Monitoring District Heating Substations via Clustering Analysis*

Shahrooz Abghari[1], Veselka Boeva[1], Jens Brage[2],
Christian Johansson[2], Håkan Grahn[1], and Niklas Lavesson[3]

[1] Blekinge Institute of Technology, 371 79, Karlskrona Sweden shahrooz.abghari@bth.se
[2] NODA Intelligent Systems AB, 374 35, Karlshamn Sweden
[3] Jönköping University, 551 11, Jönköping Sweden

Abstract. In this paper, we describe an ongoing study for detecting deviating behaviour of district heating (DH) substations. We propose an approach for modelling, monitoring and analyzing the DH substations operational behaviour on a weekly basis. The proposed approach combines sequential pattern mining together with clustering analysis and minimum spanning tree to identify outliers. Our goal is to detect changes in operational behaviour of substations that can decrease their efficiency.

1 Introduction

District heating (DH) system provides a number of *buildings* with heat and domestic hot water from a *central boiler plant* through a *distribution network* for a limited geographical area. The provided heat transfers through substations from the distribution network into consumers' buildings to get heat and domestic hot water on demand. DH substations consist of different components and each can be a potential source of faults. Faults in substations can arise from stuck valves, fouled heat exchangers, malfunctions in temperature transmitters, control systems and more [1].

Faults in substations cause sub-optimal behaviour and in the long term, decrease the efficiency of the DH system. Therefore, early detection of faults can reduce maintenance cost and help avoid abnormal event progression. This in return can also increase the consumers' satisfaction and minimize energy waste. Gadd and Werner [2] showed that hourly meter readings can be used for detecting faults at DH substations. They identified three fault groups: 1) low average annual temperature difference, 2) poor substation control, and 3) unsuitable heat load pattern. The results of the study showed that the first group is the most important issue to focus on while addressing the third is probably the easiest and the most cost-effective.

Xue et al. [3] applied clustering analysis and association rule mining to detect faults in substations. Cluster analysis was applied in two steps 1) to partition the substations based on monthly historical heat load variations and 2) to identify daily heat variation using hourly data. The result of the clustering analysis was used for feature discretization and preparation for association rule mining. The results of the

* This work is part of the research project "*Scalable resource-efficient systems for big data analytics*" funded by the Knowledge Foundation (grant: 20140032) in Sweden.

study showed that the method can discover useful knowledge to improve the energy performance of the substations. However, for temporal knowledge discovery advanced data mining techniques were required.

Månsson et al. [1] proposed a method based on gradient boosting regression to predict an hourly mass flow of a well performing substation using only a few number of features. The built model was tested by manipulating the well performing substation data to simulate two scenarios: communication problems and a drifting meter fault. The results of the study showed that the proposed model can be used for continued fault detection.

In this study, we propose an unsupervised method that combines different data analysis techniques for modelling and monitoring the DH substations' operational behaviours in the absence of a labelled data. We initially extract weekly frequent patterns. Notice that we look for contextual collective outliers, i.e., sequences of events that since they occurred together in a specific time period (heating season) can be identified as outliers. The extracted patterns in each week are clustered, which model the DH substation's operational behaviour. Next, we analyze and assess the similarity between substation behaviours for every two consecutive weeks. The assessed similarity and a user-specified threshold can be used to measure the discrepancy between the substation performance within the studied time periods. When the discrepancy exceeds the threshold, we can perform further analysis by integrating the produced clustering solutions into a consensus clustering. In order to identify deviating behaviours we apply the minimum spanning tree (MST) algorithm on the exemplars of the built consensus clustering solution and cut the longest edge(s) of the MST. Smallest and distant sub-trees can be interpreted as outliers.

2 Proposed Solution

We propose a hybrid method for modelling, monitoring and analyzing the DH substations' operational behavior and performance. The proposed solution combines sequential pattern mining, clustering analysis and the MST algorithm. These data analysis techniques have also been used in our previous work where we have proposed a method for identification of sequences of unexpected events in data streams [4]. The current study reproduces and evaluates these techniques in a different industrial context to generalise their application. Steps of the proposed method are as follows:

Data preprocessing: In this step, we first remove all the duplicates and impute missing values. Additionally, extreme values that are often a result of faults in measurement tools are smoothed out by a Hampel filter [5], which is a median absolute deviation (MAD) based estimation. The filter computes the median, MAD, and the standard deviation (SD) over the data in a local window. We apply the filter with the default parameters. The size of the window is seven, i.e., 3-neighbours on either side of a sample. The threshold for extreme value detection is three. This means in each window a sample with the distance three times the SD from its local median is considered as an extreme value and is replaced by the local median.

