ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 195-203
doi: 10.18178/wcse.2020.06.031

Research and Realization of A Three-Level Label System for Railway Locomotive Equipment Portrait+

Xin Li[1 +], Tianyun Shi[2], Bao Chang[2], Xiaoning Ma[2], Jun Liu[2]

[1] Postgraduate Department, China Academy of Railway Sciences, Beijing, China

[2] Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing, China

Abstract. Locomotives are important transportation production equipment for railway industry. Carrying out equipment portrait analysis and label system research is an important way to promote the development of big data technologies in railway locomotive system. Firstly, by briefly summarizing the theoretical knowledge of portrait technology and label technology, this paper researches and designs a technical framework suitable for locomotive label system and locomotive equipment portrait. Secondly, based on this technical framework, the three-level label system of locomotives is introduced in detail. Finally, an optimized clustering algorithm for centroid selection is proposed to form a complete and reasonable locomotive label system. The related research of locomotive label system can help to realize the applications of locomotive equipment portrait, improve the railway transportation production efficiency, and strengthen the safety management ability.

Keywords: big data, portrait technology, label technology, locomotive, clustering algorithm

1. Introduction

As important equipment assets of railway industry, the safety status of locomotives will directly affect the operation efficiency and safety level of railway transportation production [1-2]. Therefore, it is particularly important to make good use of data of locomotives to form the locomotive label system, and then carry out the locomotive equipment portrait research to analyse the locomotive quality reliably.

However, big data analysis based on locomotive equipment portrait is still in its infancy, and the practical application of big data technologies still faces some difficulties in the railway locomotive system [3-4]. First of all, a large amount of data is still scattered in various information systems and is not fully integrated. Secondly, big data technologies have not been deeply applied, which leads to insufficient depth of data analysis. Therefore, it is necessary to make full use of all kinds of data of locomotives to design the locomotive label system, and then research the equipment portrait technology for locomotives, so as to realize the benign interaction between data management and production practice, which is conducive to equipment management, accident prevention and safety decision-making.

The content of this paper is organized as follows. The second part briefly introduces the concept and principle of equipment portrait and label technology. Section 3 researches and designs the technical framework of locomotive label system, and makes a detailed description. The fourth part researches and puts forward a optimal clustering algorithm suitable for the generation of locomotive labels, and verifies the algorithm, then finally forms the complete locomotive label system. Section 5 summarizes the main research results of this paper and introduces the related research plans in the future.

+ Foundation item: The Scientific Research and Development Project of China Railway Corporation (K2018S007).
+ Corresponding author. Tel.: +86 13261316600.
E-mail address: florian_lee@163.com.

2. Equipment Portrait and Label Technology

2.1. Equipment Portrait

Portrait technology is one of the effective big data analysis tools [5]. By referring to the concept and application of Personas [6], it is possible to research the equipment portrait of railway equipment. The formation of equipment portrait can be summarized as follows: Based on the characteristics of the target equipment, all kinds data of the target equipment are collected actively or passively, then the useful information from these data is extracted as portrait labels [7], then the abstract equipment portrait model is built by constructing precise, fine-grained and structured label system to describe the properties, features and performance of the equipment [8], and then the data mining and analysis methods are used to analyse the state of the equipment.

In terms of locomotives, the vast amounts of data are integrated from various business areas of the railway locomotive system, such as locomotive operation, running preparation, repair, special examination, etc., and then these data are converted into the locomotive labels by a series of means of data processing and analysis to make the locomotive state easier to be grasped, so as to improve the application effect of locomotive equipment portrait, reduce the difficulties of data analysis, and promote the deep combination of big data technologies and locomotive safety management.

2.2. Label Technology

Labels are very refined features based on actual conditions [9]. The key content of portrait technology is to label the target [10-11], so as to describe the real state of the object in the form of label. Labels have three basic characteristics: “artificial definition”, “semantic information” and “short text” [12]. Based on the artificial generalization and definition, a label has the explicit and unique semantic meaning, which no longer needs the excessive pre-processing such as text analysis. This facilitates the calculation, analysis and visualization of information by computers.

