Temporal Data Mining Approaches and Green Design Implementation for Data Center Chillers Management System

Wong Hui Shin
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.
hswong@mmu.edu.my

Dr. Ho Chin Kuan
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.
ckho@mmu.edu.my

Prof. Y.P.Singh
Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia.
y.p.singh@mmu.edu.my

Abstract: - Temporal data mining refers time series data points recorded along with the time which they were observed. Time series is a sequence of data points collected at equal time interval for each point. We apply temporal data mining approaches to data center. A data center chillers management system is studied, planned and designed. In this study, we explore the 2 modules of the system: 1) chillers fault detection module and 2) chillers lifetime prediction or time for maintenance module. We propose temporal data mining approaches to solve 1) change point detection and 2) motif mining problems in the said system. The temporal data mining approaches are important to predict the chillers fault, usage lifetime and time for maintenance. They are also the main objectives of this chillers management system. This paper also illustrates the architecture and the framework of the data center chillers management system. It aims to justify the data center chillers management system from the information technology perspective.

Key-Words: - temporal data mining, chance point detection, motif mining, cost optimization, equipments fault management and equipments usage lifetime and time for maintenance management.

1 Introduction

Data center consists IT Equipments (servers, storage, networking) which are fitted in racks and they are arranged in rows. A large data center may contain a maximum of thousand of racks in a few thousands square feet space. Figure 1 shows the typical cooling infrastructure in a data centre. The energy dissipated from the IT equipments needs to be cooling down by having the Computer Room Air Conditioning (CRAC) units. The CRAC colds the exhaust hot air from the IT Equipments racks. The numbers of chillers form the CRAC. The chillers' working condition is significant affecting the data center cooling process.

Time series is a collection of data points with equal interval time for each pair of data points. In our study, we use time series datasets which is collected from a 70,000 square feet data center with 2000 racks of IT equipment. The temporal data mining tasks includes temporal data characterization and comparison, temporal clustering analysis, temporal classification, temporal association rules, temporal pattern analysis, and temporal prediction and trend analysis.

Change point detection is used to detect the obvious sudden change of the magnitude from a sequence of stable valued data points. Change point detection is also use to monitor the room temperature of the data center. When there is a faulty or shut down status for any instrument, a change point is detected. The room temperature will change to much higher temperature.

Motif mining says mining the collection of patterns in an interval time called motif. The motif indicates the temperature routine in certain period of time. The motif is useful for room temperature monitoring and it is good to predict the optimal resources usage in a period of time.

The change point detection and motif mining are important to predict the equipments fault and devices lifetime or time for maintenance module. They help in budgeting the instruments turn over. Also, they are the interesting problems to be solved for operation cost deduction of data centre.

In this paper, first, we will discuss a data center chillers management system with temporal data mining approaches and green design implementation. We focus on 2 modules: 1) chillers fault detection module and 2) chillers lifetime prediction or time for maintenance module. We justify the 2 modules from information technology perspective, i.e. applying temporal data mining approaches to solve change point detection and motif mining problems.

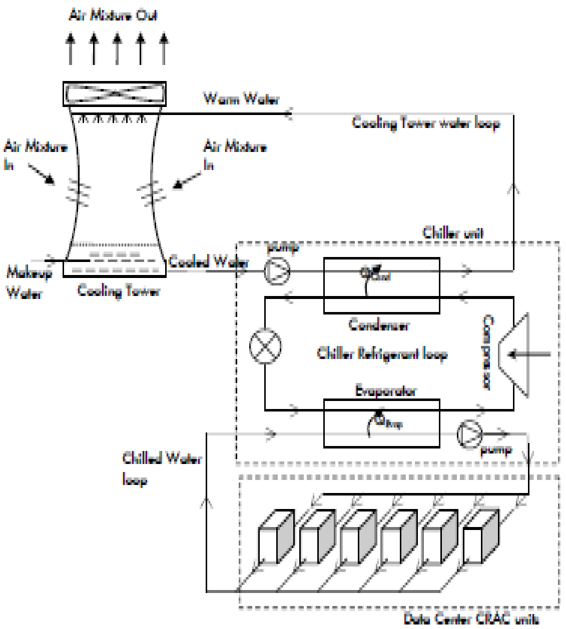

Fig 1: Typical Cooling Infrastructure in a Data Centre [18]

