Association Rules Mining for Railway Accident Causes Based on Improved HFACS

Xiaoqing Zeng[1(✉)], Haixiang Lin[1,2], Ran Lu[2], and Gu Min[3]

[1] Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China
zengxq@tongji.edu.cn
[2] School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[3] Shanghai Municipal Engineering Design Institute(Group) Co., Ltd., Shanghai 200092, China

Abstract. Aiming at the problem of full quantitative analysis of railway accident causes, an improved HFACS model is proposed. Firstly, based on the original HFACS model, an improved HFACS model for railway industry was constructed Secondly, based on the improved model and association rule algorithm, the causative factors of 504 railway accidents from 2008 to 2009 collected by a railway bureau were comprehensively quantified, the association rule base of accident causative factors was mined, and the data of accident causative factors were visualized. The results show that the four main causes of railway accidents include irregularities, inadequate safety inspection, inadequate awareness of safety responsibility and inadequate education of safety responsibility. Based on the improved HFACS causative correlation analysis method, the importance of technical factors can be enhanced. Finally, the author puts forward some solutions to the lack of safety responsibility consciousness of the key accident factors.

Keywords: Railway Accident · Improved HFACS Model · Causation Factors · Apriori Algorithm

The advantage of association rules lies in the visualization of association rules after mining the causative association rules of railway accidents. The close relationship among the causative factors of railway accidents can be shown more intuitively through images, and then further solutions can be proposed. In order to make the image easy to analyze, the first 20 strong association rules are visualized in this paper. Figure 2 shows a visualization of the top 20 strong association rules sorted by support.

1 Introduction

At present, with the rapid development of China's railway industry, more and more attention is paid to 'the occurrence of railway accidents. The occurrence of railway accidents has caused serious economic losses, casualties and so on, so it is urgent to analyze the railway accidents and draw lessons from them. The frequent occurrence of railway accidents seriously affects the competitiveness of China in the international

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
X. Zeng et al. (Eds.): ITASC 2022, LNEE 1042, pp. 196–204, 2023.
https://doi.org/10.1007/978-981-99-2252-9_18

railway. Therefore, the prevention of railway accidents is imperative. HFACS (Human Factors Analysis and Classification System), proposed by Wiegmann et al., is an accident cause model for safety accidents in aviation field. The model analyzes the failure factors at the four levels of unsafe behavior, preconditions of unsafe behavior, supervision of unsafe behavior and organizational influence in detail. Therefore, this model is not only applied in the aviation field, but also in the navigation industry, medical industry, coal mining industry and railway industry.

In the field of railway industry, Australian scholar Lisa Punzet et al. used HFACS applicable to the railway industry to analyze the investigation report (N = 35) of SPADS by the Australian Railway Fredding Organization to check the human factors involved in the incident and determine the future trend. British scholar Madigan Ruth et al. used HFACS model to understand the relationship between active factors and potential factors of railway safety accidents, as well as specific causal paths. In China, Zhan Qingjian applied the HFACS model to the analysis of railway accidents in China and identified the main causes of accidents. Chen Ruiwei uses HFACS model to conduct qualitative analysis on the hazard sources of high-speed railway traffic dispatching system from three perspectives of "man-machine-loop". Because the traditional HFACS model is not fully applicable to the railway industry, and the research on the HFACS model in the railway field is not in-depth at home and abroad. Therefore, an improved HFACS model suitable for railway field is proposed.

2 Improvement of HFACS Model

Because the railway accident is caused by the interaction of human, machine, environment, management and other factors. However, the traditional HFACS mainly analyzes the causes of accidents from the human aspect, without taking into account other important accident drivers, such as technical aspects: equipment ineffectiveness, equipment design defects and so on, which have a great impact on the occurrence of railway accidents. As a result, the traditional HFACS model is designed for accidents in the aviation field, but is not applicable to the railway field in some aspects. Therefore, it is necessary to improve the traditional HFACS.

