Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 143 (2017) 119–124

Energy Procedia

www.elsevier.com/locate/procedia

World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference, WES-CUE 2017, 19–21 July 2017, Singapore

Mining Gradual Patterns in Big Building Operational Data for Building Energy Efficiency Enhancement

Cheng Fan[a,]*, Fu Xiao[b]

[a]*Department of Construction Management and Real Estate, Shenzhen University, Shenzhen, 518000, China*
[b]*Department of Building Services Engineering, The Hong Kong Polytechnic University, Hong Kong, China*

Abstract

The advance in information technology has enabled the real-time monitoring and controls over building operations. Massive amounts of building operational data are being collected and available for knowledge discovery. Advanced data analytics are urgently needed to fully realize the potential of big building operational data in enhancing building energy efficiency. Data mining (DM) technology, which is renowned for its excellence in discovering hidden knowledge from massive datasets, has attracted increasing attention from the building industry. The rapid development in DM has provided powerful mining methods for extracting insights in various knowledge representations. Gradual pattern mining is a promising technique for identifying interesting patterns in big data. The knowledge discovered is represented as gradual rules, i.e., '*the more/less A, the more/less B*'. It can bring special interests to building energy management by highlighting co-variations among building variables. This paper investigates the usefulness of gradual pattern mining in analysing massive building operational data. Together with the use of decision trees, motif discovery and association rule mining, a comprehensive mining method is developed to ensure the quality and applicability of the knowledge discovered. The method is validated through a case study, using the real-world data retrieved from an educational building in Hong Kong. It shows that novel and valuable insights on building operation characteristics can be obtained, based on which fault detection and optimal control strategies can be developed to enhance building operational performance.

Peer-review under responsibility of the scientific committee of the World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference.

* Corresponding author. Tel.: +86-0755-26916426; fax: +86-0755-26916426.
E-mail address: fancheng@szu.edu.cn

1876-6102
Peer-review under responsibility of the scientific committee of the World Engineers Summit – Applied Energy Symposium & Forum: Low Carbon Cities & Urban Energy Joint Conference.
10.1016/j.egypro.2017.12.658

Keywords: Gradual pattern mining; Data mining; Knowledge discover; Building operational performance; Building energy efficiency.

1. Introduction

Adopting the Building Automation System (BAS) for the real-time monitoring and controls of the building services systems has become a top trend in the building sector. As reported by Waide et al. (2013), approximately 22% of the energy consumed during building operations can be saved through the use of advanced building automation technologies. Considering that building operations account for 80-90% of the total building energy consumption, reliable and robust BAS-based building energy management methods are urgently needed to achieve building sustainability.

One promising approach is to perform knowledge discovery from the massive amounts of building operational data collected by BAS. Building operational data typically record the indoor and outdoor environment, power consumptions and operating parameters of different building services systems or components. The knowledge hidden can be very helpful for improving the building energy efficiency. As reviewed by Ma and Wang (2009), previous studies mainly adopted traditional data analysis methods, such as statistics and physical principles, to analyze building operational data for predictive modeling, fault detection and diagnosis, and control optimization. Despite of the encouraging research results, such data analysis methods are neither efficient nor effective when analyzing massive amounts of data. Advanced data analytics for analyzing building operational data are urgently needed to fully realize the potential of big building operational data.

Data mining (DM) is a powerful technology in discovering potentially useful knowledge from large and noisy data. It has been successfully applied in various industries, such as retails, financial services and health care [Liao et al. 2012]. DM techniques can be generally classified into supervised and unsupervised techniques. Supervised techniques are capable of revealing the complex relationships between input and output variables. In the building field, existing studies mainly applied supervised DM to facilitate the predictive modeling of building energy consumptions, system performance indices, and indoor environment [Molina-Solana et al. 2017]. Unsupervised DM is useful for identifying the intrinsic data structures, correlations and associations. It has been applied to identify typical building operation conditions, operating behaviors, and interactions among building variables [Yu et al. 2016]. Association rule mining (ARM) is one of the most powerful unsupervised DM techniques. The knowledge discovered is highly interpretable and typically represented as a rule, i.e., $A \rightarrow B$, stating that if event A happens, B will also happens. Conventional ARM algorithms, such as Apriori and FP-growth, are suitable for mining associations between categorical variables. Such algorithms have been applied to extract associations among building operational data [Xiao and Fan 2014]. Considering that building operational data are mostly numerical data, some studies have adopted quantitative association rule mining (QARM) as the mining technique [Fan et al. 2015]. QARM adopts a data-driven approach to automatically identify the intervals for data discretization. Another research direction is to explore the temporal associations in building operational data [Fan et al. 2015].

