Safety Science 48 (2010) 1427–1435

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Application of association rules in Iranian Railways (RAI) accident data analysis

Ahmad Mirabadi [a,b,*], Shabnam Sharifian [a,b,c]

[a] Iran University of Science and Technology (IUST), School of Railway Engineering, Narmak, Tehran, Iran
[b] Iran Institute of Railway Research and Development (IRRD), Tehran, Iran
[c] Khatam Institute of Higher Education, Ferdos Boulevard, Tehran, Iran

ARTICLE INFO

Article history:
Received 16 February 2010
Received in revised form 27 May 2010
Accepted 19 June 2010

Keywords:
Data mining
Association rules
CRISP reference model
Iranian Railways (RAI)
Railway accidents
Accident analysis

ABSTRACT

The demand to travel by rail is ever increasing because it benefits both passengers and freight; therefore it is of utmost importance for railway administrators to carry passengers and freight safely to their destinations. Undergoing safety procedures and developing safety systems require awareness of what is causing unsafe conditions. This can be accomplished by learning from the past. This research has been performed to analyze the data from past accidents of the Iranian Railway (RAI) by applying association rules data mining techniques in order to discover and reveal unknown relationships and patterns among the data. By the application of CRISP-DM as the data mining methodology and utilizing Clementine 12.0 as the software tool, the mentioned objectives of this paper were fulfilled. For this research some 6500 accident records were selected from the accidents database from 1996 to 2005. The ultimate relationships and patterns extracted can been utilized to develop regulations and rules. This research considers accident conditions and relationships discovered among the most common accident factors (human error, wagon and track) with other fields of the database in order to prevent them from occurring in the future.

1. Introduction

Today we sail in an ocean of all kinds and ever increasing amounts of experimental data (i.e., examples, samples, measurements, records, patterns, pictures, tunes, observations, etc.) produced by various sensors, cameras, microphones, pieces of software and/or other human made devices. The first obvious consequence of this fact is that humans cannot handle such massive quantities of data which are usually appearing in the numeric shape as huge (rectangular or square) matrices (Huang et al., 2006).

To understand the term 'data mining' it is useful to look at the literal translation of the word: to mine in English means to extract. The verb usually refers to mining operations that extract from the Earth her hidden, precious resources. The association of this word with data suggests an in-depth search to find additional information which previously went unnoticed in the mass of data available (Giudici, 2003).

This terminology was first formally put forward by Usama Fayaad.[1] He used to refer to a set of integrated analytical techniques divided into several phases with the aim of extrapolating previously unknown knowledge from massive sets of observed data that did not appear to have any obvious regularity or important relationships. As the term 'data mining' slowly established itself, it became a synonym for the whole process of extrapolating knowledge (Giudici, 2003). Here is a more complete definition of data mining:

> *Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database* (Giudici, 2003).

According to the online technology magazine, ZDNET News, data mining is predicted to be "one of the most revolutionary developments of the next decade." In fact, the MIT Technology Review chose data mining as one of ten emerging technologies that will change the world. "Data mining expertise is the most sought after" among information technology professionals, according to the 1999 Information Week National Salary Survey. Since many companies have implemented a data warehouse strategy, they are now starting to look at what they can do with all that data (Larose, 2005).

Railway is nowadays an excellent means of transport to reduce pollution and avoid traffic congestion; a safe and economic way to reach a destination for both passengers and freight. As a result, passengers and freight owners prefer railway to other transportation modes. Therefore, to retain this conception and achieve a competitive advantage, railway transportation administrators should

* Corresponding author at: Iran University of Science and Technology (IUST), School of Railway Engineering, Narmak, Tehran, Iran. Tel.: +98 21 77491029.

E-mail addresses: mirabadi@iust.ac.ir (A. Mirabadi), shabnam.sharifian@gmail.com (S. Sharifian).

[1] The "First International Conference on Knowledge Discovery and Data Mining", Montreal, 1995.

0925-7535/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ssci.2010.06.006

Table 1
Literature review on accident analysis.