2 Data Center

Data center are rows of IT Equipment (servers, storage, networking) in racks. A data center with thousand of racks in several thousand of square feet space is considered as a large data center. Our study focus at large data center or high tier data center with chillers facilities. Chillers are the basic working units in Computer Room Air Conditioning (CRAC) units. The CRAC colds the hot air from the IT equipments racks. A CRA has many chillers. The numbers of chillers required in a data center depends on the size and thermal density of a data center. The chillers' working condition is significant affecting the thermal management for a data center.

3 Data Center Chillers Management System Overview

We propose data center chillers management system with data mining approaches and green design implementation. It contains 2 modules: 1) chillers fault detection module and 2) chillers lifetime prediction or time for maintenance module.

Chillers fault detection module identifies the chillers operation patterns for failure and predict chillers fault. It helps in doing chillers management contingency planning. It ensures no chillers operation shut down.

Chillers lifetime prediction or time for maintenance module detects the optimal operating period or duration for chillers. This module eases the data center operators in decision making for new chillers purchasing or replacement or time for maintenance. It aims to ensure no chillers operation shut down.

3.1 Information Technology Perspective

We use temporal data mining approaches and green design implementation to solve change point detection and motif mining problems which enables chillers fault detection and lifetime prediction or time for maintenance.

3.1.1 Temporal Data Mining

Temporal data mining is application of data mining techniques in time series datasets. Temporal data characterization and comparison, temporal clustering analysis, temporal classification, temporal association rules, temporal pattern analysis, and temporal prediction and trend analysis are temporal data mining properties.

Temporal association rule mining shows attributes value conditions that occur frequently in a given time series dataset. It finds associations relationships among large set of data items. It identifies collections of data attributes that are statistically related.

Classification, estimation, prediction, unsupervised clustering, and market basket analysis are data mining strategies [8]. Classification and estimation strategies generalize current outcome from building models to predict future outcome. Classification

ISBN: 978-1-61804-110-4

strategy has categorical output and the estimation strategy has a numeric outcome.

Temporal prediction, temporal classification, temporal clustering, temporal search and retrieval and temporal pattern discovery are the temporal data mining tasks [3]. Temporal prediction has to do with forecasting future values of time series based on its past projected values. A predictive model is build to forecast the next appearance of pattern in a sequence of symbols or the projected values from the existing values/ datasets. Some examples of temporal classification are speech recognition, gesture recognition and hand-writing recognition. Each new input of the sequence is identified by one of the finite classes in temporal classification. Temporal clustering is to group a collection of time series based on their similarity. Temporal clustering may indicate the clusters for navigation patterns of different user groups. The idea of temporal search and retrieval is to find the pattern matches in a time series.

