Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P158-cd

A Knowledge Graph Method for Risk Factor Analysis of Underground Gas Storage

Mingyuan Wu

College of Safety and Ocean Engineering, China University of Petroleum (Beijing), China. Key Laboratory of Oil and Gas Safety and Emergency Technology, Ministry of Emergency Management, China. E-mail: 1045913043@qq.com

Jinqiu Hu*

College of Safety and Ocean Engineering, China University of Petroleum (Beijing), China. Key Laboratory of Oil and Gas Safety and Emergency Technology, Ministry of Emergency Management, China. E-mail: hujq@cup.edu.cn,(corresponding author)

Xiaowen Fan

College of Safety and Ocean Engineering, China University of Petroleum (Beijing), China. Key Laboratory of Oil and Gas Safety and Emergency Technology, Ministry of Emergency Management, China. E-mail: fxwen412@163.com

Laibin Zhang

College of Safety and Ocean Engineering, China University of Petroleum (Beijing), China. Key Laboratory of Oil and Gas Safety and Emergency Technology, Ministry of Emergency Management, China. E-mail: cupanquan@163.com

In recent years, the data from underground gas storage stations have become more complex and scaled up. This paper proposes a knowledge graph method for risk factors analysis to use textual information such as production reports during the operation period of gas storage and underground gas storage. The technique extracts relationships from textual data of the gas storage operation period, identifies risk factors using a Bi-directional Long-Short Term Memory network and Conditional Random Field algorithm (Bi-LSTM-CRF), finds the connections among them, and builds a knowledge graph of risk factors based on the extraction results using Neo4j graph database. In addition, this paper compares Bi-LSTM-CRF with other models, and its accuracy, recall, and F1 value metrics are improved by 3.6%, 2.9%, and 3.2%, respectively. The results show that the Bi-LSTM-CRF risk identification method has the highest accuracy rate of 94.3% and the best results in unstructured text extraction from gas storage reservoirs. This paper proposes that the risk factor analysis method based on a knowledge graph can characterize the relationship between risk factors and effectively improve underground gas storage sites' risk management capability.

Keywords: risk factor analysis; knowledge graph; underground gas storage; relationship extraction; Bi-directional Long-Short Term Memory network Conditional Random Field algorithm (Bi-LSTM-CRF).

1. Introduction

An underground gas storage facility is an industrial structure designed to store large quantities of compressed natural gas (CNG) or liquefied natural gas (LNG). These storage tanks are generally built underground and used for various purposes, including commercial, industrial, and residential use. Underground Gas storage comprises collection stations, injection and extraction stations, injection and extraction pipelines, valve chambers, and other facilities. The facilities are also used for the safe transportation and distribution of natural gas across various industries, including power generation and transportation. Many technologies are involved in underground gas storage, including remote monitoring, leakage monitoring, and formation simulation. The process of underground gas storage can be divided into two stages: gas injection and gas extraction. In the gas injection stage, the reservoir receives natural gas from the outside world through pipelines. Finally, it inputs it into the formation through a series of processes, such as compression and scrubbing. In the gas extraction stage, the reservoir extracts gas from the wellhead of the injection and extraction wells. Finally, it inputs it into the pipeline for other industries through processes such as dehydration, desulfurization, and scrubbing. Underground gas storage is critical to maintaining a dynamic balance between natural gas supply and demand. Changes in store on the production or import side of natural gas, days on the consumption side, seasonal changes in the market, natural disasters, and changes in supply due to unforeseen events can all lead to fluctuations in supply and demand. Underground gas storage helps to improve the flexibility of natural gas supply and the rational planning and operation of transportation infrastructure such as pipelines. Underground gas storage facilities, as an essential energy storage infrastructure, could lead to severe consequences in accidents. Hence, it has high requirements for safety. As in Figure 1, the main types of accidents in gas storage are divided into damage to injection and extraction wells or casing, gas migration during gas injection, and failure of storage ground facilities (Xie et al. 2009).

