BUILD SIMUL (2013) 6: 207 – 222
DOI 10.1007/s12273-013-0117-8

Extracting knowledge from building-related data — A data mining framework

Zhun (Jerry) Yu[1], Benjamin C. M. Fung[2], Fariborz Haghighat[1] (✉)

1. Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, H3G 1M8, Canada
2. Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, H3G 1M8, Canada

Abstract

Energy management systems provide an opportunity to collect vast amounts of building-related data. The data contain abundant knowledge about the interactions between a building's energy consumption and the influencing factors. It is highly desirable that the hidden knowledge can be extracted from the data in order to help improve building energy performance. However, the data are rarely translated into useful knowledge due to their complexity and a lack of effective data analysis techniques. This paper first conducts a comprehensive review of the commonly used data analysis methods applied to building-related data. Both the strengths and weaknesses of each method are discussed. Then, the critical analysis of the previous solutions to three fundamental problems of building energy performance improvement that remain significant barriers is performed. Considering the limitations of those commonly used data analysis methods, data mining techniques are proposed as a primary tool to analyze building-related data. Moreover, a data analysis process and a data mining framework are proposed that enable building-related data to be analyzed more efficiently. The process refers to a series of sequential steps in analyzing data. The framework includes different data mining techniques and algorithms, from which a set of efficient data analysis methodologies can be developed. The applications of the process and framework to two sets of collected data demonstrate their applicability and abilities to extract useful knowledge. Particularly, four data analysis methodologies were developed to solve the three problems. For demonstration purposes, these methodologies were applied to the collected data. These methodologies are introduced in the published papers and are summarized in this paper. More extensive investigations will be performed in order to further evaluate the effectiveness of the framework.

Keywords

building-related data,
data mining,
framework,
influencing factor,
occupant behavior,
energy efficiency

Article History

Received: 14 August 2012
Revised: 10 January 2013
Accepted: 22 January 2013

© Tsinghua University Press and Springer-Verlag Berlin Heidelberg 2013

1 Introduction

The energy consumption in the building sector is of mounting concern. With rising living standards, building energy consumption has significantly increased over the past few decades. The current high level of consumption and the steady increase in demand for energy necessitate a thorough understanding of the major influencing factors in order to develop effective approaches to reducing building energy consumption. Factors influencing building energy consumption can be divided into seven categories (Yu et al. 2011a), as shown in Table 1.

These seven factors play an essential role in reducing building energy consumption and efforts should be made to clearly understand their influences. However, there still are significant barriers that prevent researchers and architects from completely understanding these factors. For example, researchers and architects often observe a large discrepancy between the designed/simulated and the actual building energy consumption. The reasons for this discrepancy are not well understood and often have more to do with occupant behavior than building design. Three fundamental problems of building energy performance improvement that remain significant barriers are as follows:

(1) Building energy demand models are developed mainly for the purposes of the prediction of the total building

E-mail: haghi@bcee.concordia.ca

energy consumption (Yu et al. 2010). How can we develop interpretable building energy demand models that can also help us to understand the influence of these factors on total building energy consumption and that can be easily used by those who lack advanced technical expertises in building thermal analysis?

(2) Various factors influence building energy consumption at the same time, leading to a lack of precision when identifying the individual effects of occupant behavior (Yu et al. 2011a). How can we quantitatively identify such effects without including the impact of other influencing factors such as weather conditions? Moreover, it is difficult to investigate the occupant behavior analytically due to its complicated characteristics (Yu et al. 2012). How can we identify occupant behavior that needs to be modified for energy conservation and make recommendations for such modification?

(3) Based on building automation systems (BASs), large number of heating, ventilation, and air-conditioning (HVAC) system parameters can be monitored and huge amounts of operational data can be collected. How can one examine all the associations (i.e., connections or relationships) and correlations among these parameters and acquire useful information from them to better understand building behavior and develop methodology to reduce its energy consumption?

These barriers can lead to misunderstandings of how the influencing factors affect building energy performance, and thus can add difficulties to developing plans for energy saving measures. Therefore, it is vital that these barriers are removed so that building energy performance can be improved efficiently.

To overcome these barriers, one effective method is to analyze building-related data and acquire relevant useful facts, considering that such data contain abundant knowledge about these influencing factors. In general, building-related data can be categorized into three categories (Yu et al. 2011a), as shown in Table 2.

Since the introduction and implementation of BASs in buildings, vast amounts of building-related data have been collected and stored. Moreover, for an existing building, building-related data can be surveyed through different methods (e.g., analysis of design documentation, questionnaires, and interviews). The data contain abundant information on building design, operation, and maintenance that can be extracted to help reduce building energy consumption. However, the data are rarely analyzed and translated into useful knowledge, mainly due to their complexity (especially, large volumes and poor quality) and a lack of effective data analysis techniques. Consequently, it is necessary to develop more effective data analysis techniques to deal with the challenges caused by the complexity of building-related data. Moreover, based on the proposed data analysis techniques, a data analysis process and a data analysis framework can be established to assist in analyzing building-related data more efficiently. Note that the data analysis process refers to a series of sequential steps in analyzing building-related data. The data analysis framework includes different data analysis algorithms, from which a set of efficient data analysis methodologies can be developed. Both the process and the framework are aimed at successfully extracting (mining) hidden and useful knowledge from building-related data in order to improve building energy performance.

Table 1 Seven categories of influencing factors of building energy consumption

No.	Influencing factor	Example
1	Climate	Outdoor air temperature, solar radiation
2	Building-related characteristics	Type, area, orientation
3	User-related characteristics, except for social and economic factors	User presence
4	Building service systems and operation	Space cooling/heating, hot water supplying
5	Building occupants' behavior and activities	Turn on/off lights, TVs
6	Social and economic factors	Degree of education, energy cost
7	Indoor environmental quality required	Preferred indoor air quality and comfort

Table 2 Three categories of building-related data

No.	Building-related data		Example
1	Climatic data		Outdoor air temperature, outdoor relative humidity
2	Building operational data	Operational data of HVAC systems	Supply air temperature, outdoor airflow rates
		IEQ data	Occupant thermal comfort and health
		Energy data	Monthly electricity consumption, end-use loads of household appliances
3	Building physical parameters		Floor area, window-to-wall ratio

2 Commonly used methods for analyzing building-related data

Methods for analyzing building-related data, with the goal of evaluating and improving building energy performance, can be classified into three main categories (see Table 3). Each method is reviewed and evaluated in the following sections.

Table 3 Three categories of methods of analyzing building-related data

No.	Method	Example
1	Typical indicator method	Annual total energy use, energy use intensity
2	Statistical analysis method	Regression analysis, correlation analysis
3	Building simulation method	EnergyPlus, TRNSYS, ESP-r

2.1 Typical indicator method

Typical performance indicators, particularly indicators of energy use intensity and energy use efficiency, are a simple means of analyzing building-related data and evaluating building energy performance.

2.1.1 *Indicators of energy use intensity*

Energy use intensity (EUI) is a unit of measurement that describes energy consumption of buildings or building service systems, such as space heating/cooling and lighting. Generally, a building's EUI is calculated as the ratio of annual total building energy consumption to the total floor area of the building, thus representing the energy consumed by the building relative to its size. Similarly, building service systems' EUI can also be calculated.

EUIs were mainly utilized to survey building energy use patterns and identify the underlying factors influencing building energy consumption (Chung and Hui 2009; Priyadarsini et al. 2009; Chen et al. 2009a). In particular, these indicators could be utilized to compare the building energy consumption before and after retrofitting, thereby evaluating the energy-saving potential of various energy-efficient technologies (Balaras et al. 2003).

2.1.2 *Indicators of energy use efficiency*

Indicators of energy use efficiency test the quality of consuming energy. A representative indicator in the field of building engineering is *coefficient of performance* (COP), a measure of the energy efficiency of various cooling/heating devices. COP is calculated as the ratio of the amount of energy provided by a system to the amount energy consumed by that system. Therefore, a higher COP indicates a more energy-efficient system.

COPs were used to investigate the effects of various operating conditions, as well as the design parameters, on the performance of heating/cooling systems (such as heat pumps, chillers, and heat transformers) (Balta et al. 2010; Wood et al. 2010; Chekir and Bellagi 2011). In particular, heating/cooling system performance optimization could be carried out by maximizing the COP of the systems (Waltrich et al. 2011) or by identifying the optimum design parameters of the systems with an acceptable result of COP (Bi et al. 2008; Abu Hamdeh et al. 2010). Also, COPs have been used to predict the operating energy performance of heating/cooling systems (Wang et al. 2012).

