Research on Construction Method for Automatic Classification of Group Enterprise R&D Resources Space

Lei Wang[1], Qingpeng Wang[1], Hongyu Shao[1], Li Li[2], and Sizhe Pan[1](✉)

[1] School of Mechanical Engineering, Tianjin University, Tianjin 300350, China
psizhe@tju.edu.cn

[2] AVIC Jiangxi Hongdu Aviation Industry Group Company Ltd., Nanchang 330024, China

Abstract. In order to improve the organization and management ability of R&D resources in the group enterprise, and realize the efficient utilization of R&D resources, this paper puts forward the design resource space automatic construction classification method based on the analysis of the space construction requirements of R&D resources in the group enterprise. Machine learning is used to extract keywords from R&D resources, and semantic maps are built based on the semantic similarity of keywords. Hierarchical clustering is implemented through a community detection algorithm. Then, based on the BRET-LEAM model, the mapping of R&D resources to resource space is achieved, and the construction of multidimensional R&D resource space automatic classification is finally completed. Finally, an instance of group enterprise R&D resources automatic classification is built.

Keywords: Group Enterprise · R&D Resources · Resource Space

1 Introduction

Group enterprises, as an organizational form with group headquarters as the center and multiple business subsidiaries, play an important role in promoting the economy of various countries, especially emerging markets [11] After years of development, China's aerospace, rail transit, ordnance equipment and other group enterprises have accumulated a large number of hardware, software and information R&D resources. These R&D resources are diverse and large in total, scattered in different organizations or systems, lacking unified management, low resource utilization rate and insufficient integration with product development process [13]. Therefore, it is necessary to establish the group R&D resource space to realize the integrated management, sharing and effective use of the group R&D resources.

At present, the R&D of resource management mainly focuses on the integration of resource representation, storage, index and query. Li Shaojun, Zhong Ershun [4] proposed a data grid technology, which links data resources, information resources and knowledge resources that are not evenly distributed in space into a logical whole. Liu Panpan, Wang Junyi [5] proposed an index library based on XML structure to realize educational resource retrieval. L Stojakovic [6] established the top-level ontology of online

N. Radojević et al. (Eds.): ICAID 2022, AHIS 7, pp. 521–529, 2023.
https://doi.org/10.2991/978-94-6463-010-7_54

learning resources and users, and used RDF triplet pattern to describe resource relations and form resource relation graph. Resource space model by Zhuge Hai [14] is put forward earliest, and its main idea is unified by one or more dimensions to define, share, organization and management of various resources, using semantic chain model realization of semantic interconnection network, by combining with resource space model, form a single semantic image, making all kinds of resources in unity and concise semantic space interconnection [14].

There are few researches on the centralized organization and management of group enterprise R&D resources. Based on the analysis of group R&D resource space construction requirements, this paper proposes a framework for automatic classification of R&D resource space. Through BRET-LEAM model, this paper completes the mapping of R&D resources to resource space, and realizes the automatic construction of group R&D resource space.

2 Demand of Group R&D Resource Space Construction

(1) Demand for unified and classified management of the group's R&D resources
With the R&D, production, maintenance and updating of products over the years, a large number of hardware, software and information R&D resources of the group are scattered in the heterogeneous organizational system. Its information is huge and complicated, and has dispersive, heterogeneous, diversity. Although there have been some relevant studies in classification model [8, 12], ontology model [1, 2], knowledge graph [3, 7, 10] and other fields, the lack of top-level unified classification management is not conducive to efficient and accurate positioning of R&D resources.

(2) Demand for automatic expression of group R&D resources
With the continuous growth and increasing complexity of resource information, it is more difficult to express and obtain unified R&D resources of group enterprises. It is necessary to consider the automatic extraction method of group R&D resource attributes to ensure the fast acquisition and expression of R&D resources.

(3) Demand of group enterprise R&D resources matching
The key to make effective use of R&D resources is to improve the efficiency of on-demand matching and push of R&D resources. Resource demanders can obtain needed resources quickly and efficiently when they are searching for resources, when they are dealing with new business activities or solving engineering problems.