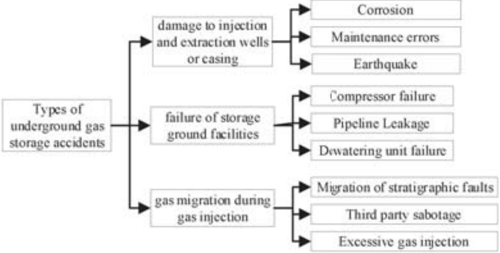

Types of underground gas storage accidents
damage to injection and extraction wells or casing
Corrosion
Maintenance errors
Earthquake
failure of storage ground facilities
Compressor failure
Pipeline Leakage
Dewatering unit failure
gas migration during gas injection
Migration of stratigraphic faults
Third party sabotage
Excessive gas injection

Fig.1 The main types of accidents in underground gas storage

In terms of risk factor analysis of underground gas storage, Ren et al. used the FTA-AHP method to identify and quantitatively assess and evaluate the main control factors of leakage risk of gas storage blocking well (Ren et al. 2018). Pu et al analyzed risk factors from the perspective of risk mechanism and risk factors of significant risks of accidental gas storage reservoirs, established a fault tree model, and conducted a series of qualitative and quantitative analyses (Pu et al. 2022). Ma et al identified the hazards of the above-ground station system regarding safety and environmental risks and establishes a basis for the hazard assessment and evaluation of gas storage reservoirs (Ma et al., 2022). Zhang et al analyzes the difficulties in completing injection and extraction wells in underground gas storage reservoirs and examines the critical risk factors in construction operations (Zhang et al. 2015). Xie et al based on the statistical analysis of the accident data of gas storage reservoirs, an accident tree analysis method is used to analyze the main risk factors that cause accidents (Xie et al. 2009). Zheng et al analyzed the risk factors present in gas storage structures by studying the causes and consequences of seal failure in gas storage reservoirs and classified them into three classes (Zheng et al., 2022). Yu et al analyzed the risk factors existing in the production and operation of gas storage from the occupational health perspective (Yu et al.). Li et al proposes a technical solution for implementing quantitative risk assessment by analyzing risk factors and other aspects. This provides a basis for improving risk control measures in gas storage reservoirs (Li et al.2010). Zhang et al used the butterfly knot and Bayesian network methods to construct a dynamic risk analysis model for the injection and extraction pipeline column thread. It used fuzzy set theory to infer its causal failure factors and analyze its potential risk factors (Zhang et al. 2021). The fuzzy multi-attribute HAZOP technique was proposed to analyze the risk factors of natural gas wellhead facilities in response to the shortcomings of the HAZOP method (Cheraghi et al. 2019). FMECA and HAZOP forms were combined to analyze the risk of LNG storage systems using the logical sequence of cause-deviation-result of process parameters. It was successfully applied to an LNG storage system in Italy (Giardina et al.). A new hazard identification technique, Dynamic Procedure for Atypical Accident Scenario Identification (DyPASI), is proposed for LNG facilities to perform the risk analysis of LNG terminals (Paltrinieri et al. 2015).

Knowledge graphs, as a data representation in computer science, have been widely used in engineering in recent years. Chen et al applied natural language processing to construct an accident evolution knowledge graph for overseas natural gas pipeline stations, which can provide some reference for risk control and accident prevention by on-site safety management (Chen et al. 2022). Li et al applied the named entity recognition model to rapidly construct a HAZOP knowledge graph using existing HAZOP reports to provide HAZOP analysis information (Li et al. 2021). Based on the knowledge graph of hydrocarbon formation logging identification, Yang et al proposes a knowledge-driven hydrocarbon formation evaluation model to facilitate efficient and high-quality identification of hydrocarbon formations (Yang et al. 2022).

