Future Generation Computer Systems 76 (2017) 495–498

Contents lists available at ScienceDirect

Future Generation Computer Systems

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

A note on exploration of IoT generated big data using semantics

Rajiv Ranjan [a,b,*], Dhavalkumar Thakker [c], Armin Haller [d], Rajkumar Buyya [e]

[a] School of Computer, Chinese University of Geosciences, Wuhan, China
[b] School of Computing Science, Newcastle University, United Kingdom
[c] School of Electrical Engineering and Computer Science, University of Bradford, United Kingdom
[d] Research School of Computer Science, Australian National University, Australia
[e] Cloud Computing and Distributed Systems (CLOUDS) Laboratory, School of Computing and Information Systems, The University of Melbourne, Australia

A B S T R A C T

Welcome to this special issue of the Future Generation Computer Systems (FGCS) journal. The special issue compiles seven technical contributions that significantly advance the state-of-the-art in exploration of Internet of Things (IoT) generated big data using semantic web techniques and technologies.

1. Introduction

Recent studies have shown that we generate 2.5 quintillion (2.5×10^{18}) bytes of data per day (Cisco and IBM) and this is set to explode to 40 yotta (40×10^{24}) bytes by 2020 – this is 5,200 GB for every person on earth. Much of these data is and will be generated from IoT [1] devices such as sensors, RFIDs, remote sensing satellites, business transactions, actuators (such as machines/equipment fitted with sensors and deployed for mining, oil exploration, or manufacturing operations), lab instruments (e.g., high energy physics synchrotron), and smart consumer appliances (TV, phone, etc.), but also social media and clickstreams. This vision has recently given rise to the notion of IoT Big Data Applications (IoTBDAs) in domains such as Healthcare, Smart Cities, Smart Manufacturing, and Smart Energy Grids. These IoTBDAs are required to have novel capability (currently non-existent) of analyzing large number of dynamic data streams [2], tens of years of historical data, and static knowledge about the physical world (i.e. city map, road network map, utility network map, etc.) to support real-time and/or near real-time decision making.

The decision making process involving such big data applications often involves exploration for meaningful patterns and connections. Despite the rapid evolution of IoTBDAs; current generation of Cloud Computing and Big Data Processing techniques/frameworks (e.g., batch processing, stream processing, and NoSQL) lack the following very important abilities to support effective exploration: Several novel interfaces and interaction means for exploration of big data are being proposed, for example, exploratory search systems, data browsers, visualisation environments and knowledge graph-based search engines. Although on the rise, the current solutions are still maturing and can benefit from computational models that aid intuitiveness and improve the effectiveness of exploration tasks. The Semantic Web and its derivatives in the form of Linked data and Web of data can play a crucial role in addressing various big data exploration challenges.

The ability to discover semantic context [3–5] is one of the primary requirements as in regards to exploring and managing big data. Further, enriching data exploration techniques with semantic context information has been proven useful in inferring knowledge about what is exactly happening in the physical environment [6], which is being monitored and/or controlled, by IoT sensors and/or actuators. For example, identifying a malfunctioning IoT sensor by comparing data from nearby functioning sensor is critical, so the data exploration engine can ignore the data items captured by the malfunctioning sensors. Semantic context data can provide information about data quality, which can also have a direct impact on the final analytics results. Moreover, semantic context information can be used to develop efficient and effective data collection plans specially, when multiple sensors available nearby offer similar information.

The main goal of this Special Issue is to explore new directions and approaches about key research topics needed to leverage innovative research aimed at tackling big data exploration challenges in IoTBDAs, based on semantic technologies [7]. We encouraged the submission of work with important theoretical and practical results, as well as case studies on existing use of semantic technologies for big data exploration.

2. Summary of contributions

In this section, we present the summary of the papers that were accepted for publication in this special issue.

