Journal of Theoretical and Applied Information Technology
31st May 2013. Vol. 51 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HETEROGENEOUS INFORMATION INTEGRITY FOR CHINESE MINORITIES WITHIN CLOUD COMPUTING PLATFORM

LIRONG QIU
School of Information Technology, Minzu University of China, Beijing, China
E-mail: qiu_lirong@126.com

ABSTRACT

The minority information resources' construction is an important part of national information construction in China. The heterogeneous distributed data semantic integration technology within cloud computing platform, mainly focus on how to use the machine automatically analyzing, understanding and handling of user needs and heterogeneous information. In this paper, the methodology of the semantic integration and dynamic management of heterogeneous distributed data is introduced firstly. And then how to build the semantic ontology for heterogeneous information are illustrated. The technology can find the data related with "cloud application", and then it can convert the chaotic data source to an orderly, easy-to-use source of information.

Keywords: *National Minority Information; Cloud Computing; Semantic Integration*

1. INTRODUCTION

All ethnic groups have created a content-rich, colorful splendor of national culture in the development process, and generated a lot of Chinese minority information resources. These resources contain many of elements, such as geographic resource information, the state of the environment information, education information, sports and cultural information, etc. These resources can cover many aspects of the national economic and social life basically; these aspects involve ethnic, religious, sports, culture, and the arts, social order, population, etc.

The minority information resources' construction is an important part of national information construction. These minority resources' collection, sorting and comprehensive utilization, can conductive to the accumulation of economic and social integration of the ethnic minority areas and promote the development of national regional economic. At the same time, it can provide real data support and timely information services for formulating our country's national policy, upgrading and restructuring the national regional industry, improving people's livelihood and comprehensive management of social. In addition, it can also provide data support for the study, heritage and conservation of minority culture.

Minzu university of China has accumulated and formatted a massive minority information survey data when sort out and count the minority information, for example, we have established some minority information databases, such as the databases about the economic development, traditional sports culture and tourism resources of minority areas.

But there are also many problems in these data. For example, the management of these data is extremely fragmented, and these data exists the excess of data redundancy and inconsistency; the sharing degree of data fail to meet requirements of the overall development and utilization of information resources; large amounts of data cannot provide an unified interface, and cannot adopt a common standard and specification; we cannot obtain shared and common data source, and this situation results in a large number of islands of data. All of these problems can't make use of these data. The reasons for these problems are that we use different equipment and information technology when we built these data at that time.

These Chinese minority information resources have the following characteristics: multilingual (involving English, Chinese, Mongolian, Tibetan, Uyghur, Yi language), wide data sources (involving downloaded resources, resources from fieldwork, purchase, intermediate generated and access to library), multi-format (text, PDF, image, audio,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

video and icon), wide varieties (including religion, sports, arts, social security, agriculture, industry, land use, soil, vegetation, hydrological, terrain, traffic, population, administrative regions), geographically dispersed (including Tibet, Xinjiang, Beijing, etc.), heterogeneous data, growing faster and existing data redundancy.

The Chinese minority information resource is an important part of the information management of the nation, and it can provide important support for the management and decision-making of the ethnic minority areas. How to make use of these heterogeneous distributed resources and provide unified data management and services is an urgent problem to be solved. The advent of cloud computing technology supplies a new way for information management and services [1].

IBM announced the plan of cloud computing at the end of 2007. As an emerging technology, cloud computing is based on open standards and service, with the internet as the center, and it can make various computing resources on the Internet work together. These resources can be built several large data center and computing center, and then we can provide some special services such as safe, fast and convenient data storage and network computing.

"Cloud applications" are to use the large –scale data centers, and robust server to run network applications and services. In the "cloud", all the data processing tasks are accomplished by a large number of distributed computers. End users can access the computer and storage system through the network according to their demand. The data center is responsible for handling the data on the client computer. This can be through a data center to provide data services to users who use a variety of different equipment. So we can realize that allowing anyone with Internet connected devices can access "cloud applications" [2].

The data center is the core of cloud computing, the scale and reliability of data center resources have an important impact on the upper layer of the cloud computing services. Some companies, such as Google, Facebook, attach great importance to the construction of the data center. In 2009, the data center of Facebook has 30,000 compute nodes. By 2010, the number of compute nodes is to reach 60,000. Google every quarter spent about 600 million dollars in data center construction [3].

Unlike traditional enterprise data centers, cloud computing data center has the following characteristics:

(1) Autonomy. Compared with the traditional data center need artificial maintenance, cloud computing data center for the scale effect of the system automatically reconfigure.

