Study on Ontology-based Semantic Extraction Method for Heterogeneous Data

Gaihai Li [a], Yan Zhang [b], Wangli Yang [c], Guiying Shi [d]

Northeast Petroleum University, Daqing 163318, China;

[a]13039863165@163.com, [b]zhangyuanyang_309@126.com, [c]ywl1008@126.com, [d]dqpisgy@126.com

Keywords: ontology, heterogeneous data, semantic extraction, data integration.

Abstract. In allusion to scientific data isomerism and diversity as well as the lack of semantics in petroleum domain, this paper puts forward ontology-based semantic extraction method for heterogeneous data. Firstly, semantic extraction framework for scientific data in petroleum domain is constructed. On this basis, corresponding semantic transformation and abstract rules are worked out in allusion to structured and semi-structured data to form local domain ontology. Then, overall ontology in petroleum domain of data is constructed through dependence mapping among heterogeneous data. Effective organization and uniform management of heterogeneous data are achieved. Finally, correlation methods are applied in petroleum domain to provide support for effective management of petroleum scientific data and service sharing.

Introduction

To a large extent, petroleum and gas industry belong to information and knowledge-based industry. Since it involves multiple disciplines and multiple businesses and the production life is very long, information types are numerous and information amount is huge. Especially in recent years, as data scale increases rapidly, big data management problem stands out increasingly. People are faced with a new challenge in data management. Besides, as market demand for petroleum increases continuously, exploitation for many years and geological complexity lead to gradually increase in exploration and exploitation difficulty. In such circumstance, new technology is required to achieve accurate exploration and exploitation of petroleum. The concept of “wisdom oil field” is to effectively integrate decentralized information and provide accurate information support for scientific decision-making for exploration and exploitation. Traditional and pure relational database management mode cannot effectively cope with such situation. To improve data utilization efficiency and meet users’ diversified needs, a new data processing method is badly needed to support and achieve uniform transformation and extraction of large-scale heterogeneous data. Thus, this paper proposes ontology-based semantic extraction method for heterogeneous data and applies it in petroleum data sharing service so as to achieve effective management and sharing of petroleum data through semantic support of ontology technology.

Related Researches

In allusion to semantic [1-2] treatment of heterogeneous data, certain research achievements have been gained correctly. Xu J. Et al. put forward decomposition aggregation query to integrate domain heterogeneous data [3]. It has certain support effect for semantic query of heterogeneous data. Vincini M. et al. propose semantic integration method of heterogeneous data source in data transformation system [4], which to some extent achieves semantic integration of Web data. Stein R et al. propose to apply SimpleDB storage technique to support semantic treatment Stratustore [5], which verifies the superiority of cloud database technology SimpleDB in processing simple semantic query response. Llora X et al. raise a method to construct data-intensive cloud application environment based on semantic Web technology [6]. It uses ontology technology to construct data model and stresses reuse and sharing of components. Singh G. et al. raise a method to construct cloud computation of uniform

joint ontology [7]. This method integrates ontology semantic technology into cloud computation, supports and achieves ontology-based efficient retrieval service. Chen H. et al. propose a practical semantic tool oriented to relational database [8] which is applied in traditional Chinese medicine data management and plays certain data integration role. Besides, Zhang Q. et al. propose a data integration framework based on SOA and ontology technology [9] to encapsulate heterogeneous data in Web Service form. Meanwhile, ontology technology is introduced. The superiority of domain concept description with ontology well solves data isomerism problem in data integration and achieves data access transparency. Jiang Z. et al. aims at semantic isomerism problem in electric system data integration to propose an ontology-based data integration framework [10]. Ontology semantic description module is added in data integration middleware module of traditional data integration framework, which to some extent solves semantic isomerism problem in data integration process. Semantic treatment researches on the above heterogeneous data to some extent solve semantic integration problem of heterogeneous data, but most lack more detailed semantic transformation and extraction description.

Semantic Extraction Framework

Problem description. Data have multiple forms of isomerism. Take petroleum and gas industrial data for example. 4 aspects are mainly involved:

Data isomerism: data resource expression forms are diversified, including structured form, semi-structured XML file;

Semantic isomerism: data are ambiguous at concept and meaning layer; the same thing may have multiple expressions. For example, the depth of a well can be expressed as Depth or TD;

Unit isomerism: data expression of the same concept adopts different units so that inconsistent exists in data analysis and comparison. For example, temperature data can be expressed with Fahrenheit degree or Celsius degree;

Accuracy isomerism: data expression adopts different accuracy ranges. The calculation results are also different.

Framework design

In allusion to scientific data isomerism in petroleum domain, this paper designs a semantic extraction framework for heterogeneous data (as shown in Fig.1). This framework formulates different transformation rules in allusion to data sources with different features. Different modes are adopted to carry out semantic packaging for heterogeneous data sources, and domain ontology is extracted based on mapping to express domain knowledge. Finally, scientific data sharing service in petroleum domain is achieved. Usually, heterogeneous data in petroleum domain mainly include 2 types: structured 2D table data and semi-structured XML file data. Different semantic transformation and extraction rules can be worked out for diverse types of data. Miscellaneous data can be uniformly transformed to OWL [11-12] element form. On this basis, data contents with the same connotation can me merged and mapped to preliminarily construct and form domain ontology.