Labels can be divided into three categories: “basic attribute labels”, “dynamic behavior labels” and “comprehensive evaluation labels” [13]. The basic attribute labels describe the intrinsic and static properties of an object, which remain unchanged or change slowly over time. The dynamic behavior labels reflect the real-time status information of the research object, such as running time, running preparation situation, fault information and so on. Such labels are the important part of a complete portrait. The comprehensive evaluation labels are obtained by summarizing the characteristics of the target with the data accumulation.

In order to realize a more accurate portrait, labels should be as rich as possible to form a label system [14]. In addition, in order to facilitate analysis and calculation, the structure of the label system should be fixed and unified in form [15]. The structure of the label system needs to be determined according to the characteristics of the research object and the volume of the data.

3. Locomotive Label System

3.1. Technical Framework of the Locomotive Label System

The technical framework of the locomotive label system is divided into three layers, namely “data fusion layer”, “label library layer” and “label application layer”, as shown in Fig. 1. With the help of big data mining algorithms, all kinds of data about locomotives are integrated to carry out the whole life cycle management of locomotive labels, so as to satisfy relevant analysis requirements of locomotive transportation production.

3.1.1. Data Fusion Layer

By means of system docking, data entry, batch import, etc., the data fusion layer collects basic information, operation information, repair information, accident and fault information, quality evaluation information, cross-domain information and other locomotive data to provide data sources for subsequent links.

The basic information is static data, which can be kept unchanged for a long time and can be easily obtained. Data such as operation information, running preparation information, repair information and fault information are dynamic data. These data basically cover all links in the daily transportation production of locomotives. These data can be obtained by means of system docking and manual dump. Safety analysis information and

quality evaluation information belongs to the comprehensive evaluation data. Cross-domain information is a kind of data related to locomotive safety that exists in other professional fields, such as weather information, geographic information, power supply information, etc.