Reference	Techniques	Findings
Chong et al. (2004)	Decision Trees, NN[a]	By modeling the severity of traffic accidents' injury, the three most important factors in fatal injury were driver's seat belt usage, light condition of road-way, and driver's alcohol usage
Chong et al. (2005)	NN, SVM,[b] Decision Trees	The paper developed models that accurately classified the severity of injuries within five categories: no injury, possible injury, non-incapacitating injury, incapacitating injury, and fatal injury
Nefti and Oussalah (2004)	NN	Presented model taking irregularities in the positioning of the rails as input and predicted the safety ratio of the rails and therefore the safety of the train on the track
Barai (2003)	Data Mining	This paper reviews the applications of data mining techniques in transportation engineering problems
Chang and Wang (2006)	CART Model	A model is developed to establish a relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables. Results indicate the most important variable is vehicle type
Solomon et al. (2006)	Decision Trees, NN, Market-Basket Analysis, K-Means	The paper demonstrates the use of data mining to evaluate the traffic safety improvement of red-light-signal controlled intersections monitored by cameras in reducing fatalities
Tesema et al. (2005)	Regression Trees	This paper applies data mining to determine interesting patterns with respect to injury severity on road accidents data
Sze and Wong (2007)	Binary Logistic Regression, Logistic Regression Diagnostics	A decreasing trend in pedestrian injury risk was reveled in this paper, controlling the influences of demographic, road environment, and other fields. Influences of pedestrian behavior, traffic congestion, and junction type on pedestrian injury risk were also subject to temporal variation
Depaire et al. (2008)	Clustering	This paper applies latent class clustering for identifying homogenous traffic accident types. The cluster analysis uses vehicle type as a basis for segmentation
Anderson (2009)	KDE,[c] Clustering	This paper studies the spatial patterns of road accident injury and applies environmental data and results from the patterns in order to create a classification of road accident hotspots
Abugessaisa (2008)	VDM,[d] Exploratory Data Analysis, Clustering, SOM,[e] Classification Trees,	This paper discovers clusters and relationships in road safety database, explores the contents and structure of the data set, and covers interactive explorations based on brushing and linking methods to detect and recognize interesting patterns in the available database
Sohn and Lee (2003)	NN, Decision Tree, Bayesian Fusion, Bagging, Clustering	In this paper, various algorithms to improve the accuracy of individual classifiers for two categories of severity of road traffic accident were applied. Results indicate a clustering based classification algorithm works best for road traffic accident classification in Korea
Lee et al. (2004)	Clustering	This paper explores the application of data mining in incident situations investigation. Data used is obtained from simulation. The demonstration shows data mining enables the user to mark the impact area of the incident temporally and spatially
Xie et al. (2007)	NN, Bayesian Methods, Negative Binomial	This study compares three types of models for predicting motor vehicle crashes

[a] Neural networks (NN).
[b] Support vector machines (SVM).
[c] Kernel density estimation (KDE).
[d] Visual data mining (VDM).
[e] Self-organized-maps (SOM).

work diligently to raise the level of safety and reduce factors that cause accidents.

Developing safety systems and undergoing safety procedures require an awareness of the nature of previous accidents on the railway network to identify potential causes. Many accidents occur because past accidents or failures were not utilized for change and growth. For this reason it is of utmost importance to analyze accident data in order to extract hidden knowledge among huge amounts of data. Effective analysis of data from a database can help in development of knowledge that can support safety management strategies and reduce future accidents.

The objective of this article is to discover meaningful new correlations, patterns and trends among rail accidents' data of the Iranian Railways (RAI). Once unnoticed and hidden relations are clarified, the outputs are further analyzed in order to present suitable solutions according to the needs of RAI.

In this article, we considered safety as "being protected against *loss of life* and *property* in all aspects and dimensions related to the railway industry for both railway and non-railway individuals." Therefore, the objective of this paper is to analyze data on accidents which have resulted in loss of life and property by applying association rules, a data mining technique.

Data mining has been an active analytical technique in many scientific areas over the years (Chang and Wang, 2006). One of these areas is transportation, and many researchers have figured out the useful role data mining plays in dealing with a mass of transportation data, and they have seen the advantages of applying data mining to retrieve or analyze the data (Chong et al., 2004).

One of the most common fields of transportation to apply data mining is accident analysis. Very little is known about the usefulness of applying data mining in railway accident analysis, although there are numerous applications of data mining in road accident analysis. One main reason for that may be the limited number of accidents happening on railway networks compared to those on roads. One significant category of events happening on railway networks is near miss incidents. Railway companies with sophisticated safety information systems, who gather information on them, can apply data mining techniques effectively in analyzing these events. By considering the number of rail accidents on the RAI network, applying data mining techniques and considering experience of similar analysis on roads will be useful. Table 1 holds a literature review performed for this purpose.

Demand for traveling by rail for both passenger and freight has increased significantly over the past years and is predicted to increase steeper than before. Railway administrators are planning to respond to the future growing demands with a wide range of software and hardware procedures. No mater what these procedures are, they all lead to increase in mobility and accidents are a consequence of this increase.

