Enterprise development forecast analysis under environmental policy based on knowledge graph

Zeyu Zhang¹, Xiaoting Wang³, Lingzhong Meng² and Qian Wang³,⁴,⁵

¹College of Software Engineering, Sichuan University, Chengdu 610207, China;
²Institute of Software, Chinese Academy of Sciences, Beijing 100190, China;
³Chinese Academy of Environmental Planning, Beijing 100012, China;
⁴School of Environment, Tsinghua University, Beijing, 100084, China
⁵Email: wangqian@caep.org.cn

Abstract. In order to fully explore the relationship between environmental policy and enterprise development, and further improve the forecasting effect of environmental enterprise development, this paper proposes a method of forecasting enterprise development status under environmental policy based on knowledge graph. This method extracts environmental policies and corporate information from the public website of the environmental department to build an environmental corporate knowledge graph, combines cosine similarity and regression algorithms to establish a prediction model for environmental policies and corporate development, and uses the model to analyse environmental policies and predict corporate future development. Experiments show that this method makes full use of the advantages of knowledge graphs in relation analysis and graphical display, has good accuracy and recall, greatly reduces the time cost of enterprise forecasting, and has strong application value.

1. Introduction

The impact of environmental policies on development of enterprises has become more and more significant. Banks need to comprehensively evaluate environmental policies and enterprises to assess the future development of the enterprises, and make decisions on whether to lend funds to them. The complexity and uncertainty of the impact of environmental policies on enterprises have brought difficulties to artificial enterprise development prediction. Building environmental policies and enterprise databases and prediction models can effectively improve the accuracy and efficiency of corporate forecasts.

Traditional enterprise forecasts mostly use description language matching methods to analyze the enterprise development. In the literature [1], the use of logistic regression to establish a prediction model for corporate financial proportions and characteristics is proposed. In the literature [2], support vector machine is used to predict corporate carbon emission risks. In the literature [3], a method of prediction model based on cognitive mapping and classification assessment is proposed. Although these studies have improved the traditional forecasting algorithms, most of them are considered from the development factors of the use cases themselves. They have not fully considered the relationship between the enterprises and the relationship between the enterprises and the environmental policies, thus reducing the specificity. In addition, traditional enterprise predictive analysis databases mostly use relational databases as the underlying data storage. When the number of companies being analyzed...
is increasing rapidly, the traditional database design architecture has gradually been unable to clearly describe the knowledge associations between different companies and pollution.

Knowledge Graph [4] has been widely used in various professional fields [5], using knowledge graphs to construct interactive knowledge retrieval and question answering systems [6], using knowledge graphs and semantic similarity to construct personalized education recommendation algorithm [7], using knowledge graphs and deep learning to predict stock price trends. These methods using knowledge graphs show better results than traditional methods.

In order to make up for the shortcomings of traditional enterprise forecasting analysis methods, this paper introduces knowledge graph in the prediction of environmental policies and enterprise development. Based on the existing website data, a knowledge graph of environmental policies and enterprises is constructed, and regression algorithm is used to carry out the prediction analysis of enterprises based on the knowledge map. This method gives full play to the advantages of the graph database in relational retrieval performance, and integrates the knowledge of environmental policy, pollutants, enterprise information and the relationship between them in visual graph form, which provides a new solution for the further study of the forecast analysis of development of enterprises.

2. Environmental enterprise prediction knowledge graph

2.1. Overview of environmental enterprise forecasting knowledge graph
The environmental policy and corporate knowledge graph constructed in this paper belongs to the Domain-specific Knowledge Graph, which is structured using the attribute graph model to predict the development of enterprises. The project construction process is showing in Figure 1.

![Figure 1. Construction process of enterprise prediction based on knowledge graph.](image)

The knowledge map divides companies into sub-graphs according to their industry, and adds attributes for each enterprise and each environmental policy. The graph associates each enterprise with its upstream and downstream enterprises, links enterprises with environmental pollutants in terms of pollutant emissions, and associates pollutants with environmental policies in a related relationship.

The environmental enterprise prediction knowledge graph re-stores environment and enterprises knowledge in the government environmental website database in the form of a complex semantic network diagram through the graph database, which is convenient for environmental analysis and enterprise forecasters to access Environmental business forecast knowledge from the perspective of “relationship” visualization.

