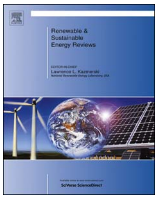
Contents lists available at ScienceDirect

Renewable and Sustainable Energy Reviews

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



Data quality of electricity consumption data in a smart grid environment

Wen Chen[a], Kaile Zhou[a,b,*], Shanlin Yang[a,b], Cheng Wu[c]

[a] School of Management, Hefei University of Technology, Hefei 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making of Ministry of Education, Hefei University of Technology, Hefei 230009, China
[c] Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Keywords:
Electricity consumption data
Data quality
Outlier detection
Outlier data
Smart grid

ABSTRACT

With the increasing penetration of traditional and emerging information technologies in the electric power industry, together with the rapid development of electricity market reform, the electric power industry has accumulated a large amount of data. Data quality issues have become increasingly prominent, which affect the accuracy and effectiveness of electricity data mining and energy big data analytics. It is also closely related to the safety and reliability of the power system operation and management based on data-driven decision support. In this paper, we study the data quality of electricity consumption data in a smart grid environment. First, we analyze the significance of data quality. Also, the definition and classification of data quality issues are explained. Then we analyze the data quality of electricity consumption data and introduce the characteristics of electricity consumption data in a smart grid environment. The data quality issues of electricity consumption data are divided into three types, namely noise data, incomplete data and outlier data. We make a detailed discussion on these three types of data quality issues. In view of that outlier data is one of the most prominent issues in electricity consumption data, so we mainly focus on the outlier detection of electricity consumption data. This paper introduces the causes of electricity consumption outlier data and illustrates the significance of the electricity consumption outlier data from the negative and positive aspects respectively. Finally, the focus of this paper is to provide a review on the detection methods of electricity consumption outlier data. The methods are mainly divided into two categories, namely the data mining-based and the state estimation-based methods.

1. Introduction

At present, with the rapid and stable development of the economy, the demand for energy is constantly expanding. As an important form of energy, electricity plays an important role in the development of modern economy and society [1,2]. With the deepening reform of energy sector and the continuous progress of energy technologies, electric power will play an increasingly indispensable role. Wind energy, solar energy and other new energy sources can be converted into electricity. The electric power industry is one of the most important basic energy industries in the development of national economy. With the development of economy and society, electricity demand continues to expand. It promotes the expansion of electricity consumption market and stimulates the development of electric power industry. In recent years, information technology for electric power industry has also been developed. The construction of smart grid and the application of emerging IT technology make the scale of electric power big data resources continues to grow [3,4]. Construction and application of the smart grid have stimulated the accumulation of electric power grid operation data, production management data and electricity consumption data [5,6]. These accumulated data contain redundant, missing and outlier data, resulting in serious issues of electric power data quality.

Electric power data quality problems are not just faced by China. Many countries in the world have these problems. At present, many countries in the world are in the reform of electricity market and vigorously promoting the construction and application of smart grid. The scale of electricity consumption data collected by advanced metering infrastructure (AMI) in smart grid is becoming increasingly huge in many countries. The quality of these electricity big data has a direct impact on the accuracy and effectiveness of electric power system management and application based on data analysis, which further affects the safety and reliability of the whole power system [7,8]. Therefore, data quality of electricity consumption data is an important research and application issue for the development of power industry in many countries.

The quality of electricity consumption data is a core issue in the process of electricity data mining, and it plays an important role in

* Corresponding author at: School of Management, Hefei University of Technology, Hefei 230009, China.
E-mail addresses: zhoukaile@hfut.edu.cn, kailezhou@gmail.com (K. Zhou).

http://dx.doi.org/10.1016/j.rser.2016.10.054
Received 18 April 2016; Received in revised form 3 September 2016; Accepted 30 October 2016
Available online xxxx
1364-0321/ © 2016 Elsevier Ltd. All rights reserved.

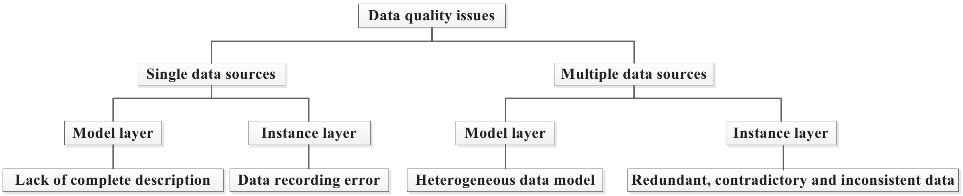


Fig. 1. Classifications of data quality issues [28].

energy big data analytics [9,10]. At present, some researches on data mining focus on the data mining algorithm and ignore the data quality processing before data analysis. Some mature algorithms have requirements on the data set, such as data integrity and data redundancy [8,11–13]. However, in real life, electricity consumption data are incomplete, redundant and ambiguous, which are inconsistent with the requirements of many data mining algorithms [14]. In addition, noise data seriously affect the efficiency of data mining algorithms and result in invalid induction [15–17]. The improvement of electricity consumption data quality has become a key issue in the realization of electric power data mining system.

