Parallel Distributed Patterns Mining Using Hadoop MapReduce Framework

Ishak H. A. Meddah, University of Sciences and Technology of Oran “Mohamed Boudiaf” (USTO-MB), Oran, Algeria
Khaled Belkadi, University of Sciences and Technology of Oran “Mohamed Boudiaf” (USTO-MB), Oran, Algeria

ABSTRACT

The treatment of large data is proving more difficult in different axes, but the arrival of the framework MapReduce is a solution of this problem. With it we can analyze and process vast amounts of data. It does this by distributing the computational work across a cluster of virtual servers running in a cloud or large set of machines while process mining provides an important bridge between data mining and business process analysis. The process mining techniques allow for extracting information from event logs. In general, there are two steps in process mining: correlation definition or discovery and process inference or composition. Firstly, the authors’ work consists to mine small patterns from a log traces. Those patterns are the representation of the traces execution from a log file of a business process. In this step, they use existing techniques. The patterns are represented by finite state automaton or their regular expression. The final model is the combination of only two types of small patterns whom are represented by the regular expressions (ab)* and (ab*c)*. Secondly, the authors compute these patterns in parallel, and then combine those small patterns using the MapReduce framework. They have two parties: the first is the Map Step in which they mine patterns from execution traces; the second is the combination of these small patterns as reduce step. The authors’ results are promising in that they show that their approach is scalable, general, and precise. It minimizes the execution time by the use of the MapReduce framework.

KEYWORDS

Business Process, MapReduce, Patterns, Process Mining, Traces

1. INTRODUCTION

Many techniques have been proposed that mine such patterns from execution traces. However; most existing techniques mine only simple patterns, or they mine a single complex pattern that is restricted to a particular set of manually selected events. Recent work has recognized that patterns can be specified as regular languages (Ammons et al., 2002). This allows the compact representation of patterns as regular expressions or finite state automata, and it allows the characterization of the pattern mining as a language learning problem; current approaches are fundamentally similar; each takes as input a static program or a dynamic traces or profile and produces one or more compact regular languages that specify the pattern representation or the workflow. However; the individual solutions differ in key ways.

In this paper, we present a new general approach to patterns mining that addresses several of the limitations of current techniques. Our insight is twofold. First, we recognize that instances of smaller patterns can be composed in parallel into larger patterns. Second, we observed also that the composition of small pattern can be in parallel; Our work is an amelioration to the existing techniques.

DOI: 10.4018/IJGHPC.2017040105
in mining patterns and parallel distributed process. We then leverage this insight to divide our work into two parties; The first one, we use a technique how we can mine two types of small patterns and we compose them by using standard algorithms for finite state automaton manipulation, and some special rules using by M. Gabel and Z. Su (Gabel & Su, 2008), the mining is also performed by symbolic mining algorithm (Gabel & Su, 2008; Zhang et al., 2015). The second one, we use the framework MapReduce in mining and composing micro patterns; those patterns have been shown as regular expressions or their finite state automatons, in this party we mine small patterns using the same symbolic mining algorithm but in parallel as Map step, and we compute this small patterns into larger pattern in parallel as reduce step.

Our approach has been implemented in the java programming language with two log files of two application; the SKYPE and VIBER applications. The size of the first log file is 10 GB, and the second is 18 GB, who are generated by log file generator; we have tested our approach in three clusters in a cloud, the first regroup five machines, and the second regroups ten machines, and the third regroups 20 machines the traces in our applications are the call, the answer, and the messages, etc.

2. RELATED WORK

Many techniques are suggested in the domain of process mining, we quote:

M. Gabel et al. (Gabel & Su, 2008) present a new general technique for mining temporal specification, they realized their work in two steps, firstly they discovered the simple patterns using existing techniques, then combine these patterns using the composition and some rules like Branching and Sequencing rules.

Temporal specification expresses formal correctness requirement of an application’s ordering of specific actions and events during execution, they discovered patterns from traces of execution or program source code; The simples patterns are represented using regular expression (ab)\(^*\) or (ab*c)\(^*\) and their representation using finite state automaton, after they combine simple patterns to construct a temporal specification using a finite state automaton.

G. Greco et al. (Greco et al., 2006) discovered several clusters by using a clustering technique, and then they calculate the pattern from each cluster, they combine these patterns to construct a final model, they discovered a workflow scheme from, and then they mine a workflow using a Mine Workflow Algorithm, after they define many clusters from a log traces by using clustering technique and Process Discover Algorithm and some rules cluster.

Then they use a Find Features algorithm to find a pattern of each cluster, finally they combine these patterns to construct a completely hierarchical workflow model.

In their clustering algorithm, clusters reflect only structural similarities among traces; they say that in future works extending their techniques to take care of the environment so that clusters may reflect not only structural similarities among traces, but also information about, e.g., users and data values.

H.R. Motahari-Nezhed et al. (Motahari-Nezhad et al., 2008) use a service conversation log; first they split a log into several partitions, 2\(^{rd}\) they discovered a model from each partition, and 3\(^{rd}\), they annotate the discover protocol model with various metadata to construct a protocol model from real-word service conversation logs.

The protocol is the specification of all possible conversations that a service can have with its partners and the conversation consists of a sequence of messages exchanged between two or more services; During the split they discovered a simple precise protocol models by analyzing messages sequences in the log, they eliminate conversations considered noisy or not presented in the log; they
Exploiting P2P and Grid Computing Technologies for Resource Sharing to Support High Performance Distributed System
[www.igi-global.com/chapter/exploiting-p2p-grid-computing-technologies/40813?camid=4v1a](www.igi-global.com/chapter/exploiting-p2p-grid-computing-technologies/40813?camid=4v1a)

Adjusting Thread Parallelism Dynamically to Accelerate Dynamic Programming with Irregular Workload Distribution on GPGPUs
[www.igi-global.com/article/adjusting-thread-parallelism-dynamically-to-accelerate-dynamic-programming-with-irregular-workload-distribution-on-gpgpus/114710?camid=4v1a](www.igi-global.com/article/adjusting-thread-parallelism-dynamically-to-accelerate-dynamic-programming-with-irregular-workload-distribution-on-gpgpus/114710?camid=4v1a)