Virtual Dataspace-A Service Oriented Model for Scientific Big Data

Wei Lin[1], Changjun Hu[2], Yang Li[3], Xin Cheng[4]
School of Computer & Communication Engineering
University of Science & Technology Beijing
Beijing, China
[1]linwei0201@gmail.com, [2]hu.cj.mail@gmail.com, [3]mailbox.liyang@gmail.com, [4]chengxin0613@gmail.com.

***Abstract*—The massive, distributed, heterogeneous and diverse features of big data have raised challenges to traditional data management systems. As the development and innovation of DataSpace, virtual dataspace (VDS) model is proposed for big data management. Local ontologies are created from data sources. Then the local ontologies are mapped and formed a global ontology. Based on this, access log and user feedback are considered for data evolution. At last, a material scientists-oriented service (materials scholar assistant) is introduced as the application case of VDS.**

Keywords-dataspace; virtual dataspace; big data; scientific data management;

I. INTRODUCTION

The explosive growth of information has taken us into the era of big data. The features of big data have raised challenges to traditional data management systems. 1）Mass: new IDC Digital Universe study show that from 2005 to 2020, the digital universe will grow from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020) [1]. 2）Distribution: databases are distributed in different regions and hard to be integrated. 3）Heterogeneity: data schema, storage format and operation system among the databases are distinct. 4）Diversity: data are available in various formats such as relational data, XML, HTML, image, video and audio.

Traditional data management systems can hardly meet the demand of big data management and evolution. In spite of the great success of relational database management system in dealing with structured data, it is particularly inadequate in managing massive, distributed, heterogeneous and diverse data. 1）For such a massive scale of data, semantic integration rather than simple storage is necessary. 2）The diversity of data schema and format hinders the unification of data model, especially for non-structured data such as images and audios. Under the circumstances, dataspace was put forward as a new service mode for big data management.

We propose virtual dataspace (VDS) model for data management according to the actual need in scientific domain [2]. The VDS model is regarded as the development and innovation of DataSpace. VDS refers to a set of subject, data, and flexible services constructed by virtual technologies. VDS masks the complexity of data integration, data relationship, format diversity and provide various flexible data services on the basis. That is what we called virtualization. There are two major issues in the VDS model: 1）Semantic integration of heterogeneous data source. 2）data relationship construction and data evolution. Local ontologies are created from data sources. Then the local ontologies are mapped and formed a global ontology. Based on this, access log and user feedback are considered for data evolution. At last, a material scientists-oriented service (materials scholar assistant) is introduced as the use case of VDS.

The paper is structured as follows. We state the challenge of data management in section 1. Section 2 provides the reader with related research while section 3 explains the technical details along with the realization of VDS. An application case of data evolution in VDS is introduced in section 4. Finally, the conclusion and remarks on future work are given in section 5.

II. RELATED WORK

As the challenges raised for traditional data management systems, a new data management technology has become a key necessity. M. Franklin proposed the idea of dataspace to tackle data management issue in 2005[3]. Since then considerable attention has been given to dataspace research concerning data model [4,5], link data[6,7] and data query[8]. Research about dataspace is still in its infancy, there are only a few prototype systems such as iMeMex[9], Semex[10] and OrientSpace[11,12,13].

J. Dittrich et al. proposed a platform for personal dataspace management [9]. They defined a unified data model that allows the integration of information in distributed and heterogeneous data sources. A new search and query language over this data model was also developed along with algorithms for the efficient processing of complex queries. This system has developed a queryable and updatable view on the user's personal information. However, semantic query is not supported by iQL.

A Domain Model-based personal information management system known as Semex was developed by Dong X. et al.[10]. Semex consists of 3 sub-modules which are domain management module, data collection module and

978-0-7695-5044-2/13 $26.00 © 2013 IEEE
DOI 10.1109/EIDWT.2013.5

data analysis module. Domain Model provides the logical view of one's personal information, based on meaningful objects and associations. The data collection module is responsible for data extraction, integration, cleaning and indexing. The data analysis module analyzes data for search and browsing. Semantic information was considered when integrating user's personal information through Domain Model. Unfortunately, a bad domain model will seriously affected systemperformance.