Currently, in the gas and oil and gas storage industry, the identification of risk factors in equipment systems is usually chosen through manual analysis of production site data, process system structures, and historical accident reports. This risk analysis method relies on expert experience to identify risk factors through a fixed analysis. However, the unstructured text data generated proliferates with the complexity of gas storage systems. The traditional risk analysis methods could be more efficient, requiring analysts to understand the analysis targets more deeply.

Therefore, this paper proposes a method based on knowledge graph, applying the Bi-LSTM-CRF model to realize the automatic extraction of relationships from unstructured text data, complete the construction of the risk factor list of gas storage, and establish the knowledge graph of gas storage risk factors through the risk factor list, to guide the field operators in risk factor ranking.

2. Fundamental Theory

2.1.*Bi-LSTM-CRF model principle*

LSTM is a special kind of recurrent neural network that can analyze input information using time series and better analyze long text data due to the introduction of the forgetting function (Chen et al. 2022). The network model mainly consists of three parts: the input layer, the hidden layer, and the output layer, and the hidden layer connects the front and back layers, which can make the recurrent network (RNN) have some "memory" ability. However, because the LSTM has a forgetting gate, capturing the contextual relationship in the input text data takes much work. The one-way LSTM can only consider the forward information in the text sequence but cannot handle the backward information. In natural language processing, the text information is cut into individual words. A specific correlation exists between each word’s front and back terms, so the contextual relationship must be considered (Yang et al., 2022). A neural network model capable of storing text, antecedent, and post-text data in real-time, the Bi-LSTM network, has also been proposed to enhance adaptability. This neural network module can perform forward and backward processing of a word successively, which in turn integrates the data derived from the model, thus improving the efficiency and correctness of the module in processing data (Gandhi et al. 2020), (Lin et al. 2019).

The conditional random field (CRF) algorithm is a discriminative algorithm that uses the Hidden Markov algorithm to achieve the construction of relationships between text labels, solving the problem of labeling bias in the Bi-LSTM model through a global normalization process to generate optimal sequences, making the output results reasonable and credible (Li et al. 2021). Therefore, applying the Bi-LSTM-CRF model, which has the features of positive and negative semantic recognition and reinforces the influence of adjacent labels, can effectively extract the relationships between risk factors within the unstructured texts of gas storage reservoirs.

2.2.*Knowledge Graph Principle*

A knowledge graph is a topological network model consisting of nodes and edges, where nodes represent entities and edges represent relationships between entities, represented by a triple = (E, R, S), where E represents entities; R represents relationships between entities; and S represents the set of triples of entities E and relationships R. The triad in the knowledge graph has two forms: {entity, attribute, attribute value} and {entity, relation, entity}. The entity is a comprehensive representation of knowledge; details mainly refer to entities' potential properties and parameter values. This way, entities and relations are formed into a structured semantic relational network through an extensive collection of triples (Liu et al., 2022).

The databases commonly used in knowledge graphs are divided into two categories: relational databases (SQL) and non-relational databases (NoSQL). Among them, the graph data storage method of NoSQL, which is simple and intuitive, and stores data in the form of graphs by constituting graphs through nodes and relationships, not only expresses the complex relationships among data more concisely and clearly but also has inherent advantages in dealing with such tasks as knowledge graphs. This paper adopts a graph database as the data storage method.

3. Risk factor analysis process during the operation period of gas storage

3.1.*Text Reconfiguration*

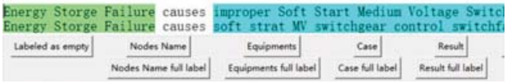
For the unstructured text data in the daily operation of gas storage, 70% of the text is selected as the training set to train the neural network model, and 30% is used as the test set to test the recognition effect of the model subsequently.

