Construction and Retrieval of New Ventures' Growth Risk Case Based on Knowledge Gene Expression Method

Bo Yang* and Yiming Liao
School of Information Management, Jiangxi University of Finance and Economics, China.
Email: 1144175053@qq.com

Abstract. Constructing a knowledge service model of risk cases to realize the internal and external risks of new ventures is an important means of its risk management and prevention. And case expression is the key to improving the efficiency of case retrieval and reasoning. Combining the basic characteristics of new ventures, this paper constructs a gene mapping of growth risk cases based on knowledge gene expression, which includes a three-layer factors structure of Risk-Countermeasures-Effectors. The gene mapping constructed in this paper has an excellent systemic and hierarchical structure, reflecting the connection between sub-factors, and also describes the risk case knowledge clearly and progressively. It makes up for the shortcomings of traditional ontology and structural representation and improves the efficiency of case storage and retrieval, while providing managers with a decision-making reference. In addition, this paper uses the longest common sub-sequence algorithm to evaluate the similarity of the risk level of cases, which is used for the preliminary retrieval of risk cases. Finally, the advantages of risk case representation based on knowledge gene expression compared with the framework and ontology model are compared, and the model of risk case retrieval is constructed.

1. Introduction

Compared with mature enterprises, new ventures have not formed an effective and reliable growth mode in terms of operating mechanism, organizational structure and value construction, etc. In addition, the large amount of risk knowledge resources inside and outside the ventures has brought about the problem of information overload. The ventures hopes to obtain comprehensive and accurate risk information and knowledge to reduce the burden of risk identification and assist decision-making.

The gene of knowledge is the core of knowledge. It is composed of core concepts and relationships between core concepts and unique ways of thinking and academic norms. It determines the adaptability, extension and development of knowledge[1]. Knowledge gene is the basic functional unit of existing knowledge collection. Knowledge genes are expressed in terms of primitive concepts, concepts, methods, skills, relationships, etc. It has stability, dominance, heritability, variability, and the ability to control the development direction of a knowledge field[10]. Yu and Li et al. (2016) draw on the related research of bio-gene mapping, and propose a case representation method based on gene map to realize the construction of grid emergency case. Zhang and Huang (2019) proposed a user recommendation model based on the gene mapping of social media. Due to the multi-agent, multi-risk sources and dynamic characteristics of the growth risk management of new ventures, it is difficult to discover the inherent evolution of growth risk events through data processing. Therefore, knowledge gene expression can be used to achieve risk case presentation for risk decision-making. This method matches the growth risk cases with the highest similarity through knowledge storage, matching, and
retrieval of past risk events and response plans, combined with the heterogeneous characteristics of current risk events, to efficiently obtain risk prevention and control strategies.

On the other hand, according to bioinformatics related research, gene sequence comparison is through double sequence and multiple sequence alignment to identify differences between sequences to find the similarity and homology of gene sequences. Huang Jiaqi (2015) believes that sequence comparison can detect the similarity relationship (structure and function) between the new sequence and the known sequence in the database, thus providing factual basis for determining the structure and function information of the new sequence. Therefore, based on the sequence characteristics of knowledge gene expression, this paper considers using the most common double sequence gene comparison method in bioinformatics-the longest common subsequence (LCS) model to calculate similarity for case matching. Xu Ke et al. (2019) borrowed the principle of the longest common subsequence model, extracted subject words from decision text, and classified them into goal and action levels according to the characteristics of the decision text. Combined with the sample population to determine the replacement matrix, gap penalty rules and objective function, a similarity comparison analysis model for high-level emergency decision text is constructed. Zhang Zipei (2018) proposed an algorithm based on the longest subsequence to solve the similarity based on the traditional algorithm to solve the similarity, and verified the superiority of the proposed algorithm through experiments. Yu Haiying and Zhao Junlan (2008) elaborated the application of the longest common subsequence algorithm in the similarity measurement of program code structure, and enumerated two algorithms for calculating the optimal value and one algorithm for obtaining the longest common identifier subsequence. According to the optimal value, the structural similarity value is obtained, and then the structural similarity program pair can be found.