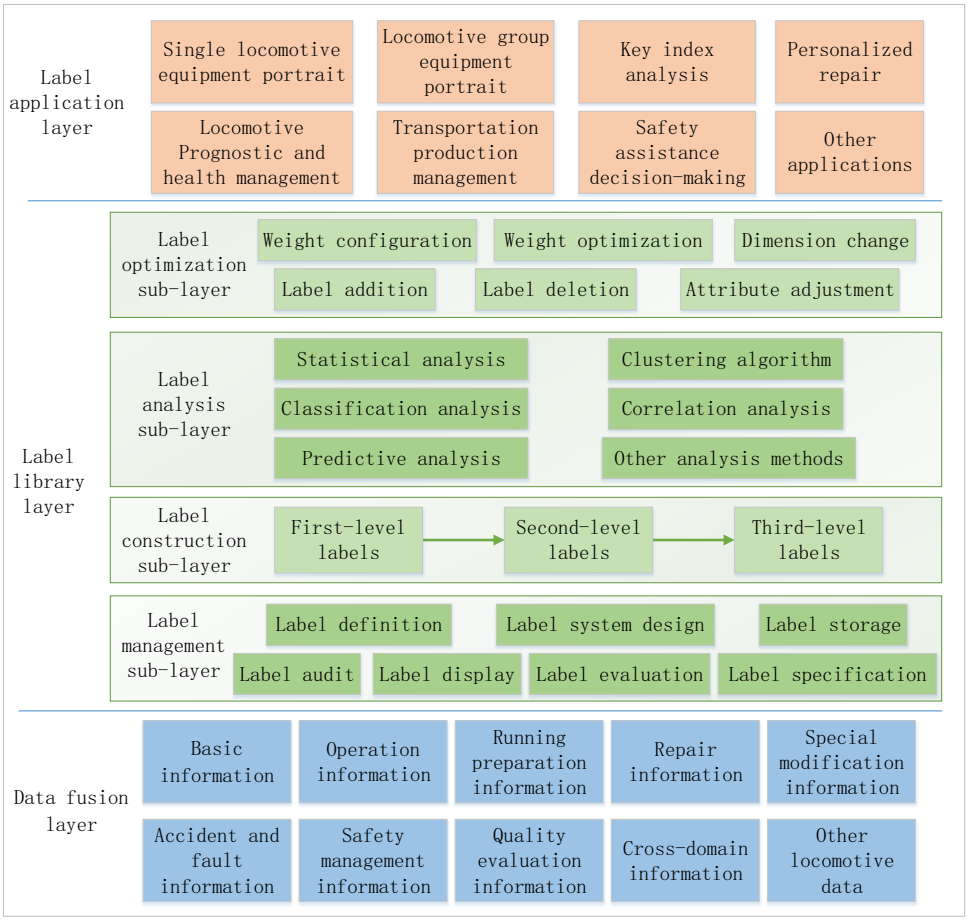


Fig. 1: The technical framework of the locomotive label system

3.1.2. Label Library Layer

The label library layer makes use of the data provided by the data fusion layer to generate labels by means of a series of methods such as data cleaning, data fusion and data mining. The label library layer contains four sub-layers: “label management”, “label construction”, “label analysis” and “label optimization”.

The label management sub-layer aims to achieve whole life cycle management of labels. Based on label specification, this sub-layer provides services for label generation, storage and modification through label definition, system design, data storage, audit, display, evaluation and verification, etc.

According to label definition and design rules, label construction sub-layer processes various data, and forms specific labels, then constructs the label system for locomotive equipment portrait. The locomotive label system can be designed as three-level label system, which is “first-level label”, “second-level label” and “third-level label”. The first-level labels and the second-level labels are abstract summaries of the third-level labels and have no specific use significance. The third-level labels contain all the actual labels used to reflect the real characteristics of locomotives.

Label analysis sub-layer includes statistical analysis, clustering analysis, classification analysis, correlation analysis, predictive analysis and other data mining methods. It provides analysis algorithms for label generation, optimization and utilization.

The label optimization sub-layer is responsible for label iterating. It mainly includes weight configuration, weight optimization, dimension change, label addition, label deletion, attribute adjustment and so on.

The above four sub-layers cooperate with each other to ensure the continuous improvement of the label system.

3.1.3. Label Application Layer

The label application layer provides a series of label application services for locomotive transportation production, such as locomotive equipment portrait, key index analysis, personalized repair, safety assistance decision-making and so on.

Single locomotive equipment portrait is the basic function of the label application layer. Based on the locomotive labels generated in the label library layer, it depicts the locomotive equipment characteristics and safety status comprehensively, objectively and accurately. On this basis, the equipment portrait of a certain locomotive group can be formed through statistics, clustering, classification and other analytical means from different analytical dimensions, such as locomotive types, whole depot locomotives, whole bureau locomotives, transportation line, transportation task, etc.

Furthermore, based on abundant labels, key index analysis can be carried out to meet production and management demands such as fault handling, running preparation scheduling, equipment modification, etc. The locomotive label system and equipment portrait analysis can provide stable data support for prognostic and health management (PHM) of locomotives, so as to improve locomotive quality management ability. In addition, it will be possible to change the planned repair of locomotives into personalized repair, which will promote accurate repair, refined control, consumption reduction and cost saving with more flexible locomotive repair plans, and finally realize the state repair of locomotives.

The locomotive label system can also provide data support for locomotive transportation production management and safety assistance decision-making. The locomotive labels are generated from the data accumulating in the daily production of locomotives, which can help to realize a benign closed loop between the production data and the locomotive management.

3.2. Three-level Label System of Locomotives

3.2.1. The First-level Label

The first-level label of locomotive reflects the basic analytical dimensions of locomotive equipment portrait. They are fixed in number and uniform in form. The first-level labels of locomotives can be sorted into several dimensions, such as basic information, operation quality, running preparation quality, repair quality, quality evaluation and so on. Among them, the basic information belongs to the “basic attribute label”, the operation quality, running preparation quality, repair quality, special rectification and data analysis belong to the “dynamic behavior label”, and the quality evaluation belongs to the “comprehensive evaluation label”.

3.2.2. The Second-level Label

As similar to the first-level labels, and the quantity and form of the second-level labels of locomotives are basically fixed, which reflect the detailed analytical dimensions. These dimensions include basic features, operational faults, running preparation downtime, repair focus, special rectification, quality evaluation, etc.

