for businesses to take their decisions based on knowledge gained. With Data mining we can detect trends in sales, develop campaigns for marketing, and predict loyalty Points for Customers. Challenges of data mining[1] are handling various data types, scalability of algorithms, usefulness of results, mining information from various sources of data, and data security.

Data analytics is the new field which of examines raw **data** in order to draw conclusions regarding that information. **Data analytics** is utilized mainly in industries to make the decisions of their organization in better way.

Big Data [2][3][4] refers to huge set of data which is collected from the sources like twitter, sensor data, face book, online applications. Various companies now a days have huge data of their transactions with respect to different customers and suppliers. To understand business better, various industries have to analyze this large amount of data for increasing their profits. With the help of big data various organizations can learn facts about their business and operations and can transform knowledge into much better decisions and improve their business performance and efficiency. Hence nowadays Big is very popular.

Transportation industry is generating large volume of data from its various sources. Also the data generated can be structured, Semi structured or unstructured data. The mining algorithms are not able to analyse the semi structured or unstructured data. Large number of research is done during past years in many countries in the area of datamining and analytics in railways Sector. But unfortunately India is not using technologies in data mining and analytics that are used in advanced countries. There are many research areas where researchers can work in the area of data mining and analytics in Indian rail sector.

1.1. Challenges of data mining and analytics in rail Sector [5]

1) **Heterogeneity:** The data which is collected in railways is not homogeneous. Hence whenever we apply algorithm to railways data set it becomes very difficult to transfer the data into one uniform format.

[1] Ph.D. Research Scholar, MPSTME, NMIMS University, Mumbai, India, *E-mail: thakurrashmik@gmail.com*

[2] Professor & Dean, School of CSIT, Symbiosis University of Applied Sciences, Indore, India, *E-mail: manojvilasrao@gmail.com*

2) **Inconsistency and Incompleteness**: There are missing values from data is collected from different geographical locations. This occurs as data is collected and stored at different databases which are geographically located far and the databases can be incomplete and inconsistent.
3) **Merging of Data**: To do the analysis, data has to be on central server which practically is not possible as the data in railways is usually scattered zonal wise. For analysis purpose there is need to merge the data from different databases.
4) **Timeliness**: It is important that real time data analytics techniques should be implemented in railways. But many times it's not possible to collect real time data for analysis purpose within time limit.
5) **Privacy and Data Ownership**: When dealing with electronic information, Privacy is an important issue on which one needs to focus upon.

1.2. Analytics Application Areas

- Freight movement and routing optimization
- Inventory Management & Capacity Optimization
- Improved Customer Experience by develop effecting communication via insights from social media, persona segmentation & preferences
- Reduced Environmental Impact and Increased Safety
- Optimized Transit schedules by predicting impact of maintenance, road-works, congestion & accidents.

II. APPLICATIONS OF ANALYTICS IN RAIL INDUSTRY[6][7]

- **Train Signal and Control Systems:** Urban railways use many signals to observe train movements. Train signals can be integrated with ordinary street traffic light systems. Bay Area Rapid Transit and Philadelphia have deployed analytics for operations for running trains automatically.
- **Route planning and Scheduling:** Due to advancement in analytical processing of the complex data laborious and manual transit scheduling tasks are made easy and more efficient. Due to development of powerful software's routing and development of time tables and train schedules have become easy.
- **Detecting Vehicle Location Automatically:** Locating Vehicle automatically and associated information system for passenger are the most popular applications of analytics. By using GPS based system, it has become very easy for control centers to detect where the particular train is running and is the train following its schedule. We can find Travel time from it.
- **Automated Fare Collection:** Automated fare collection systems uses ticket vending machine at all stations that can receive cash or process credit card swipes. Due to this central database is automatically updated. Due to this passes and discounted multi-tickets are encouraged. When credit cards are swiped, passenger information can be correlated and their travel patterns can be studied which can help in improving services.
- **Counting Passengers Automatically:** With the development of techniques which counts passengers automatically transit authorities can count how many passengers are actually boarding or deboarding the train which can update central database and help the authorities to provide better service. Also this data can help in studying, analyzing and visualizing train wise, day wise passenger usage information.

III. ADVANCES IN DATA MINING AND ANALYTICS IN RAILWAYS OUTSIDE INDIA

Youfang Lin, Huaiyu Wan, Rui Jiang, Zhihao Wu, and Xuguang Jia[8] proposed a technique for better understanding of passengers from their purposes of travel. Paper uses real dataset of Passenger travel Records of Civil Aviation in China. It focuses on group based travel which includes group composition analysis, travel motivation exploration and group behavior modeling. It focuses on finding the purposes of

travel of passenger groups according to the historical travel records of passengers which helps to identify whether it is tourist or business group. Travel times of each single passenger as well as co travel time between any 2 passengers can be analyzed. This can be used in Destination Image marketing, tourism product involvement improvement, airport/Railway Service, Airlines can adjust their flight schedules according to different types of passenger groups.