In view of this, for the construction of the knowledge service model for the growth risk of new ventures, this paper combines the biological gene mapping theory with the case of new venture growth risk, and proposes a model database of new venture growth risk cases based on the expression of knowledge genes. The risk factors are included in the identification and index measurement of the growth risk of new ventures, and at the same time, the expression and storage of unstructured risk emergency knowledge are realized. In addition, based on the characteristics of the gene arrangement of the risk case library, the longest common subsequence model is used to calculate the similarity of the risk level of the case to conduct a preliminary screening of risk cases. Based on the Euclidean distance, the risk attribute similarity of different data types is calculated, and the structural similarity of the case knowledge genes is used to retrieve the most similar source cases in the case database. Finally, this paper analyzes the advantages of case representation methods based on knowledge gene expression and traditional methods from two aspects of risk sources and case retrieval. The construction of the risk case database can provide new ventures with more efficient and reliable risk knowledge services.

2. Expression of Growth Risk Cases of New Ventures Based on Gene Mapping

2.1. Construction of Growth Risk Gene Mapping for New Ventures

According to the growth risk indicator system of knowledge-intensive new ventures based on risk sources established by Ma (2016), the growth risk case gene mapping can be divided into six gene segments under two types [1].

2.1.1. Risk genes. The data types of risk genes can be divided into numeric type and fuzzy semantic type. Deterministic values, such as the rate of doctoralization of R & D personnel in the cognitive risk factors of managers, account for 17%, etc. While interval values refer to the data that can be changed within a certain range, such that the design defects in R & D risk factors account for 5%-10% of the overall product. Fuzzy semantics describe unstructured statements. For subjective information such as less loopholes, medium and more loopholes in the contract terms, the boundaries of the index level classification are vague.
Genes on risk segments are typically made up of Attribute-weighted base pairs. Attribute names and weights are assigned to the DNA and Phosphoric Acid connected to the base pair, respectively. In this paper, the attributes are sorted according to the size of its weight value. The attribute relationship includes homogeneity, heterogeneity and juxtaposition[2]. Its base pair structure is shown in figure 1:

Figure 1. Model of base pairs of risk genes

1) The risk factors of research and development (r&d) refer to the possibility of product r&d risks caused by immature technology, changes in the environment, uncreative personnel and other factors during product research and development. Taking R & D risk factors as an example, the expression paradigms of various risk factors are as follows: Let \( R = \{ r_i | i = 1,2,...,n \} \), where \( r_i \) denotes the \( i \) th risk factor of research and development, \( n \) denotes the total number of category. They are listed in order of influence as \( R = \{ \text{Technical Risks, Environmental Risks, Staff Creativity Risks} \} \). Let \( \theta_{i1} = \frac{r_{i1}}{\theta_{11}} + \frac{r_{i2}}{\theta_{12}} + ... + \frac{r_{in}}{\theta_{1n}} \), where \( r_i \) denotes the \( i \) th DNA sequence, \( r_j \) and \( \theta_j \) denote the \( j \)th attribute of the \( i \)th gene on the research and development risk factors and the weight or the importance degree of this attribute, respectively, such that \( \sum_{j=1}^{n} \theta_j = 1, \theta_j \in [0,1], \theta_{11} > \theta_{12} > ... \).

2) The risk factors of manager cognitive refer to the possibility that the prediction or decision result caused by the subjective factors of managers deviates from the expected target. Let \( M = \{ m_i | i = 1,2,...,n \} \), where \( m_i \) denotes the category of manager cognitive risk factors, \( n \) denotes the amount of category. The factors are sequenced as \( M = \{ \text{Managers Predict Risks, Management Decision Risks} \} \). Let \( \omega_{i1} = \sum_{j=1}^{n} \omega_{ij} = 1, \omega_{ij} \in [0,1], \omega_{11} > \omega_{12} > ... \).

3) The risk factors of product adaptability refer to the risk factor brought by product performance defects or substitutes in the market to the operation process of product project. Let \( P = \{ p_i | i = 1,2,...,n \} \), where \( p_i \) denotes the category of product adaptability risk factors. The factors are sequenced by weight as \( P = \{ \text{Product Defects Risks, Product Substitution Risks} \} \). Let \( \frac{p_{i1}}{\lambda_{p1}} + \frac{p_{i2}}{\lambda_{p2}} + ... + \frac{p_{in}}{\lambda_{pn}} \), where \( p_i \) and
\( \lambda_{ij} \) denote the \( j \) th attribute of the \( i \) th gene on the product adaptability risk factors and the weight or the importance degree of this attribute, such that \( \sum_{i=1}^{n} \lambda_{ij} = 1 \), \( \lambda_{ij} \in [0,1] \), \( \lambda_{ij} \geq \lambda_{ij} \).

