Advance Access publication on 12 September 2016
doi:10.1093/comjnl/bxw073

Data Science

Sebastian Maneth[1]* and Alexandra Poulovassilis[2]

[1] *School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK*
[2] *London Knowledge Lab, Birkbeck, University of London, London WC1E 7HX, UK*
**Corresponding author: smaneth@inf.ed.ac.uk*

Keywords: data science; network mappings; graph pattern matching; database fragmentation; event detection; transaction processing; type inference; medical ontologies

Received 18 August 2016; revised 18 August 2016
Handling editor: Fionn Murtagh

1. CONTENT OF THE ISSUE

The field of Data Science concerns techniques for extracting knowledge from diverse data, with a particular focus on ‘big’ data exhibiting ‘V’ attributes such as volume, velocity, variety, value and veracity. The field of data science is becoming increasingly influential in the public, private and voluntary sectors, with its overarching aim of increasing understanding of services, products and stakeholders in all areas of human activity. Techniques from data science are being developed and used in applied and interdisciplinary research across the biological, medical and physical sciences, the social sciences and the arts and humanities.

Key research challenges in data science include: the development of computational techniques that are able to scale to the volumes and varieties of the data generated by web-based, mobile and pervasive technologies, and to the rate at which data is being produced by large-scale business, social media and scientific applications; the development of data cleansing, transformation, modelling, analysis, integration and visualisation tools that allow data scientists to understand and improve the veracity of big data and to extract value from it quickly, easily and reliably; and ensuring organisations and users data security, privacy and ownership concerns. The articles in this section of the issue describe recent work in addressing some of these challenges. The excellence of these papers happens to have originated through BICOD, the 30th British International Conference on Databases (formerly known as BNCOD) which took place in Edinburgh, UK, on July 6–8, 2015, see [1].

Yang Cao, Wenfei Fan and Shuai Ma address the volume aspect of big data in their paper entitled ‘Virtual Network Mapping in Cloud Computing: A Graph Pattern Matching Approach’. Motivated by the need to support dynamically varying workloads over distributed data centres in cloud computing, this article addresses the problem of automatically mapping a virtual network to a physical network substrate in such a way that the desired capacity, bandwidth and latency constraints on the virtual network nodes and links are satisfied. The authors explore how this problem can be approached using graph pattern matching techniques, with the virtual network being expressed as a graph pattern and the physical network substrate as a graph. Complexity bounds are derived for different mapping constraints, and simulation experiments are conducted to investigate the practical effectiveness of a range of mapping algorithms.

Lena Wiese, Tim Waage and Ferdinand Bollwein address both the volume and variety aspects of big data in their paper entitled ‘A Replication Scheme for Multiple Fragmentations with Overlapping Fragments’. Large-scale distributed datasets are fragmented and distributed over multiple servers in order to achieve data locality and faster query processing through parallelization. In order to further increase locality and to improve fault-tolerance, fragments are typically replicated over multiple servers. In this article, the authors investigate efficient replication procedures in cases when there are several fragmentations of the same database table, motivating this through the scenario of query relaxation where it may be advantageous to materialize different clusterings of the data so as to efficiently support queries that are being relaxed with respect to different attributes. Support for query relaxation is desirable in settings where the data is complex and heterogeneous, such that the user may not be fully familiar with its structure and may find it hard to pose queries that exactly match the users information seeking requirements.

Andreas Weiler, Michael Grossniklaus and Marc H. Scholl address the volume and velocity aspects of big data in their paper entitled ‘Design and Analysis of Measures to Evaluate Event Detection Techniques for Twitter Data Streams’. The authors present a framework for automatically evaluating different event detection techniques over Twitter data streams. The motivation for such techniques is the potential to leverage the large volumes of tweets for extracting news items relating to topic areas such as natural disasters, disease epidemics and political events, as well as trending topics and sentiment analysis. The computational challenges of deriving such information from tweets arise from their volumes, rate of production, brevity, imprecision and variable linguistic quality. A set of measures that rely on external ground-truth services are described for comparing different techniques with respect to task-based performance measures such as duplicate event detection rate, precision and recall. Runtime measures such as throughput and memory consumption are also considered. To assess the effectiveness of the proposed measures in comparing and discriminating between different event detection techniques, a range of techniques are implemented over the same data stream management system and their performance on real Twitter datasets is determined using the proposed measures.

Yu Liu and Peter McBrien address the veracity aspect of big data in their paper entitled ‘Transactional and Incremental Type Inference from Data Updates’. Traditional database systems support transactional updates and the ACID properties of atomicity, consistency, isolation and durability. This article explores transactional reasoning over knowledge bases that are expressed in the OWL 2 RL ontology language and stored in a relational database. The authors consider a setting where the derived facts inferable from the stored facts and the ontology axioms are materialized in the database, so that queries are answered directly from this materialized knowledge base. When insertions or deletions are made to the ontology A-Box, the materialized knowledge base is updated using a set of database triggers generated from the ontology T-Box, hence extending the ACID properties of database transactions to the materialized results of ontology-based reasoning. A performance analysis is undertaken comparing the approach to two other state-of-the-art ontology reasoners, one non-materializing and one materializing, with respect to the LUBM benchmark.

Reem Q. Al Fayes and Mike Joy address the volume, variety and value aspects of big data in their paper entitled ‘Using Linked Data for Integrating Educational Medical Web Databases based on BioMedical Ontologies’. Motivated by the advantages and increasing impact of Linked Open Data, the authors consider the challenges of acquiring and integrating information from diverse web sources into one Linked Dataset. Their specific setting is that of Medical Education, and techniques are investigated for integrating large amounts of information from PubMed articles, YouTube videos and specialist blogs. They describe a system that is able to harvest heterogeneous metadata from such sources and to enrich it with annotations drawn from concepts in biomedical ontologies such as SNOMED CT, so as to facilitate the automatic linkage of entities across different web sources. The SNOMED CT ontology can then be used an entry point for users browsing and querying the integrated dataset.

ACKNOWLEDGEMENTS

The guest editors of this special issue thank Professor Fionn Murtagh for his valuable suggestions and help in preparing this issue. We thank all authors who have submitted their work to the issue. Special thanks go to all the reviewers for their thorough comments that helped in enhancing the quality of the papers.

REFERENCE

[1] Maneth, S. (ed.) (2015) *Data Science—30th British Int. Conf. Databases, BICOD 2015, Edinburgh, UK, July 6–8, 2015, Proc.*, Lecture Notes in Computer Science, **9147**. Springer.