An application of data mining in district heating substations for improving energy performance

Puning Xue[1], *Zhigang* Zhou[1], *Xin* Chen[1], and *Jing* Liu[1,2,*]

[1]Shool of Municipal and Environmental Engineering, Harbin Institute of Technology, Harbin 150000, China
[2]State Key Laboratory of Urban Water Resource and Environment, Harbin Institute of Technology, Harbin 150000, China

Abstract. Automatic meter reading system is capable of collecting and storing a huge number of district heating (DH) data. However, the data obtained are rarely fully utilized. Data mining is a promising technology to discover potential interesting knowledge from vast data. This paper applies data mining methods to analyse the massive data for improving energy performance of DH substation. The technical approach contains three steps: data selection, cluster analysis and association rule mining (ARM). Two-heating-season data of a substation are used for case study. Cluster analysis identifies six distinct heating patterns based on the primary heat of the substation. ARM reveals that secondary pressure difference and secondary flow rate have a strong correlation. Using the discovered rules, a fault occurring in remote flow meter installed at secondary network is detected accurately. The application demonstrates that data mining techniques can effectively extrapolate potential useful knowledge to better understand substation operation strategies and improve substation energy performance.

1 Introduction

China is the second largest building energy user in the world [1]. District heating (DH) system, an efficient method of supplying heat to consumers, account for 24.0% of total building energy consumption in China [2]. With the acceleration of urbanization, the DH sector presents a tendency of sharp increasing for the past few years. As a result, improving DH energy performance is vital for building energy savings.

Recently, automatic meter reading system has been installed in DH system resulting in large amount of DH-related data. However, the data obtained is rarely fully analysed due to the poor quality of meter readings and the lack of effective data analysis techniques. Data mining is an effective technique to extract new knowledge from big data, while few studies have been performed concerning the application of data mining in DH system.

This paper aims to use data mining method for improving energy performance of DH substations. The method includes two popular data mining techniques, namely, cluster

[*] Corresponding author: liujinghit0@163.com

DOI: 10.1051/e3sconf/20172200192

analysis and association rules mining. A case study of analysing the dataset collected from a DH substation in Changchun, China, is conducted to demonstrate its applicability.

2 Methodology

2.1 Technical approach

Hand et al. [3] define data mining as: "The analysis of large observation datasets to find unsuspected relationships and to summarize the data in novel ways so that owners can fully understand and make use of the data". In the past several decades, it has been successfully applied in economics, retails, telecommunication, and financial services [4]. Recently, efforts have also been made to investigate the application of data mining in HVAC field, including building energy consumption prediction [5, 6], building energy management [7, 8], fault detection and diagnosis [9, 10], and occupant behaviour [11, 12].

From a broad view, data mining is the process of the Knowledge Discovery in Database (KDD) and involves the following seven steps [4]: (1) Data cleaning; (2) Data integration; (3) Data selection; (4) Data transformation; (5) Data mining; (6) Pattern evaluation; (7) Knowledge presentation. Steps 1−4 are different forms of data preprocessing and aim to improve data quality and transform the data into suitable format for mining. Step 5 is the essential process for data analysis. And steps 6−7 are the process of post-mining, where the truly interesting knowledge are selected and presented to users.

With the consideration of data characteristics of DH substations, a three-step method is employed to extrapolate unsuspected patterns and associations from DH-related data. In step 1, two-heating-season substation data with a resolution of 10-min are selected from the automatic meter reading system. And one-heating-season data are considered as a dataset. That is, two datasets named dataset 1 and 2 are prepared for mining. It should be mentioned that both step 2 and 3 are performed separately in each dataset. In step 2, cluster analysis is applied to identify distinct heating patterns based on the primary heat supply of the substation. Due to the heat load variation in DH system are both seasonal and daily [13]. Thus, cluster analysis consists of two parts: month clustering and hour clustering. Specifically, data within one month are regarded as a whole data object, and clustering is performed to identify the seasonal heating patterns among all objects. Then, data in an hour are considered as a whole data object, and hour clustering is applied to study the daily operation patterns based on each seasonal heating pattern. In step 3, association rule mining (ARM) is implemented to discover the unsuspected knowledge in the format of association rules based on each heating pattern identified by cluster analysis.

