Feature extraction modelling of enterprise innovation behaviour data based on morphological gradient

Shibiao Mu

School of Electro-mechanical and Information Technology,
Yiwu Industrial and Commercial College,
Yi'wu 322000, China
Email: Shibiaom@outlook.com

Abstract: Aiming at the problem of slow speed and low accuracy of traditional feature extraction model for enterprise behaviour data, a feature extraction model for enterprise innovation behaviour data based on morphological gradient is constructed. The model is divided into two parts: the virtual method is used to integrate the data of enterprise innovation behaviour, and the data are synthesised and filtered; the morphological gradient operator is used to extract the features of the integrated data of enterprise innovation behaviour. The simulation results show that using the proposed model to extract the characteristics of enterprise innovation behaviour data, the extraction process only takes 15.68 min, and the average extraction accuracy can reach 96.68%. This result is much better than the three traditional models and achieves the expected goal.

Keywords: morphological gradient; enterprise innovation behaviour; data characteristics; feature extraction model.

Reference to this paper should be made as follows: Mu, S. (2022) 'Feature extraction modelling of enterprise innovation behaviour data based on morphological gradient', *Int. J. Information Technology and Management*, Vol. 21, Nos. 2/3, pp.294–310.

Biographical notes: Shibiao Mu received his BS degree in Computer Network from Beijing University of Post Telecommunication, China, in 2006, and his ME degree in Information and Computing Science from China West Normal University, Nanchong, China, in 2010. Since 2010, he has been a teacher in Yiwu Industrial and Commercial College. His current research interests include cloud computing, and data mining.

1 Introduction

With the innovation and progress of science and technology, all walks of life have entered a period of rapid development. Large enterprises and companies have been established. During the decades of reform and opening-up and the implementation of foreign trade policies, major enterprises have developed vigorously, while the competition among them has become increasingly fierce. In this context, in order to remain invincible, major enterprises have carried out rectification and transformation of their own development model. In the process of continuous reform and development, leaders of major enterprises gradually realise that real-time innovative behaviour is the

only way out. Enterprise innovation behaviour refers to the behaviour that enterprises gain excess profits by attracting new products, adopting new production methods, opening up new markets, obtaining new sources of raw materials and implementing new forms of enterprise organisation in order to occupy a dominant position in competition. However, the enterprise's behaviour innovation is not accomplished overnight. It needs to collect and analyse a large amount of data to explore the risks and benefits of a behaviour decision. It can only be carried out after careful consideration. Otherwise, if the behaviour decision is wrong, it will not only bring huge economic losses to the enterprise, but also affect the long-term development of the enterprise. It even causes the bankruptcy of enterprises. Therefore, how to extract the characteristics of risk and benefit from limited data, so as to provide decision support for enterprise innovation behaviour, has become one of the important research topics of sustainable development of contemporary enterprises.

Under the above background, domestic and foreign experts have conducted in-depth research and put forward many solutions. Xiao and Yue (2017) based on the user data of an internet financial company, including basic information, transaction information and log behaviour information, we use recency frequency monthly (RFM) and time frequency plane domain (tfpd) to extract user behaviour characteristics and transaction characteristics, and use logic regression, random forest and support vector machine to establish an enterprise behaviour extraction model. The extraction accuracy of the model is ideal, but it takes a long time. Tung and Jordann (2017) designed an innovation behaviour extraction method based on the evolution of the pattern trend of employees' interest characteristics, improved the traditional innovation behaviour extraction method, and then improved the accuracy and diversity of the enterprise innovation behaviour extraction results. Based on the previous research, this paper makes innovation, and finds the evolution trend of enterprise innovation behaviour mode through the time series statistical prediction of tag time series and the time sequence path method of weighted keyword co-occurrence time element splicing. In practical application, it is found that the extraction accuracy of this method can not meet the current requirements in this field. Xue and Xu (2017) proposes an automatic identification method of team innovation information relationship, which can be used in information organisation and in the actual process of team innovation. In this model, we improve the word frequency feature extraction method in three aspects: 'feature word classification processing', 'small sentence number' feature and 'last comment information attitude' feature to improve the recognition accuracy. However, the steps of this method are complex and can not be widely used.

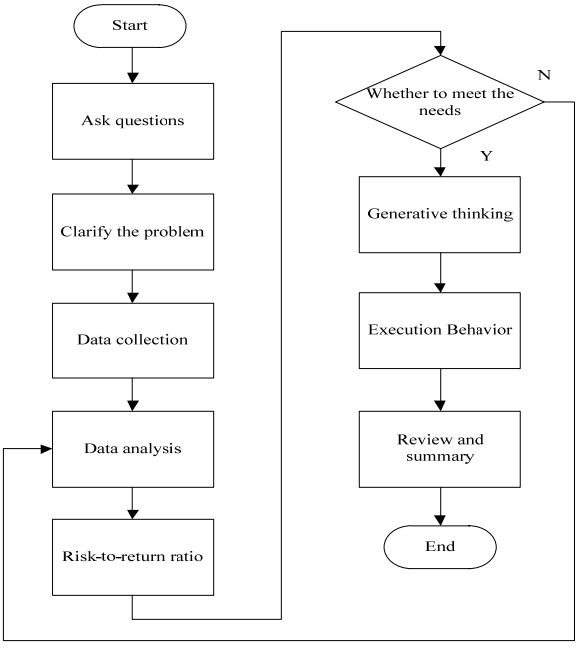
To solve these problems, a new feature extraction model for enterprise innovation behaviour data based on morphological gradient is studied. Morphological gradient is an algorithm originally used in image processing and analysis. By making a difference between the dilated original image and the corroded original image, the morphological gradient can be obtained to highlight the periphery of the highlighted area. Under the guidance of this principle, the model starts from the opposite point of view, that is, eliminating irrelevant or poorly correlated data in the data, and the remaining data is the data feature to be found. The specific process is to obtain the data gradient through morphological gradient operator, and to segment the data into feature part and flat part by taking the local neighbourhoods mean of gradient as the local adaptive threshold (Xing and Jiang, 2016). Finally, the simulation experiment results show that compared with the three traditional data feature extraction models, this model has faster extraction speed and

lower error rate, which shows that the design of the proposed model achieves the expected goal, can provide more accurate data support for the decision-making and implementation of enterprise innovation behaviour, which ensures the sustainable development of enterprises to a certain extent.