3.1.2 Change Point Detection

In statistic framework, change point detection is investigated which including CUSUM (cumulative sum) [37] and GLR (generalized likelihood ratio) [38, 39]. Change point detection detects the change of the magnitude from a sequence of data points. Yashinobu and Masashi present a novel non-parametric approach to detecting the change of probability distributions of sequence data [3]. They provide a change-point detection algorithm based on direct density ratio estimation that can computed very efficiently in an online manner [3]. Change Point Detection discovers time points at which properties of time series data change [3]. The Change Point Detection indicates the adequate models to characterize chiller behavior and also the varying interpretations that are attachable to multivariate data [9, 10, 11]. Takeuchi and Yamanishi identified a change point when a significant change of statistical regularity of the access patterns has occurred or when a burst of outliers has emerged [16]. They created ChangeFinder, a unifying framework detects outliers and change point detection from non-stationary time series by a score system with two learning processes from AR models and SDAR algorithm; first learned for outliers detection and second learned for change point detection. Each data point on the basis of learned model is given a score [16]. The high scores for data points are identified as outlier [16]. The change points are detected from those data point scores which categorized as moving-averaged scores in time series [16]. Jaxk Reeves and members reviewed and compared many of the methods which proposed to detect undocumented changepoints in climate data series [17]. The methods reviewed and compared include the standard normal homogeneity (SNH) test, Wilcoxon’s nonparametric test, two-phase regression (TPR) procedures, inhomogeneity tests, information criteria procedures and various variants thereof [17]. The group showed the common trend TPR and Sawa’s Bayes criteria procedure seem optimal for most climate series, whereas the SNH procedure and its non-parametric variant are probably best when trend and periodic effects can be diminished by using homogenous reference series [17]. Wang, Zhang and Shin presented Change-Point Monitoring (CPM) which applied a nonparametric Cumulative Sum (CUSUM) to detect denial of service (DoS) attacks [20]. Yang, Dumont and Ansermino presented a change detection scheme based on adaptive Kalman filter, an Exponentially Weighted Moving Average (EWMA) predictor and CUSUM testing [21]. The presented scheme is compared with Trigg’s approach [21]. D’Angelo, Palhares, Takahashi and Loschi proposed a novel fuzzy/ Bayesian methodology for a change point detection in time series has been applied to solve the fault detection problem in dynamical systems [23]. The proposed methodology is based on two-step formulation; 1) a fuzzy clustering generates a transformed time series with beta distribution; 2) a Metropolis-Hastings algorithm is used to detect the probability of the change point in the transformed time series [23]. Singular-spectrum analysis (SSA) evaluates the degree of the changes between two consecutive sequences using the distance defined based on singular spectrums [40]. Subspace identification (SI) identifies a subspace in which time series sequences are constrained and evaluates the distance of target sequences from the subspace. A recursive subspace identification algorithm based on an instrumental variable method is used for estimating the subspace [41]. One-class support vector machine (ONE) interactively compares the support of probability densities estimated by one-class support vector machine from samples in the reference and test intervals [42]. Adams and Mackay introduced Bayesian Online Change Point Detection (BOCPD) algorithm [43]. Turner did an extending version of BOCPD [44, 45, 46].

3.1.3 Motif Mining

Motif mining is prescribed as mining the collection of repetitive patterns in an interval time called motif. HP Labs in Pala Alto and the Department of Computer Science at Virginia Tech identify regions of time series progression that demonstrate better/ improved sustainability measures than others. In order to understand multivariate time series data about chiller utilizations, they seek to identify motifs that underlay how different chillers are involved in meeting the varying demand posed by data centers [9, 10, 11]. The overall goal for Motif Mining is to do sustainability characterization for a diagnostic tool. Firstly, transduce the continuous multivariate stream into discrete symbol stream amendable for processing by episode mining algorithm. Secondly, perform a K-mean clustering on those vectors and use the cluster labels as symbols to encode the time series to single symbol sequence. Thirdly, do a run-length encoding of the sequence symbol and noting where transitions from one symbol to another occur. Frequent episode mining is now conducted over that sequence of transitions. The motif mining algorithm is deterministic and guarantees finding repeating patterns in the discrete domain with a given support threshold. Yu-Feng, Chun-Ping and Jun-Zhou proposed an algorithm to discover self-correlation of stock price in stock price sequences by firstly search for one part of the time series motif using ordinal comparison and k-NN clustering algorithm, and then attempt to discover the correlation between motifs and subsequences connected behind them [27]. The experimental results for the proposed algorithm are benchmarked with BP neural networks and association rules [27]. Keogh, Chakrabarti, Pazzani and Mehrotra introduced Piecewise Aggregate Approximation (PAA), and it is compared with other dimensionality reduction techniques such as singular Value Decomposition (SVD), the Discrete Fourier transform (DFT), and the Discrete Wavelets Transform (DWT) [31]. Patel, Keogh, Lin and Lonardi introduced an efficient algorithm which utilizes Euclidean metrics, Minkowski metrics and Dynamic Time Warping to discover the non-trivial definition of time series motif [32]. Chiu, Keogh and Lonardi introduce an algorithm inspired by recent advances in pattern discovery in biosequences to mine the motif in time series [33]. Castro and Azevedo proposed the Multiresolution Motif Discovery in Time Series algorithm, Mr.Motif [34]. They defined the motif discovery problem as an approximate Top-K frequent subsequences discovery problem [34]. They used the iSAX representation multiresolution capability to obtain motifs at different resolutions [34]. Mueen, Keogh, Zhu, Cash and Westover have introduced the first exact motif search algorithm which is faster than brute force search [35].