After the statistics and analysis of 504 railway accidents, the traditional HFACS was improved. The improved HFACS is used to analyze the leading factors of railway accidents from four levels: unsafe behavior, premise of unsafe behavior, dereliction of duty by relevant railway departments, and organizational influence. The premise of unsafe behavior is again divided into human factors, technical factors, environmental factors. Raising the technical environment of the original environmental factors to a level reflects the importance of technical factors. The dereliction of duty by relevant railway departments has replaced unsafe supervision, which further reflects the intensity of railway management and control. In terms of organizational influence, it is further divided into three aspects: professional training, working conditions, and information communication. Figure 1 shows the improved HFACS model.

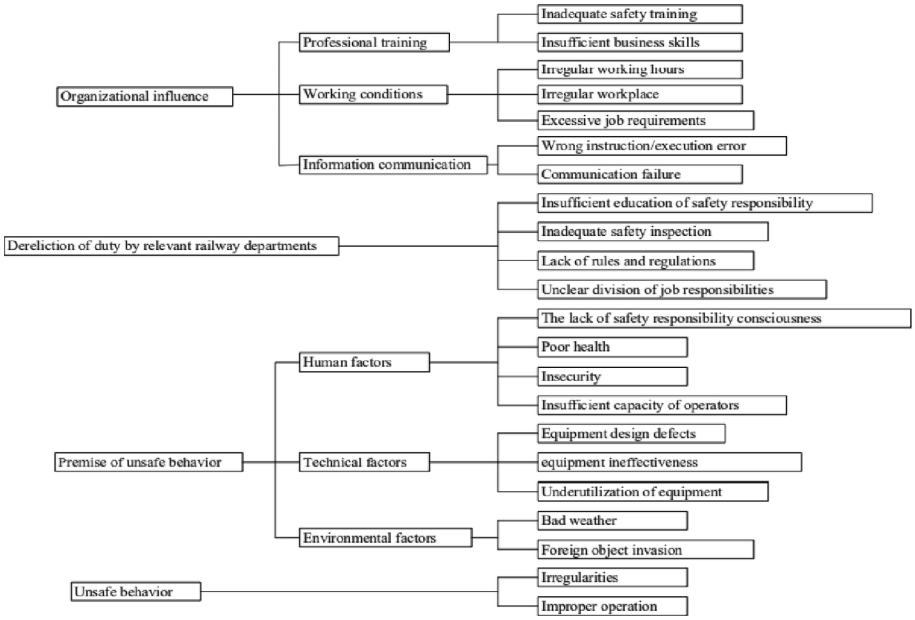

Fig. 1. Improved HFACS model

3 Correlation Analysis of Causative Factors of Railway Accidents Based on Improved HFACS

3.1 Relationship Between Causative Factors of Railway Accidents And Association Rules

The railway system is a comprehensive system composed of many links. The accidents are often caused by the interaction of some links. The occurrence of accidents includes a variety of factors, some of which are closely related, and when problems occur in some factors, the chain reaction will lead to problems in other factors, which will lead to the occurrence of a specific accident. Therefore, the association rules can be applied to the analysis of the causative factors of railway accidents, and the correlation between the causative factors of relevant accidents can be mined out. Through the visualization of association rules, the relationship between the causes of railway accidents can be visually displayed, and the key factors leading to railway accidents can be found out.

3.2 Association Rules

Association rules are a method to mine the relationship between variables in a data set. Related definitions of association rules are as follows:

Definition 1: Item and itemset.
Assume that D is the data set of railway accidents, that is, the transaction set; I is the set of all the cause factors of railway accidents in D, that is, the itemset; T is all the cause

factors of a certain railway accident. One or more items in each transaction are included in itemset I, namely T ∈ I.

The expression form of association rules is A = > B, where A and B are both contained in I, and A ∩ B = Ø, A is the Antecedent, and B is the Consequent.

