Optimization of Data Insight Tool Based on Engineering Technology Data Governance Project in Ultra-deep Oil & Gas Fields

Qiang Zhang[1], Chun-lin Hu[2(✉)], Rui Chen[3], Ke-cheng Jiang[2], Xin Li[2], Nan Xiao[3], Qing-gang Yang[2], and Bing-bing Zhou[2]

[1] PetroChina Tarim Oilfield Company, Xinjiang, China
[2] PetroChina Kunlun Digital Intelligence Technology Co., Ltd., Beijing, China
huchunlin39@cnpc.com.cn
[3] PetroChina Tarim Oilfield Company, R&D Center, Xinjiang, China

Abstract. Based on the engineering technology data in the ultra-deep oil & gas fields, this paper utilizes data insight tool to identify and extract information from various types of data stored in documents with text or tables, which meets the needs of data governance project. If the document information is about text content, the natural language processing (NLP) method is directly selected for recognition; If the document information is a table, it is necessary to convert the table into a heterogeneous data table with Date-Frame format first by Python language, and then recognize and extract it. These two processing methods can successfully convert unstructured data to structured data, solving the problem of low accuracy and low timeliness of extracting information from different documents. The NumPy & Pandas learning with Python language and other algorithms/functions play an important role in building metadata models, labeling fields, and training backend algorithms of data insight tool structure. The target trained extraction model is very crucial to the identification and extraction of various information. Relying on this and later, the qualified data generated after steps of extraction of target documents, selection of matching data for review and multi-level audit evaluation will be marked with "EDG", which is the main data source of various professional databases of Tari Oilfield and the guarantee of the capacity and

22, IFEDC Organizing Committee.
This paper was prepared for presentation at the 2022 International Field Exploration and Development Conference in Urumqi, China, 16–18 August, 2022.

J. Lin (Ed.): IFEDC 2023, SSGG, pp. 55–64, 2024.
https://doi.org/10.1007/978-981-97-0272-5_5

quality of the data lake. Examples show that the data insight tool has strong adaptability, obvious optimization effects, and superior performance compared to other extraction tools. The development and application of data insight tool have significantly improved the identification and extraction ability for engineering data of ultra-deep oil & gas fields, improved the identification accuracy and extraction speed, and met the needs of data governance.

Keywords: Data insight tool · Information extraction and metadata

1 Introduction

With the development of computer technologies such as cloud computing and big data, people from all works of life are gradually realizing the value of data. In the process of ultra-deep oil & gas exploration and production, there is also a large amount of data, which is generated in different ways at different times and stored in the form of documents. There are not only geophysical, drilling and logging data, but also gas testing, geological structure, and downhole operation data, as well as scientific research report data. For documents from the same data source, there are also storage differences such as text, horizontal tables, vertical tables, and two-dimensional tables. Overall, it appears as a long collection and survival time, discontinuous, and mixed with other data, diverse storage types, complex document structure, and large and disorderly data volume.

In order to make full use of these data, it is necessary to treat them as a whole and convert all unstructured data into structured data, so as better serve the engineering technology data governance project of ultra-deep oil & gas fields, which has practical significance [1, 2].

Unstructured data refers to data with irregular or incomplete structure, no predefined data model, and inconvenient to be represented by the two-dimensional logical tables of the database, including office documents, text, images, HTML, reports, images, and audio/video information in various formats, while structured data, also known as row data, refers to data logically expressed and realized through the two-dimensional table structure. It strictly follows data format and length specifications, mainly utilizing relational databases for storage and management.

The conventional data structuring is achieved by constructing an information extraction model. The specific method is to manually mark the information to be extracted in the sample to obtain the training sample set, then select the appropriate supervised learning algorithm, train the model with the training sample set, and finally obtain the model for information extraction.

This method may not always be effective. For example, the sample and text to be extracted are in a fixed format, such as ID cards and invoices. When the format of the text to be extracted is very consistent with the sample, this method can obtain a model with high extraction accuracy. When there are differences in format between, the accuracy of model extraction decreases, and the larger the difference, the lower the accuracy. Although the accuracy of the model can be improved by increasing the number of samples, but an increase in the number of samples often means an increase in training difficulty. In the oil and gas industry, the processes of oil and gas exploration

and production are complex, and the documents generated by different periods/methods may have significant differences. Even increasing the number of samples cannot achieve the accuracy value required for practical applications [3].

Data insight is a method that can efficiently extract information, the tool equipped with corresponding devices and storage media, especially suitable for environments where there are differences in the format of samples and documents to be extracted, significantly improving the accuracy of information extraction models.

