processes

MDPI

Article

Construction Method and Practical Application of Oil and Gas Field Surface Engineering Case Database Based on Knowledge Graph

Taiwu Xia [1], Zhixiang Dai [1], Yihua Zhang [2], Feng Wang [1], Wei Zhang [1], Li Xu [1], Dan Zhou [3] and Jun Zhou [4,*]

[1] Natural Gas Gathering and Transmission Engineering Technology Research Institute, PetroChina Southwest Oil and Gas Field Company, Chengdu 610041, China; snxuli@petrochina.com.cn (L.X.)
[2] Infrastructure Construction Engineering Department, PetroChina Southwest Oil and Gas Field Company, Chengdu 610066, China
[3] School of Intelligent Manufacturing, Panzhihua College, Panzhihua 617000, China
[4] Petroleum Engineering School, Southwest Petroleum University, Chengdu 610500, China
* Correspondence: zhoujunswpu@163.com

Abstract: To address the challenge of quickly and efficiently accessing relevant management experience for a wide range of ground engineering construction projects, supporting project management with information technology is crucial. This includes the establishment of a case database and an application platform for intelligent search and recommendations. The article leverages Optical Character Recognition (OCR) technology, knowledge graph technology, and Natural Language Processing (NLP) technology. It explores the mechanisms for classifying construction cases, methods for constructing a case database, structuring case data, intelligently retrieving and matching cases, and intelligent recommendation methods. This research forms a complete, feasible, and scalable method for deconstructing, storing, intelligently retrieving, and recommending construction cases, providing a theoretical basis for the establishment of a construction case database. It aims to meet the needs of digital project management and intelligent decision-making support in the oil and gas sector, thereby enhancing the efficiency and accuracy of project construction. This work offers a theoretical foundation for the development of an intelligent management platform for ground engineering projects in the oil and gas industry, supporting the sector's digital transformation and intelligent development.

Keywords: engineering construction cases; knowledge graph technology; intelligent retrieval; intelligent push; decision-making assistance

Citation: Xia, T.; Dai, Z.; Zhang, Y.; Wang, F.; Zhang, W.; Xu, L.; Zhou, D.; Zhou, J. Construction Method and Practical Application of Oil and Gas Field Surface Engineering Case Database Based on Knowledge Graph. *Processes* **2024**, *12*, 1088. https://doi.org/10.3390/pr12061088

Academic Editor: Youguo Yan

Received: 12 April 2024
Revised: 19 May 2024
Accepted: 24 May 2024
Published: 25 May 2024

1. Introduction

1.1. Motivation

In recent years, China's oil and gas field ground engineering construction projects have become characterized by their large scale, high quantity, and short timelines, which has significantly increased the burden of construction management. There is an urgent need to improve project management efficiency through information technology. Given the similarities among ground engineering projects, referencing typical project cases during construction is essential for improving efficiency and accuracy. However, there is currently a lack of channels for referencing typical construction cases and systematically managing case data. It is necessary to accelerate the development of methods for organizing construction case databases, establish logical frameworks for managing construction cases, and develop intelligent retrieval and recommendation systems. This will facilitate a decision-making system for compliance, schedule, quality, investment, and other aspects, thus enabling intelligent management applications for ground engineering.

Traditional oil and gas field ground engineering case databases, with their wide range of data sources and varying data quality, often contain errors, outdated or incomplete

information, which leads to low practicality and credibility. Additionally, these databases, established long ago, have limited scalability, hindering data updates and the integration of new features. Their application scope is relatively narrow and cannot meet current ground engineering demands. There is an urgent need to utilize advanced information technology to research engineering construction case database creation, data structuring, and intelligent matching retrieval and push mechanisms. These efforts will develop technical methods for case database entry, supporting the construction of intelligent retrieval and push application platforms for ground engineering, optimizing project construction, acceptance management, and enhancing analysis and decision-making assistance for clusters of oil and gas field ground engineering projects.

1.2. Contributions

(1) Research on establishing a case database using NLP, OCR, and knowledge graphs.
(2) Development of a comprehensive, feasible, and scalable methodology for deconstructing, storing, intelligently searching, and pushing construction case studies.
(3) Creation of a database for oil and gas field surface construction cases, enabling search and push functions.

2. Literature Review

A systematic investigation has been conducted into the field of oil and gas well case databases, both domestically and internationally. In developed Western countries such as the USA and Canada, major information service institutions including IHS Markit, Wood Mackenzie, and National Geological Databases have been established. These organizations possess unique insights and have made significant contributions to the field of databases, enhancing information collection processes, expanding the breadth and depth of data gathered, and effectively managing and utilizing vast information resources. They provide robust technical services for managing large volumes of data.

To date, many scholars have conducted research in different fields related to the creation of databases. Cai S et al. [1] established a spatial information database and management platform based on spatial information of oil and gas resources and related information analysis technology, effectively supporting the entire decision-making process of oil and gas resource evaluation, management, exploration, and exploitation. In the related field of wind energy, Sánchez-del Rey A et al. [2] established a database for wind resource assessment by integrating data related to wind resources with GIS tools. In research focused on carbon emission reduction, Zhu W et al. [3] developed a new carbon emission calculation model for oil and gas resources utilizing GIS 10.8 software and IPCC algorithms to effectively manage the data generated in the exploration and natural processes of oil and natural gas. This model analyzes and predicts future carbon emissions from oil and natural gas production, significantly contributing to carbon emission reduction. Zhixin Wen et al. [4] utilized commercial databases such as S&P Global and Rystad, as well as public databases, to systematically analyze global deepwater oil and gas exploration trends and make recommendations for overseas deepwater oil and gas exploration business. In the mining industry, the establishment of databases remains a key research direction. Jasansky S et al. [5] created an open database on the global production of coal and metal mines to understand production trends, ensuring the authenticity and transparency of the information. Qing Guan et al. [6] extracted knowledge from multisource heterogeneous knowledge carriers to construct a graphical knowledge base, enhancing work efficiency in oil and gas exploration and development processes. To support the digital transformation of the oil and gas industry, Su J et al. [7] analyzed and stored large amounts of information using big data analytics and artificial intelligence methods, breaking down data silos and business barriers, enhancing digital transformation data governance capabilities, and promoting the high-quality development of the oil and gas industry. Similarly, Maroufkhani P et al. [8] identified methods for applying digital technologies in the energy sector to help the resource and energy industry accelerate digital transformation and increase value

creation through these innovations. Wu L et al. [9] built a data-driven decision support system that requires multiple independent databases, utilizing artificial data and large oil and gas project cases for functional validation of the system. This system improves data quality and promotes the utilization of fragmented data.

Knowledge graph technology, as the main method of database construction in this paper, offers a wide range of application areas and good expandability. Many scholars in different fields have conducted research based on knowledge graph technology. Huang et al. [10] built various thematic knowledge graphs based on a large amount of professional oil and gas information and proposed an intelligent search engine based on knowledge graphs that better understands users' search intentions. Tang X et al. [11] proposed a method for constructing a domain knowledge graph based on the ontology of petroleum exploration and development on the basis of traditional NLP (Natural Language Processing), achieving a recognition accuracy rate of 90%, thereby providing better knowledge services for the oil and gas industry. Yuan J et al. [12] proposed a data semantic standardization methodology based on the knowledge organization model, addressing the problem of knowledge graph data sharing, and verified the application of the model in the petroleum industry through relevant examples. Compared with traditional machine learning methods, Zhou X G et al. [13] used a variety of NLP methods to build knowledge graphs from structured and unstructured data, demonstrating that the accuracy of the knowledge graph method increased by more than 7.69%. To reduce the risk of accidents in natural gas pipelines, Bai Y et al. [14] proposed a novel risk assessment model based on knowledge graphs that objectively support the safety management and risk reduction of natural gas pipelines and other process units in the digital era. Knowledge graph technology is also applicable to other fields. Pei Y et al. [15] utilized a large amount of data from mines and constructed a knowledge graph of mines based on deep learning and NLP techniques, extracting the mineralization laws and standard information for finding mines, which saves a significant amount of time and money.

