Process Mining in Big Data Scenario

Antonia Azzini, Ernesto Damiani

SESAR Lab - Dipartimento di Informatica
Università degli Studi di Milano, Italy
antonia.azzini,ernesto.damiani@unimi.it

Abstract. In the last years the management and analysis of big data generated from information systems are becoming one of the most important topics in the Business Process Intelligence (BPI). In this field researchers show how Process Mining could become very helpful in bridging the gap between data and processes. The aim of this work is to present and discuss a brief review of the literature reporting most of the Process Mining chances that meet Big Data and the challenges carried out, showing the critical aspects and the advantages of different solutions.

Keywords: process mining, big data, business process management

1 Introduction

In the last years the management and analysis of big data generated from information systems are becoming one of the most important topics in the Business Process Intelligence (BPI). In fact there is a need for data scientists to transform event data into actionable information, but, a comprehensive data analysis is required.

As reported by [7], Big data has become a board-level topic and organizations are investing heavily in related technologies, even if it is not always clear how to derive value from data.

Several companies have unused process data that can be used for Process Mining. This is a side-effect of the ongoing automation of business processes, leaving digital traces of real process executions as a byproduct. According to the process science idea, all these digital traces reflect what happens in the real world and enable the application of process mining: indeed, Business Processes can be made visible to understand how these processes are actually executed, by giving a transparency that helps organizations to re-gain control over their complex business environments. They automatically creates this transparency from existing data logs, and the analysis can be easily repeated with little effort to adapt to these changes or to validate the effects of improvement initiatives.

To enforce such an aspect, Some authors [7] argue that one needs to carefully combine process-centric and data-centric approaches. This seems obvious, yet most data science (process science) approaches are process (data) agnostic. Process Mining techniques aim to bridge this gap [18, 3, 20].

This work reports a brief review of the literature showing most of the Process Mining solutions that meet Big Data.

2 Challenges

This section presents some of the most critical issues and important aspects related to the use of Process Mining and Big Data, presented in the literature in the last few years [18].

2.1 Data Streams

In the context of Process Mining an important aspect regards the *Process Discovery*: it is concerned with deriving process-related information from event logs and, thus, enabling the business analyst to extract and understand the actual behaviour of business process. Even though they are now increasingly used in commercial settings, many of the developed process discovery algorithms are designed to work in a static fashion, but not easily applicable for processing real time event streams.

In such a scenario an important aspect regards the definition of the most suitable Process Mining methodology to handle online data [19, 13], and to collect and save them into trusted devices [16].

2.2 Process-aware Data Mining

As already reported in the literature [7], most of the open issues in the context of *Map Reduce* regard the Map function. In fact, Map Reduce implementations of Process Mining algorithms should compute keys based on data types that are available within process log entries, such as task, location and operator ID.

Another open issue regards load balancing. The overall performance of *Map Reduce* depends on the data balancing., that can sometimes be achieved by writing intelligent Map functions. The critical aspect regards the fact that estimating (e.g., from process model analysis) or learning (e.g., by preliminary sampling event streams) the cardinality distribution of keys is by no means straightforward.

2.3 Process Coordination and Data Integration

The emergence of Big Data results in a new set of challenges for *Process Discovery* on event streams, like diversity of event formats from different sources [15], stream dimensions (thousands of events per second), and less rigid processes (frequent changes have to be applied to business processes found at the operational level). The difficulties lie, for example, in data capture, storage, searching, sharing, analysis, and visualization [11].Challenges in *Big Data* analysis include data inconsistency and incompleteness, scalability, timeliness, and security [12]. Also integrity checking is a critical aspect, due to the lack of support of remote data access and to the lack of information regarding internal storage, as well as the semantic aspects, as reported in [14]. Additional critical issues also arise in the semantic when data congruency semantic is considered [5]. A typical problem of the integration of distributed data sources regards the *Semantic Lifting* procedures. In a distributed context one of the most critical aspects related to the *Semantic Lifting* regards the fact that, if locally applied, it's not possible to check the real data value.

Moreover, further deal of controversy reported in the literature [9] other issues like the *Quality of Data* (bigger data are not always better data), that concern about accessibility, *Data Integration*, and the digital divides deriving from the limited access rights to the data, between people or organizations with or without the access rights.

3 Early Integrated Studies and Discussions

Examples to summarize the Process Mining structure and to show the features are reported in [1], together with a discussion about the main Process Mining use cases. Scalable Dynamic Process Discovery (SDPD) is a process that describes the monitoring of one or more Business Process Management Systems (BPMSs) in order to provide at any time an accurate representation of the current state of the processes deployed in the systems with respect to the control flow, resource, performance perspectives and the state of still open traces.

Other works regard the Map Reduce and present a discussion over the usage of domain taxonomies, in order to improve the efficiency of Map Reduce [8, 10], while [17] presents an approach to enhance trace clustering performances. Another application of PM is the so called comparative process mining, which uses process cubes [2] to compare different sub-processes. The main advantage of this approach is that it can be easily used to identify differences between two branches of the same organization or two groups of customers [4].

Moreover, the actual increase of the interests in manufacturing process management and analysis in manufacturing leads some authors to present in [21] a manufacturing data analysis system that collects event logs based both on structured and unstructured data, and analyzes them with process mining for enhancing the process analysis results. An interesting approach presents a methodology designed for consistent process mining algorithms in a Big Data context [5]. The authors underline how a common conceptual model is typically defined to address name resolution. Such an aspect implies that each local source is tasked for applying a semantic lifting procedure in order to express the local data in term of the common model, by potentially introducing semantic heterogeneity in data. As reported by [6], the aim is to investigate solutions for handling semantic lifting without compromising the requirements imposed in a Big Data context.

