Energy and Buildings 47 (2012) 430–440

Contents lists available at SciVerse ScienceDirect

Energy and Buildings

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

A novel methodology for knowledge discovery through mining associations between building operational data

Zhun (Jerry) Yu [a], Fariborz Haghighat [a,*], Benjamin C.M. Fung [b], Liang Zhou [c]

[a] Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Quebec, H3G 1M8, Canada
[b] Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, H3G 1M8, Canada
[c] Institute for Research in Construction, National Research Council Canada, Ottawa, Ontario, K1A 0R6, Canada

ARTICLE INFO

Article history:
Received 3 October 2011
Received in revised form
10 December 2011
Accepted 13 December 2011

Keywords:
Data mining
Association
Correlation
Knowledge discovery
Building operational data
Energy conservation
Influencing factors

ABSTRACT

Nowadays, vast amounts of data on building operation and management have been collected and stored. However, the data is rarely translated into useful knowledge about building energy performance improvement, due mainly to its extreme complexity and a lack of effective data analysis techniques. This paper reports the development of a new methodology for examining all associations and correlations between building operational data, thereby discovering useful knowledge about energy conservation. The method is based on a basic data mining technique (association rule mining). To take full advantage of building operational data, both daily and annual time periods should be mined. Moreover, data from two different years should be mined, and the obtained associations and correlations in the two years should be compared. In order to demonstrate the applicability of the proposed method, the method was applied to the operational data of the air-conditioning system in a building located in Montreal. The results show energy waste in the air-conditioning system as well as equipment faults. A low/no cost strategy for saving energy in the system operation was also proposed. The results obtained could help to better understand building operation and provide opportunities for energy conservation.

1. Introduction

Energy consumption in the building sector contributes substantially to the global energy consumption and to the production of greenhouse gas emissions. Furthermore, building industry is not only energy-intensive, but also knowledge-intensive. Hence, it is highly desirable that useful information hidden in building operation be discovered to help reduce its energy consumption. An effective method to achieve this goal is to extract such information from the measured building-related data and translate this information into useful knowledge to be used in the daily building operation. Note that the real data of a building contains the actual information of building operation; and thus can reflect the building performance accurately. Moreover, vast amounts of data on building operation and management have been collected and stored, since building automation systems become a part of building design. In general, building-related data includes:

(1) Climatic data, such as outdoor air temperature and relative humidity
(2) Building operational data, mainly operational data of HVAC systems (e.g. supply air temperature and fresh air flow rates), IEQ data (e.g. indoor air temperature and human thermal comfort), and energy data (e.g. monthly electricity consumption and end-use loads of household appliances)
(3) Building physical parameters, such as floor area and window-to-wall-ratio

These data (climatic data, building operational data, building physical parameters) may have a direct/indirect influence on each other, considering they are closely related to the operation of that specific building. Specifically, there may exist strong associations (i.e. connections or relationships) and correlations between them that should be identified and used by the building managers to develop energy efficient building operation strategies. For example, the association/correlation between building energy consumption and climatic parameters (e.g. outdoor air temperature) generally reflects how building operation is affected by weather conditions. All these possible associations/correlations between building-related data need to be identified in order to develop an effective energy conserving strategy and operation. Note that the energy consumption of HVAC systems could account for over 25% of the total building energy consumption.

A number of studies have been conducted to identify the associations and correlations between measured building-related

* Corresponding author. Tel.: +1 514 848 2424x3192; fax: +1 514 848 7965.
E-mail address: haghi@bcee.concordia.ca (F. Haghighat).

0378-7788/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.enbuild.2011.12.018

Nomenclature

TA	air temperature (°C)
TG	glycol temperature (°C)
H	relative humidity (kg/kg)
Q	flow rate (L/S)
F	frequency of variable-speed drives on fans (Hz)

Subscripts and superscripts

I, II, III, IV, V	fresh air handling unit 1 (FHU1), FHU2, FHU3, FHU4, FHU5
VI, VII	return air handling unit 1 (RHU1), RHU2
VIII, IX	exhaust air handling unit 1 (EHU1), EHU2
ac	after cooling coil
ah	after heating coil
br	before recuperation
ar	after recuperation
1, 2, 3	fan1, fan2, fan3
i, ii, iii	recuperation1, recuperation2, recuperation3
o	outdoor
VA	VA part
ENCS	ENCS part

data. Traditionally, researchers utilized statistical analysis techniques, in particular regression equations, and mainly focused on the relationships between building energy consumption and the influencing factors, such as building physical parameters [1–3], occupancy patterns [4,5], building operation and management [6], social and economic factors [7], indoor air quality requirements [8], and weather conditions [9]. However, few of them examined associations and correlations between building operational data, especially operational data of HVAC systems, to better understand building operation in order to improve building performance. For example, the associations/correlations between operational data of different AHUs, such as air flow rates, or between operational data of the same AHU, such as air temperature after cooling coils and heating coils, are seldom analyzed. This is mainly due to the amount and complexity of such data and a lack of effective data analysis techniques. This is caused by a large number of HVAC system parameters and huge amounts of operational data. Moreover, poor quality of measured data (e.g. outliers and missing values) can also greatly add to the complexity.

The strength of statistical techniques lies in their simplicity and widespread familiarity. However, most statistical techniques, especially correlation analysis, are utilized with the premise that data analysts, based on their domain knowledge, "believe" that strong associations and correlations exist among two or more parameters. For example, one performs correlation analyses between building energy consumption and outdoor air temperature since s/he "believes" that outdoor air temperature may have a significant influence on the building energy consumption. The analyses mainly depend on the prior knowledge of the analyst and adopted statistical methods. As a result, a lot of useful information, particularly indirect influences between the data could be missed (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). At the same time, commonly a large number of parameters are monitored and huge amounts of operational data from the HVAC system are collected. Consequently, it is very difficult and often infeasible for data analysts to conduct statistical analyses, say the correlation analyses, on every combination of the parameters so as 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, however, impractical in practice. This paper reports the development of a methodology for examining *all* these associations and correlations between the building operational data in order to achieve a better understanding of the building operation and to provide opportunities to conserve energy.

2. Methodology

2.1. Data mining

Considering the limitations of the statistical analysis techniques, we propose data mining to analyze measured building-related data. Data mining techniques lead the way to automatically analyzing huge amounts of data. They can be used to extract interesting, useful, and previously unknown knowledge from data, and therefore fit well into the purpose of this study.

In the past decade, different definitions of data mining have been given by various researchers. For example, Hand et al. [10] define data mining as "the analysis of large observational data sets 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 [11], 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 multi-disciplinary approaches. It is often used to extract "interesting," hidden, but useful patterns from a large volume of data and to transform the data into knowledge that could benefit further work.

A basic data mining technique, association rule mining, provides a feasible solution to identify all interesting relations between data values even for large datasets. Therefore, in this study it was employed to help examine all the associations and correlations between the building operational data.

2.2. Association rule mining

In data mining, association rules are often used to represent the patterns of parameters that are frequently associated together. An example is given to illustrate the concept of association rules. Assume that 100 occupants live in 100 different rooms in the same building and each room has both a window and a door. Moreover, 40 occupants open the windows and 20 occupants open the doors. If 10 occupants open both the windows and doors 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 open windows. Then, the information that occupants who open windows also tend to open doors at the same time can be represented in the following association rule:

$$\text{open\_windows} \rightarrow \text{open\_doors}[\text{support} = 10\% \quad \text{confidence} = 25\%]$$

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 eventually. Accordingly, the association rule mining (ARM) can be defined as finding out association rules that satisfy the predefined minimum support and confidence from a given database.

