Available online at www.sciencedirect.com

ScienceDirect

Energy Procedia 158 (2019) 2481–2487

Energy Procedia

www.elsevier.com/locate/procedia

10th International Conference on Applied Energy (ICAE2018), 22-25 August 2018, Hong Kong, China

Discovering Complex Knowledge in Massive Building Operational Data Using Graph Mining for Building Energy Management

Cheng Fan[a], Mengjie Song[b], Fu Xiao[c,*], Xue Xue[d,e]

[a]*Department of Construction Management and Real Estate, Shenzhen Univeristiy, Shenzhen, 518000, China*
[b]*Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Japan*
[c]*Department of Building Services Engineering, The Hong Kong Polytechnic Univeristy, Hong Kong, China*
[d]*Department of Building Technology and Science, Tsinghua University, Beijing, 100000, China*
[e]*Shenzhen DAS Intellitech Co., Ltd., Shenzhen, 518000, China*

Abstract

Discovering useful knowledge from massive building operational data is considered as a promising way to improve building operational performance. Conventional data analytics can only handle data stored in a single two-dimensional data table, while lacking the ability to represent and analyze data in complex formats (e.g., multi-relational databases). Graphs are capable of integrating and representing various types of information, such as spatial information and affiliations. The knowledge discovery based on graph data can therefore be very helpful for revealing complex relationships in building operations. This study proposes a novel methodology for analyzing massive building operational data using graph-mining techniques. Two problems are specifically addressed, i.e., graph generation based on building operational data and knowledge discovery from graph data. The methodology has been applied to analyze the building operational data retrieved from a real building in Hong Kong. The research results show that the knowledge obtained is valuable to characterize complex building operation patterns and identify atypical operations.

Keywords: Graph mining; Knowledge discovery; Data mining; Building automation system; Building operational performance.

Introduction

Employing the Building Automation System (BAS) for the automated control and management of building energy systems has become a top trend in the building sector. Besides fulfilling the online monitoring and control functions, BAS also records a large number of measurements and control signals at short time intervals (e.g., seconds to minutes). The knowledge hidden in such massive BAS data can be very valuable for building energy management and optimization.

10.1016/j.egypro.2019.01.378

Data mining (DM) is a promising solution to the knowledge discovery from massive data. It has been successfully applied in various industries for knowledge discovery, such as retails, financial services and marketing [1]. In the building field, DM techniques have been applied for predicting the building energy consumption, system performance indices and indoor environment, extracting frequent operating patterns and detecting anomalies in building operation [2,3]. Commonly used DM techniques in the building field include statistical learning and machine learning (e.g., support vector machines, decision trees, and artificial neural networks), association rule mining, clustering analysis and outlier detection methods [4].

One essential premise of applying the abovementioned DM techniques is that the data need to reside in a single two-dimensional data table. If the data format becomes more complex (e.g., multiple data tables are used to store different types of information), data pre-processing is required to unify these tables into one before applying conventional DM techniques. Such type of data pre-processing could be time-consuming and sometimes not even possible without a significant information loss. It can be foreseen that the BAS data will become more diverse and complex due to the enrichment in the types of information that can be collected, e.g., temporal and spatial information. Therefore, advanced analytics are urgently needed to ensure the mining efficiency and effectiveness. The research gap emerges as little research has been done in this area.

To tackle this problem, this study proposes a novel methodology to discover complex knowledge from BAS data using graph mining. The paper is organized as follows: Section 2 serves as an overview on graph mining; Section 3 introduces the research methodology; A case study is shown in Section 4 and conclusions are drawn in Section 5.

1. An overview of graph mining

1.1. Basics on graphs

Graph is one of the most generic, natural, and interpretable formats for data representation. Great flexibility is provided in the knowledge discovery process as users can readily manipulate the graph layout to integrate and represent various types of information. The main elements of a graph are introduced as follows. A graph *G* consists of a set of vertices (or nodes), denoted as *V(G)* and a set of edges (or links), denoted as *E(G)*. A graph *S* is said to be a subgraph of graph *G* if $V(S) \subseteq V(G)$ and $E(S) \subseteq E(G)$.