4) The risk factors of social network refer to the possibility of uncertainty and risk in the growth process of a newly created enterprise. Let \( S = \{ S_i | i = 1,2,...,n \} \), where \( S_i \) denotes the category of social network risk factors, \( n \) denotes the total number of category. The factors are sequenced by weight as \( S|_1 \text{Government Risks}, S|_2 \text{Competitive Enterprise Risks}, S|_3 \text{Cooperative Enterprise Risks} \).

Let \( \{ S_j \} = \frac{S_{j1}}{\gamma_{a1}} + \frac{S_{j2}}{\gamma_{a2}} + ... + \frac{S_{jn}}{\gamma_{an}} \), where \( S_{ji} \) and \( \gamma_{ai} \) denote the \( j \) th attribute of the \( i \) th gene on the social network risk factors and the weight or the importance degree of this attribute, such that \( \sum_{i=1,j=1}^{n,n} \gamma_{ai} = 1 \), \( \gamma_{ai} \in [0,1] \), \( \gamma_{ai} \geq \gamma_{ai} \geq ... \).

2.1.2. Countermeasures and effector genes. The risk countermeasures and effector genes reflect the risk response strategies and effects adopted by the source cases. And their base pair structure is shown in figure 2:

![Figure 2. Model of base pairs of countermeasures and effector genes](image)

1) The countermeasure genes indicate the control and prevention measures for the description of the risk problem. The arrangement of the gene fragments is identical to the structure of the risk gene fragment. Let \( C = \{ C_i | i = 1,2,...,n \} \), where \( C_i \) denotes the risk response strategy for each source risk cases, \( n \) denotes the total number of category. Let \( \{ C_j \} = \frac{c_{q1}}{c_{s_{q1}}} + \frac{c_{q2}}{c_{s_{q2}}} + ... + \frac{c_{qn}}{c_{s_{qn}}} \), where \( \{ C_j \} \) denotes the \( i \) th DNA sequence, \( c_{qj} \) is the \( j \) th risk problem on the \( i \) th countermeasure gene, and \( c_{s_{qj}} \) is a description of the coping strategies and prevention measures corresponding to \( c_{qj} \). In addition, \( t_{ast} \) is the time factor of gene \( C_i \) indicating the action initiation time for the risk response measure. And \( \{ s_{ai} \} \) is the spatial factor of gene \( C_i \) indicating the expected rectification department or target location of the response measure. Its base molecule knowledge structure and model expression are shown in Figure 3:
2) Effector genes refer to the evaluation feedback and subsequent rectification measures of the risk events and its prevention and control behavior after the implementation of the risk response strategy. Its base molecule knowledge structure and model expression are shown in Figure 4.

Combining the above six growth risk elements, a conceptual model of the genetic map of growth risk based on knowledge gene expression is constructed as shown in Figure 5. Relevant countermeasure and effector genes, the base molecular structure models of these two genes are as described above.
2.2. Storage and Retrieval of Risk Case Based on Knowledge Gene Expression

Knowledge gene mapping of new venture growth risk cases includes three elements of risk scenario genes, response plans, and effect reviews. They formed a triad including \( \text{RS}, \text{CS}, \text{ES} = \text{RS} \lor M \lor P \lor S \lor \text{CS} \lor \text{ES} \). Let the mapping of the original case set \( O \) on the \( \text{RS} \) is \( T \). Then we can expressed it as

\[
T(O) = \{ rs \in RS \mid \exists cs \in CS, es \in ES, (rs, cs, es) \in O \}.
\]

We assume that the following theorems are followed when storing information according to the gene structure:

For any \( rs \in T(O) \) and \( cs \in CS \), there is only a unique corresponding \( es = es_r^{(rs, cs)} \).

For any \( rs \in T(O) \), there is a unique \( cs \in CS \) corresponding thereto in the non-empty set \( es \in CS \).

The risk database of new ventures is stored in units of the same type of risk cases, and cases with similar risk genes are stored in the same location. Decision makers can intuitively classify and retrieve the contents of risk knowledge gene maps, and use the experience of historical cases to guide the response to new risk events.

