Hindawi
Complexity
Volume 2019, Article ID 4184708, 3 pages
https://doi.org/10.1155/2019/4184708

Editorial

Advances in Architectures, Big Data, and Machine Learning Techniques for Complex Internet of Things Systems

David Gil,[1] **Magnus Johnsson**,[2,3,4] **Higinio Mora**,[1] **and Julian Szymanski**[5]

[1] *University of Alicante, Alicante, Spain*
[2] *Malmö University, Malmö, Sweden*
[3] *Department of Intelligent Cybernetic Systems, NRNU MEPhI, Moscow, Russia*
[4] *AI Research AB, Höör, Sweden*
[5] *Gdansk University of Technology, Gdansk, Poland*

Correspondence should be addressed to David Gil; david.gil@ua.es

Received 27 February 2019; Accepted 27 February 2019; Published 24 March 2019

The field of Big Data is rapidly developing with a lot of ongoing research, which will likely continue to expand in the future. A crucial part of this is Knowledge Discovery from Data (KDD), also known as the Knowledge Discovery Process (KDP). This process is a very complex procedure, and for that reason it is essential to divide it into several steps (Figure 1). Some authors use five steps to describe this procedure, whereas others use only four. We use the following four-step description:

(1) Generation of Data: Data is generated from a multitude of various data sources, such as sensors, social media, the web, a multitude of devices, software applications, people, and various kinds of sensors [1].

(2) Collection of Data: This step involves the storage of data into various types of databases, such as MongoDB, elastic, InfluxDB, MySQL, and NoSQL suitable for Big Data technologies [2–7]. Often the raw sensor data are collected, but it is common to also link them to contextual information [8–10]. Other tasks such as cleaning, integration, and transformation of data are essential for the optimal storage in databases [11–13]. Several tools and technologies suitable to semi-automatize tasks such as Extract, Transform, and Load (ETL) [14] exist, e.g., Apache NIFI [15, 16] and Pentaho [17, 18]. These are very useful since there are many tasks involved in this step of the KDD procedure.

(3) Machine Learning and Data Mining: Diverse machine learning and data mining methods are applied and benchmarked, and the results are compared. It should be kept in mind that though there are many machine learning methods, not all of them are suitable to use with Big Data [19, 20].

(4) Classification, Prediction, and Visualization: This step focuses on, in particular, the obtaining of visualizations that present all the classification and prediction results in a useful way. Tools such as Grafana [21] could help to interpret the data visually, but also to simplify the identification of Key Performance Indicators (KPI) [22].

This special issue received in total 19 submitted papers, and after a meticulous reviewing process the editors decided to accept eight of these for publication, which implies an acceptance rate of about 42%.

Anomaly analysis is a crucial issue since it is a significant part of many areas, such as medical health, credit card fraud, and intrusion detection (X. Xu et al.). The authors of this paper provide a complete state-of-the-art presentation of anomaly detection. High dimensionalities and mixed types of data are the focus of this study as the identification of anomalous patterns is far from trivial. The authors introduce the reader to current advances on anomaly detection, while debating the pros and cons of various detection methods.

There are areas that are well-known, though barely referenced in the literature. For example, the one presented by M. Lodeiro-Santiago et al., where the goal is to detect small boats (pateras) to help address the problem of dangerous immigration. In this paper, the authors use deep

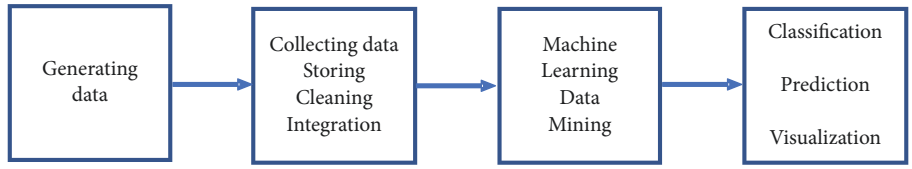

Figure 1: Procedure for the KDD.

convolutional neural networks to improve detection methods based on image processing through the application of filters. Their novel approach is able to recognise the boats through patterns regardless of where they are located. The proposed approach, which works in real-time, allows the detection of boats and people for search and rescue teams in order to plan for rescue operations before an emergency happens. The proposed method includes the use of essential cryptographic protocols for the protection of the highly sensitive information managed.

A method for the evaluation of heart rate streams in patients with ischemic heart disease is presented by M. D. Peláez-Aguilera et al. The authors present an innovative linguistic approach to manage relevant linguistic descriptions (protoforms). This provides a foundation for the cardiac rehabilitation team to identify sessions with significance indicators through linguistic summaries. As it is faced in the manuscript, cardiac rehabilitation programs are crucial to significantly decrease mortality rates in high-risk patients with ischemic heart disease.

In the work presented by Z. Marszałek et al., a fully flexible sorting method designed for parallel processing is presented. The authors describe a method based on modified merge sort designed for multicore architectures. The flexibility of the method, which is implemented for a number of processors, increases the efficiency of sorting by distributing the tasks between logical cores in a flexible way. Since powerful computer resources are often not very well exploited, their main goal is to use efficient algorithms to support the proficient use of all available resources.