2.2 Cluster analysis

Cluster analysis is the process of grouping data objects into clusters so that objects in the same cluster have high similarity, while objects in different clusters have low similarity [4]. Agglomerative hierarchical clustering is one of commonly used clustering algorithm and is adopted in this study. Using a bottom-up strategy, it begins with every object representing a singleton cluster and recursively merge the closest clusters into a single cluster until certain termination conditions are satisfied [4]. It should be mentioned that users can specify the desired clusters number k as the termination condition. The performance of clustering result is evaluated by Dunn index (DI). The index is defined as the ratio of minimum distance between clusters to maximum distance inside clusters [14]. The larger value of DI means higher inter-cluster distances and lower intra-cluster distances, and then indicates better performance of clustering.

2.3 Association rule mining

ARM aims to identify frequent correlations hidden in dataset. Let $I = \{I_1, I_2, \dots, I_n\}$ be a set of n attributes called itemset. Let $D = \{D_1, D_2, \dots, D_m\}$ be a set of transactions called the database. Each transaction in D contains a subset of the items in I. The association rule can be described by the formula $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \neq \varnothing$, $Y \neq \varnothing$, and $X \cap Y \neq \varnothing$. The itemset X is called the antecedent of the rule, while the itemset Y is called the consequent of the rule [4]. The validity and certainty of the association rule can be measured in terms of its support, confidence and lift. Support is the joint probability of the antecedent and consequent. Confidence is the conditional probability of the consequent, given the antecedent. And lift can be considered as a simple dependence and correlation measurement between the consequent and the antecedent.

In this paper, Apriori algorithm is used to discover association rules. To perform ARM, minimum support threshold and minimum confidence threshold should be pre-specified. Moreover, most of the original dataset is numeric while Apriori algorithm require the data to be categorical. Therefore, data discretization is conducted to transform numeric attributes to categorical values before ARM. Specifically, all numeric attributes are classified into three categories using equal frequency method. And the three categories can be defined as “low”, “medium”, and “high” according to original numerical magnitude.

3 Results and discussions

3.1 Dataset

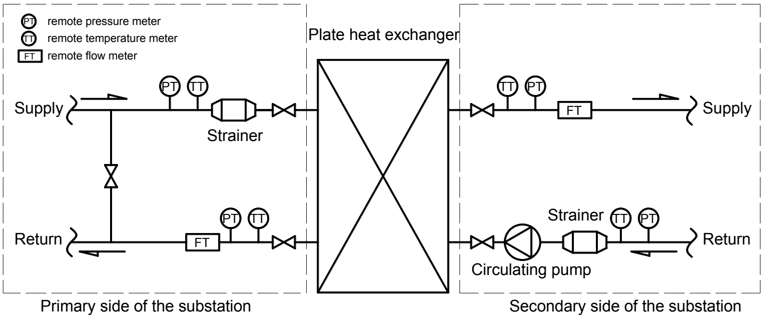


Fig. 1. Schematic view of the study substation.

The study DH substation is located in Changchun, a large provincial capital city in northeast China. Fig. 1 shows the schematic view of the substation. Automated and remote meters are mounted for monitoring and recording real-time operation parameters. Two-heating-season historical data from October 19, 2014 to April 11, 2015 and from October 19, 2015 to April 11, 2016 are collected for case study. The data of 14−15 heating season are called dataset 1 and the data of 15−16 heating season are called dataset 2. As shown in Table 1, a total of 14 attributes are monitored, and data of each attribute are collected with a measuring resolution of 10 min. It should be mentioned that there is not any remote heat meter mounted in the substation, so primary heat and secondary heat are calculated based on corresponding measured values of supply and return temperatures and flow rate separately. In order to make the association rules easier to understand, selecting the attributes under consideration is necessary. The pressure difference $\Delta P_{1\text{st}}$ and $\Delta P_{2\text{nd}}$ are calculated to express the pressure characteristics of primary network and secondary network, respectively. Similarly, the temperature difference $\Delta T_{1\text{st}}$ and $\Delta T_{2\text{nd}}$ are calculated to represent the temperature characteristics of primary network and secondary network separately. As a result, the number of attributes is reduced to 10.