2 Feature extraction model for enterprise innovation behaviour data based on morphological gradient

Competition is the internal driving force to promote the development of enterprises. Especially in the current resource-constrained environment, in order to seize market share, enterprises have to innovate their behaviour in order to reduce their own costs and expand profit margins. Enterprise innovation behaviour refers to the behaviour that enterprises obtain excess profits by adopting different production methods in order to occupy a dominant position in the competition (Feng et al., 2017). Building the data feature extraction model of enterprise innovation behaviour can further obtain the development status of enterprise innovation and enhance the competitive advantage of enterprise. However, due to the uncertainty of the market, the decision-making of enterprise innovation behaviour is not a one-step process, but a complex process. It needs to go through several important aspects, such as raising questions, clarifying problems, collecting data, data analysis, risk-to-benefit ratio, generating ideas, execution behaviour, review and summary, as shown in Figure 1 (Zollo et al., 2016).

Figure 1 Basic process of enterprise innovation behaviour

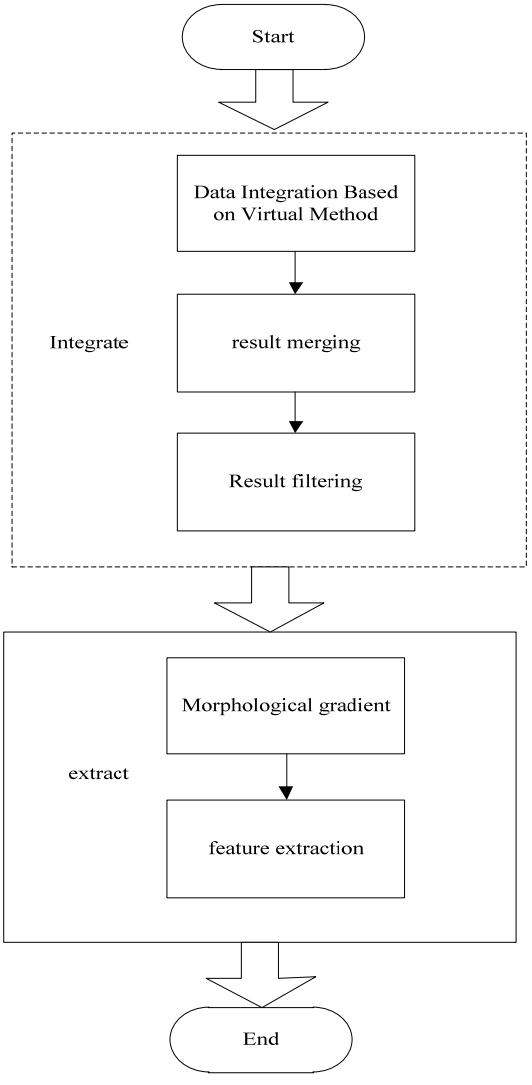


In the process of Figure 1, data analysis is the key step. In this step, the main work of the product is to extract the hidden noise features from the complex big data. In short, it is to

summarise the essential features from the big data, simplify the process of data analysis, so as to make more accurate behaviour decisions.

In the past, the feature extraction of enterprise innovation behaviour data was mainly from the perspective of positive thinking. The process was complex and the accuracy of extraction was low. For this reason, on the basis of referring to the traditional data feature extraction model, this paper studies a feature extraction model for enterprise innovation behaviour data based on morphological gradient from the perspective of reverse thinking (Crommelinck et al., 2016). Using morphological gradient theory, the model eliminates irrelevant redundant data from big data, leaving data that can reflect the essential characteristics of information. Figure 2 shows the basic process of model building.

Figure 2 Construction process of feature extraction model for enterprise innovation behaviour data based on morphological gradient



2.1 Data integration of enterprise innovation behaviour

The decision-making and implementation of enterprise innovation behaviour are based on data, so before data feature extraction, the integration of related data of the proposed problem is the first step. Data integration refers to the process of integrating distributed heterogeneous data from different sources, formats, characteristics and properties, but also interrelated (Hopf et al., 2016). There are five main problems in data integration, as shown in Table 1.

Table 1 Five problems in data integration

Problem	*Explain*
Isomerism	System heterogeneity: the differences among application systems, database management systems and even operating systems on which data depends.
	Schema heterogeneity: data storage mode is different. General storage mode includes relational mode, object mode, object relational mode and document nesting mode, among which relational mode is the mainstream storage mode.
Integrity	Data integrity refers to the correctness, consistency and consistency of data.
	Constraint integrity, which refers to the relationship between data and data, is the only feature that characterises the logic between data. Ensuring the integrity of constraints is the premise of good data publishing and data exchange, which can facilitate data processing and improve efficiency.
Integration content limitation	Data integration among multiple data sources is not to integrate all data, so how to define the scope and authority of integration constitutes the limitation of integration content.
Semantic conflict	There are semantic differences among information resources, which may lead to conflicts. Semantic conflicts will complicate data integration. So how to minimise the semantic conflict is also a hot and difficult research topic in data integration.
Conflict of authority	Since the data to be integrated may belong to different units or departments, how to ensure that the privileges of the original data are not infringed on the basis of accessing heterogeneous data has become a practical problem to be faced in integrating heterogeneous data.