Table 1. Textual reconstruction of operating procedures

Pre-processed text information	Energy storage failure due to improper adjustment of the stroke switch in the soft start medium voltage switch cabinet
Cause Node	improper adjustment of the stroke switch in the soft start medium voltage switch cabinet
Result Nodes	Energy storage failure

Before the relationship extraction, the text needs to be reconstructed, and the fault phenomena and causes are saved as plain text in a uniform format as a result of causes. The pre-processing results are shown in Table 1.

As shown in Figure 2, the BIO annotation tool is used to add labels to the unstructured text data in the daily operation of the gas storage reservoir after processing, with "B" and "I" classifying the text information of causes and consequences into two parts: object and behavior, and O indicating the part that does not belong to the causal text. For example, "Leakage at the filter flange cover due to loose bolts", "due to" and "cause" are labeled as "O", "loose bolts", and "Leakage at the filter flange cover" are marked as "B". The following chart of the labeling tool, labeled as empty output for "O", node name output for "Label1", the equipment output for " Label2", "Label3" as the reason, and "Label4" as a result.

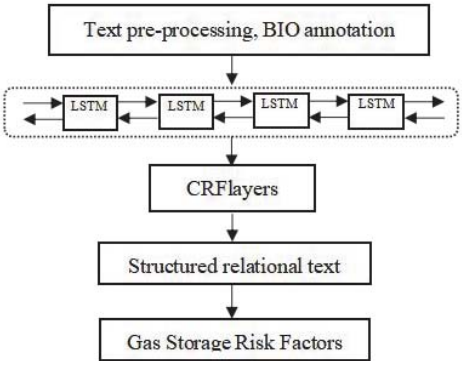


Energy Storge Failure causes improper Soft Start Medium Voltage Switc
Energy Storge Failure causes soft strat MV switchgear control switchf

Fig.2 Text annotation

3.2.*Relationship extraction based on Bi-LSTM-CRF model*

As shown in Figure 3, the processed raw dataset with labels is input into the Bi-LSTM-CRF model to obtain representation vectors with contextual relationships. The CRF layer predicts the relationship division results to receive structured relationship texts. Finally, the text data are organized to form a list of risk factors for gas storage reservoirs.



Text pre-processing, BIO annotation
LSTM
LSTM
LSTM
LSTM
CRFlayers
Structured relational text
Gas Storage Risk Factors

Fig.3 Bi-LSTM-CRF text relationship extraction process

3.3.*Knowledge graph construction of risk factors during the operation period of gas storage*

Neo4j is a more popular graph database that is widely used in the field of knowledge graph research. This database can represent text data in the form of topological networks and realize the construction of knowledge graphs by inputting the labels and names of nodes and the relationships between each node. In this paper, the extracted list of risk factors and the relationships between risk factors are imported into the Neo4j graph database to build a knowledge graph of risk factors of gas storage reservoirs.

4. Cases and Analysis

4.1.*Introduction of data sources*

In this paper, the text data of a gas storage operation period is used, such as "A Gas Storage Operation Regulations (2022 Edition)", as the input data of the relationship extraction model. The dataset belongs to unstructured text data, with more than 100,000 characters.

4.2.*Risk factor extraction during the operation of gas storage*

The Bi-LSTM-CRF model is established, and the manually labeled original dataset with labels is used as the training set of this classifier model to transform the pure text features into distributed features. The model parameters and training parameters are adjusted in the model training process. The learning rate is a hyperparameter that determines the step size at which the model learns during training. Dropout is a regularization technique used to prevent overfitting in neural networks. Epoch is one full pass through the entire training dataset. Accuracy is a measure of a model's performance.

Repeated tests improve the model training effect when the learning rate is 0.001, and the dropout is 0.5. As shown in Figure 4, with the increase in the number of, the model has the highest accuracy rate of 95.9% when the epoch number is 46 times. The accuracy rate (A) is calculated as follows.

$$A = \frac{a_1}{a_2} \times 100\% \tag{1}$$

where a_1 is the number of texts correctly judged by the model and a_2 is the total number of texts.