2.2. Framework of environmental enterprise forecast knowledge graph
The basic technical framework for constructing the environmental enterprise predictive knowledge graph mainly includes a five-layer structure: raw data, knowledge extraction, knowledge fusion, knowledge processing and knowledge application, as shown in Figure 2.
The environmental enterprise knowledge graph uses the public environmental policy and corporate data stored in the national environment department website databases as the main source of the original data of the knowledge graph. Various environmental and corporate entities, attributes and relationship data are extracted by the knowledge extraction module, and then submitted to the knowledge fusion layer to use the semantic similarity calculation model for entity disambiguation processing, which is used to establish the semantic network of the environmental enterprise prediction knowledge graph. The knowledge processing module dynamically updates the content of the environmental enterprise knowledge map, continuously evaluates and improves the quality and application effects of existing knowledge. The knowledge application layer uses graphical data display to provide environmental enterprise information retrieval and data statistical analysis application implementation.

2.3. Construction of environmental enterprise forecast knowledge graph
This paper chooses Neo4j graph database [8] to store environment enterprise prediction knowledge graph, and access data through Cypher data query language. Using natural language processing algorithms such as sentence segmentation, word segmentation, and keyword extraction [9], the collected environmental policies, pollutants, and corporate data are processed for raw data. The knowledge units of entities, attributes of entities and relationships between entities are processed and extracted. Through processing and extraction, more than 20,000 knowledge entities covering the environment and the enterprises are finally obtained.

Before new knowledge is added to the environmental policy knowledge graph, entity matching of the knowledge graph is required. This paper selects the fast similarity calculation and rule-based method in the knowledge graph matching method to quickly match the environment and enterprise entities.

The knowledge processing part mainly completes the addition of new environmental policies and enterprises knowledge and the update of original knowledge. The main content includes the use of relational recommendation technology based on knowledge expression to complete the discovery and reasoning of historical incomplete knowledge, the modification of low-quality entity feature attributes, and the correction of the wrong knowledge entity similarity relationship matching in the original knowledge graph. The constructed environmental policy enterprise knowledge map effectively integrates environmental and enterprise-related information, provides an interpretable basis for environmental enterprise predictive analysis, and provides important basic data support for further decision-making and predictive analysis. The graph display effect is shown in Figure 3. The dark pink node represents the enterprise, the yellow node represents the category of environmental pollutants, the blue node represents the category of the company, the light pink node represents the pollutant...
substance, the green node represents the environmental policy category, and the gray node represents the specific environmental policy.

Figure 3. The knowledge graph of environmental policy and enterprises.

3. Enterprise prediction under environmental policy based on knowledge graph

3.1. Environmental policy enterprise forecast design based on policy knowledge matching

When there is a large amount of environmental policies and related corporate information in the knowledge base, a large amount of information matching is required. This article gives full play to the advantages of knowledge graphs in relational network analysis. In the prediction, the information of the company to be predicted is used as the input for knowledge matching, searching and matching the sub-graph of the companies in the same industry and the environmental policy associated with the companies in that industry, and combining the company to be predicted and the relevant environmental policies to carry out enterprise development forecast analysis.

Through the analysis of the knowledge graph, it can be found that, on the one hand, environmental policies, environmental pollutants, and the enterprise’s attributes affect the development of the enterprise. On the other hand, companies with a high degree of similarity often show similar changes in corporate development when they are simultaneously affected by external events. Therefore, the cosine similarity analysis and regression model analysis between enterprises are carried out in sequence to jointly construct a risk prediction model describing the development of enterprises.

Use the cosine similarity algorithm [10] to describe the similarity between enterprises, namely \( F_{\text{similarity}} \). Then, determine the factors that affect the development of the enterprise, that is, the parameter set in the regression model, and two regression methods, gradient descent [11] and formal equation [12], are selected to construct the “environmental policy-pollution Material-Enterprise Development Regression Forecast Model”. The formula is as follows.

\[
F_{\text{regression}} = X \cdot \theta
\]

\[
X = [x_1, x_2, x_3, \ldots, x_n], \theta = [a_1, a_2, a_3, \ldots, a_n, r]
\]

\( X \) represents the description matrix of each enterprise, \( \theta \) represents the parameter matrix, including the parameter \( a_n \) corresponding to attributes of each enterprise and the intercept \( r \).
The two analysis methods are combined to avoid the contingency and uncertainty of a single method of forecasting. Through actual tests, the weights of the two methods in the model are set to 0.2 and 0.8, respectively, to obtain the final enterprise development risk prediction model. The formula is as follows.

\[ F_{prediction} = \lambda_1 F_{similarity} + \lambda_2 F_{regression} \]  
\[ \lambda_1 = 0.2, \lambda_2 = 0.8 \]  

(2)

3.2. Environmental policy enterprise forecast experiment based on knowledge graph

3.2.1. Data processing and data set construction. This paper collects public data from the national government website and environmental resource websites of various provinces, including the ecological environment data of the Ministry of Ecology and Environment, the ecological environment data of the Shanxi Provincial Department of Ecology and Environment, the key pollutant discharge enterprise data of the Shanxi Provincial Department of Ecology and Environmental policies from websites. Then, the data is sorted and classified into three categories: environmental policies, environmental pollutants, and enterprises.