Li Y introduced "core space" model and "task space" model for constructing the personal dataspace (PDS) management system OrientSpace[11,12,13]. OrientSpace is a user-centered system which takes Objective Semantic Link (OSL) and Memory-based Semantic Link (MSL) to describe PDS. Data evolution based on the analysis of user behavior facilitates better services. Although behavior of subjects has been fully considered in the system, it's only for personal information management and has very limited applicability.

As a new data management technology, research of dataspace is still in early stages. The number of prototype system is rather limited, let alone scientific domain-specific model. Dataspace model in scientific domain is a fertile field for study.

III. VIRTUAL DATASPACE MODEL

A. *Virtual Dataspace*

Dataspace is a new data management model which consists of subject, data and data relationships. The VDS model is regarded as the development and innovation of DataSpace. VDS is a set of subject, data, and flexible on-demand services constructed by virtual technologies. VDS mask the complexity of data integration, data relationship and format diversity and provide various flexible data services on the basis. That is what we called virtualization. As shown in Fig.1, VDS consists of semantic integration module, data evolution module and data service module.

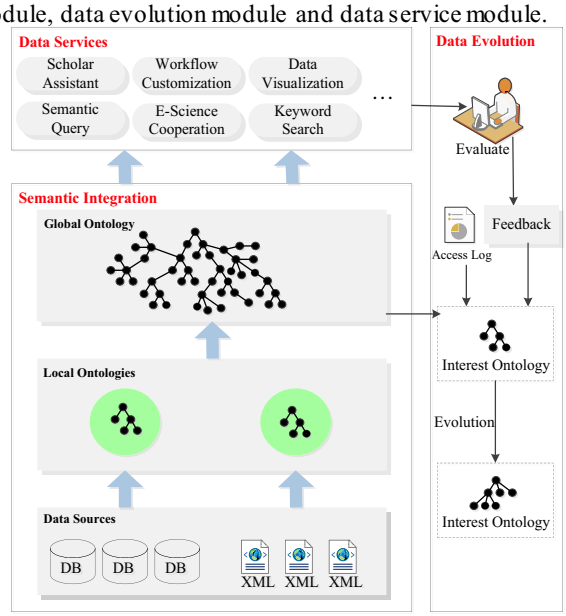

Figure 1. Virtual Dataspace Framework.

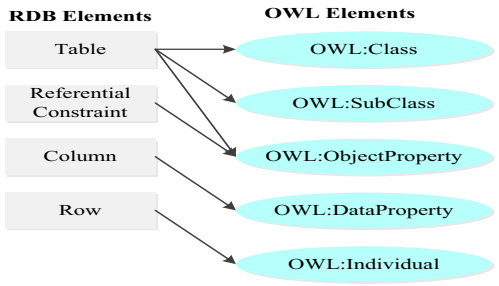

Figure 2. Mapping between Database elements and OWL elements.

B. *Semantic Integration of Virtual Dataspace*

There are several data formats in VDS such as structured, semi-structured and non-structured files. We complete data integration of diverse format based on ontology technology.

1) *From Relational Database to Local Ontology*

Relational schema is the mode taken by relational database systems to manage and organize data. Relationships between entities are represented by relations which are actually refer to two-dimensional tables. Rows in the tables are called tuples and columns in the tables are called attributes [14]. Relational database is regarded as the composition of tables, rows, columns and constrains.

Ontology is a concept introduced from philosophy area by experts of artificial intelligence domain. Asuncion[15] et al. reorganized the concept of ontology and proposed a quintuple method to describe it. The quintuple consists of class, relation, function, axiom and instance.

We define five rules to transform relational databases to OWL ontology based on the definition of them, as shown in Fig.2.

In this paper, extraction from database to OWL ontology is divided into two parts: the information extraction part and the db-ontology transformation part. The framework is shown in Fig.3.

The related schema information extracted from database is regarded as the semantic information of ontology. Both of schema and data are transferred to the ontology mapper. Related ontologies are automatic created according to the rules defined in Fig.2 and the information in the ontology mapper.

2) *From XML to Local Ontology*

XML is a standard for data exchange proposed by W3C in 1998, it's a serious of markup languages. XML has advantages such as self-describing, scalability, separation of data and representation and easy for programming. DTD

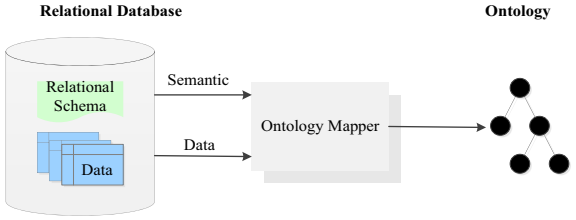

Figure 3. Transformation from Database to OWL Ontology.

