Design and Implementation of Oil and Gas Information on Intelligent Search Engine Based on Knowledge Graph

Shujun Huang1*, Yuyan Wang1* and Xiang Yu2

1 Strategic Research Center of Oil And Gas Resources, Ministry of Natural Resources, Beijing, China
2 Beijing Shida Zhengxin Mineral Resources Technology Consulting Co., Ltd. Beijing, China
Email: huangsj@sinooilgas.org.cn; wangyy@sinooilgas.org.cn; 11024556@qq.com

Abstract. With the deepening of oil and gas informatization construction, various types of oil and gas information increase geometrically. In response to the challenges brought by massive oil and gas information in the search process, this paper proposes an intelligent search engine for oil and gas information based on knowledge graph. Functional modules such as knowledge fusion, knowledge processing, semantic relevance judgment, topic classification, and search result set sorting display, using map service as an interactive interface, realize intelligent search of oil and gas information. Compared with traditional search engines, it has more advantages in the search of oil and gas information, and it is easier to understand the user’s search intent in the oil and gas field, effectively improving the quality of search.

Keywords. Oil and gas information; knowledge graph; intelligent search engine; information retrieval.

1. Introduction
In the current situations of both increasing domestic oil and gas exploration and development, and ensuring national energy security, oil and gas resources are an important energy support for national construction. With the deepening the construction of oil and gas resource informatization, oil and gas information show a exponential growth trend. Compared with the full-text search services provided by traditional search engines, intelligent search services with specified features for oil and gas has been increasingly demanded by the oil and gas industry.

Traditional search engines are mainly based on full-text matching of keywords, and lack of knowledge understanding and reasoning capabilities for query content, resulting in certain limitations in search results in accuracy, depth and breadth. With the continuous introduction of artificial intelligence technologies such as natural language processing and knowledge graphs, the industry has carried out the research and development of various intelligent information search engine products, which has greatly improved the user experience in personalized and intelligent search [1], such as Baidu’s Baidu Zhixin. The oil and gas industry are a highly specialized field that involves a variety of different disciplines in both upstream and downstream subjects. The large and complex knowledge system requires professional search engines, which may not be well satisfied by commonly used intelligent search engines such as Google, Baidu or Sogou, which pay more attention to the generality of search while lacks the professional knowledge system of the oil and gas industry. In addition, these universal knowledge bases often come from various Internet web pages and thus the data content lacks reliability.
When performing a query search, the content required by the user cannot be well obtained by these search engines.

The informatization construction of the Center of Oil and Gas (Strategic Research Center of Oil and Gas Resources, MNR, PRC) has been deepened year by year, and has successively developed information systems for strategical research of oil and gas resources, mining management support, comprehensive management and public service, while accumulating a large amount of related data on oil and gas resource. Aiming at this massive information, it is important to accelerate the research and implementation for the intelligent search engine of oil and gas information.

2. Overall Architecture Design
The key to the realization of the oil and gas information intelligent search engine is whether it can accurately construct a knowledge graph for the oil and gas field. In the process of building the oil and gas knowledge graph, we can combine the foundation of Center of Oil and Gas informatization construction, apply both top-down and bottom-up methods to extract ontologies and rules from the existing structured data of the Center of Oil and Gas, and fill them into the knowledge graph. Meanwhile, for the poorly constructed information collected from the Internet and various types of documents, we can identify entities, attributes, and relationships by pattern recognition and rule making, and then add them to the knowledge graph.

As shown in figure 1, the system can be divided into three layers, including the lower data layer, the middle service layer, and the upper display layer. The data layer mainly includes semi-structured data, unstructured data, structured data, and knowledge graph data. The construction of knowledge graph is a continuous and recurring process, whose core includes knowledge extraction, knowledge fusion and knowledge processing. Because the structured data is well standardized, it is easier to extract knowledge from it. Meanwhile, the poorly standardized semi-structured and unstructured data are difficult to be directly applied for acquiring knowledge. Therefore, a series of operations such as entity recognition, attribute extraction, and relationship extraction are required to extract ontologies and associations of knowledge. Then the data can be stored into the knowledge graph after knowledge fusion. Meanwhile, starting from existing knowledge through knowledge reasoning, we can construct new associations among entities and expand the knowledge network. The service layer mainly includes query processing and retrieval services. The query processing performs entity recognition, syntax analysis, and semantic analysis of user query sentences based on the ontology library. Then we rely on the retrieval service to complete the search based on the knowledge graph. Finally, the search results are returned to the display layer, which mainly provides intelligent search and results display for different users on multiple devices.