2. Methodology and tool

The methodology of this article is based on the CRISP-DM reference model. This model (Fig. 1) consists of six phases, their respective tasks, and the relationships between these tasks. Moving back

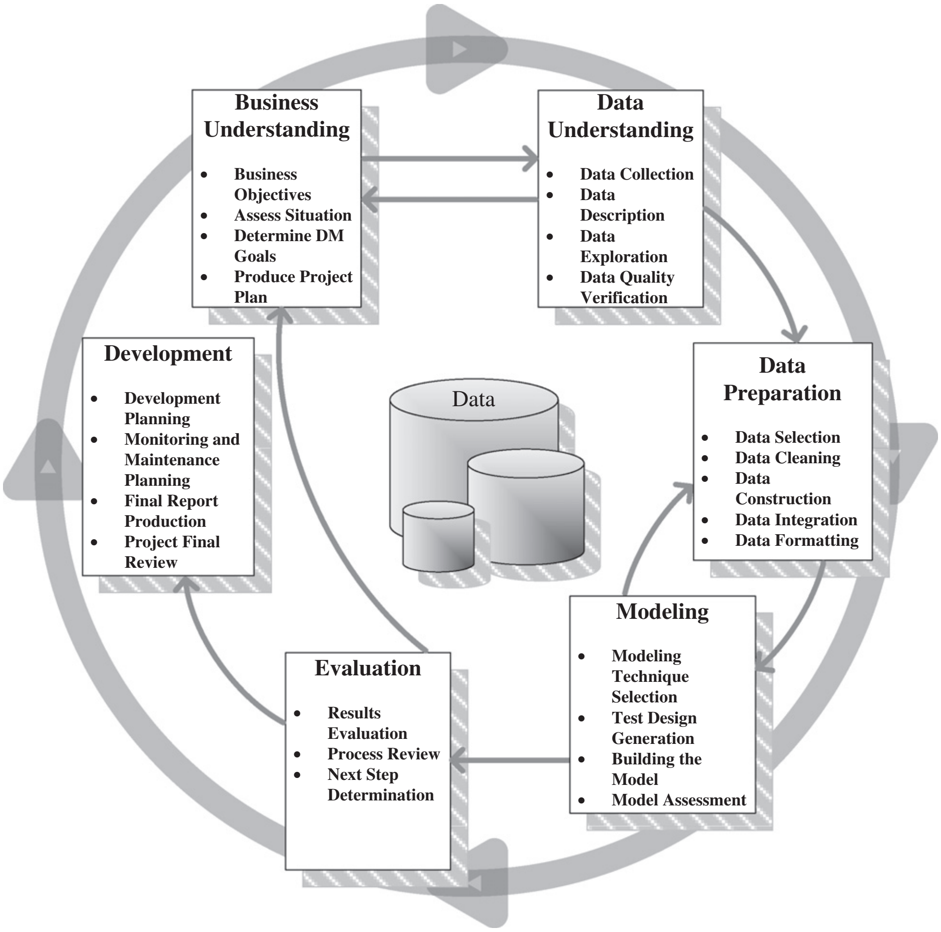

Fig. 1. Phases of the CRISP-DM reference model.

and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next (Chapman et al., 2000).

A convenient way to adopt data mining analysis is to use a software program that hosts facilities to mine the data in a variety of ways. Clementine 12.0, a product of Integral Sol., Ltd., due to its visual interface, algorithm breadth and following CRISP model made it suitable for the purpose of mining data in this research.

3. Results

In this section, results of the six phase of the research methodology are reviewed.

3.1. Business understanding: safety and accident analysis in RAI

The Iranian Railways (RAI) is the national state-owned railway system of Iran. The Iranian Ministry of Roads and Transportation (R&T) is the state agency that oversees the RAI.

In 2008, Iran with an area of 1,648,195 km^2 and nearly 70 million populations, RAI operated 11,106 km of rail network over 14 districts (RAI, 2006). The railway network expands by about 500 km per year according to the Ministry of R&T.

In 2006, RAI reported its facilities as follows (RAI, 2006):

1. Locomotives (diesel-electric, electric and shunting), numbering 565.
2. Wagons (covered, low sided, high sided, flat, well-wagon, tank wagon, mineral materials carrying wagon, bulk carrying, ballast, gas, refrigerator, etc.), numbering 16,330.
3. Different kind of passenger coaches, numbering 1192.
4. Main stations, numbering 429.
5. Bogie change installation at Jolfa and Sarakhs international border stations witch changes about 200 bogies each 24 h based on two working shifts.
6. Free Trade International stations, numbering 16.