At present, there are three main methods of data integration: federated database system, materialised method and virtual method.

1 Federal database system

Federal database system, referred to as FDBS, refers to the integration method of single metadata databases which are independent of each other but have certain connections with each other, as shown in Figure 3.

As a management centre, the federated database system can map different data sources to a unified database. In the process of mapping, it can realise schema transformation and solve the problems caused by heterogeneity (Susto et al., 2016). This integration method has some limitations, and its application scope is not very wide, and it needs a long time to establish the system.

2 Materialised method

Materialised method, also known as data warehouse method, is to use ETL tools to extract data from various data sources, and then clean, reduce and transform them. Finally, the processed data are loaded into the data warehouse uniformly, as shown in Figure 4.

This data integration method is simple to operate, but the implementation cycle is long, and there is a phenomenon of repeated storage, resulting in insufficient data space.

3 Virtual method

Virtual method, also known as middleware method, is currently the most commonly used method, but also the data integration method to be used in this study. It integrates heterogeneous source data through virtual views, as shown in Figure 5

Virtual method is totally different from the above two integration methods. It does not extract data from heterogeneous data sources and store them in a unified database, but still keeps the data in the original heterogeneous data sources. It only searches the required data by establishing a virtual integration view. The characteristic is that it will not produce a large number of interference data. The integrated data itself is a preliminary processed data (Liu et al., 2017a).

Figure 3 Federated database system

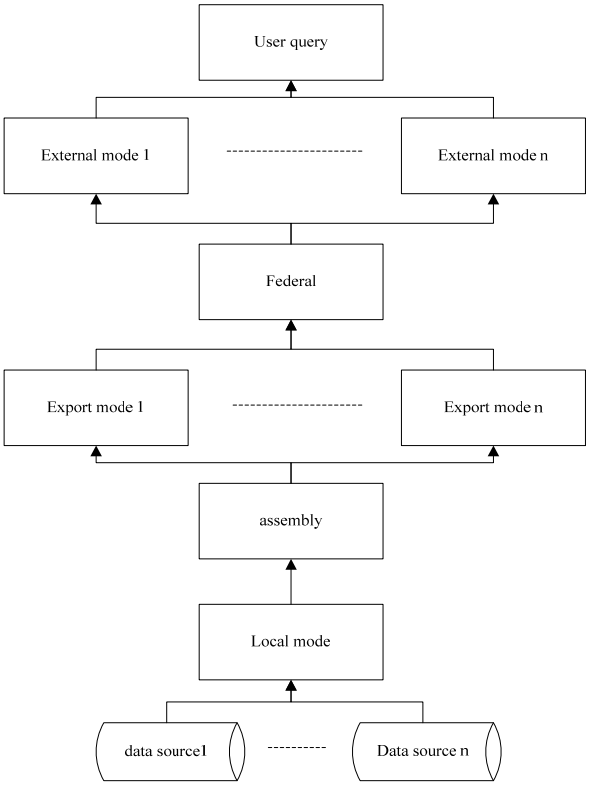


Figure 4 Materialised method

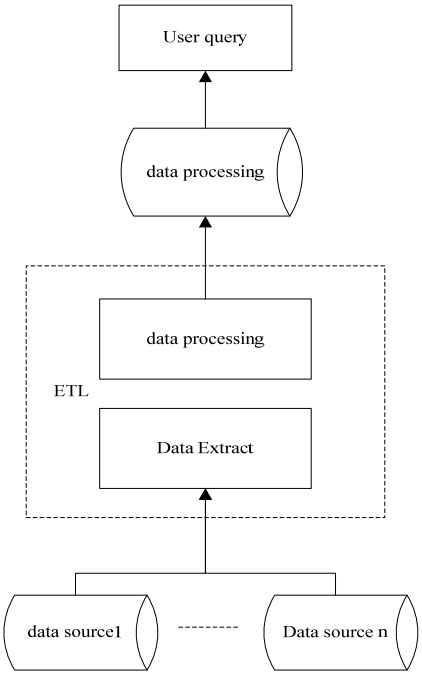
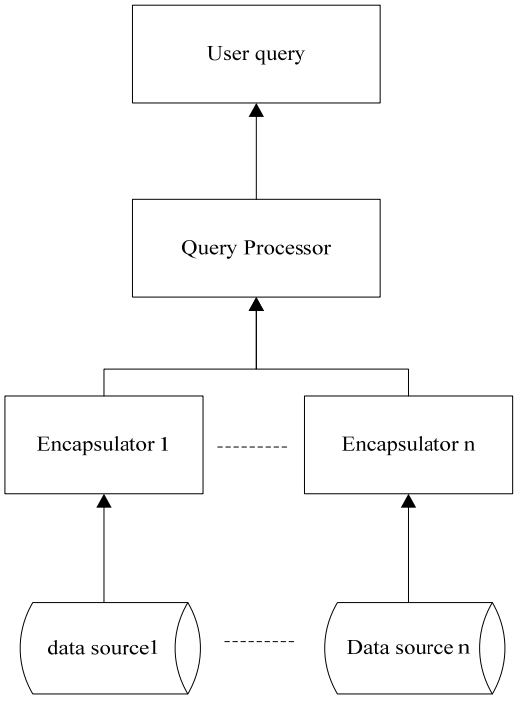
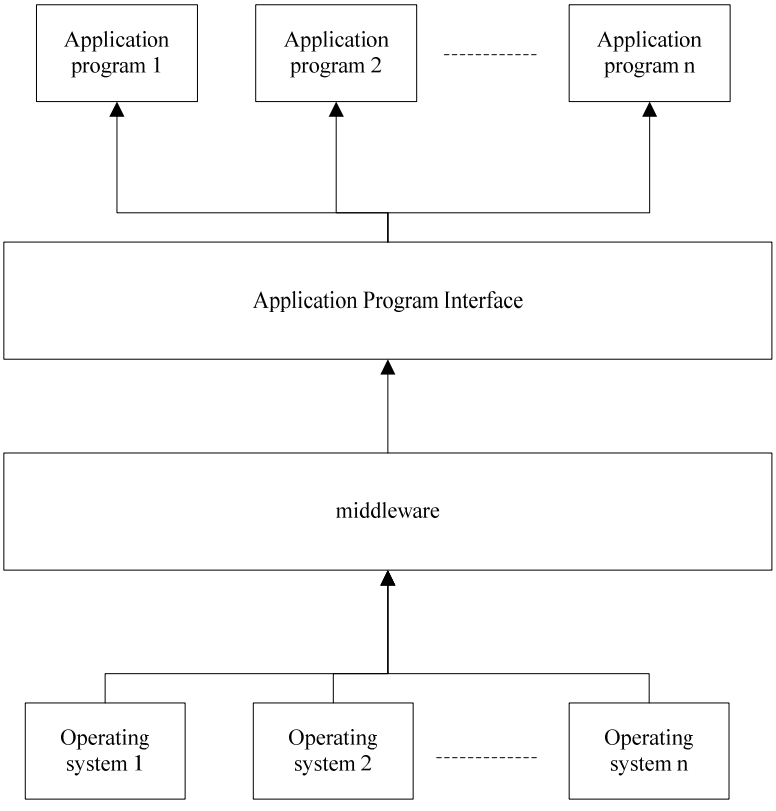