![Figure 1. Overall architecture design.](image-url)
3. Key Technical Research

3.1. The Construction of the Oil and Gas Information Ontology Library

Ontology is a specification for the expression of entities and relationships. It can describe the conceptual model of the objective world at the level of semantics and knowledge. It is also the semantic basis for shared communication between different subjects in the oil and gas field. A domain ontology usually includes concepts, attributes of concepts, relationships between concepts, and constraints on attributes and relationships. The oil and gas information ontology library are the basis for realizing the division of entity domains in the oil and gas field. Combining with the construction of the ontology database, we can effectively extract knowledge entities and entity relationships.

Combining the business characteristics of the Center of Oil and Gas, the oil and gas information involved mainly includes topics such as resource management, reserves management, mineral rights management, exploration and exploitation, strategic research, and industry information, which are intertwined and complicated with a rich vocabulary. After years of work, the Center of Oil and Gas has accumulated a large amount of professional information in the oil and gas field, including structured data and unstructured literature. These data have high accuracy and reliability, which has laid the foundation for the construction of the ontology library.

The key to establishing a domain ontology model is to establish a suitable domain ontology framework model [2]. This article uses the combination of Ontology Development 101 method (also known as the seven-step method) and Methontology method to effectively make up for the shortcomings of the two methods. As shown in figure 2, Based on business division and characteristics, our strategy analyzes the professional vocabulary in the oil and gas field and then find out related concept definitions and relationships between concepts, such as: kind-of relationship, part-of relationship and etc. Finally, we use the Protégé tool to build the oil and gas information ontology.

Figure 2. Conceptual relationship of oil and gas exploration and development (partial results).

3.2. Knowledge Entity Recognition Based on BiLSTM-CRF Model

Knowledge entity recognition is the identification and extraction of corresponding strings capable of expressing knowledge entities in text data. It is the most critical part of the entire knowledge graph construction process and the basis of subsequent semantic analysis and entity relationship extraction. The construction quality directly affects the search quality of search engines.

The oil and gas industry are a complex field involving various disciplines such as exploration and development, oil and gas geology, reservoir engineering, mining engineering, engineering mechanics, and chemical engineering, from which rich and complex data of various types are produced. To maximize the quality of the annotation, our framework annotates the keywords based on the professional library accumulated by the Center of Oil and Gas while reducing manual annotation. We apply the
BiLSTM-CRF method to model the input text by representing the text as a low-dimensional and high-density vector, and then deepen the representation for the text through the BiLSTM layer. BiLSTM fuses two sets of LSTM layers with opposite learning directions (a forward input sequence and a reverse input sequence), which can realize that the current word contains both historical information and future information, and be more conducive to labeling the current word. Finally we use the CRF layer to decode the entire labeled sequence in order to get the final result [3].

As shown in Figure 3, after bidirectional feature extraction at the BiLSTM layer, the CRF layer is accessed for global label prediction. It has both the advantages of BiLSTM’s automatic construction of context features and the global optimal result. The model calculates the optimal label sequence by formula (1). We denote the output dimensions of BiLSTM as tag size, which equals the probability of \( w_i \) mapping to tag. We then denote the output matrix of BiLSTM as \( P \), where \( P_{i,j} \) represents the prediction probability of word \( w_i \) mapping to \( \text{tag}_j \).

\[
s(X, y) = \sum_{t=0}^{T} A_{y_t, y_{t+1}} + \sum_{t=1}^{T} P_{i, j_t}
\]

Figure 3. Structure of BiLSTM-CRF.