A simple example is given to illustrate the usefulness of graph data. Table-1 presents the power consumption of a chiller and a cooling tower at time T_1 and T_2. Table-2 records the spatial location of these two components, i.e., one in basement and one on rooftop. The information is three-dimensional (i.e., time, power, location) and therefore, it is non-trivial to integrate these two tables into one without information loss. By contrast, a graph can be readily constructed for information integration as shown in Fig. 1. The top 2 vertices represent the temporal information and are labelled as "T_1" and "T_2" respectively. The edge connecting these two vertices are labelled as "$dT=1$" which indicates that the time step difference. Each of the top 2 vertices is connected with two vertices labelled as "*Chiller*" and "*CT*". The power consumption is encoded as edge labels. The bottom two vertices stand for the spatial information.

Table 1. An example data set containing the power data at two time steps

Time/Power	Chiller	Cooling tower
T1	Low	Low
T2	High	High

Table 2. An example data set containing the location of two components

Component	Location
Chiller	Basement
Cooling tower	Rooftop

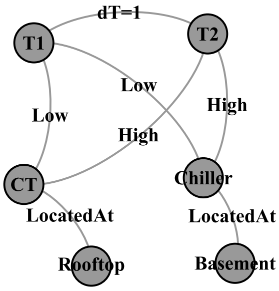

Fig. 1. An example graph for representing information in two data tables

1.2. Frequent subgraph mining

Graph mining is the most widely used techniques in analysing complex and multi-relational data [5]. It has been successfully used to discover complicated knowledge in bioinformatics, financial services, counter-terrorism, social network analysis and etc. [5, 6]. Frequent subgraph mining (FSM) is one of the most essential graph mining techniques. It mainly works on undirected graphs with labelled vertices and edges. Popular applications of FSM include finding the common substructures of chemicals and identifying the frequent patterns of terrorist attacks [6].

FSM algorithms can be classified based on two criteria, i.e., whether the search is exact or inexact, and whether the search strategy is breadth-first or depth-first [7]. Inexact search FSM algorithms, such as SUBDUE and CREW, use approximated measures to compare two graphs. The mining efficiency is higher, but it is not guaranteed to discover all frequent subgraphs. Exact FSM algorithms are more commonly used due to their ability of discovering all frequent subgraphs. The algorithms can be further classified based on the search strategies, i.e., either breadth-first or depth-first search. The depth-first search strategy is typically more computationally efficient. Some representative algorithms are MoFa, gSpan, FFSM and GASTON. A recent study compared the performance of these four exact DFS-based FSM algorithms, showing that the gSpan algorithm generally has better performance in terms of the running time and memory usage [7].

One essential challenge of FSM is that the number of frequent subgraphs discovered can be very large and the majority of them are redundant. A subgraph becomes redundant if there is a super-graph that has the same support count. To enhance the mining efficiency, Yan and Han proposed an algorithm called CloseGraph to mine closed frequent graphs based on their work of gSpan [8]. A subgraph is called closed if there exists no super-graph having the same support count. In this study, the CloseGraph is adopted to mine frequent subgraphs.

2. Research methodology

2.1. Research outline

The methodology consists of four steps, i.e., data exploration, graph generation, frequent subgraph mining and post-mining. Data exploration adopts the decision tree method to characterize the building operation patterns, based on which the whole BAS data are divided for in-depth knowledge discovery. The second step transforms the raw BAS data into graph data. The CloseGraph algorithm is applied in the third step to identify frequent subgraphs. A post-mining method is proposed for the ease of knowledge interpretation, selection and application.