Le Minh Kieu, Ashish Bhaskar, and Edward Chung[9] proposed segmentation of passengers by data from smart card. It uses dataset Translink which is transit authority of Queensland (SEQ), Australia. Paper focuses on transit passenger characterization and segmenting passengers using dynamic SC data. The segmentation is used to group same travel pattern of passengers i.e with same level of transit journeys at regular times and places. Paper focuses on analyzing temporal pattern from travel itineraries. This concept can be useful for Fare Collection, Setting new policies which can be beneficial to passengers.

Evelien van der Hurk, Leo Kroon, Gábor Maróti, and Peter Vervest[10] proposed a technique by which we can deduct passenger choice for routes using smart cartd data. It uses dataset of Netherland Railways. It focuses on Passenger Route choice information. The passenger route choice deduction is important for analyzing passenger service in terms of travel time, which is dependent on route choice. Due to the unique data set resulting from conductor checks, the passenger route choice deduction can be validated on a journey specific level. The journey validation is of higher accuracy than validation methods based on train capacity utilization. It can be used for calculating train utilization. Experienced journey of passengers can be analyzed in terms of waiting time, in-vehicle time and no. of transfers.

Adithya Thaduri, Diego Galar, and Uday Kumar[11] traces that big data analytics can be used in railways. The paper describes Big Data technologies in transportation specifically to Railways. This technique can be applied to Swedish Rail Administrator. Big Data analytics can be applied in scheduling the windows when there is less traffic, to do maintenance tasks at bottlenecks, rescheduling the assets with respect amount of traffic by the passengers.

IV. COMPARITIVE ANALYSIS OF ADVANCED TECHNIQUES USED

Table I Analysis on Various Parameters

Country	*Advantages*	*Traditional Algorithm Used*	*Disadvantages of Traditional Approach*	*Improvements in traditional Approach*	*Future Scope*
China	• Helps to better understand passengers and should bring about meaningful changes for travel service and decision making of Passenger carriers and government. • Can help carriers provide precise and personalized services or recommendations for passengers. • Helpful for government to make decisions on economic development or urban construction. • Statistical information of different types of passenger groups can help govt. evaluate the economic tendency so that they can adjust policies timely and effectively on operationalization of destination.	Logistic regression for Classification	• For building the classifier it considers only basic features for passenger group including demographic characteristics (age, gender), characteristics of current travel (group size, mileage), historical characteristics (historical travel time and mileage of group Members). • classifiers are developed on assumption of independent distribution which assumes that any two of the passenger groups is independent of each other. But in real life scenario this assumption don't work.	Paper focuses on collective inference model which exploit autocorrelation dependence relations between the variables of related entities to improve predictions by estimating the joint probability distributions over the entire graph and establishes passenger social networks. Also iterative classification technique is used which allows that there are some statistical dependence relations rather than being completely independent between the labels of target objects.	Apart from only analyzing basic features one can apply analytics for getting various advanced features like how much tickets the customer is purchasing. Above method can give details of the potentially high value travelers which can be used for increasing the profit of carrier, to give promotional benefits to the passengers and market segmentation. Also the group information can be used next time when the ticket is booked so that there is no need to reenter information gin and aga in which saves the time.