In summary, there remains a lack of research on the construction methods and applications of case databases for oil and gas field ground engineering projects. The application of ground engineering case databases is insufficient, and their practical application efficiency is low, hindering the effective support of the digital and intelligent development of oil and gas fields.

3. Research Framework and Technology Introduction

3.1. Research Framework

In addressing the construction method and functions of a database for oil and gas field ground engineering project cases, it can be divided into the following six parts: data processing, case categorization, case decomposition, case storage, intelligent retrieval, and intelligent push. Among these, data processing, case categorization, case decomposition, and case storage belong to the construction method of the case database, while intelligent retrieval and intelligent push are corresponding functions of the application platform based on the case database. The method and application of constructing engineering project cases are shown in Figure 1.

(1) Data processing

Collect materials related to oil and gas field ground engineering cases, including paper and electronic documents. Paper documents require OCR technology for text recognition. Convert text data into a computer-readable format, manually correct errors, and store it in the cloud, enriching the case database. After integrating all materials, use NLP technology for text preprocessing, providing a data basis for the case database.

(2) Case categorization

Initially, categorize individual engineering cases and typical construction cases (single projects) into 33 subcategories under the following 3 main categories: key procedures, special environments, and risky operations, based on the hierarchical classification method

and case content. Next, generate keywords using the TF-IDF algorithm and establish a storage classification coordinate system that reflects logic and correlation to simplify subsequent case decomposition processes.

(3) Case decomposition

Propose a case decomposition method guided by nine major decomposition items derived from the storage classification coordinate system. This method acquires structured data of ground engineering cases, facilitating subsequent case storage and forming the core content for storage, comprising original case files and structured data, to achieve the decomposition of oil and gas field ground engineering cases.

(4) Case storage

First, utilize knowledge graph technology to build a knowledge graph for oil and gas field ground engineering cases, providing underlying logic for case storage. Then, match the structured data from case decomposition to the established storage classification coordinate system. Integrate original files, structured fields, storage classification coordinates, and the knowledge graph database of ground engineering cases to realize case storage, thereby constructing a knowledge graph-based database for oil and gas field ground engineering cases.

(5) Intelligent retrieval

Implement intelligent retrieval in different scenarios based on NLP technology and knowledge graph matching techniques [16]. After understanding users' retrieval intentions, conduct precise or fuzzy searches based on their query methods.

(6) Intelligent push

Utilize knowledge graph technology to intelligently push structured content derived from intelligent retrieval.

The aforementioned process can provide theoretical and technical support for building an intelligent retrieval and push platform for engineering project cases.

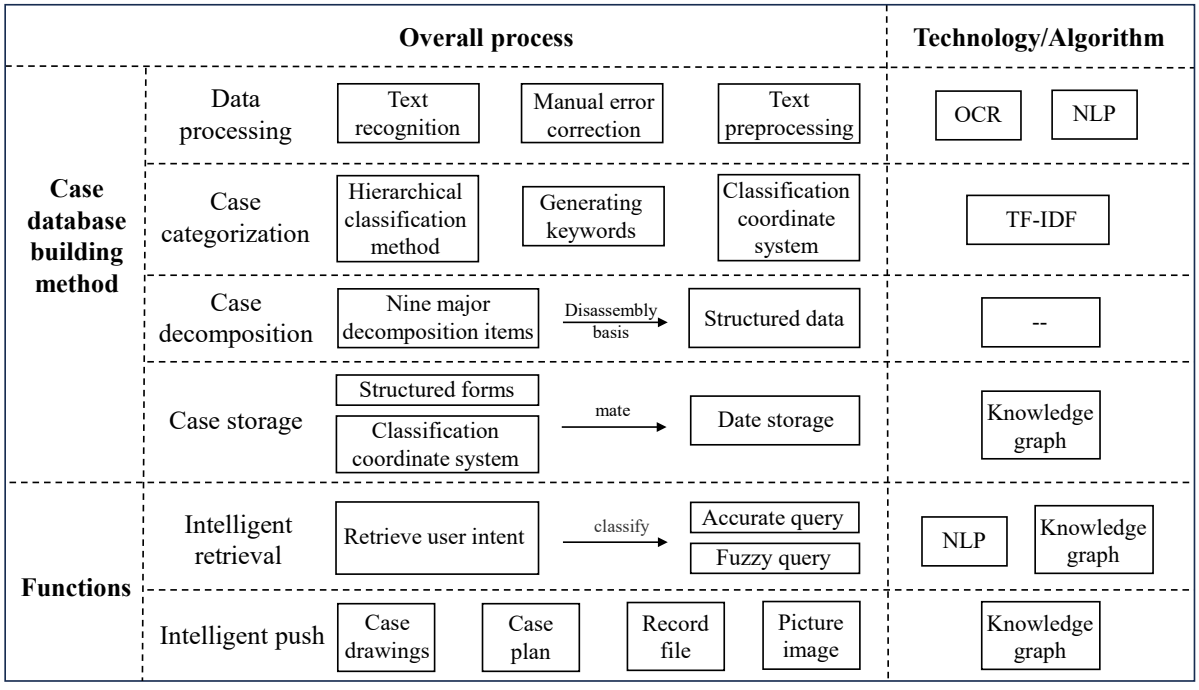

Figure 1. Engineering construction case library construction method and application.

3.2. Database Building Technology

In the construction of a database for oil and gas field engineering project cases, the complexity of its database construction mechanism necessitates corresponding technical support to refine the construction process. This primarily involves OCR technology, knowledge graph technology, and NLP technology.

(1) OCR intelligent recognition technology

OCR intelligent recognition technology can establish an intelligent scanning and recognition platform to complete text scanning tasks with high quality, constructing a more efficient business architecture, thus achieving cost reduction and efficiency increase [17,18]. In this study, its primary role is to scan and store relevant paper materials of oil and gas field ground engineering construction cases, providing the data foundation for subsequent project case structural decomposition.

(2) Knowledge graph technology

Knowledge graph technology is the most crucial technical support for constructing a case database. Building a knowledge graph based on the data of oil and gas field construction engineering cases provides a solid foundation for the intelligent retrieval and push functions of subsequent engineering case databases. The construction of a knowledge graph in the oil and gas field domain comprises the following three key technical processes: knowledge extraction, knowledge fusion, and knowledge reasoning [19].

(3) NLP technology

In the intelligent retrieval of oil and gas field ground engineering construction cases, NLP technology is used to process the natural language input by users, converting it into computer language for easier recognition by the system. This can better identify the users' retrieval intentions, improving the response capabilities and accuracy of the intelligent retrieval system [20]. Combined with the actual needs of intelligent retrieval in the engineering case database, achieving natural language understanding is sufficient.

3.3. Construction of the Knowledge Graph

The construction of a knowledge graph for oil and gas field engineering cases encounters the following challenges: (1) the field of oil and gas surface engineering is broad, covering many different areas of expertise, leading to a complex knowledge structure that includes noise data, affecting the accuracy and professionalism of the knowledge graph. Therefore, human intervention is needed to select and control the target data. (2) The uniqueness of oil and gas field surface engineering cases often means that ensuring the completeness of case data is challenging, and the differences in format, structure, and quality across various data sources complicate data processing. (3) Building a database knowledge graph for engineering cases involves handling a large volume of data, which may grow continuously over time. Additionally, the knowledge graph requires regular updates and maintenance, as well as the processing of new and changing information [21].