Definition 2: Support and confidence
Usually, support and confidence are used as the measurement standards of association rules. For itemset A, if Count (A) is equal to all transaction sets containing itemset A; at this time, the support of A is:

$$Support(A) = \frac{Count(A)}{|D|} \tag{1}$$

Similarly, A = > B's support for the number of transaction contains both A and B Count (A, B) divided by the total number of transactions |D|.

$$Support(A \Rightarrow B) = \frac{Count(A, B)}{|D|} \tag{2}$$

At this moment, the support indicates the probability that the item set A and B appear together. For the association rule A = > B, the confidence level Conf(A = > B) refers to the ratio of the itemset containing A and B to the containing item set A in the total transaction set D.

$$Confident(A \Rightarrow B) = \frac{Support(A \Rightarrow B)}{Support(A)} \tag{3}$$

Confidence indicates the probability of including B under the premise of including A.

Definition 3: Lift Since the confidence only considers the support of the antecedents of the rules and does not consider the support of the Consequents of the rules, there will be misleading association rules. Therefore, lift is introduced to remove misleading association rules.

Lift is the ratio of the probability of including B if including A to the probability of occurrence of B in transaction set D.

$$Lift(A \Rightarrow B) = \frac{P(B|A)}{P(\overline{P})} = \frac{Confident(A \Rightarrow B)}{Support(\overline{B})} \tag{4}$$

If Lift > 1, A and B are positively correlated; If Lift < 1, then A and B are negatively associated and these negative association rules are removed. The higher lift, the greater the influence of A and B.

Definition 4: Frequent itemsets
The minimum support, namely Min-sup, is the measurement standard set by the user. If the support of A itemsets is not less than the minimum support, then A itemsets are called frequent itemsets. If B itemsets are contained in A itemsets and A itemsets are frequent itemsets, then B itemsets are frequent itemsets; if B itemsets are included in A itemsets and B itemsets are not frequent itemsets, then A itemsets not frequent itemsets.

3.3 Apriori Algorithm

Apriori algorithm is a common algorithm for association rules. The mining steps of the algorithm are divided into two parts: finding out all frequent itemsets and generating association rules. The specific process of the algorithm: (1) Find out all frequent item sets: it is an iterative method of searching for exclusion layer by layer to find out all frequent item sets. There are two steps: first, the transaction set is scanned to determine the occurrence times of item sets containing the same element, and the itemsets that do not satisfy Min-sup are removed. Second, iterate until no maximum itemsets appears. For example, in the K step, the K-1 item set obtained in the K-1 step generates the candidate k-frequent set. The transaction set is scanned to determine whether the support degree of the candidate itemsets K-1 is greater than Min-sup, and the itemsets less than the minimum support is removed to find the k-frequent itemsets. The above is the connection step and pruning step.

(2) Generate association rules: the frequent itemsets mined in the previous step is used to set the minimum confidence Min-conf to mine association rules. The pseudocode of frequent itemsets found by Apriori algorithm is shown in Table 1.

Table 1. The pseudocode of frequent itemsets found by Apriori algorithm

Input:transaction set D, minimum support Min-sup	
1	$L_1 = \text{find\_frequent\_1\_itemsets(D)}$
2	$\text{For}(k = 2; L_{k.1} ! = \emptyset; k + +)\{$
3	$C_k = \text{Apriori\_gen}(L_{k\text{-}1})$
4	For each transction $t \in D\{$
5	$C_t = \text{subset}(C_{k,t})$
6	For each candidate $c \in C_t$
7	$c.\text{count} + +$
8	$\}$
9	$L_K = \{c = C \mid c.\text{count} \geq \text{Min\_sup}$
10	$\}$
11	Return $L = U_k L_k$

CK is the set of candidate item sets of length k, and LK is the set of frequent itemsets of length k.

3.4 Case Analysis

3.4.1 Establish the Database of Railway Accident Causing Factors

Based on the improved HFACS model in this paper, 504 railway accidents from 2008 to 2009 were coded and analyzed. Taking "Guangzhou East Railway Station D725 train for preparing to enter the general C category accident of trains" as an example to classify

and code the causes of the accident. It can be determined that this causal factor is coded as "1"; otherwise, it is "0". Table 2 shows the classification and coding of the case report of Guangzhou East Railway Station D725 train for preparing to enter the general C category accident of trains".