Figure 5 Virtual method



In this integration method, wrappers and middleware are the keys. Among them, the wrapper, as its name implies, plays the role of packaging data sources. When a user issues a query command, the wrapper converts it into a mode corresponding to a heterogeneous data source, and extracts the required data from it according to the decomposition command and stores it in a temporary data table (Liu et al., 2016). Middleware is a kind of software between operating system and application. Its main function is to integrate data stored in temporary data tables and solve conflicts between them, as shown in Figure 6.

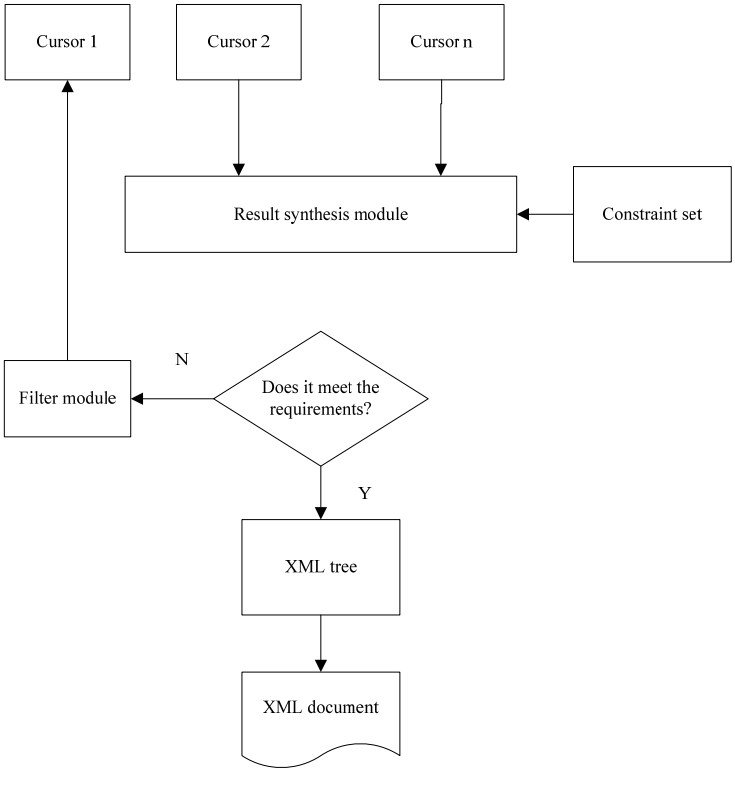
Figure 6 Middleware



After the middleware integrates the data stored in the temporary data table, each data has its own mode because it comes from different data sources. If they are integrated directly, the accuracy of feature extraction in the later period will be reduced. Based on this consideration, in the process of middleware integration, XML transformation, synthesis and filtering need to be carried out at the same time, as shown in Figure 7.

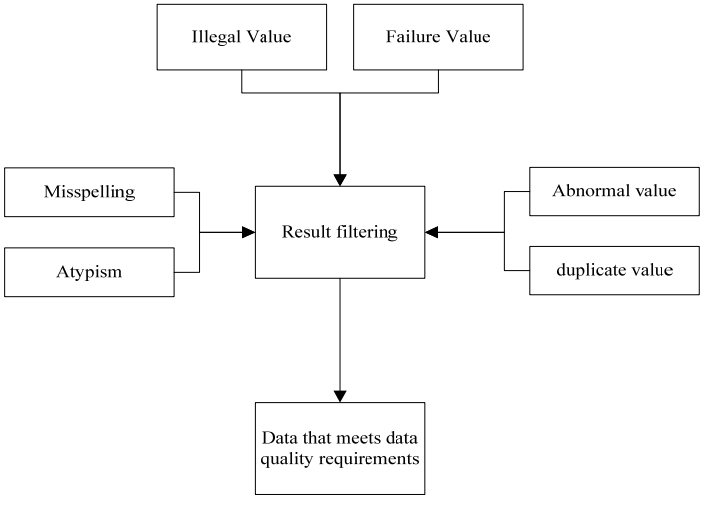
Results synthesis: firstly, according to the decomposed user query commands, the query tree and DTD are obtained. Under the guidance of appropriate rules, an XML tree is generated. Then, the generated XML tree is constrained and validated. Finally, the integrated data is transformed into a unified XML form by using the XML tree (Kavakiotis et al., 2016).