Despite of the usefulness of the abovementioned ARM techniques, the knowledge discovered can only describe the co-occurrence of different events or conditions. There lacks methodologies to examine the co-variations between building variables. Gradual pattern mining, which is a special branch of ARM, can be utilized to discover gradual patterns (or co-variations). This study investigates the power of gradual pattern mining in analyzing massive building operational data. Together with the use of decision trees and motif discovery, a comprehensive methodology is proposed to enable the discovery of applicable knowledge for building energy management.

2. Research Methodology

2.1. Research outline

Fig. 1 depicts the research outline. The left side illustrates the overall knowledge discovery process and the right side shows the techniques or methods used at each mining phase. Data cleaning is performed to handle missing values and outliers. Data partitioning is performed based on decision trees. The insights obtained are used to partition the whole building operational data into different groups for separate mining. The aim is to enhance the reliability and

robustness of the knowledge discovered. Subsequence filtering aims to identify the most frequent temporal patterns in building operations. It helps building operation staff to narrow down the amount of data to be analyzed and only focus on the most valuable or significant data subsets. The primary technique used for knowledge mining is gradual pattern mining. The knowledge discovered is represent as gradual patterns.

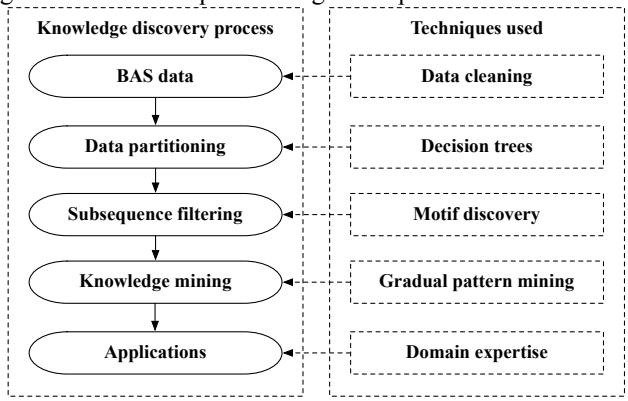

Fig. 1. Research outline

2.2. Data partitioning using decision trees

Decision tree is a supervised DM technique, which has been widely used to capture the nonlinear and complex relationships between input and output variables. The model developed is highly interpretable and the insights obtained are used to provide guidance for data partitioning.

Data partitioning should be performed based on typical building operation patterns. To achieve this target, the aggregated building cooling load is selected as the output variable, as it is a direct indicator of building operation patterns. The time variables, such as *Year*, *Month*, *Day*, *Day type*, *Hour*, *Minute*, are selected as input variables. The CART algorithm is selected for model development. It provides a data-driven approach for data partitioning.

2.3. Subsequence filtering using motif discovery

Building operational data are in essence time series data. The great dimensionality of time series data usually imposes great challenges in computing and knowledge mining. To tackle this challenge, the motif discovery is performed to identify the frequent subsequences in each data group. The benefits are two-fold. Firstly, it helps building operation staff to quickly identify the most dominant or significant temporal patterns in building operations. Secondly, the significant subsequences discovered can be used for further in-depth mining. It helps to narrow down the amounts of data to be analyzed while minimizing the risks of missing valuable knowledge.

A computationally efficient algorithm is adopted in this study for motif discovery [Chiu et al. 2003]. The method is based on the concept of symbolic aggregate approximation (SAX) and random projection. The numeric time series is firstly transformed into symbols using the SAX method. Random projection is then performed to evaluate the similarity among different subsequences, based on which significant motifs are discovered. The validity of this method in identifying meaning motifs in building operational data has been proved in our previous studies [Fan et al. 2015].