2.1.3 *Strengths and weaknesses*

The major advantage of the typical indicator method is its simplicity. Moreover, the use of these typical indicators makes it possible to draw a direct comparison of energy performance between different buildings. However, typical indicators alone are insufficient to analyze building-related data and to evaluate building energy performance. Furthermore, they cannot provide insights into building energy-use patterns and investigate the impact of each influencing factor on the total building energy performance.

2.2 Statistical analysis method

Statistical analysis techniques, particularly regression analysis (both linear regression and nonlinear regression) and correlation analysis, were extensively applied to analyze building-related data.

2.2.1 *Regression analysis*

Regression analysis was mainly used to predict building energy consumption based on environmental data or building physical parameters (Lam et al. 1997; Dong et al. 2005b). Also, regression analysis was used to predict among other parameters, such as indoor air temperature and relative humidity (Givoni and Krüger 2003; Krüger and Givoni 2004; Freire et al. 2008), the overall heat transfer coefficient (the *U*-value) (Jiménez and Heras 2005), and the energy consumption of different types of cooling plants (e.g., centrifugal chillers and ice storage systems (Kim and Kim 2007)). An additional application of regression analysis was to compare the effects of influencing factors on building energy performance. For example, Zhang (2004) compared the influence of climatic characteristics on residential building energy performance in China with that in Japan, Canada, and the United States by examining regression equations between annual energy consumption per household and heating degree-days.

2.2.2 Correlation analysis

Correlation analysis was utilized to identify the relationship between building energy consumption and influencing factors such as climate, building physical parameters, occupancy patterns, and HVAC system design and operation (Tonooka et al. 2006; de la Flor et al. 2006; Chen et al. 2009b). The relationship contributes to a clear understanding of the effects of these influencing factors on building energy consumption. Factors with major effects were given priority over other factors during building design and operation, with the ultimate goal of reducing energy consumption (Deng and Burnett 2000).

2.2.3 Strengths and weaknesses

The strength of statistical techniques is their simplicity and widespread familiarity. However, the outcome of regression analysis methods is normally complicated mathematical equations, which are not understandable and interpretable especially for common users without advanced mathematical knowledge. For example, when predicting the building energy consumption, it is difficult to ascertain the influence of the seven factors from the equations. Moreover, building operational data (e.g., operational data of HVAC systems) are usually recorded at short time intervals, which can be considered instantaneous. As a result, various random disturbances that do not usually follow a normal (Gaussian) distribution, such as occupancy, ventilation rates, and solar gains, can add bias and noise to the data, reducing the prediction accuracy (Ghiaus 2006).

Correlation analysis is mainly utilized with the premise that data analysts, based on their expertise, "believe" that strong associations and correlations exist among two or more parameters. For example, one performs correlation analysis between building energy consumption and outdoor air temperature based on "believing" that outdoor air temperature may have a significant influence on the building energy consumption. Such analysis depends mainly on the prior expertise of analysts and adopted statistical techniques. As a result, useful knowledge could be lost, especially indirect associations and correlations between data (e.g., parameters *A* and *B* do not have a direct impact on *C*, but they may have an indirect impact through parameters *D* and *E*). Moreover, commonly a large number of parameters are monitored from HVAC systems and huge amounts of operational data are collected. Consequently, it is very difficult and often infeasible for data analysts to conduct statistical analyses, the correlation analyses, for example, on every combination of the parameters in order to discover all of the associations and correlations that are crucial for achieving the optimum building performance. In this regard, consider, for example, a database with *n* parameters. A data analyst employs traditional correlation analysis to identify the associations/correlations between each pair of the parameters in this database. The number of possible combinations is C(n, 2). Suppose $n = 100$, then the analyst has to conduct 4950 correlation tests, which is impractical.

2.3 Building simulation method

Building energy simulation is another method widely employed to analyze building-related data for evaluating and improving building energy performance. The method was mainly applied to simulate building energy consumption under various conditions in order to identify the relationship between building energy consumption and different influencing factors (e.g., total building energy consumption and building relative compactness (Ourghi et al. 2007), heating/cooling loads and building control strategies (Eskin and Türkmen 2008), and annual electricity consumption and the overall heat transfer coefficient, *U* (Lam 2000)).

2.3.1 Strengths and weaknesses

Building simulation allows for the prediction of building energy performance under various conditions. However, this method does not perform well in simulating energy performance for occupied buildings, as compared to unoccupied buildings, due to a lack of sufficient knowledge about occupant behavior and the patterns of building use, which are normally not deterministic and depend on the occupant and building function.

3 Critical analyses of previous solutions to the fundamental problems

As mentioned in Section 1, there still remain a number of fundamental problems with improving thermal performance of buildings. Different solutions have been proposed and they will be discussed in the following sections.

3.1 Building energy demand models

In recent years, different models have been developed to predict building energy demand. Generally, these models can be divided into three main categories: regression models, simulation models and Artificial Neural Network models. The strengths and weaknesses of regression models and simulation models have already been reviewed in Sections 2.2.3 and 2.3.1, respectively. The application of Artificial Neural Network models is reviewed as follows.

Previous studies showed that Artificial Neural Network models have been widely applied to correlate the total building energy consumption with climatic/physical variables

(Ekici and Aksoy 2009; Escrivá-Escrivá et al. 2011). Also, both building cooling and heating demand were predicted by using Artificial Neural Network models. For example, Olofsson and Andersson (2001) investigated the potential of a neural network to predict the annual space heating demand of a building, based on the measured average daily outdoor and indoor temperatures and space heating energy consumption for a limited time period. Hou et al. (2006) developed a novel method integrating rough sets (RS) theory and Artificial Neural Network in order to predict building cooling demand, based on building operational data of HVAC systems. In addition, it was noticed that, new neural network algorithms such as Support Vector Machines (SVM) were applied to predict building energy demand (Dong et al. 2005a) and building cooling/heating demand (Qiong et al. 2009).

The most important advantage of Artificial Neural Network models, over other models, is the ability to provide predictions even for a multivariable mixed-integer problem, which involves both integers (e.g., binary values) and continuous variables (Yao et al. 2006). However, the major limitation of this method is that the network is considered a black-box model—a relationship between the individual influencing factor and output cannot be observed directly.

In summary, a review of the three main energy demand modeling methods was conducted. These modeling methods have been successfully applied to predict building energy demand. However, the models developed using these methods are not understandable and interpretable. This makes it difficult for these methods to provide useful knowledge of building energy performance improvement. New modeling methods need to be proposed in order to overcome such limitations.

3.2 Study on occupant behavior

3.2.1 The influences of occupant behavior on building energy consumption

Recently there has been mounting interest in studying the influences of occupant behavior on the total building energy consumption (Kyrö et al. 2011) or end-use loads such as lighting (Yun et al. 2012) and space/water heating (Santin et al. 2009). Generally, these studies can be divided into two categories. The first category focuses on the effects of occupant presence on building energy consumption. For example, Emery and Kippenhan (2006) reported a survey on the effects of occupant presence on energy usage in four nearly identical houses. These houses were divided into two pairs, and the building envelope of one pair was constructed with improved thermal resistance. One of each pair of houses was left unoccupied, while the other was occupied. They compared the total building energy consumption of the occupied and unoccupied houses in a heating season. Masoso and Grobler (2010) compared energy consumption during non-working hours with that during working hours in six commercial buildings. Sub-hourly power consumption profiles for different buildings have been audited and compared.

The second category of studies focuses on the effects of occupants' actions on the building energy consumption. For example, Schweiker and Shukuya (2010) investigated the quantitative effects of the occupant behavior change (frequency of using the air-conditioner units, either occasionally or frequently) and building envelope improvement on the exergy consumption for heating and cooling. The data were collected in 39 student rooms of a university dormitory. Ouyang and Hokao (2009) investigated energy-saving potential by improving user behavior in 124 households in China. These houses were divided into two groups: one group was educated to promote energy-conscious behavior and put corresponding energy-saving measures into effect, while the other group was not. Comparisons were made between monthly energy consumption for both groups.