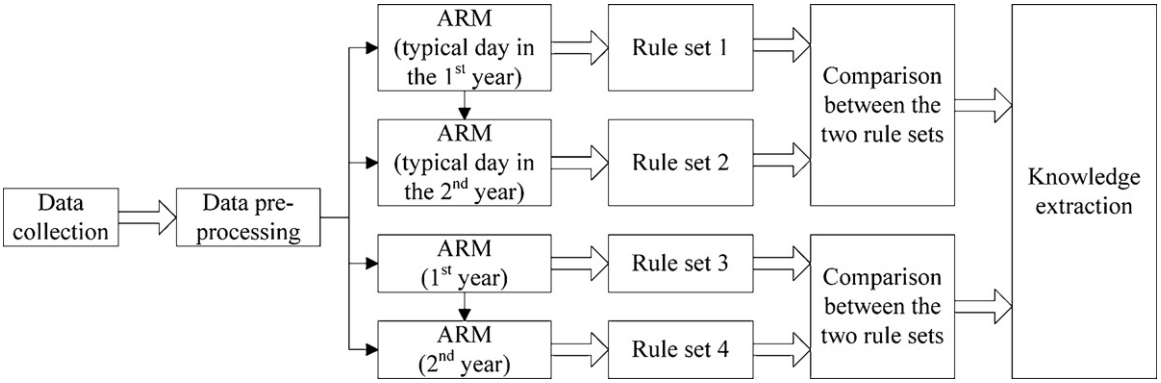

Fig. 1. Proposed methodology to examine all the associations and correlations between building operational data.

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

$$\text{support}(X \rightarrow Y) = P(X \cup Y)$$

$$\text{confidence}(X \rightarrow Y) = P(Y|X)$$

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,

$$\text{lift}(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X)P(Y)} = \frac{P(Y|X)}{P(Y)}$$

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 [12]) while a lift value less than 1 represents a negative correlation. If the value is equal to 1, i.e. $P(X \cup Y) = P(X)P(Y)$, the occurrence of X is independent of the occurrence of Y, and there is no correlation between X and Y.

Commonly used ARM algorithms include the Apriori algorithm and the frequent-pattern growth (FP-growth) algorithm [13]. In this study, we employed the FP-growth algorithm, along with the open-source data mining software RapidMiner [14], to mine association rules due to its high efficiency and wide applicability. 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. For the specific algorithm of the FP-growth the reader can refer to [13].

Additionally, in order to perform the ARM, the value of quantitative attributes generally needs to be classified into categorical values. Given that building operational data, such as supply air temperature and monthly energy consumption, is normally described as either high or low by occupants in practice, a two-interval scale, i.e. HIGH and LOW, was applied in this study. Specifically, for each quantitative attribute, data ranged from the average of the maximum and minimum to the maximum value is 'HIGH', and data ranged from the minimum value to the average of the maximum and minimum is 'LOW'.

2.3. Proposed methodology

A new methodology is proposed for examining all associations and correlations between building operational data and leading to knowledge discovery. The methodology is based on a basic data mining technique: association rule mining (ARM). In order to find and take advantage of more complete associations and correlations, building operational data in two different time periods (i.e. both a day and a year) is mined, considering associations/correlations between operational data in different time periods could significantly be different. Moreover, data in two different years is mined, and obtained associations and correlations in the two years are compared between each other. The comparison can assist in identifying marked changes in associations/correlations and also building operation, thereby uncovering useful knowledge. The proposed methodology is given in Fig. 1, and it can be divided into 8 steps and is explained as follows:

Step 1 Data collection. Two-year building operational data need to be collected and stored in a database.
Step 2 Data pre-processing. Measured data is often noisy (especially containing outlier values whose values are grossly different, i.e. much higher or lower, from others in databases), which will lead to low-quality mining results. Hence, the collected data should be processed to remove outliers.
Step 3 Perform the ARM in a typical day (e.g. the coldest or hottest day) data in the 1st year. Obtained rules are stored in rule set 1.
Step 4 Select parameters having associations in the typical day data in the 1st year; and perform the ARM in the typical day data in the 2nd year within the selected parameters, in order to remove time effects and reduce other influences, such as the change of occupant behavior and weather conditions. Obtained rules are stored in rule set 2.
Step 5 Perform the ARM in the 1st year data. Obtained rules are stored in rule set 3.
Step 6 Select parameters having associations in the 1st year data; and perform the ARM in the 2nd year data within the selected parameters. Obtained rules are stored in rule set 4.
Step 7 Compare the rule sets 1 and 2, and the rule sets 3 and 4; and highlight the similarity and difference in associations between the two different time periods (i.e. the typical day in the1st year and 2nd year, the 1st year and the 2nd year).
Step 8 Extract useful knowledge from the comparison between these rules.

3. Data collection

The EV pavilion located in Montreal, a complex building that mainly includes offices and chemical labs, was selected as data source in this study. This building consists of two parts: the ENCS part (17 floors) and the VA part (12 floors), as shown in Fig. 2.

Both of these two buildings have their own VAV air-conditioning systems. In the ENCS part, the air handling units (AHUs) are installed in the mechanical rooms on each floor except for the 17th floor (the mechanical floor), where various equipment, such

Fig. 2. EV Pavilion at Concordia University.

as the chillers and fresh air handling units (FHUs), are installed. On the 17th floor, two identical FHUs (i.e. the FHU 1 and FHU 2) are employed to process fresh air and each has two variable speed fans in parallel, as shown in Fig. 3. Due to the existence of chemical labs in the ENCS part, the fresh air is separated into two parts: part 1 is sent to the local mechanical rooms in each floor and mixed with the return air from that floor's rooms other than chemical labs. Then the mixed air is conditioned by the AHUs in that floor's mechanical room and supplied to those rooms again. Meanwhile, part 2 is mixed with the return air from the atriums in the ENCS part. Then the mixed air is conditioned by the FHU 3, which also has two variable speed fans in parallel, and sent to the chemical labs. The exhaust air from both the chemical labs and other rooms is discharged outside directly by the EHU 1, which contains two variable speed fans, as shown in the dash line square. Moreover, the dash dot line in Fig. 3 indicates a recuperation loop installed between the fresh air and the exhaust air to exchange heat in both cooling and heating seasons.

The flowchart of air-conditioning system in the VA part is shown in Fig. 4. Similarly, air handling units (AHUs) are installed in the mechanical rooms on each floor except for the 12th floor (the mechanical floor), where various equipment, such as the chillers and fresh air handling units, are installed. On the 12th floor, two identical FHUs (i.e. the FHU 4 and FHU 5) are employed to process fresh air and each of them has two variable speed fans in parallel. Given that there is no chemical lab in the VA part, the fresh air is mixed with the return air from all the VA part directly. The mixed air is sent to the local mechanical rooms in each floor to be conditioned by the AHUs, and then sent to various rooms in the same floor. Two RHUs (i.e. the RHU 1 and RHU 2) are employed to return air, and each of them has two variable speed fans in parallel. The exhaust air in the VA part is discharged to outside by the EHU 2, which contains three variable speed fans, as shown in the dash line square. Also, the dash dot line in Fig. 4 indicates a recuperation loop installed between the fresh air and exhaust air to exchange heat in both cooling and heating seasons.

