Environmental Modelling & Software 94 (2017) 100–111

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Identification and characterization of information-networks in long-tail data collections

Mostafa M. Elag [a], Praveen Kumar [a, *], Luigi Marini [b], James D. Myers [c], Margaret Hedstrom [c], Beth A. Plale [d]

[a] Ven Te Chow Hydrosystems Laboratory, Department of Civil and Environmental Engineering, University of Illinois, Urbana, IL, USA
[b] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, USA
[c] School of Information, University of Michigan, Ann Arbor, MI, USA
[d] School of Informatics and Computing Director, Data to Insight Center, Indiana University Bloomington, Bloomington, IN, USA

ARTICLE INFO

Article history:
Received 18 November 2016
Received in revised form
24 March 2017
Accepted 28 March 2017

Keywords:
Long-tail data
Information-networks
Linked-data
Cyberinfrastructure
Environmental data
Data-intensive science

ABSTRACT

Scientists' ability to synthesize and reuse long-tail scientific data lags far behind their ability to collect and produce these data. Many Earth Science Cyberinfrastructures enable sharing and publishing their data over the web using metadata standards. While profiling data attributes advances the Linked Data approach, it has become clear that building information-networks among distributed data silos is essential to increase their integration and reusability. In this research, we developed a Long-Tail Information-Network (LTIN) model, which uses a metadata-driven approach to build semantic information-networks among datasets published over the web and aggregate them around environmental events. The model identifies and characterizes the spatial and temporal contextual association links and dependencies among datasets. This paper presents the design and application of the LTIN model, and an evaluation of its performance. The model capabilities were demonstrated by inferring the information-network of a stream discharge located at the downstream end of the Illinois River.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Scientific data collected or produced by individuals and small groups are generally targeted to address specific scientific issues and often have limited geographic or temporal coverage (Foster et al., 2012; Wagener et al., 2010; Elag et al., 2011). These data are usually irreplaceable, expensive to reproduce, infrequently reused, and have been defined as long-tail data (Swan et al., 2008; Heidorn, 2008; Kumar, 2006). As the volume of long-tail data increases, the scientific community is developing and maintaining CyberInfrastructures (CIs) to manage and share their data, including the Consortium for the Advancement of Hydrologic Science, Inc., Hydrologic Information System (CUASHI-HIS) (Tarboton et al., 2009), DataOne (Michener et al., 2011; Allard, 2012), Sustainable Environmental Actionable Data (SEAD) (Myers et al., 2015), and the Critical Zone Observatory Integrated Data Management System (CZOData) (Zaslavsky et al., 2011). These CIs rely on the Service-Orientated Architecture (SOA) to advance the Data as a Service (DaaS) principle, which allows publishing and delivering data over the web via standard coding of their metadata using markup languages and web services (Hey and Trefethen, 2005). In DaaS, a data node refers to any self-contained dataset that has a unique identifier and information profile, which explicitly depicts its attributes, behavior, and relationship with other nodes (Kumar, 2006). A large number of data nodes together constitute a data collection, i.e., a virtual aggregation of data nodes coming from different CIs. Identification of the contextual association links and dependencies among long-tail data nodes has immense value for increasing their integration and reusability, which are required to address multidisciplinary science challenges and advance scientific exploration (Horsburgh, 2014; Gahegan and Adams, 2014).

Contextual Association Links (CALs) are primitive and essential types of links that refer to potential associations among data nodes due to proximity in space, time, problem context, or any other user-defined criteria. CALs are required to build geoscience information-networks among data nodes distributed across different CIs and link data nodes with related environmental events. An Information-Network (IN) describes the CALs across the many attributes of independent long-tail data nodes. For example, understanding changes in the spatiotemporal patterns of sediment and

[*] Corresponding author.
E-mail address: kumar1@illinois.edu (P. Kumar).

http://dx.doi.org/10.1016/j.envsoft.2017.03.032
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

contaminant fluxes in a river due to climate and land use change requires an information-network that links hydrologic, water quality, biological, atmospheric, sediment, and nutrient monitoring data (Kumar, 2015). Similarly, understanding the changes in cropping pattern, which refers to changes in an area cultivated with various crops over a period of time, requires identification of the CALs among current and historical hydrologic, atmospheric, and geochemical data from soil samples along with sociocultural and economic data. Currently, building information-networks among distributed data nodes for such multidisciplinary studies is hindered by heterogeneity in coordinate systems, scales, provenance, units, variables, providers, users, file formats, and scientific contexts (Elag and Goodall, 2013; Heidorn, 2008; Foster et al., 2012; Ruddell et al., 2014).

Integration of distributed data nodes is a tedious and time-consuming process because a scientist has to search across distributed CIs, analyze heterogeneous information profiles associated with data nodes, identify the types of relationships among data nodes, and annotate the data nodes with their relationships. A long-tail data collection may include data that are generated from a variety of sources such as scientific experiments, field observations, model simulations, or derived from other data nodes. Diversity in data types, formats, and data models are the primary challenges in seamless integration of distributed data (Hey and Trefethen, 2005; Kumar, 2006; Elag and Goodall, 2013; Horsburgh et al., 2016). In addition, advances in instrumentation, sensing, computational approaches, and information exchange technologies, are rapidly increasing the production rate of such data. The combination of heterogeneity and high production rates of data increases the data-to-knowledge latency, that is, the production of knowledge by harnessing data lags far behind the data production itself (Foster et al., 2012; Fayyad et al., 1996; Elag et al., 2016). This latency hinders the integration and reusability of data (Fig. 1). Thus, an efficient data integration strategy requires multiple syntactic and semantic translations to manage across different domain CIs with various data models and identify CALs among distributed data nodes.