Evidently, comparative analyses on measured data were conducted in these studies to identify the effects of occupant behavior. However, apart from occupant behavior, the other influencing factors also simultaneously contribute to the variation in building energy consumption, while this method is unable to adequately remove the effects of those factors and identify the influences of occupant behavior. Although in these studies some measures were implemented to remove the impact of those factors, such as by using nearly identical housing characteristics and by taking energy data in other years with similar climatic conditions as a reference, the effects of these measures are questionable since even a slight difference in some building parameters (e.g., heat loss coefficient, infiltration) and weather parameters (e.g., annual average outdoor air temperature) would result in remarkable fluctuations in the building energy consumption. Therefore, new methods need to be developed in order to identify the effects of occupant behavior precisely.

3.2.2 Modification of occupant behavior

A number of studies have been conducted to study the modification of building occupants' behavior. As reviewed in Section 3.2.1, Ouyang and Hokao (2009) improved occupant behavior in 124 households in China and estimated the energy-saving potential. Al-Mumin et al. (2003) simulated occupant behavior improvement (i.e., occupant behavior before and after the modification such as turning off lights when rooms are empty and setting the air-conditioner thermostat to a higher temperature) and the corresponding annual electricity consumption reduction by using the energy simulation program ENERWIN. In these studies,

two approaches were used to modify occupant behavior: energy-saving education approach and building simulation approach. Both of them can have an immediate effect on the building energy consumption reduction. However, both approaches have certain limitations.

Regarding the energy-saving education approach, commonly detailed energy-saving measures and tips on the efficient use of various household appliances should be provided for occupants. Considering that a family normally has a number of appliances, and that each appliance may require specific tips (e.g., for air-conditioners: select the highest thermostat setting within the comfort range, clean or change air filters regularly, keep indoor/outdoor coils clean, get a programmable thermostat, etc.), there could be a large number of energy-saving measures and tips for an individual family. For example, one family may have 40 household appliances, with each appliance having an average of 7 energy-saving tips. Accordingly, the occupants need to follow and implement 280 tips, which is impractical. Although a booklet of these tips can be prepared for building occupants, it is very difficult for occupants to remember them all distinctly and implement them for a long time in practice. Furthermore, occupants may not fully understand and have confidence in these tips' effectiveness because they only provide qualitative information. In addition, some energy-saving opportunities can only be initiated by building occupants themselves. For example, when occupants realize they have consumed too much energy on both TVs and computers, they can avoid using both devices simultaneously when they can really only focus on one of them, or make a conscious effort to reduce usage time. Therefore, instead of simply providing occupants with a number of general energy-saving recommendations, it is more rational and efficient to help them modify their behavior in two steps. First, it is necessary to identify the behavior that needs to be modified. This can be achieved by analyzing measured building-related data. Second, feasible recommendations to mitigate the identified behavior can be presented with the goal of reducing total building energy consumption.

With regard to the building simulation approach, current simulation tools can only imitate some typical activities in a rigid way, such as the control of sun-shading devices, while realistic building occupant behavior patterns are more complicated.

In summary, new methods are needed for evaluating occupant behavior in existing buildings and for helping occupants to efficiently modify their activities/usage.

3.3 Associations and correlations between building-related data

Building-related data may have a direct/indirect influence on each other, considering that they are closely related to the same buildings. Specifically, there may be strong associations and correlations between them. Both these associations and correlations should be examined to understand building operation, determine rules of conserving energy, and develop appropriate strategies to design buildings.

A number of studies have been conducted to identify associations and correlations between measured building-related data. Researchers utilized statistical analysis techniques, particularly correlation analysis, and focused mainly on the relationships between building energy consumption and its influencing factors, as reviewed in Section 2.1.2 (the limitation of these techniques has already been addressed). Moreover, few researchers examined associations and correlations between building operational data, especially operational data of HVAC systems, to better understand building operation in order to improve building energy performance. This is mainly due to the complexity of such data and a lack of effective data analysis techniques. Note that the energy consumption of HVAC systems can account for a large portion of total building energy consumption (Pérez-Lombard et al. 2011).

Clearly, new methods are needed for discovering all the useful and important associations and correlations between building operational data.

4 Data mining and its applications in building engineering

4.1 Data mining

Considering the limitations of the data analysis methods commonly used in building engineering, data mining is proposed as a primary tool to analyze building-related data to extract useful and hidden knowledge. Data mining techniques excel at automatically analyzing huge amounts of data and searching for useful information: these techniques fit well with the purpose of this study.

Different definitions of data mining have been given by various researchers. For example, Hand et al. (2001) define data mining as "*the analysis of large observational datasets to find unsuspected relationships and to summarize the data in novel ways so that data owners can fully understand and make use of the data*." As defined by Cabena et al. (1998), data mining is "*an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large databases*." Based on these statements, it can be concluded that data mining is essentially a combination of multidisciplinary approaches. It is often used to extract unseen patterns from a large volume of data and to transform them in turn into

useful/practical knowledge that could benefit future work, such as reducing building energy consumption.

In this study, three main data mining techniques: data classification, cluster analysis, and association rule mining are proposed. Before these techniques are introduced, basic terms and concepts in relation to data mining are explained:

- **Dataset, Attribute, and Instance**. A dataset is a set of data items. It is roughly equivalent to a two-dimensional (i.e., column and row) spreadsheet or database table, as shown in Fig. 1. Each database table consists of a set of attributes (usually in different columns or fields) and stores a large set of instances (usually in rows or records). Consider a mechanically ventilated building with 1000 monitored parameters. Each parameter can be considered an attribute, and a record of all these parameters in a specific time point can be considered an instance.
- **Target attribute, Predictor attribute**. A target attribute is the attribute predicted as a function of other attributes (i.e., predictor attributes). For example, building energy consumption (i.e., target attribute) could be predicted as a function of building-related parameters such as floor area and number of occupants (i.e., predictor attributes).

Based on the above definition of data mining terms, data classification, cluster analysis, and association rule mining are described as follows. Moreover, considering that each data mining technique has a number of algorithms, a typical algorithm of each technique is also briefly introduced.

- **Data classification**. A form of data analysis that can be used to build classification models describing important data classes. The models are constructed for *target attributes* as a function of the values of *predictor attributes*. The goal of data classification is to classify data into various predefined classes (e.g., air-conditioner operating states can be classified as "ON" or "OFF"), thereby providing the description, categorization, and generalization of given databases.

The decision tree algorithm is one of the most commonly used data classification algorithms; it uses a flowchart-like tree structure (Quinlan 1986; Han and Kamber 2006).

	Attribute 1	⋯	Attribute m
Instance 1	×	⋯	×
⋯	⋯	⋯	⋯
Instance i	×	⋯	×
⋯	⋯	⋯	⋯
Instance j	×	⋯	×
⋯	⋯	⋯	⋯
Instance n	×	⋯	×

Fig. 1 A schematic diagram of dataset, attribute, and instance

Figure 2 shows a decision tree indicating whether occupants turn lights on or off in their offices. Assume 100 instances are used to build this decision tree, and each instance has three attributes: working hours, office occupancy, and the state of lights. The target variable for the above decision tree is light states, with potential states being classified as either turning on or down. The predictor variables are working time (during normal working hours or not) and office occupancy (occupied or empty). As shown in Fig. 2, the decision tree consists of three kinds of nodes: root, internal, and leaf. A root node and an internal node denote a binary split test on an attribute, while a leaf node represents an outcome of the classification, and thus holding a categorical target label. Moreover, the numbers in the parentheses at the end of each leaf node depict the number of data records in this leaf. If some leaves are impure (i.e., some instances are misclassified into this node), the number of misclassified instances will be given after a slash. For example, (60/5) in the leftmost leaf in Fig. 2 means that among the 60 instances not during normal working hours that have been classified as turned down, 5 of them actually have the value turned on. Whether light states should be classified as being "turned on" or "turned down" can be predicted by using this decision tree. For example, if the working time is during normal working hours and the office is not empty, occupants will turn lights on; otherwise they will turn them down.

The procedure for generating a decision tree is explained as follows. Initially, all instances are grouped together into a single partition. At each iteration, the algorithm chooses a predictor attribute that can "best" separate the target class values in the partition. The ability that a predictor attribute can separate the target class values is measured based on an attribute selection criterion, which can be referred to (Yu et al. 2010). After a predictor attribute is chosen, the algorithm splits the partition into child partitions such that each child partition contains the same value of the chosen selected attribute. The decision tree algorithm iteratively splits a partition and stops when either of the following terminating conditions is met:

(1) All records in a partition share the same target class value. Thus, the class label of the leaf node is the target class value.
(2) There are no more records for a particular value of a predictor variable. In this case, a leaf node is created with the majority class value in the parent partition.