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 in the two air-conditioning systems and data of each parameter was trended at a 15-min interval. The monitored parameters are given in Table 1.

Table 1
The monitored parameters of the air-conditioning systems.

No.	Parameter	No.	Parameter	No.	Parameter	No.	Parameter
1	Q_{I1}	21	TA_{IVac}	41	TA_{IXari}	61	F_{IX3}
2	Q_{I2}	22	TA_{Vac}	42	TA_{IXarii}		
3	Q_{II1}	23	TA_{Iah}	43	$TA_{IXariii}$		
4	Q_{II2}	24	TA_{IIah}	44	TA_{oENCS}		
5	Q_{III1}	25	TA_{IVah}	45	H_{oENCS}		
6	Q_{III2}	26	TA_{Vah}	46	TA_{oVA}		
7	Q_{IV1}	27	TA_{IIbr}	47	H_{oVA}		
8	Q_{IV2}	28	TA_{IVbr}	48	TG_{ENCSar}		
9	Q_{V1}	29	TA_{Vbr}	49	TG_{VAar}		
10	Q_{V2}	30	TA_{Iar}	50	F_{I}		
11	Q_{III}	31	TA_{IIar}	51	F_{II}		
12	Q_{VI}	32	TA_{IVar}	52	F_{III}		
13	Q_{VII}	33	TA_{Var}	53	F_{IV}		
14	Q_{VIII1}	34	$TA_{VIIIbri}$	54	F_{V}		
15	Q_{VIII2}	35	$TA_{VIIIbrii}$	55	F_{VI}		
16	Q_{VIII3}	36	TA_{IXbri}	56	F_{VII}		
17	Q_{IX1}	37	TA_{IXbrii}	57	F_{VIII1}		
18	Q_{IX2}	38	$TA_{IXbriii}$	58	F_{VIII2}		
19	TA_{Iac}	39	$TA_{VIIIari}$	59	F_{IX1}		
20	TA_{IIac}	40	$TA_{VIIIarii}$	60	F_{IX2}		

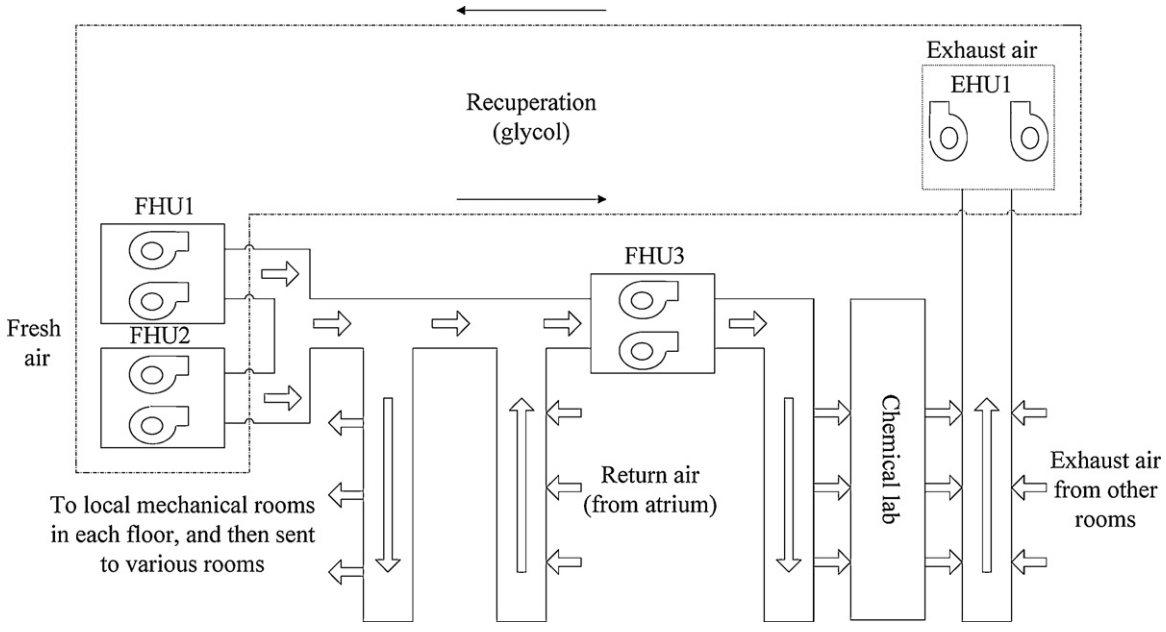

Fig. 3. Flow chart of air-conditioning system in the ENCS part.

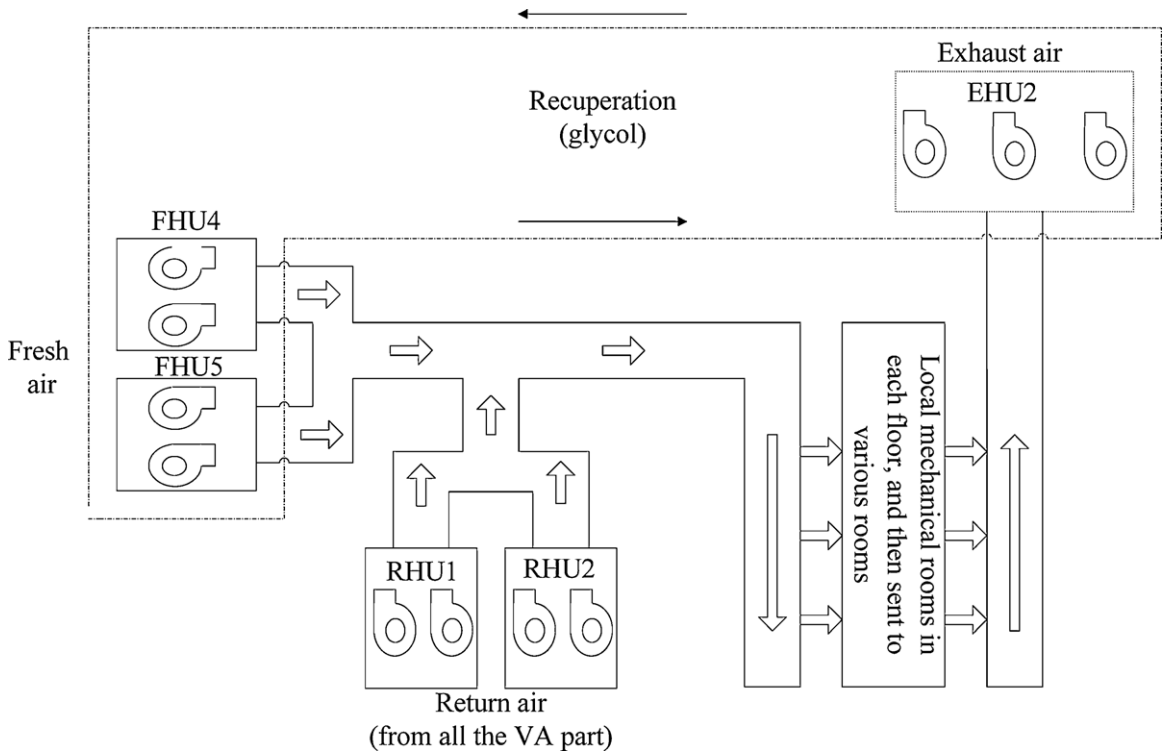

Fig. 4. Flow chart of air-conditioning system in the VA part.

