A Research on Ontology Modeling for Multi Source Heterogeneous Audit Data

Li Chunqiang[1, a], Chai Weiyan[2,b] and Chen Linan[3,c]

[1]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*

[1]*School of Information Management, Beijing Information Science & Technology University, Beijing 100101 ,China*

[2] *School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*

[3] *Network Information Center, Beijing University of Post and Telecommunications ,Beijing 100876,China*

[a]*tsiang@126.com,* [b] *wchai@northic.com,* [c]*chenlinan@bistu.edu.cn*

Abstract

In order to solve the problem of semantic heterogeneity in data integration, this paper places emphasis on establishing a good data model in the stage of expressing data. Through the ontology modeling method on audit data integration, we using the method of top-down ontology integration, put forward to the process of building Audit-ontology. This paper constructs the audit domain ontology, defines audit information ontology Audit-ontology, and analyzes the concept and semantic analysis of the audit information,establishes ontology classification framework, and formalized expression by using OWL language. Finally, we connect the ontology to data source.

Keywords database application system; heterogeneous database; information integration; access control; simplified integration

1 Introduction

Generally, the key to realize data integration technology is to make the system in different software and hardware equipment of the interconnection and communication. Although data integration technology has solved some problem of heterogeneous system and structure, but the grammar and semantic heterogeneity have not yet been resolved. Heterogeneous data integration based on XML can solve the heterogeneous problem of grammar, but it also has no good solutions to the semantic heterogeneous[1].The Ontology has a good ability of semantic expression, so the heterogeneous data integration method based on ontology emerges, and there are many research in various fields, it solves the

problem of semantic heterogeneity of data effectively,achieve the interaction and sharing of enterprise internal and external heterogeneous data[2].With the in-depth development of the audit database construction, the problem of semantic heterogeneity need to be solved urgently[3].

The content of this paper is as follows: the proposed to construct a model based on the theory of audit domain ontology, using ontology to obtain audit domain knowledge, describing the concept of knowledge in the field, and the relationships between these concepts, which can solve the problem ofsemantic heterogeneity of audit data, to realize the semantic heterogeneous data sharing and interaction.

2 The theory of Ontology

The meaning of ontology. In computer science theory, the common definition of ontology is proposed by Studer et al: ontology is sharing standard conceptual model explicit formal specification[4].

General speaking, ontology provides computer the conceptual model of ontology in essence, can expresses the relationship between expression of concepts and concept; in the form, it represents a glossary, the goal is to express explicit provide knowledge in related fields, to realize the semantic communication between the man-machine [5].

Ontology specification to describe has not a unified standard yet, representation methods are commonly used: four element and six element method. The four element represents the basic idea of the method is[6]：

⑴The four elements in an ontology are the concept, relationship, examples and axiom.

⑵Concept represents a collection of entities or things of a field; Relationship mainly describe the interaction relationship between the concept and the concept of property; Examples are concrete concept representation; Axiomatic value range is used to restrict the class and instance, contains many specific rules and constraint axioms.

Six tuple ontology representation method is put forward by Dr Naing Myo Myo[7]. He believes that the ontology is composed of six parts：

an ontology={C,A^C,R,A^R,H,X}

Among them, C represents the collection of concepts: A^C represents multiple attribute set consists of sets, where each attribute set corresponds to a concept. R is a set of relationship; A^R is a multiple attribute set consists of sets, a relationship in which each attribute set corresponds to the R; H represents the hierarchical relationships between concepts; X represents the axiom set.

Ontology description language. Create ontology must depend on some ontology description language. At present, there are many kinds of ontology representation language, can be roughly divided into three categories： Based on description logic (DL), such as KIF, CycL and CLASSIC; based on XML,such as XOL, SHOE,OML and RDF; based on XML+DL,such as OIL, DAML+OIL and OWL[8] These three types of ontology representation language have different features, they can be separately applied to the semantic Web environment, artificial intelligence analysis of environment etc. At present ,OWL is the most popular language[9].