The Gen. Dept. of Movement Safety is held responsible for safety issues in RAI. One of the functions performed by this Dept is gathering and analysis of accidents data. The analyses fulfilled on the data stored in the database are simple descriptive statistics and comparison of the statistics with previous periods. For this purpose an annual bulletin has been published every year since 2001, to reports the number of accidents categorized by accident type, grade, cause, etc. in railway districts.

As previously mentioned the objective of this paper is to discover correlations and trends that lead to loss of life and property related to railway and non-railway individual and ultimately develop solutions to break the identified accident patterns toward safety.

3.2. Data understanding: gaining insight on accidents data

In data understanding phase, data collection, codification and quality verification play critical roles in ensuring a qualified results in data mining process. An unbiased and accurate data collection leads to more accurate analysis results. Different researchers have investigated the causes of errors in data collection and codification stage and proposed models and procedures to limit such dilemmas (Tormo et al., 2009; Van der Schaaf and Kanse, 2004).

Table 2
Job – Years of Service – Districts – Accident Type.

Job	Years of Service	Districts	Accident Type	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Shunting man	16–20	Hormozgan	Fire accident	0.26	5	1	0.11	100	4	0.75	91.18
									5	0.99	93.33
Assistant driver	6–10	Khorasan	Fire accident	0.31	57.14	1	0.26	83.33	4	0.88	77.5
									5	1.59	88.89

The data stored in RAI accidents database (DB) was used to fulfill this research. This DB was in two different parts: (1) 1996–2005 and (2) 2006 and later. These two parts were developed with different software developing languages and contained different types of information fields; even similar information fields were stored in different formats or left unfilled due to changes in users of the system and lack of personnel training in using the new system. Therefore, because of the larger amount of data in the 1996–2005 DB, this part with some 6500 records was chosen for mining and analysis. Studying fields of each table, 38 out of 63 information fields were found proper for data mining. Appendix A holds titles and definitions of these fields.

Others fields were ether descriptive, not filed regularly through out these years, or anyhow irrelevant for data analysis. There was also some information fields witch were extracted during the modeling phase from those mentioned in Table 2, according to requirements of the models developed.

3.3. Data preparation: getting the data fit for mining

It is quite common users of the system enter data with errors or unusual values, even data might be stored in an inconsistent format with analysis objectives, and therefore data must be prepared before mining. In other words, the data wished to be analyzed in the real world by data mining techniques are incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data), noisy (containing errors, or outlier values that deviate from the expected), and inconsistent (e.g., containing discrepancies in the department codes used to categorize items) (Han and Kamber, 2006).

In this research missing values, noises and inconsistency data were dealt with as follows:

- Missing data: replaces by their mod (for set data type), mean value (for range data types), and in some cases they were removed or ignored (depending on the field or amount of missing data).
- Noisy data: omitted due to the small number of such records or lack of information to clean others.
- Inconsistent data: numerical and alphabetical pieces of information became coded in suitable numerical sets, ranges or flag type.

3.4. Modeling: association rules

Association rules were developed to underline groups of items that typically occur together in a set of transactions. Association rules are probably the most well-known local method for detecting relationships between variables. They can be used to mine very large data sets, for which a global analysis may be too complex and unstable (Giudici, 2003) witch take the form "If antecedent, then consequent," along with a measure of the support and confidence associated with the rule (Larose, 2005).

Among the association rule algorithms supported by Clementine 12.0; witch are Generalized Rule Induction (GRI), Apriori, and CARMA; GRI is utilized for association rules extraction.

GRI applies the *J-measure*:

$$J = p(x)\left[p(y|x)\ln\frac{p(y|x)}{p(y)} + [1 - p(y|x)]\ln\frac{1 - p(y|x)}{1 - p(y)}\right]$$

where

- $p(x)$ represents the probability or confidence of the observed value of x and it can be used as a measure of the coverage of the antecedent.
- $p(y)$ represents the prior probability or confidence of the value of y and it can be used as a measure of the prevalence of the observed value of y in the consequent.
- $p(y|x)$ represents the conditional probability, or posterior confidence, of y given that x has occurred. This is a measure of the probability of the observed value of y given that this value of x has occurred. That is, $p(y|x)$ represents an updated probability of observing this value of y after taking into account the additional knowledge of the value of x. In association rule terminology, $p(y|x)$ is measured directly by the confidence of the rule.

