processes

MDPI

Article

Establishment of Technical Standard Database for Surface Engineering Construction of Oil and Gas Field

Taiwu Xia [1], Zhixiang Dai [1], Zhan Huang [2], Li Liu [2], Ming Luo [2], Feng Wang [1], Wei Zhang [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; xiatw@petrolchina.com (T.X.); daizhixiang@petrochina.com.cn (Z.D.); w_feng@petrochina.com.cn (F.W.); zhang_wei@petrochina.com.cn (W.Z.)
[2] Infrastructure Construction Engineering Department, PetroChina Southwest Oil and Gas Field Company, Chengdu 610066, China; liliu@petrochina.com.cn (L.L.); luom@petrochina.com.cn (M.L.)
[3] School of Intelligent Manufacturing, Panzhihua College, Panzhihua 617000, China; zhoudan@pzhu.edu.cn
[4] Petroleum Engineering School, Southwest Petroleum University, Chengdu 610500, China
* Correspondence: 201599010096@swpu.edu.cn

Abstract: In recent years, oil and gas field surface engineering construction projects tend to be large in scale, large in quantity, and short in cycle. The task of surface construction management has increased significantly. In the process of project construction, corresponding standards and specifications are required to provide sufficient technical guidance and support for design, construction, and management personnel to ensure project management and control towards compliance, safety, and quality. However, the oil and gas field engineering standards are numerous and specialized, involving different levels of national standards, enterprise standards, and industry standards, which leads to the inefficiency of the actual use of standards and specifications. To solve them, this paper uses knowledge graph technology, OCR recognition, and natural language processing technology to conduct systematic research on the knowledge classification mechanism, data extraction, database construction mechanism, data structuring, and intelligent retrieval matching of oil-gas field surface engineering construction standards. In this study, the structured identification, storage, and information warehousing of standards are realized, and a highly sharable library of standards and specifications is formed, which realizes the intelligent retrieval and pushing of technical standards for surface engineering construction. This paper creates conditions for the realization of intelligent push and benchmarking management of standards and specifications, providing support for digital transformation and intelligent development of oil–gas fields.

Keywords: standard database; disassemble; knowledge graph; intelligent retrieval; intelligent push

Citation: Xia, T.; Dai, Z.; Huang, Z.; Liu, L.; Luo, M.; Wang, F.; Zhang, W.; Zhou, D.; Zhou, J. Establishment of Technical Standard Database for Surface Engineering Construction of Oil and Gas Field. *Processes* **2023**, *11*, 2831. https://doi.org/10.3390/pr11102831

Academic Editors: Wenyang Shi and Yang Wang

Received: 29 July 2023
Revised: 18 September 2023
Accepted: 19 September 2023
Published: 26 September 2023

1. Introduction

The surface engineering construction projects of oil and gas fields in recent years have shown the trend of large scale, large numbers, and short construction periods, and the task of ground engineering construction management has increased obviously. The use of information technology to achieve efficient management of project construction is imminent. In the context of the rapidly developing digital economy, enterprises need to carry out digital transformation and upgrading to adapt to digital transformation and intelligent development [1–3]. To realize effective management and efficient development, intelligent technology must be applied to oil and gas field construction to achieve professional development and trans-era development. In the process of digital construction of oil and gas field surface engineering, implementing standardization is very important for the construction of smart oil and gas fields. With the acceleration of oil and gas field development and construction, the surface engineering construction standard, as a technical basis

in the process of oil and gas field surface construction, can standardize the engineering construction and technology to ensure the safety and quality of the project, and to promote the technical development.

Oil and gas field engineering standards are numerous and wide-ranging, involving different levels of national standards, enterprise standards, and industry standards. The content of the standards between the various professions refer to each other, and there are even references not timely and contradictory content. It is difficult for technical personnel to find and master standards, and the standards are not updated in time, which increases the management difficulty of oil and gas field surface engineering. The review of applicable standards by the traditional mode to determine which standards or clauses to use not only requires a high level of competence of technicians themselves but also cannot improve efficiency. Therefore, resource sharing and information construction of engineering standards and specifications (In the following, standards or specifications are collectively referred to as standards) are particularly important. The use of intelligent technology to organize the standard can reduce the repetitive work of the oil and gas field surface engineering, improve the standardization of the construction of the project, and strengthen the efficiency of the technicians.

The database of domestic and foreign petroleum industry standards is investigated. Applications such as the safety barrier database of petrochemical plants [4] and the failure database of oil and gas pipelines have been designed and developed [5] in the petroleum industry. However, there are few standard databases developed in China. Therefore, combining the above background, this paper proposes to establish a database of standards for oil and gas field surface engineering construction. Knowledge graph technology has been used in many fields. For example, the Industry 4.0 knowledge graph [6], TCM health care knowledge graph [7], earth science data knowledge graph [8], large-scale engineering patent knowledge graph [9], measurement remote sensing application knowledge graph [10], and so on. In the field of oil and natural gas, Ruishan Du et al. [11] introduced ontology into knowledge construction in the field of oil development, designed an ontology integration platform and realized knowledge representation and knowledge sharing. Xianming Tang et al. [12] proposed a domain knowledge graph engineering construction method on the basis of traditional natural language processing based on the ontology of petroleum exploration and development to provide better knowledge services for the oil and gas industry. Jike Ge et al. [13] proposed a knowledge integration and sharing system framework based on petroleum exploration domain ontology, which minimizes the complexity of heterogeneous data and enhances the knowledge integration and information sharing capabilities among different operating units. Qing Guan et al. [14] extracted knowledge from heterogeneous knowledge carriers from multiple sources and built a graphical knowledge base to improve work efficiency in the process of oil and gas exploration and development. Knowledge graph technology [15] can provide scientific methods and powerful data for oil and gas field surface engineering construction, realize resource sharing and information integration, improve work efficiency, provide professional information retrieval for oil and gas field surface engineering construction, and improve economic benefits. In terms of data retrieval, Faming Gong et al. [16] proposed a domain ontology construction process based on the Neo4j graph database and a retrieval method based on a two-layer index architecture. Qi Zhou et al. [17] explored a new method of using keywords for semantic web queries. This method automatically transforms keyword queries into formal logic queries so that end users can conduct semantic searches by using familiar keywords. Ganggao Zhu et al. [18] proposed a knowledge graph entity search matching framework that combines natural language query processing, entity linking, entity type linking, and semantic similarity-based query extension. Shujun Huang [19] constructed a knowledge graph of various topics based on a large amount of specialized oil and gas information and proposed an intelligent search engine for oil and gas information based on the knowledge graph, which can better understand the user's search intent. Wang Ce [20] proposed a search result ranking model based on knowledge graph. Therefore, this paper

incorporates knowledge graph techniques to retrieve and push the standard database to realize the application level of the database.

In summary, there are a limited number of scholarly articles available on the standard database of the petroleum industry, both domestically and internationally. Through reviewing the information, it is understood that there is a lack of universal collection tools for oil and gas field engineering standards in the market, the application of standard database is not wide enough, and the actual use of standards is inefficient. Furthermore, the majority of oil and gas field engineering standards are retrieved using conventional search methods. Because of this, this paper analyzes the surface engineering standards, collects empirical data, and adopts knowledge mapping technology, OCR recognition, and natural language processing technology to investigate the knowledge classification mechanism, data extraction, database construction, data structuring, and intelligent retrieval and matching of standards in the process of oil and gas field surface engineering construction. It realizes structured disassembly, real-time query, matching push, and systematic management of standards, provides the required application services for engineering projects, and improves the efficiency of the utilization of standards in engineering construction.