2.4. Knowledge mining using gradual pattern mining

Gradual pattern mining is adopted in this study as the primary technique to discover a novel type of knowledge from the building operational data, i.e., the co-variations between variables. Such kind of knowledge was firstly explored as gradual dependency through the use of ARM and linear regression [Hullermeier 2002]. Various semantics have been proposed to represent such type of knowledge, such as fuzzy association rules and degree variations [Berzal et al. 2007; Molina et al. 2007]. As a scalable solution to mine large numeric data set, Di-Jorio et

al. (2009) proposed a computationally efficient algorithm based on a binary matrices-based approach. The concept of closed gradual patterns was introduced by Ayouni et al. (2010) to reduce the quantity of gradual patterns obtained and therefore, providing a concise knowledge representation for manual inspection. The state-of-the-art implementations of mining gradual patterns include the PGLCM [Do et al. 2010] and ParaMiner [Negrevergne et al. 2014]. The ParaMiner was reported to be highly efficient due to its unique data reduction technique and multi-core architecture. This study adopts the ParaMiner as the mining tool.

The knowledge discovered by gradual pattern mining is represented as gradual patterns, which consists of a set of gradual items. A linguistic expression of gradual patterns is '*the more/less A, the more/less B*'. A gradual item is denoted as A^+ or A^-, It is in essence a pair made of a variable A and a variation denoted by + or -. A^+ is linguistically expressed as '*the more, the larger, or the higher A is*'. By contrast, A^- means '*the less, the smaller, or the lower A is*'. A gradual pattern M is a set of gradual items. The length of a gradual pattern is the same as the number of variables involved. For instance, $M=A^+B^+$ indicates that '*the less A, the more B*'. A support threshold, which specifies the minimum relative frequency of a gradual pattern to be considered as frequent, should be defined for practical applications. It is defined as 60% in this study for the discovery of frequent and dominant gradual patterns in subsequences.

3. Case study

3.1. Description of building and building operational data

The building operational data retrieved from an educational building in the Hong Kong Polytechnic University are adopted for analysis. The building consists of offices, classrooms and a computer data center. It has a gross floor area of approximately 11,000m^2 and 8500m^2 are served with air-conditioning systems. One-year data with a collection interval of 30-minute are used to validate the usefulness of the proposed method. The data includes the time variables (i.e., *Year*, *Month*, *Day*, *Day type*, *Hour*, *Minute*), outdoor environment variables (i.e., outdoor dry-bulb temperature and relative humidity), operating parameters of the HVAC system (e.g., temperatures, flow rates of the chilled water and condenser water, on-off signals and frequencies of fans), power consumptions of water-cooled chillers, and the building cooling load. To evaluate the building operational performance, the coefficient of performance (COP) of the water-cooled chillers and the part-load ratios are calculated as indicators. In total, the data have 113 variables and 17,040 observations.

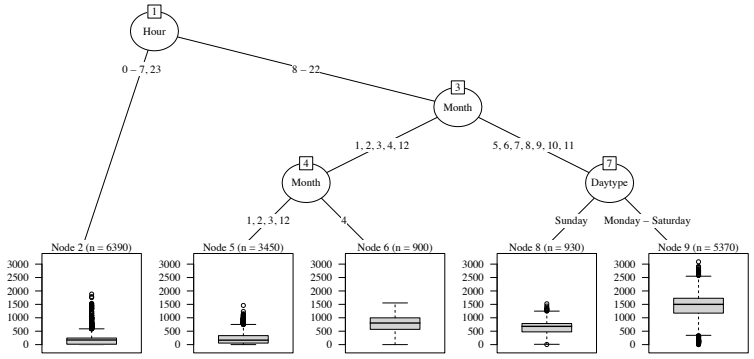

Fig. 2. Decision tee model for data partitioning

3.2. Insights on typical building operation patterns

A decision tree model is developed based on building cooling load and time variables. As shown in Fig. 2, three time variables, i.e., *Hour*, *Month* and *Day type*, are identified together with their splitting variables. The splitting values for *Hour* can be used to identify the office and non-office hour. In terms of the *Month*, May to November, which are the hotter and more humid seasons in Hong Kong, are grouped in contrast to those cooler and less humid

seasons. It should be mentioned that Node 4 further classified the cooler seasons into two groups, i.e., one is April and the other contains January, February and December. This is normal as April is the transient season. In terms of the *Day type*, Monday to Saturday are grouped as their cooling loads are typically higher than those on Sunday. This is in accordance with expectations, as conferences and seminars are often scheduled on Saturday in this building, resulting in a similar cooling load with weekdays.