Figure 7 Data processing



Result filtering: in order to improve the quality of the integrated data, data filtering is needed, that is, to remove duplicate data, incomplete data, error data and so on. The specific process is shown in Figure 8.

Figure 8 Result filtering process



2.2 Feature extraction of enterprise innovation behaviour data

After data integration, the basic data of feature extraction has been obtained. Although this data has been processed, the cardinality is still very large, so it is still a complex process to obtain data features from it. Here, it is done by morphological gradient.

2.2.1 Principle of morphological gradient

2.2.1.1 Basic operations of morphology

Mathematical morphology is a mathematical method developed on the basis of set theory, integral set and topology, which is different from space-time domain analysis and frequency domain analysis (Liu et al., 2017b). Its basic contents include corrosion and expansion, open and closed operations, skeleton extraction, limit corrosion, hit-miss transformation, morphological gradient, Top-hat transformation, particle analysis, watershed transformation, etc.

In the above content, this research mainly uses two operators, corrosion and expansion, whose principle is to remove data smaller than structural elements in scale through corrosion and expansion transformation, while retaining the essential characteristics of data, which is the data features we want to extract. The following is a detailed description of its principle (Liu and Gillies, 2016).

If $S(n)$ is the data processed by integration, the domain is defined as $x = \{1, 2, \ldots, N\}$; $z(m)$ is the structural element, and the definition is predefined as $y = \{1, 2, \ldots, M\}$, and $M \leq N$. $S(n)$ defines corrosion and expansion of $z(m)$ as:

$$\begin{cases} (S-z)(n) = \min\left[S(n+m) - z(m)\right], m \in y \\ (S+z)(n) = \max\left[S(n-m) + z(m)\right], m \in y \end{cases} \tag{1}$$

The above two calculations are described as follows: firstly, the structural elements are moved point by point until the origin of the structural elements overlaps with a point in the original data. Then, the amplitude of the structural elements is subtracted by the data amplitude in the definition domain, or the amplitude of the structural elements is added by the data amplitude in the definition domain. Finally, the minimum value and the maximum value of the amplitude are obtained, that is, the corrosion or expansion results of the data (Pölsterl et al., 2016).

On the basis of corrosion and expansion, according to the different order, it can be divided into open operation and closed operation. Open operation is a calculation method that uses structural elements to corrode data first and then expand. The mathematical description is defined as:

$$(S \circ z)(n) = (S \Theta z \oplus z)(n) \tag{2}$$

Closed operation is contrary to open operation. It is a calculation method that uses structural elements to expand data first and then corrode it. Mathematical description is defined as:

$$(S \cdot z)(n) = (S \oplus z \Theta z)(n) \tag{3}$$

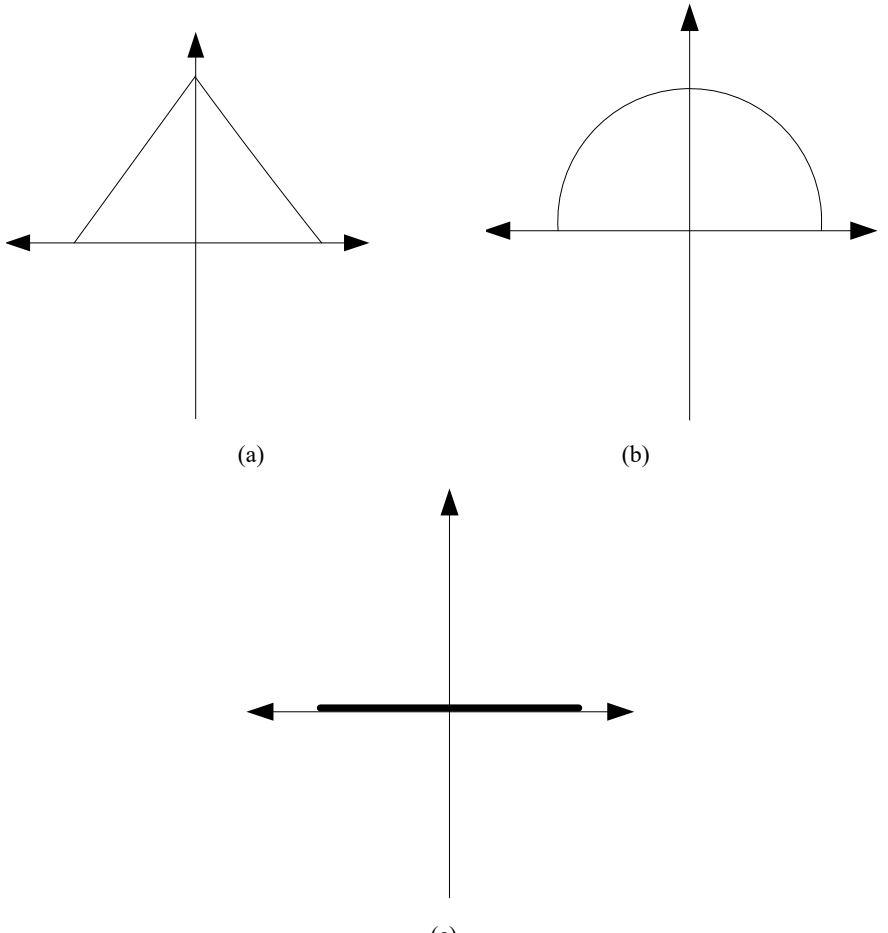
In the formula, Θ denotes corrosion, ⊕ denotes expansion, ○ denotes open operation and • denotes closed operation.

