Future Generation Computer Systems 45 (2015) 13–24

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

Automated preprocessing of environmental data

Mauno Rönkkö [a,*], Jani Heikkinen [b], Ville Kotovirta [c], Venkatachalam Chandrasekar [d]

[a] Department of Environmental Science, University of Eastern Finland, PO Box 1627, 70211 Kuopio, Finland
[b] CSC – IT Center for Science, PO Box 405, 02101 Espoo, Finland
[c] VTT Technical research Centre of Finland, PO Box 1000, 02044 VTT, Finland
[d] Electrical & Computer Engineering, Colorado State University, 1373 Campus Delivery, Fort Collins, CO 80523-1373, USA

HIGHLIGHTS

- A characterization based method for automated preprocessing of environmental data.
- A formalization of the preprocessing selection and sequencing problem.
- An algorithm solving the selection and sequencing problem.
- Simple case study implementation as a cloud service.

ARTICLE INFO

Article history:
Received 24 June 2013
Received in revised form
13 February 2014
Accepted 9 October 2014
Available online 22 October 2014

Keywords:
Environmental informatics
Workflows
Data preprocessing
Reachability analysis
Formal methods

ABSTRACT

In this article we discuss automated preprocessing of environmental data for further use. Environmental data is by default heterogeneous, as it may consist of data from sources such as weather stations, weather radars, chemical sensors, acoustic sensors, and off-line laboratory analysis. When integrating data from such heterogeneous sources, it needs to be processed in a context dependent manner. In addition, there is no single generic processing method; rather, several atomic methods need to be applied and in an appropriate sequence. Furthermore, the problem is complicated by the requirements set by the intended use of the data. The requirements influence not only the set of applicable methods but also the application sequence. In this article, we study automation of the selection and sequencing of preprocessing methods based on the user requirements. As the main contribution, we propose here the use of characterizations and a reachability algorithm to solve the selection and sequencing problem. In this article, we present the algorithm and argue for its correctness. We also discuss, how the algorithm is implemented as a cloud service, and illustrate the use of the service with simple case studies.

1. Introduction

There are currently many factors driving the design and development of environmental information systems. For instance, the global warming has set a future trend for cutting energy costs to reduce the carbon footprint. Consequently, monitoring and controlling energy consumption, greenhouse gas emissions, and energy wastage are of global interest. Another factor affecting the development of environmental monitoring systems has to do with safety and security; the modern society is vulnerable in case of natural disasters and intentional or involuntary incidents.

Not surprisingly, as pointed out by [1], "High-resolution, continuous, accurate monitoring of the environment is of great importance for many applications—from weather forecasting to pollution regulation". Environmental monitoring has, thus, become an attractive topic for state-of-the-art research. For instance, monitoring of the Fukushima incident is still relevant for impact assessment [2]. Similarly, continuous monitoring of factories [3] and power plants is critical as they may cause long term effects.

For monitoring purposes, data sources need to be reliable. When considering environmental data sources in particular, there is a significant variation in the quality. For instance, as pointed out by Williams et al. [4], data coming from weather measurement stations owned by individuals is hardly accurate. Consequently, Williams et al. propose the use of INTAMAP [5] and UncertML [6] to manage the uncertainties and to improve the accuracy of data. In many cases, however, the environmental data is corrupted or

* Corresponding author.
E-mail addresses: mauno.ronkko@uef.fi (M. Rönkkö), jani.heikkinen@csc.fi (J. Heikkinen), ville.kotovirta@vtt.fi (V. Kotovirta), chandra@engr.colostate.edu (V. Chandrasekar).

http://dx.doi.org/10.1016/j.future.2014.10.011
0167-739X/

even missing some measurement values. Consequently, there is research done on computational methods to somehow correct the measurement data. For instance, Junninen et al. [7] have studied various methods for imputing missing values for air quality data. The methods include linear interpolation, nearest neighbor interpolation, regression based imputation, and use of various neural nets. Similar methods have also been studied by others, for instance by Schneider [8] and Dixon [9].

When considering more elaborate environmental information systems, services, and assessment, the issues are more complex and instead of mere quality and reliability of data one should really start from the beginning and consider the applicability of the data to a specific need in the first place. The motivation for this is that the same data sources are used for multiple purposes. For instance, the measurement data from weather stations could be used not only for weather forecasting services, but also directly for HVAC control systems of residential buildings. Clearly, these two activities have a very different notion for data quality, and the same methods for imputing missing values, for instance, cannot be used in both cases.