Table 1. The monitored attributes of the dataset.

No.	Nomenclature	Attribute	Unit
1	t	Date	1
2	τ	Time	1
3	$P_{\mathrm{s,1st}}$	Primary supply pressure	MPa
4	$P_{\mathrm{r,1st}}$	Primary return pressure	MPa
5	$T_{\mathrm{s,1st}}$	Primary supply temperature	°C
6	$T_{\mathrm{r,1st}}$	Primary return temperature	°C
7	G_{1st}	Primary mass flow rate	t/h
8	Q_{1st}	Primary heat power	MJ/h
9	$P_{\mathrm{s,2nd}}$	Secondary supply pressure	MPa
10	$P_{\mathrm{r,2nd}}$	Secondary return pressure	MPa
11	$T_{\mathrm{s,2nd}}$	Secondary supply temperature	°C
12	$T_{\mathrm{r,2nd}}$	Secondary return temperature	°C
13	G_{2nd}	Secondary mass flow rate	t/h
14	Q_{2nd}	Secondary heat power	MJ/h

3.2 Substation heating patterns

Due to data are gathered at a 10-min interval, 4,320 observations of Q_{1st} are collected in each month. If all 4,320 observations are used, a 4,320-dimension sparse matrix is generated. Month clustering on this sparse matrix can take a long time, making such analysis impractical or infeasible. Therefore, a reduced representation of this sparse matrix is necessary. The mean and standard deviations of Q_{1st} are calculated for each month, resulting in a 2-dimension feature dataset. Month clustering on this feature dataset is more efficient yet produce the same analytical results. k, the number of clusters, is set from 2 to 3. Correspondingly, DI values are 1.08 and 1.25, respectively. Therefore, $k = 3$ is selected for clustering. Table 2 shows the clustering results of dataset 1. Three clusters are identified, indicating three distinct seasonal heating patterns. Same clustering results can be obtained from dataset 2.

Table 2. Clustering results of dataset 1.

Cluster ID	Month
A	10, 4
B	11, 3
C	12, 1, 2

Based on the three seasonal heating patterns, hour clustering is then performed. Note that, the original objects in the months partitioned into the same cluster should be analysed together. k is set between 2 and 10. Fig. 2 illustrated the evaluation results of cluster C of dataset 1. The DI is maximum when $k = 2$. Therefore, such parameter setting is selected. Fig. 3 shows the clustering memberships. It seems that a majority of objects in cluster *C1* distribute between 11:00−14:00 and 23:00−2:00. The rest of data are basically gathered from 2:00−11:00 and 14:00−23:00, and are collected in cluster *C2*. The box plot in Fig. 4 demonstrates the clustering results. Cluster *C1* and cluster *C2* indicate non-heating period and heating period, respectively. It can be seen that the substation is running under intermittent operation in each day. Same hour clustering results can be obtained from other five seasonal operational patterns of the two datasets. It can be concluded that the intermittent operation is a strategy commonly used during the two heating seasons. Table 3 summarizes the overall results of cluster analysis. By a combination of seasonal and daily operation patterns, each original dataset is disaggregated into 6 clusters. Each cluster indicates a particular operation pattern.

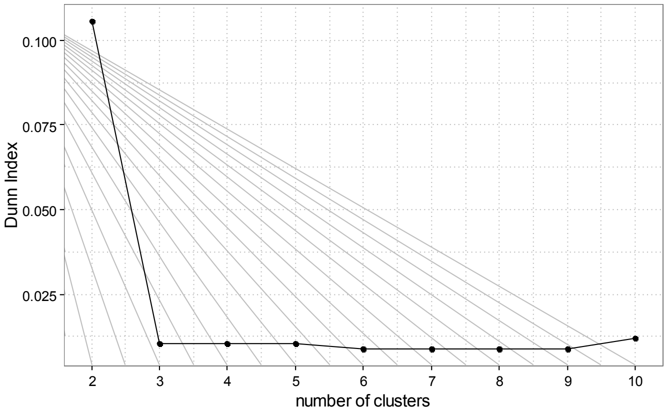


Fig. 2. Clustering evaluation of seasonal heating pattern *C* of dataset 1.

