Future Generation Computer Systems 54 (2016) 456–468

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

A Virtual Dataspaces Model for large-scale materials scientific data access

Changjun Hu [a,b], Yang Li [a,b,*], Xin Cheng [a], Zhenyu Liu [c]

[a] School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China
[b] Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, 100083, China
[c] IBM Global Business Services, Beijing, 100101, China

HIGHLIGHTS

- We define a model and architecture for the Virtual Dataspace to organise and manage large-scale, diverse materials scientific data.
- We introduce the evolution algorithm for the model based on the life cycle of the data resources.
- A scalable dataspace system is designed for managing, exploring and sharing materials scientific data.

ARTICLE INFO

Article history:
Received 14 May 2014
Received in revised form
16 December 2014
Accepted 4 May 2015
Available online 14 May 2015

Keywords:
Data access
Virtual Dataspace
Scientific data
Materials engineering

ABSTRACT

It is increasingly common to encounter materials researchers engaged in the collaborative analysis and transformation of large-scale scientific data over extended periods of time. A scalable system for managing, tracing, exploring and communicating the analysis of diverse scientific data is required for these researchers. Nowadays dataspaces systems offer a pay-as-you-go approach to data management, which offer services on the data in place, without losing the context surrounding the data. Thus, we define a model and architecture for a Virtual Dataspaces (VDS) capable of addressing this requirement. The automatic construction process of the Virtual DataSpaces Model (VDM) is described in detail for effectively organising multi-source and heterogeneous data resources. Furthermore, the dynamic evolution algorithm of VDS is analysed and designed for timely tracking of the life cycle of the data resources. Such a system could bring increases facilitating discovery, understanding and sharing of both scientific data resources. An application case in the field of materials engineering is described to evaluate the effectiveness of the proposed model.

1. Introduction

In Materials Science, data has become one of the fastest growing data sets all over the world today. Its growth are attributed mainly to theoretical study, scientific experiments, simulation, manufacturing, etc. Materials scientific data tend to be particularly heterogeneous in terms of their type and source compared with data encountered in other fields. Data sets are considered "big" when they are large, complex, and difficult to organise and process. Nowadays, Big Data has been proposed and called the cadenza of the third wave [1]. However, the concept of Big Data has arisen due to the new challenges of information evolution for large-scale data management, reusing and sharing [2,3]. Materials science data possess the typical characteristics of big data (see Fig. 1). Advances in materials science data management and analysis have placed the field on the verge of a revolution in how researchers conduct services, analyse properties and even discover new materials. The main issue of materials data service is how to discover valuable knowledge from distributed, massive and heterogeneous data and relieve the users from this information overflow problem.

In materials engineering, the bridge between effective data management and intelligent data services is needed. The data services are built based on user requirements, which include (1) reducing unrelated data and obtaining data of interest from various complex resources; (2) exploring the complex relationships and implied knowledge from data sources; (3) composing data queries using field-specific terminology; (4) collecting data according to user requirements. Thus different users need

* Corresponding author at: School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), Beijing, 100083, China.
E-mail address: mailbox.liyang@gmail.com (Y. Li).

http://dx.doi.org/10.1016/j.future.2015.05.004
0167-739X/

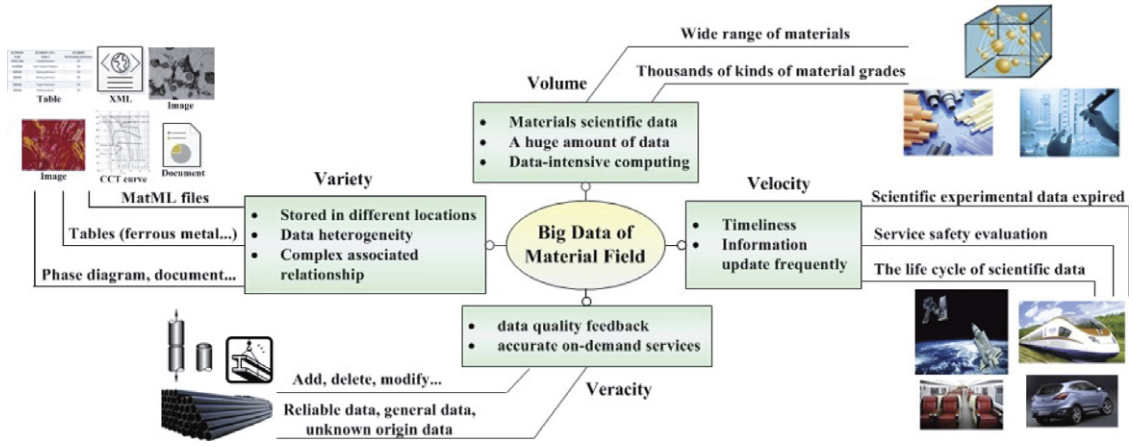

Fig. 1. Materials science data possess the typical characteristics of big data.