As usual, the user specifies desired minimum support and confidence criteria. For GRI, however, the user also specifies how many association rules he or she would like to be reported, thereby defining the size of an association rule table referenced by the algorithm. The GRI algorithm then generates single-antecedent association rules, and calculates J, the value of the *J-measure* for the rule. If the "interestingness" of the new rule, as quantified by the *J-measure*, is higher than the current minimum J in the rule table, the new rule is inserted into the rule table, which keeps a constant size by eliminating the rule with minimum J. More specialized rules with more antecedents are then considered (Larose, 2005).

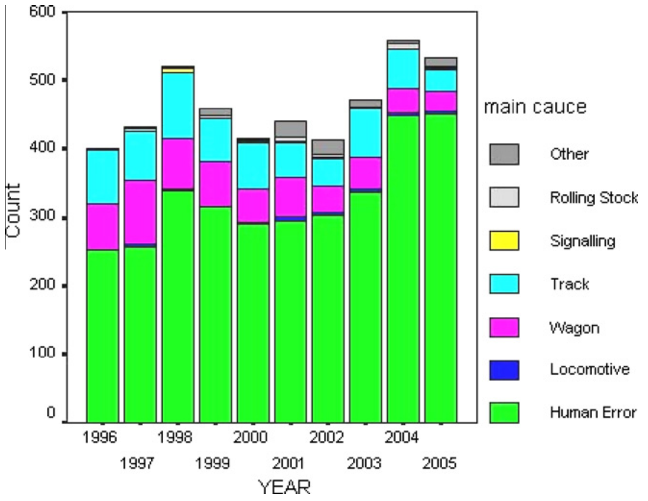

Fig. 2. Share of accident factors from 1996 to 2005.

Table 3
Job – Age – Districts – Accident Grade.

Job	Age	Districts	Accident Grade	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Driver	21–30	Hormozgan	D3	0.49	90.91	4	1.92	78.16	4	0.75	91.18
									5	0.99	100

Table 4
Accident Factor – Day of Week – Accident Grade.

Accident Factor	Day of Week	Accident Grade	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Human error	Saturday	D1	2.3	56.73	–	–	–	1	0.2	77.78
Human error	Sunday	I1	1.06	91.67	–	–	–	1	0.46	95.24
Human error	Saturday	D3	9.71	65.46	2	5.08	61.74	3	0.13	83.33
					5	0.42	52.63	4	1.02	76.09

Table 5
Accident Factor – Districts – Accident Type.

Accident Factor	Districts	Accident Type	Support	Confidence	Loss of Life	Support	Confidence	Loss of Property	Support	Confidence
Human error	Arak	Fire accident	0.46	100	3	0.15	100	–	–	–
					5	0.26	100			
Human error	Hormozgan	Fire accident	1.81	100	3	0.11	100	1	0.11	100
					4	0.75	91.18			
					5	0.99	93.33			
Human error	Esfehan	Fire accident	0.97	100	4	0.31	100	–	–	–
					5	0.71	90.62			
Human error	Tehran	Fire accident	3.89	100	4	1.99	86.67	–	–	–
					5	2.05	97.85			
Human error	South	Fire accident	0.93	100	4	0.22	90	–	–	–
					5	0.73	100			
Human error	South east	Derailment	0.77	100	2	0.07	100	–	–	–
					6	0.11	100			
Human error	Khorasan	Fire accident	2.12	100	4	0.88	77.5	1	0.26	83.3
					5	1.59	88.89			
Wagon	Arak	Collision of RS with O	0.6	77.78	–	–	–	4	0.15	100
Wagon	Hormozgan	Collision of RS with O	0.82	94.59	–	–	–	3	0.13	100
Wagon	Tehran	Collision of RS with O	1.92	73.56	–	–	–	3	0.09	100
Wagon	South	Collision of RS with O	0.57	80.77	–	–	–	3	0.07	100
Wagon	South east	Collision of RS with O	1.81	65.85	–	–	–	4	1.48	71.64
Wagon	Khorasan	Collision of RS with O	0.38	70.59	–	–	–	4	0.29	84.62
Wagon	North	Collision of RS with O	0.84	84.21	–	–	–	4	0.29	100

For the purpose of this paper, the share of accident factors during the years 1996–2005 were examined for more effective rule extractions. According to Fig. 2 among all factors, human error, wagon and track are the most accident producing ones. Therefore in this paper, the relationship between these factors and other fields of the database are examined.

In order to study meaningful relations, the experience of the researchers along with priorities of railway safety experts of RAI, resulted in a list of candidate relations for studying association rules. Table 1 holds the fields examined in relation to one another by considering accident factors. Among the association rules examined, those contributing to loss of life and property are considered.