2 Features of Data Document in Ultra-deep Oil & Gas Fields

At the initial stage of data insight tool development, the unstructured data of engineering technology in ultra-deep oil & gas fields was briefly classified according to the type of data documents, and the following rules were summarized:

(1) The same type of data documents of different wells in different oilfields may have large differences, and similarly, documents of different data types also have different structures; (2) Within the same document, there may be both text and table content, and table documents of the same type may also contain horizontal, vertical, or two-dimensional tables; (3) The probability that data simultaneously exists in three types of text, table, and image in a data document is high. Among the total data of all types of documents, table data accounts for the largest proportion, reaching over 90%. The image data is about 8%, and text data less than 2%. (4) Table documents mainly include basic table, cross page table, two-dimensional table, and transposition table. Among them, cross page table accounts for the largest proportion, reaching over 60% of the total table documents. Transposition table is about 15%, while basic tables and two-dimensional tables are about 20% and 5%, respectively.

In response to the features of data document in the ultra-deep oil & gas fields, the key task of the data insight tool primarily solves the problem of extracting information from table data documents, followed by image and text information extraction.

3 Integral Development of the Data Insight Tool

The data insight tool typically consists of two parts: hardware and software. The hardware includes mechanical and electronic apparatus /devices or storage media used to support data conversion, storage, transmission, and communication responses such as computers, circuit boards, memories, and transceivers; The software includes corresponding networks, algorithms, sample sets, training models, databases, and data communication and control.

3.1 Hardware Composition and Network Architecture of the Data Insight Tool

Specifically, the data insight tool includes three functional blocks: data processing center, database, and user terminal. Each function block contains hardware and software to complete all functions one by one, that is, trigger user information extraction requests, generate metadata and target documents, collect and call training sample sets, generate

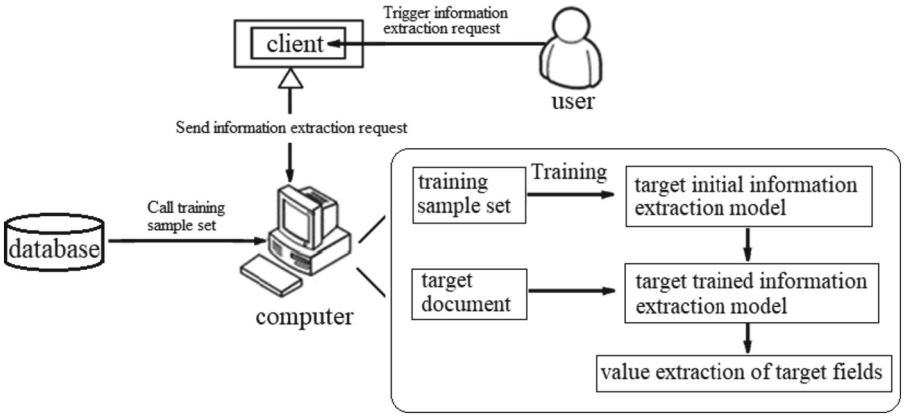

Fig. 1. Hardware composition and network architecture diagram of the data insight tool

information extraction models for target training, and extract the values of target fields. The hardware composition and network architecture diagram shown in Fig. 1.

The user terminal (shown as client in Fig. 1) is the port where users make requests and receive responses, and is the response center for data processing center exchange with the outside world. It is equipped with information extraction application software, which is used to receive information extraction requests triggered by users. Users can access the operation interface of the client or enter the website to extract web pages, sending information extraction requests to the data processing center.

The database (shown as database in Fig. 1) is the storage center of structured data, target documents and training samples/training sample sets. The data processing center calls the training samples from the database to form a training sample set according to the user's information extraction request.

The data processing center (shown as computer in Fig. 1) is the center for data reception and transmission, algorithm collection, training models, and communication control. It mainly includes various algorithms for constructing metadata models, training sample sets for completing data recognition and extraction, and executing instructions. It is the most core part of the data insight architecture.

The data processing center adopts a modular design concept and is divided into modules such as reception, training, determination, conversion, and acquisition. Briefly, the receiving module receives extraction requests to form initial metadata and target documents; The determination module obtains the training sample set corresponding to the initial metadata; The training module generates a target initial information extraction model and trains it to generate a target trained extraction model corresponding to the initial metadata model; The conversion module is used to determine and convert the file format of training samples; The acquisition module extracts target fields from the target document using an information extraction model based on target training.