The company type data is segmented, and the company’s attributes include: company name, company size, company category, company location, company emission pollutants, company operation status, and whether the company is closed or not, etc. The enterprise category is divided into eighteen industry categories: chemical raw materials and chemical products manufacturing, non-metallic mineral products Industry, food manufacturing industry, paper and paper products industry, beverage manufacturing industry, etc. The enterprise scale is divided into seven types: super large, large first grade, large second grade, medium one, medium two, small and others.

Pollutants are divided into seven categories: air pollution, water pollution, soil pollution, ecological environmental pollution, and solid waste pollution, acoustic pollution and radiation pollution. The attributes of environmental policies include: policy issuance time, policy implementation time, policy implementation area, policy content and environmental indicators.

1864 enterprises in Shanxi were selected in the knowledge graph, including 1844 normal operating enterprises and 20 closed enterprises. According to the preprocessing results of the enterprise and environmental data, the data attribute description of the enterprise is expressed in the form of a matrix.

3.2.2. Environmental policy enterprise forecast model training. In the process of model training, 95% of the data set, that is, 1769 companies, are randomly selected as the training set, 5% of the data set, that is, 95 companies are used as the validation set. The model parameter matrix is initialized to 0.

Set the Loss Function [13] as the input data matrix \( X \) point multiplied by the model parameter matrix \( \theta \) subtract the standard output matrix \( Y \). The loss function formula is as follows.

\[ Loss = X \cdot \text{dot}(\theta) - Y \]

\[ \text{Total Loss} = \text{Loss}.\text{transpose()} \cdot \text{dot}(\text{Loss})/(2 \cdot m) \]  \( m \) means Number of samples \( (3) \)

The strategy of “reduce learning rate on plateau” is adopted, and the learning rates of training periods 0, 50, and 100 are set to 0.1, 0.01, and 0.001 respectively, and the total training period is set 400.

3.2.3. Environmental policy enterprise model evaluation index. In the enterprise risk prediction, the enterprise risk prediction results are divided into true positive, false positive, false negative and true negative. In addition, Precision, Recall [14] and F1-Measure [15] are selected to evaluate the performance of model prediction.

4. Prediction result and case analysis

4.1. Case analysis

The experiment selected data from 8 test companies as input for experimental analysis to evaluate the application performance of the model. Among them, types of enterprises, types of enterprises, types of
pollutants discharged and cities and districts that appears in the location of enterprises are all represented by numbers. The operating status of each enterprise is represented by 1 and -1, where 1 means the enterprise is normally developed, and -1 means the enterprise is closed.

The company attributes, company industries, and model output results of these companies are shown in the following Table 1.

| Enterprise | Scale | Location | Industry | Pollutants | Development status | Output |
|------------|-------|----------|----------|------------|--------------------|--------|
| P1         | 5     | 15       | 10       | 1          | 1                  | 0.872  |
| P2         | 5     | 62       | 8        | 2          | 1                  | 0.786  |
| P3         | 3     | 54       | 11       | 1          | 1                  | 0.706  |
| P4         | 2     | 6        | 12       | 2          | 1                  | 0.914  |
| P5         | 1     | 27       | 3        | 2          | 1                  | 0.527  |
| P6         | 1     | 35       | 7        | 1          | -1                 | -0.626 |
| P7         | 5     | 39       | 8        | 2          | -1                 | -0.504 |
| P8         | 5     | 17       | 1        | 2          | -1                 | -0.996 |

4.2. Environmental policy enterprise forecast model evaluation

For the output result of the enterprise risk prediction model, set the output result between 0.6 and 1 as the model predicts that the enterprise is developing normally, and the output result between -1 and -0.6 is set as the model prediction Enterprises will face the risk of shutting down.

In the process of verifying the model, 95 companies in the verification group are verified, of which 88 groups are normal operating companies and 7 groups are shut down companies. The model successfully predicts 88 companies among them, including 82 normal operating companies and 6 closed companies, with an accuracy rate of 0.929 and a recall rate of 0.932.

We test the traditional similarity algorithm and regression algorithm to predict the development of the enterprise, and compare it with the enterprise prediction method based on the knowledge graph (EPKM) proposed in this paper. The results are showing in Table 2.

| Precision | Recall | F1-Measure | Time |
|-----------|--------|------------|------|
| Linear Regression | 0.853 | 0.875 | 0.863 | 0.026 |
| EPKM | 0.926 | 0.932 | 0.929 | 0.016 |

The results show that the enterprise prediction model based on the knowledge graph not only improves the accuracy of prediction, but also obtains a better recall rate compared with the traditional linear regression algorithm. The precision rate, recall rate and F1-Measure index are all better than traditional algorithms and the precision of this model is 92%. Moreover, in terms of prediction time, the prediction algorithm based on the knowledge graph saves nearly 61% of the time compared with the traditional method and obtains better time efficiency.