The construction of the knowledge graph is crucial, acting as a bridge between processes and as one of the foundational projects in building a case database [22]. The construction process of the knowledge graph, as shown in Figure 2, involves knowledge extraction, knowledge fusion, and knowledge reasoning steps, culminating in the construction of the engineering case database knowledge graph.

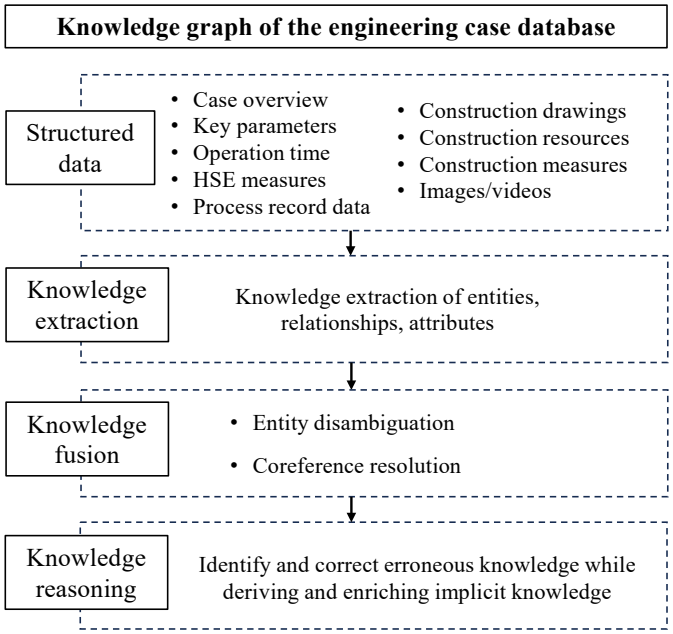

Figure 2. Schematic diagram of construction process of knowledge graph of engineering case library.

3.3.1. Knowledge Extraction

Knowledge extraction aims to automatically obtain entities, relationships, attributes, and other knowledge from structured, semi-structured, or unstructured data. After data processing and decomposition, the case data are structured, allowing for faster knowledge extraction [23]. The Lattice LSTM entity extraction model is used for information collection, effectively utilizing Chinese lexical information and avoiding errors in Chinese word segmentation. The structure of knowledge extraction for the engineering construction cases is shown in Figure 3.

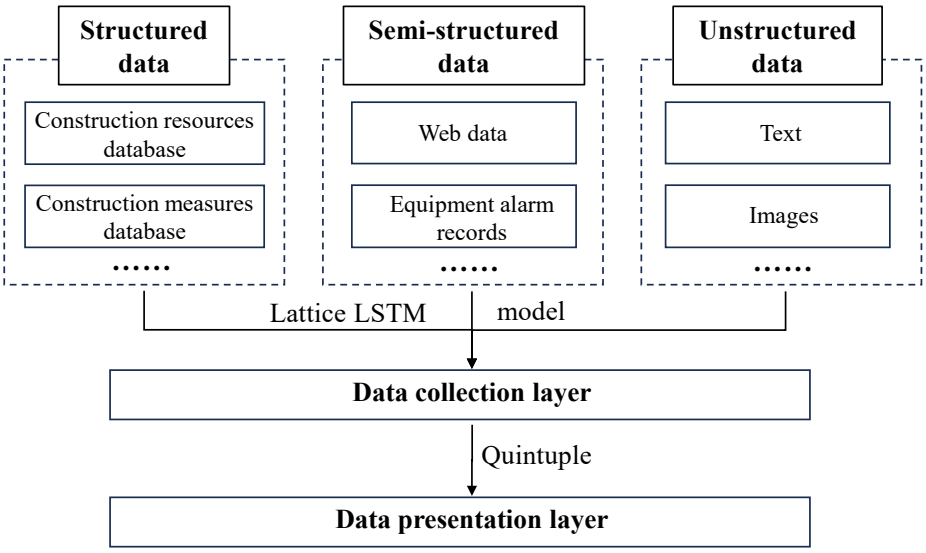

Figure 3. Schematic of the knowledge extraction structure for construction cases.

(1) Entity extraction

In this study, the Lattice LSTM entity extraction model is used for the entity extraction phase. It can be considered an extension of the character-based Named Entity Recognition model, with words added as input vectors, such as "oil and gas field" in Figure 4, and additional gates to control the flow of information.

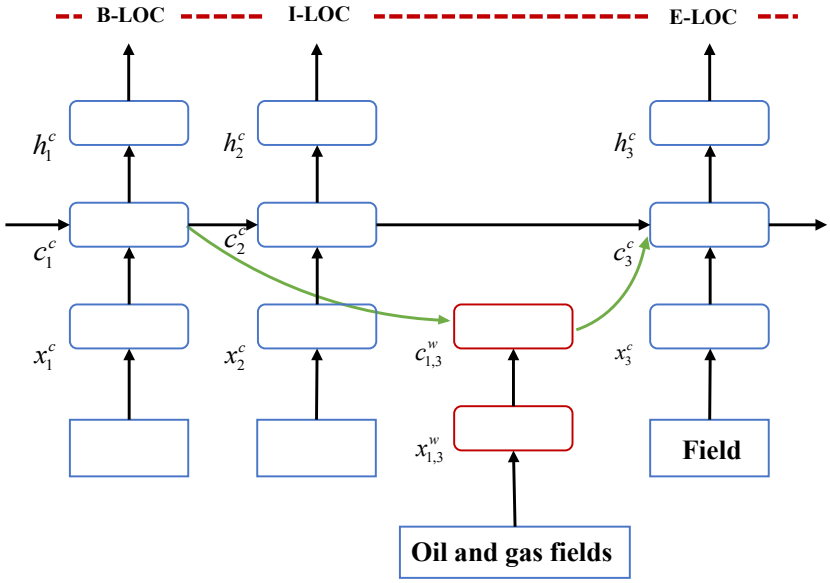

Figure 4. Lattice LSTM model.

As shown in Figure 4, the model's inputs are a character sequence $c_1, c_2, \ldots, c_m$, and all character subsequences in the case database D that match words. The subsequence $w_{b,e}^d$ starts with character index b and ends with character index e, where $w_{1,2}^d$ is "oil" and $w_{1,3}^d$ is the "oil field".

For character c_j, character representation is obtained through a character embedding matrix. Then, using the input gate i_j^c, the forget gate f_c, output gate o_j^c, the unit vector c_j^c and hidden vector h_j^c are derived. c_j^c records the cyclic information flow of the sentence, while h_j^c is used for sequence labeling. LSTM recurrent calculations are then performed, and the basic LSTM function is shown in Equation (4). Next, softmax calculations are performed based on the input gates of characters and words to obtain weighted coefficients, followed by weighted summation, laying the foundation for subsequent training [24].

$$\begin{aligned} \begin{bmatrix} \mathbf{i}_j^c \\ \mathbf{o}_j^c \\ \mathbf{f}_j^c \\ \widetilde{\mathbf{c}}_j^c \end{bmatrix} &= \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\mathbf{W}^{c\top} \begin{bmatrix} \mathbf{x}_j^c \\ \mathbf{h}_{j-1}^c \end{bmatrix} + \mathbf{b}^c \right) \\ \mathbf{c}_j^c &= \mathbf{f}_j^c \odot \mathbf{c}_{j-1}^c + \mathbf{i}_j^c \odot \widetilde{\mathbf{c}}_j^c \\ \mathbf{h}_j^c &= \mathbf{o}_j^c \odot \tanh\left(\mathbf{c}_j^c\right) \end{aligned} \tag{1}$$

In the equation, $\mathbf{i}_j^c$, $\mathbf{f}_j^c$, and $\mathbf{o}_j^c$ represent a set of input gates, forget gates, and output gates, respectively, while $\mathbf{W}^{c\top}$ and $\mathbf{b}^c$ are model parameters, and σ represents the sigmoid function.

