ISSN 2320 -2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11, Pages : 01-06 (2014)
Special Issue of ICCECT 2014 - Held during 01-02 December, 2014,Bangkok, Thailand

Research on semantic-based domain data integration technology

Xin Liu[1], Changjun HU[2], Yang Li[3], Lina Jia[4]

[1]University of Science and Technology Beijing, Beijing, China, ustb.liuxin@gmail.com

[2]University of Science and Technology Beijing, Beijing, China, chungjin.hu@gmail.com

[3]University of Science and Technology Beijing, Beijing, China

[4]University of Science and Technology Beijing, Beijing, China

Abstract: For Mass, distributed, heterogeneous data sources of petroleum engineering, this paper presents an petroleum engineering semantic-based data integration technology based on building global semantic mode. PSDT builds a global semantic data model which appropriate to the domain of Petroleum engineering by ontology extraction, ontology mapping, ontology merging, ontology evolution and constraint reasoning for domain data model. Users and upper applications can have a direct access to underlying complex data sources through the global semantic data model. This study integrates the Ontology to build the global semantic data model for the distributed, heterogeneous and complex semantic correlation data source and to provide comprehensive, real-time data services. Experiments have shown that this method is feasible and effective.

Key words: domain data engineering; ontology; data semantic integration; distributed data processing.

INTRODUCTION

Oil and gas production is an important domain of petroleum exploration and development, Oil and gas well production design, decision analysis, diagnosis and management is the key to improve the production efficiency, reduce cost and improve the benefit. Optimal design on oil and gas well production system involves a large amount of data, including production data, well structure, equipment data, geological structure, seismic data and reservoir data. The data has huge quantity, many different types and complex relationships. The features of the data are as follows[1]:

•Distribution: Oil field is composed of a number of exploration institutes, oil production plants, geophysical research institutes and other units. Different units collate, collect, process, apply and analysis various types of data, and store corresponding data in their own database. It means that different types of data are stored in several different physical databases.

•Heterogeneity: Each database has its own specialized data structure and naming conventions, leading to four kinds of heterogeneity consist of system heterogeneity, syntax heterogeneity, structure heterogeneity and semantic heterogeneity:System Heterogeneity means operating system and hardware environment of data are various in different oil fields. Syntax Heterogeneity indicates that oil fields take different storage methods for different types of data. For example, some data are stored in relational databases, and some are stored in forms of text files. Structure Heterogeneity means that different oil fields represent the same type of data with different data schemas. Semantic Heterogeneity mainly intends that different words with the same meaning or the same words have different meanings.

•Instantaneity: Petroleum engineering data are dynamic, updated in real time, and in critical

ISSN 2320 -2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11, Pages : 01-06 (2014)
Special Issue of ICCECT 2014 - Held during 01-02 December, 2014,Bangkok, Thailand

instant need. Each oil field creates a large amount of production data every day, constantly updated basic data and regularly updated equipment data. So it is important to ensure the real-time of data for upper applications.

•Complex semantic associations: It mainly refers to the complex associations between different data. For example, the liquid production is equal to the sum of the water production and the oil production. Another one is the oil production divided by the liquid production is the water cut.

The unique character of petroleum engineering domain data brings a big challenge to traditional data management methods. On one hand the data storage of each oil field company is dispersed, the data lack of logical organization relationship, each oil field company has different data storage model and Data naming rules and data management model. Thus it is urgent to establish a global semantic data model which is suitable for multiple oil fields to achieve the unification of data management platform. On the other hand, data of oil companies are considerable autonomy, which increases the difficulty of data exchange and sharing. But data from different professional databases are increasingly need to work together to support upper applications of the domain. So semantic data integration and building uniform interfaces directly accessing to the underlying data resources is of great significance.

RELATED WORK

As the complexity of data leads to a new challenge for traditional data management, it is of utmost importance to generate a new way of data integration.

PA Bernstein et al. [2] survey and research the current enterprise data integration methods, steps and tools, points out that each step of the current information integration method need human intervention because of the complexity of the integration steps. It suggests more possibility of automation. Apparently, their work is not for specific domain of data services, particularly in the petroleum domain.