Since we are monitoring the operational behaviour of substations based on outdoor temperature, five features that have a strong negative correlation with outdoor temperature are selected. These features are as follows: 1) primary temperature difference, 2) secondary temperature difference, 3) primary mass flow rate, 4) primary heat, and 5) substation efficiency. Using sequential pattern mining requires converting the continuous features to categorized features, *data discretization*. This process can be performed in a supervised or an unsupervised fashion [6]. Due to unavailability of labelled data, k-means-based discretization is used. The size of k is set to be four, the same as the number of season periods in Sweden.

Data segmentation and pattern extraction: Proper size of the time window for pattern extraction can lead us to monitor operational behaviour of the substations rather than the residents' behaviour. Therefore, after performing some preliminary experiments and having discussions with domain experts, the time window size is set to be a *week*. The PrefixSpan algorithm [7] is used to find frequent sequential patterns with the length of five in each week. Those sequential patterns that satisfy the user-specified support are considered as frequent patterns. In order to capture hourly operational behaviours of the substations, the size of the user-specified support threshold is set to be 1, i.e., any patterns that appear at least once will be considered.

Data analysis: The data analysis step can be further broken down into three sub-steps: 1) clustering the extracted patterns, 2) assessing a substation's behaviour by comparing the clustering solutions produced for every two consecutive weeks, and 3) applying further analysis and evaluation of the observed behaviour by building a minimum spanning tree and detecting the potential outliers. Sub-steps 2 and 3 can help the domain experts to better understand the DH individual substations' behaviour and also in comparison among the substations from the same head load category.

1. *Clustering frequent sequential patterns:* At this step, we model the substation operational behaviour by clustering the extracted patterns based on their similarities into groups. We use the *affinity propagation* (AP) [8] algorithm due to its ability to adapt the number of clusters from the data. Additionally, AP considers an actual pattern to be the exemplar in each cluster. The similarity between the patterns are calculated using Levenshtein distance [9].
2. *Assessing a substation's behaviours:* We can analyze and assess the similarity between substation's behaviours for every two consecutive weeks. This can be done through pairwise comparison of the exemplars of every two weeks clustering solutions. The assessed similarity can be used to measure the discrepancy between the substation performance within every two weeks period. All the assessed similarities can be considered as the substation's operational behaviour signature profile for the whole heating period. Additionally, this profile can be used for comparing the performance of substations with the same heat load category. When the discrepancy is significant (above a given threshold, e.g., 25%) and when the outdoor temperature is below 10 °C (heating season) a further analysis can be preformed by integrating the produced clustering solutions into a *consensus clustering*.

3. *Minimum spanning tree building and detecting outliers:* We can further apply the MST algorithm, which builds an MST by considering the exemplars of the built consensus clustering solution as nodes and the distance between them as edges. Notice that an MST is a tree with a minimum traversing cost. In order to identify deviating behaviours, the longest edge(s) of the MST is removed. Smallest and distant sub-trees created by the cut can be interpreted as outliers.

3 Summary

We present a hybrid method for modelling and monitoring the DH substations' operational behaviour. The proposed method can facilitate domain experts to better understand the DH substations' behaviour individually and in comparison with substations from the same head load category. We have applied the proposed methods on 10 randomly selected buildings (4 school buildings and 6 residential buildings). The collected data covers a two year period (2017 and 2018). The initial results of the study show that the proposed method enables to identify deviating behaviour in the DH substations. For example, we have found that the performance of some substations become suboptimal when the outdoor temperature is fluctuating between above and below 10 °C. This requires further analysis and interpretation by the domain experts in order to qualify whether it can be related to the occurrence of specific faults. The next step will be to apply the method on richer datasets to evaluate further its performance and scalability.

References

1. Månsson, S., Kallioniemi, P.O.J., Sernhed, K., Thern, M.: A machine learning approach to fault detection in district heating substations. Energy Procedia **149** (2018) 226–235
2. Gadd, H., Werner, S.: Fault detection in district heating substations. Applied Energy **157** (2015) 51–59
3. Xue, P., Zhou, Z., Fang, X., Chen, X., Liu, L., Liu, Y., Liu, J.: Fault detection and operation optimization in district heating substations based on data mining techniques. Applied Energy **205** (2017) 926–940
4. Abghari, S., Boeva, V., Lavesson, N., Grahn, H., Ickin, S., Gustafsson, J.: A minimum spanning tree clustering approach for outlier detection in event sequences. In: 17th IEEE Int'l Conf. on Machine Learning and Applications. (2018) 1123–1130
5. Hampel, F.R.: A general qualitative definition of robustness. The Annals of Mathematical Statistics (1971) 1887–1896
6. Kotsiantis, S., Kanellopoulos, D.: Discretization techniques: A recent survey. GESTS Int'l Transactions on Computer Science and Engineering **32**(1) (2006) 47–58
7. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: Proc. of the 17th Int'l Conf. on Data Engineering. (2001) 215–224
8. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814) (2007) 972–976
9. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. Volume 10. (1966) 707–710