2. Research Framework

Aiming at the database construction mechanism of standards for oil and gas field surface engineering construction, this paper divides the research mechanism into six modules: identification and audit, disassembly, storage, intelligent retrieval, intelligent push, and knowledge graph technology. Among them, the intelligent retrieval module and intelligent push module are the application fields of knowledge graph technology. The module framework is shown in Figure 1. The intelligent construction of oil and gas field ground engineering is the process of applying advanced information technology and automation technology in the operation and management of oil and gas field ground facilities so as to realize the process of improving production efficiency, reducing costs, and improving safety and environmental protection. Therefore, the development of the standard database enables standardized application in multiple scenarios, such as design, review, and acceptance during the construction of oil and gas field ground engineering projects, which improves the standardization level and efficiency of project construction.

(1) Identification and audit module: collecting and analyzing oil and gas surface engineering standard information, collecting scanned PDF files through OCR intelligent recognition technology, using text recognition OCR engine to scan the content of the document, converting it into a structured document that can be pulled for recognition and data extraction (editable PDF or WORD file), then reviewing and correcting it manually.

(2) Disassemble module: The classification and grading of standards is the basis for realizing structured identification and storage. Research on the classification mechanism is carried out according to the basic attributes and keywords of standards; the classification includes manual classification and intelligent classification. And the grading includes the first-level title, second-level title, and third-level title. Then, the structured documents are manually disassembled, system-assisted disassembled, or intelligently disassembled, following the basic attributes of standards, clauses of standards, and keywords to form structured data. The standard keyword database formed after disassembly is used as the basic database of intelligent retrieval and matching.

(3) Storage module: Establish a database of standards to realize the storage of structured standard forms.

(4) Intelligent retrieval module: Based on knowledge graph technology, intelligent retrieval under different scenarios is realized, including two retrieval scenarios, three user-oriented retrieval methods, two retrieval mechanisms of the system, and the formation of different standard lists and clause lists. Natural language processing technology (NLP) is used to understand user intentions, reduce ambiguity, and improve retrieval matching degrees.

(5) Intelligent push module: Based on the knowledge graph, intelligent push can be carried out from three dimensions of keyword matching degree, release time, and terms requirements, and can be applied to different application scenarios.
(6) Knowledge graph technology module: The basic construction process of a knowledge graph includes data acquisition, knowledge extraction, knowledge fusion, knowledge processing, and Knowledge update. And we apply knowledge graph technology to intelligent retrieval and intelligent push of standards.

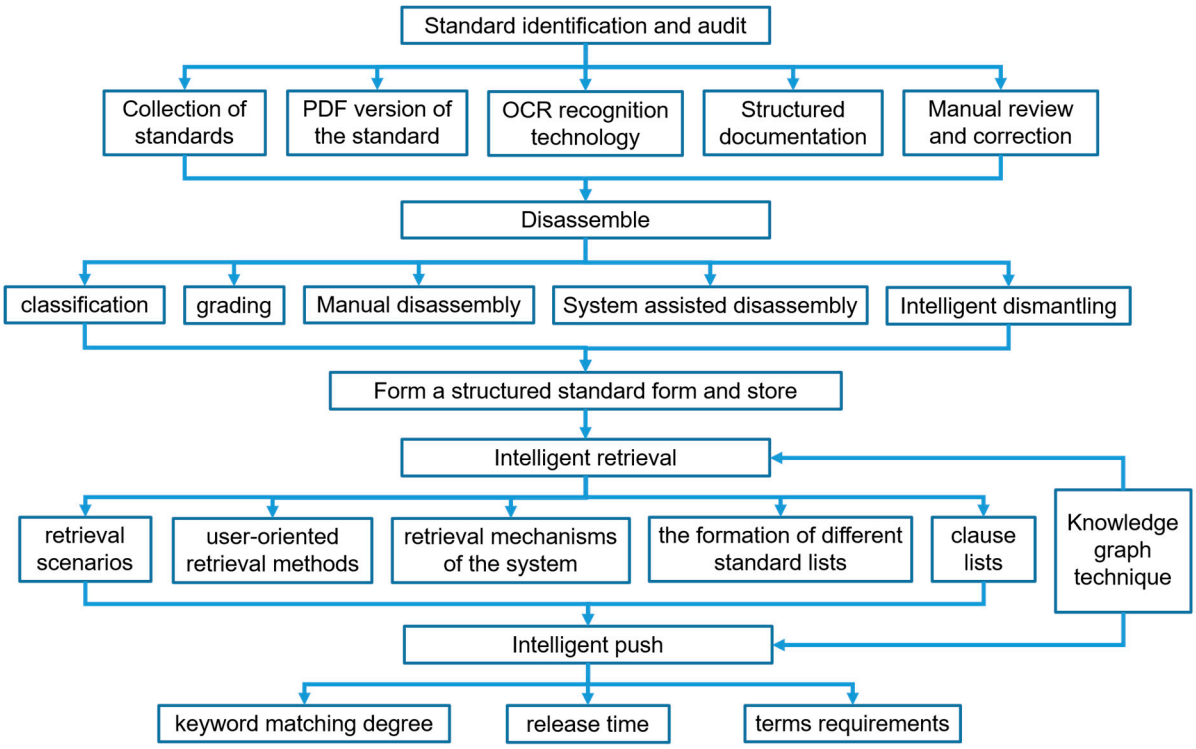

Figure 1. Research framework of standard database construction mechanism.

3. Basic Theory of Knowledge Graph

The concept of a knowledge graph was published by Google in May 2012 [21]. It is a data representation form that uses graph structure to model things and the relationships between them. In general, knowledge graphs and knowledge bases are the same concept and can be used interchangeably [22]. Knowledge graphs have recently been widely used in various artificial intelligence systems. At present, automatic knowledge acquisition [23,24], knowledge reasoning [25], knowledge representation [26], and knowledge fusion [27] related to knowledge graph have become powerful assets for search question answering [28], big data analysis [24], intelligent recommendation, and data integration [27]. The knowledge graph is mainly used to extract entities, relationships between entities, and attributes of entities from different types of complex data and then integrate the three through the data model or topology of the graph structure to display the abstract and scattered knowledge clearly and visually. The knowledge graph is formally defined as $G = \{E, R, F\}$, where E represents the set of entities, R represents the set of relations, and F represents the set of facts. Each fact element in F is represented by triples (h, r, t), where h and t represent the head entity and the tail entity, respectively, and r represents the relationship between the head entity h and the tail entity t, which is logically divided into data layer and pattern layer (core), and the construction methods are divided into top-down construction and bottom-up construction. The knowledge graph construction framework of standards is shown in Figure 2.