In each dimension of the first-level labels, different second-level labels can be obtained. The basic information is divided into basic feature and operation feature. The operation quality is made up of 10 second-level labels, such as locomotive daily running distance, locomotive break number, locomotive break type, operation fault number, temporary repair fault type, odd repair number, etc. The running preparation quality includes 6 second-level labels, such as running preparation downtime, running preparation duration, special inspection problem number, special inspection problem type, safety device problem number and so on. The repair quality covers 9 second-level labels, including repair classification, quality identification problem number, quality identification problem type, performance test problem number, trial operation problem number, etc. The special rectification includes 4 second-level labels, such as spring appraisal grade, other rectification projects and so on. The data analysis is consist of 6A analysis times and running gear analysis times. The quality evaluation has 4 second-level labels, such as overall grade, key device and safety trend, etc.

3.2.3. The Third-level Label

The third-level labels are the personalized information that reflects the actual quality status of a locomotive and represents the specific content of the second-level labels of a locomotive. Therefore, the number and the

content of the third-level labels varies from locomotive to locomotive. The construction process of the locomotive label system is the obtaining process of the third-level labels.

There are three ways to get the third-level labels of a locomotive:

The first method is direct acquisition. These labels are mainly used to describe the inherent attributes of locomotives, which can be directly obtained from the database without too much calculation, such as service date, manufacturer, running line, etc. Most of these labels belong to the first-level label "basic information".

The second method is statistical calculation. Such labels can be obtained by simple logic operations, such as fault type, fault number, key parts, etc. These labels mostly belong to dynamic behavior labels, which reflect the real-time state of locomotives.

The third method is mining analysis. Such labels cannot be obtained through simple calculation, but can be generated with the help of data mining algorithms, such as clustering, classification, prediction, correlation analysis, etc. This method is used to produce most of the third-level labels for locomotives.

4. Generation Method of the Third-level Labels Based on Clustering

4.1. An Optimized Clustering Algorithm

K-means algorithm is a distance-based clustering algorithm [16], which uses distance as the similarity index. K-means algorithm divides the data into pre-set K clusters based on the minimization error function, and takes compact and independent clusters as the ultimate goal.

The process of k-means algorithm is as follows. Firstly, K samples are randomly selected as the initial centroids $\mu_i (i \leq K)$ from all samples. Then the distance between each sample x and the K centroids is calculated separately, and the sample is assigned into the corresponding cluster C_i of the clustering centroid with the minimum distance. Furthermore, the centroid of each cluster is recalculated as the next new centroid $\mu_i = \frac{1}{|\mu_i|}\sum_{x \in C_i} x$. Finally, the process keeps repeating until the centroids of the clusters stop changing.

For the samples of Euclidean space, the sum of square errors (SSE) is taken as the objective function of clustering, and is taken as the index to measure the clustering effect [17]. The smaller the SSE is, the more similar the samples in the cluster are. The optimal clustering result should produce the minimum SSE.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

Although k-means algorithm has the advantages of fast and efficient, it is a locally optimal clustering algorithm [18], and its clustering effect depends heavily on the initialization centroids [19-20]. This problem can be solved by optimizing the selection of initial centroids.

In order to avoid the locally optimal clustering, the distance between the centroids should be as far as possible when the initial centroids are selected. The basic steps to select the optimized centroids are as follows.

a. A sample x_i is randomly selected from the data set $U = \{x_1, x_2, \ldots, x_n\}$ as the first cluster centroid μ_1.

b. The shortest distance $D(x)$ between each sample and the selected cluster centroids is calculated. Then, the sample x_i with the largest distance value $D(x_i)$ is selected as the next cluster centroid.

c. Repeat step b until all cluster centroids are selected. Then, the subsequent clustering processes are the same as K-means algorithm.

Making the distance between the initial centroids as large as possible can significantly improve the final clustering result. Although it takes more time to select the initial centroids, the convergence speed of the clustering process and the stability of the algorithm can be improved.

4.2. Comparison of the Two Algorithms

The classic dataset IRIS is selected to compare the clustering effects of the two algorithms described above, as shown in Fig. 2.