The insights obtained from the decision tree model are used to divide the whole building operational data into 12 groups. The grouping is based on the following partitioning criteria: *Hour* in {0-7, 23} and {8-22}, *Month* in {1, 2, 3, 12}, {4} and {5-11}, *Day type* in {Monday to Saturday} and {Sunday}.

3.3. Insights on temporal building operation patterns

Motif discovery is performed for each data group separately to identify significant temporal operation patterns. The group corresponding to office hours (i.e., 8 a.m. to 10 p.m.) on Monday to Saturday in transient seasons (i.e., April) is utilized for illustration. Each daily time series is firstly transformed into symbols using the SAX method. The window size for piecewise aggregation is set as 2 (i.e., 1-hour) and the alphabet number is set as 3 (i.e., possible symbols are *A*, *B* and *C*). Temporal operation patterns for each variable can be successfully identified. Fig. 3 presents 4 significant temporal patterns considering the building cooling load. Figs. 3(a) and 3(b) have similar cooling load between 8 a.m. and 6 p.m. The major difference is observed for the rest time span. The cooling load shown in Fig. 3(a) does not decrease dramatically until 10:30 p.m., while a significant drop is observed at 6:30 p.m. in Fig. 3(b). By contrast, Figs. 3(c) and 3(d) have smaller cooling load between 8 a.m. and 6 p.m. These two temporal patterns can be distinguished according to the cooling load between 6:30 p.m. and 11:30 p.m.

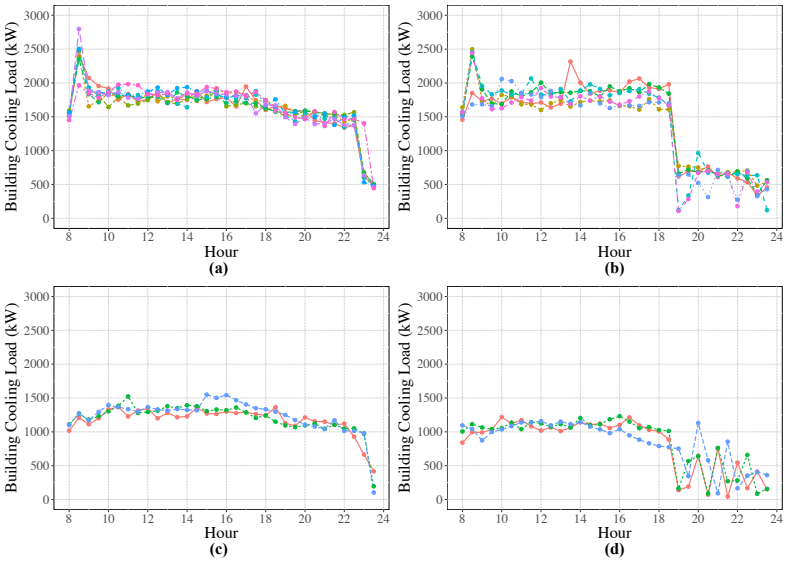

Fig. 3. Examples of significant temporal building operation patterns identified

3.4. Insights on gradual patterns in building operations

The ParaMiner algorithm is applied to discover gradual relationships within different temporal operation patterns. Table 1 presents some example gradual patterns extracted from the motifs discovered above. These patterns are in accordance with domain expertise and therefore, can be used to construct a knowledge database for detecting abnormal operating conditions. It is worth mentioning that the patterns are of special interests when performance evaluation metrics are involved. For instance, Patterns 5 and 6 describe two contradicting patterns. Pattern 5 states that the higher the part-load ratio, the higher the COP, while Pattern 6 holds the opposite. Further investigation shows that highest COPs are achieved when PLR ranges from 0.8 to 0.9. Future studies will be conducted to perform more comprehensive analysis for practical applications.