In consideration of the needs of audit data constraints, expressiveness and reasoning ability, this paper will use the OWL language to express ontology model.

Ontology construction. There are many design method of ontology now, but in the actual construction process, the majority of people are follow the 5 rules proposed by Tom Gruber[10]:

⑴Clarity and objectivity, namely using natural language give the clear, objective semantic definition to terms defined by ontology.

⑵Consistency, namely the term's inference and term itself are compatible, there are no contradictions between them.

⑶Scalability, i.e. the ontology design should not only consider the existing data or data,but also have good scalability. An ontology provides a shared vocabulary, it should provides the basic of concept in the expected tasks, at the same time its expression should enable people to monotone expansion and specialized illustration of this vocabulary.

⑷Minimal encoding bias, that is to add a general or special terms to the body, do not need to modify the existing content.

⑸Minimal ontological commitment, namely modeling objects are treated as little as possible constraints.

Under the guidance of different principle, according to the research question or different areas, the researchers create a lot of ontology building method.Commonly used methods contain top-down generation ontology method, generated bottom-up ontology method, the skeleton method, enterprise modeling method, structured modeling method and so on.

Protégé is a tool to acquire knowledge developed by Stanford University, is a free software development, mainly for the acquisition of knowledge and the existing ontology merging and alignment.Protégé has become one of the most popular tools for building a domestic body[11].

This paper will use the Protégé as tools to build ontology model of multi-source heterogeneous audit. With the tool's support, we can focus research on concepts, relations, attributes and other core elements that ontology construction required,without the need to know too much of the ontology description language syntax and usage,thus saving ontology construction time,reduces the ontology construction difficulty, improve the ability of the body information expression.

2 Domain ontology construction

The audit work has obvious industry characteristics, therefore, we use domain ontology to construct the audit ontology. Research on audit data ontology multi-source heterogeneous, like other fields' ontology, required clear expression of ontology, the concept of audit data, from the perspective of semantics and logic, construct a set of clearly describing the formal system of audit data by introducing ontology and description logic, and form reflect of the relationship and interaction between the audit data from different sources the structure characteristics.

Domain ontology consists of several elements: ①the concept or class; ②the class relationship; ③attribute; ④ attribute constraint; ⑤the example. The process of constructing domain ontology is the expression of these elements of the process[12].

The essence of audit information ontology modeling is abstract concept lattice or semi formal concept tree from non formal audit field glossary, then form the logic model formalized by description logic modeling, finally finish the process of audit information ontology. This process has 4 steps: the formal concept analysis, semantic analysis, the description of logic modeling and coding.

The process of building audit domain ontology is as follows[13]:

⑴Determine the research category. This step needs to be covered with clear domain ontology, ontology construction objective, development, use and maintenance of ontology.

⑵construct term table. In formal concept analysis to the field, as far as possible to list the areas in need of various concepts, terminology used. Without requiring detailed classification of concepts, but to ensure the full coverage.

⑶Construction of ontology classification framework. Based on the glossary, combining semantic analysis, divide concept of non organization structure in the field into different branch according to certain rules. In the process of classification, we need to eliminate the concept which repeated, error, beyond the scope of the screen, leaving the expression is widely recognized, clear concept, the formation of classification framework.

⑷Construct the logical model. According to the built classification, based on the descriptionlogic,absorption of domain knowledge and experience,definite the relationship between the concepts in the field. Specific construction steps:

●Establishing the concept of data representation, definitions, etc.. If the concept is belong to the class, need to define class and classes hierarchy relationship;

●The definition of the concept of (class) attribute and attribute constraints;

●Established the relationship between concepts or categories;

●Create the instance

⑸The formation of domain ontology. Based on the logical model, combining the ontology description language, code, formalization of domain ontology.