2.2. Graph generation

This study proposes a variable-based transformation method to generate graph data from BAS data. The method is developed with two considerations, i.e., the computation efficiency and the compatibility with FSM algorithms. The former requires the number of vertices and edges used to describe a certain amount of information is the minimum. The latter requires the graph to be labelled, connected (i.e., there is always a path from one vertex to another), but not weighted. The variable-based transformation method is inspired by the graph representation used in social network analysis, where each individual is represented as a vertex and their associated relationships are shown as edges. The general idea is that each BAS variable is in analogy to an individual person and denoted as a vertex. The interactions between BAS variables during a certain time period are encoded as edge labels.

The main challenge is to come up with a to describe the interaction between two variables. An intuitive way to describe the interaction between two variables during a certain time period is to calculate their correlation. However, the information conveyed in the resulting graph can be too abstract to provide insightful knowledge for practical applications. For instance, a high correlation between two numeric variables does not provide any indication on the actual operating conditions, e.g., whether the power consumption is at a low or high level.

This study developed a novel edge labelling method to represent the interaction between two variables in the BAS data. The method only works with categorical variables and therefore, discretization should be performed for numeric variables. Assuming that the BAS data has *N* observations, the first step is to determine a window size (denoted as *w*), which is used to divide the BAS data into $\frac{N}{w}$ non-overlapping temporal segments. The dominant, or the most frequent interaction modes between two variables in these temporal segments are identified and used as edge labels. An interaction mode is defined as a vector containing the categorical values of both variables. A notation is created based on the dominant interaction mode between two variables during each temporal segment. Table 3 presents an example of such notation assuming both variables have two levels (denoted as “Low” and “High”). If there is a tie in the dominant interaction mode, a longer notation is created with each end surrounded by zeros, e.g., denoted as “0120” when {*Low*, *Low*} and {*Low*, *High*} are tied as the dominant interaction mode. The edge label between two variables can be obtained by combining the notations in different temporal segments.

Table 3. Notations for different dominant interaction modes

Variable A	Variable B	Interaction mode	Notation
Low	Low	{Low, Low}	1
Low	High	{Low, High}	2
High	Low	{High, Low}	3
High	High	{High, High}	4

2.3. Post-mining methods

To facilitate the knowledge post-mining, a method is proposed to automatically output anomalies based on the frequent subgraphs discovered. Assuming that *Y* frequent subgraphs are discovered based on *X* graphs, the method outputs an anomaly score for each of the *X* graphs. The general idea is that a graph is abnormal if it has no subgraphs that perfectly match any of the frequent subgraphs discovered. For a given graph G_i, the anomaly score is defined as $A_i = \frac{1}{Y}\sum_{j=1}^{Y}\frac{D_{i,j}}{N_{s,j}}$, where $D_{i,j}$ is the minimal number of differences in vertices and edges between any subgraphs of G_i and the j^{th} frequent subgraph, $N_{s,j}$ is the number of vertices and edges of the j^{th} frequent subgraph. If there exists a perfect match between any subgraphs of G_i and a frequent subgraph discovered, A_i is assigned as infinity. A larger A_i indicates that G_i is less close to any of the frequent subgraphs discovered. The closer the A_i approaches to zero, the more interesting or potentially useful the anomaly could be, since it indicates a well-disguised anomaly. Therefore, it is recommended to manually inspect graphs with small anomaly scores.

3. Research results

3.1. BAS data description

The BAS data used in this study were retrieved from the zero-carbon building in Hong Kong, known as the ZCB. ZCB has a total site area of 14,700m^2. Most of the site is a landscaped area for public use. The main building is a 3-storey building with a footprint of 1,400m^2. More detailed information can be found in [9].

Around one-year BAS data (from April 2013 to March 2014) are adopted for analysis. In total, the data contains 8304 hourly recorded observations and 38 variables, including the year, month, day, hour, day type, the power consumption of 3 water-cooled chillers (*WCC*), 4 chilled water pumps (*CHWP*), 3 condenser water pumps (*CDWP*), 3 cooling towers (*CT*), 5 air-handling units (*AHU*) and 1 primary air-handling unit (*PAU*); the power consumption of outdoor landscape lighting (*LandLight*), normal power and lighting of the eco-office, basement area (*Base*), G/F common area (*GF*), multi-purpose room (*MPR*), mezzanine area (*Mezz*); the power generation the biodiesel tri-generator (*BDG*).