4 Research Methodology

This project uses Software Development Life Cycle (SDLC) for data center chillers management system's development. The waterfall approach works on Requirement Phase, Design Phase, Implementation Phase, Verification Phase and Maintenance Phase. In requirement phase and design phase, we have a few stages. Stage 1: Temporal data mining problem identification and stage 2: Temporal data mining problem verification, stage 3: Temporal data mining solution experiments and temporal data mining application in system development, stage 4: Temporal data mining solution testing. In temporal data mining problem identification, we identify 2 problems: 1) change point detection and 2) motif mining.

We plan to solve the change-point detection problem based on Bayesian machine learning, two-phase regression (TPR) procedures, adaptive Kalman filter, an Exponentially Weighted Moving Average (EWMA) predictor and CUSUM testing.

In order to solve the motif mining problem, firstly, transduce the continuous multivariate stream into discrete symbol stream amendable for processing by episode mining algorithm. Secondly, perform a K-mean clustering on available vectors and use cluster labels as symbols to encode the time series to single symbol sequence. Thirdly, do a run-length encoding of the sequence symbol and noting where transitions from one symbol to another occur.

5 Conclusion

A temporal data mining approaches and green design implementation for data center chillers management system is studied, planned and designed. In our study, we propose the said system with only focusing in 2 modules: 1) chillers fault detection module and 2) chillers lifetime prediction or time for maintenance module. We will incorporate predictive analytics techniques to solve the problems of the system: 1) change point detection and 2) motif mining. We are looking forward for a government grant to develop the system.

References:

[1] Cláudia M. Antunes , Arlindo L. Oliveira , "Temporal Data Mining: an overview", Lecture Notes in Computer Science.

[2] Weiqiang Lin, Mehmet A. Orgun, Graham J. Williams, "An Overview of Temporal Data Mining", The Australasian Data Mining Workshop (2002), Sydney, Australia, Pg. 83-90.

[3] Srivatsanlaxman and P. S. Sastry, "A survey of temporal data mining." Sadhana A vol. 31, Part 2, April 2006, pp. 173-198.

[4] The Sun Daily Malaysia news, "The 19 new Entry-Point Projects (EPPs) under the Economic Transformation Programme (ETP)" on 12nd January 2010.

[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, "Introduction To Data Mining", Addison Wesley, ISBN: 0-321-32136-7.

[6] Olivia Parr Rub, "Data Mining Cookbook", Wiley, ISBN: 0-471-38564-6

[7] Daniel T. Larose, "Data Mining Methods and Models", Wiley, ISBN: 0-471-66656-4

[8] Richard J. Roiger, Michael W. Geatz, "Data Mining: A Tutorial-Based Primer", Addison Wesley, ISBN: 0-321-32136-7

[9] Debprakash Patnaik, Manish Marwah, Ratnesh K. Sharma, and Naren Ramakrishnan, "Data Mining for Modeling Chiller Systems in Data Centers", 9th International Symposium on Advances in Intelligent Data Analysis, IDA 2010, Tucson, AZ, USA, May 19-21, 2010: 125-136.