3 The audit information classification and Concept analysis

We divide audit by information content and purpose, mainly contains:

⑴Financial audit

⑵The forensic audit of Finance and Economics

⑶The economic benefit audit

⑷The economic responsibility audit

Because the original record data is first-hand information in the collected by audit department, is the original source of the audit information system data, independent of the specific business projects, with general purpose, belonging to the audit information primitive information, so we use the original record data type as the research object, for establishing the audit information ontology model

and the multiple heterogeneous, mapping ontology to the original records in the database, finally complete the audit ontology from construction to realize the whole process.

According to the audit work methods and means, combined with original data collection process by the audit department, based on industry general terms, we initially established following conceptual words:

⑴The budget implementation audit

⑵Financial revenues and expenditures of enterprises Audit

⑶The social security fund and other financial revenue and expenditure audit

⑷The financial revenues and expenditures of financial institutions audit

⑸The execution of the budget and final accounts of construction project audit

Audit data source is complex, structural diversity, each of the following categories can be divided into many small classes, each small class may be referred to a research direction, this paper analyses the concept of audit information using the top-down method, starting from the top in thefield of general concepts,stepwise refinement,such as the three major links of social security Chinese security auditing field, fund collection, fund management, fund payment. Almost all of the concept are related to this three concepts.

In order to reduce semantic ambiguity of the audit information description and clarify the text description of nature, the essence of audit object require analysis, understand ontology semantic of audit object. Provides a favorable theoretical foundation for the concept of audit domain ontology formalization method.

We will define the conceptualization as three tuple[14]:

$$C=<D,W,R>$$

Among them, R is collection of relational conception and connotation defined in the field of space <D, W>; domain space <D, W> is a complete structure, D is domain, W is the largest state of D set. The key to concept of relationship, is like conceptualization to ontology.

The concept of the relationshipρdefined from W to D all the epitaxial relationship mapping (or function)

$$\rho: W \rightarrow 2^D$$

Given a language L and its vocabulary V, a ontological commitment O to conceptualization of the C=<D, W, R> is

$$O=(C,\zeta)$$

Among them,ζ is a constant that mapping $V \rightarrow D \cup R$ give element of D to V, the elements in the R is assigned to the predicate symbols in V.

Based on the principle of the formal concept, to health care for the object, define its conceptualization to:

D={ Insurance object domain }

W={ Insurance object state of all possible }

$R=\{\rho_{\text{Auditing Concepts}}\}$

Later in this paper will use medical insurance as an example. Health care concept analysis as shown in figure 1.

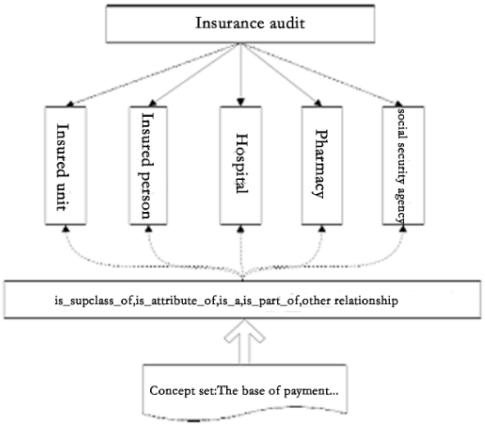

Fig. 1 Health care concept layer graph

Through the object feature refers to the concept of audit information, we analyse the audit information expression domains and complete, complete structure of the audit information, buildinformation model based on audit information classification, study on audit information like domain conceptualization process,from concept to complete audit possible mapping relationship between the epitaxial relationship, determined the relationship of each a concept.

4 semantic analysis of audit information

The expression of semantic attribute of audit information ontology.

Corresponding of concept space, there are common method like analogy, algebra method, the concept of integrated / binding method etc.. List the concept of the attribute is a concept describing the semantic way, namely attribute enumeration, this is a kind of method which is easy to understand and effectively, it uses each attribute with concepts to characterize the concept, Through meeting or to have these attributes characteristic express concept connotation.

There are many kinds of description of the concept, the concept of audit information is usually expressed in the description is based on the description of language. From these words, which reflect the essence of the concept, does not depend on the epitaxial relationship is the core to determine its semantic meaning. According to the characteristics of concept description, expressing concern about the connotation of the concept, attributes extracted representation of concepts of ontology, is the effective way to formal concept.

