Research and Development of B/S-Based Data Mining System for Petroleum Information

Qian Zhang[1(✉)], Shiyun Mi[1], and Xinchao Liu[2]

[1] Research Institute of Petroleum, Exploration and Development, Beijing, China
{zhangvqian, symi}@petrochina.com.cn
[2] China Petroleum Technology & Development Corporation, Beijing, China
liuxc@cptdc.cnpc.com.cn

Abstract. With continuous enhancements in level of informatization, petroleum industry accumulated massive exploration and development data. How to clarify underlying messages through data mining to provide predictive, decisive and regular information is very urgent problem at the present stage. During implementation of the concerned research, prevailing data mining technologies have been reviewed to develop B/S-based petroleum data mining system with consideration to actual conditions in petroleum industry. The innovative system include designated database, algorithm invoke and result presentation sections. Through application of data fusion techniques, and with consideration to specific mining demands of petroleum industry, the System created designated database; by constructing B/S-based algorithm invoking and presentation platform, MATLAB, R, PYTHON, and other prevailing mining algorithmic languages can be invoked and relevant results can be presented.

Keywords: B/S-based · Data mining system · Petroleum information

J. Lin (ed.), *Proceedings of the International Field Exploration and Development Conference 2018*, Springer Series in Geomechanics and Geoengineering, https://doi.org/10.1007/978-981-13-7127-1_68

1 Introduction

With fast progresses and technologies, levels of informatization of oil/gas enterprises became higher and higher [1]. At the same time, these enterprises accumulated more and more exploration and development data through daily operations of various kinds. But it is worth noticing that these massive electronic data generated through daily operations can hardly be used for decision-making processes of these enterprises. With fast progresses in information technologies, statistics, database, machine learning and other technologies have been used jointly to process massive data through data mining to acquire useful information. Conventional statistics are often simple quantitative statistics, whereas large quantities of previously unknown, but potentially useful knowledge petroleum data were often neglected. The reason for this is conventional statistics and inquires can only derive simplified results similar to quantitative summation without satisfactory presentation of data distribution and other features. Consequently, these statistics often have value of data statistics without knowledge mining significance. Through in-depth mining of these massive petroleum information, important statistics for assessments of oil and gas resources, exploration risk assessments and decision-making processes of petroleum enterprises can be derived. Under such circumstances, in-depth mining of database of petroleum information system to extract useful data as much as possible can be seen as a technical subject of great importance.

In existing technologies [2–4], the process from data preparation to data mining algorithm and further to assessment algorithm can be achieved only through coding. In other words, existing technologies cannot provide algorithm researchers with unified, interactive, and high-efficiency interfaces for algorithm realization sand tests. It is quite challenging to develop data mining systems by using existing technologies. Currently, there are multiple data mining tools with significantly different functions and performances. Popular data mining tools include IBM Intelligent Miner, SAS Enterprise Miner, SPSS Clementine, and Weka. To develop data mining systems specifically for petroleum industry, it is necessary to consider specific conditions of petroleum enterprises (such as background, data, mining technology, etc.) to construct mining systems suitable for petroleum data. Data mining platform technologies can provide clients with data processing, high-efficient expandable data mining algorithm, data presentation and other functions. In this way, investments of petroleum enterprises in data mining can be minimized. In addition, it is possible to accelerate promotion of data mining operations, to shorten R&D periods and to enhance profitability of relevant products.

2 Development of Data Mining Systems for Petroleum Information

Petroleum data have relatively long life cycles. Drilling data acquired decades ago are still highly useful nowadays to highlight conditions related to structures and formations. Through data mining, it is possible to extract values from data, to guide

