J. Parallel Distrib. Comput. ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

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

A parallel metaheuristic data clustering framework for cloud

Chun-Wei Tsai *, Shi-Jui Liu, Yi-Chung Wang

Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, ROC
Taiwan Information Security Center at NCHU (TWISC@NCHU), Taichung, Taiwan, ROC

ARTICLE INFO

Article history:
Received 31 July 2017
Received in revised form 27 October 2017
Accepted 30 October 2017
Available online xxxx

Keywords:
Metaheuristic algorithm
Internet of things
Data clustering problem

ABSTRACT

A high performance data analytics for internet of things (IoT) has been a promising research subject in recent years because traditional data mining algorithms may not be applicable to big data of IoT. One of the main reasons is that the data that need to be analyzed may exceed the storage size of a single machine. The computation cost of data analysis tasks that is too high for a single computer system is another critical problem we have to confront when analyzing data from an IoT system. That is why an efficient data clustering framework for metaheuristic algorithm on a cloud computing environment is presented in this paper for data analytics, which explains how to divide mining tasks of a mining algorithm into different nodes (i.e., the Map process) and then aggregate the mining results from these nodes (i.e., Reduce process). We further attempted to use the proposed framework to implement data clustering algorithms (e.g., *k*-means, genetic *k*-means, and particle swarm optimization) on a standalone system and Spark. The experimental results show that the performance of the proposed framework makes it useful to develop data clustering algorithms on a cloud computing environment.

1. Introduction

Since internet of things is aimed to connect everything (e.g., sensors, information appliances, or devices) to the Internet to provide a more accurate and convenient service to people, the possibilities of IoT have been discovered and discussed for years [3,18,48,78,6,69,8,49]. According to the report [22], the market of internet of things and machine to machine technologies was forecasted to be up to USD 883.55 billion in 2022. From the study [1] that depicted the concept of IoT up to now, the applications of smart home [10,44], smart city [43], smart grid [17], or even business intelligence all further explain that the number and diversity of IoT applications will gradually increase at different succession stages today. Of course, we all know that the possibilities of internet of things are not limited to these applications; on the contrary, it will be part of our daily life, and it will provide more better applications that simply cannot be realized before. For example, forecasting the traffic collision and forecasting the traffic jam are two difficult tasks for intelligent transport systems, both of them, however, can be mitigated once all the cars are connected to the Internet and the statuses of them or around them are reported via the technologies of IoT and vehicular ad hoc networks (VANETs), called the internet of vehicles (IoV) [77].

From the perspective of history, the Internet has undergone several transformations from the so-called "Internet of Data" to the "Internet of Content" and then to the "Internet of People" [55]. But today, the term "Internet of Data" (IoD) in some recent studies [12] can be regarded as the extension of internet of things, which contains much more information than just data collection by objects,[1] such as virtual tags for the collected data from RFID tags. No matter what the definition of IoD is, there is no doubt that the data contains a lot of information and knowledge, especially data from IoT, which apparently may support us to develop a better service. That is why how to find out useful information from these data has become an important research issue [64]. But the amount of data of internet of things is typically very large; thus, they cannot be stored in a single machine. Such data also are collected from different kinds of devices and are created very quickly. The data deluge caused by IoT can be regarded as a kind of big data. That is why several recent studies [25,20,75,37] took into account IoT, big data, and cloud computing together. In these researches, internet of things is considered as a large integrated system for which cloud computing is used to provide the computing power and storage space it needs and big data is used to provide a new way to find out useful information from such a large scale of data.

Since a possible solution is to use a cloud computing platform to support data mining algorithms to find out hidden information

* Corresponding author at: Department of Computer Science and Engineering, National Chung Hsing University, Taichung, Taiwan, ROC.
E-mail address: cwtsai@nchu.edu.tw (C.W. Tsai).

[1] In this paper, the things or objects represent sensors, devices, or appliances belonging to the edge layer, and they play the role of monitoring and collecting the data by internet of things system.

https://doi.org/10.1016/j.jpdc.2017.10.020
0743-7315/

from the data of IoT, how to use free machine learning libraries or how to develop parallelized mining algorithms for a cloud computing platform has become a critical research topic. For the free machine learning libraries, Mahout [38] for Hadoop or MLlib [58,41] for Spark are two representative solutions. It does not matter either free machine learning libraries (e.g., Mahout or MLlib) are used or parallelized mining algorithms with the MapReduce model are developed, both of them have pros and cons. Although free machine learning libraries can be used to achieve data mining task, its development has some restrictions. Even though we can decide how to develop parallelized mining algorithms, it typically has a high barrier, especially redesigning the data mining or machine learning algorithms for the MapReduce model.