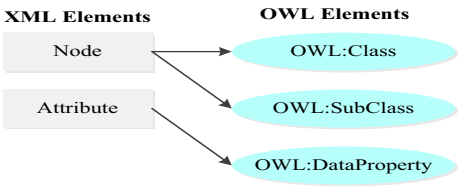

Figure 4. Mapping between XML Elements and OWL Elements.

(Document Type Definition) and XML Schema are two methods to describe meta data in XML. We adopt XML Schema to describe XML in VDS.

XML element includes node and attribute. Mapping from XML to OWL ontology is divided into content mapping and relationship mapping. Fig.4 shows the mapping between XML elements and OWL elements.

Content mapping: XML nodes are mapped into OWL classes and XML attributes are mapped into OWL data properties. Relationship mapping: parent-child relationships in XML are mapped into class-subclass relationships of OWL ontology and element-attribute relationships in XML are mapped into class-data property relationships of OWL ontology.

3) From Local Ontology to Global Ontology

The global ontology solves the problem of heterogeneity among data sources or local ontologies and provides a unified view for user. It unifies semantic expression of domain vocabulary based on the mappings among local ontologies.

In VDS, transformation from local ontologies to global ontology is complete by ontology mapping[16]. Ontology mapping is an iterative process which mainly includes feature extraction, user interaction, similarity computing and mapping discovery.

Feature extraction: Information such as concepts, individuals and relations are extracted from heterogeneous ontologies.

User interaction: Some concept pairs are mapped manually before or after the mapping process.

Similarity computing: Similarities of concept pairs are calculated by the compositive measure. More details for similarity computing is described in [16].

Mapping discovery: The threshold is set to 0.7. The concept pairs of which the similarity value in the compositive matrix is greater than 0.7 will be mapped.

Iteration: The accuracy of ontology mapping will be improved by the iteration of the process.

Fig.5 shows the mapping process.

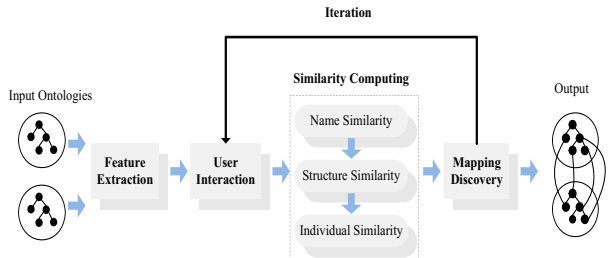

Figure 5. Ontology Mapping Process.

Figure 6. ITEMS OF ACCESS LOG

Item	Explanation
ID	register ID of the user
IP	IP of the user
URL	visiting URL
Time	visiting time

C. Data Evolution in VDS

Being trapped in the ocean of data, the users are bothered with many unrelated information. Existing data service has poor user experience as they simply provides data to users without semantic. Thus in VDS the semantic relationships are constructed considering the information in access log. Besides, data evolution is realized according to user feedback.

1) Access Log-based data relationship

In order to reduce interference of unrelated data, the user's interests are mined by access log analyzing. Take visits to the website of materials as an example, the user must be interested in 'ferrous material' if most visits were made to 'ferrous node'. In such a case more ferrous material and related information would be recommended to him. That's part of the personalized service we provided in VDS.

Access log records the behaviors of the users for every visit. In VDS, ID, IP, URL and visit time are recorded. Table I shows the detail explanations of each item.

Relevant information would be recommended to the user based on access log analyzing. As shown in Fig.6, information recommendation consists of pre-processing, similarity computing, record discovery and ontology creating.