(2) Unified standard. By managing large-scale clusters with the unified and standard method, we can reduce the cost of data management.

(3) Size scalability. Cloud computing data centers usually use large-scale cost-effective equipment, and provide the expansion of space. At last, realizing the data resources handle.

According to these features, clouding computing data centers ask that the underlying distributed heterogeneous data can customize the underlying data resources flexibility for a particular "cloud application". Providing the integration of distributed heterogeneous data resources is the key problem that cloud computing data center platform needs to solve [6].

The isomerism, semantic fuzziness and ambiguity in data itself increase the difficulty of the machine analysis. Data (for the computer in terms of binary data) is only to convey the semantic of the media, and the semantic expression is the core and key of the communication [7].

The heterogeneous distributed data semantic integration technology which is based on cloud computing platform, studies how to use the machine automatically analyzing, understanding and handling of user needs and heterogeneous mainly. The technology can find the data related with "cloud application", and then it can convert the chaotic data source to an orderly, easy-to-use source of information.

2. RELATED WORK

Examples of three industry specific cloud computing are Google's cloud computing platform and cloud computing network applications, IBM's "Blue Cloud" platform products as well as Amazon's Elastic Compute Cloud. Google cannot share the cloud computing's internal infrastructure with external users. IBM's "Blue Cloud" computing platform is available for sale soft and hardware collection, users build their own cloud computing applications based on these software and hardware products. Amazon's Elastic Compute Cloud is a hosted cloud computing platform, users can directly use this platform through the remote operator interface, but they cannot see the actual physical node [4].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The U.S. National Science Foundation (NSF) in cooperation with Google, IBM, Hewlett-Packard, Intel and Yahoo to promote cloud computing research. In2008, NSF established the Cluster Exploratory initiative (CluE Cluster Exploratory initiative), its goal is to make cloud computing easier to use and more reliable for researching, and let more academic researchers be able to access the large-scale, distributed computing resources provided by IBM and Google [5].

University of Chicago (UC) and University of Florida (UFL) have started the scientific research projects of clouds, their goals have two: One is to make science and teaching program can use EC2 model of cloud computing to test; another is to understand the potential challenges and coping strategies brought by cloud computing.

In China's "cloud computing" plans, Tsinghua University is the first university to participate in the cooperation. It will open"mass data processing" course with Google cooperation. Among them, Google provide course materials to the professors at Tsinghua University, and provides laboratory equipment, and assists school to build the "cloud computing" experiment environment which is based on the existing operation resources.

In early 2008, IBM in cooperation with the Wuxi municipal government established Wuxi software park cloud computing center.

In July 2008, Rising launched a "cloud security" plan.

In 2009, the VMware vForum User Conference held in China, this conference brought the concept of open cloud computing into China.

"The decision of the State Council on speeding up the cultivation and development of strategic emerging industries", published in October 18, 2010, regarded the cloud computing as one of the strategic emerging industry of the 12th Five Year Plan. On the same day, the Ministry of Industry, Development and Reform Commission jointly issued the "Notice on cloud computing service innovation and development of pilot and demonstration work". This notice identified in the five cities of Beijing, Shanghai, Shenzhen, Hangzhou, Wuxi, first to carry out the development of cloud computing service innovation pilot and demonstration work.

3. METHODOLOGY

The semantic integration and dynamic management of heterogeneous distributed data (involving computing resources and storage resources) aims to manage scattered resources centrally, and provide secure, transparent access to resources and management strategy for specific application services [8].

Using field descriptive ontology knowledge to realize the effective organization of disorderly heterogeneous resources and this can reflect the semantic structure of information resources. People can archive that resource location based on knowledge, query and management, to support semantic interoperability.

A. Metadata Representation

Metadata is data about data, its basic purpose is to manage data, so as to realize the query, reading, exchange and sharing. The data constitute only by Arabic numerals, which doesn't have the definition and description, can't be used in the practical study. The metadata base which needs to follow the definition of metadata combines system, reports, indicators, reporting directory, grouping and other kinds of metadata organic which are related with survey project. And then, the metadata base provides effective service for all types of users, and also facilitates the information resources' integration, management and preservation.

The metadata can be used as the auxiliary data of the database information, and this is helpful when users need to reprocess data [9]. The establishment of the metadata base can avoid the blindness when users who don't know much about ethnic affairs select data. The meaning of establishing metadata base is:

(1) Being able to provide a unified standard for different Ethnic Affairs Commission having the established National Information Services database to join the information platform.

(2) To develop the metadata specifications of the physicochemical database. To develop and publish the corresponding metadata.