4. Data pre-processing

Outliers are data objects whose values are grossly different (i.e. much higher or lower) from others in the database. Outliers regularly occur in building energy consumption measurement and they are often indicative of measurement errors, and thus must be removed. Removal of outliers plays a crucial role in preparing for the ARM, since the outliers will skew and thus alter the grouping of data. For example, suppose an attribute ranges from 0 to 10, and can be discretized into two intervals, [0,5) and [5,10] (or LOW and HIGH), by using the methods mentioned previously. If there exists an outlier (e.g. 30), then the two intervals are [0,15) and [15,30] (or LOW and HIGH) by using the same method. Accordingly, all the data are defined as LOW except the outlier, which is not actually true.

Various methods can be used for effective detection and removal of the outliers. In this study, a method based on the lower quartile (Q_1) and the upper quartile (Q_3) of the standard boxplot was used due to its simplicity [15]. Specifically, outlying values can be distinguished using the following two rules:

Rule 1: Data values that are less than $Q_1 - 1.5 \times (Q_3 - Q_1)$ are defined as outliers
Rule 2: Data values that are larger than $Q_3 + 1.5 \times (Q_3 - Q_1)$ are defined as outliers

With consideration of the seasonality of building energy consumption, the ARM was performed based on the seasonal data instead of the annual data in this study (refer to steps 5 and 6 in Section 2.3). Given that the EV building is located in Montreal which has cold winters, the winter data in both 2007 and 2009 was mined to generate association rules (as mentioned earlier, the winter data in 2008 was unavailable). Furthermore, only the data in working days/hours were used when mining seasonal data, considering that building energy consumption is significantly different between working days/hours and non-working days/hours due to occupant behavior (for the EV building, non-working days include weekends and holidays; and working hours are from 8 AM to 5 PM). The resulting data in 2007 and 2009 were stored in dataset_1 and dataset_2, respectively. Fig. 5 shows the distribution of two intervals of the entire ARM attributes in the dataset_1after the removal of outliers and discretization. Note that the numbers in the abscissa represent the ARM attributes, and correspond to the numbers in Table 1. Clearly, it can be observed that most of the percentages range from 30% to 70%, indicating roughly a uniform distribution.

5. Results and discussion

5.1. ARM on the coldest day in the dataset_1 and dataset_2

The initial rule mining was carried out with the dataset_1 and dataset_2 on the coldest day in both 2007 and 2009. After experimenting with various combinations of support and confidence values, a support of 80% and a confidence of 95% were set as minimum thresholds. The thresholds mean that, for each generated association rule, at least 80% of all the data records under analysis contain both premise and conclusion; and the probability that a premise's emergence leads to a conclusion's occurrence is 95% or more. In addition, the minimum threshold of lift value was set 1 to find positive correlations. The mining in the dataset_1 generated 476 rules (i.e. the rule set 1) and 43 parameters were involved. Then, the association rules were mined in the dataset_2 and only the data records of these 43 parameters were used. Such mining generated 169 rules (i.e. the rule set 2). Among the generated rules, many of them are obvious and uninteresting; and truly interesting rules need to be further identified based on domain knowledge. Also, the two rule sets (i.e. the rule sets 1 and 2) were compared with each other. As a result, three potentially useful association rules were found and they are given in Table 2.

Clearly, the premise and conclusion of the first two rules are reversed, and thus shows that the following four facts frequently occurred at the same time in winter 2007:

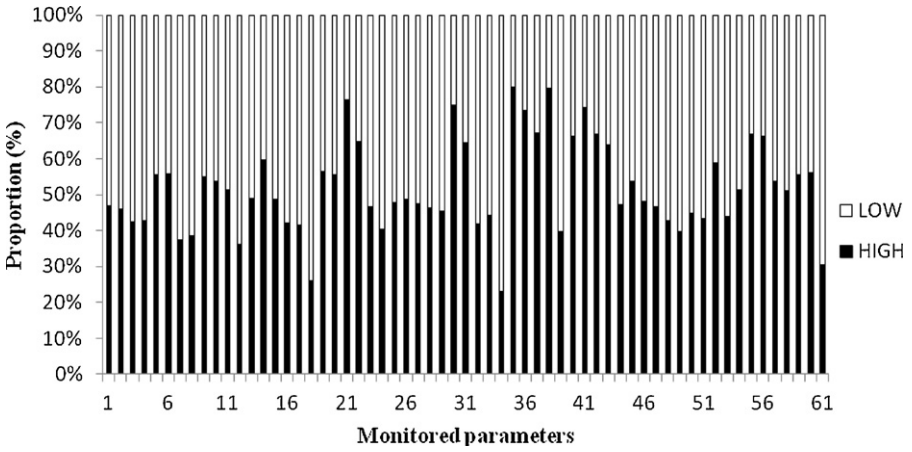

Fig. 5. Distribution of two intervals of all monitored parameters in the dataset_1.

Table 2
Three best rules generated.

No.	Premise	Conclusion	Sup	Conf	Lift	Dataset
Rule 1	TA_{IVah} [high] TA_{IVac} [low]	F_{IV} [high], TA_{Vac} [low]	0.81	0.99	1.21	1
Rule 2	F_{IV} [high] TA_{Vac} [low]	TA_{IVah} [high] TA_{IVac} [low]	0.81	0.99	1.21	1
Rule 3	TAI_{Vac} [low]	TA_{IVah} [high]	0.78	1.00	1.12	2

(1) The fresh air temperature after the heating coil in the FHU 4 was 'HIGH'
(2) The fresh air temperature after the cooling coil in the FHU 4 was 'LOW'
(3) The fresh air fan frequency of the FHU 4 in the VA side was 'HIGH'
(4) The fresh air temperature after the cooling coil in the FHU 5 was 'LOW'

Also, *Rule 3* shows that the following two facts frequently occurred at the same time in winter 2009:

(5) The fresh air temperature after the cooling coil in the FHU 4 was 'LOW'
(6) The fresh air temperature after the heating coil in the FHU 4 was 'HIGH'

Based on facts 1, 2, 5, and 6, it was observed that, in winter, the fresh air temperature in the FHU 4 usually increased first and then significantly decreased, which indicates a possible waste of energy. In order to illustrate this observation clearly, the screenshot of the FHU 4 control panel is shown in Fig. 6. In this diagram, the components in △, □, ◯, ▽ are the heat recovery (recuperation), heating coil, humidifier and cooling coil system, respectively.

The heating coil system was always on while the cooling coil system was always shut down in winter[1]. Hence, after the heating coil system, the temperature of fresh air drops only because of the humidification system that uses municipal water at about 2 °C. Site visit confirmed that this water was drained directly to sewage after humidification process. The heating and humidifying process is plotted in Fig. 7 (left).

