Energy and Buildings 75 (2014) 109–118

Contents lists available at ScienceDirect

Energy and Buildings

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

Data mining in building automation system for improving building operational performance

Fu Xiao*, Cheng Fan

Department of Building Services Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 11 July 2013
Received in revised form 18 October 2013
Accepted 5 February 2014

Keywords:
Data mining
Building automation system
Feature extraction
Clustering analysis
Association rule mining
Recursive partitioning

ABSTRACT

Today's building automation system (BAS) provides us with a tremendous amount of data on actual building operation. Buildings are becoming not only energy-intensive, but also information-intensive. Data mining (DM) is an emerging powerful technique with great potential to discover hidden knowledge in large data sets. This study investigates the use of DM for analyzing the large data sets in BAS with the aim of improving building operational performance. An applicable framework for mining BAS database is proposed. The framework is implemented to mine the BAS database of the tallest building in Hong Kong. After data preparation, clustering analysis is performed to identify the typical power consumption patterns of the building. Then, association rule mining is adopted to unveil the associations among power consumptions of major components in each cluster. Lastly, post-mining is conducted to interpret the rules. 457 rules are obtained in association rule mining, of which the majority can be easily deduced from domain knowledge and hence be ignored in this study. Four of the rules are used for improving building performance. This study shows that DM techniques are valuable for knowledge discovery in BAS database; however, solid domain knowledge is still needed to apply the knowledge discovered to achieve better building operational performance.

1. Introduction

Buildings have great impacts on human life and global sustainability. They consume a large amount of energy to build a comfortable, healthy, safety and productive environment for human beings. Buildings consume 41% of primary energy in the United States, which exceeds the transportation sector (29%) and the industry sector (30%) [1]. In Hong Kong, buildings contribute to nearly 90% of the total electric energy consumption and around 60% of greenhouse gas emissions [2]. Buildings consume energy in their whole life cycles. Normally, the energy use during the operation stage accounts for 80–90% of their lifecycle energy use [3]. Improving building operational performance is of significant importance for energy saving in the building sector.

Modern buildings are usually integrated with a diversity of advanced technologies. Building automation system (BAS) is a typical example which integrates technologies from information science, computing science, control theory and etc. BAS enables modern buildings to be more intelligent through real-time automatic monitoring and control. A huge number of records of temperature, humidity, flow rate, pressure, power, control signals, states of equipment and etc., are stored in BAS database. However, the data in BAS are rarely fully interpreted and utilized. The reason is twofold: the poor quality of the data and the lack of effective and convenient tools for analyzing the large data sets. BAS data usually contain significant missing values and outliers. Modern BASs can only perform simple data analysis and visualization functions, such as historical data tracking, moving averages and alarming of simple abnormalities. They are not capable of systematically analyzing the massive data sets in their database. The building automation industry needs powerful tools to analyze the massive operational data to obtain knowledge for improving building operational performance.

Data mining (DM) is an emerging powerful technology with great potential to discover hidden knowledge in large data sets. In recent years, DM has been gaining increasing interest in various industries, such as banking and financial services, retails, healthcare, telecommunication and counter-terrorism [4]. The use of data mining techniques in the building field also yields encouraging outcomes in energy saving and improving indoor environment. Generally speaking, DM has been used for load prediction, fault detection and diagnosis as well as optimal controls in the building field. Dong et al. [5] used support vector regression models to predict the monthly building energy bills. The research results validated the feasibility and applicability of support vector regression

* Corresponding author. Tel.: +852 2766 4194; fax: +852 2765 7198.
E-mail address: linda.xiao@polyu.edu.hk (F. Xiao).

http://dx.doi.org/10.1016/j.enbuild.2014.02.005
0378-7788/

in the building load forecasting area. Amin-Naseri and Soroush [6] presented a hybrid neural network model combined with the clustering analysis algorithm to predict the daily electrical peak load. It was shown that, compared to statistical methods, such as linear regression, DM-based methods had significant superiority in prediction accuracy. Kusiak et al. [7] developed the ensemble models of the energy consumption of the heating, ventilation and air conditioning (HVAC) components and adopted the particle swarm optimization algorithm to search the optimal set points of HVAC components. It was reported that 7% of HVAC energy consumption could be achieved by using the proposed method. Ahmed et al. [8] investigated the impacts of building characteristics and climate conditions using classification techniques on indoor thermal comforts and indoor illuminance level. Three methods, i.e., the Naïve Bayes, decision tree, and support vector machine, were developed. It was claimed that DM could be used as a decision aid to facilitate building operational processes. Yu et al. [9] established a decision tree model to predict the building energy use intensity. The results showed that the proposed method was able to accurately predict the energy use intensity (93% for training data, and 92% for testing data). Yu et al. [10] adopted association rule mining to save building operational energy. The frequent-pattern growth algorithm was used to generate rules among variables of HVAC air-side system. It was shown that the discovered association rules can be used to identify energy waste, detect equipment faults, and gain insight into building operation. Cabrera and Zareipour [11] demonstrated the capability of association rule mining in identifying lighting energy waste patterns. The research results showed that effective energy saving measures could be generated using the association rules discovered. The simulation results showed that significant savings, as high as 70% of current energy use, were achievable.

