A scientific data extraction architecture using classified metadata

Yue-Shan Chang · Hsiang-Tai Cheng

Published online: 13 August 2010
© Springer Science+Business Media, LLC 2010

Abstract Data extraction and information retrieval from a great volume of data set always is a tedious and difficult work. Therefore, an effective and efficient technology for searching for desired data becomes increasingly important. Since metadata with certain attributes may characterize data files, to extract data with the help of metadata can be expectably to simplify the work. Metadata Classification has been proposed to improve significantly the performance of scientific data extraction. In this paper, a scientific data extraction architecture based on the assistance of metadata classification mechanism is proposed. The architecture is built by utilizing mediator/wrapper architecture to develop a scientific data extracting system to help oceanographer analyzing ocean’s ecology. The result of performance evaluation shows that the architecture with the help of metadata classification can extract user’s desired data effectively and efficiently.

Keywords Metadata · Metadata classification · Data extraction · Argo · Ocean data

1 Introduction

Many large scientific data archives manage and store huge quantities of data, deal with this data throughout its life cycle, and focus on particular scientific domains. Scientists can utilize these data in relevant research. Data extraction from a great volume of data set always is a tedious and difficult work. Therefore, an effective technology for searching for and extracting desired data becomes increasingly important.

Y.-S. Chang (✉) · H.-T. Cheng
Department of Computer Science and Information Engineering, National Taipei University, 151, University Road, Sanhsia, Taipei County, 237, Taiwan
e-mail: ysc@mail.ntpu.edu.tw
url: http://web.ntpu.edu.tw/~ysc/
H.-T. Cheng
e-mail: edisoncheng3918@gmail.com
There was representative research proposed in last decade [1, 3, 11, 15, 19]. Most of proposed approaches utilized mediator/wrapper architecture to access various information sources. These works offered effective technologies and tools for extracting desired data.

It is well known that metadata is “data about other data,” of any sort in any media. An item of metadata [18] may describe an individual datum, or content item, or a collection of data including multiple content items and hierarchical levels, for instance, a database’s schema. Metadata, in general, can be used for assisting data extraction or information retrieval, especially in the domain of multimedia retrieval [12], digital archive project [10], and scientific data search [18]. Using metadata to represent the file system also minimizes the processing required to handle operations. Metadata management in scientific domain is always an active and important research issue [8, 13, 18, 20]. Obviously, taking advantage of metadata management for retrieving user concern data from scientific dataset is an effective and efficient approach in recent years.

Recently, climate variability and change raise increased attention internationally. Understanding and predicting climate variability becomes an important work in the era. A global ocean observation project, name Argo,\(^1\) can help the climatologist to understand and to predict the climate variability [14] and help the oceanographer to analyze the ocean’s ecology by means of temperature, salinity, and other information; Undoubtedly, these needs to access and analyze massive data during a long duration and in a large range. Therefore, efficient accessing Argo data [5–7, 14] can speed up the predication and analysis.

Many existing systems\(^2,\(^3\) [8, 13, 15, 17, 20] have been proposed and have provided similar functionalities for data extraction. Most, however, share one or more of the following problems:

1. They do not support flexible query and extraction interface for users.
2. They do not provide flexible, extensible, and scalable architecture for scientific data extraction.
3. They do not provide metadata management by classifying the metadata and offer the help for data extraction based on the managed metadata.
4. They do not support efficient data extraction mechanism for scientific data.

In order to improve significantly extracting performance of scientific data, we have proposed a Metadata Classification (MC)\(^4\) to accelerating scientific data search and retrieving. The approach has been applied to Argo.

In this paper, we first present a Modified Metadata Classification (MMC) approach, which is the base of fast searching for the location of desired data. In order to provide a flexible and effective architecture to extract scientific data, we, based on the MMC, propose a scientific data extraction architecture that can tie the metadata classification mechanism to extract efficiently the searched file. The architecture is applied to the Argo project and builds an Argo data extraction system. The architecture is

\(^{1}\)http://www.argo.ucsd.edu/.
\(^{2}\)http://www.usgodae.org/cgi-bin/argo_select.pl.