Research on the quality risk assessment model for water conservancy projects based on the semi-supervised classification of text

Zhenyu Zou2, Jingchun Feng2, He Wei1, Sheng Li1 & Ke Zhang2*  
1 Business School, Hohai University, Nanjing 211100, China  
2" World Water Valley" and Water Ecological Civilization Collaborative Innovation Center of Jiangsu Province, Nanjing 211100, China  
*Corresponding author’s e-mail: zhangke@nuaa.edu.cn

Abstract. This paper systematically analyzes the quality supervision data system of water conservancy projects. Then, according to the characteristics of water conservancy project quality supervision text, the Word2vec algorithm and TFIDF algorithm are combined to construct a feature extraction system of water conservancy project quality supervision text suitable for short length and few samples. Finally, a semi-supervisory model system consisting of logical regression, simple Bayes, and SVM is introduced to solve the problem of incomplete quality supervision risk data for water conservancy projects. To sum up, on the basis of the three parts, i.e. data system, feature extraction, and semi-supervisory text classification, we build a water conservancy project quality risk assessment framework and provide a data processing tool for machine learning of hydraulic engineering quality risks.

1. Introduction
China's water conservancy industry is in a new period of rapid development, and water conservancy project construction should strengthen the quality of the project itself, to ensure that it can withstand the test. A water conservancy project is an important foundation for economic and social development, and strengthening quality supervision and management is the guarantee to improve the quality level of a water conservancy project. In the National Conference on Product Quality and Safety Supervision of the Fourth Plenary Session of the 19th Central Committee, "How to Prevent Quality and Safety Risks" is an important part of the government supervision and management. The quality risk management work of water conservancy projects is composed of identifying, evaluating, controlling, and preventing quality risk, in which quality risk assessment is the basis of quality risk control and prevention.

At present, the research scope of water conservancy project risk management mainly includes planning risk, engineering design risk, engineering economic risk, engineering cost risk, technical risk in the construction process, project operation risk, environmental risk. Hreinnson[1] used Monte Carlo to assess the risk of hydroelectric power system expansion. Schoustra[2] proposed a new design method from the perspective of project risk, which considers the acceptable risk level of the project to conduct a risk assessment study. Goodfellow[3] combined the bi-county water tunnel project to carry out an efficiency analysis of adding the risk register to the tender documents. Frone[4] used a qualitative risk assessment matrix to analyze the main economic risks of a regional water supply and sanitation project in Romania. Rybak[5] designed questionnaires and analyzed the cost risk factors of water and sewerage...
systems construction by issuing questionnaires. The above studies show that at present what is needed is to study the risk management based on quality supervision data. The existing method of risk assessment of water conservancy projects generally obtains the basic data of risk factors by subjective quantification, neglecting the collection, processing, and use of objective data sources (unstructured data, etc.). There are few ways to evaluate the data with incomplete and uneven risk categories. On account of this, this paper conducts an in-depth study on the quality risk of water conservancy projects by means of a risk assessment model based on the semi-supervised classification of texts.

2. Quality supervision data system for water conservancy projects
In the era of big data, the value of data is highly concerned and valued, and the concept of quality supervision and management of water conservancy projects has ushered in the opportunity of transformation and innovation. With the strengthening of information construction for the quality supervision of water conservancy projects in China in recent years, the quality supervision system of water conservancy projects has been applied in some places, and the registration of enterprise supervision has been realized as well. The on-line operation in areas such as supervision of enterprise behavior, entity quality supervision, record of bad behavior, completion acceptance and record-keeping, has improved the efficiency of quality supervision to a certain extent. The traditional engineering quality supervision information system focuses on the recording process of information, which achieves the goal and effect of lifelong responsibility through information systems to some degree, but it ignores the use of the information itself, resulting in a certain amount of waste. Therefore, this paper establishes the water conservancy project quality supervision data system (see Fig.1), with a purpose to collect quality supervision data for analysis, and serve the quality supervision.