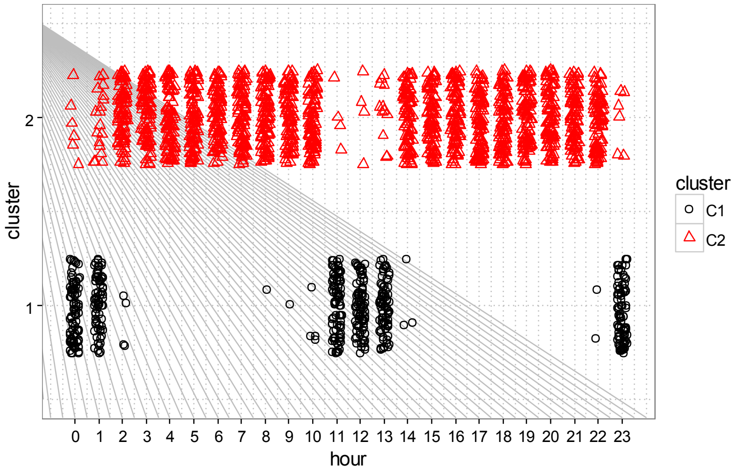


Fig. 3. Clustering results of seasonal heating pattern *C* of dataset 1.

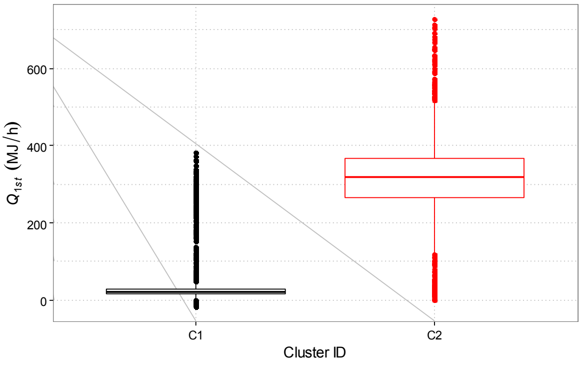


Fig. 4. Cluster features of daily heating pattern *C1* and *C2* of dataset 1.

Table 3. Summary of the clustering results of dataset 1 and 2.

Data set	**Heating season**	**Seasonal operation pattern**	**Cluster membership**	**Daily operation pattern**	**Cluster membership**		**Cluster ID**
1	14–15: 19/10/2014–11/04/2015	*A*	10, 4	Non-heating	7:00-16:00	20:00-4:00	*A1*
				Heating	4:00-7:00	16:00-20:00	*A2*
		B	11, 3	Non-heating	9:00-16:00	21:00-4:00	*B1*
				Heating	4:00-9:00	16:00-21:00	*B2*
		C	12, 1, 2	Non-heating	11:00-14:00	23:00-2:00	*C1*
				Heating	2:00-11:00	14:00-23:00	*C2*
2	15–16: 19/10/2015–11/04/2016	*A*	10, 4	Non-heating	8:00-16:00	21:00-4:00	*A1*
				Heating	4:00-8:00	16:00-21:00	*A2*
		B	11, 3	Non-heating	9:00-15:00	21:00-3:00	*B1*
				Heating	3:00-9:00	15:00-21:00	*B2*
		C	12, 1, 2	Non-heating	11:00-13:00	23:00-2:00	*C1*
				Heating	2:00-11:00	13:00-23:00	*C2*