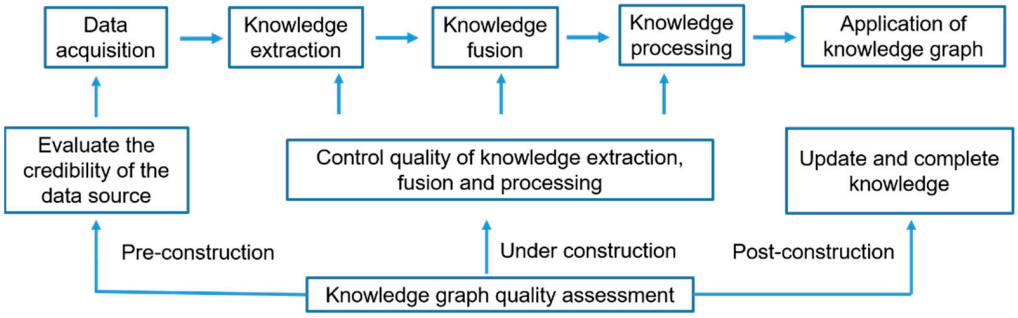

Figure 2. Knowledge graph construction framework of standards.

3.1. Data Acquisition

Data is the source of knowledge graph construction. According to the different structures, the data acquisition of the knowledge graph can be divided into three categories: structured data, semi-structured data, and unstructured data. Structured data are the data in the existing third-party knowledge base. Semi-structured data are webpage data. Unstructured data are standard PDF files, pictures, and audio files. Different extraction methods are used for different data types. In this study, the standards in the field of oil and gas field surface engineering construction are downloaded from the website, converted into Word documents with software tools, extracted with text extraction tools, and then data cleaning is carried out. The online labeling platform label-studio is adopted to carry out entity marking for standard names, standard numbers, terms, reference files, keywords, clauses, and other contents in the text, and relational marking for reference, modification, replacement, association, inclusion, and so on. The tool then returns a JSON file. We process the JSON file to obtain the annotated corpus in BIEOS format.

3.2. Knowledge Extraction

Knowledge extraction is the automatic or semi-automatic process of standard knowledge graph construction to obtain entities, relationships, attributes, and other available knowledge units from original data. It is a key hub connecting data acquisition and knowledge fusion and also a core step of knowledge graph construction. It can be divided into entity extraction, relation extraction, and attribute extraction.

(1) Entity extraction, also known as entity recognition (NER), aims to extract entity elements from the text and form knowledge (structured data) into the knowledge graph. The methods of entity extraction can be divided into rule-based entity extraction, supervised entity extraction, and unsupervised entity extraction. Alternative models include the hidden markov model (HMM), conditional random field (CRF) model, and neural network model. Commonly used methods include rule matching, machine learning, and deep learning. In this study, we employ a convolutional neural network (CNN) algorithm for entity recognition tasks using a dataset containing 300 standards. We divide the labeled data set into a training set, a validation set, and a test set at a ratio of 8:1:1. A multi-level convolutional neural network model with an embedding dimension set to 200, a dropout rate of 0.3, and batch size of 32 is constructed. The core feature of the model is a three-layer convolution operation, which uses convolution kernel sizes of 3, 4, and 5, respectively, to capture local features at different scales in the text. Each convolution is followed by a maximum pooling layer to preserve the most significant feature information. In addition, we set the initial learning rate to 0.001 and adopted the learning rate attenuation strategy to ensure smooth convergence of the model during training. To reduce the risk of overfitting, we introduce dropout regularization techniques and use ReLU activation functions to enhance the nonlinear capabilities of the model. During the training process, we use the cross entropy loss function to measure the prediction error of the model and use the stochastic gradient descent (SGD) optimizer to adjust the model parameters. The key parameters can be

tested and adjusted many times to achieve the best performance. We use F1 scores, recall rates, and accuracy rates to evaluate the validity of the model. After model training, we achieved results of an F1 score of 0.6239, a recall rate of 0.5574, and an accuracy rate of 0.9729 on the training set. On the test set, we achieved results with an F1 score of 0.5849, a recall rate of 0.5137, and an accuracy rate of 0.9644. Finally, the extracted entities are mapped to the nodes of the knowledge graph.

(2) Relation extraction (RE) is the core content of knowledge extraction. By obtaining a certain semantic relation or category of relation between entities, the entity pair is automatically identified, and the triplet formed by the relation between this pair of entities is connected. Recent studies on RE are mainly based on neural network methods, including CNN, recurrent neural network (RNN), attention mechanism (ATT), graph convolutional network (GCN), adversarial training (AT), reinforcement learning (RL), and entity-relationship joint extraction (JERE). In this study, convolutional neural networks are also used to extract the relationship between text content. The attention layer is added to improve the CNN model and improve the performance of relation extraction. Set the number of convolutional nuclei to 6, the embedding dimension to 300, the learning rate to 0.0005, and the dropout parameter to 0.5. In addition, this paper uses SAM as the weight distribution model in the Attention layer. The obtained data set with relational annotation is trained on the modified CNN model. After model training, we achieved the results of an F1 score of 0.8641, a recall rate of 0.9264, and an accuracy rate of 0.9348 on the training set. On the test set, we obtained the results of an F1 score of 0.9021, a recall rate of 0.8875, and an accuracy rate of 0.9176. The result shows that it has a good extraction effect and stability. After the model is trained, the extracted information is mapped to the "edge" of the knowledge graph.

(3) Attribute extraction is the foundation of knowledge base construction and application, extracting attribute names and attribute values of entities from raw data of different information sources, constructing attribute lists of entities, forming complete entity concepts, and completing the entities. After completing entity identification and knowledge extraction by the above methods, this paper will manually extract the attributes of the entity identified by the algorithm and the extracted relationship, that is, by reading the standard text, if there is an attribute corresponding to the entity or the attribute corresponding to the relationship in the text, it will be extracted to make the knowledge expressed by the triplet more perfect.

3.3. Knowledge Fusion

Knowledge fusion is the integration of multifaceted knowledge in knowledge graph construction, including the same entity from different knowledge bases, multiple different knowledge graphs, and external knowledge from multiple heterogeneous sources. It determines the equivalent instances, equivalent classes, and equivalent attributes in the knowledge graph in order to finally realize the updating of the existing knowledge graph. The main tasks of knowledge fusion include entity alignment and entity disambiguation. Entity alignment is the main task of the knowledge fusion stage; the purpose is to find semantically identical entities. Entity disambiguation is to remove the ambiguity of entity indicators in different texts and map them to actual entities according to a given text.

3.4. Knowledge Processing

Knowledge processing is based on knowledge extraction and knowledge fusion, which processes basic facts to form a structured knowledge system and high-quality knowledge to realize unified knowledge management. The concrete steps of knowledge processing include ontology construction, knowledge reasoning, and quality assessment.

(1) Ontology construction refers to the construction of the concept template of knowledge at the pattern layer, which standardizes the description of concepts and their relationships within a specified domain. The process includes two parts: concept extraction and inter-concept relationship extraction. According to the degree of automation of

the construction process, it can be divided into manual construction, semi-automatic construction, and automatic construction. The purpose of ontology construction is to build a knowledge data model and hierarchical system; the main methods are manual editing, entity similarity, entity relationship automatic extraction, and so on.

(2) Knowledge reasoning is to mine or infer unknown or implicit semantic relations because of the incompleteness of existing facts or relations in the knowledge graph. The objects of knowledge reasoning can be entities, relationships, and the structure of a knowledge graph. The main methods of knowledge reasoning include the method based on logical rules, the method based on distributed representation, and the method based on neural networks [25].