[10] D. Patnaik, M. Marwah, R. K. Sharma, N. Ramakrishnan, "Temporal Data Mining Approaches for Sustainable Chiller Management in Data Centers", in ACM Transactions on Intelligent Systems and Technology (TIST), 2011 (In Press).

[11] Debprakash Patnaik, Manish Marwah, Ratnesh Sharma, Naren Ramakrishnan, "Sustainable operation and management of data center chillers using temporal data mining", Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 2009.

[12] Yoshinobu Kawahara, Masashi Sugiyama, "Change-Point Detection in Time-Series Data by Direct Density-Ratio Estimation", Proceedings of 2009 SIAM International Conference on Data Mining SDM2009, pp.389~400.

[13] Ramos, L. and Bianchini, R., "C-Oracle: Predictive thermal management for data centers", in Proc. of IEEE 14th International Symposium on High Performance Computer Architecture, pp.111-122, 2008.

[14] Sharma, R. K., Bash, C. E., Patel, C. D., Friedrich, R. J. and Chase, J. S., "Balance of Power: Dynamic Thermal Management for Internet Data Centers", IEEE Internet Computing archive, 9(1), pp.42-49, 2005.

[15] Filani, D., He, J., Gao, S., Rajappa, M, Kumar, A., Shah, R. and Nagappan, R, "Dynamic Data Center Power Management: Trends, Issues, and Solutions", white paper, Intel, 2008.

[16] Jun-ichi Takeushi, Kenji Yamanishi, "A unifying Framework for Detecting Outliers and Change Points from Time Series", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 4, April 2006.

[17] Jaxk Reeves, Jien Chen, XiaoLan L. Wang, Robert Lund, QiQi Lu, "A Review and Comparison of Changepoint Detection Techniques for Climate Data", Journal of Applied Meteorology and Climatology, Vol. 46, September 2006.

[18] Menne, J.M., and C. N. Williams, "Detection of undocumented changepoints using multiple test statistics and composite reference series", Journal of Climate, Vol. 18, pg. 4271-4286, 2005.

[19] Wang, X. L., "Detection of undocumented changepoints: A Revision of the Two-phase Regression Model", Journal of Climate, Vol. 16, pg. 3383-3385, 2003.

[20] H. Wang, D. Zhang, K. G. Shin, "Change Point Monitoring for the Detection of DOS Attacks", IEEE Transactions on Dependable and Secure Computing, Vol. 1, Issues 4, pg. 193-208, 2004.

[21] P. Yang, Dumont, G., Ansermino, J. M., "Adaptive Change Detection in Heart Rate Trend Monitoring in Anesthetized Chidren", IEEE Transactions on Biomedical Engineering, Vol. 53, Issues 11, pg. 2211-2219, November 2006.

[22] E. Rafajlowicz, M. Pawlak, A. Steland, "Nonparametric Sequential Change-Point Detection by a Vertically Trimmed Box Method", IEEE Transactions on Information Theory, Vol. 56, No. 7, pg. 3621-3634, July 2010.

[23] D'Angelo, M. F. S., Palhares, R,M., Takahashi, R.H.C., Loschi, R. H., "Fuzzy/ Bayesian Change Point Detection Approach to Incipient Fault Detection", IET Control Theory &

Applications, Vol. 5, No. 4, pg. 539-551, March 3, 2011.

[24] Tartakovsky, A.G., Rozovskii, B. L., Blazek, R. B., Hong Joong Kim, “A Novel Approach to Detection of Intrusions in Computer Networks via adaptive Sequential and Batch-sequential Change Point Detection Methods”, IEEE Transactions on signal Processing, Vol. 54, No. 9, September 2006.

[25] E. Keogh, K. Lin, & A. Fu, "HOT SAX: Finding the Most Unusual Time Series Subsequence: Algorithms and Applications". International Conference on Data Mining (ICDM), 2005.

