information

MDPI

Editorial

Foreword to the Special Issue: "Semantics for Big Data Integration"

Domenico Beneventano and Maurizio Vincini *

Dipartimento di Ingegneria "Enzo Ferrari", Università di Modena e Reggio Emilia, Via Vivarelli 10, 41125 Modena, Italy; domenico.beneventano@unimore.it

* Correspondence: maurizio.vincini@unimore.it; Tel.: +39-059-205-6249

Received: 15 February 2019; Accepted: 15 February 2019; Published: 18 February 2019

Abstract: In recent years, a great deal of interest has been shown toward big data. Much of the work on big data has focused on volume and velocity in order to consider dataset size. Indeed, the problems of variety, velocity, and veracity are equally important in dealing with the heterogeneity, diversity, and complexity of data, where semantic technologies can be explored to deal with these issues. This Special Issue aims at discussing emerging approaches from academic and industrial stakeholders for disseminating innovative solutions that explore how big data can leverage semantics, for example, by examining the challenges and opportunities arising from adapting and transferring semantic technologies to the big data context.

Keywords: big data dimensions; big data technology; big data integration; semantic data design; information visibility

In recent years, there has been an enormous diffusion of big data technologies, usually oriented to data processing, omitting an equally important aspect related to the transformation of data to be ready for this process. In fact, it is increasingly urgent to address the issue of heterogeneity, diversity, and complexity of data, and how to normalize, integrate, and transform the data from many sources into the format required to run large-scale analysis [1].

This Special Issue addresses the research about big data management with semantic technologies as a unified data access layer and a consistent approach to analytic execution. Semantic technologies have been used to create domain models describing mutually relevant datasets and the relationships between them.

Among the submissions received, all of which went through a rigorous peer-review process, five papers have been selected for publication. In "A Hybrid Information Mining Approach for Knowledge Discovery in Cardiovascular Disease (CVD)", the authors combine different techniques of data exploration with the aim of extracting information and knowledge from a semantically rich dataset in the healthcare context [2]. In "High Performance Methods for Linked Open Data Connectivity Analytics" [3], the authors introduce scalable methods and algorithms to find all the connections among the Linked Open Data (LOD) Cloud datasets. In "LOD for Data Warehouses: Managing the Ecosystem Co-Evolution" [4], the authors propose and implement a management approach for the evolution of data warehouses with internal and external data published in LOD formats. In "Chinese Microblog Topic Detection through POS-Based Semantic Expansion" [5], the authors propose a topic detection method for Chinese microblogs, based on the semantic description of the microblog post according to a thesaurus. Finally, in "Integration of Web APIs and Linked Data Using SPARQL Micro-Services—Application to Biodiversity Use Cases" [6], the authors introduce methods to exploit the Semantic Web standards to enable automatic combination of disparate resource representations coming from both linked data interfaces and non-RDF (Resource Description Framework) Web APIs (Application Programming Interface).

The papers are described in more detail below.

1. A Hybrid Information Mining Approach for Knowledge Discovery in Cardiovascular Disease (CVD), by Stefania Pasanisi and Roberto Paiano

With the development of big data, there are huge amounts of semantically rich data in many domains that can be analyzed in order to provide useful knowledge using business intelligence techniques. This paper combines different techniques of data exploration with the aim of extracting information and knowledge from a rich dataset in the healthcare context. The results were obtained according to three distinct perspectives—descriptive, local, and predictive—as follows: (1) the stratification of the patient population into nine patterns obtained through a descriptive method, represented by a clustering algorithm; (2) relationships between attributes identified in the general dataset and within clusters through a local method—the association rule algorithm; and (3) the prediction of patients with aggravated conditions through a predictive method. An interesting outcome of the paper is to show that the combination of clustering techniques, neural networks, and association rules is an efficient strategy for information discovery and management of chronic disease.

2. High Performance Methods for Linked Open Data Connectivity Analytics, by Michalis Mountantonakis and Yannis Tzitzikas

The main objective of linked data is linking and integration, and a major step for evaluating whether this target has been reached—especially in a big data context—is to find all the connections among the Linked Open Data (LOD) Cloud datasets. Connectivity among two or more datasets can occur due to equivalence relationships between URIs, but, generally, it is not an easy task to find these connections due to the large number of LOD datasets. For this reason, the paper introduces scalable methods and algorithms for performing the computation of transitive and symmetric closure for equivalence relationships and for constructing dedicated global semantics-aware indexes that cover the whole content of datasets; moreover, the paper proposes connectivity measurements among two or more datasets. Finally, the speedup of the proposed approach and comparative results for over two billion triples are evaluated.