Note that pre-processing refers to the process of information extraction from access log. The concept which the user is interested in was extracted from the access log. Then similarity between domain ontology concepts and interest concepts is computed. The concept pair whose similarity value is greater than the threshold is adopted to generate interest ontology. Similarity computing is based on (1).

$$Sim(Ci, Wi) = \frac{\omega_1 Sim_1(C_i, W_i) + \omega_2 Sim_2(C_i, W_i) + \cdots + \omega_n Sim_n(C_i, W_i)}{\omega_1 + \omega_2 + \cdots + \omega_n} \quad (1)$$

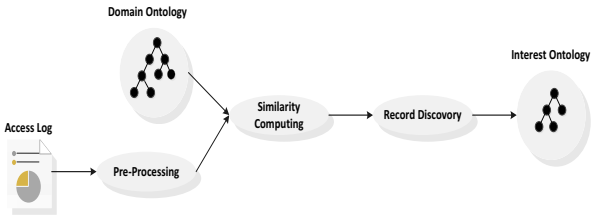

Figure 7. Access Log Processing Procedure.

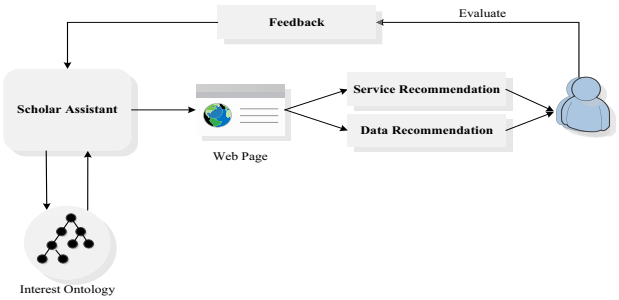

Figure 8. User Feedback Process.

Note that Ci refers to the concept of domain ontology and Wi refers to the concept extracted from access log. Sim_1 (C_i,W_i) 、 Sim_2 (C_i,W_i) ⋯ Sim_n (C_i,W_i) represent similarity of Ci and Wi computed based on different rules. ω_1 、 ω_2⋯ ω_nis computed by the sigmod function[sigmod]. Four rules are defined for similarity computing.

Rule 1: if Wi is similar to Ci's parent concept, Wi is similar to Ci.

Rule 2: if Wi is similar to Ci's children concept, Wi is similar to Ci.

Rule 3: if Wi is similar to Ci's sibling concept, Wi is similar to Ci.

Rule 4: if Wi is similar to Ci's individual, Wi is similar to Ci.

2) User Feedback-based data evolution

User feedback is the motive power and important basis of data evolution in VDS. Data and services are provided to the users and user feedbacks are accepted at the same time. User feedback process is shown in Fig.7.

IV. APPLICATION CASE

Considering the application requirement of material scientific domain, we developed VDS-based material scientific data sharing platform. The platform has integrated massive, heterogeneous and diverse data from more than 30 distributed organizations. Unified data sharing service and flexible services is constructed on that basis. Below we'll discuss the material scholar assistant as an application case.

Figure 9. Initialization of Interest Ontoloy.

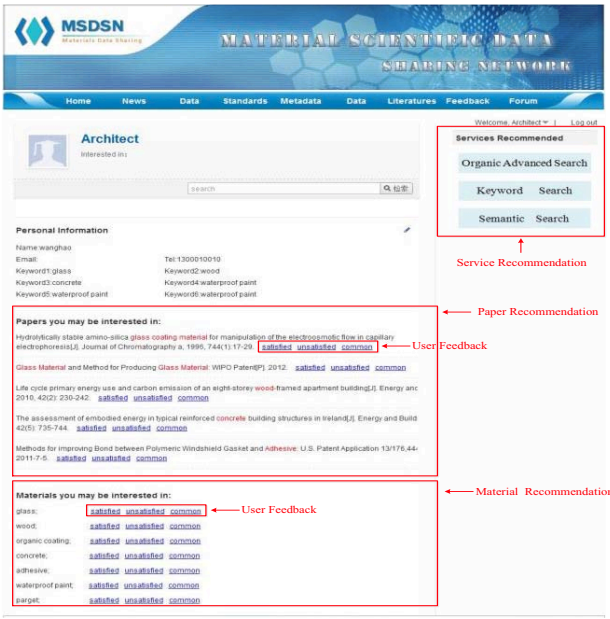

Figure 10. Data Evolution.

Material scientists need to look for desired materials in the ocean of data as soon as possible. For an architect, he may be interested in building materials and organic materials. A domain knowledge base should be constructed at first by integrating massive, distributed, heterogeneous and diverse data to ontologies. On that basis, a user-centered material scholar assistant service is completed according to access log and user feedback analyzing.

Firstly, the information provided while registering is collected and formed an interest ontology, as shown in Fig.8.