Loss of life and property are weighted as follows:

– Non-RAI people killed = 5, RAI personnel killed = 4, non-RAI people injured = 3, RAI personnel injured = 2.
– Loss of 0–50 million Rials[2] = 5, Loss of 50–500 million Rials = 4, loss of 500–1500 million Rials = 3, loss of 1500–2000 million Rials = 2, loss of 2000 million Rials and higher = 1.

Tables 2–7 hold samples of association rules extracted from GRI algorithm. Each table holds confidence and support for each relation, and losses of life and property along with their confidence and support. List of all rules examined in the research are presented in Appendix B.

3.5. Evaluation: evaluating model outcomes

Regardless what some software vendor advertisements may claim, data mining software just cannot be purchased and installed, and you cannot sit back and watch it solve all the problems. Data mining is not magic. Without skilled human supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. The wrong analysis is worse than no analysis, since it leads to policy recommendations that will probably turn out to be just and expensive failure. Therefore results of software must be evaluated by human experts.

The data miner created models with the software and sets features of the module based on research hypothesis. Once the model execution terminates, and results are evaluated by human experts.

[2] Iranian currency.

Table 6
Job – Accident Type.

Job	Accident Type	Support	Confidence
Driver	Collision of RS with O	24.42	51.36

Table 7
Accident Factor – Accident Type.

Accident Factor	Accident Type	Support	Confidence
Human error	Collision of RS with O	46.29	59.9
Human error	Derailment	6.36	100
Human error	Fire accident	15.54	98.58
Wagon	Collision of RS with O	13.64	69.9

Table 8
Two rules defined initially and omitted later as a result of evaluation phase.

Emp. Condition	Level of Edu.	District	Accident Type/Grade
Governmental Official	Graduate	Khorasan	Derailment
Governmental Official	Graduate	Hormozgan	D3

For example Table 8 represents an example of two rules found by Clementine witch were omitted in this phase. When RAI safety experts analyzed the relationship of accidents with the level of education of employees, no reasonable explanations was found. Seeking such relationship with employee safety trainings is much more rational, but no information in this area was in hand.

In this phase, further review on related accident records archived in RAI Safety Dept and consultation with safety experts were applied to evaluate the rationality of the results found. Development in the next phase is based on outcomes of Evaluation Phase.

3.6. Development: safety regulations determination

If a system would be considered as a chain and its links as the system elements, once the chain is put under pressure, it tears from the weak link. The railway system is somehow like a chain and when accidents happen, we have got a weak link at that part of the chain. Once we recognize these points, they can be replaced with stronger links and help the system perform more efficiently in the future.

With the application of data mining, we have recognized such links by discovering repetitive patterns within the past accidents data. Up to this point, we have recognized areas where problems have lied in a general point of view. At the develop phase rules and regulations have been developed and suggested to RAI according to past findings to prevent past trends to be repeated. Some of these rules might already exist in RAI's rule books, and some may not. For those who exist, their repeat counts as emphasis and to those who do not, they can be considered suggestions. There are three main areas for the mentioned rules:

(1) Human resource
   - Workers uniforms, protective equipment and safety kits.
   - Work and rest hours.
   - Training.
   - Attaining qualified and promotion.
   - Reward and welfare.
   - Other facilities.

(2) Track
   - Inspections.
   - Maintenance.

(3) Wagons and freight
   - Carrying dangerous goods.

4. Conclusion

This research work on data mining has been followed with the objective of identifying hidden relationships of the most common accidents, their potential causes such as human factors, rolling stock, tracks and signaling systems with other fields of the RAI accidents database. Following CRISP-DM reference model, some 6500 records and 38 fields of data from RAI's accident database throughout the years 1996–2005 were applied for mining. Association rules was chosen as the data mining technique. Ultimate discovered patterns where extracted from Clementine 12.0 software. Results of final analysis where used to define safety rules and regulations in three areas of human resource, track and wagon as suggestions to RAI (Rules and Regulations are not mentioned in this paper) to prevent the accidents patterns from happening in future.

5. Discussion

Studies definitely do not stop here. There are other data mining techniques witch their application would be useful. One of them is time series. A time series is an ordered collection of measurements taken at regular intervals. The measurements may be of anything that might be of interest. Methods of modeling time series assume that history repeats itself – if not exactly, then closely enough that by studying the past, you can make better decisions in the future. By this technique you predict the values of one or more series over time. For example, you may want to predict the expected occurrence of specific accidents threw out the year in a particular district or on the entire network.