- **Cluster analysis**. The process of merging data into different clusters so that instances in the same cluster have a high similarity, while instances in different clusters have a low similarity. The similarity between the instances is evaluated based on their attribute values, and it is normally computed based on the distance between

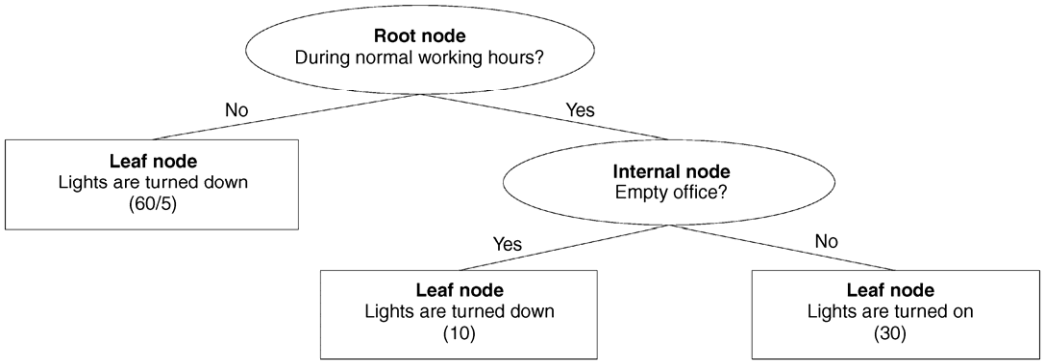

Fig. 2 Schematic illustration of a simple hypothetical decision tree

each pair of instances. One popular distance measure is Euclidean distance (Han and Kamber 2006):

$$d(U,V) = \sqrt{(x_{u1} - x_{v1})^2 + (x_{u2} - x_{v2})^2 + \cdots + (x_{um} - x_{vm})^2}$$

where $U = (x_{u1}, x_{u2}, \cdots, x_{um})$ and $V = (x_{v1}, x_{v2}, \cdots, x_{vm})$ are instances. $x_{u1}, \cdots, x_{um}$ are m attribute values of U and $x_{v1}, \cdots, x_{vm}$ are m attribute values of V.

Figure 3 shows a clustering schema based on a hypothetical building data table. It contains various energy-related attributes such as outdoor air temperature (T) and building heat loss coefficient (HLC).

The data table consists of m attributes and n instances. Each attribute represents a variable and each instance denotes a building. All the instances are grouped into w clusters. Accordingly, these w clusters are homogeneous internally and heterogeneous between different clusters (Han and Kamber 2006). Such internal cohesion and external separation are based upon the m attributes as well as their influences; it implies that these attributes have the most similar holistic effects on the building energy performance of the same cluster buildings, while the effects are significantly distinct for buildings in different clusters.

		Attribute 1 (T)	...	Attribute m (HLC)
Cluster 1	Instance 1	×	×	×
	...	×	×	×
	Instance i	×	×	×
⋮	...	×	×	×
	Instance j	×	×	×
Cluster w	...	×	×	×
	Instance n	×	×	×

Fig. 3 Clustering schema

The *K-means algorithm* is one of the simplest partition methods to solve a clustering problem. Given a dataset (D) containing l instances, the K-means algorithm aims to partition these l objects into k clusters with two restraints: (1) the center of each cluster is the mean position of all instances in that cluster; and (2) each instance has been assigned to the cluster with the closest center. This algorithm consists of five steps: (1) randomly select k instances from the dataset as the initial cluster centers; (2) calculate the distance between each remaining instance and each initially chosen center; (3) assign each remaining instance to the cluster with the closest center; (4) recalculate the mean values, i.e., the cluster centers, of the new clusters; and (5) repeat steps 2 to 4 until the algorithm converges, meaning that the cluster centers do not change.

- **Association rule mining (ARM)**. It is a method to identify all associations and correlations between attribute values. The output is a set of association rules that are used to represent patterns of attributes that are frequently associated together (i.e., frequent patterns). For example, assume that 100 occupants live in 100 different rooms in a building, and each occupant has both a computer and a table lamp. Assume 40 occupants turn on their computers and 20 occupants turn on their table lamps. If 10 occupants turn on both computers and table lamps during the same period of time, it can be calculated that these 10 occupants account for 10% of all the building occupants, and 25% of the occupants who turn on computers. Then, the information that occupants who turn on computers also tend to turn on table lamps at the same time can be represented in the following association rule:

$$\text{turn\_on\_computers} \rightarrow \text{turn\_on\_table\_lamps}$$
$$[support = 10\%, confidence = 25\%]$$

In this statement, *support* and *confidence* are employed to indicate the validity and certainty of this association rule. Different users or domain experts can set different

thresholds for *support* and *confidence,* according to their own requirements, in order to discover useful knowledge in the end. Accordingly, association rule mining can be defined as discovering association rules that satisfy the predefined minimum *support* and *confidence* from a given database.

Mathematically, *support* and *confidence* can be calculated by probability, $P(X \bigcup Y)$[1], and conditional probability, $P(Y|X)$, respectively (X denotes the premise and Y denotes the consequence in the sequence). That is,

$$support(X \rightarrow Y) = P(X \bigcup Y)$$

$$confidence(X \rightarrow Y) = P(Y \mid X)$$

Another concept, *lift,* which is similar to *confidence,* is commonly used to demonstrate the correlation between the occurrence of X and Y when conducting the ARM. Mathematically,

$$lift(X \rightarrow Y) = \frac{P(X \bigcup Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

Particularly, a *lift* value greater than 1 represents a positive correlation (the higher this value is, the more likely that X coexists with Y, and there is a certain relationship between X and Y (Han and Kamber 2006)), while a *lift* value less than 1 represents a negative correlation. If the value is equal to 1, i.e., $P(X \bigcup Y) = P(X)P(Y)$ the occurrence of X is independent of the occurrence of Y, and there is no correlation between X and Y.

A commonly used *ARM algorithm* is the *frequent-pattern growth (FP-growth) algorithm* (Han and Kamber 2006). The FP-growth algorithm adopts a "divide-and-conquer" strategy to further improve the efficiency of examining association rules in a database. A frequent-pattern tree is first constructed to represent the database. Based on this tree, the database is divided into a set of sub-databases that will be mined separately.

Data classification, Cluster analysis, Association rule mining have been extensively applied in various fields such as industrial and medical (Delgado et al. 2001; Jiao and Zhang 2005; Georgilakis et al. 2007; Pan et al. 2007; Hsu 2009). However, their applications in the field of energy are still sparse. It should be mentioned that, due to the fact that several classification methods (e.g., Artificial Neural Network method, Genetic Algorithm, Rough Set approach, and Fuzzy Set approach) were less commonly used for data classification in commercial data mining systems, in this study these methods were not assigned to data classification (but they are included in the data mining system).

[1] In data mining, the notation $P(X \bigcup Y)$ indicates the probability that an instance contains both X and Y, but not either X or Y.

4.2 Current applications of data mining in building engineering

Previous work seldom studied how to use these data mining techniques to process building-related data and extract useful and hidden knowledge. To the best of our ability, no literature was found regarding the association rule mining technique.

Tso and Yau (2007) used the data classification technique to compare the accuracy of regression analysis, the Artificial Neural Network method, and the decision tree method (i.e., one typical data classification method) in predicting the average weekly electricity consumption for both summer and winter in Hong Kong.

Santamouris et al. (2007) applied cluster analysis technique to classify and rate the energy performance of school buildings. Based on the cluster analysis and Principal Component Analysis (PCA) techniques, Gaitani et al. (2010) proposed an approach to rating the energy performance of space heating and evaluating the potential energy savings in the school sector. Also, Lam et al. (2009) combined cluster analysis and PCA to identify climatic influences on chiller plant electricity consumption. Wu and Clements-Croome (2007) applied the cluster analysis technique to analyze noisy indoor environmental data measured from wireless sensor networks. They used cluster analysis first to discover outliers and then to estimate the distribution of indoor temperature.

In summary, data mining is a relatively new concept/tool applicable to energy conservation in buildings. By definition (see Section 3.1), there is a distinct possibility that, in order to improve building energy performance, data mining can be employed to extract hidden useful knowledge from huge amounts of building-related data. To achieve this goal, in the following sections a data analysis process and a systematic data mining framework are proposed that enable building-related data to be analyzed more efficiently.