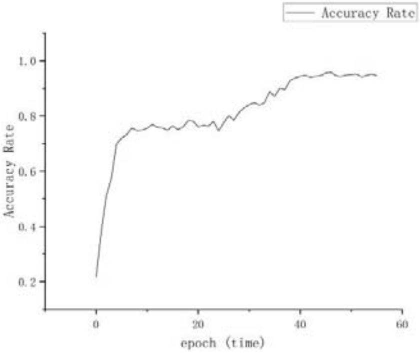


Fig.4 Model accuracy

After obtaining the optimal model, the new samples are sequentially labeled, and the unlabeled test set is input to complete the relation extraction. The model can learn the annotation method and automatically annotate the cause and effect nodes in the sentence. Among them, "due to" and "caused by" reflect the risk relationship, "improper adjustment of travel switch" is the cause node, and "energy storage failure " is the result node.

It is worth noting that the aforementioned program was executed on a system equipped with a 3060 graphics card, a 10700 CPU, and 16GB of RAM. The total running time for the program was approximately 10 minutes.

4.3.*Knowledge graph construction of gas storage risk factors*

4.3.1.*Introduction of risk factors*

The list of risk factors and the relationships between risk factors are imported into Neo4j as a list of data to form a graph database. The name, node type, and label of each node are set to facilitate the management of the database.

After importing all data tables, the overall network of risk factor diagrams is formed. The node-risk relationship and deviation-relationship tables with the relationship type "impact" are combined to obtain a risk factor relationship graph network with 355 nodes, including 248 risk nodes and 107 deviations. There are 822 "influence" relationships.

4.3.2.*Case description*

The Neo4j graph database supports the querying of all relationships for a node. The risk factor graph network is analyzed by taking the collection station emptying system as an example. Nodes represent risk elements; connecting lines represent relationships; each complete line indicates complete risk information and the number of branches, that is, how many problems are risk identification and control focus.

The risk factor node table (part) of the air release system of the collecting station is shown in Table 2.

Table 2 Table of risk factor nodes of the collector station release system (partial)

Nodes	Risk	ID
Collector station emptying system	Air cannot be discharged from clogged pipes	2177
	Fluid accumulation at the bottom of the venting tube	2178
	Excessive throttling of the venting valve leads to full-station venting	2179
	Inability to effectively relieve potential fire and explosion risks	2099
	Pipeline ice plugging	3100
	Artificial fluid drainage is not timely	2180

4.4.*Analysis of risk factors*

Figure 5 presents a comprehensive knowledge graph delineating the risk factors associated with the exhaust system of a gathering station. The analysis reveals that several subsystems, including the instrument air system, the collector station drainage system, the self-gas system within the collector station, the collector station exhaust system, and the dewatering unit (spanning from glycerine input to glycerine output), as well as the dewatering unit's solution regeneration phase (which receives rich liquid from the dewatering unit and regenerates refined glycerine through high temperature thermal oil), function as relatively independent nodes. These subsystems exhibit minimal interdependence.

Contrastingly, each valve chamber exhibits a profound interconnection with each injection and extraction station, presenting numerous common risk nodes. These shared nodes could potentially instigate a multitude of similar risk scenarios. The most significant correlation is observed between systems at the gathering and injection stations. Critical systems such as the inlet measurement system, the inlet system of the gas extraction trunk line, the external transmission measurement and pressure regulation system, and the outlet system of the gathering and injection station show a high degree of interdependence. Any functional failure in these interconnected systems could potentially yield multiple similar risk outcomes.