3.2. Identification of frequent operation patterns

The variable-based graphs are used as high-level abstractions of the BAS data. The BAS data during the office hours (i.e., 9 a.m. to 6 p.m.) in the working days of hot seasons are transformed into variable-based graphs with structural, temporal and level information embedded. An example graph is shown in Fig. 2. Each numeric variable is discretized into 3 levels, denoted as *Idle*, *Low* and *High*. The *Load Demand* is designed as the central vertex and connected with seven vertices representing the HVAC subsystems, i.e., *WCC*, *AHU*, *CT*, *CDWP*, *CHWP*, *PAU* and *BDG*. Some subsystems contain multiple components and such affiliation information is showed using edges. The normal power and lighting consumptions at different locations in ZCB are also recorded in the graph. The edge labels are created to summarize the interactions between two variables in three temporal segments, i.e., 9 a.m. to 11 a.m., 12 p.m. to 3 p.m. and 4 p.m. to 6 p.m. For instance, the edge label between *CT* and *CT3* is "*699*". It means that the dominant modes are "*CT=Low, CT3=High*", "*CT=High, CT3=High*" and "*CT=High, CT3=High*" in each temporal segment respectively. The minimum support threshold for FSM is set as 10%, which meets the common definition of anomalies [10]. In total, 1082 frequent subgraphs are discovered and used as a knowledge database. The post-mining method proposed in Section 3.3 is used to find atypical operations. Examples are shown as follows.