3. Feature extraction technologies for quality risk assessment of water conservancy projects
3.1 The application of Jieba
Jieba is one of the most common Chinese word-split tools, which is used in this paper for word-splits of quality supervision text. The core word-splitting algorithms involved in Jieba mainly include:
(i) Achieve efficient word scanning based on the tree structure (Trie), and construct the directed acyclic graph (DAG) with full-word path.
(ii) Use dynamic planning (DP) to find the largest probability path to find the most likely singing combination.
(iii) Recognize uncovered words based on the Hidden Markov Model (HMM), that is, solve the problem of word-splits decoding.
3.2 Combination of TFIDF and Word2vec

It is difficult for TFIDF to represent the contextual characteristics of each Chinese word, which makes the characteristic meaning of each word isolated and inconsistent with the logic of the language. However, Word2vec bridges the logical relationship between words well when expressing word vector characteristics of text, but the weight of a single word in the text is not taken into account. Under this circumstance, this paper combines Word2vec and TFIDF to enhance the characteristics by weighting the word vector. Four combination modes are adopted to enhance the representation effect of feature vector on text, which are as follows:

(i) Softmax processing of the feature vector

While constructing the feature for water conservancy text, calculate the TFIDF values of the words in the text firstly, then process the values with Softmax to convert them into relative probabilities, and finally, multiply the relative probability directly with the Word2vec word vector to get the word vector of the text.

(ii) Standardization processing of the feature vector

Min-max normalization is also called deviation standardization. Map data to interval \([0,1]\) by linear transformation. This can be achieved by Formula (1).

\[
    x^* = \frac{x - \min}{\text{max} - \min}
\]

Where max is the maximum value of TFIDF data, while min is the minimum value of TFIDF data and max-min represents range in water conservancy text.

(iii) Product processing of the feature vector

Use TFIDF as the weight of each word, and then multiply this word vector directly with Word2vec word vector without any processing, to construct a new word vector. The result is showed by Formula (2).

\[
    \text{TFIDF} \times \text{Word2vec} = [T_1, T_2, \ldots, T_n] \times \begin{bmatrix}
    w_{11} & w_{12} & \cdots & w_{1n} \\
    w_{21} & w_{22} & \cdots & w_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    w_{n1} & w_{n2} & \cdots & w_{nn}
\end{bmatrix}
\]

Where \([T_1, T_2, \ldots, T_n]\) are TFIDF weights of words from the 1st one to the nth one respectively, and \([w_{11}, w_{12}, \ldots, w_{nn}]\) are their word vectors. The word vector for each word is made up of dense data with n dimensions.

(iv) Geometric average processing of the feature vector

In order to enhance the feature representation effect of each word, each dimension of each word in the text is processed with geometric average as the result of the final word vector value of the sentence. For instance, regarding the sentence "rectification, work, not in order", after the product processing of Word2vec and TFIDF, the features of the three, rectification, work, and not in order are represented by \([w_{11}, w_{12}, \ldots, w_{1n}]\), \([w_{21}, w_{22}, \ldots, w_{2n}]\), and \([w_{31}, w_{32}, \ldots, w_{3n}]\) respectively. Where 1 in \(w_{1n}\) represents the 1st word, and \(n\) represents the dimension of the word vector. The final feature is expressed by (3).

\[
    \sqrt[n]{w_{11}^2 + w_{21}^2 + \cdots + w_{1n}^2} \sqrt[n]{w_{12}^2 + w_{22}^2 + \cdots + w_{2n}^2} \cdots \sqrt[n]{w_{1n}^2 + w_{2n}^2 + \cdots + w_{nn}^2}
\]

4. Semi-supervised Learning (SSL)

4.1 Semi-Supervised Clustering of K-means

The core idea of the K-means algorithm is that the samples within the cluster are similar and the samples between clusters are different. Suppose the sample set is \(D = (x_1, x_2, \ldots, x_n)\). K-means algorithm divides the original dataset \(D\) into \(k\) clusters, represented by \(C = (C_1, C_2, \ldots, C_k)\), and it makes the sum of the squares of errors minimal. The definition is represented by Formula (4).
\[ E = \sum_{i=1}^{k} \sum_{x \in C_i} \| x - \mu_i \|^2 \] (4)

Where \( \mu_i \) is the cluster heart, and it is defined as (5).

\[
\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x
\] (5)

4.2 **Construct a collaborative semi-supervised algorithm**

In this paper, the model is constructed by the method of collaborative semi-supervision. The basic principle of collaborative training is based on several weak classifier training, and finally the voting method is used for decision-making, which is a thought of collaborative training. Because Support Vector Machine has a good classification generalization effect, Logistic Regressive has the characteristics of efficient processing of industrial-grade data, and Naive Bayes has natural advantages in text classification, weak classifiers such as Naive Bayes, Support Vector Machine, and Logistic Regressive are selected to support the collaborative training for the research of risk assessment of water conservancy quality in this paper.

