Knowledge-Based Systems 132 (2017) 215–225

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Complete tolerance relation based parallel filling for incomplete energy big data

Jingling Yuan[a,*], Mincheng Chen[a], Tao Jiang[a], Tao Li[b]

[a] School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China
[b] Department of Electrical and Computer Engineering, University of Florida, Florida 32611, USA

ARTICLE INFO

Article history:
Received 23 May 2016
Revised 15 June 2017
Accepted 20 June 2017
Available online 23 June 2017

Keywords:
Green data center
Incomplete energy big data
Parallel filling on Spark
Complete toleration class
Adaptive management architecture of incomplete energy big data

ABSTRACT

With the approaching of cloud and big data computing era, renewable energy such as solar energy is increasingly integrated into data center power provisioning systems. Nevertheless, the power statistics collection may not be possible or available due to the fact that renewable energy supply exhibits intermittency, time varying behavior (e.g. shortage or failure), resulting in missing data. In this paper, we propose a filling algorithm based on complete tolerance class to solve the missing of energy big data issue. Note that traditional method based on rough sets will likely to fail when there is severe missing data, and its solution on tolerance relation and tolerance class is more complex, which is not suitable for the large scale and the time varying energy big data. Our proposed algorithm expands the tolerance relation into the complete tolerance relation to partition the complete tolerance class. Moreover, our algorithm fills the missing attribute values of the energy big data in data center, which ensures the data integrity and improve the classification accuracy. We further parallelize and optimize our algorithm on state-of-the-art Spark cluster computing framework.

In addition, we propose the adaptive management architecture that handles incomplete energy big data in green data centers. Our proposed architecture integrates the techniques for preprocessing energy data, filling incomplete energy data and building decision model. It increases the power assignment efficiency between solar power and utility, while enhancing load performance and service availability. As a result, it can provide better service for green data centers. We perform comprehensive experiments on an energy data set and the results show the Completing Incomplete Big Data (CIBD) algorithm can guarantee the completeness of data while improving the filling accuracy by 10% compared to general filling algorithms such as MEAN or ERS. The proposed algorithm and architecture show more benefit as the data missing rate increases. We further utilize the filled data to establish the random forest model and yield desirable results. Compared to the Hadoop based filling algorithm, the processing speed of the CIBD algorithm improves by 50% on the 4GB data size.

1. Introduction

As cloud computing and the era of big data is coming, large-scale data centers have been widely deployed around the world and the global data center power demand increases rapidly [1]. The energy consumption of data centers accounted for 0.5% of the world total energy consumption in 2013. It is estimated that by 2020, the average annual energy consumption of the data centers will account for 1% of the global annual total energy consumption [2]. The huge energy consumption of data centers not only increases its operation cost, but also yields negative impact on the environment. Leveraging renewable energy is one of the most desirable ways to amortize data center energy cost and reduce green house gas emission. A recent survey [3] shows that the green data center is the trend for future data center development.

Green data centers utilize solar or other new renewable energy sources as power supply [4]. Note that although renewable energy can relieve the power stress caused by data centers, it usually cannot be exclusively used without the traditional power grid. It is challenging to (1) coordinate the traditional power grid with the renewable energy sources, and (2) adjust the power grid supply according to the real-time variability of both renewable energy production and data center power dissipation [5]. To achieve the above goals and efficient energy source management, we need to efficiently process the energy big data, which consist of energy consumption statistics and other related data in green data centers.

* Corresponding author.
E-mail addresses: yuanjingling@126.com (J. Yuan), wester589@163.com (M. Chen), jany@whut.edu.cn (T. Jiang), taoli.ece.ufl@gmail.com (T. Li).

http://dx.doi.org/10.1016/j.knosys.2017.06.027
0950-7051/

Energy big data exhibit characteristics such as *Volume*, enormous scale and *Velocity*, due to the unstable renewable energy. In addition, energy big data manifest *Variety* since their forms are complex, often mixing with incomplete data caused by power-off and other factors. While the energy big data rarely involve unstructured data, they still contain semi-structured data such as current, voltage and other signals, system logs, renewable energy monitoring data and so on. Furthermore, energy big data have great *Value* for reducing energy consumption if they can be effectively analyzed and processed timely. However, renewable energy such as solar power in green data centers is intermittent, which causes supply instability or shortage. Worse, the power failure can cause server, switch, rack or other equipment failure and the missing of the corresponding power profiling statistics. Meanwhile, the quality of the energy big data will be affected by Remote Terminal Unit (RTU) collection [6], power meter acquisition, channel status, parameters setting, and other factors. Incomplete energy big data, caused by these issues, can severely affect the classification and other models in green data center. For example, if one builds the green data center load prediction model without considering the incomplete energy big data, which may contain important information on load changes, the prediction accuracy of the model cannot be guaranteed. Therefore, it can be a new challenge for the green data center to deal with the incomplete energy big data and build more accurate energy management models.