It can be seen from the above formula that the score of the entire sequence is the sum of the scores of each character mark, and the score of each character mark is obtained by adding the prediction probability of BiLSTM and the transition probability of CRF, which makes up for the shortcomings of traditional BiLSTM, thus ensuring the effectiveness of the sequence of the marking results.

3.3. Query Error Correction Based on N-Gram Model

This article uses the N-Gram model to evaluate whether the search statement is reasonable, analyzes whether the query content has errors and corrects them, and improves the user experience of the search engine.

The N-Gram model evaluates the probability of each word in the query based on the premise under the Markovian hypothesis that the appearance of the Nth word in a query only depends on the N-1th word [4]. The probability of the entire query is the product of the probability of each word appearing, for example, the query \( Q = (W_1, W_2, ..., W_n) \), which is further applied to determining whether the query is reasonable. The formula in the Bi-Gram model is as follows:

\[
P(Q) = P(W_1)P(W_2|W_1)P(W_3|W_2) ... P(W_n|W_{n-1})
\]

\[
P(W_n|W_1W_2...W_n) \text{ can be calculated by Maximum Likelihood Estimation as formula (3):}
\]

\[
P(W_n|W_1W_2...W_n) = \frac{c(w_1w_2...w_n)}{c(w_1w_2...w_{n-1})}
\]
In formula (3), \( c(w_1 w_2 \ldots w_n) \) represents the number of times that the phase \( w_1 w_2 \ldots w_n \) occurs in the lexicon:

\[
P(Q) = P(W_1) \frac{c(w_1 w_2)}{c(w_1)} \frac{c(w_2 w_3)}{c(w_2)} \ldots \frac{c(w_{n-1} w_n)}{c(w_{n-1})} \tag{4}
\]

Similarly, the formula of the Tri-Gram model can be derived as follows:

\[
P(Q) = \prod_{i=1}^{n} \frac{c(w_{i-2} w_{i-1} w_i)}{c(w_{i-2} w_{i-1})} \tag{5}
\]

Through the analysis of search logs and the word segmentation of query sentences, a training corpus required for query error correction is formed. This article uses Bi-Gram for calculations with a query word length of 2 and Tri-Gram for calculations larger than 2.

When the query word does not exist in the corpus, the probability calculated according to the N-Gram model is 0, which causes typical data sparseness problems. Therefore, data smoothing needs to be done to avoid this problem. This article uses Add-one Smoothing algorithm to deal with the data sparseness, by adding 1 to the time of occurrence for each query term [5]. Finally, we can obtain the formula for the Bi-Gram model with Add-one Smoothing:

\[
P(W_i|W_{i-1}) = \frac{c(w_{i-1}w_i)+1}{c(w_{i-1})+v_1} \tag{6}
\]

Similarly, the formula of the Tri-Gram model with Add-one Smoothing can be derived as follows:

\[
P(W_i|W_{i-1}W_{i-2}) = \frac{c(w_{i-2}w_{i-1}w_i)+1}{c(w_{i-2}w_{i-1})+v_2} \tag{7}
\]

In the above formula, \( v_1 \) and \( v_2 \) represents the number of sets \( c(w_{i-1}) \) and \( c(w_{i-2}w_{i-1}) \), respectively.

3.4. Sorting Correction Mechanism Based on Implicit User Feedback

The quality of search result ranking is an important indicator of the quality of a search engine. This article optimizes and refines the search result ranking based on implicit user feedback to achieve the personalized search needs of different users. Compared to explicit user feedback, collecting implicit feedback data does not require additional work or cause dissatisfaction of users. Therefore, it has the characteristics of lower collection cost, wider application scenarios, and larger data size [6].

Use the log system to collect various characteristic data of users in the system, including browsing records, number of clicks, click time, download records, favorite records, sharing records, to form characteristic database for a specific user.