deployment of exploration and development wells and to enhance performances of exploration and development operations. Generally speaking, petroleum data have following characteristics: (1) Petroleum data have massive volumes with high dimensions and high relevance. For example, drilling operations can generate drilling engineering data, formation data, together with accompanying logging, well logging, analyses and testing data. Massive geologic and engineering parameters may interact with each other to affect productivities and other core data; (2) Petroleum data are often incomplete. Different parameters may be acquired at different frequencies and these data may cover different extents. In addition, difficulties in data acquisition may frequently lead to missing of data; (3) Petroleum data have a large number of types. Petroleum data include character data, such as formation lithology, basin classification, structure and others. Petroleum data also include data of value type, such as productivities oil/gas wells, volumes of water injection, reservoir attributes. In addition, there are logical data, and data of range types (0–1). With consideration to specific features of petroleum data, it is necessary to redesign and unify petroleum data standards and codes of interfaces from infrastructure to analysis application. By studying data model, system architecture and specific technologies, framework for data mining systems of petroleum data with "Decision-making driven by data" can be established eventually.

2.1 Platform Framework of Data Mining System for Petroleum Information

To achieve effective mining of petroleum exploration and development data, general procedures for data mining must be followed. At the same time, specific features of petroleum data shall be analyzed thoroughly. With combination of geologic and engineering factors, development and realization of algorithms can be achieved. Petroleum exploration and development involves data related to geology, reservoir attributes, productivities of oil/gas wells, engineering procedures and various subjects. Relevant analyses may involve key indexes of various types, but majority of data mining technologies are based on following classical principles. The plateform framework for data mining of petroleum information was shown in Fig. 1.

To derive prediction results of shale gas EUR abundance by using porosity, shale thickness, buried depths, TOC contents, maturity of organic matters, and other factors. First all, it is necessary to provide data source of shale gas fields with such attributes. Then, it is necessary to construct multiple mining models, such as regression models for shale gas EUR abundance and porosity, and TOC contents, or regression models for shale gas EUR abundance and shale thickness, buried depths, organic matters' maturity and others. Step 3, it is necessary to assess these regression models to identify the most accurate one. Step 4, it is necessary to clarify correlation among shale gas EUR abundance and the six influencing factors to highlight weighting of these factors. Step 5, algorithm development module of the application data mining platform will be used for development of relevant algorithms. Finally, results related to prediction, classification, clustering and other aspects can be presented.