The four risk factors of r&d, managerial cognition, product adaptability and social network risk (Ma, 2016) are represented by \( \delta, \varepsilon, \eta \) and \( \tau \). And they can be classified into three levels: low, medium and high risk. We calculate the risk level similarity by the genetic sequence similarity optimization algorithm[9]. The longest sub-sequence whose two sequences are discontinuous is called the Longest Common Sequence (LCS)[9], and the similarity formula is:

![Figure 5. Knowledge gene mapping of new ventures' growth risk case](image-url)
\[ \text{sim}(t, s) = \frac{2 \cdot \text{length}(\text{LCS}(R_t, R_s))}{\text{length}(R_t) + \text{length}(R_s)} \]

where \( t \) and \( s \) denotes the target case and the source case, respectively. \( \text{length}(\cdot) \) is a function of the length of a sequence, if its gene sequence is as follows:

\[ R_1 = \delta, \delta, \varepsilon, \eta, \tau, \xi \]

Then we can calculate that \( \text{LCS}(R_1, R_2) = \delta, \varepsilon, \eta, \xi \) and \( \text{length}(\text{LCS}) = 4 \). The calculated similarity is 0.57.

Threshold \( \alpha \) can be set for case retrieval. Only when \( \text{sim}(R_t, R_s) \geq \alpha \), we believe that the source risk case size is comparable to the target case and incorporate it into the set of available cases.

3. Characteristics and Comparison of Knowledge Gene Expression

This part mainly explains the comparison between the knowledge gene expression method and the main method of present case representation, namely ontology representation and frame representation. The knowledge gene mapping model of new ventures mainly draws on the structural characteristics of biological genes, and the different arrangements and combinations of risk knowledge genes determine the diversity and specificity of risk cases. It effectively combines the advantages of ontology and framework in expression, and overcomes the deficiencies in risk case expression and prevention and control decisions.

3.1. The Degree of Structure of Risk Cases in New Ventures

The basic idea of the structured method is to decompose complex risk case problems, and realize the modular expression and coding of risk problems through a top-down, stepwise refinement method. The risk case ontology model mainly considers the structure of risk cases from the perspective of modeling primitives. Therefore, the lack of hierarchical decomposition in expressing risk cases reflects weaker structural capabilities. Compared with the case framework model, the case representation method based on knowledge gene expression not only inherits the advantages of the framework model but also improves its deficiencies.

On the one hand, similar to the multi-layered organizational structure of Frame system-Subframe-Slot value-Side of the framework model, the knowledge gene expression atlas model divides risk cases into hierarchical models of Knowledge gene mapping-Risk gene fragment-Risk gene unit-Risk prevention and control base pair. It demonstrates the structured ability to gradually refine risk cases from top to bottom. The structure has strong adaptability to different risk cases, especially when expressing new risk event scenarios, which can reflect the correlation between the evolution of risk scenarios and risk genes. On the other hand, the framework model is poorly organized, especially when analyzing specific risk scenarios contained in risk cases, risk managers need to combine different subframes multiple times. The knowledge gene atlas model can clearly express the risk event through the structure of the DNA chain itself, and then connect the risk gene through the base molecular model of the risk gene, which has better connectivity and clarity. It fully expresses the whole process of occurrence, development, evolution and disappearance of risk events, including the interaction between four risk factors and risk response measures.

3.2. Expression of Risk Prevention and Control Knowledge

The expression of risk prevention and control knowledge mainly includes declarative knowledge of risk case scenarios and process knowledge of prevention and control measures. In expressing declarative knowledge, risk factors describe the static prevention and control scenarios of risk cases. It is highly generalized with both ontology and framework models. The essential characteristics of the things with the same attributes in the risk event are extracted and extended to the objects of these attributes. For example, the common attributes of different r&d risk factors are extracted, and r&d risk gene fragments are constructed to express knowledge about risk events related to product r&d.
However, when constructing declarative knowledge, ontology models are mainly based on conceptual ontology of risk domain knowledge. This process requires expertise in many fields to assist in the construction, and it is more difficult to apply to the field of risk cases. In expressing procedural knowledge, the knowledge gene expression method can better describe the dynamics of procedural knowledge. Using the one-to-one correspondence characteristics of the base molecule model and the molecular association structure of the analog ontology network, the expression of procedural knowledge is realized. In addition, it clearly expresses the mutual acceptance of risk case knowledge and prevention and control tasks.