The PageRank algorithm was proposed by Lawrence Page and Sergey Brin in 1998. It mainly determines the ranking (PR value) through hyperlinks in the network. The standard calculation is shown in formula (8):

\[
PR(u) = (1 - \alpha) + \alpha \sum_{v \in \text{Out}(u)} \frac{PR(v)}{\text{Out}(v)} \tag{8}
\]

This article combines user characteristic data and introduces an improved PageRank algorithm. Based on the correlation between \( PR(u) \) against Click(u), Time(u), and Content (u), we add the respective weights of Click(u), Time(u), and Content (u) into the formula of PageRank algorithm to obtain a new PR value calculation formula [7]:

\[
PR(U) = PR(u) \times \frac{\text{Click}(u) \times \text{Content}(u)}{\text{Time}(u)} \tag{9}
\]

In order to avoid the omission of user feature data collection, which leads to unavailable situations in the subsequent system expansion process, the amount of user characteristic data collected in this paper is larger than the amount needed by the improved PageRank algorithm.
4. Realization of Intelligent Search Engine for Oil and Gas Information

4.1. Using Gensim to Implement Word2vec Model Training

Word2vec is a model proposed by Google in 2013 that can convert words in languages into vectors. Using Word2vec for model training, words with similar meanings in the training set will be close to each other in the vector space, which can better express different words similarities and analogies. Word2vec has two models: Continuous Bag-of-Words (CBOW) and Skip-Gram [8]. The former predicts the target word based on the original sentence, while word is predicted by the target word with the latter model, which is used in this article.

In this paper, the third-party Python toolkit Gensim is used to implement Word2vec word vector training. The vector space dimension is set to 200 and the training window size is set to 5, along with the Hierarchical Softmax algorithm. The specific training steps are as follows:

1. Preprocess the corpus to implement one sentence per line, remove the Chinese punctuation marks in the text, and use the word segmentation tool HanLP to segment the Chinese corpus;
2. Convert the processed corpus into a sentence iterator, and iteratively return a list of words, using the LineSentence () method in word2vec.py;
3. Input the above processing results into the word2vec object of Gensim for training.

4.2. Knowledge Graph Storage Based on Neo4J

After the knowledge extraction is completed, the knowledge map needs to be stored persistently. In this paper, the popular open source graph database Neo4j is used for storage. Compared with the RDF storage form, graph storage is more general. Neo4j is a high-performance, highly available, and highly scalable NoSQL database. It provides multiple data import methods and graph query language Cypher. It supports knowledge graph visualization and various graph mining algorithms. It is conducive to the realization of knowledge reasoning and is applicable to the oil and gas field.

As shown in figure 4, the underlying data storage types of Neo4j are divided into nodes and relationships, both of which have corresponding properties and are stored in the form of key-value pairs to form a network structure. There can be several relationships between any two nodes, but a relationship has only one start node and one end node.

![Figure 4. Neo4j data structure.](image)

As shown in figure 5, after completing the import of the data, the Cypher language is used to operate and visualize the knowledge graph of the oil and gas field.

4.3. Integrate Elasticsearch to Index Massive Data

ElasticSearch is a distributed search engine developed based on the Java language. It can realize the rapid storage and search of massive data. Considering the complex characteristics of data in the oil and gas field, we need the rapid retrieval analysis of massive oil and gas information from different sources. On the basis of Neo4j, this article integrates ElasticSearch to improve the efficiency of data storage and...
retrieval. The combination of Cypher search and Elasticsearch search is used to meet the rapid search of oil and gas massive data.

Figure 5. Visualization of knowledge map in oil and gas field (partial results).

This article is based on the Neo4j-Elasticsearch plug-in to achieve integration and synchronization between Neo4j and Elasticsearch to maximize the use of Elasticsearch’s fast indexing capabilities. Neo4j mainly implements the storage of knowledge maps. Except for some hot data, it does not directly provide query services to users, but provides knowledge query services facing Elasticsearch instead, which can achieve the semantic analysis of query terms, relational queries, and ranking calculation of query results. Elasticsearch further queries and returns the final results based on the content returned by Neo4j. As shown in figure 6, in practice, combined with the topical content (such as natural gas supply and marketing) that users are currently paying attention to or researching, some hot data can be stored in Neo4j’s node data, which can be directly obtained from Neo4j when inquiring, in order to further improve the query efficiency.