In October 1990, petroleum companies like BP Exploration, Chevron Corporation, Elf Aquitaine, Mobil Corporation, Texaco Inc sponsor Petrotechnical Open Standards Consortium (POSC) to solve the problem of computer standardization of oil and gas exploration and development[3]. Epicentre data model is one of the essential criteria of POSC. Epicentre is the oil data model in international standard, but it is not entirely suitable for E&P industry in our country[4]. To be in step with international applications, we should come up with petroleum data standards with characteristics of our country on the basis of Epicentre.

In view of distributed data integration, the circumstances of rich data lead to the emergence of technology of data warehouse and multi-database. Relational data is best served by Localized\Distributed mapping, middleware, data extraction, ETL (Extract-Transform-Load) or other technologies. About semantic integration, some technology like domain ontology and concept semantic layer can shield the difference among the distributed heterogeneous database and implement the semantic-based uniform resource access service.

Domain data are the foundation of conducting scientific research work. It is a key aspect to build data models based on semantic and carry out data integration and applications in specific fields[5].

THE IMPLEMENTATION OF PSDT

Architecture

PSDT provides a rich semantic view of the underlying data and interfaces enabling users and upper applications to access data. The architecture of PSDT is shown in Figure 1.The bottom of the architecture is data sources storing in different databases, such as A1 production DB, A2 production DB, Downhole DB etc. The middle layer is local ontologies extracting from

ISSN 2320 -2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11, Pages : 01-06 (2014)
Special Issue of ICCECT 2014 - Held during 01-02 December, 2014,Bangkok, Thailand

the data sources below. And then, the global ontology is formed as the result of combining local ontologies in the upper layer.

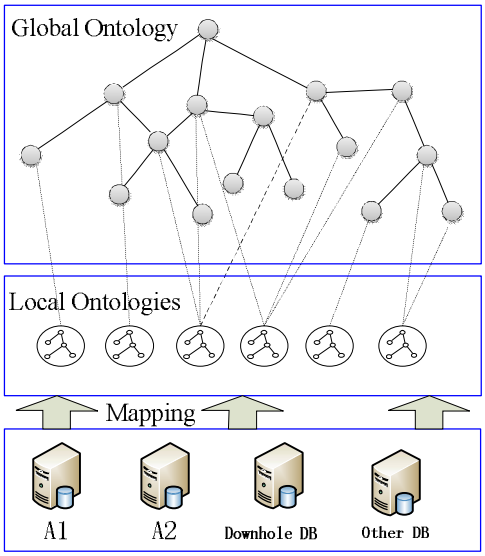

Fig 1. Architecture of PSDT

Construction of global ontology

Petroleum engineering domain databases are widely dispersed over the country. These databases always have different storage schemes and data structures and the data in these databases exist complex association relationship between them, so the data is not good for the use for domain experts and the upper application. Therefore, it is an urgent need for a data integration system which can hide the heterogeneity of data resources, so that the users can access to the underlying data through the domain global ontology.

Adopting a hybrid strategy to build the global ontology is a worthwhile exercise. For one thing, a top-down approach is used to filter the demand data. Entities, relationships and attributes between data entities can be got by organizing and classifying the data. For another, take a bottom-up method to build local ontologies, which are results of extracting schemas of databases and items of synonym list. And then the global ontology is established according to ontology evolution, ontology mapping and imposed semantic constraints.

From relational database to local ontologies

Since the majority of petroleum engineering data are stored in relational databases, we are here to study mapping from Relational Database to OWL ontology[6].

A relational database is composed of a set of relational schemas, including basic table structures and integrity constraints. An OWL ontology consists of properties, classes, axioms and individuals[7]. The mappings between data models and ontology, classes and properties are considered in this step.

The synonym list of petroleum engineering is built by domain experts and DBAs by reference to exploration and development handbooks of oil fields. The synonymous items with different names and same meaning in the handbooks are gathered together in the synonym list to solve the phenomena of semantic heterogeneity.