(3) Quality assessment is usually carried out at the stage of knowledge extraction or fusion, and the confidence of knowledge is evaluated to retain the knowledge with high confidence and effectively guarantee the quality of the knowledge graph. The purpose is to improve the quality of knowledge samples, enhance the effect of knowledge extraction, and increase the effectiveness of the model.

3.5. Knowledge Update

Knowledge update is a standard that iteratively updates the content of the knowledge graph with the passage of time or the increase of new knowledge to ensure the timeliness of knowledge. The new knowledge of the standards can be added to the original knowledge graph only after knowledge fusion. At the same time, the following considerations should be taken into account: for the new entity or relationship, whether the same entity or relationship already exists in the existing graph, and if so, the duplicate should be carried out. If there is any invalid entity or relationship after knowledge updating, it should be removed.

4. Disassemble and Store

The data structuring, disassembling, and storing of standards are important links in the establishment of a standard database. Disassemble the identified and audited standards. First, the standards are classified according to their basic attributes and keywords; then, the standards are graded according to first-level titles and second-level titles. Finally, the standards are disassembled by manual dismantling, system-assisted dismantling, or intelligent dismantling strategies, and the structured storage of the standard is completed. The disassembly and storage block diagram of standards is shown in Figure 3, which mainly includes three parts: classification and grading of standards, disassembly of standards, and storage of standards.

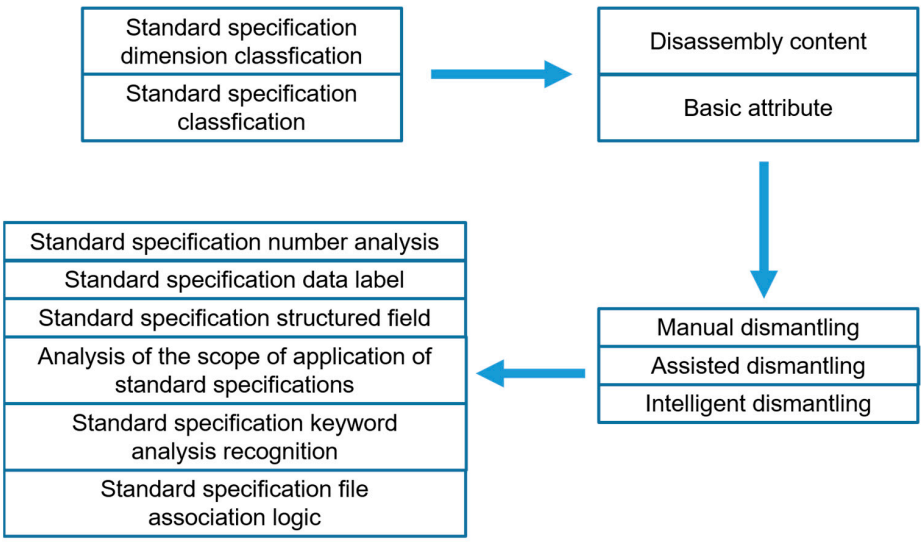

Figure 3. Standard disassembly and storage general block diagram.

4.1. Classification and Grading of Standards

Classification of standards is the first step to realizing standard disassembly and storage. Only by classifying different technical standards can we realize the structured identification and storage of standards, establish the mechanism for storing structured information of standards, and explore the intelligent matching method of technical standards in typical application scenarios. Through the research on the classification mechanism of oil and gas field surface engineering standards, it is known that the knowledge classification of oil and gas field surface engineering standards can be realized mainly by manual analysis or computer technology.

The main principles of classification are (1) Scientificity: Classification should be combined with the actual situation, mature standard data characteristics, formulate data classification standards, and build a stable classification system. (2) Systematicness: Formulate a unified classification standard and follow it, and the classification should be coherent and in-depth layer by layer to ensure that the classification is systematic. (3) Scalability: The classification system should ensure that enough space is reserved for the addition of new categories, can not destroy and disrupt the original classification system, and consider the extension and refinement of lower-level subclasses in an all-around way. (4) Practicability: Based on taking into account the above principles, it is necessary to combine the actual situation of the enterprise to make the practicality and operability of the classification stronger. (5) Simplicity: The number of classification layers of data resources should be as few as possible to facilitate the management and query of data resources.

According to the classification principle of standard, and combined with the construction characteristics of oil and gas fields, six dimensions are divided into multiple attributes of standards for standard classification, as shown in Figure 4. The standard classification involves standard stage, standard level, oil and gas field type, medium conditions, facility type, and specialty type, which are used to guide the classification of standards into repositories and support the rapid positioning of specific standards during retrieval. The application stages include design, construction, and acceptance. The standard levels include international standards, national standards, and industry standards. The oil and gas field types include oil fields, gas fields, and oil and gas fields. The medium conditions include oil, natural gas, shale gas, and tight gas. The facility types include pipelines, stations, and treatment plants. The specialty types include process, structure, corrosion prevention, and equipment. After the standard is divided, the associated standards are extracted from the standard database by combining basic attributes such as standard number, standard name, publication time, and keywords to form a list of associated standards.

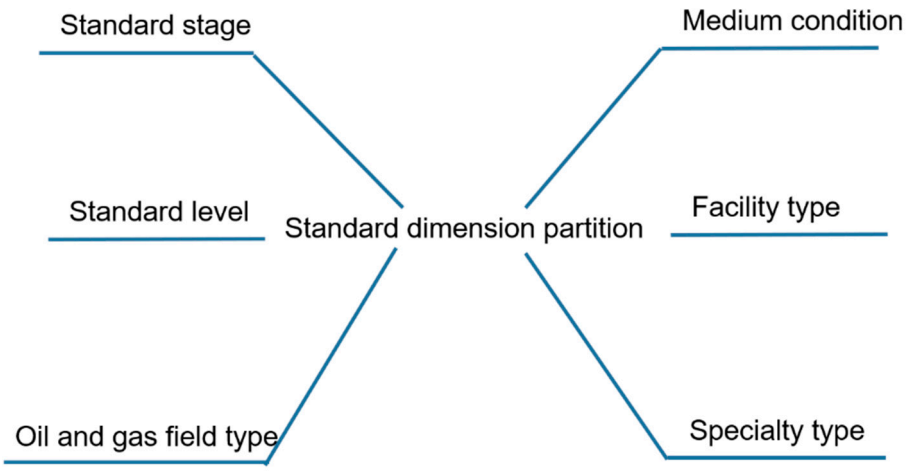

Figure 4. Dimension division of standards.

The grading of standards is based on the classification database of standards, and the method of decision tree is adopted to stratify variable files and data generated in the

project. Generally, the standards are graded according to first-level headings, second-level headings, and third-level headings.

4.2. Disassembly of Standards

After the standards are classified and graded, we can use manual disassembly, system-assisted disassembly, or intelligent disassembly to complete the structured disassembly of the standards to realize data entry.