Large and complex spatial data can be leveraged for different purposes and in different IoTBDAs, however, traditional processing and analytical methods may not be fit-for-purpose. Such methods

* Corresponding author.
E-mail address: rranjans@gmail.com (R. Ranjan).

http://dx.doi.org/10.1016/j.future.2017.06.032
0167-739X/

generally use mathematical causal analysis techniques, which seek to find causes and effects of everything. In spatial big data, however, users only wish to obtain all possible related resources and data rather than why or how these correlations occur or the underlying classical theories. There are also many unknowns between cause and effect. Therefore, imprecise and non-classical methods have been used to realize similar and fuzzy retrieval of spatial big data based on semantics. Associative retrieval, on the other hand, has been identified as a potential technique for big data. In the paper titled "*Associative retrieval in spatial big data based on spreading activation with semantic ontology* (http://dx.doi.org/10.1016/j.future.2016.10.018)", Sun et al. integrate the spreading activation (SA) algorithm and the ontology model in order to promote the associative retrieval of big data. In their approach, constraints based on variant weights of semantic links are considered with the aim of improving the spreading-activation process and ensuring the accuracy of search results. Semantic inference rules are also introduced to the SA algorithm to find latent spreading path and help obtain results, which are more relevant. Their theoretical and experimental analysis demonstrate the utility of this approach.

With the development of mobile technology, the users browsing habits are gradually shifting from information retrieval to recommendation. The classification-mapping algorithm between a user's interests and web contents has become more and more difficult with the volume and variety of web pages. Some big news portal sites and social media companies hire more editors to label these new concepts and words, and use the computing servers with larger memory to deal with the massive document classification, based on traditional supervised or semi-supervised machine learning methods. Li et al. in the paper titled "*An optimized approach for massive web page classification using entity similarity based on semantic network* (http://dx.doi.org/10.1016/j.future.2017.03.003)", provide an optimized classification algorithm for massive web page classification using semantics networks, such as Wikipedia, WordNet. In this paper, they used Wikipedia dataset and initialized few category words as class words. A weight estimation algorithm based on the depth and breadth of the Wikipedia network is used to calculate the class weight of all Wikipedia words. A kinship-relation association based on content similarity is proposed to optimize the unbalance problem when a category node inherited the probability from multiple fathers. The keywords of web page are extracted from the title and main text using an N-gram with Wikipedia words, and a Bayesian classifier is used to estimate the page class probability. Experimental results shows that the proposed method has very good scalability, robustness and reliability for massive web pages.

Multimedia big data is difficult to handle because of its enormous amount and the elusive property of underlying information. To study how to explore valuable information among multimedia big data with low complexity, in the paper titled "*Object detection among multimedia big data in the compressive measurement domain under mobile distributed architecture* (http://dx.doi.org/10.1016/j.future.2017.03.004)", Guo et al. propose an object detection method of big data, which is in the compressive measurement domain under a mobile distributed computing architecture. It includes the sparse representation and object detection processes. Considering the unbalanced computation capacity between a mobile center cloud and mobile edge sites, they shift large storage burden into the cloud, while performing the dictionary learning by using compressive measurements in the mobile edge sites. Specifically, after getting the measurements at the edge sites, they perform dictionary learning to obtain the sparse representation in the pixel domain, then select significant images and their feature vectors to be stored in the center cloud. In addition, they also analyze the trained dictionary in the measurement domain employing measurements. In order to reveal the two kinds of dictionaries' relationship, they conduct a formulation process into each of them and find that the relationship depends on the uniqueness relation between the original signal and the sparse coefficient in the measurement domain. At the same time, they keep coefficients for a certain time period at the mobile edge sites in order to realize real-time object detection, taking advantage of low latency of the mobile edge computing ends. Since the sparse coefficients and the original signal have a one-to-one correspondence relationship, they can just search for the matched coefficients of the image block for detecting object. Experimental results show that Hadamard measurement matrix can better preserve the characteristics of the original signal than Gaussian matrix and that the proposed method can achieve a favorable detection performance. Meanwhile, the computation cost and storage cost of the proposed detection process can be significantly reduced compared with traditional methods, which is suitable for the multimedia big data.