We can use two methods to define Ontology attribute: rigor and integrity[15]:

Rigor, any object in this property $\boldsymbol{\phi}$ must keep such properties, can be expressed as:

$$\forall x\boldsymbol{\phi}(x)\rightarrow\Box\boldsymbol{\phi}(x)$$

Integrity , an attribute $\boldsymbol{\phi}$ with integral only if there is an equivalence relation ω, makes all elements in each $\boldsymbol{\phi}$ in ω is connected to the elements only can through ω.

Starting from the ontology meaning, combined with the characteristics of the concept of audit information, extracting ontology attribute concept of audit information should abide by the following rules: (1) Audit concept can be

expressed by many attributes, but domain ontology requires defining a set of minimum number of properties as axioms to express its semantic properties, namely in the premise of semantic integrity removed excess. ⑵Determine the attribute concept of semantic screening the whole concept refers to all the possible properties of the object,not to the extension and change, do not change with the change of state of the object, the attribute values are kept constant.⑶Not every concept can cover all the attributes set of ontology, an attribute is the ontology attribute to one concept but does not guarantee that it is ontology attribute to another concept. ⑷Thecombination of different concepts of ontology attribute values , should be able to showa corresponding concept, and can distinguish different audit concept, does not produce the ambiguity.

Based on the above rules, combined with " 3 The audit information classification and Concept analysis ", here to insurance audit case expression, as shown in Figure 2 of the ontology semantic attributes of medical insurance.

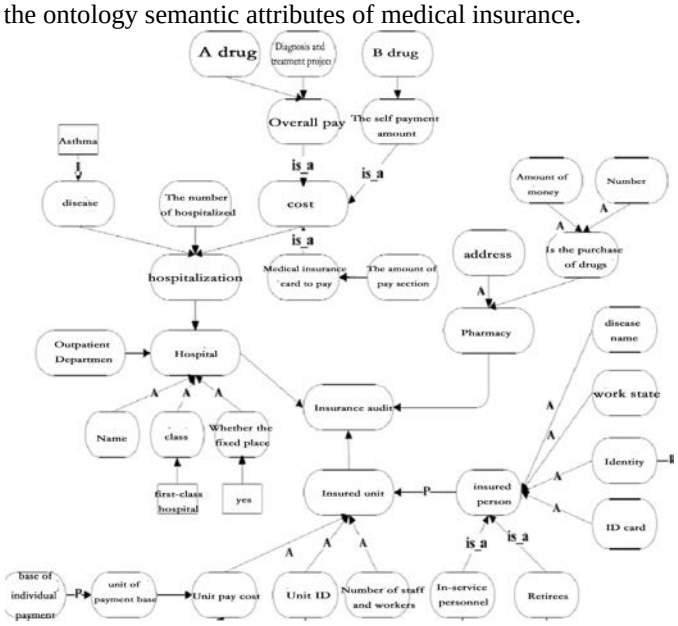

Fig. 2 Medical ontology concept semantic relation graph

The audit information ontology classification framework. According to the characteristics of the audit data, it can be divided into grade one,grade two and grade three in three different classification. The first classification is as follows: (1)The budget implementation audit; (2)Enterprises financial revenue and expenditure audit; (3)The social security fund and other financial revenue and expenditure audit; (4)The financial revenue and expenditure of financial institutions audit; (4)The execution of the budget and final accounts audit. Combined with the "3 audit field information classification and analysis of the concept of" picking up the social security fund here and other financial revenue and expenditure audit is further illustrated, second level categories include: endowment insurance, unemployment insurance,industrial injury insurance,

maternity insurance and medical insurance.The medical insurance include the participating units and individuals, hospitals,pharmacies and the social security agencies and other three level classification. As shown in figure 3

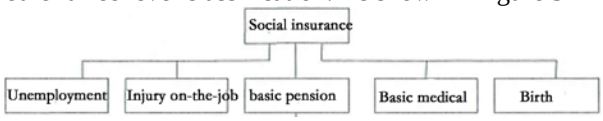

Fig. 3 The field of social ontology framework

5 Construction and data association logic model of audit information ontology

The establishment of the audit information ontology. In this paper, we use Protégé software to achieve audit ontology construction. The audit information ontology includes class, instance, attribute three parts, in Protégé, they are represented by three tuple in the relation, i.e. A→B,C={A,B}. A represents a class (class), B represents an attribute (property), C represents a class or instance (individual). Clearly the above concept, ontology formalization construction process is as follows.