contd. table 1

Country	*Advantages*	*Traditional Algorithm Used*	*Disadvantages of Traditional Approach*	*Improvements in traditional Approach*	*Future Scope*
Australia	• Passenger Segmentation can bring various profits to authorities to serve their passengers in better way. • Incentives and services can be given to Passengers of regular usage to encourage passengers to use public transport. • The travel pattern can be observed which benefits strategies such as transfer coordination and origin-destination demand management by monitoring and inferring passenger movements through their travel habits. • Transit authorities could pay special attention to passengers who were not irregular passengers but have recently become irregular. These are potential customers whose behavior has changed due to certain reasons.	DBSCAN-density based clustering algorithm	DBSCAN was able to give only Spatial and Temporal travel pattern.	Apriori algorithm is used along with DBSCAN.	Paper considers only data of working days. Data from Smart card of other time periods can also be used to evaluate the policies and measure their to transit passengers. Analysis of coming-back home can be studied. One can observe the behavior and how the trip is segmented to understand different passenger types.
Netherlands	It can be used for measuring of passenger service.	Dijkstra's or Belman Ford	• Traditional algorithms aims to find cost of routes which is minimum but passengers are known to not always travel along unique minimum cost path. Other routes may be more attractive because they contain fewer transfers, are carried out by different train types, have lower fares or have departure or arrival time that is convenient for passengers. • Most passenger's choice models for routes are based on utility maximization or regret minimization. But in the case of sudden changes in the timetable these models could not be valid. Due to unavailability of up-to-date information, One passenger may not travel as predicted by the existing model. Second, the urgency to make quick decisions may result in unexpected travel routes. Thus, route choices based on the traditional models may be incorrect in these specific situations.	Paper proposes a new method for route deduction and validates this method through additional data resulting from conductor checks. Validation of the method is possible as conductor checks provide partial information on route choice for significant subset of all journeys made by smart card. As a result of the case study commonly made assumptions on route choice behavior that is that passengers take the first departing route or first arriving route don't hold true.	• Learning algorithm can be used to reduce the set of routes. • Development of statistical arguments that define when to eliminate a route or where to gather additional data to validate the elimination of the route. Results of route deduction can be used to fine tune the parameter settings of extended network, possibly aiming to set them in such a way that based on single setting, all possible routes can be found which can increase computational speed, reduce no. of candidate routes and possibly increase performance of the method.

contd. table 1

Country	*Advantages*	*Traditional Algorithm Used*	*Disadvantages of Traditional Approach*	*Improvements in traditional Approach*	*Future Scope*
Swedan	• By analyzing the data, researchers and users working in business can make decisions from the developed Model. Many tools are applied in the research that can also be used for mining the data and drawing conclusions. We can use data mining, machine learning, predictive analytics • It can identify behavior of bottlenecks, maximum loads, variation in traffic, unplanned delay timings, inspection timings and accidents that will impact customer's comfort, business's losses and asset's reputation. It can enhance overall efficiency that can lead to reduce in operating costs.	All Algorithms in Data Mining	Data Mining Algorithms.	Traditional approaches applied to big data	Algorithms used in the paper are limited to lab conditions but not Applicable to real life conditions. There can be improvement in algorithms which support scenario of real time.

V. SELECTED EXAMPLES IN RAIL INDUSTRY

1) **Bay Area Rapid Transit (BART)[12]**: BART provides automated rail rapid system (RRT) for San Francisco Bay Area. BART keeps record of supervision over all steps of their system like including services related to passengers, operations of train etc Various analytical features such as delay analysis, passenger flow modelling and system performance analysis is done. Also BART team can monitor train arrival and departure timings.

2) **Salt Lake City TRAX[13]:** LRT system of Salt Lake City's shows how tremendous volume of data which is real time can be utilized for analysis purpose. It implements train tracking and dispatching system which relies on GPS based network to locate trains in peak time and also to inform the problems which train faces to the control center, located at TRAX's Jordan River Service Center. It is very good example of application of big data analytics.

3) **Philadelphia — SEPTA Regional Rail[14]:** Southeastern Pennsylvania Transportation Authority (SEPTA) has implemented analytics and big data throughout its system. Different modules are implemented like controlling of operations, collecting fares automatically, counting passengers automatically etc. Analytics plays a role in current planning and scheduling. SEPTA's passenger information system provides PIDs with train arrival/departure updates in some of the system's larger stations. It also includes an app that provides bus and train status information to passengers' smartphones (Android and I-phone platforms).

4) **Philadelphia — SEPTA Suburban Trolley Lines[14]:** Suburban Troll lines is good example of Analytics particularly in Signaling-Control functions. GPs is used for train scheduling and measuring on time performance. Google maps are integrated with system so that change in route can be mapped in Trapeze.

5) **Seattle — Sound Transit's Link and Sounder[15]:** The Seattle provides various services to reach surrounding areas of metros. Analytics and Big data are involved in signal-control-dispatching; passenger information with online and smartphone train statistics information, Automated fare collection, GPS

functionalities etc. Analytics is involved in maintaining passenger information online, detecting status of train using smart phones, signaling-control-dispatching; APC; GPS and AVL capabilities; and AFC with TVMs in stations. In automated fare collection, A contactless, stored-value "smartcard" containing a microprocessor, the ORCA (One Regional Card for All) card is used for the payment of public transportation fares thus providing a virtually "seamless" fare-payment (in effect, a prepaid pass) among these multiple systems and agencies. Many discount schemes are offered to the card holders taking the packages and also to disabled and senior citizens.