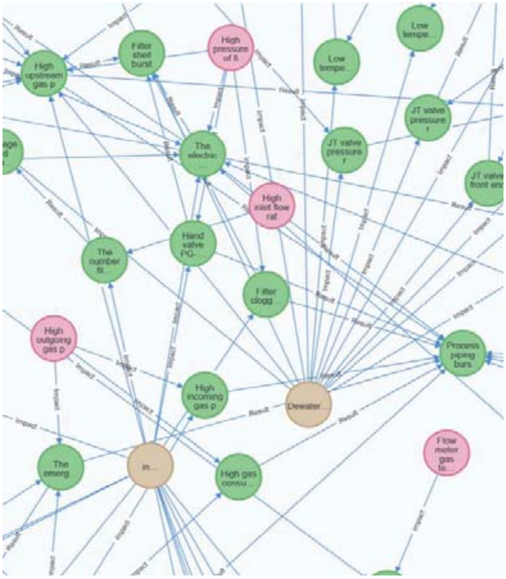


Fig.5 Risk Factor Knowledge Graph for the Gathering Station Exhaust System

4.5.*Comparative analysis of different models*

This paper extracts unstructured texts during the operation of gas storage based on the Bi-LSTM-CRF model, analyzes the types of risk factors, including personnel's misoperation and equipment's failure. Finally constructs a knowledge graph of risk factors. To further verify the superiority of the Bi-LSTM-CRF model for unstructured text extraction of gas storage, the model is compared with the Bi-GRU (Gated Recurrent Unit)-CRF model and the LSTM-CRF model. And the precision rate (P), recall rate (R), and F1 value are used as indicators to evaluate the model's performance, as shown in Figure 6. The formulas are as follows.

$$P = \frac{c_1}{c_2} \times 100\% \tag{2}$$

$$R = \frac{c_1}{c_3} \times 100\% \tag{3}$$

$$F1 = \frac{2 \times (P \times R)}{P + R} \times 100\% \tag{4}$$

Where c_1 is the number of correct results returned, c_2 is the number of all results returned, and c_3 is the number of results that should be returned.

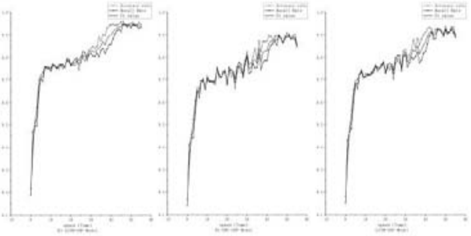


Fig.6 Comparison of the performance of different models

As shown in Table 3, Bi-LSTM-CRF is higher than Bi-GRU-CRF in accuracy, recall, and F1 value metric, indicating that the LSTM model is more effective than the GRU model for unstructured text extraction from gas storage reservoirs; Bi-LSTM-CRF improves 3.6%, 2.9%, and 3.2%, indicating that the introduction of contextual relationship recognition has a more noticeable improvement on the effect of unstructured text extraction from gas storage reservoirs.

Table 3 Table of risk factor nodes of the collector station release system (partial)

Models	Accuracy （%）	Recall （%）	F1 Value （%）
Bi-LSTM-CRF	94.3	93.2	93.7
Bi-GRU-CRF	90.7	90.3	90.5
LSTM-CRF	93.1	92.5	92.8

In addition, because the traditional method of analyzing risk factors is not only limited to the available information but also relies on the expert experience of analysts, the conventional way is higher in the breadth of risk factor coverage. In contrast, the risk factor analysis method based on knowledge graph has a higher depth and refinement of risk factor analysis, which consumes less human, material, and time costs. It can analyze the connection between risk factors, which can improve the risk management capability of the underground gas storage site.

5. Conclusions

This paper uses the BIO annotation method to annotate the text of gas storage operating procedures. The Bi-LSTM-CRF model is used to extract relationships from the annotated text, providing a more refined list of risk factors for gas storage sites. And the Bi-LSTM-CRF model is compared with Bi-GRU-CRF and LSTM-CRF, and it is found that the model has the highest accuracy rate of 94.3%. The Bi-LSTM model structure introduces the identification of contextual relationships, which is the best in unstructured text extraction from gas storage reservoirs.