In order to solve the above problems, it is necessary to study the automatic construction method of group enterprise R&D resource space. In order to realize efficient organization and management of resources, integration of resources and research and development process, play the core value of R&D resources in the product lifecycle.

3 Construction Method of the R&D Resource Space Automatic Classification

3.1 Framework for Automatic Classification of the R&D Resource Space

The framework for automatic classification of enterprise resource space is shown in Fig. 1. First, all R&D resource description files are integrated to form the original corpus, and then the axis and coordinates of resource space are extracted from the corpus. Then each R&D resource is mapped to the resource space to form a complete resource space classification containing R&D resource instances.

3.2 The Method of R&D Resource Space Generation

(1) Keyword extraction
Firstly, the resource description text in the R & D resource corpus is pre-processed to generate a pre-processed corpus. The pre-processing process includes word segmentation, stop word removal and special character removal using jieba word segmentation tool. And then used to generate the pretreatment of the corpus training theme LDA model, and by calculating the degree of confusion, determine the optimal number of topics K each topic below contains the theme of the weight of keywords and keywords, given the need to extract the keywords for the corresponding term resource areas, field terms of nouns, so only keep nouns under each topic, Select the top 20 nouns of each topic weight as the keywords of that topic. After LDA topic model processing, because each resource has a certain topic and keyword and keyword weight, the range transformation method is used to standardize the weight of 20 keywords, where each resource can be expressed as $S = T_k\{t_1, t_2, \ldots, t_{20}, w_1, w_2, \ldots, w_{20}\}$ and $\sum_{i=1}^{20} w_i = 1$ (Fig. 2).

(2) Semantic graph construction
After keyword extraction, all the keywords of all resources in the corpus can be obtained. The semantic graph is an undirected weighted graph. Node is the keyword, edge is the relation between keywords, node weight is the weight of keywords, edge weight is the semantic similarity of two keywords. Semantic similarity is calculated based on semantic distance of domain ontology, and only keywords in the same semantic ontology have correlation. There are edges in the semantic graph

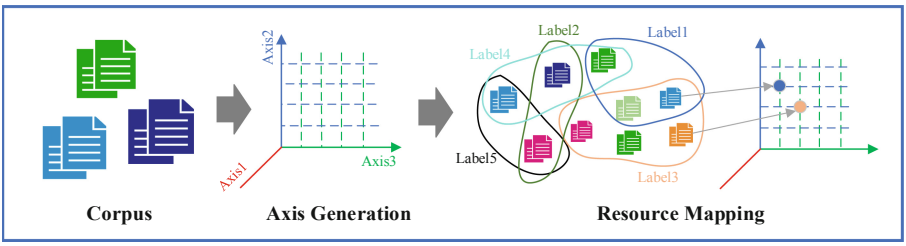

Fig. 1. The framework for automatic classification of the R&D resource space

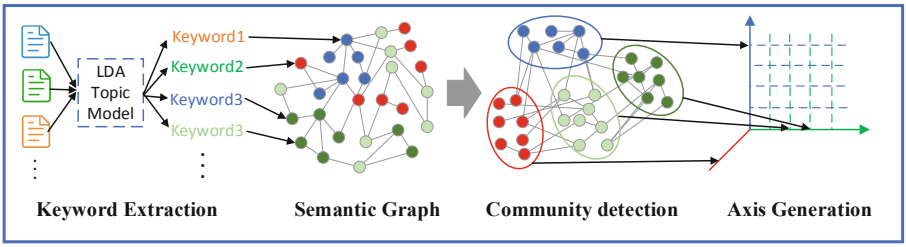

Fig. 2. R&D resource space generation process

to connect, and the calculation method of semantic similarity between two keywords is as follows:

$$relatedness(t_1, t_2) = -\log\left(\frac{ShortestPath(t_1,t_2)}{2D}\right) \quad (1)$$

$ShortestPath(t_1, t_2)$ represents the minimum semantic distance between keywords t_1 and t_2 in domain ontology. D is the depth of ontology in the domain where keywords t_1 and t_2 belong.