This paper mainly discusses the data quality issues of electricity consumption. The structure of this paper is as follows. Section 2 introduces the meanings and classifications of data quality. Section 3 describes the data quality of electricity consumption, including the characteristics and classifications of electricity consumption data quality. We focus on the description of three kinds of electricity consumption data quality issues, i.e., noisy data, incomplete data and outlier data. Then in Section 4, we analyze the reasons and significances of electricity consumption outlier data. The detection of electricity consumption outlier data are introduced in Section 5, which can be divided into two categories, namely the data mining-based and the state estimation-based electricity consumption outlier data detection methods. Finally, Section 6 summarizes the full text.

2. Data quality

With the rapid development of Information and Communication Technologies (ICTs), large amounts of data have been accumulated from all walks of life. High data quality is the basic condition of data analysis and mining. In order to discover knowledge from data to support decision-makings, we must ensure the data quality. Therefore, it is necessary to understand the meanings of data quality deeply.

2.1. Data quality management

There are many definitions of data quality. Saha [18] pointed out that data quality refers to the recognition of outlier data and elimination of error data before data is loaded into data center. Huang et al. [19] believed that data quality is the degree of data suitable for use. Roger and Mangiameli [20] focused on the four properties of data quality, namely accuracy, integrity, consistency and timeliness. They pointed out that data quality is the consumer of information. Alizamini et al. [21] believed that data quality is a complex non-structural concept and data refinement process. As can be seen from the above, there is no one uniform consensus on the definitions of data quality. From different point of view, the definitions of data quality are also different. In this paper, we define data quality management as the process of removing the noise, identifying the outlier and processing the incomplete from raw data. Eventually, data quality is higher, and the results of data analysis are more accurate.

Data quality is a comprehensive concept of multidimensional factors, which can be described and evaluated from six aspects, i.e., accuracy, integrity, consistency, self-consistency, availability and timeliness [22]. The data accuracy is the core of data quality, which refers that data must correctly and truly reflect the actual business. Data business descriptions are also reasonable and accurate. The data integrity means that there are no field and record deletion in the recorded data. The recorded data can fully describe the recorded business. The data consistency means that data is logically consistent [23]. It mainly includes three sub-indexes, namely concept, range and format of consistency [24]. The data self-consistency is that data should meet the constraints, which describes the relationships between data [25,26]. Data must satisfy the relationships to each other and cannot be contradictory. The data availability refers to the available degree of data. That is, data should be easy to access, understand and use [27]. The data timeliness implies that data can play a role in the required time. Outdated data has a negative impact on the data quality and indirectly affects the results of data analysis.

2.2. Classifications of data quality issues

Aebi [28] proposed that data quality issues can be divided into four categories, namely the single data source model layer, single data source instance layer, multiple data source model layer and multiple data source instance layer, as shown in Fig. 1.

The single data source issues can be analyzed on two aspects, i.e., the model and instance layer. From the perspective of model layer, lacking complete descriptions and low-grade model designs are main issues. Although database has a complete description of data and model design, it may also have some data quality issues, such as lacking unique or referential constraints. The main issue of instance layer is human errors, such as spelling errors, duplicate records and so on [29–31].

The multiple data source issues are more complex. The multiple data source model layer has heterogeneous data model issues, such as name and structure conflicts [32]. For the multiple data source instance layer, it has redundant, contradictory and inconsistent data.

3. Data quality of electricity consumption data

3.1. Electricity consumption data

At present, large amounts of electricity consumption data have been accumulated. It is difficult for people to directly discover hidden knowledge behind these data which are heterogeneous and inconsistent [33]. In order to discover knowledge, how to ensure data quality is the most important step. To improve the quality of electricity consumption data, its characteristics should be fully considered. With the development of information technology and smart grid, electricity consumption data present the following characteristics.

(1) The volumes of data have increased rapidly. With the development of smart grid, many electric power companies have established Big Data and cloud computing centers to process the data. So it makes the available data increased.
(2) Data transmission and processing speed have greatly accelerated due to the establishment of intelligent equipment and system [34].
(3) Management is more precise. The data mining technology makes the analysis and management of electric energy meters data more precise.