(3) Management tools. To assist and guide each professional data center to complete their professional database metadata's construction.

(4) Develop metadirectory-based high availability and efficiency of data application service system platform. Set up the data entry collection, management, query, and the corresponding rights management mechanism.

(5) To achieve unified management of existing distributed data and services through the advanced metadata directory technology.

Journal of Theoretical and Applied Information Technology
31st May 2013. Vol. 51 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(6) Database management system can realize that loading the database dynamic. When the structure of data change, we can be as simple as to implement this change, and the corresponding data application system doesn't need to be reconstructed. So the versatility of the system is good, and it is easy to master for scientific and technical personnel, and this system can provide an ideal soft environment for various scientific and technical data retrieval query and its management.

B. The Semantic Model Of Data

Generally, if we want to use and operate data reasonable, we first need to understand the meaning of data. In fact, in the field of data management, many researchers have already started to study the problem of data's semantic since the first birth of database. Because the traditional database application is always in a relative closed and stable environment. Semantic problem can get reasonable control, so has not become a key problem. Although the meaning of the data in the database is not fully described, in this relatively closed user group of relatively stable environment, we can do some correct operation and meaningful treatment about data through the agreed meaning of data between users. Programmer can also develop successful database applications which are based on the agreed semantic. Program can control and process data meaningful, this is because programmers have the semantic information of data hard-coded into the program code.

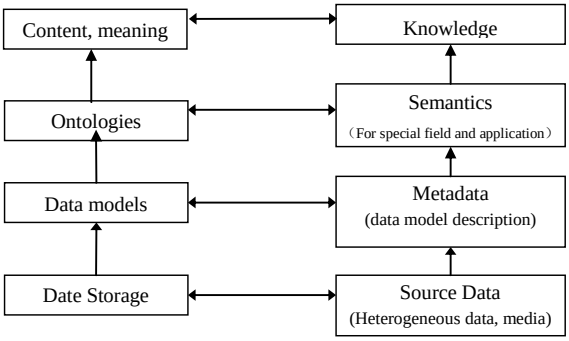

Figure 1. National Information Resource Modeling Demand

Analyzing the features of metadata, we can find that metadata is generally complicated format except for such a simple format of the data dictionary. Metadata's format is changeable, and most of them have a complex hierarchy and relationship. Metadata general data volume is not large, during the system operation for read-only; the metadata is used dispersed generally with the features of cross-process, cross-platform.

We can organize the field data or resources by establishing data models, so as to share with others. Compared with some data which is described through some simple relation model, metadata modeling is much more complicated which is offered or used by different organization. So if we want to share complex cross-platform metadata resources, we need to organize and associate these data according to the semantic way. Obviously, the traditional object-oriented model cannot achieve this goal. We will study the ontology-based metadata resources. And the mapping relationship between metadata and domain ontology is shown in figure 1.

C. Semi-Automatic Semantic Relevance Framework

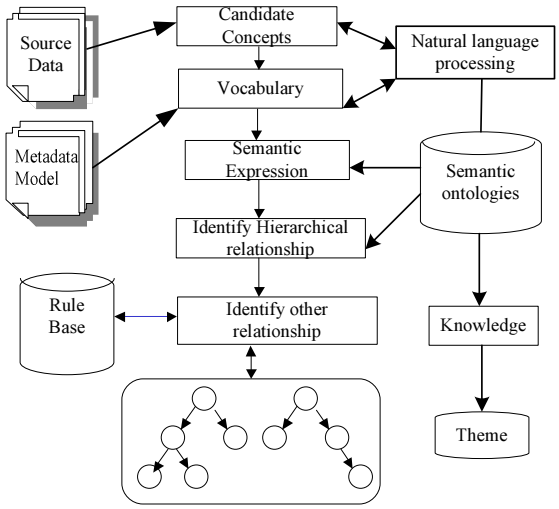

Figure 2. Semi-Automatic Semantic Relevance Frameworks

Figure 2 proposes the semi-automatic semantic relevance framework between heterogeneous data sources. This framework uses database, semi-structured documents (Web page, XML documents, etc.) and unstructured documents such as text files as importation. This framework can realize these importations' vectorization through the method of shallow natural language processing, and then use the methods of machine learning and data mining to analyze the concept of implied semantic relations.

The framework has the following features, as shown in figure 2:

Domain experts can intervene at any time, and some formalization of knowledge like user experience can be embedded into this framework.

The learning framework is a general process of iterative, so we can improve the quality of ontology continuously.