Based on the risk factor analysis, this paper inputs the list of risk factors into the Neo4j database to construct a knowledge graph of risk factors during the operation period of gas storage, characterizing the relationship between risk factors, which can better target the risk factors during the operation period of gas storage and can effectively improve the risk management capability of gas storage sites.

Acknowledgment

This work was supported by the Natural Science Foundation of China (52074323).

References

Chen Chuan-Gang,Hu Jin-Qiu,Han Z-Con,et al (2022). Knowledge graph modeling and early warning method for overseas natural gas pipeline station accident evolution under harsh environmental conditions. *Journal of Tsinghua University (Natural Science Edition) 62*(6), 1081-1087.

Cheraghi M , Baladeh A E , Khakzad N (2019). A fuzzy multi-attribute HAZOP technique (FMA-HAZOP): application to gas wellhead facilities.*Safety Science 114*, 12-22.

Gandhi H, Attar V (2020). Extracting Aspect Terms using CRF and Bi-LSTM Models. *Procedia Computer Science, 167*:2486-2495

Giardina M, Morale M (2015). Safety study of an LNG regasification plant using an FMECA and HAZOP integrated methodology. *Journal of Loss Prevention in the Process Industries 35*, 35-45.

Li LF, Luo JH, Zhao XW, et al. (2010). Risk assessment techniques and control measures for underground gas storage in salt caverns. *Oil and Gas Storage and Transportation 29*(9), 648-651.

Lin C W, Shao Y, Zhou Y, et al. (2019). A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. *Engineering*

Applications of Artificial Intelligence 85(10), 175-181.

Liu GQ, Gong RB, Shi YJ, et al. (2022). Knowledge graph construction of oil and gas formation logging and its intelligent identification method. *Petroleum Exploration and Development 49*(3), 502-512.

Ma Zenghui, Xu Changfeng, Wang Mingfeng, et al (2022). Integrity management system and risk identification of gas reservoir-type underground gas storage. *China Safety Production Science and Technology 18*(03), 76-81.

Paltrinieri N , Tugnoli A , Cozzani V (2015). Hazard identification for innovative LNG regasification technologies. *Reliability Engineering & System Safety 137*(5), 18-28.

Pu H B, Wang X, Jiao J (2022). Identification of risk factors and analysis of structural importance of wellbore in Jintan gas storage reservoir. *Oil and Gas Storage and Transportation 41*(09),1014-1020.

Ren B, Chen Yuanqiang, and Lei Junjie, et al (2018).Risk evaluation of gas leakage from gas reservoir plugging wells based on FTA-AHP method. *Safety, Health, and Environment 18*(5), 44-47.

Xie LH, Zhang H, Li HELIN (2009).Accident analysis and risk identification of depleted oil and gas reservoir-type underground gas storage. *Natural Gas Industry 29*(11)*,* 116-119.

Yang W , Envelope P , Brij B, et al (2022). An improved entity recognition approach to cyber-social knowledge provision of intellectual property using a CRF- LSTM model. *Pattern Recognition Letters 163*,145-151.

Yu Chunlei. Identification and control of occupational hazard factors in underground gas storage. *Proceedings of the 2014 Symposium on Underground Gas (Oil) Storage and LNG Technology*. 2014, 112-114.

Zhang Qiang, Song Guihua, Zhu Guanghai, et al. (2015). Risk identification and technical requirements for re-completion operations of injection and extraction wells in underground gas storage reservoirs. *Proceedings of the 2015 National Natural Gas Academic Conference*, 24-27.

Zhang Wuming, Yang Shangxiao, Yan Xiangzhen, et al. (2021).Dynamic risk assessment of thread failure of gas reservoir injection and extraction pipe columns. *China Safety Production Science and Technology 17*(3), 84-90.

Zheng YAL, Qiu STS, Lai X, et al. (2022). Risk classification and control of gas reservoir-type underground gas storage geological body injection and extraction operation. *Natural Gas Industry 42*(3), 114-119.