It is very difficult to process large amount of structured and unstructured big data generated by IoTBDAs with traditional sequential programming methods. To this end, the paper titled "*A MapReduce-based Scalable Discovery and Indexing of Structured Big Data* (http://dx.doi.org/10.1016/j.future.2017.03.028)", by Singh et al., proposes a parallel B-Tree index and its implementation in the MapReduce framework for improving efficiency of random reads over the existing approaches. The benefit of using the MapReduce framework is that it encapsulates the complexity of implementing parallelism and fault tolerance from users and presents these in a user friendly way. The proposed index reduces the number of data accesses for range queries and thus improves efficiency. The B-Tree index on MapReduce is implemented in a chained-MapReduce process that reduces intermediate data access time between successive map and reduce functions, and improves efficiency. Finally, five performance metrics have been used to validate the performance of the proposed index for range search query in MapReduce, such as, varying cluster size and, size of range search query coverage on execution time, the number of map tasks and size of Input/Output (I/O) data. The effect of varying Hadoop Distributed File System (HDFS) block size and, analysis of the size of heap memory and intermediate data generated during map and reduce functions also shows the superiority of the proposed index. It is observed through experimental results that the parallel B-Tree index along with a chained MapReduce environment performs better than the default non-indexed dataset of the Hadoop and B-Tree like Global Index in MapReduce.

The representation, management and application of continuously increasing amounts of heterogeneous data generated by IoTBDA remains an open research challenge specially in context of smart home application. To this end, Tao et al. in their paper "*Ontology-based data semantic management and application in IoT- and cloud-enabled smart homes* (http://dx.doi.org/10.1016/j.future.2016.11.012)", propose a scheme for ontology-based data semantic management and application. Based on the smart home system model abstracted from the perspective of implementing users' household operations, a general domain ontology model is designed by defining the correlative concepts, and a logical data semantic fusion model is designed accordingly. Subsequently, to achieve high-efficiency ontology data queries and updates in the implementation of the data semantic fusion model, a relational-database-based ontology data decomposition storage method is developed by thoroughly investigating existing storage modes, and the performance is demonstrated using a group of elaborated ontology data query and update operations. The work attempts to provide accurate and personalized home services, and the efficiency is demonstrated through experiments conducted on the developed testing system for user behavior reasoning.

Just like other web-based information systems, IoTBDAs must also deal with the plethora of Cyber Security and Privacy threats

that currently disrupt organisations and can potentially hold the data of entire industries and even countries for ransom. To realise its full potential, IoTBDAs must deal effectively with such threats and ensure the security and privacy of the information collected and distilled from IoT devices. However, IoT presents several unique challenges that make the application of existing security and privacy techniques difficult. This is because IoTBDAs encompass a variety of security and privacy solutions for protecting such IoT data on the move and in store at the device layer, the IoT infrastructure/platform layer, and the IoT application layer. Therefore, ensuring end-to-end privacy across these three IoT layers is a grand challenge in IoT. In the paper titled "*Privacy preserving Internet of Things: From privacy techniques to a blueprint architecture and efficient implementation* (http://dx.doi.org/10.1016/j.future.2017.03.001)", Jayaraman et al. tackle the IoT privacy preservation problem. In particular, they propose innovative techniques for privacy preservation of IoT data, introduce a privacy preserving IoT Architecture, and also describe the implementation of an efficient proof of concept system that utilises all these to ensure that IoT data remains private. The proposed privacy preservation techniques utilize multiple IoT cloud data stores to protect the privacy of data collected from IoT. The proposed privacy preserving IoT Architecture and proof of concept implementation are based on extensions of OpenIoT — a widely used open source platform for IoT application development. Experimental evaluations are also provided to validate the efficiency and performance outcomes of the proposed privacy preserving techniques and architecture.