Subject database is an information system database which is established for the service main

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

body, with the only source and stable structure. If we can convert the original application database which faces the application system to subject database when we integrate data, this will be a good kind of data management schema which can develop sustainable.

4. INFORMATION INTEGRATION BASE ON ONTOLOGY

Ontology demand analysis is the basis of ontology's development and design. The demand analysis is mainly according to the requirements of the application to determine appropriate ontology requirement. The Ontology Requirement Specification (ORS) represent the result of analyzing, and this document should include the following information:

A. The Domain And Target Of Ontology

First of all, we should determine the domain and target of ontology. The domain of ontology is that the scope which ontology should contain and involve, and the target of ontology is that the features which ontology should own. For example, some explanation like the abstracts levels of concept and relationship, the degree of formalization, the way of reasoning and the representation language is involved in the domain and target of ontology. A good method is to refer to the existing ontology, to analyze its inadequacies, and then we can extend and modify these ontologies. We can also analyze the use case to determine the major and related areas. Determining the target of ontology is related with the use of ontology, for example, ontology is regarded as a structured means of knowledge or as reusable field knowledge base, usually the latter requires higher.

B. Users and Use Case

In software engineering, the Use Case is a good method to capture the requirements of software. Every Use Case summarizes the different views to ontology of the users, such as what concepts should be included in the ontology, what kind of reasoning should be supported, what shortages exist in the existing ontology, what are the mainly hinder factors, etc. For different individual users, the difference may be very large. Especially, the hinder factors have an important impact on the development of new domain ontology.

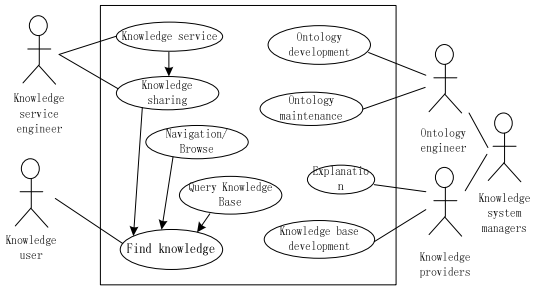

Figure 3. Use Case Based On Ontology Knowledge System

Figure 3 is a figure about the Use Case which is based on the ontology knowledge system. Four typical users of the knowledge system and its Use Case are cited in the picture.

Ontology engineer: Developing and maintaining the ontology, and building field conceptual model.

Knowledge provider: Developing and maintaining the knowledge base, some general knowledge source including Web document, the electronic document which has different format, database. This knowledge must be annotation and stored in the knowledge base.

Knowledge users: The end users of knowledge base, which implement knowledge query work mainly.

Knowledge service engineer: Developing knowledge service which is based on application, and promoting the knowledge system to be used in the environment of dynamic, heterogeneous.

The Use Case describes the knowledge system and the application's real usage which is supported by the knowledge system. The concepts and relationships which are involved in the Use Case must be reflected in the ontology. In order to gain these concepts and relationships from the Use Case, ontology engineers can use the method of capacity table to get the content of the domain ontology.

C. Domain's Analysis And Modeling

In order to have a clear understanding of the areas of modeling from the perspective of a comprehensive, objective and overall, and to guide the latter building part of the ontology, it is necessary to analyze the things in the domain and determine the modeling method of domain. This process is called field analysis. Field analysis mainly consists of three tasks: to determine the source of knowledge, to analyze the field from the perspective of macroscopic and to determine the modeling methods of the domain.

Determining the source of knowledge is the premise of the analysis of the field. Some common knowledge sources include: questionnaire, the dictionary of domain, the rules of domain, standard templates, a variety of statistical material,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

organizational charts, product manual, technical write paper, project proposal, the experience of experts in the field, Web pages, indexing vocabulary, other electronic documents, database and so on. Ontology engineers can select the appropriate amount of typical field knowledge source to analyze according to certain criteria.

Things in the field are very complex, in order to describe these things fully and correctly, these things need to be detailed analyzed. In general, we can analyze things from the perspective of "inside and outside", "surface and inside", "dynamic and static", "general and special", "whole and part", etc.

If we can clear these relationships, we can use them to guide the latter of the domain ontology.

D. Modeling Method Of The Domain

The modeling of domain knowledge includes three aspects, namely, the static field knowledge model, reasoning knowledge model and task knowledge model. The static knowledge model describes the fact and knowledge in the field which we are concerned. Reasoning knowledge model describes the basic reasoning process which uses the static knowledge. Task knowledge model describes the application's target, and how to decompose task into subtasks to achieve these targets. The modeling methods usually have three options: Top-down method, Bottom-up method and Intermediate expansion method. They are introduced as follows:

(1) Top-down method

It first models the domain concepts at the higher levels, and then gains the bottom concepts through gradually thinning. The method needs more artificial intervention, but can produce higher quality domain ontology, but also supports the existing top domain ontology's reuse.