Table 2. The classification and coding of the case report of Guangzhou East Railway Station D725 train for preparing to enter the general C category accident of trains"

Improve the classification of HFACS causative factors	Coding
Inadequate safety training X1	1
Insufficient business skills X2	1
Irregular working hours X3	0
Irregular workplace X4	0
Excessive job requirements X5	0
Wrong instruction/execution error X6	0
Communication failure X7	1
Insufficient education of safety responsibility T8	1
Inadequate safety inspection T9	1
Lack of rules and regulations T10	0
Unclear division of job responsibilities T11	0
The lack of safety responsibility consciousnessR12	1
Poor health R13	0
Insecurity R14	0
Insufficient capacity of operatorsR15	1
Equipment design defectsR16	0
Equipment ineffectivenessR17	0
Underutilization of equipmentR18	0
Bad weatherR19	0
Foreign object invasionR20	0
IrregularitiesO21	1
Improper operationO22	1

As can be seen from Table 1, this case can be represented by a Boolean matrix of 1 × 22 dimensions, as shown in Formula (5):

$$\left[1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\right] \tag{5}$$

Similarly, a 1 × 22-dimensional matrix of 504 cases of all railway accidents can be obtained. Finally, all the Boolean matrices of 1 × 22 dimensions are combined into a Boolean matrix of 504 rows × 22 columns, namely the railway accident data set RA-D,

as shown in Formula (6):

$$RA-D=\begin{bmatrix} X_{1,1}\mathrm{L} & R_{1,12} & \mathrm{L} & O_{1,22} \\ \mathrm{MO} & \mathrm{M} & \mathrm{O} & \mathrm{M} \\ X_{504,1}\mathrm{L} & R_{504,12}\mathrm{L} & O_{504,22} & \end{bmatrix} \tag{6}$$

This data set clearly represents the causative factors of each railway accident, which provides a basis for mining association rules later.

3.4.2 Mining the Relationship Between the Factors Causing Railway Accidents with Small Data

The Apriori algorithm is applied to the data set RA-D composed of Boolean matrices, where each row represents an accident case and each column represents an accident causative factor item. Through the iterative test, the minimum support is 0.03, the minimum confidence is 0.1, and the minimum lift is 1. For the sake of analysis, only consider the maximum Antecedent term to be 2. Finally, a total of 211 rules were generated from the data set of the causes of railway accidents. Among them, 91.8% of the rules have a support between 0.03 and 0.09. {Improper operation} = > {The lack of safety responsibility consciousness}" has the largest support, with a value of 0.11; There is 82.6% confidence between 0.1 and 0.8, and the greater the confidence, the fewer the rules, "{The lack of safety responsibility consciousness, Inadequate safety training} = > {Inadequate safety inspection}" the highest confidence, the value is 1. At the same time, 79.4% of lift were between 1 and 11. Table 3 shows the top 5 association rules of lift.

Table 3. The top 5 association rules of lift

Aules	Support	Confident	Lift
{Wrong instruction/execution error, Irregularities} = > {Unclear division of job responsibilities}	0.037	0.354	13.528
{Improper operation, Communication failure} = > {Insufficient capacity of operators}	0.055	0.461	13.006
{Inadequate safety inspection, Communication failure} = > {Improper operation}	0.0408	0.381	12.803
{Inadequate safety inspection, Insufficient capacity of operators} = > {Improper operation}	0.037	0.400	12.803
{Bad weather} = > {Inadequate safety inspection}			

The advantage of association rules lies in the visualization of association rules after mining the causative association rules of railway accidents. The close relationship among the causative factors of railway accidents can be shown more intuitively through images, and then further solutions can be proposed. In order to make the image easy to analyze, the

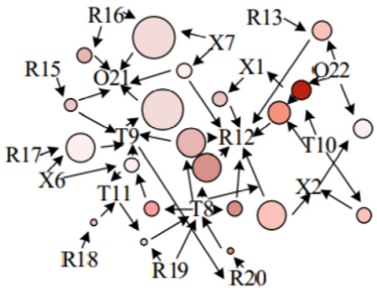

Fig. 2. A visualization of the top 20 association rules sorted by support

first 20 strong association rules are visualized in this paper. Figure 2 shows a visualization of the top 20 strong association rules sorted by support.