Studying the past behavior of a series will help you identify patterns and make better forecasts. When plotted, many time series exhibit one or more of the following features:

- Trends.
- Seasonal and non-seasonal cycles.
- Pulses and steps.
- Outliers.

Because planning decisions take time to implement, forecasts are an essential tool in many planning processes. The continuation of this research will concentrate on discovering seasonal and non-seasonal cycles, pulses and steps, and Outliers.

Acknowledgements

The authors would like to thank the Safety Dept of the Iranian Railways (RAI) to support this study by extensive comments from their safety experts and engineers, and also acknowledge AF Infrastructure to inspirit arranging this study.

The other important area for further investigation is on improving the procedure of accident data collection and codification, which will be effective in avoiding biased results in data analysis process.

Appendix A. Chosen fields of the RAI accidents database for data mining

#	Field title	Definition
1	Accident key	Unique key for each accident
2	Date	YY/MM/DD on which accident happened
3	Time	HH/MM on which accident happened
4	Day of week	Day of which accident happened starting from Saturday
5	District	RAI District on which accident happened in
6	Station before	One end of the block accident happened on
7	Station after	One end of the block accident happened on[a]
8	Kilometer	Kilometer from Tehran on which accident happened in
9	No. of RAI Personnel killed	Number of railway personnel killed in the accident
10	No. of RAI Personnel injured	Number of railway personnel injured in the accident
11	No. of people killed	Number of non-railway personnel killed in the accident (passengers, etc.)
12	No. of people injured	Number of non-railway personnel injured in the accident (passengers, etc.)
13	Losses of track properties	Financial damage to track (in Rials)
14	Losses of locos properties	Financial damage to locomotives (in Rials)
15	Losses of wagons properties	Financial damage to wagons (in Rials)
16	Losses of other properties	Other financial damages (in Rials)
17	Accident type	Type of the accident happened including: collision of rolling stocks with other rolling stocks, collision of rolling stocks with other obstacles, derailment, harm to passenger, harm to RAI personnel, fire accident (in locomotives, wagons, coaches), other
18	Accident factor	Cause of the accident consisting of: Human Error, Loco, Wagon, Track, Signaling, Rolling Stock, Other
19	Accident grade	Severity of the accident consisting of: I1, I2, D1, D2, D3 (from worst to least).
20	Accident class	Class of the accident consisting of: Railway, non-railway, fatality and injury.
21	Gradient	Steepness of the track at the accident location
22	Curve	Curve Radius at the accident location
23	Tunnel	If there is or not a tunnel at the accident site
24	Point	If there is or not a point at the accident site
25	Train type	Type of the train consisting of: passenger, freight, mixed, service, locomotive, shunting, other
26	Train no.	RAI assigned unique number to the crashed train
27	Wagon no.	RAI assigned unique number to the crashed wagon
28	Locomotive no.	RAI assigned unique number to the crashed locomotive
29	Train weight	Weight of the crashed train
30	Braked weight	Braked weight of the train
31	No. of wagons	Number of the wagons involved in the accident
32	Train speed	Speed of the train before accident
33	Age	Age of the accident culprit
34	Years of service	Years of service of the accident culprit
35	Marital situation	Marital situation of the accident culprit including single or maries.
36	Level of education	Level of education of the accident culprit including: illiterate, elementary school, guidance school, high school, graduate, AA, BA, MA, PhD
37	Job	Job of the accident culprit (witch includes 52 job titles)
38	Emp. condition	Employment condition of the accident culprit consisting of: governmental official, daily-paid, railway official, hourly, contracted

[a] Accidents which happened in stations have the same station before and after.

Appendix B. Relations examined within the accident database

Factor	Examined relation
Human error	– Job – Accident Type – Job – Site (Station/Block) – Job – Years of Service – Districts – Accident Type – Job – Years of Service – Districts – Accident Grade – Job – Years of Service – Day of Week – Accident Type – Job – Years of Service – Day of Week – Accident Grade

(*continued on next page*)

Appendix B (*continued*)