The main contributions of this paper are as follow:

(1) Our proposed architecture preprocesses the energy big data acquired from the green data centers. Then, it fills the incomplete energy big data, and builds the decision model according to the filled data. Finally, it provides decision feedback to data centers. The architecture increases the power assignment efficiency between solar power and utility, while enhancing load performance and service availability. As a result, it can provide better service for green data centers.
(2) The instable and intermittent nature of renewable energy will result in incomplete data. We proposed the filling algorithm, which based on the complete tolerance relation for Completing Incomplete Big Data (CIBD). The algorithm avoids the limitation of the traditional method based on rough set, and expands the tolerance relation into the complete tolerance relation to partition the complete tolerance class. Moreover, the algorithm fills the missing attribute values of the energy big data in data centers, which ensures data integrity and improves the classification accuracy.
(3) Facing the large scale of energy data in data centers, this paper utilizes the distributed computing framework Spark to optimize the CIBD algorithm, in order to improve the processing efficiency. Experimental results show that our CIBD algorithm improves filling accuracy by 10% compared to general filling algorithms such as MEAN and ERS. The proposed algorithm and architecture show more benefit as the data missing rate increases. We further utilize the filled data to establish the random forest model and yield desirable results. Compared to the Hadoop based implementation, the processing speed of the CIBD algorithm improves by 50% on the 4GB data size.

The rest of this paper is organized as follows: Section 2 provides related work. Section 3 presents the adaptive management architecture of incomplete energy big data in green data centers. The filling algorithm based on complete tolerance class for missing data is described in Section 4. Section 5 performs experiments to analyze the filling results and compares MapReduce vs. Spark implementation. Section 6 concludes the paper and highlights our future work.

2. Related work

Affected by renewable energy instability, intermittency, equipment failures, information acquisition errors and other factors, the energy consumption data of data centers are often inaccurate, inconsistent, and incomplete. Therefore, the incomplete data processing methods are crucial. The processing methods are generally classified to direct and indirect categories.

The direct processing method neither deletes the record containing missing data, nor fills the missing data. Generally, the direct method utilizes extensional rough set model or neural networks for processing. Chen et al. [7] proposed the algorithms for incremental updating approximations of an upward union and downward union of classes by extending dominance characteristic relation based rough sets in the Incomplete Ordered Decision Systems. Their approach presents the principles of dynamically updating approximations w.r.t. attribute values' coarsening. Shu and Qian [8] developed the corresponding attribute reduction algorithms in incomplete decision systems based on indiscernibility relation and discernibility relation, respectively. Clark et al. [9] defined consistency for incomplete data sets using rough set theory, and showed that there exist four distinct types of consistencies of incomplete data. Eirola et al. [10] presented the EM algorithm for direct estimation of pair wise distances in a data set with missing data. Wang [11] utilized the Hopfield neural network based model of classification for incomplete survey data. This model translates an incomplete pattern into fuzzy patterns, along with patterns without missing values, are used as the exemplar set for teaching the Hopfield neural network.

The indirect processing method generally fills the data. It uses some approaches (such as based on Probability Statistics) to transform incomplete information system into complete information system. Zhang et al. [12] proposed a nonparametric iterative imputation algorithm. By selecting some missing values, the algorithm used complete instances to estimate the selected missing values, and iteratively impute missing target values. Gebregziabher and DeSantis [13] utilized a latent class multiple imputation for analyzing missing categorical covariate data. Hong et al. [14] proposed the method to fill the missing value and extract the rules simultaneously by using the upper and lower approximation. We have also proposed a supplement method based on automatic threshold adjustment for dynamic quantization of non-symmetric similarity relation model [15].

Furthermore, the parallel methods based rough sets have been proposed. Qian et al. [16,17] proposed several parallelization strategies of attribute reduction for enhancing the computational efficiency. Chen et al. [18] studied the algorithms for attribute reduction in parallel using dominance-based neighborhood rough sets, which considered the partial orders among numerical and categorical attribute values. Wang et al. [19] utilized the rough set for spatial data mining with the improved parallel algorithm. Zhang et al. [20] compared the parallel large-scale rough set based methods process on the different runtime systems such as Hadoop, Phoenix and Twister, and introduced three different parallel matrix-based approaches to deal with large-scale data [21]. Zhang et al. also proposed a parallel implementation of computing composite rough set approximations on GPUs [22] and Multi-GPU [23].

3. Energy big data management model

3.1. Definition of energy big data

Energy big data consist of two parts: one is the environmental monitoring data, which originate from the internal environment of the monitoring data center and auxiliary equipment, including the measurement and collection for temperature, humidity and other