Semantic extraction of heterogeneous data

Semantic extraction is conducted with different methods in allusion to diverse types of data. Semantic transformation and extraction of domain data can be conducted mainly from two aspects: structured and semi-structured:

Semantic extraction of structured data. Aiming at structured 2D table data, the following transformation rules are adopted to extract and construct OWL ontology from relational database (as shown in Fig.2):

Rule 1 Ordinary table is transformed to OWL class (OWL: Class) or subclass (OWL: SubClass)

Rule 2 Connection table is transformed to object property (OWL: Object Property)

Rule 3 Row is transformed to class instance of OWL (OWL: Class Instance)

Rule 4 If the Column is foreign key of the table, transform it to object property of OWL (OWL: Object Property); meanwhile, add Domain restraint to the table where foreign key is and add Range restraint to the reference table of foreign key.

Rule 5 If the Column is not foreign key of the table, transform it to data property of OWL (OWL: Data Property); if the Column is main key (i.e. own non-null and uniqueness reference constraint, it is necessary to add MinCardinality constraint to the data property and Function Property constraint.

Rule 6 the data value in the table (i.e. data property owned by class instance) is transformed to data value of OWL (OWL: Data Value).

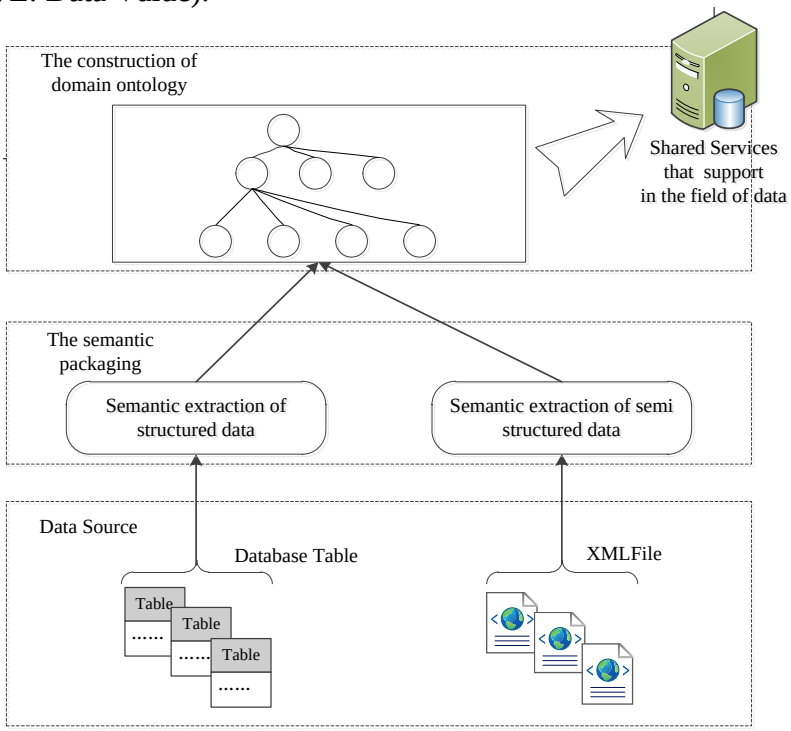

Fig.1 Semantic extraction framework of scientific data in petroleum domain

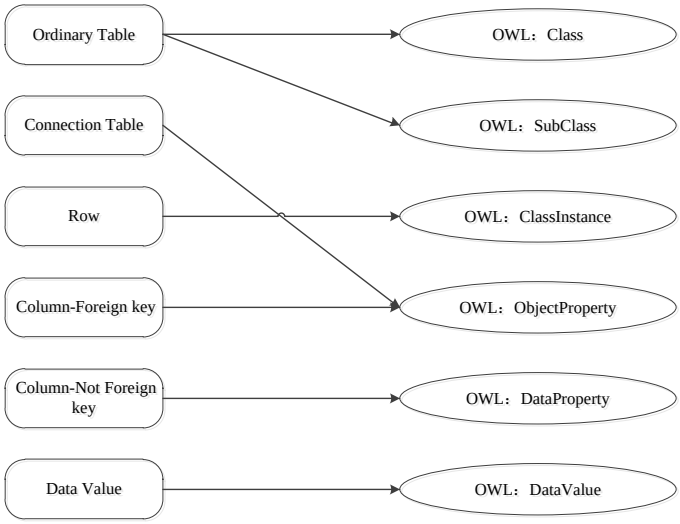

Fig.2 Semantic extraction of structured data

To explain the above rules, examples are listed. Assume there are 2 ordinary tables (Table A and Table B) and a Connection Table C. Table A and Table B are connected through inclusion relation of Table C. Table A is transformed to class of OWL; Table B is transformed to subclass of OWL; Table C is transformed to object property between Class A and Subclass B. If the main key of Table A is foreign key of Table B, main key of Table A is transformed to data property, and MinCardinality and Function Property constraints are added. Meanwhile, foreign key of Table B is transformed to object

property between Table A and Table B. Table B is subclass of Table A, where domain value is Table B and range value is Table A.