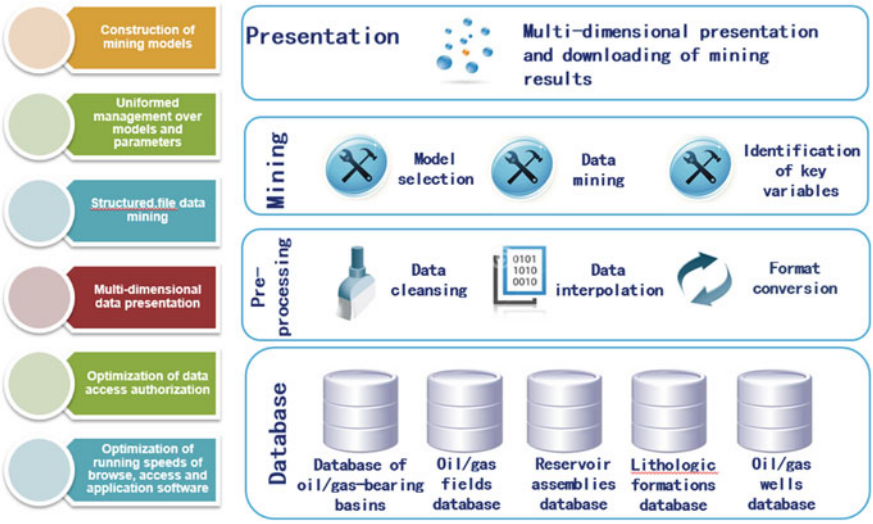

Fig. 1. Platform framework for data mining of petroleum information

2.2 Preprocessing of Data

Since petroleum data have a great varieties of origins and significantly different data quality, these data shall be cleaned, integrated and converted in accordance with their specific features. Under such circumstances, it is necessary to extract required data for preprocessing before introduction to the data mining system. Key functions of the module include data access to import data from data source and convert them structured data with unified format. These functions may support ASCII texts with separator, Excel files and others. Through ODBC, connections can be made with the Oracle database. In addition, the module has the function of data cleansing to process missing values through deletion, replacement, interpolation, and to detect abnormal multidimensional data. Correspondingly, processing interfaces for missing values and noise data have been developed. Key functions of processing interface for missing values include: neglect the data record, manual completion, utilization of a global constant to fill missing values, utilization of average attributes to fill missing values, application of the average value of all samples in the same category of the specific element and utilization of the most probable values to fill missing values; Key functions of processing interface for noise data include data classification and replacement of original data by using rational values to remove noses from the original data. In addition, the interface can be used to determine concept hierarchy of relevant data in accordance with "distance" and other criteria to transmit to higher hierarchy. In some other cases, regression model, or model predicted values can be used to replace original data.

2.3 Algorithm Invoke

Generally speaking, petroleum data conform to specific pattern of distribution, such as logarithm normal distribution, normal distribution, triangular distribution, or others. Accordingly, to construct mining model some prior information can be introduced on bases of the universal model to construct more rational and more accurate models. Through hybrid programming technologies, the system can be used for algorithm invoke to call MATLAB, R, PYTHON, and other prevailing mining algorithmic languages. At the same time, algorithm editing and debugging functions have been developed to support Java debugging and editing capacities. In addition, other functions such as single-step execution, breakpoint setting, stack monitoring, local variables, grammatical errors positioning, secondary development of existing algorithms and others.

2.4 Presentation of Data Mining Results

Online Analytical Processing (OLAP) and other diverse front-end analyses and presentation tools can be used for analyzing and processing mining results. Eventually, smart deductions, prediction, optimization and auxiliary decisions can be made for exploration and development operations. During the course, OLAP presentation, 2D charts, 3D charts, and other data presentation capacities for mining results. OLAP representation can be used for in-depth analyses and presentation of relevant mining results. Basin data can be classified in accordance with years of discovery, basin types and other dimensional data to present data related to oil and gas resources, the number of basins and some other data. At the same time, data related to oil/gas fields can be classified in accordance years to highlight productivities, recovery rates, recoverable reserve and other information related to such oilfields. Furthermore, lithologic formation data can be classified in accordance with types to highlight the quantity, ages, reserves and other information of such basins. In this way, resulting charts may have no fixed pattern. Instead, researchers may interactively propose presentation demands according to specific demands, whereas the system may response in real time to generate required charts.

3 Application of the Data Mining System for Petroleum Information

Data mining systems for petroleum information have been deployed for processing of data from 425 basins, 27,980 oil/gas fields and 693,920 oil/gas wells.

3.1 Analyses and Presentation of Distribution of Formation Systems in Accordance with Data from Oil/Gas Fields Worldwide

Jointly with spatial–temporal analyses, systematic analyses have been conducted for 27,980 oil/gas fields worldwide to highlight features in distribution of these oil/gas fields to provide necessary supports for hydrocarbon exploration. During the course,

spatial–temporal analyses were used to analyze total recoverable reserve, geological positions, geological age of reservoirs, buried depths, reserves, and other parameters of these oil/gas fields. Eventually, oil/gas fields worldwide can be classified in accordance with their geological ages and buried depths. In addition, these patterns can be presented through data visualization techniques (Fig. 2).