3.2 Procedure of Information Extraction

The procedure of information extraction for data insight sequentially involves the following steps:

(1) Receive user triggered information extraction requests, generate target fields to be extracted based on their needs and selections, record the target documents to which these target fields belong, their document types, and document locations, and construct an initial metadata model.
(2) Select training sample documents that meet the conditions and convert their format into a unified and recognizable format. Collect them together to form a training sample set, and label the target fields and field values of all documents in the sample set.
(3) Referring the metadata model and initial information of the target to build a target initial extraction model, and configure the parameters in the model.
(4) Extract the labeled target fields and field values from each training sample in the training sample set, and use them to train the target initial extraction model.
(5) Set the threshold for extraction recall and extraction accuracy. If the target initial extraction model does not reach the threshold for extraction recall or extraction accuracy, adjust the model parameters and keep training with the sample set until these two thresholds are exceeded.
(6) Now the target initial extraction model can be considered as a target trained information extraction model and shines upon the corresponding metadata model.
(7) Using the target trained information extraction model to extract the target fields in the target document, the values of the target fields can be obtained.

4 Data Insight Analysis and Strategies

In order to accurately and quickly complete information extraction, specific analysis and testing were conducted on each process and step, and corresponding strategies were summarized.

Metadata and Training Sample Set Selection Strategy
A few of metadata and corresponding training sample sets are preconfigured in the database, so that documents of different structural types correspond to different metadata. When a user requests information extraction, appropriate metadata and training sample sets are filtered out, and the types of samples are then distinguished. During the training process of the target trained information extraction model, there is no need to filter the training samples anymore, just learn all the samples in the selected training sample set. Therefore, the target trained extraction model can have a high accuracy, and requires a small number of samples, resulting in fast extraction speed.

Format Transformation Strategy for Training Sample Documents
Check the file header flag to clarify the format of the training sample. If the sample is in PDF format, use common format conversion tools to directly convert the sample format to Excel format; If it is a Word document, first convert the format of the Word document to PDF format, and then convert it to Excel format. This strategy can unify

the document format of training samples and ensure that the content in the document will not be disordered after format conversion. Meanwhile, compared to other methods such as converting each page of the sample document into images and then performing image recognition, format conversion can improve the speed of model training and usage, especially when the number of pages in the document is large, the speed improvement is more significant.

Extraction Strategies for Text and Table Documents

Using NumPy and Pandas in Python language, convert all data in the document into Date-Frame format data, and separate the text content and table content in each training sample by presetting the data length of heterogeneous data tables.

Various algorithms are set in the target initial information extraction model, such as N-gram language model for table name query, cosine similarity algorithm for obtaining similar table names, OCR for obtaining information in pictures, named entity recognition, short-term memory LSTM, conditional random field CRF and Glove algorithm, etc. for DataGrid Cell segmentation.

If it is text content, use natural language processing NLP to extract text content. Sequentially perform extraction methods and strategies such as word segmentation, word embedding, and named entity recognition, mark the extraction relationship between the target field and its field values, and extract the feature values of the target field. If it is a table content, use the concatenation function to concatenate the row data in the worksheet into a string, use the judgment function to determine whether the tables in each worksheet are continuation tables, and use the preset deletion function to delete unmarked irrelevant items and items without data in the table.

These algorithms and functions have improved the efficiency of learning and training, reducing the probability of errors.

Difference Sample Determination and Strategy

When there are difference samples in a training sample set, specific analysis pointing at the target fields of the difference samples is necessary.

If the difference is only about in the text feature values of the difference target field, it indicates that an error may have occurred during the text label process, and the difference target field in the difference sample needs to be discarded. If the differences happened both in position and contextual feature values, it indicates that the position of the target field in the difference sample may be different from that in the other samples. The feature values of the difference target field can be stored as the second feature values of the target field in the target initial information extraction model, and keep going on extracting target field and its field values from the training sample. This strategy can store multiple feature values for the same target field, and then calculate the feature value by the way of weight average or linear regression, which has a significant impact on sample reinforcement. If the differences simultaneously appear in position, text, and contextual feature values, it indicates that there may be errors in the selection of training samples and the difference sample need to be discarded.

These strategies above could ensure that during the training process, the initial target information extraction model has fewer training samples and the most types, making it more accurate and faster.

5 Application Examples

In order to verify the advantages of data insight tool in governance time, the data insight tool and other popular extraction tools were selected to synchronously extract logging data of ultra deep oil fields with "inclination" as the target field. The comparison of extraction effects is shown in Fig. 2 and Fig. 3. 50 Wells are selected in Fig. 2, while 500 Wells selected in Fig. 3 to demonstrate the extraction capability in more details. Here are three extraction tools, namely: Data Insight Extraction in the first row, Baidu Table Analysis Extraction in the second row, and Manual Extraction in the third row. In Fig. 2 and Fig. 3, blue represents the time for labeling samples, orange represents the training time, and gray represents the extraction (governance) time. The horizontally distributed numbers "0", "200", "400", "600", "800", as well as "0", "2000", "4000", and "6000" in the figures represent the time expenditure for governance, in minutes. Their specific values are related to the size of the "Inclination" data.