(3) Community detection and axial coordinate mapping

After building the semantic graph, it is necessary to find closely related keyword communities in the graph to further aggregate resources. In order to detect the closely related keyword groups in the semantic graph, it is necessary to apply appropriate graph clustering algorithm in the semantic graph. Hierarchical clustering algorithm focuses on discovering the hierarchy of communities in the network when the number of community groups is not clear. GN (Girvan-Newman) algorithm [9] is a split hierarchical clustering algorithm. It uses the ratio of edge interspaces and edge weights in the graph (edge weight ratio) as a measure of similarity. Every time the edge with high edge weight ratio is selected to delete, a hierarchical split tree is finally formed. Each level of the split tree is the divided keyword community.

Based on the hierarchical relationships in the resource model of multidimensional classification as a benchmark. The GN algorithm to the resources of the community at all levels keywords multidimensional classification model mapping, mapping a large community dimension for the resource classification, further divided the concrete category of the small community mapping for each dimension, Realize the construction of each axis of the resource space and the coordinates on the axis.

(4) R&D resource space completeness test and resource mapping

In an automated build resource space after the completion of the need for resources spatial completeness inspection, check whether there is not space are mapped to resources classification categories, if there is not mapped to resource category of multidimensional classification model of the space, also need to be tested to designed.The GN community and axis coordinate mapping, to select keywords community mapping to the multidimensional classification model, Until the constructed resource space satisfies the completeness. The resource space is constructed based on BERT-LEAM (Bidirectional encoder Representational from Transformers——label attentive) Model maps R&D resources to resource space.

3.3 Research and Development Resource Mapping Method Based on BERT-LEAM Model

Since resource space is a multi-dimensional and multi-coordinate spatial structure, each dimension corresponds to a classification standard, and each coordinate in the dimension corresponds to a specific category, mapping R&D resources to resource space is essentially a multi-label text classification task.

The multi-label text classification method based on BERT-LEAM model maps and R&D resources to multiple dimensions and coordinates of resource space.

As shown in Fig. 3, the BERT-LEAM model is composed of the BERT text feature representation layer and LEAM label semantic embedding network layer. The combination of label set and text is converted into a form of single label plus the text. After the BERT layer is converted into vector text, the LEAM network model is used to calculate the final label probability distribution of text.

The BERT model uses a bidirectional transformer structure to capture the context relationship in the statement. Transformer is a network based on self-attention mechanism, which is a kind of encoder. Since the BERT model uses token level for input, that is, input based on character level, text can be input into BERT model without word segmentation operation, and input labels and R&D resource text to be classified into