Factor	Examined relation
	– Employment Condition – Level of Education – Districts – Accident Type
	– Employment Condition – Level of Education – Districts – Accident Grade
	– Job – Age – Districts – Accident Type
	– Job – Age – Districts – Accident Grade
	– Job – Age – Day of Week – Accident Type
	– Job – Age – Day of Week – Accident Grade
	– Job – Districts – Day of Week – Accident Type
	– Job – Districts – Day of Week – Accident Grade
	– Job – Level of Education – Districts – Accident Type
	– Job – Level of Education – Districts – Accident Grade
Wagon	– Train Type – No. of Wagons (Train Length) – Districts – Accident Type
	– Train Type – No. of Wagons (Train Length) – Districts – Accident Grade
	– Train Type – Wagons Type – Districts – Accident Type
	– Train Type – Wagons Type – Districts – Accident Grade
	– Wagons Type – Tunnel – Districts – Accident Type
	– Wagons Type – Tunnel – Districts – Accident Grade
	– Wagons Type – Curve Radius – Districts – Accident Type
	– Wagons Type – Curve Radius – Districts – Accident Grade
Track	– Train Type – Gradient – Districts – Accident Type
	– Train Type – Gradient – Districts – Accident Grade
	– Train Type – Curve Radius – Districts – Accident Type
	– Train Type – Curve Radius – Districts – Accident Grade
	– Train Type – Tunnel – Districts – Accident Type
	– Train Type – Tunnel – Districts – Accident Grade
	– Train Type – Point – Districts – Accident Type
	– Train Type – Point – Districts – Accident Grade
	– Curve Radius – Speed – Districts – Accident Type
	– Curve Radius – Speed – Districts – Accident Grade
General	– Accident Factor – Districts – Accident Type
	– Accident Factor – Districts – Accident Grade
	– Accident Factor – Day of Week – Accident Type
	– Accident Factor – Day of Week – Accident Grade
	– Accident Factor – Day of Week
	– Accident Factor – Accident Type
	– Districts – Accident Type
	– Accident Factor – Site
	– Accident Type – Site
	– Districts – Site
	– Accident Factor – Speed
	– Speed – Accident Type
	– Districts – Accident Grade

References

Abugessaisa, I., 2008. Knowledge discovery in road accidents database – integration of visual and automatic data mining methods. International Journal of Public Information Systems 1, 59–85.

Anderson, T.K., 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis and Prevention. doi:10.1016/j.aap.2008.12.014.

Barai, S.K., 2003. Data mining applications in transportation engineering. Transport 18 (5), 216–223 (Vilnius: Technika).

Chang, L.Y., Wang, H.Y., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention 38, 1019–1027.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0: Step-by-step Data Mining Guide. SPSS Inc.

Chong, M.M., Abraham, A., Paprzycki, M., 2004. Traffic accident analysis using decision trees and neural networks. In: Isaias, Pedro et al. (Eds.), IADIS International Conference on Applied Computing, Portugal, vol. 2. IADIS Press, pp. 39–42. ISBN: 9729894736.

Chong, M., Abraham, A., Paprzycki, M., 2005. Traffic accident analysis using machine learning paradigms. Informatica 29, 89–98.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accident Analysis and Prevention 40, 1257–1266.

Giudici, P., 2003. Applied Data Mining: Statistical Methods for Business and Industry. John Wiley & Sons Inc.

Han, J., Kamber, M., 2006. Data Mining: Concepts and Techniques, second ed. Elsevier Inc.

Huang, T.M., Kecman, V., Kopriva, I., 2006. Kernel Based Algorithms for Mining Huge Data Sets. Springer, Berlin, Heidelberg, New York.

Larose, D.T., 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons Inc., Berlin.

Lee, D.H., Jeng, S.T., Chandrasekar, P., 2004. Applying data mining techniques for traffic incident analysis. Journal of the Institution of Engineers Singapore 44 (2).

Nefti, S., Oussalah, M., 2004. A neural network approach for railway safety prediction. In: 2004 IEEE International Conference on Systems, Man and Cybemetics.

RAI, 2006. <http://www.rai.ir/>.

Sohn, S.Y., Lee, S.H., 2003. Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea. Safety Science 41, 1–14.

Solomon, S., Nguyen, H., Liebowitz, J., Agresti, W., 2006. Using data mining to improve traffic safety programs. Industrial Management & Data Systems 106 (5), 621–643 (q Emerald Group Publishing Limited).

Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis and Prevention 39, 1267–1278.

Tesema, T.B., Abraham, A., Grosan, C., 2005. Rule mining and classification of road traffic accidents using adaptive regression trees. International Journal of Simulation 6 (10–11).

Tormo, M.T., Sanmartin J., Pace J., 2009. Update and improvement of the traffic accident data collection procedures in Spain: the METRAS method of sequencing accident events. In: Proceedings of the 4th IRTAD Conference. September, 2009, Korea.

Van der schaaf, T., Kanse, L., 2004. Biases in incident reporting databases: an empirical study in the chemical process industry. Journal of Safety Science 42 (1), 57–67.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. Accident Analysis and Prevention 39, 922–933.