Based on the basic framework of the Lattice LSTM model and considering the objectives and requirements of the intelligent search system, a set of entity extraction steps suitable for the construction case knowledge graph model has been summarized. The process is as follows: ① Selection of Historical Data: considering the balance and accuracy of evaluation, it is divided into a training set and test set according to the ratio of 7:3; ② Entity Annotation: the BIO (Beginning, Inside, Outside) tagging method is used to annotate entities in the case information; ③ Model Training: the training set is inputted into the Lattice LSTM entity extraction model based on the process described above; ④ Model Testing and Validation: after training, the test set is used to validate the effectiveness of the entity extraction model. If the accuracy rate is greater than 90%, the entity extraction based

on the LSTM algorithm is considered complete. If the accuracy is below 90%, training continues until the accuracy reaches or exceeds 90% [25].

This methodical approach ensures a high level of accuracy in entity extraction, which is crucial for constructing an effective knowledge graph for engineering construction cases.

(2) Relation extraction

Most entities obtained from entity extraction have not formed associations and need to be connected through relation extraction. Based on the entity relationship table, sentences in the case text containing entity relationships are matched to generate a quintuple template, as follows: Left + Entity1 + Middle + Entity2 + Right. The elements of the template can be set according to actual conditions, with Left, Middle, and Right being word vectors. The generated template undergoes similarity calculations, detailed as follows: Convert the target and comparison templates into specific formats and calculate similarity. If the entity types match (T1 = T1′ or T2 = T2′), calculate the similarity as Sim = W1L1L2 + W2M1M2 + W3R1R2, where W1, W2, W3 are weights, with W2 usually being the largest due to the significant influence of the middle word vector M1M2.

If the similarity exceeds 0.85, the relationship is added to the entity relationship table; if it is below 0.85, the template relationship is discarded. This process is repeated until all the text contents of oil and gas field surface engineering cases are processed, completing the knowledge extraction task and providing a stable data foundation for subsequent intelligent search and push phases.

3.3.2. Knowledge Fusion

In the diverse and complex field of surface engineering projects in oil and gas fields, the data structures of different project cases vary significantly. Utilizing knowledge fusion technology to integrate and merge multi-source knowledge to establish a more comprehensive, consistent, and accurate knowledge graph is crucial. It is also an effective method to address heterogeneity issues in knowledge graphs [26]. The main tasks of knowledge fusion include entity disambiguation and coreference resolution. (1) Entity disambiguation involves aligning different names or identifiers that describe the same entity across multiple data sources or contexts. In this study, string matching is primarily used for entity disambiguation. (2) Coreference resolution involves normalizing words that refer to the same entity across different expressions into a unified representation. In this study, coreference chain resolution is primarily used for coreference resolution [27].

3.3.3. Knowledge Inference

Knowledge inference, following extraction and fusion, identifies and corrects erroneous knowledge while deriving and enriching implicit knowledge based on existing information, enhancing the target knowledge base. This process is crucial for content correction and quality control of knowledge graphs. In this article, correction and mining are achieved primarily through logical reasoning and graph structure reasoning [28]. (1) Logical reasoning is divided into reasoning based on first-order predicate logic and reasoning based on description logic, both of which are utilized in the knowledge graph of engineering construction cases. (2) Graph structure-based reasoning involves using the inherent structure of the knowledge graph as a feature to complete reasoning tasks [29].

By integrating these reasoning techniques, the study effectively enhances the depth and breadth of knowledge representation in the engineering case knowledge graph. This comprehensive approach allows for more accurate and meaningful insights to be derived from the data, supports intelligent search and push functions, and aids the decision-making process in oil and gas field surface engineering projects.

3.4. Knowledge Graph-Based Matching Technique

Knowledge graph matching is crucial for developing case retrieval functionality, especially given the complexity and vast number of oil and gas field engineering projects. To address the challenges of large-scale knowledge graph matching, a cluster-based large

ontology matching algorithm has been adopted. This algorithm is particularly suitable for knowledge graphs with many concepts and properties, such as those in oil and gas field surface engineering. The process involves ontology segmentation, block matching, and discovering matching results. Concepts in the ontology are clustered into smaller clusters, and blocks are constructed by describing the relationships between them. Blocks from different knowledge graphs are matched based on pre-calculated structural similarities, and highly similar mapping blocks are selected for matching [30].

For the extensive knowledge graph of oil and gas field surface engineering cases, this method clusters and specifies a large number of concepts or properties, ensuring the stability of the clusters. The clusters are generally based on structural similarity. If c_i and c_j are two classes in the case database, their structural similarity is defined as shown in Equation (2) below:

$$prox(c_i, c_j) = \frac{2 \times depth(c_{ij})}{depth(c_i) + depth(c_j)} \quad (2)$$

In the formula, c_{ij} is the common parent class of c_i and c_j; $depth(c_i)$ and $depth(c_j)$ are the depths of c_i and c_j in the original inheritance relationship, respectively.

The structural similarity formula is used to compare classes and select the most similar ones for matching. To enhance accuracy and speed, language-based matcher V-Doc and structure-based matcher GMO are used as auxiliary matchers. This approach ultimately achieves higher matching accuracy and faster processing speeds.

4. Case Data Processing and Classification

Data processing and classification for oil and gas field ground engineering cases is a preliminary step in constructing the case database. It aims to reduce data noise, improve data purity, and provide an accurate storage classification coordinate system for case decomposition and storage. This process simplifies the construction of the case database. The flowchart for case data processing and classification is shown in Figure 5. It involves collecting data, recognizing text from paper documents to enhance completeness, followed by text preprocessing, keyword generation, and establishing a storage classification coordinate system for efficient data processing and categorization [31].

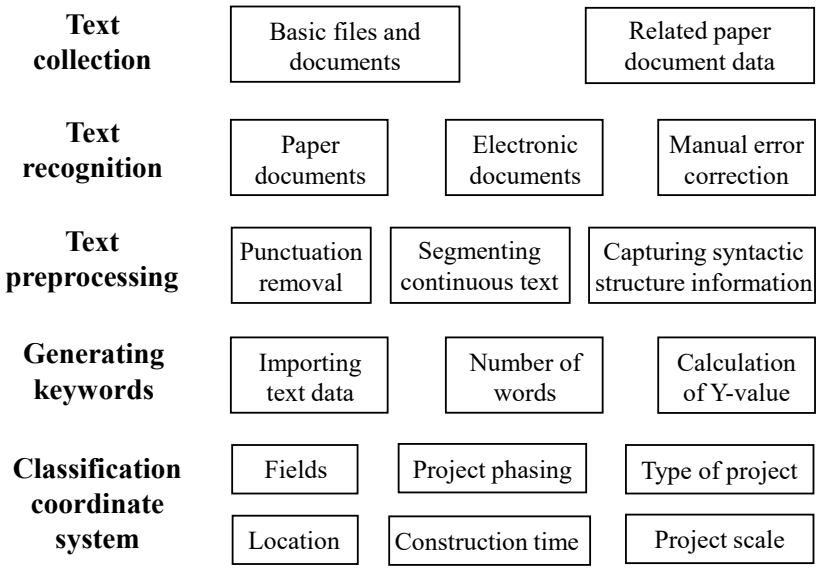

Figure 5. Flowchart of data processing and classification for engineering construction cases.

4.1. Case Data Processing

Case data processing includes the following three main components: text collection, text recognition, and text preprocessing. Initially, a vast amount of data related to surface engineering cases in oil and gas fields is collected, including basic files, documents, and

paper documents, with over 260 cases amassed in the past three years. These paper documents are converted into editable electronic formats using OCR technology, ensuring data accuracy and completeness [32]. This step involves content and structural analysis to identify the unique characteristics of each case. Finally, the text is preprocessed using NLP techniques. As shown in Figure 6, NLP transforms impure, disordered, and non-standard natural language texts into structured, manageable, and standardized text. The preprocessing steps include (1) using regular expressions to clean up punctuation in the text; (2) employing string-matching techniques for tokenization that segments continuous natural language text into semantically sound and complete sequences of words; (3) analyzing the dependency relationships between words in sentences to capture and represent syntactic structure information using a tree-like format [33]. This preprocessing results in structured, manageable, and standardized case data, providing a clear and reliable data structure for subsequent steps.