Different from using free machine learning libraries to develop data mining tools on a cloud computing environment, another research trend is to develop parallelized mining algorithms to be run on a cloud platform. The same data mining algorithms can be applied to a cloud platform in many different ways [47,80,15,35]. It is confusing for researchers new to this research domain which way is better. A general and high level description will be useful to reduce such heavy burdens. A simple data clustering framework for cloud computing platforms is presented in this study via the MapReduce model. The main contributions of this paper can be summarized as follows:

1. This paper provides a brief review for the data mining algorithms (i.e., clustering, classification, and association rule) from the Hadoop to the Spark platform.
2. This study presents a unified framework to ease the application of population-based metaheuristic-based data clustering algorithms to a cloud computing platform (i.e., Spark) for analyzing the data from the edge layer of an internet of things system.
3. In addition to the descriptions on the design details of the proposed framework, we implement several clustering algorithms using this framework for the Spark environment to show its possibilities.

The remainder of the paper is organized as follows. Section 2 gives a brief review for the data of an internet of things system, followed by a discussion on data mining algorithms for cloud computing platforms. Section 3 first gives the basic idea of the proposed framework and then provides its details. A simple example is then given to show how to apply metaheuristic algorithm for data mining problem to a cloud computing platform. Section 4 begins with a description of the simulation environment. The comparison between centralized and distributed data mining algorithms is given to show the possibility and performance of the proposed framework. Section 5 gives the conclusion and future works of this research.

2. Related work

2.1. Data of IoT

Since the development of internet of things from the study of Atzori [2] from 2009 to 2018 is about a decade, several early studies [4,3,42,18] in the first decade presented mature products to support us constructing an IoT system that contains sensors of appliances for the edge layer, cloud computing platforms and big data analytics systems for the middle layer, and dashboard or visualization tools for the application layer. As shown in Fig. 1, Bandyopadhyay and Sen [5] presented a generic five-layer architecture to explain what kind of components an IoT system has to contain from the edge layer (i.e., sensed entity) up to the application layer.

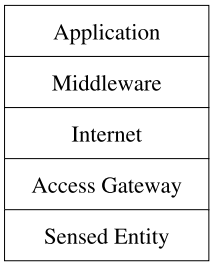

Fig. 1. The architecture of IoT [5].

Now, we are entering the second decade of internet of things, its importance and possibilities of data analytics for IoT system could be found in recent studies [55,12,45,74]. In this stage, data and their meaning for internet of things have gradually become an essential research topic of IoT. That is why recent studies have trended to interdisciplinary collaboration, such as using data mining algorithm to analyze the data of IoT or using cloud computing platform to provide the computing power or storage space for an IoT system. With these integrated systems, a critical task in the next step is how to manage the data. In [55], Santucci used a simple example to explain why the data processing and mining will become critical tasks. This example assumes that there are 6.6 billion human beings on Earth, and they have around 50 billion machines and perhaps around 50,000 billion things, i.e., objects. A tremendous amount of data will then be created from these things in different ways very quickly every time and everywhere. A later study by Fan and her colleagues [12] further defined that the internet of data (IoD) today can be regarded as the recorded activities of all the data entities in the network. In addition to the data management, the data integration and fusion are also critical tasks for finding out useful things from the data of internet of things relevant applications. For example, to develop a modern transportation system [74], today, we need to take into account various data[2] together to understand their relationships so that we are able to find out better ways to enhance the performance of traffic.

To handle and keep such a large amount of data from internet of things is, of course, a challenge because the data own the 3Vs[3] of big data. Several studies [32,37] attempted to come up with a better storage solution to deal with this problem, i.e., the problem of storing data the size of which exceeds that a single machine can handle. Although we can handle such a massive amount of data, an efficient database management system for these data is a promising research subject. That is why Li et al. [32] employed the procedures of preprocessing (e.g., data cleaning and de-duplicating) to reduce the complexity of data. In the same study, the data are stored in a key–value format to enhance the performance of queries. Since the IoT data are large and typically will be updated frequently, Ma et al. [37] used the data partitioning and tree-index to develop an update and query index framework based on HBase to make the data management more efficiently. A later study [81] based not only on the partitioning of dimensions (i.e., attributes) technologies to enhance the efficiency of IoT data management, but it also employed rough set theory to compute and determine the core attribute sets of all the data to reduce the number of dimensions. Jiang and her colleagues [25] presented

[2] These data are from the sensors, CCTVs, and relevant systems that are responsible for collecting the statuses of subways, buses, or even bicycles.

[3] The 3V's of big data are volume, variety, and velocity, which indicate that a large amount of data will be created by different devices with various formats very quickly [31].

Please cite this article in press as: C.W. Tsai, et al., A parallel metaheuristic data clustering framework for cloud, J. Parallel Distrib. Comput. (2017), https://doi.org/10.1016/j.jpdc.2017.10.020.