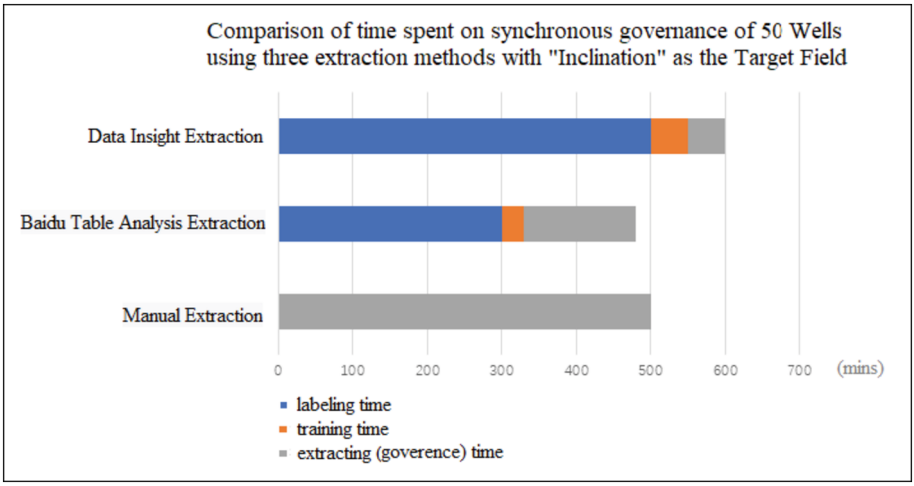

Fig. 2. The data insight tool and other popular extraction tools were selected to synchronously governance the logging data of 50 Wells in ultra deep oil fields, with "Inclination" as a target field

From Fig. 2 we can see that when the number of Wells is small, such as less than 50 wells, the Manual Extraction is the fastest method due to the time expenditure for data insights and other extraction tools to label samples. As the number of Wells increases, such as 500 wells shown in Fig. 3, the advantages of artificial intelligence are reflected. In Fig. 3, the governance time using Data Insight Extraction is the shortest, only half of the time used by Baidu Table Analysis Extraction, and 1/5–1/6 of the time used by Manual Extraction.

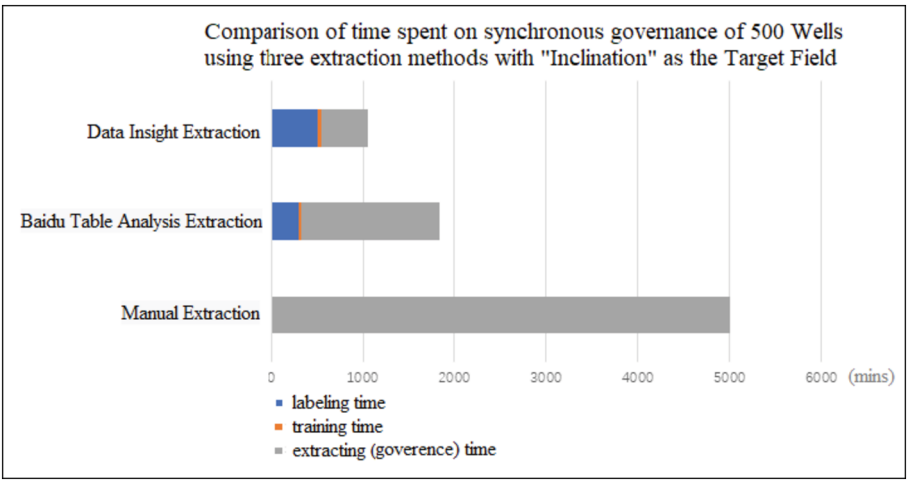

Fig. 3. The data insight tool and other popular extraction tools were selected to synchronously governance the logging data of 500 Wells in ultra deep oil fields, with "Inclination" as a target field.

It can be foreseen that for the huge amount of data in the ultra-deep oil fields, data insight tool will have significant advantages in governance time.