According to the above description, mathematical morphology only involves addition and subtraction operations, so the greatest advantage is the speed of problem processing, which is also one of the objectives of this study.

2.2.1.2 Determination of structural elements

In the calculation of corrosion and expansion mentioned above, the concept of structural element is repeatedly mentioned. The rationality of its basic operators and structural elements selection will directly affect the efficiency and quality of data feature extraction. The determination of structural elements mainly includes two aspects: one is the determination of structural elements. The second is to determine the size of structural elements (Shen et al., 2017).

Figure 9 Structured elements, (a) triangle (b) semi-circular (c) flat type



For determining the shape of structural elements, there are three main choices: triangular, semi-circular and flat, as shown in Figure 9.

Among the above structured elements, the third one is the flat structured element. Because of its simple structure and high efficiency, this structure is also chosen as the operation operator in this study (Garcia-Gasulla et al., 2018).

The size of structural elements can not be too large or too small. Over-large will cause over-adherence of closed operation, resulting in false fracture of open operation; too small will result in difficulty of edge connection of closed operation fracture, resulting in larger false fracture of open operation and dirty removal of adhesion (Zeng et al., 2019). Therefore, the size determination of structural elements needs to be calculated by the following formula.

$$L = \frac{(r-1)}{K} + 1 \tag{4}$$

In the formula, L denotes the size of structural elements, r denotes the size of small structural elements, and K denotes the number of times of corrosion or expansion.

2.2.1.3 Morphological gradient

Data characteristics can be obtained by morphological corrosion and expansion, but in the practical application process, it is difficult to ensure the accuracy of extraction due to the lack of prior knowledge. Therefore, it is necessary to construct morphological gradient on the basis of corrosion and expansion (Ji et al., 2017).

Morphological gradient refers to the difference between the post expansion data and the corrosion data, which is based on the extension and supplement, and processed by the gradient structure element, so as to obtain the data characteristics. This method can effectively filter the characteristics of the required data in big data. The mathematical formula is as follows:

$$d = S \oplus z(m) - S\Theta z(m) \tag{5}$$

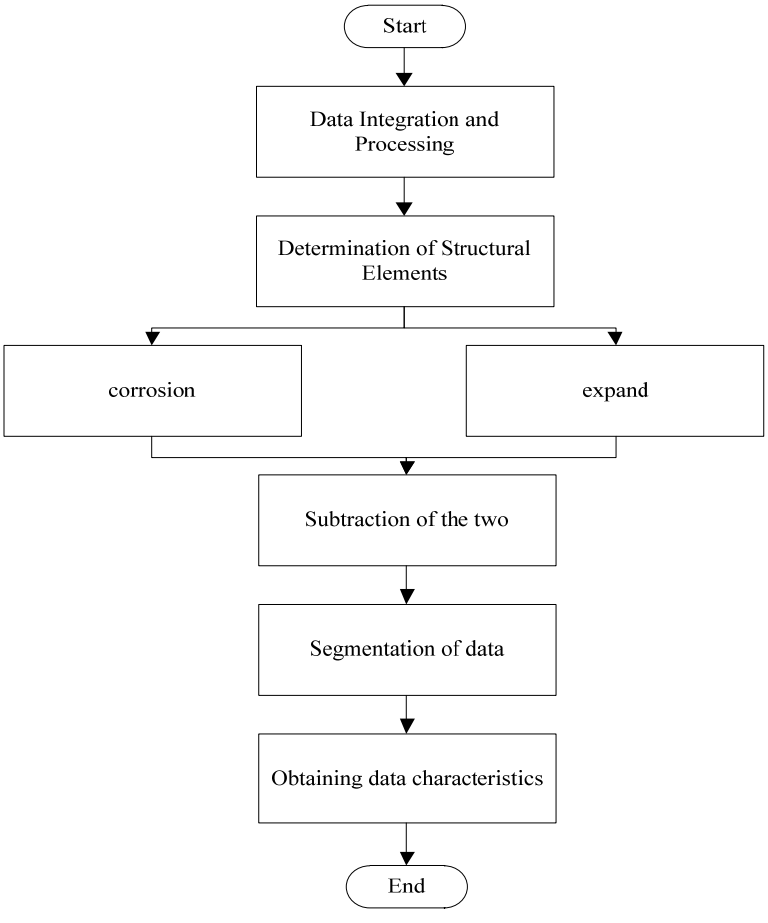
In the formula, d represents the difference between expansion data and corrosion data.

In the process of data feature extraction, when the d value is less than 1, the model will automatically filter out the data, and finally, there will be a part of the data left, and the remaining data is the representative data of the big data, that is, the data feature to be extracted (Eliseev et al., 2018).

2.2.2 Implementation of data feature extraction

Based on morphological gradient, feature extraction of enterprise innovation behaviour data is realised. The specific process is shown in Figure 10.

Figure 10 Implementation flow of data feature extraction based on morphological gradient