As seen in the left diagram of Fig. 7, outdoor air is at state point A. Process A–B represents sensible pre-heating and heat recovery,

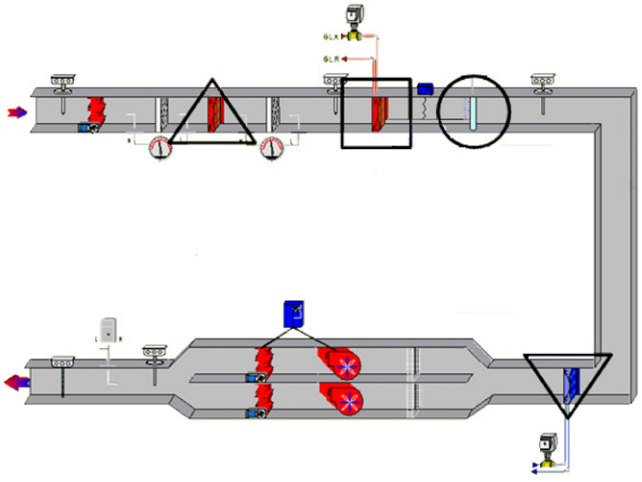

Fig. 6. Screenshot of the FHU 4 control panel.

which can be characterized by a horizontal line. After this, the heating and humidification processes are carried out successively, shown as processes B–C and C–D. Based on the monitored data, the actual air temperature after the heating coil system (point C) and the air temperature after the humidification system (point D) are plotted in Fig. 8.

Fig. 8 indicates that the air temperature after the heating coil system is around 14 °C higher than that after the humidification system. Clearly it is the low temperature of municipal water that caused the dramatic temperature drop (from state C to state D) in the conditioned fresh air, and such temperature drop can lead to a significant energy waste. That means the heat added to the fresh air during A–B and B–C processes is simply discharged with exhaust municipal water after the humidification process.

One possible remedy for such an issue would be decreasing the air temperature after the heating coil system. More specifically,

[1] Information provided by the building operators.

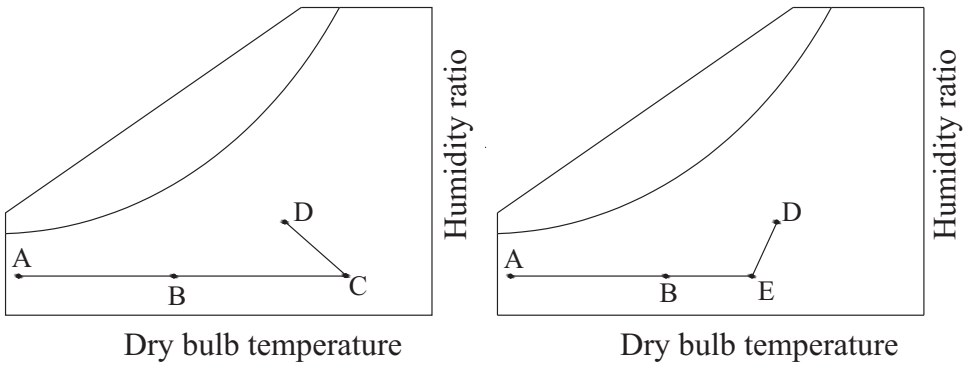

Fig. 7. Heating and humidification processes in psychrometric chart.

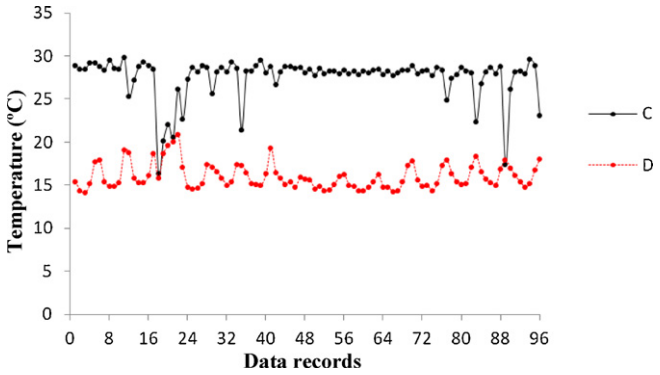

Fig. 8. Air temperature after heating coil (state C) and humidifier (state D).

shift point C to the left (to point E), as shown in the right diagram of Fig. 7. Correspondingly, one possible method in reality could be recycling and reusing (instead of discharging) the municipal water after it is warmed up after passing through the humidifier. In order to describe this process clearly, based on the monitored data and heat transfer theory, two schemas of hypothetical air/water temperature in the FHU 4 in winter before and after the remedy are given in Figs. 9 and 10.

In Fig. 9, the outdoor air temperature, air temperature after the heat recovery, air temperature after the heating coil, and air temperature after the humidifier are assumed to be −9 °C, 3 °C, 28 °C and 15 °C, respectively. At the same time, municipal water before and after the humidifier are assumed to be 2 °C and 12 °C.

In Fig. 10, the recycled high temperature municipal water (at 15 °C) and fresh municipal water (at 2 °C) could be mixed and then supplied to the humidifier again, considering the water loss during humidifying. The temperature of the mixed water is assumed to be at 8 °C and the water left the humidifier at 12 °C (or even higher). With this method, it would be enough to heat the fresh air up to a lower temperature (e.g. 21 °C as shown in Fig. 10) instead of 28 °C in the heating coil. Accordingly, a huge amount of energy can be saved in the heating coil. However, it should be mentioned that it would be necessary to treat the water before it is reused[2] to prevent microbial issues.

5.2. ARM in winter in the dataset_1 and dataset_2

Association rule mining was also carried out in winter for the dataset_1 and dataset_2. After experimenting with various combinations of support and confidence values, a support of 50% and a confidence of 80% were set as minimum thresholds. In addition, the minimum threshold of lift value was set 1 to find positive correlations. Specifically, association rules were first mined in the dataset_1. Such mining generated 461 rules (i.e. the rule set 3), and 32 parameters were involved in these rules. Then, association rules were mined in the dataset_2 and only the data records of these 32 parameters were used. Such mining generated 262 rules (i.e. the rule set 4). After that, the two sets of generated rules were compared with each other to further identify truly interesting rules. As a result, the obtained interesting rules were grouped into three categories in order to discover useful knowledge, as follows:

Category 1: same rules generated in the both datasets

From *Rules 1* and *2*, it can be observed that, the air flow rates of fan 1 in the FHU 1 and FHU 2 have a strong association and correlation. At the same time, *Rules 3* and *4* show that the air flow rates of fan 2 in the FHU 1 and FHU 2 also have a strong association and correlation (this is reasonable since the two fans in the same FHU are identical and controlled by one VSD). Therefore, it can be inferred that the total air flow rates of the FHU 1 and FHU 2 are strongly associated and correlated (Table 3).

The air flow rates of the FHUs 1 and 2 in both dataset_1 and dataset_2 are plotted in Fig. 11. It can be seen that the variation of air flow rates of these two FHUs follows the same trend. Furthermore, the values of air flow rates between these two FHUs are close to each other in both datasets. This indicates that the total air flow rates of the FHU 1 and FHU 2 are always strongly associated and correlated. Accordingly, if a continuous significant difference between them is observed, it can be inferred that either of the FHUs could have a fault. Therefore, the rules can help to understand FHU operation and also be applied to online fault detection.

Category 2: similar rules generated in both the datasets but are opposite in premise/conclusion

Six potentially useful rules in Category 2 are found and given in Table 4. *Rules 1* and *2* show that between these two years, the air flow rates of fan 1 in the FHU 4 and FHU 5 have opposite associations and correlations. Similarly, *Rules 3* and *4* can also be explained.

In order to provide an insight into the association opposition, the air flow rates of fan 1 in the FHUs 4 and 5 in these two years are plotted in Figs. 12 and 13, respectively. Considering that fan 1

Table 3
Four rules in Category 1.