Table 1. Example gradual patterns

No.	Gradual patterns	Explanations
1	CHW_ST_Main$^+$, CHW_RT_Main$^+$	The higher the supply chilled water temperature, the higher the return chilled water temperature
2	Cooling_Load$^+$, CHW_Flow_Main$^+$	The higher the cooling load, the higher the chilled water flow rate
3	CDW_RT_Main$^+$, T_ambient$^+$	The higher the return condenser water temperature, the higher the outdoor temperature
4	Cooling_Load$^+$, COP$^+$	The higher the cooling load, the higher the COP
5	PLR$^+$, COP$^+$	The higher the part-load ratio, the higher the COP
6	PLR$^+$, COP$^-$	The higher the part-load ratio, the lower the COP

4. Conclusions

This study proposes a data mining-based methodology to extract a novel type of knowledge, i.e., gradual patterns, from building operational data. The methodology is deliberately designed considering the mining efficiency and the quality of knowledge obtained. It serves as a generic solution for practical applications. A case study has been conducted to validate the methodology. Research results show that the knowledge obtained can bring novel and valuable insights for building energy management. The data analysis is performed based on the open-source software *R* and *ParaMiner*.

Acknowledgements

The authors gratefully acknowledge the support of this research by the Natural Science Foundation of SZU (grant no. 2017061).

References

[1] Waide, Ure, Karagianni, Smith, and B. Bordass. (2013) “The scope for energy and CO_2 savings in the EU through the use of building automation technology.” *Final report for the European Copper Institute 2013*.
[2] Ma, Wang (2009) “Building energy research in Hong Kong: A review.” *Renewable and Sustainable Energy Reviews* 13.8 (2009): 1870-1883.
[3] Liao, Chu, and P.Y. Hsiao. (2012) “Data mining techniques and applications – A decade review from 2000 to 2011.” *Expert Systems with Applications* 39.12 (2012): 11303-11311.
[4] Molina-Solana, Ros, Ruiz, Gomez-Romero, and M.J. Martin-Bautista. (2017) “Data science for building energy management: A review.” *Renewable and Sustainable Energy Reviews* 70 (2017: 598-609.
[5] Yu, Haghighat, and C.M. Fung. (2016) “Advances and challenges in building engineering and data mining applications for energy-efficient communities.” *Sustainable Cities and Society* 25 (2016): 33-38.
[6] Xiao, Fan (2014) “Data mining in building automation system for improving building operational performance.” *Energy and Buildings* 75 (2014): 109-118.
[7] Fan, Xiao, and C.C. Yan. (2015) “A framework for knowledge discovery in massive building automation data and its application in building diagnostics.” *Automation in Construction* 50 (2015): 81-90.
[8] Fan, Xiao, Madsen, and Dan Wang. (2015) “Temporal knowledge discovery in big BAS data for building energy management.” *Energy and Buildings* 109 (2015): 75-89.
[9] Chiu, Keogh, and S. Lonardi. (2003) “Probabilistic discovery of time series motifs.” *Proceedings of the ACM SIGKDD* (2003): 493-498.
[10] Hullermeier. (2002) "Association rules for expressing gradual dependencies." *Proceedings of the International Conference PKDD* (2002): 200-211.
[11] Berzal, Cubero, Sanchez, Vila, and J.M. Serrano. (2007) “An alternative approach to discover gradual dependencies.” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 15.5 (2007): 559-570.
[12] Molina, Serrano, Sanchez, and M. Vila. (2007) “Measuring variation strength in gradual dependencies”. *Proceedings of the International Conference EUSFLAT* (2007): 337-344.
[13] Ayouni, Laurent, Yahia, and Pascal Poncelet. (2010) “Mining closed gradual patterns.” *Proceedings of the ICAISC* (2010): 267-274.
[14] Di-Jorio, Laurent, and Maguelonne Teisseire. (2009) “Mining frequent gradual itemsets from large databases.” *Proceedings of the Advances in Intelligent Data Analysis* (2009): 297-308.
[15] Do, Laurent, and Alexandre Termier. (2010) “PGLCM: efficient parallel mining of closed frequent gradual itemsets.” *Proceedings of IEEE Intenational Conference on Data Mining* (2010): 138-147.
[16] Negrevergne, Termier, Rousset, and Jean-Francois Mehaut. (2014) “ParaMiner: A generic pattern mining algorithm for multi-core architectures.” *Data Mining and Knowledge Discovery* 28.3 (2014): 593-633.