Semantic extraction of semi-structured data. Elements of semi-structured XML file mainly include node property and property value. The transformation from XML to OWL ontology can be classified into content transformation and layer relation transformation. Thus, the following transformation rules are adopted to extract and construct OWL ontology from XML file (as shown in Fig.3):

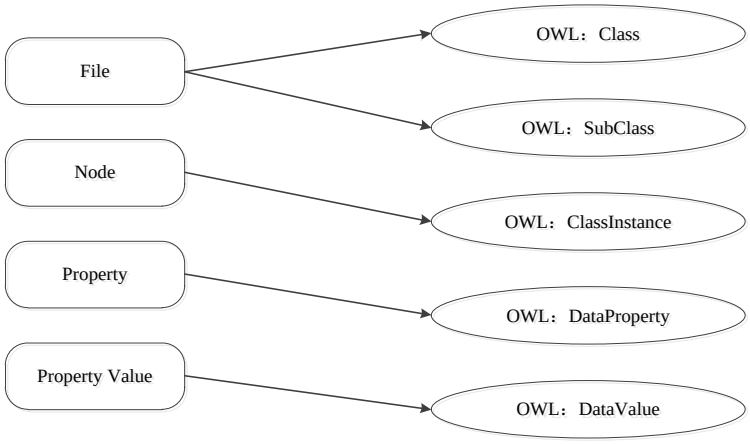

Fig.3 Semantic extraction of semi-structured data

Rule 7 XML file is transformed to class of OWL (OWL: Class) or subclass (OWL: SubClass)

Rule 8 XML file node is transformed to class instance of OWL (OWL: ClassInstance)

Rule 9 Property of XML file is transformed to data property of OWL (OWL: DataProperty)

Rule 10 Property value of XML file is transformed to data value of OWL (OWL: DataValue)

Rule 11 parent node – child node relationship in XML file is transformed to parent class – child class relationship in OWL (OWL: Parent Child Class Relationship)

Rule 12 Node – property relationship in XML file is transformed to class – data property relationship in OWL (Class Data Property Relationships).

Conclusions

This paper proposes ontology-based semantic extraction method for heterogeneous data. Based on semantic extraction framework of scientific data in petroleum domain, corresponding semantic transformation and extraction rules are worked out in allusion to structured and semi-structured data. On this basis, incidence relation among data is mapped to domain ontology. Thus, uniform data management is achieved. Relevant methods are applied in petroleum domain to provide effective support of petroleum domain data sharing service at semantic level.

Acknowledgements

This work was supported by the Provincial Department of Education of Heilongjiang province under Grant Nos”12541087”

References

[1] J. van der Geer, J.A.J. Hanraads, R.A. Lupton, The art of writing a scientific article, J. Sci. Commun. 163 (2000) 51-59.

[1] Inderjit Dhillon, Jacob Kogan, Charles Nicholas.4 Feature Selection and Document Clustering ［DB /OL］.http: / /callisto.nsu.ru /documentation /CSIR /selected /doc_ clustering /kogan.pdf, 2014-03-20.

[2] Sebastiani F.Text CategorizationDB/OL］. http: / /nmis. isti.cnr. it/sebastiani/Publications /TM05. pdf, 2014-03-20.

[3] Zobel J,Moffat A, Inverted files for text search engines [J]. ACM Computing Surveys, 2006,38(2) :1-56.

[4] Guyon I,Elisseeff A. An introduction to variable and feature selection[J].The Journal of Machine Learning Research, 2003, 3 (3 /1) :1157-1182.

[5] YanLi Qi. Introduction to Information Retrieval [M]. Beijing: Beijing University of Press, 2006(In Chinese).

[6] Dong Wu. Chinese information retrieval engine segmentation and retrieval technology [J]. Computer Applications, 2004, 24 (7): 128-131(In Chinese).

[7] Shu Li,Guo Li. Principle Analysis and development of Chinese search engine implementation techniques [J]. Application Research of Computers, 2001, 18 (11): 96-99(In Chinese).

[8] Yun Su. Search engine Google Search Tips study [J]. Gansu Science and Technology, 2005, 21 (2): 69-71, 56(In Chinese).

[9] Bin Zhang. For the Chinese Internet information retrieval system design and automatic segmentation algorithm [D]. Shanghai: East China Normal University, 2007(In Chinese).

[10] Xiaoyang He, Zhirong Wu, Lihong Lian. Analysis of the domestic search engine research situation [J]. Modern intelligence, 2005, 25 (2): 165-167,173(In Chinese).

[11] Ling Wan, Yang Xiudan. Analysis of the evaluation criteria Chinese search engine [J]. Information Science, 2000, 18 (1): 28-31, 38(In Chinese).

[12] Tan Sun, Jingyi Zhou. Advances in information retrieval research model abroad in recent years [J]. Library, 2008 (3): 82-85(In Chinese).