No.	Premise	Conclusion	Sup	Conf	Lift	Dataset
Rule 1	Q_{I1} [low]	Q_{II1} [low]	0.52	0.98	1.70	1
Rule 2	Q_{I1} [low]	Q_{II1} [low]	0.55	1.00	1.63	2
Rule 3	Q_{I2} [low]	Q_{II2} [low]	0.52	0.97	1.70	1
Rule 4	Q_{I2} [low]	Q_{II2} [low]	0.57	0.95	1.65	2

[2] Through discussion with the building operators, this energy waste was confirmed and they planned to fix this problem using an appropriate method.

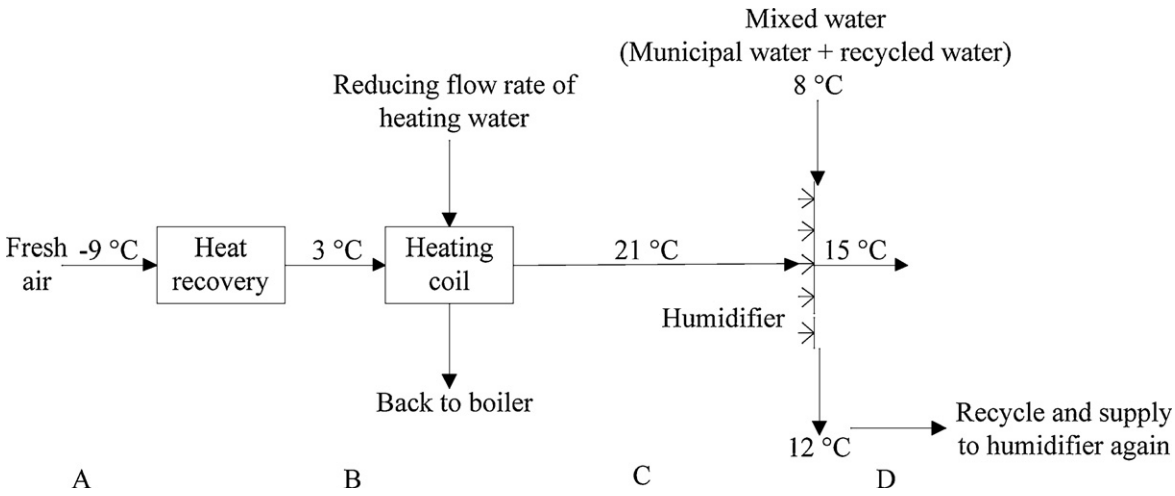

Fig. 10. Hypothetical air/water temperature in the FHU 4 after the remedy.

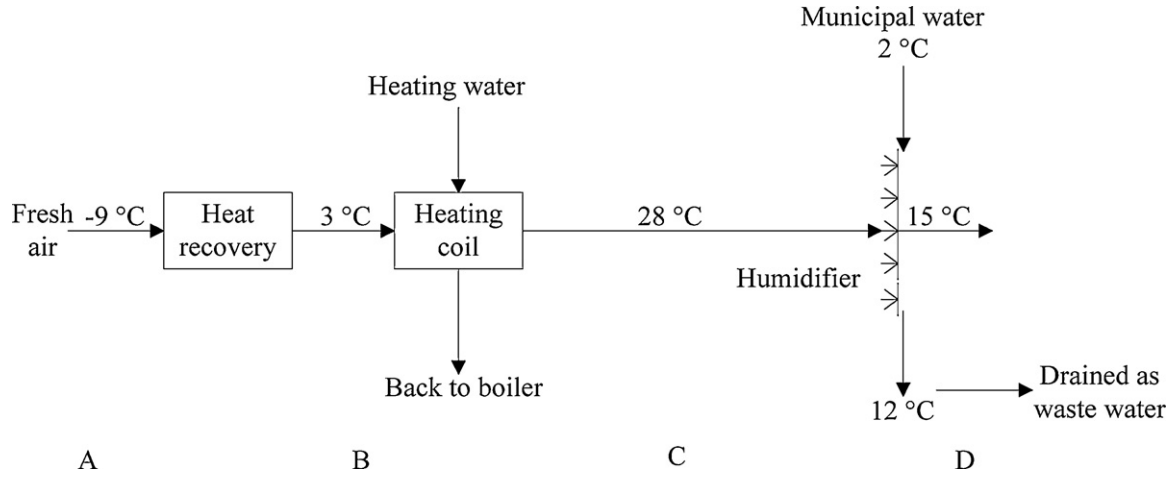

Fig. 9. Hypothetical air/water temperature in the FHU 4 before the remedy.

Table 4
Six rules in Category 2.

No.	Premise	Conclusion	Sup	Conf	Lift	Dataset
Rule 1	Q_{V1} [low]	Q_{IV1} [low]	0.59	0.92	1.49	1
Rule 2	Q_{V1} [high]	Q_{IV1} [low]	0.51	0.81	1.31	2
Rule 3	Q_{V2} [low]	Q_{IV2} [low]	0.57	0.91	1.50	1
Rule 4	Q_{V2} [high]	Q_{IV2} [low]	0.54	0.99	1.31	2
Rule 5	Q_{IX3} [low]	TA_{IXbri} [high]	0.60	0.82	1.12	1
Rule 6	Q_{IX3} [high]	TA_{IXbri} [high]	0.52	0.90	1.51	2

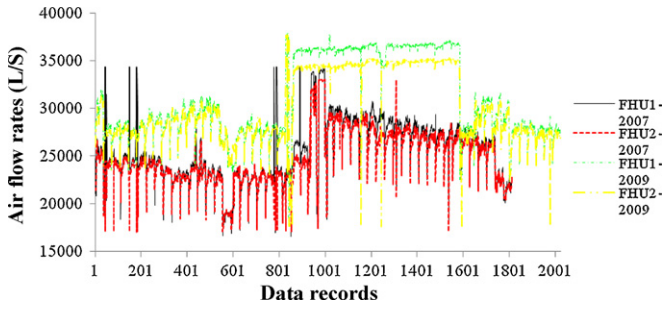

Fig. 11. Air flow rates of the FHUs 1 and 2 in the dataset_1 and dataset_2.

and fan 2 in the same FHU are identical and controlled by the same VSD, their air flow rates are approximately the same, and thus only the air flow rate of the fan 1 is plotted.

Clearly, Fig. 12 shows that the values of air flow rates of fan 1 in these two FHUs are very close in 2007. This is reasonable since

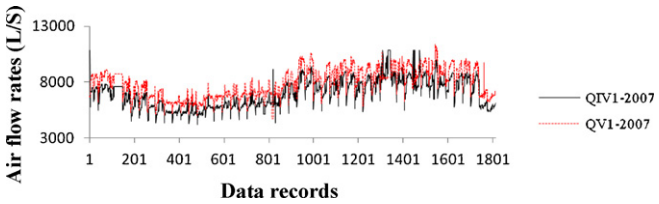

Fig. 12. Air flow rates of fan 1 in the FHUs 4 and 5 in dataset_1.

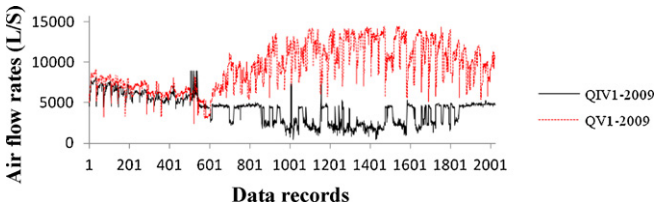

Fig. 13. Air flow rates of fan 1 in the FHUs 4 and 5 in dataset_2.