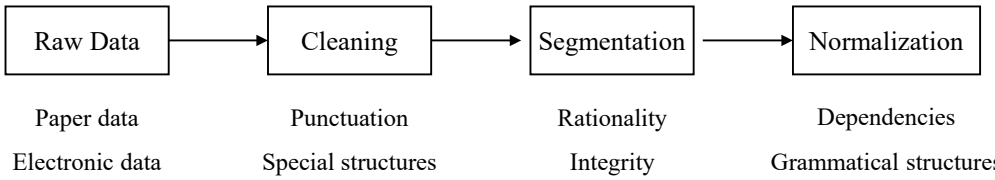

Figure 6. Text preprocessing process.

4.2. Case Data Classification

The main objective of categorizing case data is to create a storage classification coordinate system. This system aids in the convenient placement of structured data derived from case decomposition into storage. It ensures rapid and accurate storage of structured data and, during subsequent retrieval and push application scenarios, allows the coordinate system to precisely locate the required data, optimizing the user experience of retrieval and push functionalities.

Principles of case data categorization include the following: Scientific Nature: develop a stable classification system that aligns with the actual content of the project; Systematic Approach: establish uniform and coherent classification standards that progress in a logical order to ensure systematic categorization; Expandability: ensure enough space for adding new categories without disrupting existing principles and consider the expansion and refinement of subcategories; Practicality: adapt to user needs, enhancing the usability and operability of the classification [34].

The categorization process starts by dividing cases into the following two levels based on data features identified during initial processing and expert insights: individual engineering construction cases and typical construction cases. Individual engineering construction cases consist of multiple single projects, further categorized into the following three major classes totaling 33 items: special environments, key procedures, and risky operations. The specific classifications are shown in Figure 7.

(1) Special environments: This includes winter construction, rainy season construction, high-temperature construction, night-time construction, hydrogen-sulfide-rich environments, typhoon weather construction, major exhibitions, holiday period construction, and emergency construction during water or power outages.
(2) Key procedures: This encompasses processes such as assessment and procedure compliance, non-destructive testing, anti-corrosion insulation, vessel production safety, purging pressure testing, commissioning operation, pile foundation, deep excavation, and blasting.
(3) Risky operations: This category includes operations like hot work, earthmoving, circuit-breaking, working at heights, equipment maintenance, blind board plugging, pipeline and equipment opening, temporary electricity use, excavation, entry into confined spaces, lifting operations, high-voltage electricity work, climbing, refrigeration and air conditioning operations, and drilling operations.

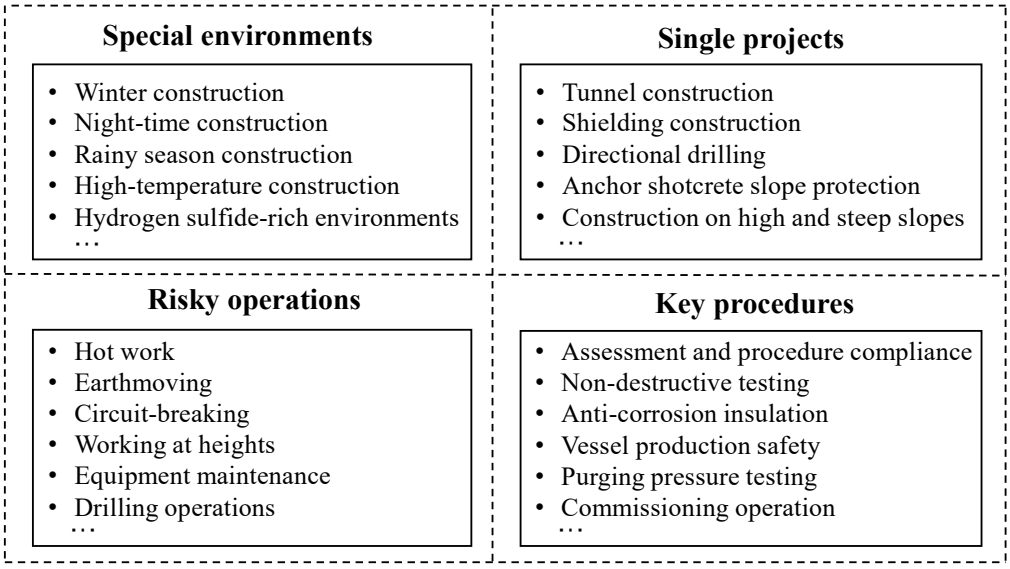

Figure 7. Schematic diagram of individual engineering case classification.

After categorizing the cases, the next step involves generating keywords from the preprocessed text data. The keyword generation process uses the TF-IDF algorithm and includes the following steps: import the preprocessed text data; count all the words that appear and the number of documents they appear in; calculate each word's TF, IDF, and Y values, and then rank them in descending order; select the highest-ranking words as keywords for output [35]. The TF-IDF algorithm's formula is as follows.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{3}$$

$$IDF_i = \log \frac{|D|}{|j : t_i \in d_j| + 1} \tag{4}$$

$$Y = TF \cdot IDF \tag{5}$$

In the formula, n_i represents the number of occurrences of term t_i in the document d_j; $n_{k,j}$ denotes the total number of terms in document d_j; D signifies the total number of documents in the case studies; and j: $t_i \in d_j$ indicates the number of documents containing term t_i. The addition of +1 to the denominator prevents division errors when the number of documents containing t_i is zero.

Finally, by choosing relevant keywords and combining them with the classification of individual engineering projects, a logical and associative storage classification coordinate system is established. This system facilitates the storage of structured data from decomposed cases. The details of the storage classification coordinate system, shown in Figure 2, are divided into the following six main categories: case field, project phase, type of engineering, project location, construction time, and scale of engineering. The details of the storage classification coordinate system, as illustrated in Figure 2, are divided into the following six main categories: the field of the case, project phase, type of engineering, project location, construction time, and the scale of the engineering.

Taking a particular oil and gas field surface engineering case as an example, keywords generated by the TF-IDF algorithm include, but are not limited to, project locations such as Central Sichuan, Northwest Sichuan, Northeast Sichuan; fields like oil fields, oil and gas fields; and types of engineering like IT projects, shale gas projects, storage projects, station, pipelines, crossing engineering, and others. It is evident that the keywords closely align with the storage classification coordinate system, thereby streamlining the process of decomposing and storing case data.

5. Case Data Decomposition and Storage

Although the processed case data are relatively standardized, it remains intricate, and some data can be categorized under multiple coordinates, making data storage challenging. Therefore, it is necessary to decompose the case data to enable data structuring, which facilitates the storage of case data. Before storing the structured case data, it is crucial to build a knowledge graph for construction cases to further optimize the storage process. The storage content primarily comprises the decomposed structured fields, resulting in a comprehensive, well-categorized, and structurally complete database of construction cases that supports rapid retrieval and intelligent push of applications [36].

5.1. Case Data Decomposition

The mechanism for decomposing construction cases involves "nine major items and two levels", where the nine items include case overview, key parameters, operation time, construction drawings, construction resources, construction measures, HSE measures, process record data, and images/videos. It is important to note that images and videos, being unstructured data, do not require decomposition and can be directly categorized for storage.