5 Proposed data analysis process

A step-by-step data analysis process of extracting useful knowledge from building-related data is proposed (see Fig. 4).

- Problem definition and objective setting;
- Data source selection: select buildings available to collect building-related data;
- Data collection: collect building-related data through BASs, field surveys, etc., and then construct a database;
- Data pre-processing/preparation: detect and remove outliers and noise, handle missing values, deal with inconsistencies and complexity through data transformation (e.g., transforming daily data into monthly data) and

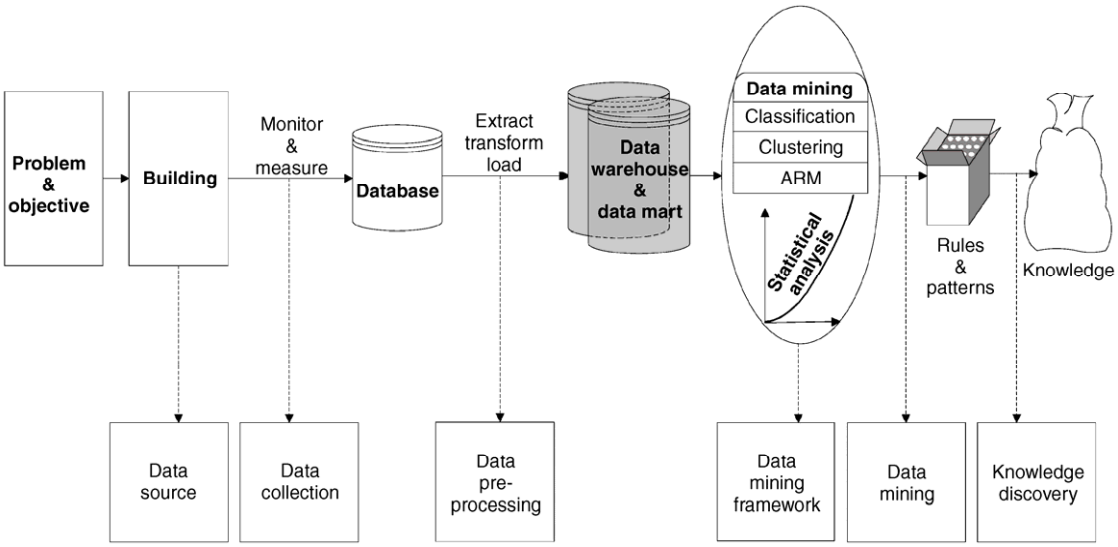

Fig. 4 Process for data analysis in building engineering

integration (e.g., combining data from multiple sources), etc.;

- Data warehouses (DWs) or data marts construction: construct DWs or data marts so as to model multidimensional data and provide *on-line analytical processing*. The gray block in Fig. 4 denotes that the step is unnecessary when the database is relatively small and there is no need to build a high-dimensional DW;
- Data mining and model construction: perform data mining based on the proposed data mining framework;
- Results analysis and evaluation: identify the most useful rules and patterns from the data mining results;
- Knowledge discovery and presentation: discover useful knowledge based on both expertise in building engineering and obtained rules/patterns.

6 Data mining framework

Figure 5 shows the proposed data mining framework consisting of four components, which are introduced below.

6.1 Component 1: Data analysis techniques/algorithms

Component 1 indicates the data analysis techniques used in this framework, including both the three data mining techniques (Data classification, Cluster analysis, Association rule mining) and traditional statistical analysis (e.g., correlation analysis and confidence levels). Typical algorithms of each data mining technique are also provided in this framework. For example, data classification can be conducted by using the *decision tree* algorithm (classify and/or predict categorical parameters) or *regression tree* algorithm (classify and/or predict numerical parameters); cluster analysis can be conducted using the *K-means* algorithm (clustering low-dimensional data) or *CLIQUE* algorithm (clustering high-dimensional data); and association rule mining can be conducted using the *Apriori* algorithm (mining Boolean association rules in small datasets) or *FP-growth* algorithm (mining Boolean association rules in large datasets) (Cios 2007; Lior and Oded 2008; Cao et al. 2009). Furthermore, different data mining techniques can be combined to mine building-related data, such as cluster analysis and data classification (e.g., clustering-then-classification), or cluster analysis and association rule mining (e.g., association rule clustering system and frequent pattern-based clustering analysis). It should be mentioned that data mining techniques/algorithms can be implemented using the open-source data mining program RapidMiner, which provides a simple and friendly graphical user interface (GUI) (RapidMiner 2012).

6.2 Component 2: Potential applications of data mining in building engineering

Component 2 indicates the potential applications of data mining in the field of building energy conservation. Generally, data mining can be applied to help accomplish three categories of tasks:

- Construct data analysis models to classify and/or predict building-related attributes, thereby benefiting the design and operation of energy-efficient buildings. Both numerical attributes (e.g., cooling/heating loads) and categorical attributes (e.g., building energy consumption classified as either 'HIGH' or 'LOW') can be used as *target attributes*.

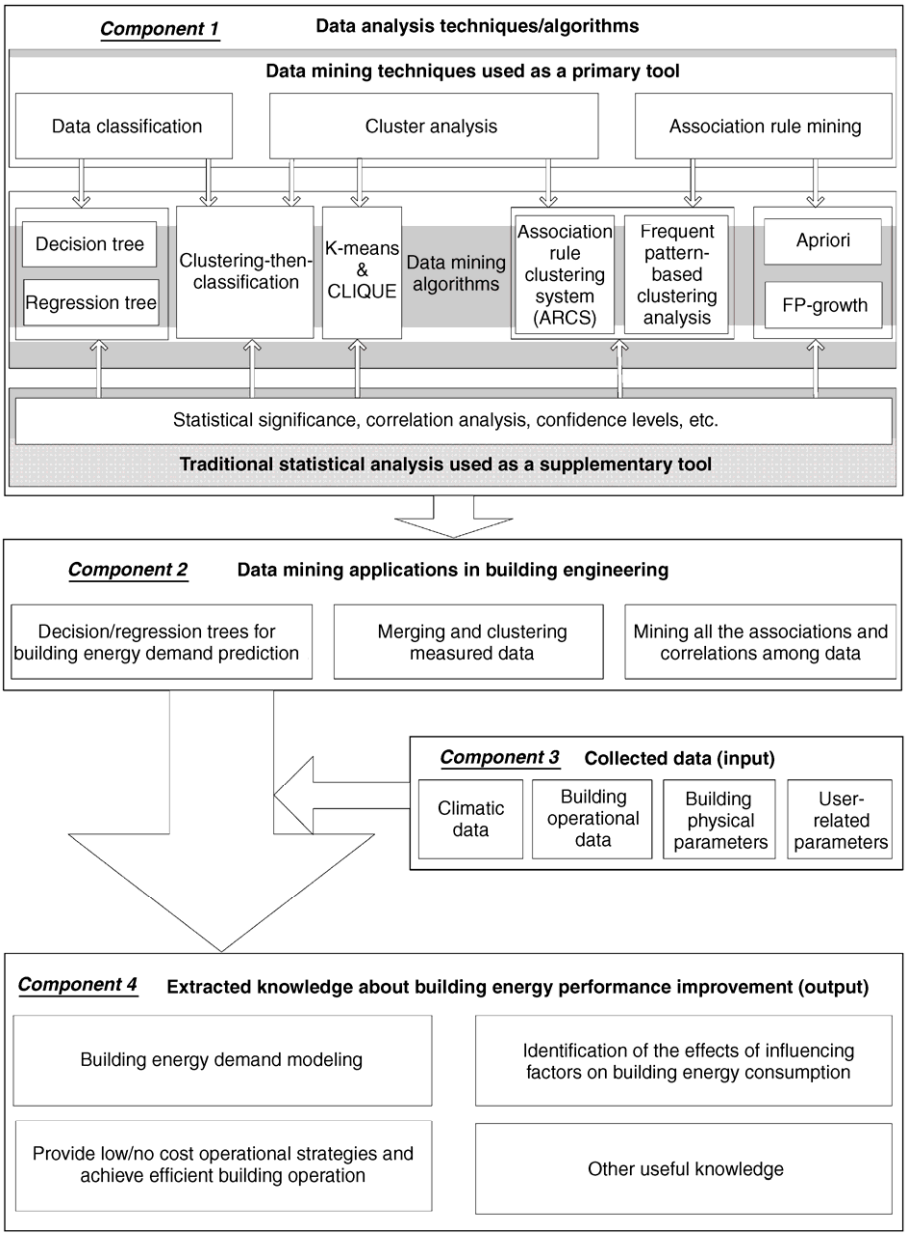

Fig. 5 Overview of the proposed data mining framework

This task is mainly performed using *data classification*.