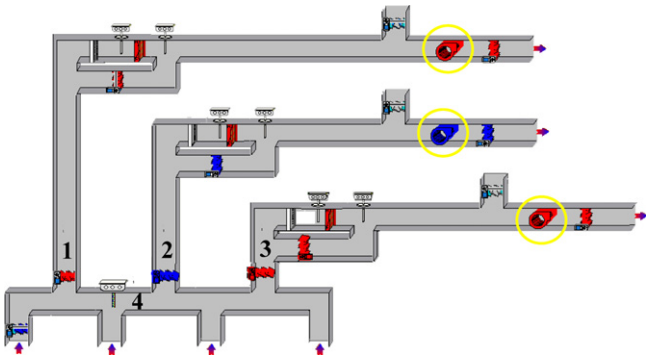

Fig. 14. Screenshot of the EHU 2 control panel.

these two FHUs are identical, and clearly their air flow rates should always be almost the same. However, Fig. 13 shows that, in 2009, the air flow rates of fan1 in the FHU 5 are much larger than that in the FHU 4 most of the time. Accordingly, it can be inferred that a fan fault occurred in the FHU 4 in 2009. Therefore, the rules can be used as a guide of fault diagnosis on the fans and FHUs.

Based on *Rules 5* and *6*, it can be found that these two rules' premises (i.e. the air flow rate of fan 3 in the EHU 2) are opposite. Fig. 14 shows the screenshot of the EHU 2 control panel. Clearly, the three exhaust air ducts (refer to 1, 2 and 3 in this diagram) are connected to each other. Furthermore, exhaust air from different parts of the VA part will be mixed in duct 4 before being distributed to the three fans (refer to three yellow circles in this diagram). A further analysis of operational data on these three fans in both years shows that two of them were always turned on to extract exhaust air while the other one was turned off. Moreover, two different control strategies were implemented in the two different years respectively: in 2007, the fans 1, 2, and 3 were turned off alternatively; in 2009, the fan 2 was always turned off while the fans 1 and 3 were always turned on. However, from the point of view of energy consumption, there is no difference between these two strategies, and it is highly desirable that a new control strategy can be proposed to save energy. Given that these three fans are identical and controlled by individual VSD, one possible energy-saving method is to use all these three fans instead of two of them. A comparison between the current and proposed strategy is made to show the energy conservation. For current strategy, assume the actual air flow rate of each fan is M, the actual fan speed is V, and the actual power required by each fan is P. Table 5 shows the results of comparison between the two strategies.

From Table 5, it is obvious that $(2P - 8P/9) = 10P/9$ can be saved if the proposed strategy is used. However, before this strategy is adopted in practice, it should be checked whether the fans will operate in the range of high efficiency, but not the dangerous unstable (surge) region at low air flow rates.Category 3: rules generated in only one dataset (either dataset_1 or dataset_2)

One potentially useful rule in Category 3 was found and given in Table 6. The rule shows that the fan frequency in the RHU 1 and RHU 2 has a strong association and correlation. The frequency of the two fans is plotted in Fig. 15, and it can be seen that F_{VI} is almost equal to F_{VII} all the time. Given that the RHU 1 and RHU 2

Table 6
One rule in Category 3.

No.	Premise	Conclusion	Sup	Conf	Lift	Dataset
1	F_{VI} [high]	F_{VII} [high]	0.60	0.97	1.60	1

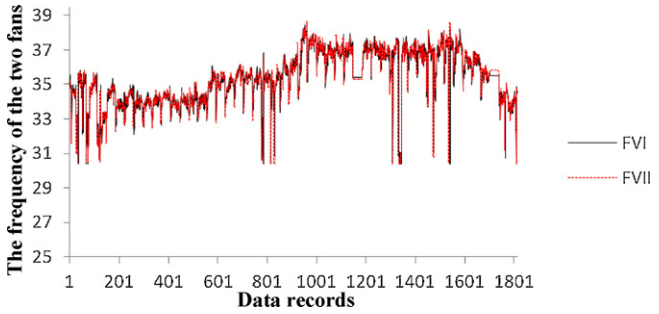

Fig. 15. Frequency of VSD on the fan in the RHU1 and RHU2 in dataset_1.

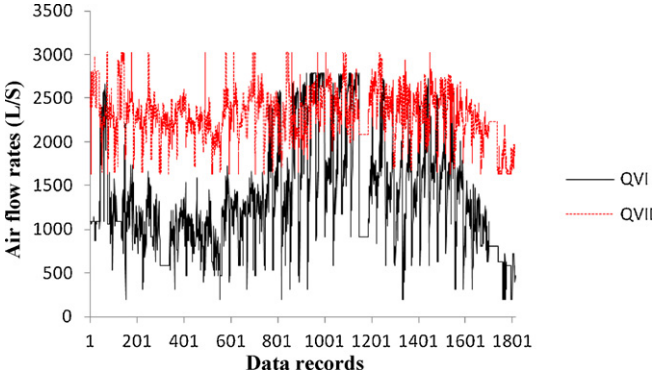

Fig. 16. Air flow rates of the fan in the RHUs 1 and 2 in dataset_1.

are identical, it can be inferred that these two RHUs' air flow rates (i.e. Q_{VI} and Q_{VII}) should be approximately identical. Accordingly, there should exist a strong association and correlation between Q_{VI} and Q_{VII}. However, no rule between Q_{VI} and Q_{VII} has been found in both dataset_1 and dataset_2. For this reason, air flow rates of the fan in the RHUs 1 and 2 in the dataset_1 are plotted in Fig. 16. Clearly, a significant difference can be found between Q_{VI} and Q_{VII}, which indicates that either RHU 1 or RHU 2 has a fault. Further, data shows that the RHU 1 did not operate in 2009 (Q_{VI} is zero in the dataset_2). Therefore, it can be concluded that the RHU 1 has a fault.

5.3. Association map

Besides association rules in the form of text, RapidMiner also provides a graphical view of an association map, representing all generated association rules. For simplicity, the association map in the dataset_2 instead of the dataset_1 is given in Fig. 17, considering that only the parameters showing up in both the rule set 3 and the rule set 4 are involved.

In this map, each line represents one association rule, and thus the amount of lines quantitatively indicates the amount of

Table 5
Comparison between the two control strategies.

Strategy	Number of fans used	Air flow rate of each fan	Total air flow rate	Fan speed	Power required by each fan	Total power required
Current	2	M	2 M	V	P	2P
Proposed	3	2 M/3	2 M	2V/3[a]	8P/27[b]	8P/9

[a] According to the fan laws, the capacity is directly proportional to the fan speed.
[b] According to the fan laws, the power required is proportional to the cube of fan speed.

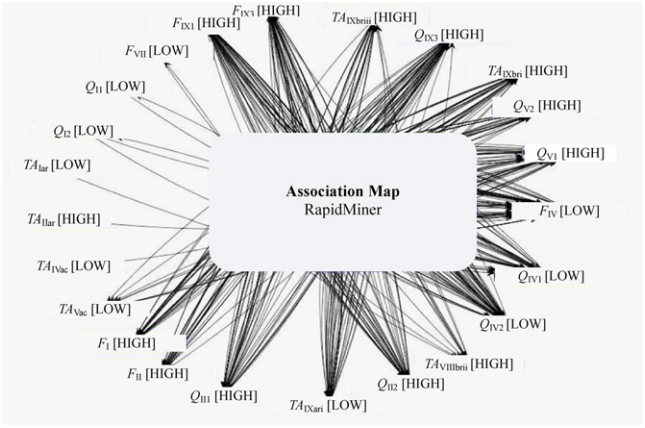

Fig. 17. Association map in the dataset_2 provided by RapidMiner.