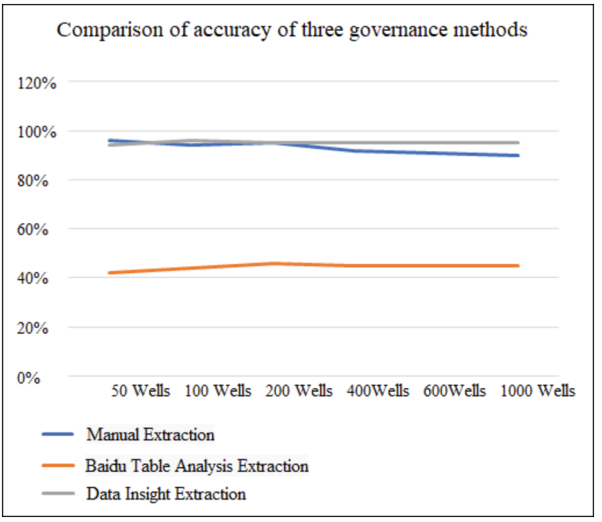

Fig. 4. The data insight tool and other popular extraction tools were selected to synchronously governance the same number of Wells of the logging data in ultra deep oil fields, with "Inclination" as a target field.

For the sake of verifying the advantages of data insight tool in governance accuracy, data insight tool and other popular extraction tools were also selected to synchronously

extract data in the same number of Wells in the logging data of ultra deep oil fields with "Inclination" as a target field. The comparison of extraction effects is shown in Fig. 4. Here the blue line represents extraction accuracy of Manual Extraction, orange line represents extraction accuracy of Baidu Table Analysis Extraction, and gray line represents the accuracy of Data Insight Extraction. The horizontal characters "50 wells", "100 wells", "200 wells", "400 wells", "600 wells", and "1000 wells" in the figure represent the number of governance Wells, while the vertical characters represent a sketch map of accuracy percentage. Likewise, their specific values are related to the size of the "Inclination" data.

From Fig. 4 we can see that both Data Insight Extraction and Manual Extraction have an ideal information extraction accuracy (as shown by the gray and blue lines) of over 90%, regardless of whether the number of Wells is small (such as 50 Wells) or large (such as 1000 Wells). However, as the number of Wells, meaning the amount of governance data increases, the extraction accuracy of Manual Extraction (as shown by the blue line) slightly decreases. While using Baidu Table Analysis Extraction, regardless the number of governance Wells large or small, the extraction accuracy (orange line) is below 50%.

The above examples indicate that the data insight tool has high extraction accuracy and fast extraction speed, and their information extraction ability far exceeds that of other common products on the market, meeting the audit requirements of business departments.

6 Conclusion

(1) The metadata of data insight tools covers a wide range, including different target fields that will be extracted, different documents storing these target fields, and their respective types and locations. These documents are all technical data generated during the ultradeep oil & gas exploration and production. When users trigger information extraction requests, multiple metadata can be generated simultaneously for backup. The target trained model is trained using a training sample set, and the type of samples in the training sample set is the same as that of the target document. All training samples are labeled with the target fields and their field values. Therefore, there is not much difference between the format of the target document and the format of the samples, and a high accuracy information extraction model can be quickly obtained.

(2) The data insight tool converts unstructured data into structured data by means of establishing metadata models, marking fields, background algorithm training etc., which plays a key role in the construction of various professional databases in ultra deep oil & gas fields. Qualified data generated after the steps such as extracting target documents, selecting matching data for review and multi-level audit and evaluation will be marked with "EDG", which is the main source of the database of ultra deep oil & gas fields and the guarantee of the capacity and quality of the data lake [4].

(3) The example verification shows that both the accuracy and speed of data insight tool used for information extraction remain at a high level, which can meet the audit requirements of business departments. It has strong adaptability, obvious optimization effects, and superior performance, far surpassing other common extraction tools in the market.

(4) From design to development, the data insight tool absolutely conforms to the current situation of technical data of ultra deep oil fields, and its accuracy and governance speed meet the requirements of practical application. It is of great significance to promote the implementation of engineering technology data governance project of Ultra Deep Oil & Gas Fields.

References

1. Gao, H., Li, J., Cao, Y., et al.: Construction of expeditor seismic data management system in Tarim oilfield. Petrol. Geophys. Explor. **43**(supplement1), 182–196 (2008)
2. Lu, Z., Chen, R., et al.: Discussion on real time data quality management of drilling engineering in Tarim Oil Field **31**(3), 118–121 (2020)
3. Li, X.: Tarim oilfield surface engineering database system. Oil Gas Field Surface Eng. **21**(2), 118 (2002)
4. Luo, C., Tang, Y., et al.: Application and prospect of collaborative research on E&P dream cloud in Tarim oilfield. China Petrol. Explor. **25**(5), 50–55 (2020)