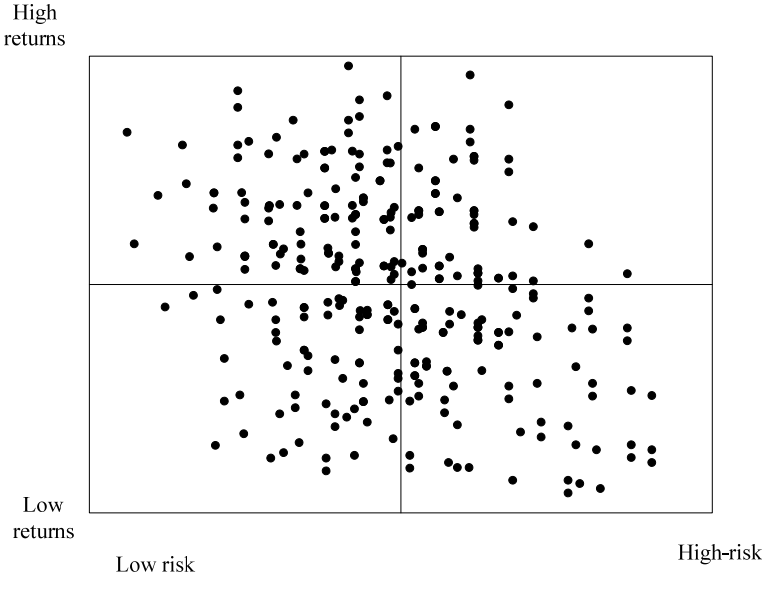
3 Testing of model performance

In order to ensure the authenticity and reliability of the experimental results. The experimental data come from a food processing enterprise, whose production efficiency can not meet people's demand through market reaction. Therefore, the enterprise wants to adopt a new production method. Using JDK 1.6.0 program development software, the data feature extraction model of this paper and three traditional data feature extraction models (the feature extraction models based on data mining, neural network, principal component analysis) are written into the system operation program, and the extraction speed and accuracy are tested by the powerful computing ability of the computer. In this experiment, the computer software environment is Windows 7 operating system, and the hardware environment is Intel Core i74670-3.4ghz, 8.ogb memory, 500GB hard disk. In the experiment, the method of Xiao and Yue (2017) and Tung and Jordann (2017) were

used as the experimental control group, and the accuracy and extraction time were used as the evaluation indexes of the method. When extracting the innovative behaviours of enterprises, their accuracy and time-consuming performance are the main defects of the traditional methods, and they are also the main problems solved in this paper.

In order to ensure the accuracy of decision-making, enterprises need to extract data features from big data. The risk and profit of this production method are analysed. Figure 11 shows the distribution map mapped to the computer after data integration.

Figure 11 Distribution of the original integrated dataset



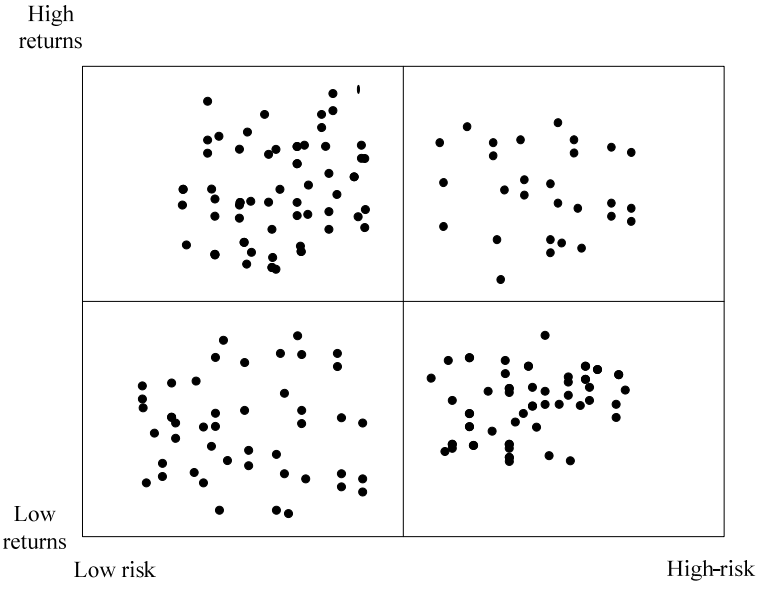
In the original integrated data, the data distribution is shown in Table 2.

Table 2 Data distribution

Key feature codes	*Number of risk data*	*Number of revenue data*
1	457	784
2	427	687
3	369	547
4	284	663
5	358	503
6	571	539
7	555	614
8	326	698

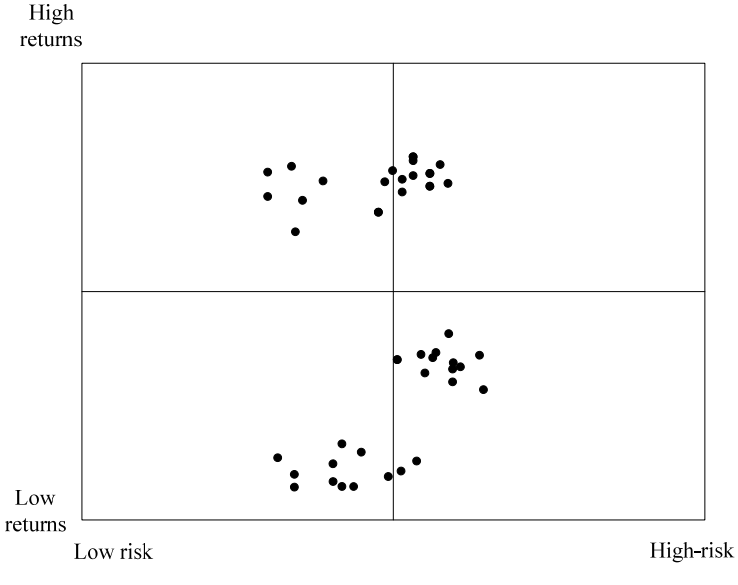
Data processing: the original integrated datasets are processed, including synthesis, filtering and so on. The final results of the processing are shown in Figure 12.