- Compute similarities/dissimilarities between objects (e.g., buildings and occupant behavior) represented by various building-related attributes (e.g., floor area and building age) and user-related attributes (e.g., number of occupants). The goal is to provide insights into building energy consumption patterns and identify the effects of influencing factors on building energy consumption. This task is mainly performed using *cluster analysis*.
- Discover useful associations and correlations among measured data (e.g., various HVAC system parameters such as supply air temperature, supply air flow rates, fan pressure drop, etc). The goals are to better understand building operation and provide opportunities for energy conservation. This task is mainly performed using *association rule mining*.

6.3 Component 3: Input

Component 3 indicates the data that can be input to the framework. Both building-related data and user-related data can be measured, collected, and input to the proposed framework for knowledge discovery. The collected data can be stored in the *dataset* illustrated in Fig. 1.

6.4 Component 4: Output

Component 4 indicates the output of the framework. According to the three categories of tasks in *Component 2*, the framework mainly outputs the following useful information about building energy performance improvement:

- building energy demand modeling;
- the effects of influencing factors on building energy consumption;
- low/no cost operational strategies for building operation.

Other useful knowledge can also be output in terms of the objective setting, available data, and the usage of data mining techniques/algorithms.

7 Demonstration of the applicability of the process and framework

To demonstrate the applicability of the proposed process and framework, the three fundamental problems listed in Section 1 were addressed in sequential steps in the process. Two available data sources were selected and from them two sets of data were collected. Then a database was constructed. The two datasets are briefly introduced as follows:

Dataset 1 To evaluate and improve residential buildings' energy performance, a project entitled "Investigation on Energy Consumption of Residents All over Japan" was carried out by the Architecture Institute of Japan from December 2002 to November 2004 (Murakami et al. 2006). For this project, field surveys on energy-related data and other relevant information were carried out in 80 residential buildings located in six different districts in Japan: Hokkaido, Tohoku, Hokuriku, Kanto, Kansai, and Kyushu. Table 4 shows the survey items and corresponding investigation methods.

Dataset 2 The EV pavilion in Montreal, a complex building that mainly includes offices and wet labs, was selected as another data source in this study. This building consists of two parts: the ENCS part (17 floors) and the VA part (12 floors). Both parts have their own VAV air-conditioning systems. In order to conduct the case study, the historical data of the air-conditioning systems in both parts were collected from December 2006 to May 2009. However, since the online monitoring program was updated from November 2007 to January 2008, data reports were not generated during this period. In total, 61 parameters were monitored at a 15-minute interval (Yu et al. 2012).

To solve the three fundamental problems, four data analysis methodologies were developed based on the framework and applied to the collected data. The four methodologies are introduced in (Yu et al. 2010; 2011a,b; 2012) and summarized as follows. Interested readers can refer to (Yu et al. 2010; 2011a,b; 2012) for more detailed descriptions.

(1) Classification analysis (*decision tree*) was applied to develop a methodology for establishing building energy-demand predictive models (i.e., decision tree-based models) (Yu et al. 2010). To demonstrate its applicability, the methodology was applied to *Dataset 1* in order to estimate residential building energy performance indexes by modeling building energy use intensity (EUI) levels (either high or low). The results indicate that the methodology's competitive advantage over other widely used modeling techniques, such as regression methods and Artificial Neural Network methods, lies in its ability to generate accurate predictive models with interpretable flowchart-like tree structures that enable users to quickly extract useful information. The accuracy of predicting the EUI levels is 92% (for comparison, prediction models by using regression methods and Artificial Neural Network methods were also developed based on the same data set. The accuracy of the obtained regression model and Artificial Neural Network model is 72% and 88%, respectively. However, it should be mentioned that the decision-tree model can only predict the EUI levels while the regression model and Artificial Neural Network model can predict the EUI values). Moreover, a lot of useful information on building energy performance improvement can be extracted from the developed model. For example, it can automatically identify and rank significant influencing factors of building EUI, as shown in Table 5. Note that outdoor air temperature was found as the most important factor and, for clarity, influencing factors were ranked in high and low temperature regions in Japan separately.

Table 4 Investigation items and methods

Method	Survey item		Measuring time
Field measurement	Different end-use loads of all kinds of fuel	Electricity	Measured every minute
		Gas	Measured every 5 minutes
		Kerosene	Measured every 5 minutes
	Indoor air temperature (1.1 m above floor)		Measured every 15 minutes
Questionnaire survey	Lifestyle, utilization of equipment, annual income, etc.		Once only
Inquiring survey	Other issues, such as basic building information		Once only

Table 5 Influencing factors and their rank

Rank	Influencing factor	
	High temperature region	Low temperature region
1	Heat loss coefficient	Space heating mode[c*]
2	Equivalent leakage area	Number of occupants
3	Hot water supply mode[a*]	House type[d*]
4	Kitchen energy mode[b*]	Heat loss coefficient

[a*] Electric or non-electric; [b*] Electric or gas; [c*] Electric or non-electric; [d*] Detached or apartment

Based on such information, designers can clearly realize which parameter deserves extra attention when designing energy-efficient buildings. Also, the model can provide the combination of significant factors as well as the threshold values that will lead to high building energy performance. For example, in high temperature regions, a lower building heat loss coefficient than 3.89 W/(m^2·K), together with a high equivalent leakage area (> 4.41 cm^2/m^2), will normally cause a low EUI. The detailed discussion on such information can be found in (Yu et al. 2010). Another advantage is that it can be utilized by users without requiring a lot of computation knowledge. The generated model, and the derived information, could greatly benefit building owners and designers; one crucial benefit is the reduction of building energy consumption.

(2) Cluster analysis (*K-means*) was used to develop a methodology for examining the effects of occupant behavior on building energy consumption (Yu et al. 2011a). Such effects can be shown by "removing" the effects of the first four factors in Table 1. Note that the first four factors are unrelated to occupant behavior. The last two factors which represent occupants' influences affect building energy consumption indirectly. Their influences are already contained within the effects of occupant behavior, and there is no need to take them into consideration when identifying the effects of occupant behavior. The methodology is realized by clustering similar buildings into various groups based on the first four influencing factors, so that for each building in the same group the four factors have similar effects on building energy consumption. Accordingly, the effects of occupant behavior can be identified accurately in these groups. The identification of building groups is the most important element of this methodology and it is achieved mainly via cluster analysis. To demonstrate its applicability, the methodology was applied to *Dataset 1*. The effects of occupant behavior, as well as behavior patterns, were identified through examining different end-use loads associated with occupant behavior. The following data analysis was conducted:

(a) Analysis of the average annual EUI of different end-use loads for each cluster—this mainly indicates the degree to which various behavior influences the total building energy consumption.

(b) Analysis of the variability in annual EUI of different end-use loads for each cluster—a boxplot of annual EUI of each end-use load was drawn. A large variability implies that there still remains great potential for energy saving by improving occupant behavior related to the end-load uses.

(c) Analysis of monthly variations of average end-use loads for each cluster—this mainly indicates the effects of occupant behavior over both time and buildings.

(d) A reference building for each cluster is defined, and then the energy-saving potential of buildings in each cluster can be evaluated by comparison with the reference building.

(e) Analysis of monthly average indoor temperature of air-conditioned room of three typical buildings (i.e., the reference building, buildings with the maximum and minimum annual EUI) for each cluster—the effects of occupant behavior should be understood and interpreted in conjunction with the investigation of indoor climate. Occupant behavior, especially those associated with HVAC, can significantly affect indoor climate, which in turn will have an influence on occupant behavior, thereby causing dramatic differences in building energy consumption.

The results show that the methodology facilitates the evaluation of building energy-saving potential by improving the behavior of building occupants, and provides multifaceted insights into building energy end-use patterns associated with the occupant behavior.