First establishes the ontology of all classes and subclasses, create "Medicare" classes in the owl:Thing node, and then create all of the class based on this class according to the above classification framework; then, create a property on the Properties page, add all leaf nodes of health classification in the domain of the properties; The third step, create an instance, create health care instance in the Individuals page, enter the property of the added, thus complete creating an instance of ontology medical insurance classification.

After the establishment of audit data ontology class framework, a owl file will be generated, all of the classes, attributes, examples are in the form of owl file expressed by OWL language. Finally, using the Protégé's OntoGraf to draw the organization chart, final audit data ontology classification and instantiation as shown in figure 4.

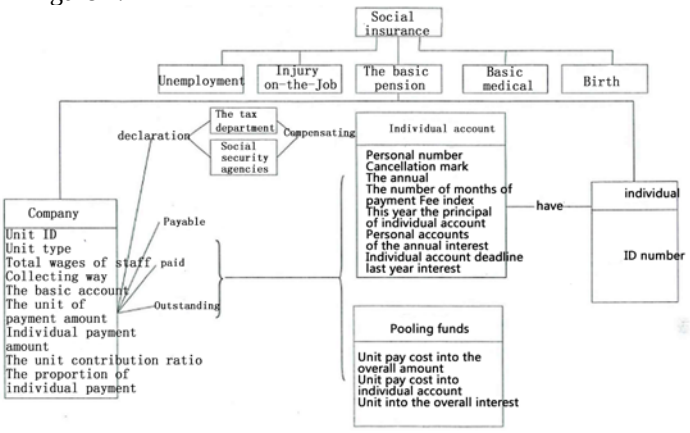

Fig. 4 the field of social ontology framework

The audit information ontology formalization description. In form, audit information ontology and classification and standard of audit data are similar, but

the existence of audit ontology concepts is not isolated, they are connected with the constraints, can express the domain knowledge without ambiguity. In order to make ontology be easily converted to structure body can be processed by computer, the body needs a description tight logic structure,support the use of formal description language expression. The ontology provide theoretical and technical support to the audit data classification and to set up the database system,in order to facilitate the integration of multi-source heterogeneous audit data. According to the characteristics of the audit data, and experience of the existing research, we define audit information ontology as follows:

Audit-ontology=<C,R,H,P,P^R>

Among them, the parameters are described below:

C represents a class (Class), representing a set of object shared some of the same attributes;

R represents the semantic relation (Relationship), transverse relationship between classes,including equivalence,joint,disjoint,and so on,transverse relationship between classes can implement semantic heterogeneous data conversion and integration;

H represent hierarchical relationship (Hierarchy), mainly refers to the parent - child relationshipclass, corresponding to R,H can be understood as the vertical relationship, single out is to emphasize the importance of hierarchical relation in the ontology, the hierarchy relationship is also a kind of semantic relation;

P represents the attribute (Property), can be divided into object attributes and data attributes. Object attributes refers to the relationship between the individual and individual, data attributes refers to the relationship from individual to numerical;

PR represents restriction of property， are constraints to the attribute value types, scope and limits of the maximum minimum number;

I represents Individual, a concept instance, object or individual.