3. LOD for Data Warehouses: Managing the Ecosystem Co-Evolution, by Selma Khouri and Ladjel Bellatreche

In the context of the extraction and integration of knowledge from big data on the Web, one of the most promising aspects is the development of linked open data (LOD) as external sources to create added value and enrich the analytical capabilities of data warehouses (DWs). Building a DW with internal and external data published in LOD formats requires managing the evolution of DWs that deal with "open world" sources and their specific characteristics. Starting from the fact that conventional evolution approaches are adapted to neither this new type of source nor to semantic constructs underlying the LOD sources, the paper proposes a method that guarantees the traceability of DW constructs for the entire design cycle. The authors propose and implement a co-evolution management approach—based on consistent track maintenance—that identifies the impact of changes semantically for all the DW design phases; the goal of the approach is to provide a complete and semantic view of the real impact of the evolution event on the DW. The approach was successfully tested using the LUBM (Lehigh University Benchmark) and different LOD datasets (DBpedia, YAGO, etc.).

4. Chinese Microblog Topic Detection through POS-Based Semantic Expansion, by Lianhong Ding, Bin Sun, and Peng Shi

Recent progress made with web applications has witnessed the rapid development of microblogs, a new type of social media for the publishing, acquiring, and spreading of information. Microblogging represents one of the most important methods of people's online communications, especially for

sharing information. Finding the significant topics of a microblog is therefore an important task, in particular for popularity tracing and public opinion following. Since traditional methods showed a low performance with respect to a short text from a microblog, this paper proposes a topic detection method based on the semantic description of the microblog post; the proposed method is based on the semantic expansion of the microblog post according to a thesaurus and on clustering techniques for topic detection. The method was developed in the context of Chinese microblogs; the evaluation performed on the dataset from Sina Weibo has shown the scalability of the semantic method and how the method can lead to better results both for post-clustering and for the detection of arguments in Chinese microblogs.

5. Integration of Web APIs and Linked Data Using SPARQL Micro-Services—Application to Biodiversity Use Cases, by Franck Michel, Catherine Faron Zucker, Olivier Gargominy, and Fabien Gandon

The architecture proposed in this paper has the goal of exploiting the Semantic Web standards to enable the automatic combination of disparate resource representations coming from both linked data interfaces and non-RDF Web APIs. Consequently, the proposed system is based on the so-called micro-services—a lightweight SPARQL endpoint that provides access to a small, resource-centric virtual graph—and exploits the architectural principles of micro-services to define an architecture, aimed at querying the Web APIs using SPARQL. The successful use of SPARQL micro-services in several real-life use cases related to the biodiversity domain is then shown. We agree with the authors that this approach could promote the emergence of an ecosystem of SPARQL services published by independent service providers, allowing linked data-based applications to glean pieces of data from a wealth of distributed, scalable, and reliable services.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vincini, M. Analyzing mappings and properties in Data Warehouse integration. *Int. J. Eng. Technol. Innov.* **2017**, *7*, 4.
2. Pasanisi, S.; Paiano, R. A Hybrid Information Mining Approach for Knowledge Discovery in Cardiovascular Disease (CVD). *Information* **2018**, *9*, 90. [CrossRef]
3. Mountantonakis, M.; Tzitzikas, Y. High Performance Methods for Linked Open Data Connectivity Analytics. *Information* **2018**, *9*, 134. [CrossRef]
4. Khouri, S.; Bellatreche, L. LOD for Data Warehouses: Managing the Ecosystem Co-Evolution. *Information* **2018**, *9*, 174. [CrossRef]
5. Ding, L.; Sun, B.; Shi, P. Chinese Microblog Topic Detection through POS-Based Semantic Expansion. *Information* **2018**, *9*, 203. [CrossRef]
6. Michel, F.; Faron Zucker, C.; Gargominy, O.; Gandon, F. Integration of Web APIs and Linked Data Using SPARQL Micro-Services—Application to Biodiversity Use Cases. *Information* **2018**, *9*, 310. [CrossRef]