The decomposition process involves a detailed breakdown of project content based on the nine items, resulting in structured data fields and the creation of structured forms, which prepares for subsequent case data storage. The two levels include first-level directory decomposition and second-level directory decomposition. At the first level, basic attributes like case overview, key parameters, and operation time are addressed. The second level includes detailed information such as construction drawings, resources, measures, HSE measures, process data, images, and videos. For specific construction cases, basic project information includes an overview, key parameters, operation time, etc. Guided by the nine items, the case is decomposed, resulting in structured data. These data form the basis for building the knowledge graph of individual engineering cases, which is linked to the source files and presented as key structured data associated with the case knowledge graph. This supports case content retrieval and positioning, intelligent push, and smart matching of typical cases.

5.2. Case Data Storage

To systematically and coherently store the structured data obtained after decomposition into the case database, it is crucial to establish a case storage mechanism. This mechanism includes two key components, a storage classification coordinate system and a knowledge graph. The storage classification coordinate system provides precise coordinates for storing structured fields and supports rapid, accurate data location in subsequent push and retrieval scenarios. This enhances the user experience for push and retrieval functions. It includes coordinates for case project location, engineering type, project scale, domain, project phase, and construction time.

After establishing the storage classification coordinate system and the knowledge graph, the next step is to leverage the knowledge graph as the foundation. The primary storage content is structured fields, tagged with coordinates from the storage classification system. Correlating storage content with respective tags efficiently completes the task of storing engineering construction cases. This process ensures each piece of information is accurately placed within the database, facilitating ease of access and utility in various applications, particularly in enhancing decision-making processes for oil and gas field surface engineering projects.

6. Intelligent Retrieval and Push

Given the outdated core of the existing oil and gas field surface engineering case retrieval systems, which are characterized by low accuracy, limited search scope, and require significant time for validation and error correction, there is a clear need for advanced systems that can meet the current demands for rapid response, robustness, and accurate

results in oil and gas fields. Consequently, an intelligent retrieval and push system for oil and gas field surface engineering cases has been developed, enhancing traditional case retrieval systems. This new system is divided into two main aspects, intelligent retrieval and intelligent push, comprising the following six components: the case database, input module for retrieval information, case retrieval module, display module for retrieval results, retrieval log generation module, and intelligent case push module. The process flow of intelligent retrieval and push for engineering construction cases is illustrated in Figure 8.

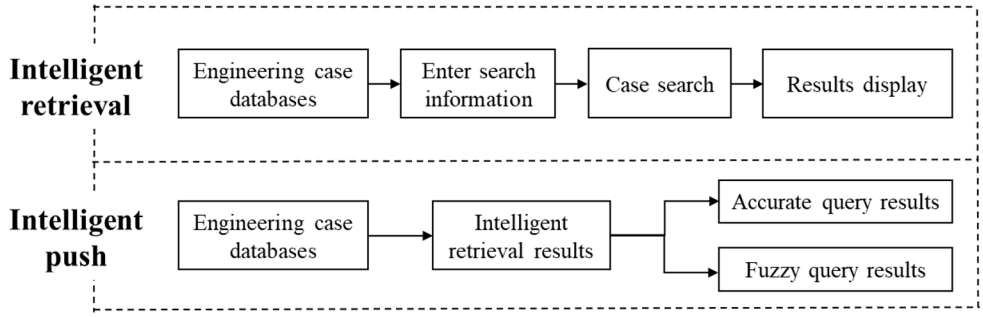

Figure 8. Flowchart of intelligent retrieval and push process for engineering construction cases.

6.1. Intelligent Retrieval

Intelligent retrieval, based on the oil and gas field surface engineering case database, selects appropriate retrieval methods according to different database structures, data volumes, and system requirements. The goal is to rapidly respond to query requests, form a unique query matching process, optimize the model based on query results, and improve retrieval speed and reliability. Intelligent retrieval methods can be classified into fuzzy queries using keywords and precise queries using exact information like case names. The query subjects for case retrieval are based on the nine major disassembly categories, divided into seven types, including searches for engineering construction case projects, construction drawings, construction measures, HSE measures, construction resources, process record data, and images.

The primary focus of intelligent retrieval is keyword searches within the knowledge graph. During case retrieval, NLP technology analyzes the input information and understands the user's intent, reducing text ambiguity and improving retrieval accuracy and speed. If the input query exists in the database, it can be quickly and accurately retrieved; if the input is vague and not present, knowledge graph-based matching calculates and matches cases according to structural similarity, ranking them by relevance. The system also supports historical data queries and search log generation.

6.2. Intelligent Push

Intelligent push, guided by the case database knowledge graph, provides project information and related content based on fuzzy keyword queries and precise information queries. It also supports tracing related projects from a single retrieval result. There are four main components of case push, which are as follows:

(1) Project case association push

Based on the project case knowledge graph, it recommends similar project cases to the search object by tracing upwards in the graph. The ranking is based on the number of matching dimensions.

(2) User habit association push

A user habit knowledge graph is built, recommending engineering projects most frequently viewed or focused on by the user, based on their history.

(3) Case file association push

In the detailed view pages for construction drawings, measures, HSE measures, resources, process records, and images, files with similar attributes are pushed, ranked by the number of matching dimensions.

(4) High-frequency search standard association push

Based on the frequency of projects searched by all users of the system, high-frequency projects are recommended on the homepage.

The intelligent retrieval and push system represents a significant advancement over traditional retrieval systems. By using knowledge graphs and sophisticated matching algorithms, it can accurately and efficiently handle large-scale, diverse oil and gas field surface engineering cases. This system provides precise and rapid retrieval capabilities, enhances the overall user experience by offering more relevant and targeted results, and facilitates better decision-making and management in oil and gas surface engineering.

7. Applications

By using the classification mechanism for construction engineering cases, database creation methods, data structuring, intelligent retrieval and matching, and intelligent push techniques proposed in the article, we have successfully constructed a highly shared engineering construction case database. This was achieved by integrating the collection and consolidation of engineering cases with NLP, OCR intelligent recognition, and knowledge graph technologies. We have also developed search and push functionalities for the case database, achieving intelligent construction of the oil and gas field surface engineering case library. The database can intelligently match and retrieve relevant cases, displaying them with key structured data and case knowledge graphs. This enhances the searchability and pinpointing of case content, assists in intelligent matching of typical cases, and has led to the development of an intelligent platform for oil and gas field surface engineering construction cases shown in Figures 9 and 10.

Using the keyword “shale gas” as an example, the platform’s functions are demonstrated as follows: (1) Case Retrieval: by entering the keyword “shale gas”, the platform lists individual engineering case information related to shale gas. The left side displays categories derived from nine disassembly projects, providing preview and download options. The right side presents related individual engineering projects in a list format, ranked by relevance. (2) Intelligent Search: the system precisely locates target documents using the keyword “shale gas”. Users can switch file types on the left side and query or trace-related engineering cases on the right side. (3) User-Based Collaborative Recommendation: based on user characteristics, including their department, position, frequently queried engineering project types, and bookmarked project types, the platform makes default recommendations on the engineering project page. The recommendations are sorted by matching degree (calculated by recommendation algorithms) from high to low and by publication date from recent to older, displayed as an engineering project list on the user’s homepage. (4) Content-Based Collaborative Recommendation: based on the characteristics of engineering cases reviewed by the user, including project location, keywords, oil and gas field types, engineering types, and special environments, the platform recommends relevant cases during the user’s query. The recommendations are sorted by matching degree (calculated by recommendation algorithms) from high to low and by publication date from recent to older, displayed as an engineering project list on the user’s homepage.

Case Maintenance and Management include the following: (1) Managing Engineering Types: on the left side of the maintenance page, users can add and delete engineering types. (2) Adding New Projects: users can add new individual engineering projects and maintain and enhance related information. (3) Improving Case Details: users can add new individual engineering projects and choose to maintain and improve related information and key measures.