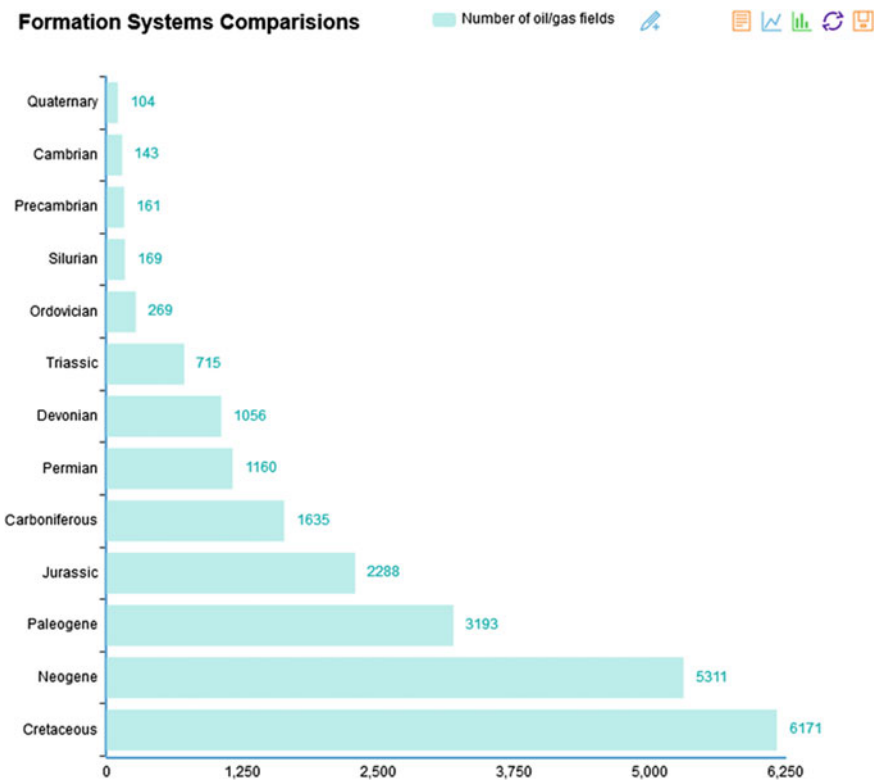

Fig. 2. Distribution of oil/gas fields worldwide in accordance with formations systems

Relevant results show oil and gas resources are highly enriched in some formations, whereas other layers may hardly contain any such resources. In earlier stages of exploration, quantitative relationships in distribution of discovered oil and gas resources in different formation systems were used to predict the number of oil/gas fields to be discovered in specific formation groups. In addition, targeted drilling operations were conducted in accordance with identified patterns. To determine distribution pattern of oil/gas fields information systems worldwide, data from 27,890 oil/gas fields worldwide were used in the concerned study. A few peaks can be observed in distribution pattern of distribution of oil/gas fields in global formation systems. These charts were generated with series as the minimum formation unit. During assessments of quantitative relationships in distribution of oil/gas fields, these

peak values show: oil and gas resources distributed predominantly in Tertiary, Mesozoic, Paleozoic formations of Cenozoic system. In addition, volumes of resources reduced information systems with elder ages. Figure 1 shows quantities of oil/gas fields in reservoir formations with different geological ages with series as the minimum formation unit. Distribution patterns of oil/gas fields show Cretaceous formations contain the largest number of oil/gas fields, followed by Late Tertiary formations. In Mesozoic formations, the number of oil/gas fields in Triassic formations is the lowest. As far as geographic locations are concerned, Mesozoic oil/gas fields distributed predominantly in Middle East, West Siberia, North America, West Europe, Australia and some other regions.

Oil/gas fields distributed predominantly in Mesozoic formations in Middle East due to development of an abnormally wide shallow-sea continental shelf on the northeastern passive continental margins of the Mesozoic Arabian Plate. Huge area of the shallow-sea area led to deposition of sediments with extremely high uniformity. Horizontally, potential generation, preservation and capping formations displayed continuous distribution of lithologic features in vast extents.

3.2 Multiple-Dimension Presentation of Statistics Related Historical Success Rates of Exploration Wells in Different Basins in the World

Success rates of exploration wells are subject to impacts of geologic types, exploration techniques, standards used, understanding to petroleum generation and distribution, specific features of petroleum and various other factors. The application was based on 693,920 oil/gas wells worldwide. Divided in accordance with basins, the data can be used to count the number of dry holes and oil/gas wells to analyze and calculate success rate of exploration wells. First of all, preprocessing of data shall be performed to remove useless and invalid data. Then well data will be counted again in accordance with the extents of specific basins to highlight success rates of explorations in the basin historically.