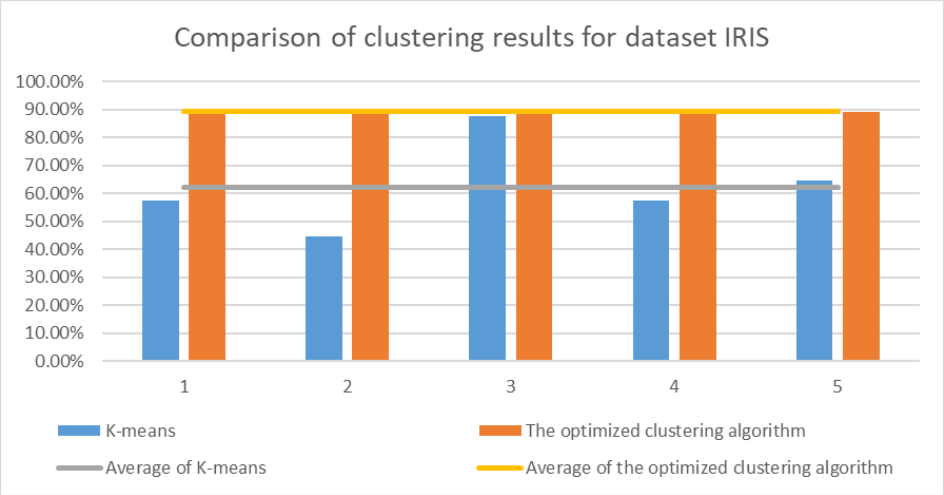


Fig. 2: Comparison of clustering results between K-means and the optimized clustering algorithm.

After 5 times of clustering calculation, the accuracy of K-means algorithm is 57.33%, 44.67%, 87.64%, 57.33% and 64.67% respectively, while the accuracy of the improved clustering algorithm can always be maintained at 89.33%.The average accuracy was 62.33% and 89.33%, respectively.

It can be seen that the optimized clustering algorithm which ameliorates the selection of the initial centroids can improve the accuracy and stability of the clustering algorithm, and the clustering effect is better.

4.3. A Selection Method for K

When K-means algorithm or the optimized clustering algorithm is used to analyze the data, the cluster number K should be determined first. However, it is often difficult to ascertain the suitable K at the beginning of clustering, that is, the exact number of the third-level labels for a label dimension is uncertain. Therefore, it is necessary to compare the clustering effects of different K and then select the appropriate one.

The value of SSE of the cluster samples will gradually decrease with the increase of the number K [21]. When K is less than the actual number of clusters, SSE will decrease rapidly with the increase of K. When K is larger than the actual number of clusters, SSE will decrease slowly with the increase of K. In fact, the number of clusters can be selected as the K that causes SSE to decline slowly instead of rapidly.

The “locomotive temporary repair number” is taken as the example to test the relationship between SSE and K, as shown in Fig. 3. In consideration of the actual production situation and the locomotive equipment portrait update cycle, the dataset is made up of the number of locomotive temporary repair per locomotive per month from January 1, 2019 to December 31, 2019 of a railway bureau, and the data volume is 735.

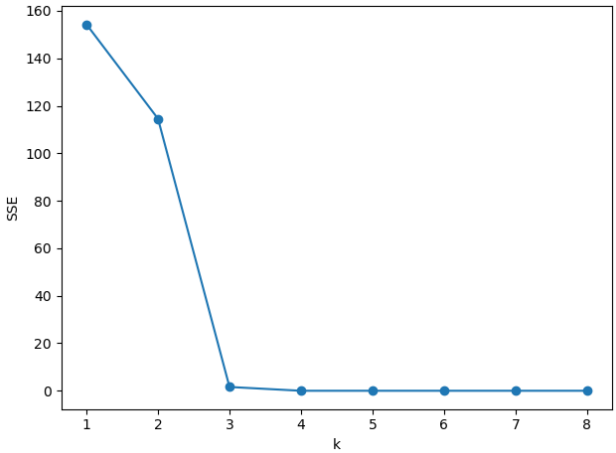


Fig. 3: Relation curve between SSE and K of the locomotive temporary repair number of a railway bureau.

As can be seen from Fig. 3, when K=3, SSE converts to a slow downward trend. Then, the “elbow method” [22] can be used to take 3 as the clustering number. The clustering results are shown in Tab. 1.