6) **Austin — Capital Metro's MetroRail[16]:** In Austin Authority in Transportation is operating its MetroRail using diesel multiple-unit. Though MetroRail is small in size it is using analytics in line's operations particularly in its ABS system. Trains are equipped with GPS. Passenger statistics are used to improve overall services and study service performance so that operations can be improved.

VI. CONCLUSION

Data mining and analytics are the focus gaining branches which are used in every sector as data is growly exponentially which cannot be handled by traditional algorithms. Due to the competition there is the need to develop the decision making systems and visualize the data for better insight of the organization.

Sentiment analysis and text analytics are fame gaining branches where researchers are focusing as data which is arriving in the organization is not only the internal generated data but also from the social networking sites like Facebook, twitter and which is in the textual format. There are many reviews or complaints that are sent to various sectors specifically railways where they are not analyzed and classified with the help of data mining classification techniques nor the positive or negative sentiments are found from them. Hence there is a need of text analytics which can handle huge reviews coming into the organization or the rail sector. Sentiment analysis focuses on the positive, negative and neutral sentiments. Study and classification of this sentiment help the rail industry to improve its services. Indian Railways is the world's largest railway network. But very less research has been done in the field of datamining, data analytics and big data in Indian Railways. There is the need for researchers to work on live problems faced by Indian railways and provide the solutions for improving the services and satisfaction level of Passengers.

REFERENCES

[1] Han, Chen, M. S., J., & Yu, P. S., (1996), "Data mining: An overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, 8(6), 866–883.

[2] HAN HU, YONGGANG WEN, TAT-SENG CHUA1,AND XUELONG LI," Toward Scalable Systems for Big Data Analytics:A Technology Tutorial",2014,IEEE Access.

[3] Chun Wei Tsai, Chin Feng Lai, Han Chieh Chao and Athanasio V. Vasilakos "Big data analytics: a survey"2015,Springer.

[4] Xindong Wu, Xingquan Zhu,Gong-Qing Wu, and Wei Ding "Data mining with Big Data", 2014, IEEE Transactions on knowledge and data Engineering.

[5] Nii Attoh-Okine "Big Data Challenges in Railway Engineering",2014,IEEE International Conference on Big Data

[6] Christopher Turner, Ashutosh Tiwari, Andrew Starr, Keven Blacktop "A Review of key Planning and scheduling in rail industry in Europe and UK", 2015,Journal of Rail and Rapid Transit

[7] António A. Nunes, Teresa Galvão Dias, and João Falcão e Cunha "Passenger Journey Destination Estimation From Automated Fare Collection System Data Using Spatial Validation", 2015, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.

[8] Youfang Lin, Huaiyu Wan, Rui Jiang, Zhihao Wu, and Xuguang Jia "Inferring the Travel Purposes of Passenger Groups for Better Understanding of Passengers", 2015, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16, NO. 1

[9] Ashish Bhaskar, Le Minh Kieu, and Edward Chung "Passenger Segmentation Using Smart Card Data", 2014, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS

[10] Evelien van der Hurk, Leo Kroon, Gábor Maróti, and Peter Vervest "Deduction of Passengers' Route Choices From Smart Card Data", 2015,IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, VOL. 16

[11] Adithya Thaduri, Diego Galar, and Uday Kumar "Railway assets: A potential domain for big data analytics", Elsevier, Volume 53, 2015, Pages 457–467, 2015 INNS Conference on Big Data

[12] Center for Urban Transportation Research (CUTR) staff. Case Study — Bay Area Transit District (BART) — San Francisco, California; CUTR, University of South Florida (USF); document #FTA-FL-26-71054-03. http://www3.cutr.usf.edu/security/documents/UCITSS/BART.pdf

[13] Henry, Lyndon. Analytics Keep SLC's Light Rail on Track. *All Analytics* (online), 28 December 2012. *http://www.allanalytics.com/messages.asp?piddl_msgthreadid=260931*

[14] Calnan, John F. (Manager, Suburban Service Planning & Schedules, SEPTA). Phone conversation, 10 April 2013. Email message, "Signals, APC, GPS etc. on SEPTA Suburban LRT ", 11 April 2013.

[15] ORCAcard.com website editors. About ORCA. Accessed 9 April 2013. *http://www.orcacard.com/ERG-Seattle/p3_001.do?m=3*

[16] Lindblom, Mike. Is Big Brother watching your ORCA card? *Seattle Times*, 17 December 2009 (updated 18 December 2009). *http://seattletimes.com/html/localnews/2010537022_orcacard18m.html*