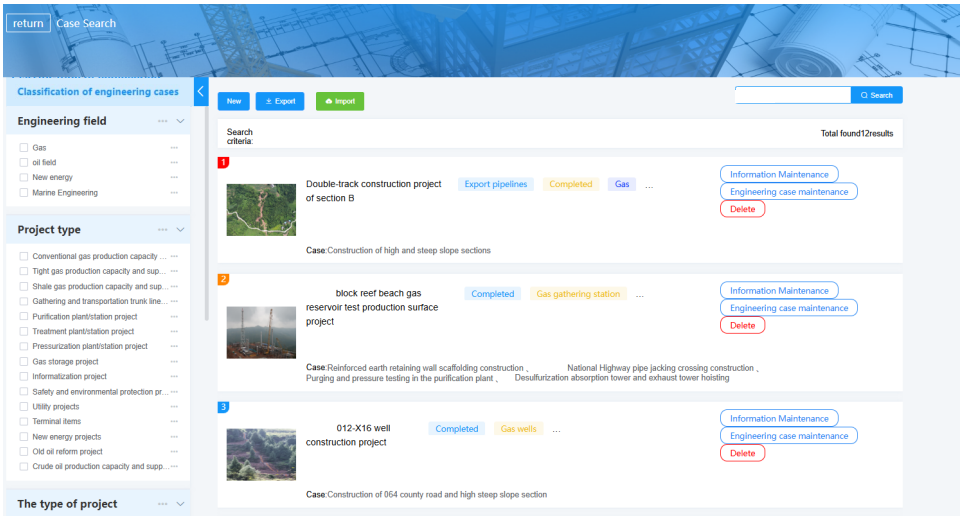

Figure 9. Illustration of the single engineering project retrieval results on the intelligent platform.

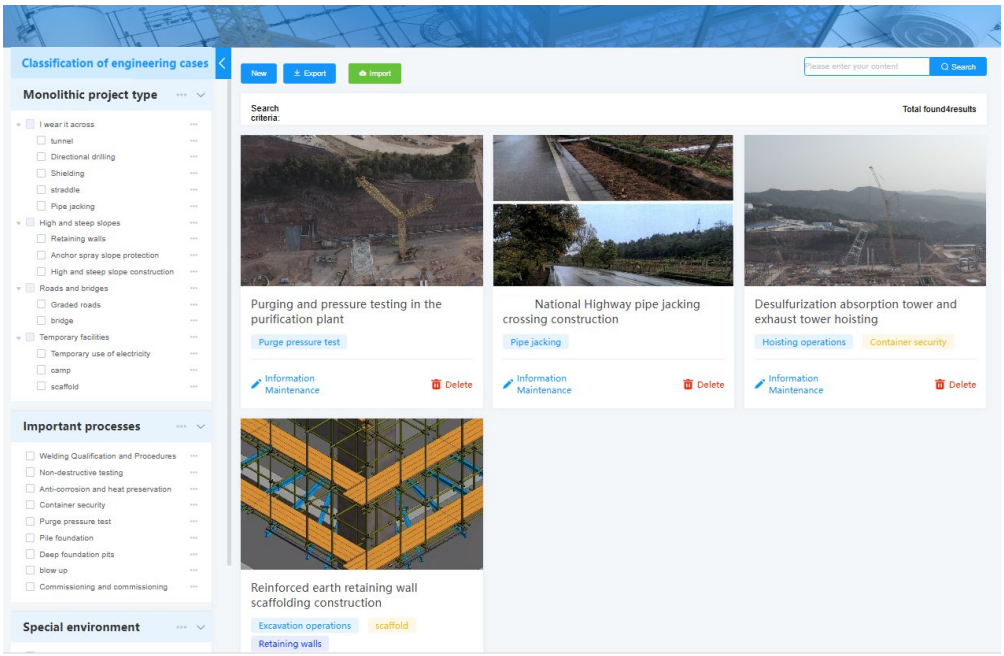

Figure 10. Illustration of the individual engineering project retrieval.

8. Conclusions

(1) In response to the need for efficient management of oil and gas field surface engineering projects, this paper proposes a case database construction method using OCR intelligent recognition, knowledge graph, and NLP technologies. It includes the study of engineering case data processing technology, knowledge classification, database construction methods, data structuring techniques, and intelligent retrieval and matching push technology. This approach generates the logic for engineering case inclusion, achieving the full process of intelligent construction of engineering project case databases.

(2) Theoretically, this method resolves the difficulty of swiftly and efficiently finding relevant management projects in engineering construction. By integrating knowledge graph technology with search engines and combining it with knowledge graph matching technology, users and managers can quickly and accurately locate target case files from a vast database. This results in a complete, feasible, and expandable set of methods for disassembling, storing, intelligently retrieving, and pushing construction

cases, supporting the construction of an intelligent management platform for surface engineering.

(3) Research on the construction case database method lays a theoretical foundation for developing an intelligent platform for oil and gas fields. It refines digital project management functions of oil and gas field surface construction engineering and provides a clear direction for constructing intelligent platforms. This enables continued advancement in intelligent assistive decision-making functionalities, leading to the creation of a general intelligent management platform for oil and gas field surface engineering. This meets the demands for digital management, digital delivery, and intelligent assistive decision-making in oil and gas fields.

Author Contributions: Validation, D.Z.; formal analysis, W.Z.; investigation, F.W.; resources, Y.Z.; data curation, Z.D.; writing—original draft, T.X.; writing—review and editing, L.X. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported and funded by the Natural Gas Gathering and Transmission Engineering Technology Research Institute, PetroChina Southwest Oil and Gas Field Company (Chengdu, China). This research came from the project of Oil and Gas Field Surface Engineering Construction Technical Standards and Engineering Case Storage Mechanism and Theoretical Algorithm Research.

Data Availability Statement: The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Conflicts of Interest: Authors Taiwu Xia, Zhixiang Dai, Yihua Zhang, Feng Wang, Wei Zhang and Li Xu were employed by the PetroChina Southwest Oil and Gas Field Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The PetroChina Southwest Oil and Gas Field Company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Nomenclature

Abbreviations	
OCR	Optical Character Recognition
NLP	Natural Language Processing
GIS	Geographic Information System
IPCC	Intergovernmental Panel on Climate Change
TF-IDF	Term Frequency-Inverse Document Frequency
LSTM	Long Short-Term Memory
BIO	Beginning, Inside, Outside
GMO	Graph Matching for Ontologies
HSE	Health, Safety and Environment
Symbol	
D	The set of all character subsequences
d_j	Document
$n_{i,j}$	The number of occurrences of term t_i in the document d_j
$n_{k,j}$	The total number of terms in document d_j
$c_1, c_2, \ldots, c_m$	Character sequence
c_i, c_j	Two classes in the case database
c_{ij}	Common parent class of c_i and c_j
c_j^c	Unit vector
h_j^c	Hidden vector
f_j^c	Forget gate
i_j^c	Input gate
o_j^c	Output gate

$w_{b,e}^{d}$	Represents a subsequence starting with character index b and ending with character index e
W1, W2, W3	Weights
Sim	Similarity of target template
$\mathbf{W}^{c\top}$ and $\mathbf{b}^{c}$	Model parameters
σ	Sigmoid function
$depth(c_i)$, $depth(c_j)$	The depths of c_i and c_j in the original inheritance relationship, respectively
j: $t_i \in d_j$	The number of documents containing term t_i