Tab. 1: The clustering results of the temporary repair number of a railway bureau when K=3.

Cluster	Number of samples	Centroid	Data range
1	585	1	[1,1]
2	140	2	[2,2]
3	10	3.2	[3,4]

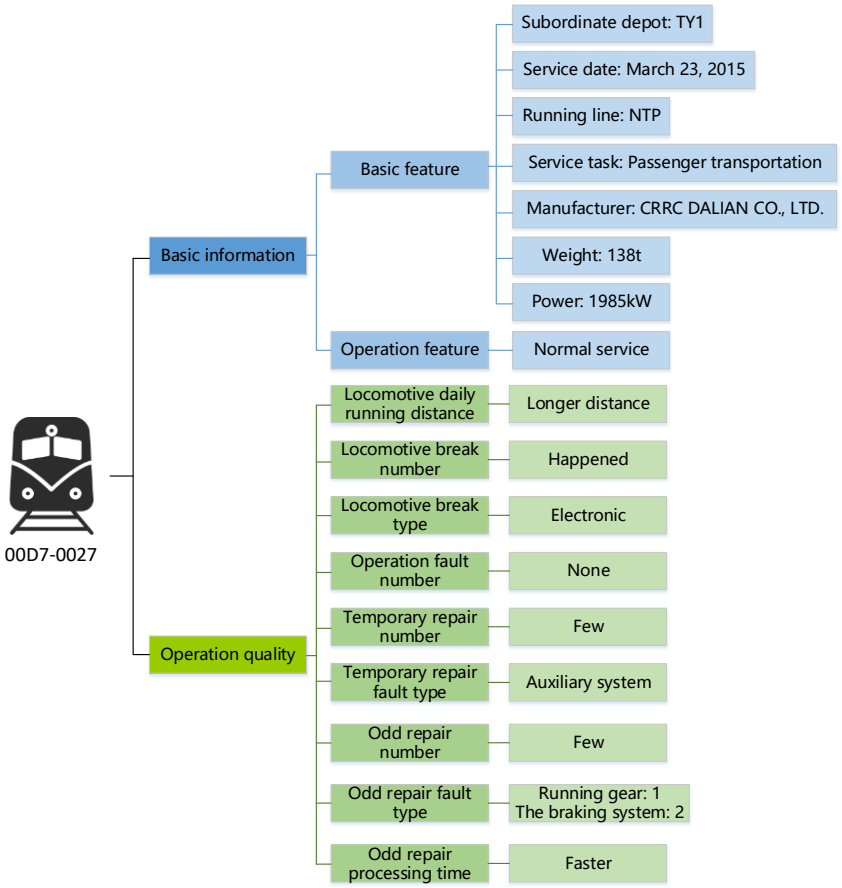
The data volume of the first cluster is 585, the data proportion is 79.6%, and the cluster centroid is 1.0. Based on this feature, the first label can be denoted as “Few”, indicating that it is a very small amount when only 1 temporary repair occurs. Then, the data volume of the second cluster is 140, and the cluster centroid is 2.0. Therefore, the second label can be denoted as “More”, implying that the number of the temporary repair is relatively large. Thirdly, there are 10 samples in the third cluster and the cluster centroid is 3.2. Based on this feature, the third label can be denoted as “Many”, representing that there is a large quantity of temporary repair, and the operation quality of locomotives is relatively poor in this dimension. In addition, a label “None” should be added to signify that the number of locomotive temporary repair is 0.

With the help of the improved clustering algorithm and the selection method of K, the third-level labels obtained by clustering can be better generated. Furthermore, the data range and meanings of the third-level labels of locomotives will be continuously adjusted with the accumulation of locomotive data.

4.4. Example of the Locomotive Label System

Based on the locomotive label system and the label acquisition method introduced above, the locomotive data of a railway bureau generated from January 1, 2019 till December 31, 2019 were analyzed, and the three-level label system applicable to the locomotive equipment portrait can be obtained.

A locomotive was randomly selected to show its label system, as shown in Fig. 4. For the sake of data security, part of the data has been desensitized.