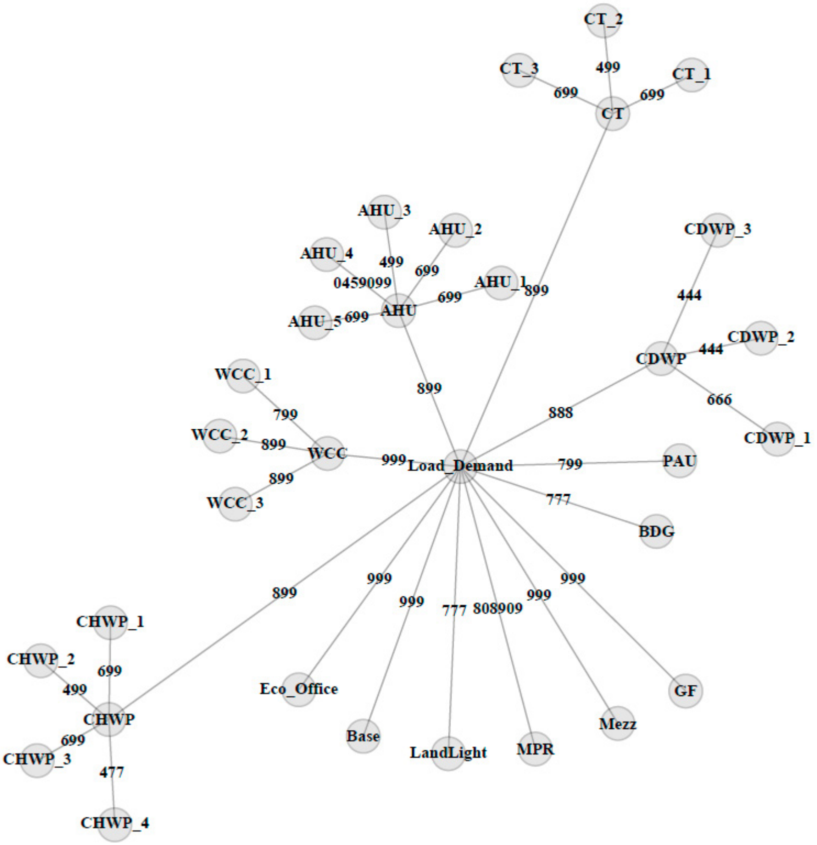

Fig. 2. An example graph generated using the variable-based approach

3.3. Discovering atypical operations

An anomaly graph is identified with a score of 0.51, indicating that on average, it is different from all the frequent subgraphs discovered with a mean proportion of 51%. Fig. 3 shows the anomaly graph with reference to its closest frequent subgraph. It is created in such a manner that the matched and unmatched portions are shown in blue and pink respectively, and the rest is shown in grey. It is apparent that the main difference is the Load Demand in the third temporal segment (i.e., 4 p.m. to 6 p.m.), which is "*High*" in the frequent subgraph and "*Low*" in the atypical operation. Further inspection reveals that the atypical graph represents the building operation during office hours on September 20, 2013 (Friday), which is a public holiday in Hong Kong. It is found out that on normal working days, ZCB are open for indoor tours during three time slots, i.e., 10 a.m. to 11:30 a.m., 2 p.m. to 3:30 p.m. and 4 p.m. to 5:30 p.m. The last tour is cancelled on Wednesdays and public holidays. The resulting load demand during that time period will be smaller than usual. The atypical operation identified is therefore an infrequent but normal operation.

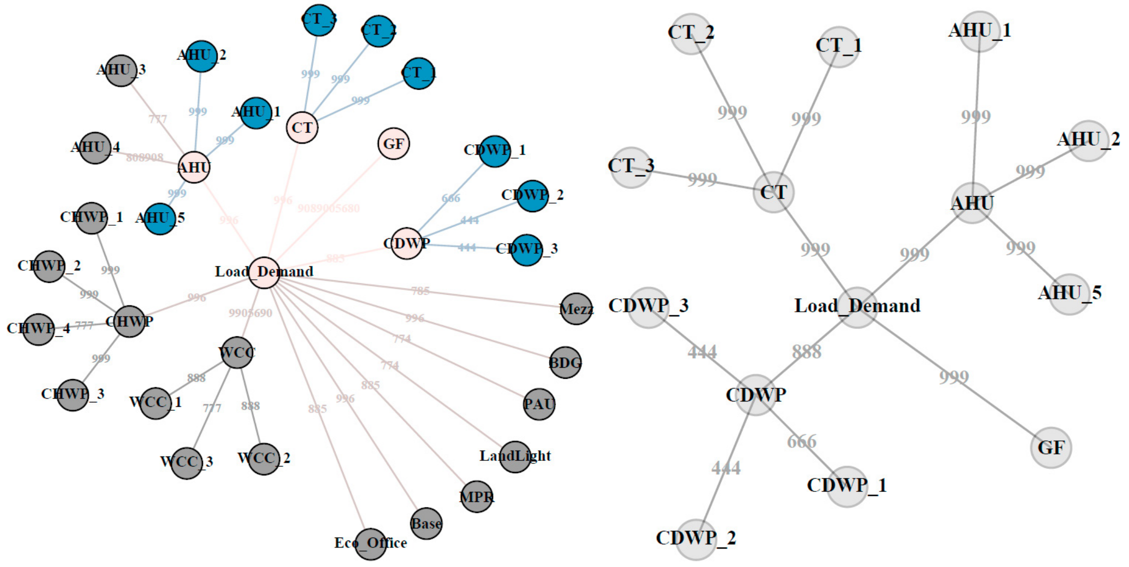

Fig. 3. (a) An atypical graph on September 20, 2013 (Friday); (b) A reference frequent subgraph for anomaly detection.