[26] E. Keogh, S. Lonardi, W. Chiu, "Finding Surprising Patterns in a Time Series Database in Linear Time and Space”, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining KDD 02, 2002.

[27] Yu-Feng Jiang, Chun-Ping Li, Jun-Zhou Han, “Stock Temporal Prediction Based on Time Series Motifs”, Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[28] Davis Minnen, Charles Isbell, Irfan Essa, Thad Starner, “Detecting Subdimensional Motifs: An Efficient Algorithm for Generalized Multivariate Pattern Discovery”, Seventh IEEE International Conference on Data Mining, pg. 601-606, 2007.

[29] Pawan Nunthanid, Vit Niennattrakul, Chotirat Ann Ratanamahatana, “Discovery of Variable Length Time Series Motif”, the 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference, pg. 472-475, 2011.

[30] A. Mueen, E. Keogh, N. Bigdely-Shamlo, “Finding Time Series Motifs in Disk-Resident Data”, Ninth IEEE International Conference on Data Mining, pg. 367-376, 2009.

[31] E. Keogh, K. Chakrabarti, M. Pazzani, S. Mehrotra, “Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases”, Journal of Knowledge and Information Systems, Vol. 3, Issue 3, 2001.

[32] P. Patel, E. Keogh, J. Lin, S. Lonardi, “Mining Motifs in Massive Time Series Databases”, IEEE International Conference on Data Mining Proceedings , Pages: 370-377, 2002.

[33] B. Chiu, E. Keogh, S. Lonardi, “Probabilistic Discovery of Time Series Motifs”, Proceedings of KDD, pg. 493-498, 2003.

[34] N. Castro, P. Azevedo, “Multiresolution Motif Discovery in Time Series”, Proceedings of the SIAM International Conference on Data Mining SDM, pg. 665-676, 2010.

[35] A. Mueen, E. Keogh, Q. Zhu, S. Cash, B. Westover, “Exact Discovery of Time Series Motifs”, Proceeding of SIAM International Conference on Data Mining: SDM, 2009, pp. 1-12.

[36] P. G. Ferreira, P. J. Azevedo, C. G. Silva, R. M. M. Brito, “Mining Approximate Motifs in Time Series”, Proceedings of Discovery Science, pp.89-101, 2006.

[37] Basseville, M. and Nikiforov, V., Detection of Abrupt Changes: Theory and Application, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1993.

[38] Gustafsson, F., The Marginalized Likelihood Ratio Test for Detecting Abrupt Changes, IEEE Trans. On Automatic Control, 41(1): 66-78, 1996.

[39] Gustafsson, F., Adaptive Filtering and Change Detection, John Wiley & Sons Inc., 2000.

[40] Moskvina, V. and Zhigljavsky, A. A., An Algorithm Based on Singular-Spectrum Analysis for Change-Point Detection, Communication in Statistics: Simulation and Computation, 32(2): 319-352, 2003.

[41] Kawahara, Y., Yairi, T. and Machida, K., Change-Point Detection in Time-Series Data Based on Sub-space Identification, in Proc. Of the 7th IEEE Int’l Conf. on Data Mining (2007), pp. 559-564.

[42] Desobry, F., Davy, M. and Doncarli, C., An Online Kernel Change Detection Algorithm, IEEE Trans. On Signal Processing, 53(8): 2961-2974, 2005.

[43] Adam, R. P., Mackay, D.J.C., Bayesian Online Change Point detection, Technical Report, University of Cambridge, Cambridge, UK., (2007).

[44] Saatci, Y., Turner, R., Rasmussen, C.E., Gaussian Process Change Point Models, Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel.

[45] Turner, R., Saatci, Y., Rasmussen, C.E., Adaptive Sequential Bayesian Change Point Detection, Temporal Segmentation Workshop at NIPS 2009, Whistler, BC, Canada, 2009.

[46] Turner, R. And Rasmussen, C.E, Supervised Bayesian Online Change Point Detection, 2010.

ISBN: 978-1-61804-110-4