3.3. Retrieval of Risk Cases
The purpose of risk case knowledge gene expression is to facilitate the retrieval and reuse of historical risk cases in the prevention and control of risk events to obtain risk knowledge and experience to assist in decision-making. The knowledge gene expression method inherits the advantages of the ordered structure of the framework model. It describes the attributes of the same object, arranges the risk case knowledge according to the attribute weight, excludes the attributes that affect the similarity of risk events, and calculates the similarity for the attribute values of different data types through the traditional similarity algorithm. In addition, the risk gene is similar to the ontology knowledge element, which has the ability to express risk case information and retrieve risk cases by factors.

Conflicts are likely to exist when searching and matching the frame model. The theorem 1 and 2 of knowledge gene expression ensure the uniqueness of the corresponding risk genes in the same case and avoid this problem. The ontology model is mainly oriented to the matching of conceptual semantics, and the knowledge gene expression method avoids the difficulty of retrieval of risk cases of its numerical symbolic attributes. Due to the uncertainty and complexity of risk events in new ventures, and the correlation between risk factors, it is only one-sided to consider the similarity of risk genes. This article uses the LCS method to calculate the overall similarity of risk cases when conducting risk case retrieval, thereby improving the availability of risk knowledge acquired through retrieval.

![Figure 6. Retrieval process of growth risk case based on knowledge gene expression](image)

The steps for retrieving risk cases of new ventures are shown in Figure 6. Corporate decision makers make risk management decisions in response to current risk events. After the current risk event is mitigated, the risk prevention and control measures are evaluated to construct the target risk case countermeasures and effect genes. It forms a complete genetic map of growth risk cases with the aforementioned four risk genes, and is stored as new case knowledge in the risk case gene bank of the company.

4. Conclusion
This paper puts forward a case base construction and retrieval model based on knowledge gene mapping, which can promote the application of case in the field of enterprise growth risk management.
and prevention and control. The knowledge gene mapping is expressed as the vertical structure level of gene fragment, sub-risk factor and base pair. The case knowledge system has excellent integrity and logic. And the horizontal knowledge structure of risk, countermeasure and effector genes is more intuitively and specifically expressed. It optimizes the connectivity of case expression and the rationality of historical cases’ reuse. The risk gene sequence similarity is calculated based on the longest common subsequence to select source cases with the same risk level. At the same time, the retrieval model obtains the comprehensive similarity of risk cases by calculating the similarity of attribute values and risk gene structures, which further enhances the comprehensiveness and usability of the source case set. As cases grow, we can retrieve similar cases using a more intelligent approach to artificial neural networks. This can further improve its retrieval efficiency.

5. Acknowledgment
This work was financially supported by Social Science Planning Project of Jiangxi Province (19TQ01); Jiangxi Provincial Department of Education Science and Technology Research Key Project (Grant No:GJJ180249); Jiangxi university humanities and social science research project (GL18103).

6. References
[1] Wu Liqun 2005 Knowledge genes, knowledge evolution and knowledge services J. Modern Information 06 177-179.
[2] Yu F, Li X Y and Wang S Y 2016 Structural structure and retrieval method of grid emergency case based on genetic map J. Journal of Systems Management 25(02) 282-292.
[3] Huang J Q 2015 Analysis of sequence alignment algorithm of bioinformatics J. Biotechnology World 11 279.
[4] Zhang J D and Huang Y 2019 Research on User Recommendation Model Based on Social Media Gene Atlas J. Journal of Information 38(08) 188-194.
[5] Xu K, Gai W M and Deng Y F 2019 Similarity comparison analysis model of emergency decision text based on LCS J. China Safety Production Science and Technology 05 11-16.
[6] Zhang Z P 2018 Research on collaborative filtering algorithm based on LCS similarity J. Information and Computer (Theory Edition) 21 51-53.
[7] Yu H Y and Zhao J L 2008 Application of the longest common subsequence algorithm in program code similarity measurement J. Journal of Inner Mongolia University (Natural Science Edition) 02 225-229.
[8] Hu Q, Li J and Li M T 2020 Dynamic mechanism and realization model of information service fusion J. Modern Intelligence 02 11-17.
[9] Wang Y L, Yang B R, Song Z F, Chen Z and Tang J J 2007 Study on the LCS algorithm of gene sequence similarity J. Computer Engineering and Applications 31 45-47.
[10] Liu Z H 2003 New development of knowledge gene theory J. Information Science 12 1243-1245.