How to obtain personalized quality of services from IoTBDAs and assist users selecting appropriate application instance has become a pressing issue with the explosion of different types of IoTBDA applications (e.g., smart home monitoring to remote healthcare monitoring) on the Internet. Collaborative QoS prediction is recently proposed addressing this issue by borrowing ideas from recommender systems. Going down this principle, Wu et al. in their paper "*Deviation-based neighborhood model for context-aware QoS prediction of cloud and IoT services* (http://dx.doi.org/10.1016/j.future.2016.10.015)", propose novel deviation-based neighborhood models for QoS prediction by taking advantages of crowd intelligence. Different from existing works, their models are under a two-tier formal framework that allows an efficient global optimization of the model parameters. The first component gives a baseline estimate for QoS prediction using deviations of the services and the users. The second component is founded on the principle of neighborhood-based collaborative filtering and contributes fine-grained adjustments of the predictions. Also, contextual information is used in the neighborhood component to strengthen the predicting ability of the proposed models. Experimental results, on a large-scale QoS-specific dataset, demonstrate that deviation-based neighborhood models can overcome existing difficulties of heuristic collaborative filtering methods and achieve superior performance than the state-of-the-art prediction methods. Also, the proposed models can naturally exploit location information to ensure more accurate prediction results.

With the growing popularity of IoTBDAs and sensors deployment, more and more cities are leaning towards smart cities solutions that can leverage this rich source of streaming data to gather knowledge that can be used to solve domain-specific problems. A key challenge that needs to be faced in this respect is the ability to automatically (i.e., context aware) discover and integrate heterogeneous sensor data streams on the fly for applications to use them. To provide a domain-independent platform and take full benefits from semantic technologies, in the paper titled "*Automated discovery and integration of semantic urban data streams: The ACEIS middleware* (http://dx.doi.org/10.1016/j.future.2017.03.002)", Gao et al. present an Automated Complex Event Implementation System (ACEIS), which serves as a middleware between sensor data streams and smart city applications. ACEIS not only automatically discovers and composes IoT streams in urban infrastructures for users' requirements expressed as complex event requests, but also automatically generates stream queries in order to detect the requested complex events, bridging the gap between high-level application users and low-level information sources. The paper also demonstrates the use of ACEIS in a smart travel planner scenario using real-world sensor devices and datasets.

Besides the abundant potential for the IOTBDAs, there are also challenges to security due to complexity and unpredictability of the Internet, clouds, and big data. One of the challenges is information and data exchange, for example, identifying untrustworthy cloud users and analyzing abnormal user behavior during information exchange. Hence in the paper titled " *A game-theoretic model and analysis of data exchange protocols for Internet of Things in clouds* (http://dx.doi.org/10.1016/j.future.2016.12.030)", Tao et al. address exchange mechanism, which is a useful theoretic basis to make secure electronic commerce and electronic business transactions possible. To ensure and verify the property of fairness, a crucial property of exchange mechanism, this paper proposes a specific model for behavior analysis based on the extensive game with imperfect information. Rationality and fairness properties are built in the corresponding game and the game tree. To verify the properties, a tree analysis method is proposed, and a linear time algorithm is given. As a case study, some flaws of the ASW protocol are found.

We hope that the readers will find the articles of this special issue to be informative and useful.

References

[1] Lizhe Wang, Rajiv Ranjan, Processing distributed internet of things data in clouds, IEEE Cloud Comput. 2 (1) (2015) 76–80. http://dx.doi.org/10.1109/MCC.2015.14.

[2] Yan Ma, Lizhe Wang, Albert Y. Zomaya, Dan Chen, Rajiv Ranjan, Task-tree based large-scale mosaicking for massive remote sensed imageries with dynamic dag scheduling, IEEE Trans. Parallel and Distrib. Syst. 25 (8) (2014) 2126–2137. http://dx.doi.org/10.1109/TPDS.2013.272.

[3] Lizhe Wang, Ke Lu, Peng Liu, R. Ranjan, Lejiao Chen, IK-SVD: dictionary learning for spatial big data via incremental atom update, Comput. Sci. Engrg. 16 (4) (2014) 41–52. http://dx.doi.org/10.1109/MCSE.2014.52.

[4] George Okeyo, Liming Chen, Hui Wang, Combining ontological and temporal formalisms for composite activity modelling and recognition in smart homes, Future Gener. Comput. Syst. (ISSN: 0167-739X) 39 (2014) 29–43. http://dx.doi.org/10.1016/j.future.2014.02.014.