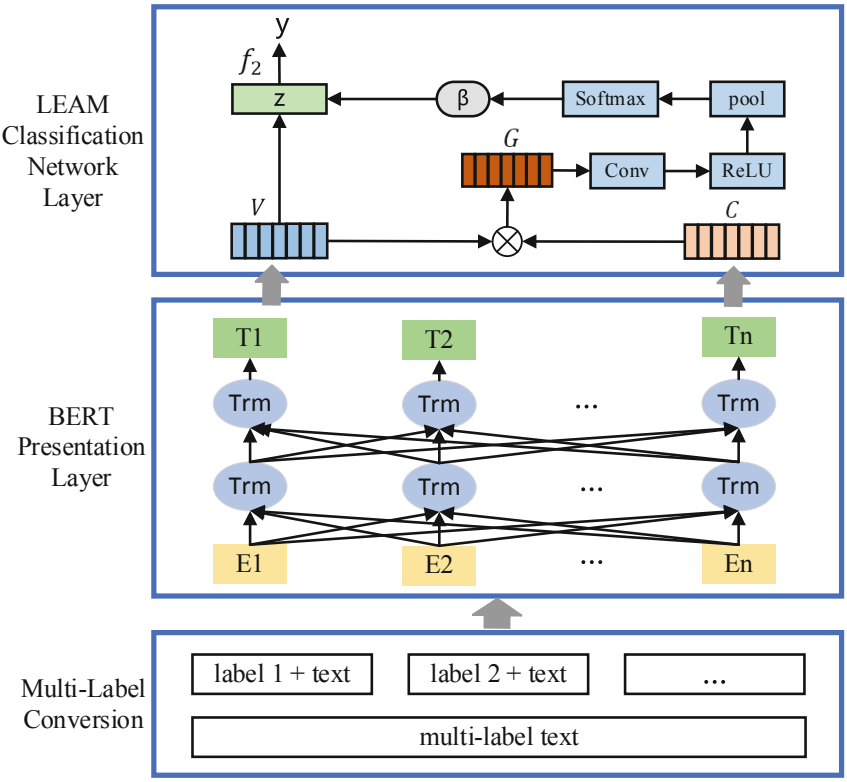

Fig. 3. Structure of BERT-LEAM model

BERT model, respectively. The label-embedding matrix C and Word-embedding matrix V after the encoding of Label and resource text are obtained, and the two matrices are input into the LEAM classification network layer for further classification operations.

LEAM network model is different from the traditional classification model. It learns the influence of label-embedding on word-embedding and uses the correlation between Label and Word to perform the aggregation calculation of label-embedding and word-embedding. The calculation method is as follows:

$$G = \left(C^T V\right) \emptyset \hat{G} \tag{2}$$

G indicates the aggregation result, and $\emptyset$ is calculated as the product of paraplegic elements. $\hat{G}$ assumes the $K \times L$ normalized matrix. K indicates the number of categories of *Label* and L indicates the number of *Word*.

Assume that each element in $\hat{G}$ is the product of l_2 regularization of the c label embedding vector and the l word embedding vector, which can be calculated as follows:

$$\hat{g}_{kl} = ||c_k|| \times ||v_l|| \tag{3}$$

$\hat{g}_{kl}$ is the element in $\hat{G}$, c_k is the c label embedding vector, and v_l is the l word embedding vector.

In order to better obtain the correlation between the local semantic information of words and tags, convolution and ReLU activation functions are introduced to further process the aggregation vector of tags and words. $G_{l-r,l+r}$ is used to measure the correlation between the phrase and label whose center is around the l word extending the range of r. The similarity vector of the l phrase and the k label is calculated as follows:

$$u_l = \mathrm{ReLU}\left(G_{l-r,l+r} W_1 + b_1\right) W_1 \in R^{2r+1}, b_1 \in R^k \tag{4}$$

u_l is the similarity vector obtained by calculation. R is the shared vector space of tag and word. $G_{l-r,l+r}$ is the fragment centered on the l word. W_1 and b_1 are parameters obtained by learning.

Then the maximum correlation coefficient is obtained by max pooling, that is, $m_l = \max\_\mathrm{pooling}(u_l)$. After maximum pooling, the vector $m = (m_1, m_2, m_3, \ldots, m_L)$, this vector represents the maximum correlation between words containing local semantics and tags.

Finally, the attention weight $\beta = (\beta_1, \beta_2, \beta_3, \ldots, \beta_L)$, Word-embedding matrix V gets the final representation Z under the weight of attention mechanism. The calculation method is as follows:

$$z = \sum_{l} \beta_l v_l \tag{5}$$

For the multi-label classification problem, LEAM model dissolves it into M single-label problems, and the training objective formula is as follows:

$$\min_{f \in F} \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} CE(y_{nk}, f_2(z_{nk})) \tag{6}$$

N is the number of categories. K is the number of labels. y_{nk} is the probability representation of the category label. z_{nk} is the text feature representation of the category tag, $f_2(z_{nk}) = \frac{1}{1+\exp(z'_{nk})}$. CE is the cross-entropy operation.

In order to increase the weight of category judgment, the distance between the text representation of the same category is smaller than the distance between the text representation of different categories. This paper introduces a label regularization term for the punishment of training targets, and the formula is as follows:

$$\min_{f \in F} \frac{1}{K} \sum_{k=1}^{K} CE(y_k, f_2(z_k)) \tag{7}$$

After the training is completed, the R&D resources that need to be mapped are input into the BERT-LEAM model in the form of single label plus text, and then all the categories corresponding to the R&D resources can be obtained, so as to map them to the corresponding positions in the resource space (Fig. 4).