All the different approaches above reported are then categorized according to their common topics and summarized in Table 1.

Further research should address confidentiality, data encryption, computation power reduction, and application of different encryption algorithms to heterogeneous data. Recently, some controversies have revealed how some security agencies are using data generated by individuals for their own benefits without permission. Therefore, access rights violations should be identified and user data should not be misused or leaked.

Table 1. Summary of the Presented Early Studies

Work	Topics	Description
[1] [7]	Process Mining, Process Discovery Business Process Management	Process Mining: a link between processes and data, and between performance and compliance.
[13] [16] [19]	Process Discovery, Scalability, Dynamic Business Process Discovery, Construct Competition Miner	Run time monitoring of BPMSs to provide current status of control flow, resource, performance perspectives, open traces.
[7] [8] [10] [17]	Process Science, Map Reduce, Trace Clustering	Application of Process Science algorithms based on Map Reduce, Trace Clustering, Intel. Map Functions, to optimize performances in big data analysis, and the handling with semantic-rich context information.
[14] [5] [9] [6] [21] [11]	Semantic Lifting, Quality of Data Data Integration, Big Data	Big Data analysis and development of process mining methodologies based on semantic issues and data integration.

4 Conclusion

The management and analysis of Big Data generated from information systems are becoming one of the most important topics in the Business Process Intelligence. The literature underlines the importance of the connection between data science (that combines techniques from statistic data mining, machine learning and high performance computing) and process science, that considers process-centric approaches of Business Process Management.

This work reports some of the recent approaches of the literature reporting Process Mining solutions that meet the arising actual Big Data challenges and critical issues.

References

1. Van der Aalst, W.: Mine your own business: Using process mining to turn big data into real value (2013)
2. Van der Aalst, W.: Process cubes: Slicing, dicing, rolling up and drilling down event data for process mining. In: Asia Pacific Business Process Management, Lecture Notes in Business Information Processing, vol. 159, pp. 1–22. Springer Int. Publishing (2013)
3. Van der Aalst, W.: Process mining as the superglue between data science and enterprise computing (2014)
4. Van der Aalst, W., Zhao, J.L., Wang, H.: Editorial: Business process intelligence: Connecting data and processes. ACM Trans. on Management Information Systems 5(4), 1–7 (2015)
5. Azzini, A., Ceravolo, P.: Consistent process mining over big data triple stores (2013)
6. Ceravolo, P., Zavatarelli, F.: Knowledge acquisition in process intelligence (2015)
7. Damiani, E., Van der Aalst, W.: Processes meet big data: Connecting data science with process science. IEEE Trans. on Services Computing 1(1), 1–10 (2015)
8. Evermann, J., Assadipour, G.: Big data meets process mining: Implementing the alpha algorithm with map-reduce (2014)

9. Fan, W., Bifet, A.: Mining big data: Current status, and forecast to the future. SIGKDD Explorartion Newsletter 14(2), 1–5 (2013)
10. Jahani, E., Cafarella, M.J., Ré, C.: Automatic optimization for mapreduce programs. Proc. of the VLDB Endoment. 4(6), 385–396 (2011)
11. Khan, N., Yaqoob, I., Abake, I., Hashem, T., et al.: Big data: Survey, technologies, opportunities, and challenges. The Scientific World Journal pp. 1–18 (2014)
12. Labrinidis, A., Jagadish, H.: Challenges and opportunities with big data. Proc. of the VLDB Endowment 5(12), 2032–2033 (2012)
13. Leemans, S., Fahland, D., van der Aalst, W.: Scalable process discovery with guarantees. In: Enterprise, Business-Process and Information Systems Modeling, Lecture Notes in Business Information Processing, vol. 214, pp. 85–101. Springer Int. Publishing (2015)
14. Mendling, J., Leopold, H., Pittke, F.: 25 challenges of semantic process modeling. Int. Journal of Information Systems and Software Engineering for Big Companies 1(1), 78–94 (2014)
15. Redlich, D., Gilani, W., Molka, T., Drobek, M., Rashid, A., Blair, G.: Introducing a framework for scalable dynamic process discovery. In: Advances in Enterprise Engineering VIII, Lecture Notes in Business Information Processing, vol. 174, pp. 151–166. Springer Int. Publishing (2014)
16. Redlich, D., Molka, T., Gilani, W., Blair, G.S., Rashid, A.: Scalable dynamic business process discovery with the constructs competition miner (2014)
17. Song, M., Yang, H., Siadat, S.H., Pechenizkiy, M.: A comparative study of dimensionality reduction techniques to enhance trace clustering performances. Expert Systems with Applications 40(9), 3722–3737 (2013)
18. Van Der Aalst, W.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Springer (2011)
19. Van Der Aalst, W., Van Hee, K.: Workflow management: models, methods, and systems. MIT press (2004)
20. Van Der Aalst, W., Adriansyah, A., de Medeiros, A.K.A., Arcieri, F., Baier, T., Blickle, T., Bose, J.C., van den Brand, P., Brandtjen, R., Buijs, J., et al.: Process mining manifesto. In: Business process management workshops. pp. 169–194. Springer (2012)
21. Yang, H., Park, M., Cho, M., Song, M., Kim, S.: A system architecture for manufacturing process analysis based on big data and process mining techniques (2014)