5. **The framework design of the quality risk assessment model for water conservancy projects**

The evaluation model framework is designed on the basis of the water conservancy project quality risk assessment objectives discussed above, the water conservancy project quality supervision data, and the key technologies involved in the risk assessment (see Fig.2), which consists of the following three main parts:

(i) **Quality supervision data system.** This part provides data support for the evaluation model and is the basis of the model. The traditional engineering quality supervision information system focuses on the recording process of information, but it neglects the use of information itself, resulting in a certain amount of waste. Therefore, this paper establishes the water conservancy project quality supervision data system, with the purposes to collect quality supervision data for analysis, and serve the quality supervision.

(ii) **Feature extraction of quality supervision text.** Firstly obtain the quality supervision text for processing, then use Word2vec and TFIDF algorithms to train the quality supervision word vector, and combine the TFIDF and Word2vec word vector to get the new quality supervision word vector, that is, four kinds of new word vectors for Softmax processing, standardized processing, geometric average processing, and product processing are constructed to represent quality-monitored text features, thus improving the effect of feature representing.

(iii) **Semi-supervised text classification.** Construct the model framework of water conservancy quality risk assessment by the collaborative semi-supervisory model formed by Naive Bayes, Logistic Regressive, and Support Vector Machine as well as the semi-supervised clustering model based on K-means, and input the data of the feature layer into the model for training, to construct a good semi-supervised prediction model of water conservancy quality risk. Finally, use the model to predict the risk level of water conservancy quality supervision text.
Quality supervision data system → Quality supervision text → Quality risk data

Feature extraction of quality supervision text
- SG-NS
- CBOW-NS
- SG-HS
- CBOW-HS

Word2vec → Word2vec + TFIDF

The combination of feature
- Production processing
- Standard text processing
- Softmax processing
- Geometric average processing

Quality risk assessment model
- Semi-Supervised Clustering
- Collaborative semi-supervision

The text of quality supervision of water conservancy projects

Figure 2. The framework design of the quality risk assessment model.

6. Conclusions
This paper is about the theory research of quality risk assessment of water conservancy projects. Firstly, the quality supervision data system of water conservancy project is constructed. Then, the feature extraction theory of quality supervision text is studied and the feature extraction technologies of TFIDF and Word2vec are introduced, on which the feature extraction method of the combination of TFIDF and Word2vec is constructed. At the same time, the application of semi-supervised clustering method and collaborative semi-supervision method in water conservancy project quality risk assessment are studied, and the model framework of water conservancy project quality risk assessment is designed, which lays the foundation for the follow-up research.

This paper constructs a quality risk assessment model for water conservancy projects. In the future, we will analyze the effect of semi-supervised clustering and collaborative semi-supervision on the quality supervision text risk prediction, to construct a quality risk assessment model more suitable for water conservancy projects according to the information characteristics.

Acknowledgments
This work involves Science and Technology Project Plan "R&D and Application of Smart Platforms for Urban-Rural Construction Projects" in 2018 of Ministry of Housing and Urban-Rural Development (2018-K8-023) and "Research on the Socialization Supervision Mechanism of Engineering Construction Market Subjects" supported by the National Social Science Foundation of China (17BGL156) and Water Science and Technology Program Innovation Project "Research on Post-Regulation Mode for Water Resources Engineering Construction in Guangdong" of Guangdong Province (2017-04).

References
[1] Hreinsson, B E L. Monte Carlo based risk analysis in hydroelectric power system expansion planning in the presence of uncertainty in project cost and capacity[Z]. Thessaloniki: UPECGR: 2003.
[2] Schoustra F, Mockett I, van Gelder P, et al. A new risk-based design approach for hydraulic engineering[J]. Journal of Risk Rsearch. 2004, 7(06):581-597.
[3] Goodfellow R, Headland P J. Transfer of A Project Risk Register from Design Into Construction: Lessons Learned from the Wssci Bi-County Water Tunnel Project[A]. 2009 Rapid Excavation
and Tunneling Conference, Proceedings[C]. USA: Soc Mining Metallurgy & Exploration Inc, 2009: 854-860.

[4] Frone S, Frone D F. Economic risks to a regional water supply and sanitation project in Romania[A]. Emerging Markets Queries in Finance and Business 2014[C], Netherlands: Elsevier Science Bv, 2015(32): 550-557.

[5] Rybka I, Bondar-Nowakowska E, Polonski M. Cost Risk in Water and Sewerage Systems Construction Projects[A]. World Multidisciplinary Civil Engineering-

[6] Architecture-Urban Planning Symposium 2016[C]. Amsterdam: Elsevier Science Bv, 2016(161):163-167.