(1) Disassemble the content

The disassembly content of the standard mainly includes the basic attributes of the standard, keywords, text bars, graphs, tables, and others. The disassembly content is shown in Figure 5. The basic attributes of the standard include standard serial number, standard type, release time, the first drafting unit, and the scope of application. According to the content of the standard clauses, the standard clauses can be divided into mandatory clauses, recommended clauses, and optional clauses. Classified by the most demanding requirements within the same terms. The setting of keywords includes (1) Standard names, nouns, and verbs of titles and articles at all levels. (2) Nouns and verbs of the names of drawings, tables, schedules, and attached drawings. (3) Conventional keywords such as the scope of adaptation, quality, safety, inspection, measures, design, construction, start, test, and acceptance. According to the basic attributes, first-level headings, second-level headings, third-level headings, keywords, and their corresponding terms, the content of the standard is disassembled, and the structured storage of the standard is realized. According to the elements such as text, charts, and tables in the standard, the text section and chart name are decomposed into editable fields. The fields after the chart name form an association with the chart to support later standard retrieval. And the content of the chart is not decomposed.

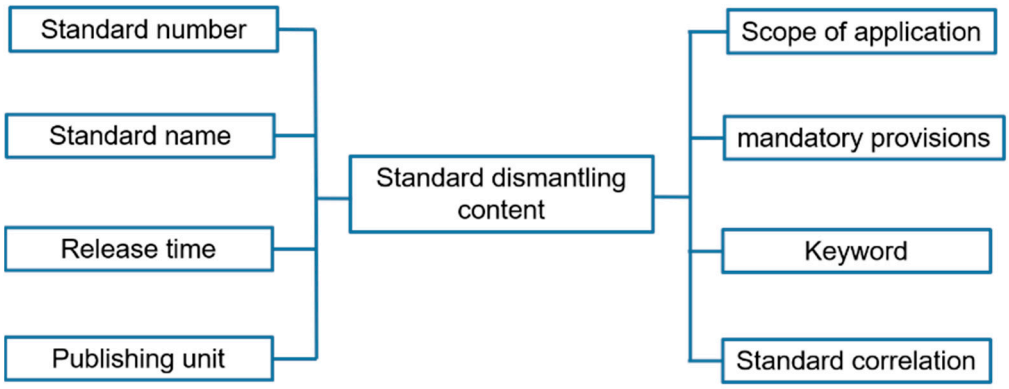

Figure 5. Standard dismantling content.

(2) Manual disassembly

After obtaining an editable standard structure document through manual proofreading by utilizing OCR recognition technology, manual disassembly can be carried out. The process of manual disassembly is that technicians utilize professional knowledge and experience to manually extract and disassemble the first-level title, second-level title, basic attributes, keywords and their corresponding mandatory provisions, recommended provisions, optional provisions, scope of application, terminology and other information, thereby ultimately forming a structured form. The manual disassembly process is very time-consuming and requires technicians to copy and paste the contents of the standard to form a file for database storage.

(3) System assisted disassembly

Load the corrected structured documents into the intelligent platform and utilize the knowledge processing annotation tool of the standard documents of the intelligent

platform. The technician manually selects the corresponding content in the standard document, and the content obtained from the selection can be corrected manually; at the same time, different labels are added to the selected content, such as standard name, scope of application, keywords, and mandatory provisions. After adding labels, the standard text and labeling information matching into the database is realized.

(4) Intelligent disassembly

Intelligent disassembly is to effectively realize the process of structured processing, automatic content identification, effective knowledge extraction, and storage of standards through advanced technical means. We upload standard documents at the platform entrance, and the system adopts natural language processing technology to automatically identify the basic information of the documents and process the text, then utilizes the convolutional neural network method in the above knowledge graph technology to identify and extract the entities, relationships, and attributes of the standard, and finally stores the data into the graph database. In the intelligent disassembly, the information to be extracted and filled in includes the name of the standard, application stage, standard grade, and oil and gas field type. The identification of keywords includes "acceptance", "industry standard", and "oil and gas field". This keywords are automatically extracted according to the keyword database. Eventually, we will complete the intelligent disassembly of structured documents on the system.

4.3. Storage of Standards

The oil and gas field standard database is a computer database system platform to solve a series of problems in the current standard adopted in the work. The database system mainly realizes the centralized storage and structured disassembly of standard documents and the professional application of standard information through the systematic management of standards and clauses. Conduct demand analysis of oil and gas field surface engineering technical standard information database, including the type, scope, and quantity of standard data and the application interaction of data, to determine the use requirements and various constraints of the information database, to form the demand specification, determine the entity object, relationship, and then form the standard information database. The oil and gas field standard database is the core part of management information systems, office automation systems, and other information systems, which is an important technical means for management.

After the standard file is disassembled and marked, entity definition and logical relationship definition are carried out for the disassembled content, and a large number of simple text contents are transformed into "three-element knowledge nodes" composed of multiple entities, relationships, and attributes to complete the construction of knowledge graph. In this paper, a graph database (Neo4j) is used to store the data in the knowledge graph. After the "three-element knowledge node" data are stored in the graph database (Neo4j), the dimension, breadth, and depth of knowledge information will be expanded, which will lay the knowledge foundation for the intelligent retrieval query and intelligent recommendation system. The database entry spectrum of standards mainly includes the storage of data knowledge and the construction of a standard knowledge graph.

(1) Data knowledge storage

Neo4j adopts the attribute graph model, which is a data management system based on the storage unit of points and edges and the design principle of efficient storage and query of graph data. The graph database is non-relational. Compared with the relational database, Neo4j is more flexible and efficient, storing the association between data as part of the data. Labels, directions, and attributes can be added to associations, while queries for relationships in other databases must be externalized at runtime, mainly for online transaction processing (OLTP). Neo4j uses Cypher declarative query language, while the database driver supports a variety of popular programming languages and can realize the

visual display of knowledge graphs on the front-end page through Cypher language or other programming languages.

(2) Construction of a knowledge map of standards

Based on the disassembled structured knowledge, we use the above knowledge graph technology to realize named entity recognition, relation extraction, and attribute extraction so as to obtain the entity, relation, and attribute of the standard. We take the relationship node as the medium to establish the relationship connection and complete the construction of the knowledge graph.

5. Intelligent Retrieval

Intelligent retrieval is the most direct way to apply the standard database. When the user queries and searches, the user's natural language query enters the retrieval system after semantic analysis and processing and then matches the content in the knowledge base. The integrated feedback results are presented to the user in a visual form. It improves the efficiency of keyword search and query. The intelligent retrieval system is divided into three parts: the bottom layer, middle layer, and presentation layer. The hierarchical diagram of the intelligent retrieval system is shown in Figure 6.

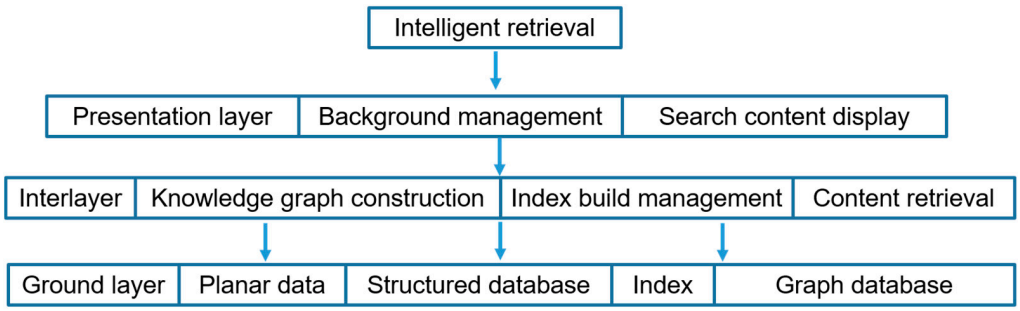

Figure 6. Schematic diagram of intelligent retrieval system.