Although DM techniques have already been used in the building field, the previous research seldom took full advantage of DM techniques in discovering knowledge underlying massive data sets. The main purpose of using DM techniques in previous work was to develop more accurate, reliable and computationally efficient models, for example, SVM [5], neural networks [6], and decision trees [9]. The number of input variables of the models in the studies above mentioned was still relatively small and the inputs were usually pre-defined based on domain knowledge. For example, domain knowledge in the building field tells us that the building cooling load is affected by the outdoor temperature, humidity, solar radiation, and etc. Therefore, these influential variables were selected as inputs of the model and the output was the cooling load [7]. DM techniques, such as neural network and support vector machine [4,12], were then used to map the inputs to the output, which produced the black-box models. Meanwhile, the previous research work seldom targeted at the massive data sets in the building automation systems. One obstacle to applying DM in BASs is that most of the DM techniques are so sophisticate that few building automation professionals are able to acquire them. The other obstacle is that DM itself cannot tell the value or the significance of the knowledge discovered, and therefore domain knowledge in the building field is still needed to interpret the knowledge for applications in BASs. Therefore, interdisciplinary research should be performed to make a breakthrough in effectively utilizing the massive data sets in BASs by using advanced DM techniques.

A number of DM techniques are available nowadays and more are emerging with the development of technology. There is a lack of generic approach to mining BAS database using DM techniques. This paper proposes an applicable framework for mining BAS database using typical DM techniques. The widely used DM techniques, including clustering analysis [4,12], association rule mining [4,12] and recursive partitioning [13], are adopted in the framework. Besides, the framework is extensible so that other DM techniques can be integrated into the framework gradually. A case study of implementing the framework in mining the data sets retrieved from the BAS of the tallest building in Hong Kong is conducted. The DM algorithms adopted in this study was performed on a Macintosh computer (Mac OS X 10.6.8), a processor of 2.2 GHz (Intel Core i7), and a memory capacity of 8GB. The resulting computation time is shorter than 120 s for each step.

2. Overview of the framework and DM techniques

Fig. 1 illustrates the framework proposed for mining BAS data sets using DM techniques. The framework consists of five steps, i.e. data preparation, clustering analysis, association rule mining (ARM), post-mining and application. Data preparation, or data preprocessing is performed to clean the data sets, reduce data dimension and transform them to suitable formats for data mining. Clustering analysis classifies the large data sets into several clusters according to typical patterns identified in the clustering analysis. Clustering analysis helps to reduce the distance among the data sets and enhance the similarity of the data sets in each cluster. Performing clustering analysis can enhance the reliability of the knowledge discovered in the next step.

Fig. 1 ARM is employed in this study to discover knowledge in the format of rules. A rule generally takes the form of “If *A*, then *B*”, while *A* is called antecedent and *B* is the consequent. Rules are established by exploring relationships between variables in data sets. As an example, a rule “{Shoes = High heels} → {Gender = Female}” states that if one wears high heels, then its gender should be female. Other DM techniques may be adopted in the framework for various purposes as shown in Fig. 1. The post-mining stage focuses on the rule selection and rule interpretation. Finally, the knowledge discovered is used to improve the building operational performance. Data preparation, post-mining and application of discovered knowledge usually require rich domain knowledge; however, the clustering analysis and ARM are almost independent of the system concerned and mainly involve mathematical algorithms. The framework and DM techniques adopted in this study are described as follows.

2.1. Data preparation

Data preparation, or data preprocessing, is an essential step in the process of knowledge discovery using DM. Previous experiences showed that data preparation might take 80% of the total data mining effort [14]. Actually, data preparation is equally important in the model-based methods due to unreliable measurements [15] and process dynamics [16]. The accuracy and reliability of the mining results are largely determined by the data quality. Data preparation is essential to mining BAS data sets because, firstly, the data quality in BAS is usually low due to measurement noises, uncertainties, sensor faults, and insufficient calibration. Two typical problems with BAS data sets are missing values and outliers and they may negatively affect the data mining performance. It was reported that when more than 15% of the training data are outliers, the decrease in neural network model’s accuracy is statistically significant even with small magnitudes of outlierness [17]. Secondly, most DM techniques have special requirements on the data format. BAS data consist of both numerical (quantitative) and categorical (qualitative) data. The typical examples of numerical data are the measurements of temperature, flow rate, power and pressure. The typical examples of categorical data are the ON/OFF control and state signals as well as date and time. Normally, ARM is applied to handle categorical data and therefore numerical data should be transformed to categorical data before conducting ARM. For instance, the power consumption of chiller is numeric data. Discretization methods can be applied to transform the numeric data