3.3 Application of association rules

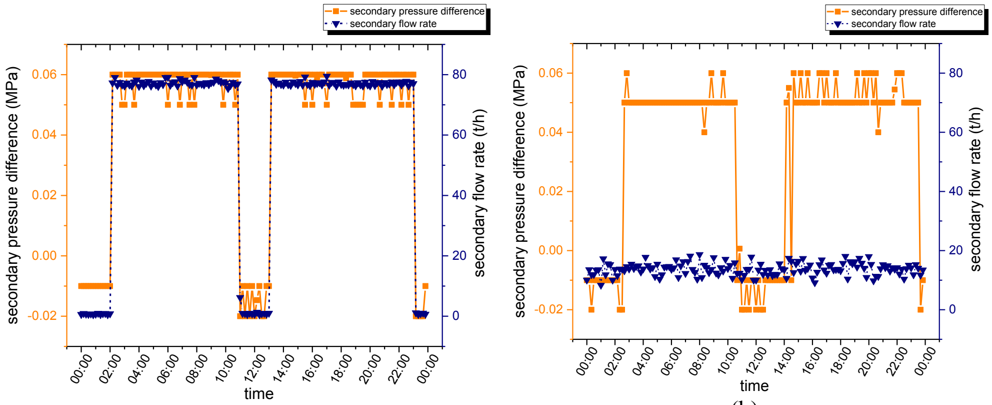
Apriori algorithm is then performed for discovering rules hidden in the 12 patterns respectively. For each pattern, a support of 0.1 and a confidence of 0.6 are set as the minimum thresholds. The support is relatively low to generate the association rules that maybe infrequent but interesting, and the confidence is relatively high to ensure the rules obtained are strong. Besides, the minimum threshold of lift is set as 1 to find positive correlations between the antecedent and consequent. In view of easy interpretation for rules extracted, the number of items in a rule is set to be 2, namely, both the antecedent and consequent of the rule have only one item. Many rules are obvious and can be easily interpreted by domain knowledge. Truly interesting rules need to be further selected.

The rules mined from dataset 1 and 2 with pattern $\Delta P_{2nd} \rightarrow G_{2nd}$ are potential interesting and are given in Table 4. From rules 1−6, it can be observed that the secondary pressure difference and secondary flow rate have a strong association in the substation. In order to provide an insight into these rules, the variation of secondary pressure difference and flow rate in one typical day of heating pattern *C* of 15−16 heating season is plotted in Fig. 5 (a). It can be seen that the variation of secondary pressure difference and flow rate follows the same trend in 15−16 heating season. Moreover, this rule pattern can help to understand the hydraulic condition of substation and be used as a guide of fault detection on the remote meters. This indicates that the secondary pressure difference and flow rate are always strongly associated. Accordingly, if the variation tendency of pressure difference and flow rate is dramatically different, then it can be deduced that the automated and remote meters may have faults.

Table 4. Summary of the rule pattern $\Delta P_{2nd} \rightarrow G_{2nd}$ of dataset 1 and 2.

No.	**Antecedent**	**Consequent**	**Support**	**Confidence**	**Lift**	**Cluster ID**	**Dataset**
	ΔP_{2nd}	G_{2nd}					
1	*low*	*low*	0.499	0.752	1.137	A1	2
2	*high*	*high*	0.818	0.895	1.095	A2	
3	*low*	*low*	0.440	0.893	1.085	B1	
4	*high*	*high*	0.820	0.852	1.040	B2	
5	*low*	*low*	0.658	0.990	1.041	C1	
6	*high*	*high*	0.953	0.989	1.038	C2	
7	*high*	*low*	0.851	1.000	1.000	C2	1

For example, rule 7 is mined from dataset 1. When substation supplies heat to consumers in heating pattern *C* of 14−15 heating season, the secondary pressure difference is high while the secondary flow rate is low. Moreover, the lift value is equal to 1. It means that there is no correlation between the secondary pressure difference and flow rate. The rule disobeys the knowledge extracted from rules 1−6, so it is reasonable to deduce that there is an anomaly in either remote flow meter or remote pressure meter. As illustrated in Fig. 5 (a) and (b), the variation of secondary pressure difference of two heating seasons is identical in heating pattern *C*. It can be inferred that the remote pressure meters work well and data transmission is accurate. Correspondingly, the variation trend of secondary flow rate should be identical to that of secondary pressure difference. However, Fig. 5 (b) shows that there is a continuous irregular oscillation in secondary flow rate. Hence, it can be deduced that a fault occurred in remote flow meter installed at secondary network in heating pattern *C* of 14−15 heating season.