(1) The bottom layer is an intelligent retrieval storage system, including two-dimensional plane data, structured database, inter-entity index, and graph database of search objects.
(2) The middle layer is the retrieval processing layer, including the construction of standard and normative knowledge graphs, index construction management, and content retrieval engine.
(3) The presentation layer is the display of content retrieval results, which are sorted according to the keyword matching degree, release time, and terms requirements in the knowledge graph structure to reduce data overburden and improve the efficiency of data query in the system.

5.1. Similarity Retrieval

Cypher is a descriptive graph query language specially developed and designed for the data model of the graph database Neo4j. It only needs simple code to realize efficient retrieval of the Neo4j database. If the user enters the standard name, keyword, and other entities in the Neo4j database, the Cypher query statement is directly constructed for the query, and all the associated information or a specified relationship information connected to the entity is obtained, then the returned results are displayed to the user in the form of knowledge graph and term list.

When there is no current search term in the thesaurus, the semantic similarity calculation is carried out to obtain the word with a similar meaning as the concept of the search term. In this paper, the semantic similarity calculation method of ontology concept and vocabulary is selected. The formula for calculating semantic similarity is shown in Equation (1):

$$Sim(c, w) = \beta \cdot Sim_1(c, w) + (1 - \beta) \cdot Sim_2(c, w) \quad (1)$$

$Sim_1(c,w)$ is the minimum edit step required to transform a string in one vocabulary into a string in another vocabulary. For example, if the minimum edit distance to convert "xyz" to "$xyzd$" is 1, a new lowercase letter d needs to be added. The specific minimum calculation editing method is shown in Equation (2).

$$Sim_1(c,w) = 1 - \frac{ED(c,w)}{MLen(c,w)} \tag{2}$$

$ED(c,w)$ is the minimum number of edit operations required to convert from one to the other between two strings. $MLen(c,w)$ is the maximum semantic cosine length between c and w.

$Sim_2(c,w)$ is the maximum semantic cosine distance between two words. Its similarity is equal to the semantic cosine similarity of the distance and embedding representation of two synonyms, and the formula is defined as shown in Equation (3):

$$Sim_2(c,w) = cos(e_c, e_w) \tag{3}$$

e_c, e_w are the embedding representations of the words c,w.

5.2. Search Dimensions

The dimensions of standard intelligent retrieval include precision retrieval and full-text fuzzy retrieval. Based on the attribute label and keyword library, it can be searched accurately by information such as name and standard number and also can be searched by keywords and other information for full-text fuzzy search queries. According to the six dimensions of the standard and other basic attributes and keywords, the associated standards are retrieved from the standard database, and the standard list is formed. At the same time, according to the specific standards in the standard list, the standard clause is quickly located from the standard structured form database to form the clause list. It can also be described as a search mechanism within the system that allows for quick location of standards and location-specific terms to retrieve related standards, regulations, charts, and tables. Quickly locating a standard is built on six dimensions, other attributes of the standard, and keywords. The second type of search mechanism matches keywords with the title or content of a standard to position it to specific terms. The combination of "and" and "or" is adopted to realize multi-element search. There are three search methods for users on the client side: (1) Find specific criteria by project name and number. (2) Find a list of standards in any combination of six categories. (3) Find standard terms by six categories and keyword combinations.

5.3. Search Method

The retrieval method is based on the stored data of the standard knowledge graph to realize intelligent retrieval in different scenarios. When the input content matches the category name, text name, keyword, text content, and attribute node, the priority of the content display is the category name, text name, keyword, text content, and attribute node, respectively. After the user searches and filters the standard files, the system provides a list of the standard files for matching query, which is displayed in the form of lists and graphics, and the matching results from high similarity to low similarity are arranged in the list. After entering the specific standard file, we can search the query keywords, locate the mandatory clauses and key provisions, and support direct access to the structured file of the standard for viewing. After analyzing the input search content by natural language processing (NLP) technology, we can understand the questions expressed by users, reduce text ambiguity, and quickly locate the search results with a high matching degree.

(1) Search based on classification

Classification-based retrieval is based on the knowledge graph of standard classification, and all the standards under a certain classification can be output together. For

example, entering "electrical" can list all the standards belonging to electrical according to the sorting algorithm. Input "electrical, line label" will meet both electrical and line label attributes of the standard according to the sorting algorithm to display the list. After selecting a search result, open the standard to read.

(2) Search based on text name

Text name-based retrieval is based on the standard classification knowledge graph, inputs the standard text keywords, traverses the standard classification knowledge graph when it can match the attribute node of the standard classification knowledge graph, displays the standard retrieval result, and opens it for reading.

(3) Search based on keyword

Keyword-based retrieval is based on the knowledge graph of standards. Keywords are entered to match the keyword attribute nodes in the knowledge graph, the relevant standards that can be matched by the keywords are displayed in a list, the content of the keyword in the text is highlighted, and the standard text is opened for standard reading.

(4) Search based on mandatory clause

Retrieval based on mandatory clause is based on a standard knowledge graph. The input text matches the content attribute node in the knowledge graph to display the standard that can match the content. The content appearing in the text is highlighted in the search list, and the standard text is opened for annotated reading.

(5) Search based on text attribute nodes

The retrieval based on text attribute nodes is based on the knowledge graph of standards. Input the name of the text attribute node to match the content attribute node in the knowledge graph, display the standards that can match the content, highlight the content in the text in the search list, and open the standard text for annotated specification reading.

6. Intelligent Push

The recommendation system is an effective way to solve the problem of information explosion and is applied in various fields to enhance user experience [29]. Intelligent push of standards is the ability to provide key information push of standards. Intelligent push can intelligently match the required standards in scenarios such as construction drawing design review, form a strong checklist of applicable standards, and provide instantly accessible associated standard items.

6.1. Recommendation Algorithm

Through the concept of "project profiling", we can deliver technical standard content in all the dimensions required by the project. At present, various types of recommendation algorithms emerge in an endless stream, and there are multiple standards for their classification. A commonly used classification method roughly divides recommendation algorithms into three categories: collaborative filtering-based, content-based, and hybrid. Among them, recommendation algorithms relying on collaborative filtering are the most widely used. Its basic principle is to recommend standards to users according to their previous preferences or the choices of other users with similar interests. This kind of algorithm can be subdivided into two categories: memory-based and model-based. No matter what kind of collaborative filtering algorithm, the system needs to obtain the historical interaction information of the user on the item, including the user's collecting, browsing, and other behavior records. If behavior (u) is used to represent the user's behavior record, then the matching value P of standard i calculated by the collaborative filtering algorithm can be expressed as

$$P(i \mid \text{ behavior } (u)) \tag{4}$$

In this paper, we use an item-based collaborative filtering algorithm to find the similarity between standards by analyzing the historical behavior of the user's retrieval of

the standards, and then the user will be recommended. The dataset contains the user's browsing length of the standards, and the cosine-based similarity is calculated as follows:

$$w_{i,j} = \cos(\theta) = \frac{\sum_{k=1}^{len}\left(n_{ki} \times n_{kj}\right)}{\sqrt{\sum_{k=1}^{len} n_{ki}^2}\sqrt{\sum_{k=1}^{len} n_{kj}^2}} \quad (5)$$

n_{ki} is the browsing duration of user k for standard i.