(3) Association rule mining (*FP-growth*) was employed to develop a methodology for examining all associations and correlations among building operational data, thereby discovering useful knowledge about energy conservation (Yu et al. 2012). Building operational data in two different time periods (i.e., both a day and a year) are mined; associations and correlations between operational data in different time periods could be significantly different. Hence, it can help us find and take advantage of more complete associations and correlations. Moreover, data in two different years are mined, and the associations and correlations in the two years are compared. Such comparison can assist in identifying a discernible change in associations and correlations, and also in building operation, thereby uncovering useful information. To demonstrate its applicability, the methodology was applied to *Dataset 2*. The procedure was able to:

(a) identify the energy waste in the air-conditioning system (e.g., it was found that, in the fresh air handling units, the heat added to the outdoor air was first transferred to humidifier, and then simply drained to municipal sewage. This energy waste was confirmed through the discussion with the building operator);

(b) detect the equipment faults (e.g., it was found that, either fan 1 or fan 2 (or both of them) in a fresh air handling unit has a fault);

(c) propose low/no cost strategies for saving energy in system operation (e.g., it was found that, the existing operating strategy of extracting exhaust air from the building was to use two of three fans while the other one was turned off. Given that these three fans are identical and controlled by individual variable-speed drive, one possible energy-saving method is to use all these three fans instead of two of them).

The results obtained could help us to better understand building operation and provide opportunities for energy conservation.

(4) Cluster analysis (*K-means*), classification analysis (*decision tree*), and association rule mining (*FP-growth*) were combined to formulate a methodology for identifying and improving occupant behavior in buildings (Yu et al. 2011b). In this study, end-use loads were divided into two levels (i.e., main and subcategory), and they were used to deduce corresponding two-level user activities (i.e., general and specific occupant behavior) indirectly. Cluster analysis and classification analysis were combined to analyze the main end-use loads, in order to identify *energy-inefficient general occupant behavior*. Then, association rules were mined to examine end-use loads at both levels to identify *energy-inefficient specific occupant behavior*. To demonstrate its applicability, the methodology was applied to *Dataset 1,* and one building with the most comprehensive household appliances was selected as the *case building*. The results show that, for the *case building*, the methodology was able to identify the behavior that needed to be modified, and to provide occupants with feasible recommendations so they could make required decisions. For example, it was found that, the usage of TV (in the master bedroom in the second floor) would quite possibly lead to the usage of lamp in the second floor. This may have occurred since the building occupants always turned the lights on when they were watching TV. An effective way of reducing energy consumption in this building is to watch TV with dim light.

The results could help building occupants modify their behavior, thereby significantly reducing building energy consumption. Moreover, given that the proposed method is partly based on comparison with similar buildings, it could motivate building occupants to modify their behavior.

8 Summary and concluding remarks

Vast amounts of building-related data are measured and collected. The data can provide abundant useful knowledge about the interactions between building energy consumption and its influencing factors. Such interactions play a crucial role in developing and implementing control strategies to improve building energy performance. It is highly desirable for hidden useful knowledge to be extracted from the data in order to gain a clear and thorough understanding of such interactions.

Commonly used data analysis methods for extracting useful knowledge from building-related data are summarized and evaluated. Our comprehensive review indicates that three general categories of data analysis techniques were used: typical indicator method, statistical analysis method, and building simulation method. Both the strengths and the weaknesses of these methods are addressed. Considering the increased size of building historical databases and the diversity of the influencing factors, these commonly used data analysis methods are insufficient to take full advantage of the data and extract useful information about the interactions and to help improve building energy performance.

In this study, data mining technique (classification analysis, cluster analysis, and association rule mining) is proposed to extract useful facts from the data. Moreover, a data analysis process and a data mining framework are proposed, enabling building-related data to be analyzed more efficiently. The applications of the process and framework to two sets of collected data demonstrate their applicability and usefulness. Accordingly, four data analysis methodologies were developed and applied to the collected data:

Classification analysis was applied to develop a methodology for establishing building energy-demand predictive models. The results demonstrate that the methodology can generate interpretable building energy-demand models that can help us understand the influence of the seven influencing factors on total building energy consumption.

Cluster analysis was used to develop a methodology for examining the influences of occupant behavior on building energy consumption. The results show that the methodology can quantitatively identify the effects of occupant behavior without including the impact of other influencing factors.

Association rule mining was employed to develop a methodology for examining all associations and correlations among building operational data, thereby discovering useful knowledge about energy conservation. The results show there are possibilities for saving energy by modifying the operation of mechanical ventilation systems and by repairing equipment.

Cluster analysis, classification analysis, and association rule mining were combined to formulate a methodology for identifying and improving occupant behavior in buildings. The results show that the methodology is able to identify the behavior that needs to be modified, and to provide

occupants with feasible recommendations so they can make required decisions to modify their behavior.

The proposed data analysis process and data mining framework provide an opportunity for standardizing the process of data mining in the field of building. By using them, researchers and designers can develop efficient data analysis methodologies and extract useful knowledge from monitored data. However, it should be mentioned that, while a lot of building-related data are sensory stream data, the current framework does not address the demand of real-time detection and response to (unexpected) events and incidents. To provide a real-time (or close to real-time) response, the current framework has to be extended to perform data mining operations on sensory stream data. The main focus of future research should be placed on applying the proposed process and framework to various building sectors, climates, and building automation systems in order to further evaluate their effectiveness and to help account for the interactions between building energy consumption and its influencing factors.

Acknowledgements

The authors would like to express their gratitude to the Public Works and Government Services Canada, and Concordia University for the financial support.

References

Abu Hamdeh NH, Al-Muhtaseb MTA (2010). Optimization of solar adsorption refrigeration system using experimental and statistical techniques. *Energy Conversion and Management*, 51: 1610 – 1615.

Al-Mumin A, Khattab O, Sridhar G (2003). Occupants' behavior and activity patterns influencing the energy consumption in the Kuwaiti residences. *Energy and Buildings*, 35: 549 – 559.

Balaras CA, Dascalaki E, Gaglia A, Droutsa K (2003). Energy conservation potential, HVAC installations and operational issues in Hellenic airports. *Energy and Buildings*, 35: 1105 – 1120.

Balta MT, Dincer I, Hepbasli A (2010). Performance and sustainability assessment of energy options for building HVAC applications. *Energy and Buildings*, 42: 1320 – 1328.

Bi Y, Chen L, Sun F (2008). Heating load, heating-load density and COP optimizations of an endoreversible air heat-pump. *Applied Energy*, 85: 607 – 617.

Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A (1998). Discovering Data Mining: From Concept to Implementation. Upper Saddle River, USA: Prentice Hall.

Cao LB, Yu PS, Zhang CQ, Zhang HF (2009). Data Mining for Business Applications. New York: Springer.

Chekir N, Bellagi A (2011). Performance improvement of a butane/octane absorption chiller. *Energy*, 36: 6278 – 6284.

Chen S, Yoshino H, Levine MD, Li Z (2009a). Contrastive analyses on annual energy consumption characteristics and the influence mechanism between new and old residential buildings in Shanghai, China, by the statistical methods. *Energy and Buildings*, 41: 1347 – 1359.

Chen S, Yoshino H, Li N (2009b). Statistical analyses on summer energy consumption characteristics of residential buildings in some cities of China. *Energy and Buildings*, 42: 136 – 146.

Chung W, Hui YV (2009). A study of energy efficiency of private office buildings in Hong Kong. *Energy and Buildings*, 41: 696 – 701.

Cios KJ (2007). Data Mining: A Knowledge Discovery Approach. New York: Springer.

de la Flor FJS, Lissén JMS, Domínguez SÁ (2006). A new methodology towards determining building performance under modified outdoor conditions. *Building and Environment*, 41: 1231 – 1238.

Delgado M, Sánchez D, MartIn-Bautista MJ, Vila M (2001). Mining association rules with improved semantics in medical databases. *Artificial Intelligence in Medicine*, 21: 241 – 245.

Deng SM, Burnett J (2000). A study of energy performance of hotel buildings in Hong Kong. *Energy and Buildings*, 31: 7 – 12.

Dong B, Cao C, Lee SE (2005a). Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37: 545 – 553.

Dong B, Lee SE, Sapar MH (2005b). A holistic utility bill analysis method for baselining whole commercial building energy consumption in Singapore. *Energy and Buildings*, 37: 167 – 174.

Ekici BB, Aksoy UT (2009). Prediction of building energy consumption by using artificial neural networks. *Advances in Engineering Software*, 40: 356 – 362.