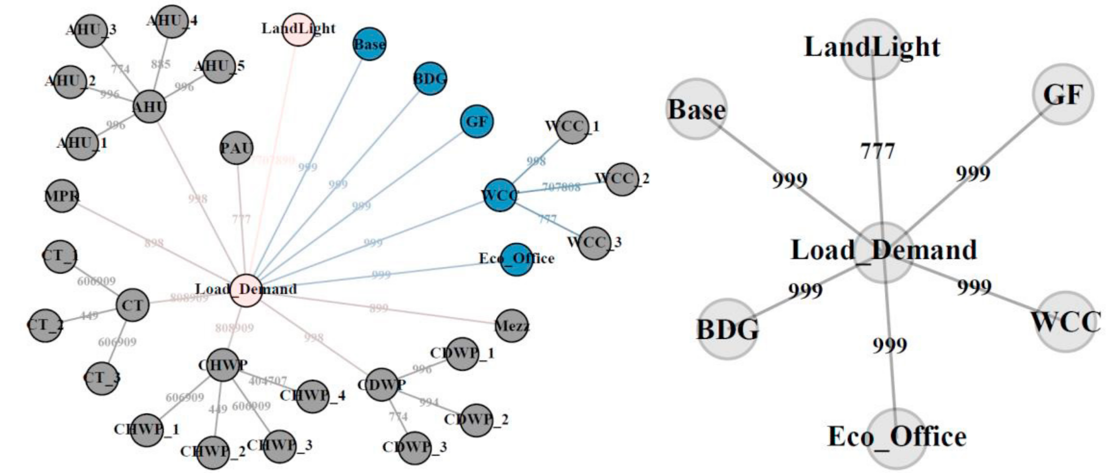

Fig. 4. (a) An atypical graph on September 2, 2013 (Monday); (b) An reference frequent subgraph for anomaly detection.

Fig. 4 presents another atypical operation on September 2, 2013 (Monday). It is observed that the power consumption of *Landscape Lighting* was *Low* and *High* at 5 p.m. and 6 p.m. while "Idle" in the frequent subgraph considered. Further inspection shows that the landscape lighting during hot seasons generally operates between 7 p.m. to 7 a.m. Such atypical operation can be caused by faults in manual control or poor outdoor visibility.

4. Conclusions

This study proposes a graph-based methodology to discover complex knowledge from massive BAS data. A variable-based transformation method is proposed to generate graphs describing complex interactions among BAS variables. The frequent subgraph mining is adopted as the primary mining technique for knowledge discovery. A graph-based anomaly detection method is developed for knowledge post-mining. The methodology has been applied to mine the BAS data retrieved from a building in Hong Kong. Atypical operations due to either accidents or faults have been successfully identified. The method proposed enables the extraction of multi-relational relationships in building operations while providing an effective visualization tool for the ease of building energy management. The open-source software *R* and *Gephi* were used to perform the mining and visualization tasks. Further study will focus on exploring the potential of graph data mining in extracting spatial and temporal knowledge in building operations.

Acknowledgements

The authors gratefully acknowledge the support of this research by the Natural Science Foundation of SZU (grant no. 2017061) and the Research Grant Council (RGC) of the Hong Kong SAR (152181/14E).

References

[1] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications-A decade review from 2000 to 2011. Expert Syst Appl 2012; 12: 11303-11.

[2] Fan C, Xiao F, Zhao Y. A short-term building cooling load prediction method using deep learning algorithms. Appl Energ 2017; 195: 222-33.

[3] Fan C, Xiao F, Yan CC. A framework for knowledge discovery in massive building automation data and its applications in building diagnostics. Automat Constr 2014; 50: 81-90.

[4] Molina-Solana M, Ros M, Ruiz MD, Gomez-Romero J, Martin-Bautista MJ. Data science for building energy management: A review. Renew Sust Energ Rev 2017; 70: 598-609.

[5] Cook DJ, Holder LB. Graph-based data mining. IEEE Intell Syst App 2000; 15: 32-41.

[6] Cook DJ, Holder LB. Mining graph data. 1st ed. New Jersey: Wiley; 2006.

[7] Worlein M, Meinl T, Fisher I, Philippsen M. A quantitative comparison of the subgraph miners MoFa, gSpan, FFSM and Gaston. Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2005, 392-404.

[8] Yan XF, Han JW. CloseGraph: Mining closed frequent graph patterns. The 9th ACM SIGKDD, August, 2003.

[9] ZCB fact sheet. Construction Industry Council, 2002. Hong Kong.

[10] Lazarevic A, Srivastava J, Kumar V. Data mining for analysis of rare events. PAKDD Tutorial, 2004.