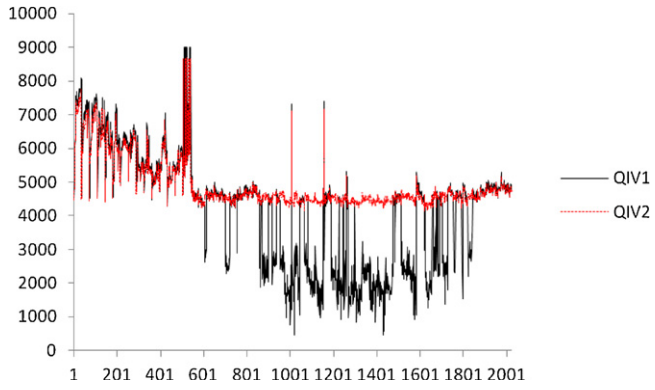

Fig. 18. Air flow rates of fans 1 and 2 in the FHU4 in the dataset_2.

associations between various parameters. Moreover, an arrow towards the parameter shows that this parameter appears in the conclusion of the association rule, and vice versa.

The map provides a holistic pattern of associations between various parameters. Clearly it can be seen that there exists a significant difference between the parameters on the amount of associations with other parameters. For example, TA_{Iar} and TA_{IIar} (i.e. fresh air temperature after the recuperation glycol in the FHUs 1 and 2) has only one association with other parameters and both of them appear in the premise. This indicates that these two parameters' values may be purely random or remain relatively stable throughout the whole winter and thus no association with other parameters can be found. It may have occurred since these two parameters are partly decided by outdoor air temperature, which is uncontrollable and relatively irregular. On the contrary, Q_{IV2} (i.e. the fresh air flow rate of fan 2 in the FHU 4) has the most associations with other parameters, and appears in both premises and conclusions. This indicates the parameter has the highest possibility of influencing or being influenced by other parameters and thus deserves extra attention.

In addition, between similar parameters (e.g. air flow rates of two fans in the same FHU), difference in the amount of associations with other parameters should not be huge. However, it is noticed that, between TA_{IVac} and TA_{Vac} (i.e. the fresh air temperature after the cooling coil in the FHUs 4 and 5), such difference is significant: TA_{IVac} only has one association with other parameters while TA_{Vac} has eight. This implies that the FHU 4 may have a fault. Accordingly, data analysis was performed on various parameters of the FHU 4; and the air flow rates of fans 1 and 2 in the FHU 4 are plotted in Fig. 18. Clearly, the air flow rates between these two fans are completely different most of the time. Considering fan 1 and fan 2 in the same FHU are identical and controlled by the same VSD, it can be inferred that, either the fan 1 or the fan 2 (or both of them) in the FHU 4 have a fault. This conclusion is in accordance with the conclusion drawn from *Rules 1–4* in Category 2 (Section 5.2).

The acquired knowledge could help building operators and owners better understand HVAC system operation and detect faults.

6. Summary and conclusions

In this paper, a new methodology is proposed for examining all the associations and correlations between building operational data. Accordingly, useful knowledge will be uncovered to help improve HVAC system performance and reduce energy consumption. The methodology is based on a basic data mining technique: association rule mining. In order to use this methodology, two-year building operational data needs to be collected. Data pre-processing should be performed before the ARM to remove outliers, so as to improve the quality of data and, consequently, the mining results. Furthermore, to take complete advantage of building operational data, data in different period length (e.g. both a day and a year) should be mined. Also, the obtained associations and correlations in different years should be compared between each other.

In order to demonstrate its applicability, this methodology was applied to the EV building located in Montreal, which is very cold in winter. Accordingly, the winter data of the air-conditioning system in this building in both 2007 and 2009 was mined. A waste of energy in the air-conditioning system was identified through mining association rules for the coldest day. Also, based on the comparison between winter association rules in the different years, possible faults in equipment were detected, and a low/no cost strategy for saving energy in system operation was proposed. Moreover, the association map was used to provide a holistic view of all the generated rules. This map could help explain how various parameters associate one with each other, and detect faults in equipment.

The proposed methodology allows for addressing the special challenges caused by the complexity of large volume of building operational data. By using this methodology, building operators and owners can discover all the useful associations and correlations between building operational data. Based on domain expertise, they can translate the obtained associations and correlations into useful knowledge, thereby better understanding building operation, identifying energy waste, detecting faults in equipment, and proposing low/no cost strategies for saving energy.

The main focus of future research should be placed on applying the proposed methodology to building operational data collected in different building sectors, climates, and building automation systems, in order to further evaluate its effectiveness and help understand the impact of different elements influencing building energy consumption. Once the methodology is generally accepted, it can be integrated into online data analysis and online fault detection to reduce building energy consumption efficiently. The software RapidMiner can be employed to perform the ARM and to help realize this methodology. Moreover, it can serve as a data mining engine for the integration and can automatically report interesting rules/patterns without requiring human intervention. However, data analysts are still necessary to compare obtained association rules to discover useful knowledge about building energy performance improvement.

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

[1] S. Deng, Energy and water uses and their performance explanatory indicators in hotels in Hong Kong, Energy and Buildings 35 (8) (2003) 775–784.
[2] P.C.H. Yu, W.K. Chow, Energy use in commercial buildings in Hong Kong, Applied Energy 69 (4) (2001) 243–255.
[3] Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy and Buildings 42 (10) (2010) 1637–1646.
[4] R. Priyadarsini, W. Xuchao, L.S. Eang, A study on energy performance of hotel buildings in Singapore, Energy and Buildings 41 (12) (2009) 1319–1324.
[5] Z. Yu, B.C.M. Fung, F. Haghighat, H. Yoshino, Edward Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy and Buildings 43 (6) (2011) 1409–1417.
[6] W. Chung, Y.V. Hui, A study of energy efficiency of private office buildings in Hong Kong, Energy and Buildings 41 (6) (2009) 696–701.
[7] Y. Tonooka, J. Liu, Y. Kondou, Y. Ning, O. Fukasawa, A survey on energy consumption in rural households in the fringes of Xian city, Energy and Buildings 38 (11) (2006) 1335–1342.
[8] S. Chen, H. Yoshino, N. Li, Statistical analyses on summer energy consumption characteristics of residential buildings in some cities of China, Energy and Buildings 42 (1) (2010) 136–146.
[9] F.J.S. de la Flor, J.M.S. Lissén, S.Á Domínguez, A new methodology towards determining building performance under modified outdoor conditions, Building and Environment 41 (9) (2006) 1231–1238.
[10] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, MIT Press, Cambridge, MA, 2001.
[11] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ, 1998.
[12] Association Rule Mining with WEKA http://maya.cs.depaul.edu/~Classes/Ect584/Weka/associate.html.
[13] J. Han, M. Kamber, Data Mining Concepts and Techniques, Elsevier Inc., San Francisco, 2006.
[14] Rapid–I–RapidMiner, http://rapid-i.com/content/view/181/190/.
[15] D.R. Helsel, R.M. Hirsch, Statistical Methods in Water Resources, U.S. Department of the Interior, 2002.