Indeed, overcoming the data-to-knowledge latency is essential for large multidisciplinary science projects such as the Critical Zone (CZ) science, where scientists are required to mange diverse data nodes from different CIs. Because many of the existing CIs have their data structure and data models implemented, one solution to overcome the challenge of data-to-knowledge latency is to build information-networks. The inclusion of new data nodes in a long-tail collection and/or modification of old data nodes continuously changes the IN, thereby, creating emergent characteristics with the potential to identify novel dependencies among the data nodes. Analysis of the emergent IN enables us to immediately identify the new data node in a broader context of other data nodes, while at the same time enhancing the value of existing data nodes. We argue, along with others (Bajcsy, 2008; Horsburgh et al., 2011; Zaslavksy et al., 2011), that building an IN of long-tail data will minimize the data-to-knowledge latency, potentially alleviate the problem of underutilized data resources, and foster the emergence of useful knowledge. This research focuses on two specific contextual association links —space and time—which identify and characterize the spatial and temporal relationships among scientific long-tail data using their information profiles. However, the model presented here is applicable for building CALs among data based on a broader set of their attributes such as information about data providers.

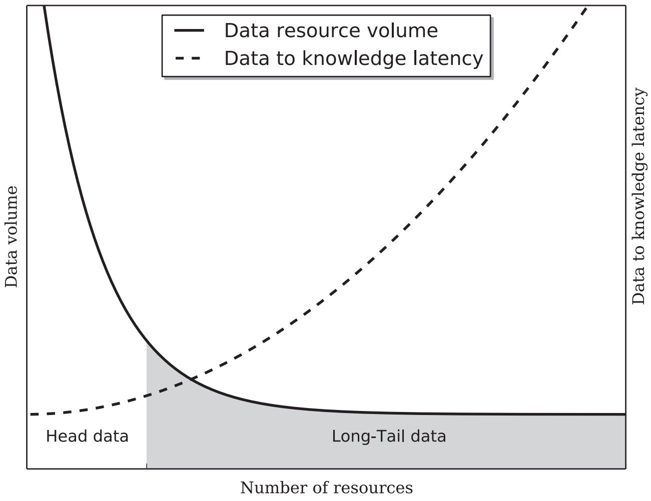

Fig. 1. The Grey zone refers to the scientific long-tail data that make up the bulk of the scientific research output. It has an exponential production rate compared to “Big data”, which are produced and maintained by very large projects and agencies. The knowledge required to reuse long-tail data is lagging far behind the data production (adapted from Foster et al. (2012) and Heidorn (2008)).

Prediction of a CAL between two data nodes is key to characterizing their emergent dependencies and establish an IN. Typical approaches for link prediction follow a content-based paradigm that relies on statistical inference of data connectivity based on the analysis of their numerical values. This paradigm ranges from feature-based grouping, which trains a binary classifier using a set of predefined features to classify data, to probabilistic graphical models, which use a Bayesian model to compute the joint-probability of a connection between data nodes (Al Hasan and Zaki, 2011). Many real networks including social and biological networks adopt this paradigm (Liben-Nowell and Kleinberg, 2007; Wang et al., 2007). Applying the content-based approaches to predict long-tail data linkages requires a considerable amount of effort and time to manually reformat large volumes of heterogeneous data nodes, where the source of heterogeneity cannot be anticipated in advance. In addition, sophisticated statistical techniques are required to minimize uncontrolled errors resulting from the interference of many factors in identifying the statistical dependency among data nodes such as sample size, missing data, and type of inter-variable dependencies (linear or nonlinear) (Newman, 2001). Therefore, there is a pressing need for a new context-based approach to solve these issues through rapid understanding of the spatial and temporal CALs among long-tail data.

In this research, we combined Information Retrieval (IR) methods, which are provided by scientific CIs with the Context-Aware Recommender System (CARS) to build the Long-Tail Information-Network (LTIN) model. First, information retrieval methods retrieve information profiles relevant to a specific search context. Then, CARS tailors the recommendation based on analyzing the retrieved information profiles using rules (Adomavicius and Tuzhilin, 2011). The present work is primarily focused on the challenge of building a recommendation model (engine), which is aware of the information profiles of distributed data nodes and can identify their CALs based on spatial and temporal rules. The model is introduced with two flavors: Long-Tail Information-Network (LTIN) full and lite. LTIN-full registers the information retrieval method of each CI and builds an IN among data nodes, which are available in the data collection. On the other hand, LTIN-lite wraps the LTIN model as a web service to build an IN between a single data node (given by the user) and nodes in the data collection on the fly. The model provides a semantic information graph that annotates its data nodes with their CALs, and characterizes the strength of the predicted CALs based on a strength matrix. The annotations are compliant with the Semantic Web standard format to allow consistent interaction among the data nodes distributed over the Internet (Fensel et al., 2011).

Download English Version:

https://daneshyari.com/en/article/4978172

Download Persian Version:

https://daneshyari.com/article/4978172

Daneshyari.com