[5] Rodrigo Bonacin, Olga Fernanda Nabuco, Ivo Pierozzi Junior, Ontology models of the impacts of agriculture and climate changes on water resources: scenarios on interoperability and information recovery, Future Gener. Comput. Syst. (ISSN: 0167-739X) 54 (2016) 423–434. http://dx.doi.org/10.1016/j.future.2015.04.010.

[6] Silvio D. Cardoso, Flor K. Amanqui, Kleberson J.A. Serique, José L.C. dos Santos, Dilvan A. Moreira, Swi: a semantic web interactive gazetteer to support linked open data, Future Gener. Comput. Syst. (ISSN: 0167-739X) 54 (2016) 389–398. http://dx.doi.org/10.1016/j.future.2015.05.006.

[7] Idafen Santana-Perez, Rafael Ferreira da Silva, Mats Rynge, Ewa Deelman, María S. Pérez-Hernández, Oscar Corcho, Reproducibility of execution environments in computational science using semantics and clouds, Future Gener. Comput. Syst. (ISSN: 0167-739X) 67 (2017) 354–367. http://dx.doi.org/10.1016/j.future.2015.12.017.

Rajiv Ranjan is a Reader in the School of Computing Science at Newcastle University, UK; chair professor in the School of Computer, Chinese University of Geoscience, Wuhan, China; and a visiting scientist at Data61, CSIRO, Australia. His research interests include grid computing, peer-to-peer networks, cloud computing, Internet of Things, and big data analytics. He has published about 200 research papers (including 120+ journal papers). His papers have received 7770+ citations in total, he has an h-index and i10-index of 36 and 74 respectively. His papers have also received 1700+ citations and h-index of 16; according to Thomson Reuters Journal Citation Report (goo.gl/mJVphW). He also has an Scopus (Author ID:22980683700) h-index of 19 and total citations >2900. Ranjan has a Ph.D. in computer science and software engineering from the University of Melbourne (2009). Contact him at raj.ranjan@ncl.ac.uk or http://rajivranjan.net.

Dhaval Thakker is a Lecturer in the School of Electrical Engineering and Computer Science, University of Bradford. He has over ten years of experience in the European Union(EU) and industrial projects delivering innovative solutions. Prior to joining Bradford, he worked as a Research Fellow at the University of Leeds from 2011 to 2015 and was leading semantic web related research in several EU projects like the EU FP7 NeTTUN, ImREAL and DICODE. Before Leeds, he worked in the industry with UK's national news agency (Press Association) as a Research & Development Consultant to provide strategic and technical leadership in implementing Semantic Web and Linked data related projects to improve access to their media repositories. Dhaval has published 50+ papers on various journals, conferences and other international forums on the topics of Semantic Web, IoT, Web services and data exploration.

Armin Haller is a Senior Lecturer at Australian National University with a dual appointment at the Research School of Computer Science and the Research School of Management. Previously, he was a Research Scientist (2010–2015) in the Digital Productivity Flagship of the CSIRO. Prior to that he was a Research Associate at DERI, Ireland, where he worked in several EU-funded research projects, such as SUPER, KnowledgeWeb and DIP. He is currently co-chairing the Semantic Sensor Network Ontology working group in the W3C. His research interests are in Ontology Engineering, Linked Data, the Internet-of-Things and Web services, where he has published over 50 research papers.

Rajkumar Buyya is a Professor of Computer Science and Software Engineering; and Director of the Cloud Computing and Distributed Systems (CLOUDS) Laboratory at the University of Melbourne, Australia. He is also serving as the founding CEO of Manjrasoft Pty Ltd., a spin-off company of the University, commercializing its innovations in Grid and Cloud Computing. He has authored over 525 publications and seven text books including "Mastering Cloud Computing" published by McGraw Hill, China Machine Press, and Morgan Kaufmann for Indian, Chinese and international markets respectively. He is one of the highest cited authors in computer science and software engineering worldwide. Recently, Dr. Buyya is recognized as "2016 Web of Science Highly Cited Researcher" by Thomson Reuters. For further information on Dr. Buyya, please visit his cyberhome: http://www.buyya.com.