Figure 6. Smart search function screenshot.

4.4. Search Accuracy Evaluation

We evaluate the intelligent search engine proposed in this article by calculating the precision and recall ratio. Four definitions for the binary classification problem are given as following: True Positives (TP) represents the amount of retrieval for related information. False Positive (FP) represents the amount of retrieval for non-related information, True Negative (TN) represents the amount of non-relevant information not retrieved, and False Negative (FN) represents the amount of relevant information not retrieved. The precision and recall ratio can be calculated as follows:

\[
Precision = \frac{TP}{TP + FP} \tag{10}
\]

\[
Recall\ Ratio = \frac{TP}{TP + FN} \tag{11}
\]
To benchmark the performance of our intelligent search engine, we selected 2000 related documents in the field of oil and gas as the test set, covering resources management, reserves management, mineral rights management, exploration and mining, and industry information. After data processing, each document contains an average of 3,000 words. Our intelligent search engine and the ElasticSearch full-text search engine are benchmarked. Based on the user’s habits of using the search engine, we examined the the top 100 search result, and calculated the precision and recall ratio of the two algorithms. The results can be seen here in figure 7.

![Figure 7. Comparison between precision and recall ratio.](image_url)

As shown in figure 7, from the above benchmark results, it can be seen that our intelligent search engine proposed in this paper is superior to the ElasticSearch full-text search engine in both Precision and Recall Ratio in terms of information retrieval in the oil and gas field, especially in using Recall Ratio as metrics, indicating its great superiority in the sensitivity of searching.

5. Conclusion

Based on the massive oil and gas professional information accumulated by the Center of Oil and Gas, this article established a knowledge graph that includes topics such as resource management, reserves management, mineral rights management, exploration and exploitation, strategic research, and industry information. Based on the knowledge graph, the knowledge model and intelligent semantic retrieval of oil and gas information has been realized. It has been proved by practice that compared with traditional search engines, our engine can better understand the user’s search intent and provide more accurate and user-friendly search results.

Facing the future, the knowledge graph of the oil and gas field needs to be further expanded and improved to ensure the comprehensiveness and accuracy of oil and gas information search. The mining and extraction of entity relationships need to be further studied and can be combined with deep learning to achieve knowledge reasoning. For the search engine, with the continuous increase of massive data, it is necessary to continue to optimize the search mechanism and strategy to ensure search efficiency. In view of the sensitivity of oil and gas information, it is necessary to further strengthen the hierarchical management of data to lay the foundation for the widespread application of intelligent search engines in the oil and gas field.

References

[1] Zheng W, Liang Z and Liang J 2014 Research on user intention oriented intelligent search engine framework New Technology of Library and Information Service 244 (3) 65-68.

[2] Yuan G, Chen S, Xin Y and Deng X 2011 Ontology construcion theory in oil fild of applied research Computing Technology and Automation 30 (3) 113-114.

[3] Chen W, Wu Y, Chen W and Zhang M 2018 Automatic keyword extraction based on BiLSTM-CRF Computer Science 45 (6A) 93-94.
[4] Guan J and Gan J 2007 Design and implementation of web search engine based on Lucene
Computer Engineering and Design 28 (2) 65-72.
[5] Chen Z, Lu Y and Liu H 2009 Chinese spelling correction in search engines based on n-gram
model Journal of China Academy of Electronics and Information Technology 3 (4) 324-325.
[6] Yin J, Wang Z, Li Q and Su W 2014 Personalized recommendation based on large-scale
implicit feedback Journal of Software 25 (9) 1953.
[7] Fang S 2012 Basd on user feedback PageRank algorithm Computing Technology and
Automation 31 (1) 90-92.
[8] Zhao M, Du H, Dong C and Chen C 2017 Diet Health Text Classification Based on word2vec
and LSTM Transactions of the Chinese Society for Agricultural Machinery 48 (10) 203.