Expression of audit information ontology formalization. Based on the above research, we use OWL language to express the audit information ontology description and formal expression, the specific contents are as follows:

(1) Definition of a class

OWL language uses class to express the concept , class represents the collection of some examples, namely all classified information in the audit ontology, we use the tag owl:Class as formal expression to the class description. Relationship between classes here are mainly hierarchical relationship of longitudinal. Firstly, we create a Medical insurance class, it is sub class of social security class, and then create its sub class: endowment insurance class, unemployment insurance class, inductrial injury insurance class,birth insurance class.The following is the definition for social security and its sub class:

< Medical insurance rdf: ID=" Medical insurance _1">
<rdfs: subClassOf><owl: Class rdf: ID="Social security"/></rdfs: subClassOf>
</owl: Class><owl: Class rdf: ID="Endowment insurance">
</rdfs: subClassOf rdf: resource="Social security"/></owl: Class>
<owl: Class rdf: ID="unemployment insurance">
<rdfs: subClassOf rdf:resource="Social security"/></owl: Class>
<owl: Class rdf:ID="industrial injury insurance">

<rdfs: subClassOf rdf:resource=”Social security”/></owl: Class>
<owl: Class rdf:
ID=”maternity insurance”><rdfs: subClassOf rdf:resource=”Social security”/></owl: Class>
<owl: Class rdf: ID=” Units”>
<rdfs: subClassOf><owl: Class rdf: ID=”Medical insurance”/></rdfs: subClassOf></owl: Class>
<owl: Class rdf: ID=”Individuals”>
<rdfs: subClassOf><owl: Class rdf: ID=”Medical insurance”/></rdfs: subClassOf></owl: Class>
<owl: Class rdf: ID=”Hospitals”>
<rdfs: subClassOf><owl: Class rdf: ID=”Medical insurance”/></rdfs: subClassOf></owl: Class>
<owl: Class rdf: ID=”Pharmacies”>
<rdfs: subClassOf><owl: Class rdf: ID=”Medical insurance”/></rdfs: subClassOf></owl: Class>
<owl: Class rdf: ID=”Social security agencies”>
<rdfs: subClassOf><owl: Class rdf: ID=”Medical insurance”/></rdfs: subClassOf></owl: Class>
(2) Description of attribute

The basic attribute of the medical and health products, including "Insured units", "Insured individuals" and "hospital","pharmacy", "Medical insurance agency",attribute definition they are shown as follows:

<owl:DatatypeProperty rdf:ID=" Medical insurance content"〉
<rdfs:rangerdf:resource="http://www.w3.org/2013/XMLSchenia#string"/>
<rdfs:domain><owl:Class>
<owl:unionOf rdf:parseType="Collection"><owl:Class rdf:about="# Insured units "/>
<owl:Class rdf:about="# Insured individuals "/><owl:Class rdf:about="# hospital "/>
<owl:Class rdf:2ibout="# pharmacy "/><owl:Class rdf:about="# Medical insurance agency "/>
</owl:unionOf></owl:Class></rdfs:domain></owl:DatatypeProperty>
……
<owl:DatatypeProperty rdf:about="# Medical insurance entity">
<rdfs:doniam rdf:resource="# Medical insurance "/></owl:DatatypeProperty>

（3）Create instance

Create an instance of a " Medical insurance " category in accordance with the above attributes, an instance can be understood as a record in a database table, the OWL language is described as follows:

<owl: Class rdf: ID=”Medical insurance”>
< Insured units rdf:datatype= "http://www.w3 .org/2013/XMLSchema#strmg">CBDW01 </ Insured units >
< Insured individuals rdf:datatype="http://www.w3 .org/2013 /XMLSchema#strmg">CBGR01 </ Insured individuals >
< hospital rdf:datatype="http://www.w3 .org/2013 /XMLSchema#strmg">YY01 </ hospital >
……
< Medical insurance agency rdf:datatype="http://www.w3 .org/2013 /XMLSchema#strmg">CBJBDW01 </ Medical insurance agency >
</ Medical insurance >

A ssociate audit information ontology to data. In order to construct better domain ontology , this paper uses the bottom-up generation method, combine the local ontology and the global ontology, form a larger information and knowledge pool. Most of the data are stored in a relational database,if it can be mapped to the ontology, it will be very useful for constructing ontology. We can make the rules of the relationship between model to build ontology mapping.