References

1. Cai, S.; Yao, W.; Gao, Y.; Yang, Z.; Xue, Y.; Li, L. The construction method of oil and gas sources data management platform based on spatial information. *Fresenius Environ. Environ. Bull.* **2022**, *31*, 593–599.
2. Sánchez-del Rey, A.; Gil-García, I.C.; García-Cascales, M.S.; Molina-García, Á. Online Wind-Atlas Databases and GIS Tool Integration for Wind Resource Assessment: A Spanish Case Study. *Energies* **2022**, *15*, 852. [CrossRef]
3. Zhu, W.; Bi, J.; Wang, X.; Zhu, Z.; Pang, W. The Evaluation System Design of GIS-Based Oil and Gas Resources Carbon Emission Database Management. In *IOP Conference Series: Earth and Environmental Science*; IOP Publishing: Bristol, UK, 2014; Volume 17, p. 012032.
4. Wen, Z.; Wang, J.; Wang, Z.; He, Z.; Song, C.; Liu, X.; Zhang, N.; Ji, T. Analysis of the world deepwater oil and gas exploration situation. *Pet. Explor. Dev.* **2023**, *50*, 1060–1076. [CrossRef]
5. Jasansky, S.; Lieber, M.; Giljum, S.; Maus, V. An open database on global coal and metal mine production. *Sci. Data* **2023**, *10*, 52. [CrossRef] [PubMed]
6. Guan, Q.; Zhang, F.; Zhang, E. Application prospect of knowledge graph technology in knowledge management of oil and gas exploration and development. In Proceedings of the 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 25–28 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 161–166.
7. Su, J.; Yao, S.; Liu, H. Data Governance Facilitate Digital Transformation of Oil and Gas Industry. *Front. Earth Sci.* **2022**, *10*, 861091. [CrossRef]
8. Maroufkhani, P.; Desouza, K.C.; Perrons, R.K.; Iranmanesh, M. Digital transformation in the resource and energy sectors: A systematic review. *Resour. Policy* **2022**, *76*, 102622. [CrossRef]
9. Wu, L.; Li, Z.; AbouRizk, S. Automating Common Data Integration for Improved Data-Driven Decision-Support System in Industrial Construction. *J. Comput. Civ. Eng.* **2022**, *36*, 04021037. [CrossRef]
10. Huang, S.; Wang, Y.; Yu, X. Design and Implementation of Oil and Gas Information on Intelligent Search Engine Based on Knowledge Graph. In *Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2020; Volume 1621, p. 012010.
11. Tang, X.; Feng, Z.; Xiao, Y.; Wang, M.; Ye, T.; Zhou, Y.; Meng, J.; Zhang, B.; Zhang, D. Construction and application of an ontology-based domain-specific knowledge graph for petroleum exploration and development. *Geosci. Front.* **2023**, *14*, 101426. [CrossRef]
12. Yuan, J.; Li, H. Research on the standardization model of data semantics in the knowledge graph construction of Oil & Gas industry. *Comput. Stand. Interfaces* **2023**, *84*, 103705.
13. Zhou, X.G.; Gong, R.B.; Shi, F.G.; Wang, Z.F. PetroKG: Construction and application of knowledge graph in upstream area of PetroChina. *J. Comput. Sci. Technol.* **2020**, *35*, 368–378. [CrossRef]
14. Bai, Y.; Wu, J.; Ren, Q.; Jiang, Y.; Cai, J. A BN-based risk assessment model of natural gas pipelines integrating knowledge graph and DEMATEL. *Process Saf. Environ. Prot.* **2023**, *171*, 640–654. [CrossRef]
15. Pei, Y.; Chai, S.; Li, X.; Samuel, J.C.; Ma, C.; Chen, H.; Lou, R.; Gao, Y. Construction and Application of a Knowledge Graph for Gold Deposits in the Jiapigou Gold Metallogenic Belt, Jilin Province, China. *Minerals* **2022**, *12*, 1173. [CrossRef]
16. Patil, R.; Boit, S.; Gudivada, V.; Nandigam, J. A survey of text representation and embedding techniques in nlp. *IEEE Access* **2023**, *11*, 36120–36146. [CrossRef]
17. Wu, K. Research Progress of Intelligent Image Recognition. In Proceedings of the 2017 2nd International Conference on Machinery, Electronics and Control Simulation (MECS 2017), Taiyuan, China, 24–25 June 2017; Atlantis Press: Amsterdam, The Netherlands, 2016; pp. 91–95.
18. Rizvi, M.; Raza, H.; Tahzeeb, S.; Jaffry, S. Optical character recognition based intelligent database management system for examination process control. In Proceedings of the 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 8–12 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 500–507.
19. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Philip, S.Y. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 494–514. [CrossRef] [PubMed]
20. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [CrossRef] [PubMed]
21. Lauriola, I.; Lavelli, A.; Aiolli, F. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing* **2022**, *470*, 443–456. [CrossRef]

22. Omran, P.G.; Wang, K.; Wang, Z. An Embedding-based Approach to Rule Learning in Knowledge Graphs. *IEEE Trans. Knowl. Data Eng.* **2019**, *33*, 1. [CrossRef]
23. Buchgeher, G.; Gabauer, D.; Martinez-Gil, J.; Ehrlinger, L. Knowledge Graphs in Manufacturing and Production: A Systematic Literature Review. *IEEE Access* **2021**, *9*, 55537–55554. [CrossRef]
24. Zhuohao, W.; Dong, W.; Qing, L.I. Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF. *Chin. J. Electron.* **2021**, *30*, 652–657. [CrossRef]
25. Al-Tawil, M.; Dimitrova, V.; Thakker, D.; Abu-Salih, B. Emerging Exploration Strategies of Knowledge Graphs. *IEEE Access* **2023**, *11*, 94713–94731. [CrossRef]
26. Wang, Y.; Goridkov, N.; Rao, V.; Cui, D.; Grandi, D.; Goucher-Lambert, K. Embedding experiential design knowledge in interactive knowledge graphs. *J. Mech. Des.* **2023**, *145*, 041412. [CrossRef]
27. Oh, D.; Lim, J.; Lim, H. Neuro-Symbolic Word Embedding Using Textual and Knowledge Graph Information. *Appl. Sci.* **2022**, *12*, 9424. [CrossRef]
28. Guo, Q.; Zhuang, F.; Qin, C.; Zhu, H.; Xie, X.; Xiong, H.; He, Q. A survey on knowledge graph-based recommender systems. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 3549–3568. [CrossRef]
29. Millar, D.; Braines, D.; D'Arcy, L.; Barclay, I.; Summers-Stay, D.; Cripps, P. Embedding dynamic knowledge graphs based on observational ontologies in semantic vector spaces. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, Online, 12–16 April 2021; SPIE: Bellingham, WA, USA, 2021; Volume 11746, pp. 404–413.
30. Zhu, Y.; Wang, G.; Karlsson, B.F. CAN-NER: Convolutional attention network for Chinese named entity recognition. *arXiv* **2019**, arXiv:1904.02141.
31. Nahar, K.M.O.; Alsmadi, I.; Al Mamlook, R.E.; Nasayreh, A.; Gharaibeh, H.; Almuflih, A.S.; Alasim, F. Recognition of Arabic Air-Written Letters: Machine Learning, Convolutional Neural Networks, and Optical Character Recognition (OCR) Techniques. *Sensors* **2023**, *23*, 9475. [CrossRef] [PubMed]
32. Yalniz, I.Z.; Manmatha, R. A fast alignment scheme for automatic ocr evaluation of books. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 754–758.
33. Liu, Y.; Yang, S.; Xu, Y.; Miao, C.; Wu, M.; Zhang, J. Contextualized graph attention network for recommendation with item knowledge graph. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 181–195. [CrossRef]
34. Shamshiri, A.; Ryu, K.R.; Park, J.Y. Text mining and natural language processing in construction. *Autom. Constr.* **2024**, *158*, 105200. [CrossRef]
35. Paulheim, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semant. Web* **2017**, *8*, 489–508. [CrossRef]
36. Sabir, E.; Rawls, S.; Natarajan, P. Implicit language model in lstm for ocr. In Proceedings of the 2017 14th IAPR international conference on document analysis and recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; IEEE: Piscataway, NJ, USA, 2017; Volume 7, pp. 27–31.