different application services, different services require different data, and one service even requests different data in different stages. The traditional database management mode has been unable to meet the variety of demands for data services. Therefore, a new data organisation service model is needed for providing data that user is interested and intelligent services require.

In this paper, we mainly focused on the modelling and evolution method for Virtual DataSpace (VDS) [4], the concept of which has been introduced in our previous work. VDS is developed based on dataspace [5], and it is used for complex data management and dynamic service demand in the field of materials engineering. In our work, we proposed a new model based on VDS to organise and process big data. Furthermore, a dynamic evolution method is introduced to manage constantly changing data according to intelligent services. The conceptual model of VDS is constructed based on the data life cycle. The process and algorithm of dynamic evolution for VDS are explored based on user feedback. Finally, the effectiveness of the model and the evolution algorithm are verified by the case study of the domain application and the comparison of different models.

The remaining part of this paper is arranged as follows. In Section 2, the previous work about dataspace is investigated. In Section 3, the modelling theory and method for VDM are described. In Section 4, the dynamic evolution process of VDS is elaborated and analysed. In Section 5, the application case for materials scientific data access in the field of materials engineering is described. Finally, the conclusion and further work are presented in Section 6.

2. Related work

The theoretical studies and practical studies on big data management have become increasingly important given the background of new challenges. An overview of the research issues and achievements in the field of big data analysis has been provided and the multidimensional data analysis problems about big data have been discussed [6]. A new algorithm with the help of a semantic graph for the more efficient and intelligent processing big data has been proposed [7]. However, because the semantics of big data are only described with RDFs, it has limited semantic representation ability. Active data [8] as a programming model has been proposed to alleviate the complexity of the data lifecycle and automatically improve the expressiveness of data management applications. However, this model does not consider the semantic characteristics and does not form a refined ideology about dynamic evolution.

Methods based on dataspace [5] as a new mode of data service have been explored by many researchers. A variety of dataspace models and prototype systems have been designed in different domains.

2.1. Dataspace model

The main purpose of dataspace is to collect and organise big data. Dataspace technology could be used in a wide range of areas, such as Personal Information Management (PIM), organising and processing scientific or engineering data, social network, etc. Dataspace model is the foundation for dataspace construction. The existing dataspace models include:

1. **iDM (iMeMex Data Model):** The iDM is the dedicated data model of the iMeMex system [9]. It organises and expresses all of the personal data resources in the form of a resource view and a resource view class. The iDM is the first dataspace model that is able to describe heterogeneous personal data resources in a unified form. However, this model uses a new query language, iQL, which is based on XPath and an SQL-like query language. For ordinary users, it is difficult to get started quickly.
2. **UDM (Unified Data Model):** UDM is a data model that is suitable for desktop search systems [10]. UDM adopts the database/information retrieval (DB/IR) integration approach that is able to dive into data items to retrieve the desktop dataspace. However, its new query language TALZBRA is very complex, and this model does not support shortcut queries.
3. **Probabilistic Semantic Data Model (P-DM):** P-DM is a dataspace model completely based on probability [11]. P-DM uses the probabilistic mediate schema [12] and probabilistic semantic mapping to achieve the semantic integration of heterogeneous data sources. This model addresses the problem of uncertainty [13,14] in different levels of dataspace and supports the top-k query response that could improve the quality of queries. However, its schema matching probability is not very accurate, and the model is difficult to extend.
4. **Domain Model:** The Domain Model is a dataspace model that is similar to the ontology method [15,16]. It supports a simple semantic query operation, but the mapping between the domain model and the data sources needs to be constructed manually.
5. **CoreSpace Model and TaskSpace Model:** The CoreSpace Model and the TaskSpace Model are the core parts of the personal dataspace management system, OrientSpace [17–19], which automatically constructs a personal dataspace. This system considers the behaviour characteristics of the subject, and