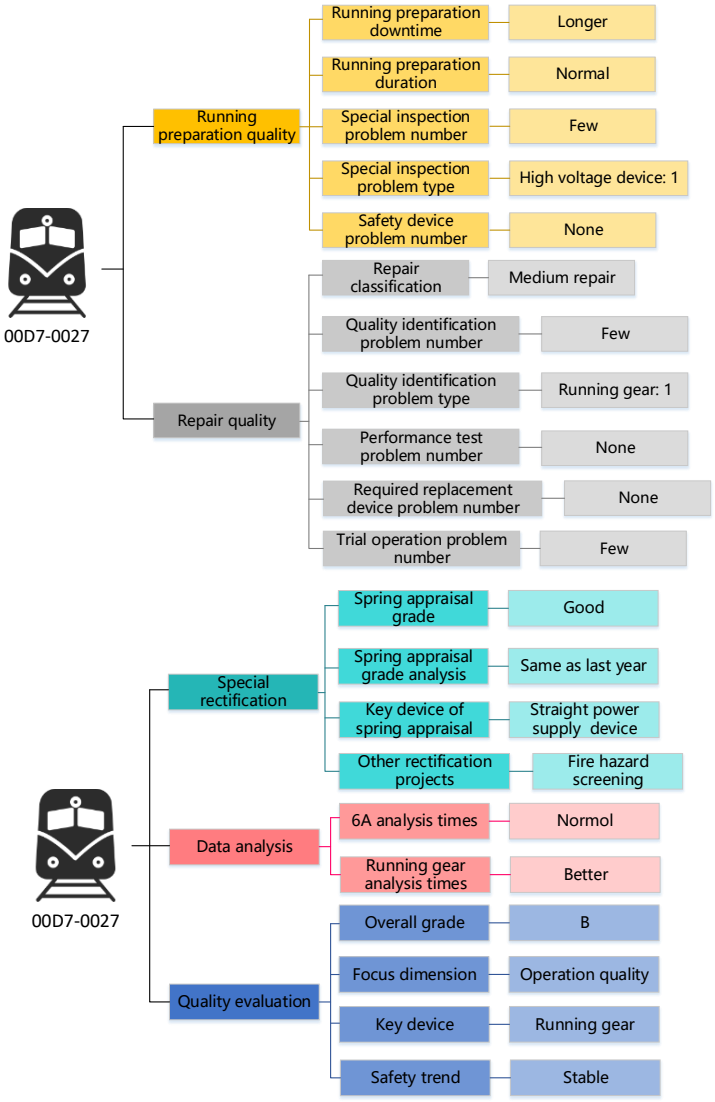


Fig. 4: The label system of a locomotive

5. Conclusion

The locomotive transportation capacity and equipment management level of the railway locomotive system is of great significance to railway transportation production. The construction of the locomotive label system and the analysis of the locomotive equipment portrait is conducive to better grasp the locomotive quality state and better meet the big data analysis demands of locomotive operation, running preparation, repair, scheduling and other businesses, which can provide technical support for the realization of efficient transportation, accurate repair and safety management of locomotives. This paper designs the technical framework of the locomotive label system for locomotive equipment portrait and other application scenarios, firstly. The internal logic of the technical framework is elaborated. Then, based on this technical framework, the three-level label system and label acquisition methods are introduced in detail. Moreover, an improved clustering algorithm using the optimized selection method of the initial centroids is researched, and the clustering effect is satisfactory. Finally, a selection method for cluster number K is introduced, and the definite third-level labels of a locomotive from a railway bureau are generated to form its complete locomotive label system. This lays a foundation for the analysis and further applications of locomotive equipment portrait.

On the basis of the technical framework of the locomotive label system formed in this paper, the next work will constantly focus on optimizing the locomotive label system, improving the application effects of the locomotive equipment portrait, and researching the realization methods of the locomotive group equipment portrait, key index analysis, personalized repair, etc., so as to provide powerful technical support for efficient and safe railway transportation production.

6. References

[1] ASITHYA THADURI, DIEGO GALAR, UDAY KUMAR. Railway assets: A potential domain for big data analytics [J]. Procedia Computer Science, 2015(53): 457-467.