Based on the schemas of tables in the specialized databases, we analyze characteristics of tables and constraints between tables, and then define a petroleum-engineering data source ontology (PDS-On), which maps synonyms in the synonym list and schemas of tables to classes and properties in the ontology. The local ontologies can be generated automatically through the program. Getting innovations from Relational.OWL[8][9], OWL-RDBO[10] and Pro/Innovator[11], we design PDS-On to describe tables, columns relations of tables and synonymy. Then extraction rules are defined as follows:

Rule 1. Convert tables and columns in databases to classes PDS-On: Table or PDS-On: Column (owl: Class), which express main concepts of the domain.

Rule 2. Hierarchical relationships between tables and columns are presented by PDS-On: hasParent and PDS-On: hasChild (owl: ObjectProperty) with owl:inverseOf constructs. PDS-On: hasChild has a direction from domain Table to range Column, while PDS-On: hasParent has an opposite direction.

ISSN 2320 -2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11, Pages : 01-06 (2014)
Special Issue of ICCECT 2014 - Held during 01-02 December, 2014,Bangkok, Thailand

Rule 3. Relationships between columns in one table are presented by PDS-On: hasBrother, which is defined in owl: ObjectProperty.

Rule 4. If a column C in table A is the foreign key to table B, PDS-On: hasChild represents the foreign key constraint, from domain column C to range Table B, while PDS-On: hasParent is the reverse semantic association.

Rule 5. Datatype Properties of classes are defined, such as PDS-On: isPK, PDS-On: isFK, PDS-On: isNullable, PDS-On: dataType, to describe the primary key, the foreign key, nullable and data type of the individual.

Rule 6. Extract the items which express the same meaning from the synonym list to convert into classes, and the relationships between classes are defined as PDS-On: hasSynonymy, which is built in owl: ObjectProperty.

The process of ontology extraction, which is shown in Fig 2.

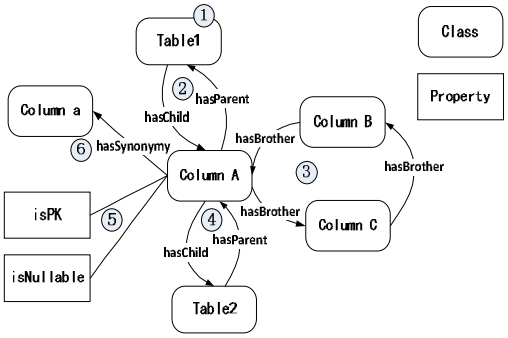

Fig 2. The steps of local ontologies extraction

The number in Fig 2 corresponds to the rule number. No.1 indicates that convert table names Table1 and Table2 and column name Column A into classes Table1, Table2 and Column A. No.2 means the relational schema of Table1 and Column A is turned to a parent-child relationship in the local ontologies. No.3 is converting the two columns Column A and Column a from the same table into a hasBrother relation. No.4 represents that the foreign key constraint of Column A in table Table1 and table Table2 is converted into a parent-child relationship. Rule 5 defines datatype properties of class Column A, while Rule 6 extracts synonyms of Column A from synonym list and defines relationships between synonyms as hasSynonymy.

In this paper, according to the definition of semantic association relationship, we consider some more detailed relationships like parent-child relationship, sibling relationship and synonymous relationship.

From local ontologies to global ontology

From local ontologies to the global ontology is divided into two steps, the evolution of the local ontology and local ontologies transform into the global ontology. The evolution of the local ontologies means if two classes have different parent node describe the same type of information, that is, the two classes correspond to different attributes of one entity formed in the step of data consolidation, the two classes evolve to a relation of hasBrother, parents of the two classes evolve to a relation of hasSynonymy[12].

When the local ontologies combining, there must exist definite relationships between the local ontologies. The global ontology can be built by mapping corresponding local ontologies. First analyzing the relationship and the schemas of different databases in the domain between the two local ontologies that have the same class. Then we can establish the two ontologies by certain rules. The parent node of the class with isPK property is mapped to the subclass of the class without isPK property.