F. M. Pérez et al. present a theoretical framework based on a generalisation of rough sets theory. This allows the establishment of a stochastic approach to solving the problem of outliers within a specific universe of data. An algorithm based on this theoretical framework is developed to make it suitable for large data volume applications. The experiments carried out validate the proposed algorithm in comparison to various algorithms analysed in the literature.

The work proposed by P. S. Szczepaniak and A. Duraj concerns the problem of outlier detection through the application of case-based reasoning. The authors argue that while this method has been successfully applied in an extensive variety of other domains, it has never been used for outlier detection.

In the manuscript presented by Q. Gu et al., the authors propose a Hybrid Genetic Grey Wolf Algorithm in order to improve the disadvantage of Grey Wolf Optimizer when solving Large-Scale Global Optimization problems.

Finally, D. Gil et al. provide a review highlighting the fact that the complexity of managing Big Data is one of the main challenges in the developing field of the Internet of Things (IoT). The review divides the discovery of knowledge into the four general steps sketched above and evaluates the most novel technologies involved. These include IoT data gathering, data cleaning and integration, data mining and machine learning, and classification, prediction, and visualization.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors acknowledge the support of the Internet of Things and People (IOTAP) Research Center at Malmö University in Sweden. This work was also supported by the Spanish Research Agency (AEI) and the European Regional Development Fund (ERDF) under the project CloudDriver4Industry TIN2017-89266-R. This work has also been funded by the Spanish Ministry of Economy and Competitiveness (MINECO/FEDER) under the granted Project SEQUOIA-UA (management requirements and methodology for Big Data analytics) TIN2015-63502-C3-3-R.

David Gil
Magnus Johnsson
Higinio Mora
Julian Szymanski

References

[1] C. A. Zaslavsky and D. G. Perera, "Sensing as a service and big data," https://arxiv.org/abs/1301.0159, 2013.

[2] J. Han, E. Haihong, G. Le, and J. Du, "Survey on NoSQL database," in *Proceedings of the 6th International Conference on Pervasive Computing and Applications (ICPCA '11)*, pp. 363–366, Port Elizabeth, South Africa, October 2011.

[3] R. Cattell, "Scalable SQL and NoSQL data stores," *ACM SIGMOD Record*, vol. 39, no. 4, pp. 12–27, 2010.

[4] Ishwarappa and J. Anuradha, "A brief introduction on big data 5Vs characteristics and hadoop technology," *Procedia Computer Science*, vol. 48, no. C, pp. 319–324, 2015.

[5] S. Patni, *Pro RESTful APIs: Design, Build and Integrate with REST, JSON, XML and JAX-RS*, Apress, Berkeley, CA, USA, 2017.

[6] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

[7] C. L. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: a survey on big data," *Information Sciences*, vol. 275, pp. 314–347, 2014.

[8] S. Satpathy, B. Sahoo, and A. K. Turuk, "Sensing and actuation as a service delivery model in cloud edge centric internet of things," *Future Generation Computer Systems*, vol. 86, pp. 281–296, 2018.

[9] J.-P. Calbimonte, H. Jeung, O. Corcho, and K. Aberer, "Semantic sensor data search in a large scale federated sensor network," in *Proceedings of the 4th International Workshop on Semantic Sensor Networks 2011, SSN 2011 - A 10th International Semantic Web Conference, ISWC 2011*, pp. 23–38, Germany, October 2011.

[10] S. Li, L. D. Xu, and X. Wang, "Compressed sensing signal and data acquisition in wireless sensor networks and internet of things," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 4, pp. 2177–2186, 2013.

[11] Z. Doan, A. Halevy, and A. Ives, *Principles of Data Integration*, Elsevier, 2012.

[12] H. Gonzalez, A. Halevy, C. S. Jensen et al., "Google fusion tables: web-centered data management and collaboration," in *Proceedings of the the 1st ACM symposium*, p. 175, June 2010.

[13] A. Y. Halevy, "Answering queries using views: a survey," *The VLDB Journal*, vol. 10, no. 4, pp. 270–294, 2001.

[14] P. Vassiliadis, "A survey of extract-transform-load technology," *International Journal of Data Warehousing and Mining*, vol. 5, no. 3, pp. 1–27, 2009.

[15] Apache NiFi.

[16] J. N. Hughes, M. D. Zimmerman, C. N. Eichelberger, and A. D. Fox, "A survey of techniques and open-source tools for processing streams of spatio-temporal events," in *Proceedings of the 7th ACM SIGSPATIAL International Workshop on GeoStreaming, IWGS 2016*, USA.

[17] R. Bouman and J. Van Dongen, *Pentaho Solutions: Business Intelligence and Data Warehousing with Pentaho and MySQL*, 2009.

[18] M. Casters, R. Bouman, and J. Van Dongen, *Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration*, John Wiley & Sons, 2010.

[19] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014.

[20] F. Chen, P. Deng, J. Wan, D. Zhang, A. V. Vasilakos, and X. Rong, "Data mining for the internet of things: literature review and challenges," *International Journal of Distributed Sensor Networks*, vol. 2015, no. i, 2015.

[21] Grafana, "The open platform for analytics and monitoring".

[22] E. Betke and J. Kunkel, "Real-time I/O-monitoring of HPC applications with SIOX, elasticsearch, Grafana and FUSE," in *Proceedings of the ISC High Performance 2017: High Performance Computing*, pp. 174–186, Springer, Cham, Switzerland, 2017.