This type of algorithm is significantly dependent on behavior (u) data, which leads to the following key issues affecting the effectiveness of recommendations: (1) The cold start problem: New users or new standards do not have sufficient historical interaction data in the system to analyze the similarity between users and standards, which makes it difficult for new standards to be recommended. As for new users, the system is also unable to make effective recommendations due to the difficulty of understanding their interests and preferences. (2) Data sparsity problem: In the recommendation system, many standards have no or few user interaction records, and the sparse interaction data makes it difficult to paint accurate portraits of these standards, which leads to the decline of many model-based methods (such as matrix decomposition).

Recommender systems for knowledge graphs have been shown to compete with state-of-the-art collaborative filtering systems and can effectively address issues such as new items and data sparsity [30]. Thus, the above problems are alleviated, and the performance and accuracy of the recommendation system are improved.

6.2. Push Dimensions

As shown in Figure 7, push dimensions can be divided into three categories according to keyword matching degree, release time, and terms. The standard list is sorted by keyword matching degree from high to low and publication time from near to far. Standard terms are sorted by keyword matching degree from high to low, release time from near to far, and terms requirements (by mandatory terms, recommended terms, optional terms order). It can be sorted by a single dimension or multiple-dimension associations.

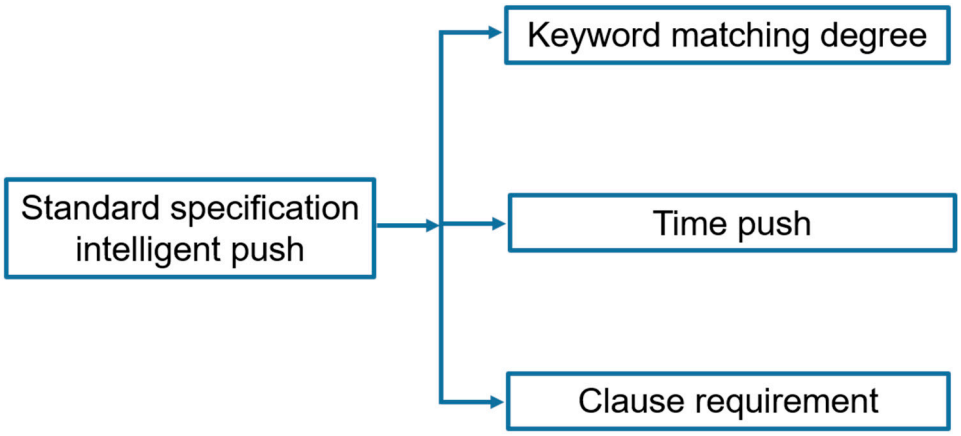

Figure 7. Push sort schematic.

6.3. Push Mode

Standard and normative intelligent push methods mainly include standard and normative association push, user habit association push, keyword association push, text bar association push, and high-frequency search standards association push.

(1) Standard association push: Based on the standard classification knowledge graph, the relevant attributes of the retrieval object are searched upward according to the graph, and other standards that have "reference and supporting use" and the same classification attributes with the retrieval object are associated and recom-

mended, and the recommendation ranking is comprehensive based on the number of matching dimensions.

(2) User habit association push: Build the user habit knowledge graph, view the standard situation according to the user history, comprehensively recommend the standards with the most historical views and the most concerned types, and display the results in the form of a list, open the standards for reading.

(3) Keyword association push: When the standard is read, other standards containing the keyword will be associated and pushed, and the results will be displayed in the form of a list. The standard can be read when opened.

(4) Text mandatory clause association push: When the standard is read, other standards corresponding to the search object mandatory clause will be associated and pushed, and the results will be displayed in the form of a list. When reading standards, other standards containing the keyword will be associated and pushed, and the results will be displayed in the form of a list, which can be opened to read the standards.

(5) The associated push of high-frequency search standards: According to the frequency of search standards and norms of all users of the system, the high-frequency standards and norms are associated and recommended.

7. Applications

With the method of disassembling into the library, intelligent retrieval, and intelligent push for knowledge graph, this paper develops an intelligent platform for technical standards of surface engineering and construction of oil and gas fields, forms a highly shareable standard library, and realizes the intelligent retrieval and intelligent push of technical standards of surface engineering and construction. Taking the standard "Environmental Noise Emission Standards for Industrial Enterprises at Plant Boundaries" as an example, the system-assisted disassembly method is used to disassemble the content of the standard document by data-structured knowledge processing annotation tools according to the scope of application, keywords, mandatory provisions, and so on. After the standard file is disassembled and annotated, entity definition and logical relationship definition are carried out on the disassembled content, which is transformed into the "three-element knowledge node" data composed of multiple entities, relationships, and attributes and stored in the graph database (Neo4j) to complete the construction of the knowledge graph, as shown in Figure 8. Under the established database, programming languages such as JAVA and C++ are used to connect to the database to build a search and matching model based on the knowledge graph and to validate the feasibility. The search and matching model is optimized according to the query results and applied on the Web site; users can search and query the standards and standard terms on their own needs.

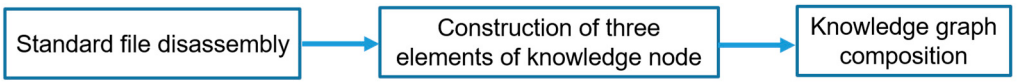

Figure 8. Standards are stored to form a knowledge graph.

The interface of the intelligent platform of technical standards for the construction of oil and gas field surface engineering is developed. In the standard database interface, standard search can be carried out by standard name and other attributes, intelligent search can be carried out by inputting keywords and so on, and condition screening of standard categories can be carried out by classification of six dimensions on the left side of the interface. In the standard maintenance interface, it can add and modify the standard classification, upload the standard, identify and download the OCR in the system, upload the structured file after manual review, mark and disassemble it on the system, and finally complete the construction and storage of the knowledge graph. At the same time, the platform provides a manual update of standards repeal announcements and can track and pay attention to the applicable standards selected in the implementation project and provide new standard update tips and reminders. For example, the B-feature project has

confirmed the application of standard 2, the new version of standard 2 is updated during the project execution, the system automatically replaces the original standard 2 information, replaces the new standard information concern items, and pushes the standard update prompt to the project team to remind the project team to pay attention to the standard changes in time.

8. Conclusions

In this paper, knowledge graph technology, OCR recognition, and other technologies are used to research the standard knowledge classification mechanism, data extraction, database construction mechanism, data structuring, intelligent retrieval method, and intelligent push method.

(1) This paper innovatively applies knowledge graph technology to the construction of a standard database. At the same time, the steps of constructing a knowledge graph and the main algorithms used in this paper are introduced. Through the construction of a knowledge graph, the retrieval efficiency and recommendation effect of standards are improved.
(2) According to the properties and contents of the standards for oil and gas field surface engineering, this paper classifies the standards in six dimensions and introduces disassembly methods of structured documents. Finally, the methods and applications of intelligent search and intelligent push are introduced.
(3) The research can provide technical support for the project management and quality control of oil and gas field surface construction projects and support the unified intelligent management of surface engineering in oil and gas field enterprises. It has good value in popularization and application and provides a reference for the construction of industry-standard databases and intelligent search and push.