Relevant results show according to available data, success rates of exploration wells distributed within 0–0.7 globally. The highest success rate of exploration wells can be observed in Malay Basin in Central and Southern Asia. Failed wells were counted in two aspects: well depths and basin types. Relevant data show wells with medium depths (500–2000 m) have the highest failure rates, followed by wells with significant depths. These features can be observed in basins of all sizes. As far as resource abundances in these basins are concerned, basins with higher resource abundance also involve higher success rates of exploration wells; with decreases in resource abundance of basins, failure rates may increase dramatically (Fig. 3).

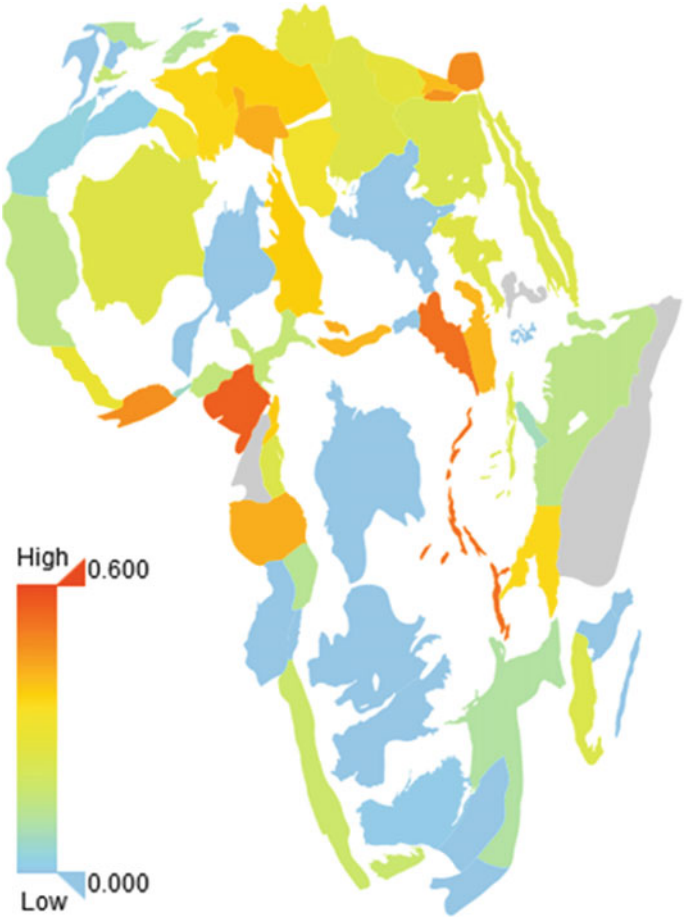

Fig. 3. Success rates of exploration wells in different basins in Africa

4 Conclusions

With consideration to unique features of petroleum data, researches have been conducted to establish overall framework for data mining platforms for petroleum information. Mining models can be constructed rapidly for different mining objectives. The system can be used for: (1) Fast invoke (or secondary development) corresponding data mining algorithm, or exploration and development operation models to generate data mining applications in accordance with specific dements of clients; (2) Detect abnormal conditions in various indexes generated during hydrocarbon exploration and development; (3) Correlation among various attributes and productivity elements in these oil/gas fields; (4) Highlight risk alerts and optimized cost control for exploration and development activities in overseas oil/gas fields. Eventually, smart deductions, prediction, optimization and auxiliary decisions can be made for exploration and development operations.

Acknowledgements. This work was supported by National Key Scientific and Technological Project "Research on Global Petroleum Resources Data System (2016ZX05029004)".

References

1. Tan C, Zhang H, Ma Y. Research and application of big data mining system for oil-gas production. Peak Data Sci. 2016;5(1):49–53.
2. Li Y, Yao D, Li Z, Hou J. Research on data mining system for big data in medical field based on R platform. J Harbin Univ Sci Technol. 2016;21(2):38–43.
3. Zhang Q, Mi S. A geostatistical approach to finding relationships between reservoir properties and estimated ultimate recovery in shale gas system. In: Proceedings of the international field exploration and development conference; 2017.
4. Zhang Q, Xu X, Mi S. A generalized p-value approach to inference on the performance measures of an M/Ek/1 queueing system. Commun Stat-Theor Methods. 2016;45(8): 2256–67.