The original record in audit information database generally includes: (1) The budget implementation data. (2)Enterprises financial data. (3) The social security fund and other financial data.(4)Financial institutions financial data. (5)The construction project budget implementation and final accounts data. Each kind of data can be divided into a some small classes, each small class corresponds to a table in a relational database. The whole body is a tree, each concept (or class) is a

node in the tree, then the relational database representation for each leaf node ontology, we use rdfs:subClassOf or WOL:ObjectProperty to express their relationships ,thus we mapping relational database table to the ontology classification frame.

6 Conclusion

Multi source heterogeneous data ontology modeling is a dynamic field, it not only by the engineering innovation driven, but also by the evolution of the problem itself. This paper apply ontology modeling method to audit data integration domain. This paper uses top-down ontology modeling , first formalized the conception analysised by audit data to obtain the glossary,then obtain ontology classification framework by semantic analysis of domain ontology, finally descript logic and code, establish ontology model. Thus finish the construction of the audit domain ontology,and uses the OWL language to realize formal expression, at the same time, the correlation between ontology and data source is mapping relationship.

Acknowledgement

It is a project supported by National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2012BAH08B02). The corresponding author is Li Chunqiang.

References

[1]Jlsabel F. Cruz,Huiyong Xiao. An Ontology-based Framework for XML Semantic Integration[C], Washington. IDEA Workshop, Proceedings of the international Database Engineering and Applications Symposium, IEEE Computer Socity, 2004: 217- 226

[2] Lee Rubao, Xu Zhiwei. Exploiting Stream Request Locality to Improve Query Throughput of a Data Integration System. IEEE TRANSACTIONS ON COMPUTERS , 2009, 58(10): 1356-1368.

[3] Di Lorenzo Giusy, Hacid Hakim, Paik Hye-young, Benatallah Boualem. Data Integration in Mashups. SIGMOD RECORD, 2009, 38(1): 59-66.

[4] Jason McHugh, Serge Abiteboul,et al. Lore: A Database Management System for Semistructured Data. ACM SIGMOD Record, 1997, 26(3): 54-66.

[5] Robert McCann, AnHai Doan, et al. Building Data Integration Systems: A Mass Collaboration Approach. Sixth International Workshop on Web and Databases(WebDB 2003), 2003, 25–30.

[6] P. Ziegler. Data Integration Projects[J]. World-Wide. 2006.

[7] Marc Friedman, Alon Levy, Todd Millstein. Navigational plans for data integration . AAAI'99, Orlando: American Association for Artificial Intelligence,1999, 67-73.

[8] Motik B, Grau B C, Horrocks I, et al. OWL 2 Web Ontology Language: Proflles[J]. W3C Recommendation (27 October 2009). 2012.

[9] Robert McCann, AnHai Doan, et al. Building Data Integration Systems: A Mass Collaboration Approach. Sixth International Workshop on Web and Databases(WebDB 2003), 2003, 25–30.

[10] Serge Abiteboul, Omar Benjelloun, Tova Milo . Web Services and Data Integration. Third International Conference on Web Information Systems Engineering(WISE 2002), IEEE Computer Society, 2002, 3–7.
[11] Zhao H, Zhang S,Zhao J. Research of Using Protégé to Build Ontology[C]//Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on. IEEE, 2012:697-700.
[12]Andrea Cal, Diego Calvanese.Data integration under integrity constraints . Information System , 2004, 29 : 147-163.
[13] Jeffrey D. Ullman. Information integration using logical views. LCNS, 1997,Volum 1186: 19-40.
[14]Marcelo Arenas, Leopoldo Bertossi, et al. Consistent Query Answers in Inconsistent Databases. InProc. PODS 2003. San Diego: ACM Press, 2003, 285-291.
[15] Andrea Calì, Domenico Lembo. On the Decidability and Complexity of Query Answering over Inconsistent and Incomplete Databases. Proc. PODS 2003.San Diego: ACM Press, 2003, 260-271.