Author Contributions: Methodology, J.Z.; Software, M.L.; Formal analysis, Z.D.; Investigation, Z.H. and F.W.; Resources, D.Z.; Data curation, L.L. and W.Z.; Writing—original draft, T.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

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

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, T. Practice and thinking of oil and gas industrial digitalization transformation. *Oil Forum* **2020**, *39*, 29–33.
2. Su, J.; Yao, S.; Liu, H. Data Governance Facilitate Digital Transformation of Oil and Gas Industry. *Front. Earth Sci.* **2022**, *10*, 861091. [CrossRef]
3. Jia, A.; Guo, J. Key technologies and understandings on the construction of smart fields. *Pet. Explor. Dev.* **2012**, *39*, 118–122. [CrossRef]
4. Fang, Z.; Huan, Z.; Wang, Z.; Li, G.; Chen, R. Research on Action Layers and Application and Database Design of Safety Barrier in Petrochemical Plant. In Proceedings of the Pressure Vessels and Piping Conference, Online, 13–15 July 2021; American Society of Mechanical Engineers: New York, NY, USA, 2021; Volume 85314, p. V001T01A043.
5. Wang, T.; Xuan, W.; Wang, X.; Ren, K. Overview of oil and gas pipeline failure database. In *ICPTT 2013: Trenchless Technology*; American Society of Civil Engineers: Reston, VA, USA, 2013; pp. 1161–1167.
6. Bader, S.R.; Grangel-Gonzalez, I.; Nanjappa, P.; Vidal, M.E.; Maleshkova, M. A knowledge graph for industry 4.0. In Proceedings of the Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Greece, 31 May–4 June 2020; Proceedings 17. Springer International Publishing: Cham, Switzerland, 2020; pp. 465–480.
7. Yu, T.; Li, J.; Yu, Q.; Tian, Y.; Shun, X.; Xu, L.; Zhu, L.; Gao, H. Knowledge graph for TCM health preservation: Design, construction, and applications. *Artif. Intell. Med.* **2017**, *77*, 48–52. [CrossRef] [PubMed]
8. Wang, C.; Ma, X.; Chen, J.; Chen, J. Information extraction and knowledge graph construction from geoscience literature. *Comput. Geosci.* **2018**, *112*, 112–120. [CrossRef]
9. Siddharth, L.; Blessing, L.T.; Wood, K.L.; Luo, J. Engineering knowledge graph from patent database. *J. Comput. Inf. Sci. Eng.* **2022**, *22*, 021008. [CrossRef]

10. Hao, X.; Ji, Z.; Li, X.; Yin, L.; Liu, L.; Sun, M.; Liu, Q.; Yang, R. Construction and application of a knowledge graph. *Remote Sens.* **2021**, *13*, 2511. [CrossRef]
11. Du, R.; Li, Y.; Shang, F.; Wu, Y. Study on ontology-based knowledge construction of petroleum exploitation domain. In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 23–24 October 2010; Volume 2, pp. 42–46.
12. 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]
13. Ge, J.; Li, Z.; Li, T.; Qiang, B. Petroleum exploration domain ontology-based knowledge integration and sharing system construction. In Proceedings of the 2011 International Conference on Network Computing and Information Security, Guilin, China, 14–15 May 2011; Volume 1, pp. 84–88.
14. 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; pp. 161–166.
15. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G.D.; Gutierrez, C.; Kirrane, S.; Gayo, J.E.L.; Navigli, R.; Neumaier, S.; et al. Knowledge graphs. *ACM Comput. Surv. (Csur)* **2021**, *54*, 1–37. [CrossRef]
16. Gong, F.; Ma, Y.; Gong, W.; Li, X.; Li, C.; Yuan, X. Neo4j graph database realizes efficient storage performance of oilfield ontology. *PLoS ONE* **2018**, *13*, e0207595. [CrossRef] [PubMed]
17. Zhou, Q.; Wang, C.; Xiong, M.; Wang, H.; Yu, Y. Spark: Adapting keyword query to semantic search. In Proceedings of the International Semantic Web Conference, Busan, Korea, 11–15 November 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 694–707.
18. Zhu, G.; Iglesias Fernandez, C.A. Sematch: Semantic entity search from knowledge graph. In Proceedings of the SumPre 2015—1st International Workshop on Summarizing and Presenting Entities and Ontologies, Portoroz, Slovenia, 1 June 2015.
19. Huang, S.; Wang, Y.; Yu, X. Design and Implementation of Oil and Gas Information on Intelligent Search Engine Based on Knowledge Graph. *J. Phys. Conf. Ser.* **2020**, *1621*, 012010. [CrossRef]
20. Wang, C.; Yu, H.; Wan, F. Information retrieval technology based on knowledge graph. In *Proceedings of the 2018 3rd International Conference on Advances in Materials, Mechatronics and Civil Engineering (ICAMMCE 2018)*; Atlantis Press: Amsterdam, The Netherlands, 2018; pp. 291–296.
21. Ehrlinger, L.; Wöß, W. Towards a definition of knowledge graphs. *SEMANTiCS (Posters Demos SuCCESS)* **2016**, *48*, 2.
22. Peng, C.; Xia, F.; Naseriparsa, M.; Osborne, F. Knowledge graphs: Opportunities and challenges. *Artif. Intell. Rev.* **2023**. [CrossRef] [PubMed]
23. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Philip, S.Y. A survey on knowledge graphs: Representation, acquisition and applications. *arXiv* **2020**, arXiv:2002.00388. [CrossRef] [PubMed]
24. Wang, J.; Zhang, W.; Wang, Y.; Sun, Z. Constructing and inferring event logic cognitive graph in the field of big data. *Sci. Sin. Informationis* **2020**, *50*, 988–1002. (In Chinese)
25. Chen, X.; Jia, S.; Xiang, Y. A review: Knowledge reasoning over knowledge graph. *Expert Syst. Pplications* **2020**, *141*, 112948. [CrossRef]
26. Yao, S.; Wang, R.; Sun, S.; Bu, D.; Liu, J. Rule-guided joint embedding learning of knowledge graphs. *J. Comput. Res. Dev.* **2020**, *57*, 2514–2522. (In Chinese)
27. Zhao, X.; Jia, Y.; Li, A.; Chang, C. A survey of multisource knowledge fusion technology. *J. Yunnan Univ. Nat. Sci. Ed.* **2020**, *42*, 65–79. (In Chinese)
28. Vakulenko, S.; Fernandez Garcia, J.D.; Polleres, A.; de Rijke, M.; Cochez, M. Message passing for complex question answering over knowledge raphs. In Proceedings of the 28th ACM Int Conf on Information and Knowledge Management, Beijing, China, 3–7 November 2019; ACM: New York, NY, USA, 2019; pp. 1431–1440.
29. Quijano-Sánchez, L.; Cantador, I.; Cortés-Cediel, M.E.; Gil, O. Recommender systems for smart cities. *Inf. Syst.* **2020**, *92*, 101545. [CrossRef]
30. Catherine, R.; Cohen, W. Personalized recommendations using knowledge graphs: A probabilistic logic programming approach. In Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 325–332.