[2] Bin Ning, Li Zhu, Yige Wang. Big Data Analytics in Intelligent Transportation Systems: A Survey [J]. IEEE Transactions on Intelligent Transportation Systems, 2019(1): 383-398.

[3] SHI Tianyun, LIU Jun, LI Ping, et al. Overall scheme and key technologies of big data platform for China Railway [J]. Railway Computer Application, 2016, 25(9): 1-6.

[4] WANG Tongjun. Overall framework and development prospect of intelligent railway [J]. Railway Computer Application, 2018, 27(7): 1-8.

[5] CAI Yi, LI Qing, XIE Haoren, et a1. Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy [J]. Neural Network, 2014, 58(10): 98-110.

[6] Zeng Hong, Wu Su Ni. User image and precision marketing on account of big data in Weibo [J]. Modern Economic Information, 2016, 24: 306-308.

[7] Golder Scott A., Huberman Bernardo A.. Usage patterns of collaborative tagging systems [J]. Journal of Information Science, 2006, 32(2): 198-208.

[8] ZHOU Huyong. A Design of User Portrait Tag System for Fusion Media [J]. Digitization User, 2019, 25(12): 212-214.

[9] ZHAO Yongzhu, MA Ji'ou, ZHANG Kexin. Research on the Label Portrait Technology Based on Life Cycle of Electricity Assets [J]. Advances of Power System & Hydroelectric Engineering, 2018, 34(1): 51-58.

[10] Wang Ye, Guo lingli, Song Wenchao, et al. Research on Personas Recommendation Algorithm Based on Big Data Technology [J]. Computer Measurement & Control, 2018, 26(12): 225-229.

[11] LI Na, FAN Zhengjie, HAO Chuanzhou, et al. A Method for Building Tag Systems Based on Semantic Feature Analysis [J]. Journal of Xi'an Jiaotong University, 2019, 53(1): 169-174.

[12] L Deng, J Gao, C Vuppalapati. Building a Big Data Analytics Service Framework for Mobile Advertising and Marketing [C]. IEEE First International Conference on Big Data Computing Service & Applications, 2015: 256-266.

[13] Wang Yang, Ding Zhigang, Zheng Shuquan DESIGN AND IMPLEMENTATION OF USER PROFILE SYSTEM [J]. Computer Applications and Software, 2018, 35(3): 8-14.

[14] Al Hasib Abdullah, Natvig Lasse, Cebrian Juan M.. A vectorized k-means algorithm for compressed datasets: design and experimental analysis [J]. Journal of supercomputing, 2018, 74(6): 2705-2728.

[15] WANG Jianren, MA Xin, DUAN Ganglong. Improved K-means Clustering k-Value Selection Algorithm [J]. Computer Engineering and Applications, 2019, 55(8): 27-33.

[16] YANG Junchuang, ZHAO Chao. Survey on K-Means Clustering Algorithm [J]. Computer Engineering and Applications, 2019, 55(23): 7-14.

[17] Hung C H, Chiou H M, Yang W N. Candidate groups search for K-harmonic means data clustering [J]. Applied Mathematical Modelling, 2013, 37(24): 10123-10128.

[18] Fengbo Kong, Hong Lai, Hailing Xiong. Quantum Hierarchical Clustering Algorithm Based on the Nearest Cluster Centroids Distance [J]. International Journal of Machine Learning and Computing, 2017, 7(5), 100-104.

[19] Mohamed Nour Elsayed, Monzer Mohamed Qasem. Analysis of Some Algorithms for Clustering Data Objects [J]. International Journal of Machine Learning and Computing, 2014, 4(1), 99-105.

[20] Pavani Y. De Silva, Chiran N. Fernando, Damith D. Wijethunge, et al. Recursive Hierarchical Clustering Algorithm [J]. International Journal of Machine Learning and Computing, 2018, 8(1), 1-7.

[21] CHENG Weiqing, LU Yanhong. Adaptive clustering algorithm based on maximum and minimum distances, and SSE [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2015, 35(2): 102-107.

[22] M A Syakur, M A Syakur, B K Khotimah, et al. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster [J]. IOP Conference Series: Materials Science and Engineering, 2018, 336(1).