Figure 12 Final results after data processing



Data feature extraction:

Figure 13 Results of data feature extraction



Result analysis:

Table 3 Comparison of extraction speed and accuracy

Data feature extraction model	*This model*	*Data mining*	*Neural network*	*Principal component analysis*
Extraction time(min)	15.68	25.87	18.94	18.65
Accuracy of extraction (%)	96.68	88.35	90.57	95.14

From Table 3, using the model based on data mining to extract the data characteristics of enterprise innovation behaviour, it takes 25.87 min in total, the comprehensive extraction accuracy is 88.35%; the total time of extracting data characteristics based on neural network is 18.94 min, the comprehensive extraction accuracy is 90.57%; the total time of extracting data characteristics based on principal component analysis is 18.65 min, the comprehensive extraction accuracy is 95.14%;Using the morphological gradient model of this study to extract the data characteristics of innovation behaviour of enterprises, it takes 15.68 minutes to complete the whole extraction, and the comprehensive extraction accuracy is 96.68%. This result is much better than that of three traditional models. Because this model determines the structure and size of structural elements, the extraction results are more accurate. This shows that the proposed model can not only quickly extract the characteristics of target data, but also ensure the accuracy of its extraction, which is conducive to the decision-making and implementation of enterprise innovation behaviour.

4 Conclusions

In summary, with the gradual strengthening of market competition, in order to occupy an advantage in it, major enterprises have carried out innovative behaviour. However, this behaviour is not a simple decision-making and implementation process, it needs a large number of data analysis to improve its scientific and rational, so as to protect the profits of enterprises. In this context, the feature extraction model in this study based on morphological gradient for enterprise innovation behaviour data can simplify the data analysis process. Finally, the test results show that the extraction speed of the model is faster and the extraction accuracy is improved, which provides an assistant tool for decision-making and implementation of enterprise innovation behaviour. In the future, we will take the stability of model operation as the research direction, and further improve the application of research methods in the data feature extraction of enterprise innovation behaviour.

References

Crommelinck, S., Bennet, R., Gerke, M. et al. (2016) 'Review of automatic feature extraction from high-resolution optical sensor data for UAV-based cadastral mapping', *Remote Sensing*, Vol. 8, No. 8, p.689.

Eliseev, E.A., Fomichov, Y.M., Yevhen, M., Kalinin, S.V. et al. (2018) 'Labyrinthine domains in ferroelectric nanoparticles: a manifestation of gradient-driven morphological transition', *Physical Review B*, Vol. 98, No. 5, pp.472–492.

Feng, C., Shi, B. and Kang, R. (2017) 'Does environmental policy reduce enterprise innovation? – Evidence from China', *Sustainability*, Vol. 9, No. 6, p.872.

Garcia-Gasulla, D., Parés, F., Vilalta, A. et al. (2018) 'On the behavior of convolutional nets for feature extraction', *Journal of Artificial Intelligence Research*, Vol. 61, No. 31, pp.459–472.

Hopf, K., Sodenkamp, M., Kozlovkiy, I. et al. (2016) 'Feature extraction and filtering for household classification based on smart electricity meter data', *Computer Science – Research and Development*, Vol. 31, No. 3, pp.141–148.

Ji, T.Y., Shi, M.J., Li, M.S. et al. (2017) 'Current transformer saturation detection using morphological gradient and morphological decomposition and its hardware implementation', *IEEE Transactions on Industrial Electronics*, Vol. PP, No. 99, p.1.

Kavakiotis, I., Xochelli, A., Agathangelidis, A. et al. (2016) 'Integrating multiple immunogenetic data sources for feature extraction and mining somatic hypermutation patterns: the case of 'towards analysis' in chronic lymphocytic leukaemia', *BMC Bioinformatics*, Vol. 17, No. Suppl 5, pp.375–387.

Liu, J.X., Wang, D., Gao, Y.L. et al. (2017a) 'A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis', *Neurocomputing*, Vol. 228, No. C, pp.263–269.

Liu, L., Yin, R., Song, W. et al. (2017b) 'An effective feature extraction method on protein secondary structure class prediction', *Journal of Bionanoscience*, Vol. 11, No. 5, pp.446–454.

Liu, R. and Gillies, D.F. (2016) 'Overfitting in linear feature extraction for classification of high-dimensional image data', *Pattern Recognition*, Vol. 53, No. C, pp.73–86.

Pölsterl, S., Conjeti, S., Navab, N. et al. (2016) 'Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection', *Artificial Intelligence in Medicine*, Vol. 72, No. 44, pp.1–11.

Shen, J., Xia, J., Shan, Y. et al. (2017) 'Classification model for imbalanced traffic data based on secondary feature extraction', *IET Communications*, Vol. 11, No. 11, pp.1725–1731.

Susto, G.A., Schirru, A., Pampuri, S. et al. (2016) 'Supervised aggregative feature extraction for big data time series regression', *IEEE Transactions on Industrial Informatics*, Vol. 12, No. 3, pp.1243–1252.

Tung, W.F. and Jordann, G. (2017) 'Crowdsourcing social network service for social enterprise innovation', *Information Systems Frontiers*, Vol. 19, No. 6, pp.1–17.

Xiao, Z. and Yue, J. (2017) 'Measurement model and its application of enterprise innovation capability based on matter element extension theory', *Procedia Engineering*, Vol. 174, pp.275–280.

Xing, L. and Jiang, S. (2016) 'Bank equity connections, intellectual property protection and enterprise innovation – a bank ownership perspective', *China Journal of Accounting Research*, Vol. 9, No. 3, pp.207–233.

Xue, C. and Xu, Y. (2017) 'Influence factor analysis of enterprise IT innovation capacity based on system dynamics', *Procedia Engineering*, Vol. 174, pp.232–239.

Zeng, Q., Adu, J., Liu, J. et al. (2019) ;Real-time adaptive visible and infrared image registration based on morphological gradient and C_SIFT', *Journal of Real-Time Image Processing*, Vol. 39, No. 17, pp.1–13.

Zollo, M., Bettinazzi, E.L.M., Neumann, K. et al. (2016) 'Toward a comprehensive model of organizational evolution: dynamic capabilities for innovation and adaptation of the enterprise model', *Global Strategy Journal*, Vol. 6, No. 3, pp.225–244.