Then related data and services are recommended to the user based on the interest ontology. Finally, evaluations of the recommended data are adopted as the foundation of data evolution in VDS, as shown in Fig.9.

V. CONCLUSION

Dataspace management system is increasingly attracting attention as an area of study. The VDS proposal is for big data management in science domain. Firstly, local ontologies are created from data sources. Secondly, the local ontologies are mapped and formed a global ontology. Based on this, access log and user feedback are considered for data evolution. At last, a material scientists-oriented service (material scholar mate) is introduced as the use case of VDS. The experimental result has verified the effectiveness of VDS for data management in material domain. In addition, our work will be developed further in the future: more optimizations will be taken on VDS model and the algorithms for better personalized service.

ACKNOWLEDGMENT

The work in this paper is supported by the R&D Infrastructure and Facility Development Program under Grant No. 2005DKA32800, the 2012 Ladder Plan Project of Beijing Key Laboratory of Knowledge Engineering for

Materials Science under Grant No. Z121101002812005, the Key Science-Technology Plan of the National 'Twelfth Five-Year-Plan' of China under Grant No.2011BAK08B04, and the National Key Basic Research and Development Program (973 Program) under Grant No. 2013CB329606.

REFERENCES

[1] Gantz, John, and David Reinsel, "The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," Technical report, IDC, 2012.

[2] Zhenyu Liu, Changjun Hu, Yang Li and Yi Huang, "DSDC: A Domain Scientific Data Cloud Based on Virtual Dataspaces," Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International. IEEE, 2012.

[3] Franklin, Michael, Alon Halevy and David Maier, "From databases to dataspaces: a new abstraction for information management," ACM Sigmod Record, vol.4 , pp.27-33, 2005.

[4] Dan Yang, Derong Shen, Tiezheng Nie, Ge Yu and Yue Kou. "Layered graph data model for data management of dataspace support platform." Web-Age Information Management. Springer Berlin Heidelberg, 2011, pp.353-365.

[5] Sarma, Anish Das, Xin Luna Dong, and Alon Y. Halevy, "Data modeling in dataspace support platforms," Conceptual Modeling: Foundations and Applications. Springer Berlin Heidelberg, 2009, pp.122-138.

[6] Curry Edward, "System of systems information interoperability using a linked dataspace," System of Systems Engineering (SoSE), 2012 7th International Conference on. IEEE, 2012.

[7] Heath, Tom and Christian Bizer, "Linked data: Evolving the web into a global data space," Synthesis lectures on the semantic web: theory and technology , vol.1 (2011), pp.1-136.

[8] Li, Yukun and Xiaofeng Meng, "Supporting context-based query in personal DataSpace," Proceedings of the 18th ACM conference on Information and knowledge management. ACM, 2009.

[9] Dittrich, Jens-Peter, M. Salles and S. Karaksashian, "iMeMex: A platform for personal dataspace management," SIGIR PIM Workshop. 2006.

[10] Dong, Xin Luna, and Alon Halevy, "A platform for personal information management and integration," Proceedings of VLDB 2005 PhD Workshop, pp.26.

[11] Li, Yukun and Xiaofeng Meng, "Exploring Personal corespace for dataspace management," Semantics, Knowledge and Grid, 2009. SKG 2009. Fifth International Conference on. IEEE, 2009.

[12] Li, Yukun, Xiaofeng Meng and Yubo Kou, "An efficient Method for constructing personal dataspace," Web Information Systems and Applications Conference, 2009. WISA 2009. Sixth. IEEE, 2009.

[13] Li, Yukun and Xiaofeng Meng, "Research on personal dataspace management," Proceedings of the 2nd SIGMOD PhD workshop on Innovative database research. ACM, 2008.

[14] Shan W and Shixuan S, "Introduction to database systems," Beijing: Higher Education Press, 2008, pp.46-47.

[15] Gómez-Pérez, Asunción and V. R. Benjamins, "Overview of knowledge sharing and reuse components: Ontologies and problem-solving methods," IJCAI and the Scandinavian AI Societies. CEUR Workshop Proceedings, 1999.

[16] Wei Lin, Changjun Hu, Yang Li and Xin Cheng, "Similarity-based Ontology Mapping in Material Science Domain," 9th International Conference on Web Information Systems and Technologies, Germany, 2013, unpublished.