5. Conclusions

This paper designs a method for predicting the development of enterprises under environmental policies based on knowledge graph, and establishes a domain knowledge graph describing environmental policies, environmental pollutants and enterprise entities, using the knowledge graph semantic network and the “relationship” attribute of graph database, and constructs an environmental policy enterprise forecasting model based on knowledge graph. While improving the accuracy of enterprise forecasting, it greatly reduces the time spent on forecasting and improves the efficiency of enterprise forecasting.
Because the large amount of enterprise data used to construct the knowledge graph is stored in structured and semi-structured on the government environmental website, knowledge extraction is relatively simplified. In fact, there is a large amount of non-structure data derived from enterprise quarterly data reports and business analysis, and the knowledge extraction and processing of these data need further research. In addition, how to use environmental policies and enterprises knowledge graph to develop more applications, as well as to optimize and improve predictive regression algorithms, still need continue research.

References
[1] Ciampi, Francesco 2015 Corporate governance characteristics and default prediction modeling for small enterprises. An empirical analysis of Italian firms JOURNAL OF BUSINESS RESEARCH 68(5) 1012-1025
[2] Zhifang Zhou, Tian Xiao, Xiaohong Chen, Chang Wang 2015A carbon risk prediction model for Chinese heavy-polluting industrial enterprises based on support vector machine International Conference on Nonlinear Dynamics and Complexity 2015
[3] Oliveira Manuel D N T, Ferreira Fernando A F, Perez-Bustamante Ilander Guillermo O, Jalali Marjan S 2017 Integrating cognitive mapping and MCDA for bankruptcy prediction in small- and medium-sized enterprises JOURNAL OF THE OPERATIONAL RESEARCH SOCIETY 68(9) 985-997
[4] Nickel Maximilian, Murphy Kevin, Tresp Volker, Gabrilovich Evgeniy 2016 A Review of Relational Machine Learning for Knowledge Graphs PROCEEDINGS OF THE IEEE 104(1) 11-13
[5] Weiguo Zheng, Hong Cheng, Yu, Jeffrey Xu, Lei Zou, Kangfei Zhao 2019 Interactive natural language question answering over knowledge graphs INFORMATION SCIENCES 481 141-159
[6] Shubing Shan, Buyang Cao 2017 Follow a guide to solve urban problems: the creation and application of urban knowledge graph IET SOFTWARE 11(3) 126-134
[7] Yang Liu, Qingguo Zeng, Huanrui Yang, Adrian Carrio 2018 Stock Price Movement Prediction from Financial News with Deep Learning and Knowledge GraphEmbedding KNOWLEDGE MANAGEMENT AND ACQUISITION FOR INTELLIGENT SYSTEMS 11016 102-113
[8] Hongcheng Huang, Ziyu Dong 2013 Research on architecture and query performance based on distributed graph database Neo4j INTERNATIONAL CONFERENCE ON CONSUMER ELECTRONICS, COMMUNICATIONS AND NETWORKS 533-536
[9] Aytuğ Onan, Serdar Korukoğlu, Hasan Bulut 2016 Ensemble of keyword extraction methods and classifiers in text classification EXPERT SYSTEMS WITH APPLICATIONS 57 232-247
[10] Guiwu Wei, Yu Wei 2018 Similarity measures of Pythagorean fuzzy sets based on the cosine function and their applications INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS 33(3) 634-652
[11] Bottou Leon 2010 Large-Scale Machine Learning with Stochastic Gradient Descent 19th International Conference on Computational Statistics 2010
[12] Arioli M., Gratton S 2012 Linear regression models, least-squares problems, normal equations, and stopping criteria for the conjugate gradient method COMPUTER PHYSICS COMMUNICATIONS 183(11) 2322-2336
[13] Hurlin Christophe, Leymarie Jeremy, Patin Antoine 2018 Loss functions for Loss Given Default model comparison EUROPEAN JOURNAL OF OPERATIONAL RESEARCH 268(1) 348-360
[14] Walters William H 2011 Comparative Recall and Precision of Simple and Expert Searches in Google Scholar and Eight Other Databases PORTAL-LIBRARIES AND THE ACADEMY 11(4) 971-1006
[15] Zachary C. Lipton, Charles Elkan, Balakrishnan Naryanaswamy 2014 Machine learning and knowledge discovery in databases ECML PKDD 2014