(2) Bottom-up method

This approach assumes that domain documents already contain most of the concepts and conceptual structure terminology in the field. We can use the semi-automated tools to extract related words, and then we can generalize and merger these words to get general concepts.

(3) Intermediate expansion method

It can be seen as a compromise of the above two methods. We general first identify the most important concepts and relationships, and then through the generalization for the upper field concept, through the specialized for bottom concept.

The above methods have their advantages and disadvantages. Top-down method can obtain more detailed description about domain, and get high quality domain ontology through more artificial participation. The quality of ontology is often lower through the bottom-up method, but we can get more complete concept model. At the same time, this method need less manual intervention, and the efficiency of it is higher than other methods. Along with the progress of the natural language processing, machine learning, data mining, information retrieval and other areas, and with the improvement of the quality of ontology, bottom-up method will be a very promising field modeling methods.

5. CONCLUSION AND FUTURE WORK

The minority information resource is an important part of national information construction of China, and it can provide important support for the management and decision-making of the ethnic minority areas. Minzu university of China has accumulated and formatted a massive minority information which are multilingual, heterogeneous and geographically dispersed. How to make use of these heterogeneous distributed resources and provide unified data management and services is an urgent problem to be solved.

The heterogeneous distributed data semantic integration technology within cloud computing platform, mainly focus on how to use the machine automatically analyzing, understanding and handling of user needs and heterogeneous.

The semantic integration and dynamic management of heterogeneous distributed data (involving computing resources and storage resources) aims to manage scattered resources centrally, and provide secure, transparent access to resources and management strategy for specific application services.

In this paper, the methodology of the semantic integration and dynamic management of heterogeneous distributed data is introduced firstly. And then how to build the semantic ontology for heterogeneous information are illustrated.

The work of this paper is a part of our ongoing research work, which aims to provide an open platform for supporting information integration for those minority data resources. Various experiments and applications have been conducting in our current research. Future work includes how to increase the semantic annotation information for semantic mapping, information retrieval model based on the semantic ontologies.

ACKNOWLEDGMENT

Our work is supported by the National nature science foundation of China (No. 61103161) and the "Science Fund for Youths" project of Minzu University of China (No. 1112KYQN39).

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

REFERENCES

[1] Foster, C. Kesselman. The Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, 2004.

[2] Cloud computing. http://en.wikipedia.org/wiki/Cloud_computing. access on Nov. 24, 2011

[3] L. Wang, J. Tao, M. Kunze, D. Rattu. The Cumulus Project: Build a Scientific Cloud for a Data Center. *Cloud Computing and Its Applications (CCA'08)*. Chicago, Illinois 2008.

[4] Amazon Elastic Compute Cloud (Amazon EC2). http://aws.amazon.com/ec2/. access on Nov. 24, 2011

[5] D. Nurmi, R. Wolski, C. Grzegorczyk, G. Obertelli, S. S. L. Youseff, D. Zagorodnov. The Eucalyptus Open-source Cloud-computing System. *Cloud Computing and Its Applications (CCA'08)*. Chicago, Illinois 2008.

[6] K. Keahey, R. Figueiredo, J. Fortes, T. Freeman, M. Tsugawa. Science Clouds: Early Experiences in Cloud Computing for Scientific Applications. *Cloud Computing and Its Applications (CCA'08)*. Chicago, Illinois 2008.

[7] M. C. Murphy, M. K. McClelland. Computer Lab to Go: A "Cloud" Computing Implementation. *21st Annual Information Systems Education Conference (ISECON 2008)*. Phoenix, 2008.

[8] X. Llora, B. Acs, L. Auvil, B. Capitanu, M. Welge, D. Goldberg. Meandre: Semantic-Driven Data-Intensive Flows in the Clouds. *4th IEEE International Conference on e-Science*. Indianapolis, Indiana, USA, 2008.

[9] Leopoldo Bertossi et al, Semantics in Databases, L. Bertossi et al. (Eds.): *Semantics in Databases*, LNCS 2582, pp. 1–6, 2003. Springer-Verlag Berlin Heidelberg.

[10] Zhiwei Xu, Huaming Liao, Haiyan Yu, Li Cha. Notes on Classifying Network Computing System. *Chinese Journal of Computers*. 31(9):1509-1515, 2008