Emery AF, Kippenhan CJ (2006). A long-term study of residential home heating consumption and the effect of occupant behavior on homes in the Pacific Northwest constructed according to improved thermal standards. *Energy*, 31: 677 – 693.

Escrivá-Escrivá G, Álvarez-Bel C, Roldán-Blay C, Alcázar-Ortega M (2011). New artificial neural network prediction method for electrical consumption forecasting based on building end-uses. *Energy and Buildings*, 43: 3112 – 3119.

Eskin N, Türkmen H (2008). Analysis of annual heating and cooling energy requirements for office buildings in different climates in Turkey. *Energy and Buildings*, 40: 763 – 773.

Freire RZ, Oliveira GHC, Mendes N (2008). Development of regression equations for predicting energy and hygrothermal performance of buildings. *Energy and Buildings*, 40: 810 – 820.

Gaitani N, Lehmann C, Santamouris M, Mihalakakou G, Patargias P (2010). Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy*, 87: 2079 – 2086.

Georgilakis PS, Gioulekas AT, Souflaris AT (2007). A decision tree method for the selection of winding material in power transformers. *Journal of Materials Processing Technology*, 181: 281 – 285.

Ghiaus C (2006). Experimental estimation of building energy performance by robust regression. *Energy and Buildings*, 38: 582 – 587.

Givoni B, Krüger EL (2003). An attempt to base prediction of indoor temperatures of occupied houses on their thermo-physical properties. In: Proceedings of the 18th International Passive and Low Energy Architecture Conference (PLEA'03), Santiago, Chile.

Han J, Kamber M (2006). Data Mining Concepts and Techniques, 2nd edn. San Francisco: Elsevier.

Hand D, Mannila H, Smyth P (2001). Principles of Data Mining. Cambridge, USA: MIT Press.

Hou ZJ, Lian ZW, Yao Y, Yuan XJ (2006). Cooling-load prediction by the combination of rough set theory and an artificial neural-network based on data-fusion technique. *Applied Energy*, 83: 1033 – 1046.

Hsu CH (2009). Data mining to improve industrial standards and enhance production and marketing: An empirical study in apparel industry. *Expert Systems with Applications*, 36: 4185 – 4191.

Jiao J, Zhang Y (2005). Product portfolio identification based on association rule mining. *Computer-Aided Design*, 37: 149 – 172.

Jiménez MJ, Heras MR (2005). Application of multi-output ARX models for estimation of the *U* and *g* values of building components in outdoor testing. *Solar Energy*, 79: 302 – 310.

Kim YS, Kim KS (2007). Simplified energy prediction method accounting for part-load performance of chiller. *Building and Environment*, 42: 507 – 515.

Krüger EL, Givoni B (2004). Predicting thermal performance in occupied dwellings. *Energy and Buildings*, 36: 301 – 307.

Kyrö R, Heinonen J, Säynäjoki A, Junnila S (2011). Occupants have little influence on the overall energyconsumption in district heated apartment buildings. *Energy and Buildings*, 43: 3484 – 3490.

Lam JC, Hui SCM, Chan ALS (1997). Regression analysis of high-rise fully air-conditioned office buildings. *Energy and Buildings*, 26: 189 – 197.

Lam JC, Wan KKW, Cheung KL (2009). An analysis of climatic influences on chiller plant electricity consumption. *Applied Energy*, 86: 933 – 940.

Lior R, Oded M (2008). Data Mining with Decision Trees: Theory and Applications. Singapore: World Scientific.

Schweiker M, Shukuya M (2010). Comparative effects of building envelope improvements and occupant behavioural changes on the exergy consumption for heating and cooling. *Energy Policy*: 2976 – 2986

Masoso OT, Grobler LJ (2010). The dark side of occupants' behaviour on building energy use. *Energy and Buildings*, 42: 173 – 177.

Murakami S, Akabayashi S, Inoue T, Yoshino H, Hasegawa K, Yuasa K, et al. (2006). Energy consumption for residential buildings in Japan. Architectural Institute of Japan, Maruzen Corp., Available online at http://tkkankyo.eng.niigata-u.ac.jp/HP/HP/database/index.htm.

Olofsson T, Andersson S (2001). Long-term energy demand predictions based on short-term measured data. *Energy and Buildings*, 33: 85 – 91.

Ourghi R, Al-Anzi A, Krarti M (2007). A simplified analysis method to predict the impact of shape on annual energy use for office buildings. *Energy Conversion and Management*, 48: 300 – 305.

Ouyang J, Hokao K (2009). Energy-saving potential by improving occupants' behavior in urban residential sector in Hangzhou City, China. *Energy and Buildings*, 41: 711 – 720.

Pan H, Li J, Zhang W (2007). Incorporating domain knowledge into medical image clustering. *Applied Mathematics and Computation*, 185: 844 – 856.

Pérez-Lombard L, Ortiz J, Coronel JF, Maestre IR (2011). A review of HVAC systems requirements in building energy regulations. *Energy and Buildings*, 43: 255 – 268.

Priyadarsini R, Xuchao W, Eang LS (2009). A study on energy performance of hotel buildings in Singapore. *Energy and Buildings*, 41: 1319 – 1324.

Li Q, Meng Q, Cai J, Yoshino H, Mochida A (2009). Applying support vector machine to predict hourly cooling load in the building. *Applied Energy*, 86: 2249 – 2256.

Quinlan JR (1986). Induction of decision trees. *Machine Learning*, 1: 81 – 106.

RapidMiner (2012): http://rapid-i.com/content/view/181/190/

Santamouris M, Mihalakakou G, Patargias P, Gaitani N, Sfakianaki K, Papaglastra M, Pavlou C, Doukas P, Primikiri E, Geros V, Assimakopoulos MN, Mitoula R, Zerefos S (2007). Using intelligent clustering techniques to classify the energy performance of school buildings. *Energy and Buildings*, 39: 45 – 51.

Santin G, Itard L, Visscher H (2009). The effect of occupancy and building characteristics on energy use for space and water heating in Dutch residential stock. *Energy and Buildings*, 41: 1223 – 1232.

Tonooka Y, Liu J, Kondou Y, Ning Y, Fukasawa O (2006). A survey on energy consumption in rural households in the fringes of Xian city. *Energy and Buildings*, 38: 1335 – 1342.

Tso GKF, Yau KKW (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32: 1761 – 1768.

Waltrich PJ, Barbosa Jr JR, Hermes CJL (2011). COP-based optimization of accelerated flow evaporators for household refrigeration applications. *Applied Thermal Engineering*, 31: 129 – 135.

Wang EY, Fung AS, Qi CY, Leong WH (2012). Performance prediction of a hybrid solar ground-source heat pump system. *Energy and Buildings*, 47: 600 – 611.

Wood CJ, Liu H, Riffat SB (2010). An investigation of the heat pump performance and ground temperature of a piled foundation heat exchanger system for a residential building. *Energy*, 35: 4932 – 4940.

Wu S, Clements-Croome D (2007). Understanding the indoor environment through mining sensory data—A case study. *Energy and Buildings*, 39: 1183 – 1191.

Yao Y, Lian ZW, Hou ZJ, Liu W (2006). An innovative air-conditioning load forecasting model based on RBF neural network and combined residual error correction. *International Journal of Refrigeration*, 29: 528 – 538.

Yu Z, Haghighat F, Fung BCM, Yoshino H (2010). A decision tree method for building energy demand modeling. *Energy and Buildings*, 42: 1637 – 1646.

Yu Z, Fung BCM, Haghighat F, Yoshino H, Morofsky E (2011a). A systematic procedure to study the influence of occupant behavior on building energy consumption. *Energy and Buildings*, 43: 1409 – 1417.

Yu Z, Haghighat F, Fung BCM, Yoshino H, Morofsky E (2011b). A methodology for identifying and improving occupant behavior in residential buildings. *Energy*, 36: 6596 – 6608.

Yu Z, Haghighat F, Fung BCM, Zhou L, Morofsky E (2012). A novel methodology for knowledge discovery through mining associations between building operational data. *Energy and Buildings*, 47: 430 – 440.

Yun GY, Kim H, Kim JT (2012). Effects of occupancy and lighting use patterns on lighting energy consumption. *Energy and Buildings*, 46: 152 – 158.

Zhang Q (2004). Residential energy consumption in China and its comparison with Japan, Canada, and USA. *Energy and Buildings*, 36: 1217 – 1225.