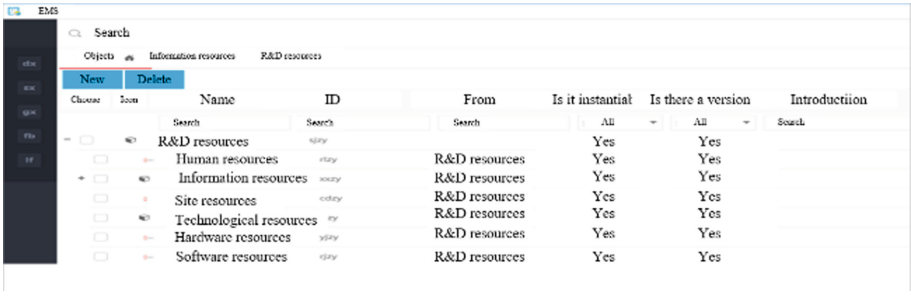

Fig. 4. Interface of R&D resources automatic classification (Figure is original)

Table 1. Operation and Function of R&D resources automatic classification

Operation	Function
Name, identifier	Adding category properties (B1-02-01)
Functional description	Add new properties to existing categories in the system.
Input	Category ID and attribute information
Operation sequence	1. Locate the classification according to the classification ID. 2. Determine whether the system has existing attributes (existing attributes can be added). 3. Create new attributes for the classification.
Output	Add new properties

4 Instance of Automatic Classification of R&D Resource Space

Dynamic modeling is carried out through EMS modeling tools, mainly establishing model classification and attributes, and the inputs and outputs are as shown in Table 1.

5 Conclusions

(1) This paper uses machine learning method to extract and cluster the R&D resources of the group enterprise, and then maps the clustering results to the multi-dimensional classification model of the classification layer of the R&D resource model, and realizes the mapping of each R&D resource to the resource space with the help of BRET-LEAM methods. Finally, a semantic data model for standardizing, storing, managing and locating resources is formed, that is, resource space model. Finally, an instance of group enterprise R&D resources automatic classification is built.

(2) The R&D resource space model constructed in this paper is beneficial to realize the integrated management and effective utilization of R&D resources of the group, and provides a theoretical basis for solving the problem of distributed heterogeneity of R&D resources, effective utilization of resources and improvement of design process efficiency.

Acknowledgment. This work was supported in part by the National Key R&D Program Project "Research on the Construction and Integration and Sharing Mode of R&D and Design Resource Space of Group Enterprises" (2018YFB1701800).

References

1. Fan L (2020) Virtual visualization design and implementation of workshop resources in cloud manufacturing environment. University of Electronic Science and Technology of China
2. Gu F, Liu YSY, Gu X (2020) Knowledge management forum. 5(2):69–81

3. Huang Y, Huang L, Chen H et al (2021) Science and technology for development. 17(5):964–971
4. Li SJ, Zhong ES et al (2015) design and implementation of peer-to-peer navigation grid platform. Sci Surv Mapp 40(5):148–153
5. Liu PP, Wang JY (2012) Design and implementation of education resource retrieval system based on XML. Inner Mongolia University
6. Ljiljana S, Steffen S, Rudi S (2001) eLearning based on the Semantic Web. Http: WebNet2001-World Conference on the WWW and Internet
7. Qi ES, Tian Y, Liu L (2018) Visualization analysis of Internet business model based on knowledge graph. Sci Technol Manag Res 38(4):190–196
8. Qi F, Shi YB, Li Y (2018) Information Technology and Informatization, 41–42
9. Qian J, Wang CK, Guo GY (2018) An algorithm for updating the centrality of node mediations in dynamic networks based on community. J Softw 29(3):853–868
10. Qiu L, Zhang AS, Li SB, et al (2022) Summary of knowledge map construction of aeronautical manufacturing. Appl Res Comput 1–10
11. Wang C, Sun Q, Xu J et al (2019) Research on the value creation mechanism of corporate headquarters from the perspective of double embedding: based on the case study of Times Group. Manage Rev 31(3):279–294
12. Xia CM (2021) Research on automatic classification model of power grid construction resources based on Naive Bayes algorithm. Bonding 48(12):93–97
13. Zhan DC, Zhao XB, Wang SQ et al (2011) Comput Integr Manuf Syst 17(3):487–494
14. Zhuge H (2007) Autonomous semantic link networking model for the knowledge grid. Concurr Comput: Pract Exp 7(19):1065–1085