Currently, the problem of using a single data source for multiple purposes is not well addressed. It is assumed that the user of the data is also responsible for preprocessing it, to fit the requirements. Consequently, there are multiple varying implementations of the same preprocessing methods for the same data. This leads to a cascading problem, where changes in the data source cause upgrading or reimplementation of all the services using the preprocessing methods. For instance, for environmental measurement systems, sensors may be replaced with other, more accurate ones. This is immediately reflected in the quality of the dataset. The data provider, however, may not inform about such a change, as it only improves the quality of the data. The problem here, however, is that some users of the system may use models assuming worse data quality, having thus issues with the improved data quality. Upgrading the models and services may not be trivial either, as they may include computational and ontology-based methods that require substantial training or revision for adaptation. Moreover, future services, such as those developed for residential building monitoring [10] or those based on proactiveness [11], cannot simply be manually adapted each time the data sources somehow change.

Thus, there is a growing need for the data sources to become themselves adaptive, to conform to the requirements set by the users of the data. This, however, is non-trivial as the users of the data need not know what preprocessing methods are available and applicable to the data. Thus, the users of the data should be able to express their requirements in a more abstract manner, so that the stated requirements are method independent. In this way, if the data sources change, there is no cascading effect; instead, changes can be fully managed by the providers of the data sources.

In this article, we propose the use of characterizations for expressing the user requirements on environmental data. For clarity, we limit characterizations here to Boolean expressions. Thus, we explicitly assume that the user knows what constraints must be set on the data in order to be applicable. In the case of environmental data, such an assumption is often plausible, as the user is often either a researcher or a consultant who needs the data for a specific purpose. For instance, the user may state that weather data for a specific region needs to be processed in such a way that it is not averaged, but all the values are limited to a given range and resolution and all outliers are removed, and that the data is given in a specific XML format. This requirement can be expressed as a Boolean expression in a method independent manner. However, this approach also implies that the data provider must not only support use of characterizations, but also provide a catalog of supported characterizations. As for environmental data, the data provider is often the domain expert and has the best knowledge of what preprocessing methods can be applied on the data and how those methods should be implemented. Thus, a catalog of preprocessing methods would also help the data provider to document their services. Furthermore, all this can be provided as a cloud service by using the standard interfacing technologies such as SOAP and REST.

As the main contribution, we study how characterizations on data and on the processing methods can be used to automate the selection and sequencing of the processing methods to meet the user requirements. For this purpose, we study the use of a simple reachability algorithm. Reachability algorithms are typically used in model-checking [12,13], to detect faulty sequences of events. Here, however, we reverse their purpose to detect a correct sequence of processing method applications. By automating the selection and sequencing of processing methods, both the providers and the users of the data sources are freed from considering how specific requirements are met for each individual case. In particular, the data provider can then focus on considering which methods are applicable to the data and how such methods are implemented efficiently.

We also exemplify the use of the algorithm with simple case studies. In the case studies, we consider preprocessing of indoor temperature measurements. For the case studies, we have implemented the algorithm as a cloud service by using the SOAP technology [14]. Thus, the implementation supports interaction not only with human users but also with computational services.

The rest of the article is organized as follows. In Section 2, we discuss environmental data and sequencing of computational processes in more detail. In Section 3, we discuss the characterization of data and processing methods along with the use of a reachability algorithm to solve the selection and sequencing problem. In Section 4, we present the case studies and the conclusion follows finally in Section 5.

2. Background and related work

Environmental data is mostly observational data. It is typically either raw or processed measurement data. It may also be computational data that is based on some environmental model, like a weather forecast model, for instance. Environmental data can also be aggregate data combining measurements either from the same phenomenon or from some different phenomena. Environmental data is also always spatio-temporal, meaning that it is always location and time dependent.

Because environmental data is inherently heterogeneous, the importance of using standards for representation cannot be overstated. Regarding spatial data, Open Geospatial Consortium provides many standards for representation and access of the data. In particular, the OGC standards provide a standardized terminology for representing specific aspects of the data. The newest proposal to the OGC standards tackles the core issue of any measurement data: uncertainty representation. For this purpose, UncertML [6] is proposed. Without uncertainty estimates, environmental data is of no practical use, as there is no reasonable way to estimate how reliable the obtained results are.

When considering environmental data, UncertML addresses the uncertainty of the dataset as a whole. For statistical methods this is enough; however, for model based methods, this may not be sufficient as model based methods may require some quality flagging on the level of individual measurement values. Such flagging is used, for instance, when gathering weather measurement data for forecasting [15].

Environmental data is not just multidimensional data containing measurement values of, for instance, temperature, wind speed,