After local ontologies combination and local ontologies evolution, a global ontology can be combined by the local ontologies.

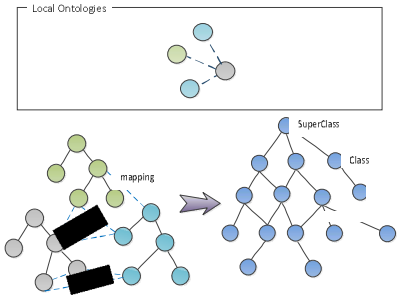

Fig 3. The process of building global ontology

ISSN 2320 -2602

International Journal of Advances in Computer Science and Technology (IJACST), Vol.3 , No.11, Pages : 01-06 (2014)
Special Issue of ICCECT 2014 - Held during 01-02 December, 2014,Bangkok, Thailand

Add semantic constraints

Some semantic constraints are given to strengthen the relationships of terminology. By using the concepts, inference engine can deduce and reason the global ontology to reorganize the concepts. Therefore we can get the implied semantic information and provide value-added services to users. Semantic constraints are defined as follows:

[Rule1: (?x PDS-On：has Child ?y)，
(?y PDS-On：hasSynonymy ?z) ->
(?x PDS-On：has Child ?z)]
[Rule2: (? xPDS-On：hasSynonymy ?y)，
(?y PDS-On：hasBrother ?z) ->
(?x PDS-On：hasBrother ?z)]
[Rule3: (?x PDS-On：hasSyonymy ?y)，
(?y PDS-On：hasParent ?z) ->
(?x PDS-On：hasParent ?z)]

Based on building the global ontology which is mentioned before, we add some semantic constraints, take well as research object and then build a petroleum engineering domain global ontology which contains more semantics. We have shown part of its content in Fig 4:

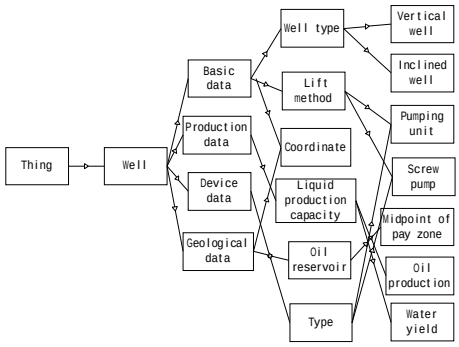

Fig 4. Petroleum engineering domain global ontology

APPLICATION EXAMPLE

China is a huge energy consumption country. It is of great significance in improving oil production and energy saving. PSDT can make the system efficiency higher by realize the data integration about complex petroleum engineering domain data. Therefore, our work plays an important role in the field of petroleum engineering.

At present, PSDT can provide a comprehensive and real-time data service in production monitoring and measure evaluation for more than 30000 oil and gas wells covering five oil fields. In the field of petroleum engineering, pump inspection period refers to the time interval from start pumping oil normally with sucker rod to stop pumping with device failure. PSDT can obviously lengthen pump inspection period by making full use of existing data resources. The pump inspection period of one oil well named MuH3-3 is less than 100 days, and sucker rod is frequently broken in the 950m and 1800m. The system finds out that the side force reaches 6kN through querying historical data of the well and calculating the existing data. After optimizing centralizer configuration and stem length, pump inspection period extends to 122 days. Besides, the system regulates pumping unit parameters of Dong3-10, and discovers that the efficiency has improved by 8.8% from 9.8% to 18.6% through reducing balanced current, equivalent torque and current change. Adjustment of Dong3-10 parameters is shown in Table 1.

Table 1. Adjustment of Dong3-10 parameters

	Before turning	**After turning**
Current balance	97%	95%
Equivalent torque	47kN·M	11.6kN·M
Current change	61A	30A
System efficiency	9.8%	18.6%

PSDT has been popularly used in the oil fields of Daqing, Jilin, Jidogn, Dagang and Huabei. On the spots of Jidong and Huabei, 356 wells participate in the test of optimization design and diagnosis. System efficiency has increased by 5.6% from 21.3% to 26.9% after pumping unit optimization design for 95 wells. And system efficiency can improve 6.5% from 30.2% to 36.7% by screw pump optimization for 50 wells. Before and after comparing histogram is shown in Fig 5.