It can be seen from Fig. 2 that among these 20 rules, the four railway accidents that cause O21 irregularities, T9 inadequate safety inspection, R12 the lack of safety responsibility consciousness, and T8 insufficient education of safety responsibility are in the middle of the visual image. Five rules point to irregularities, four rules point to inadequate safety inspections, six points to the lack of safety responsibility consciousness, and four points to insufficient education of safety responsibility. It shows that the above four accident-causing factors are closely related to other accident-causing factors and have a high degree of support, so they can be determined as the four source factors, which basically run through all railway accident cases. When these four source factors exist, it is very likely that unsafe behaviors, confusion and inadequacy of management, and related unsafe psychology will occur before and during train operation, leading to accidents. For example, The lack of safety responsibility consciousness often leads to improper operations. The occurrence of Foreign object invasion is often accompanied by factors such as inadequate security inspections. Therefore, taking corresponding measures for these four types of railway accidents can effectively reduce the occurrence of railway accidents from the source and ensure the safety of railway operation. The lack of safety responsibility consciousness, as a relatively important one of all key factors, occupies an extremely important position in the prevention and control of railway accidents. Therefore, relevant railway departments should strengthen safety education, improve corresponding rules and regulations, and give deep criticism to those responsible for safety. Only when safety education is implemented can the safety awareness of all parties be improved and the occurrence of accidents can be reduced.

4 Conclusion

(1) Based on the HFACS classification of the causes of aviation safety accidents, combined with the reality of the railway industry, the original HFACS model was improved, and a new HFACS model conforming to the railway field was established. Raise technical factors to a relatively important position. And under the improved HFACS classification framework, the collected 504 railway accident cases were coded, and the railway accident data set RA-D was established.

(2) Use Apriori algorithm of the association rules to data mine the causes of railway accidents, and visualize part of strong association rules set. Four major factors were discovered, which are irregularities, inadequate safety inspection, lack of safety responsibility consciousness, and insufficient education of safety responsibility. These four accident-cause factors are closely related to other causes of accidents, and are the root factors of other causes. Propose corresponding improvement measures for the lack of a strong sense of safety responsibility for the key cause of the accident.

References

1. Li, C., Tang, T., Chatzimichailidou, M.M., Jun, G.T., Waterson, P.: A hybrid human and organisational analysis method for railway accidents based on STAMP-HFACS and human information processing. Appl. Ergon. **79** (2019)
2. Punzet, L., Pignata, S., Rose, J.: Error types and potential mitigation strategies in Signal Passed at Danger (SPAD) events in an Australian rail organisation. Safety Sci. (2018)
3. Qingjian, Z: HFACS-RAs based railway accident casual factor modeling and hybrid learning approach. Beijing Jiaotong University (2017)
4. Ruth, M., David, G., Richard, M.: Application of human factors analysis and classification system (HFACS) to UK rail safety of the line incidents. Accid. Anal. Prev. **97** (2016)
5. Hardianto, I., Fitri, I.Z.: Indonesian railway accidents--utilizing human factors analysis and classification system in determining potential contributing factors. Work (Reading, Mass.) 41 Suppl 1 (2012)
6. Guochen, Z.: Research on EW-LDA railway accident contributor based method. Beijing Jiaotong University (2019)
7. Liu, Y., Liu, Y., Ma, X., Qiao, W.: A comprehensive model for human factors evaluation in maritime accident: HFACS and FAHP. In: International Informatization and Engineering Associations. Proceedings of 2019 2nd International Conference on Financial Management, Education and Social Science(FMESS 2019), pp. 247–253. International Informatization and Engineering Associations: Computer Science and Electronic Technology International Society, China (2019)
8. Ruiwei, C.: Hazard identification on operation system of high-speed railway. Southwest Jiaotong University (2014)