(a)

(b)

Fig. 5. The variation of secondary pressure difference and flow rate in one typical day of heating pattern *C*: (a) 15−16 heating season; (b) 14−15 heating season.

4 Conclusions

Using data mining techniques, a novel method was proposed to extrapolate useful knowledge from substation operation data. A three-step technical approach is applied to fully analyse the massive data. Two popular data mining techniques, agglomerative hierarchal clustering and Apriori algorithm, are integrated in the method. The main conclusions are: (1) Cluster analysis identifies six distinct heating patterns based on the primary heat of the substation. The heating patterns can help users to further understand the operation strategies of the substation. During the whole heating season, the substation supplies heat to users with intermittent operation in each day. (2) Association rule mining discovers the interesting rules between secondary pressure difference and secondary flow rate, which can provide an insight to fault detection on the remote meters. According to the rules mined, a fault occurring in remote flow meter installed at secondary network is detected accurately. (3) An important aspect of this paper demonstrates the successful use of data mining techniques to discover useful knowledge about substation operation, which can provide essential guidance for improving energy performance of DH substation.

One limitation of this method is that data mining techniques cannot tell the value of the rules discovered, and domain knowledge is still needed to select and interpret the rules for practical applications. Therefore, further study is still necessary and the main focus should be placed on applying more advanced data mining techniques for extracting rules more efficiently. In addition, DH systems have important roles to play in future sustainable energy systems. To fulfil such roles, the current DH systems must evolve to be smart DH systems, in which more information exchanges between interfaces are required. This study demonstrates that data mining can bring significant benefits in data processing and failure detection for DH systems. To implement future smart DH systems, it is highly recommended that future research pay more attention to develop new methodologies based on data mining techniques for energy management, failure detection, operation optimization, and decision-making.

This study was sponsored by Key Projects of Natural Science Foundation of Heilongjiang Province of China (Grant No. ZD2016010). The authors would like to thank Changchun City Heating (Group) Co. Ltd. and Heilongjiang LONG DIAN Electric Co. Ltd. for providing measured data.

References

1. IEA, *Energy Balances of OECD/non-OECD Countries 2014* (2014)
2. Building Energy Research Center of Tsinghua University, *Annual report on China building energy efficiency* (China Architecture & Building Press, Beijing, China, 2015)
3. D. Hand, H. Mannila, P. Smyth, *Principles of data mining* (MIT Press, 2001)
4. J. Han, M. Kamber, J. Pei, *Data mining concepts and techniques*, *3rd edition* (Morgan Kaufmann Publishers Inc., 2012)
5. Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, Energ. Buildings **42**, 1637–1646 (2010)
6. C. Fan, F. Xiao, S. Wang, Appl. Energ. **127**, 1–10 (2014)
7. Z. Yu, F. Haghighat, B.C.M. Fung, L. Zhou, Energ. Buildings **47**, 430–440 (2012)
8. F. Xiao, C. Fan, Energ. Buildings **75**, 109–118 (2014)
9. C. Fan, F. Xiao, C. Yan, Automat. Constr. **50**, 81–90 (2015)
10. Z. Du, B. Fan, X. Jin, J. Chi, Build. Environ. **73**, 1–11 (2014)
11. Z. Yu, B.C.M. Fung, F. Haghighat, H. Yoshino, E. Morofsky, Energ. Buildings **43**, 1409–1417 (2011)
12. S. D'Oca, T. Hong, Build. Environ. **82**, 726–739 (2014)
13. H. Gadd, S. Werner, Appl. Energ. **106**, 47–55 (2013)
14. Z. Zhou, *Machine learning* (Tsinghua University Press, Beijing, China, 2016)