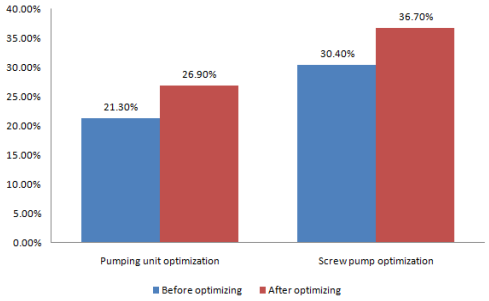

Fig 5. Comparing histogram of system efficiency

CONCLUSION

Due to the complexity of domain data , ontology technology is utilized to realize the semantic data integration. At present, this technology is the research focus. This study integrates the Ontology to build the global semantic data model for the distributed, heterogeneous and complex semantic correlation data source and to provide comprehensive, real-time data services. Experiments have shown that this method is feasible and effective.

In this paper, although this technology solves some key problems, semantic association work only develops based on Protégé or Eclipse that are still in researching. The further research is about automatic identification. And PSDT will be using in other oil fields, and extending to other domains as exploration, earthquake and so on.

REFERENCES

[1]. B. Ludäscher, K. Lin, S. Bowers et al. Managing Scientific Data: From Data Integration to Scientific Workflows[J]. *GSA Today, Special Issue GSA Today, Special Issue.*

[2]. P. A. Bernstein and L. M. Haas. Information Integration in the Enterprise[J]. *Communications of the Acm.2008,9,51(9):72-79.*

[3]. Liu X, Hu C, Li Y, et al. The Advanced Data Service Architecture for Modern Enterprise Information System[C]. *Information Science and Applications (ICISA), 2014 International Conference on. IEEE, 2014:1 - 4.*

[4]. J.F. Rainaud. A Short History of the Last 15 year's Quest for IT Interoperability in the Petroleum E&P Industry. Oil & Gas Science and Technology, *Rev. IFP, Vol. 60 2005,60(4),pp. 597-605.*

[5]. N. F. Noy. Semantic Integration: A Survey Of Ontology-Based Approaches. *SIGMOD Record, 33(4),65-70(2004).*

[6]. Jia L, Hu C, Li Y, et al. A Semantic-based Data Service for Oil and Gas Engineering[J]. *WEBIST 2014 - Proceedings of the 10th International Conference on Web Information Systems and Technologies*, v 2, p 131-136, 2014

[7]. M. Dean, G. Schreiber, S. Bechhofer et al. OWL Web Ontology Language Reference. *W3C Recommendation.* Available: http://www.w3.org/TR/owl-ref/(2004).

[8]. de Laborda C. P. and S. Conrad. Relational.OWL - A Data and Schema Representation Format Based on OWL. *In Second Asia-Pacific Conference on Conceptual Modelling (APCCM2005), volume 43 of CRPIT, pages 89 -96, Newcastle, Australia, 2005. ACS.*

[9]. de Laborda C. P. and S. Conrad. Bringing Relational Data into the Semantic Web using SPARQL and Relational.OWL. *In International Workshop on Semantic Web and Database at ICDE 2006.(2006)55-60.*

[10]. Q. Trinh, K. Barker and R. Alhajj. RDB2ONT: A Tool for Generating OWL Ontology From Relational Database Systems. I*n Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services(AICT/ICIW), Guadeloupe, French Caribbean.*

[11]. R. Shi, M Zhao and C. Sun. Knowledge Engineering and Innovation. *China Machine Press.*

[12]. GE J, HU C, LI Y, et al. An Intermediate View for Data Integration, Management in Cloud Computing[J]. *Journal of Computational Information Systems, 2013, 9(9): 3611-3618.*