[Extended Abstract] An Overview of Methods for Detecting Contexts in Workload Data

Dusan Okanovic
dusan.okanovic@novatec-gmbh.de
Novatec Consulting GmbH, Germany

Thomas Sievering
thomas.sievering@novatec-gmbh.de
Novatec Consulting GmbH, Germany

Context

The variation in workload intensity of modern enterprise applications has to be taken into account when executing load-tests [1]. There are many contexts [2], e.g., special offers, public events, or weather, which can cause the workload to change. Common examples are Black Friday and Cyber Monday - in 2019., e-commerce websites registered an increase of 137% and 112%, for these days retrospectively, in comparison to normal traffic [3]. To allow for the reduction of testing time and to test only for special scenarios that are of interest, as a part of the ContinuITy project [4], we aim to automatically generate load tests tailored to these specific scenarios.

Motivation

However, to be able to create such a load-test, we usually rely on manual detection and labeling of the workload data with contextual information. We propose to use time series segmentation and change point detection methods [5] to support the process. The benefit of splitting the time series into subsequences also goes beyond context labeling - it also helps with many different tasks like anomaly detection and forecasting. In this talk, we present an overview of these methods, with the aim to help in deciding when to choose which method, e.g., depending on the type of time series.

Approach

The detection comprises several steps. First, we use Piecewise Aggregate Approximation (PAA) and Symbolic Aggregate Approximation (SAX) for time series decomposition [6] to detect trends and seasonalities, as well as to reduce the dimensionality. Also, we compare different ways to measure the similarity between subsequences of time series and use those distances to apply clustering methods [7]. Those clusters can provide insights on the segments by, e.g., detecting consecutive days that are in the same cluster. We also try using smaller clusters to detect anomalous days. Furthermore, we use Matrix Profile [9] to discover Motifs, Discords, and Changepoints. To detect a change of context, we use different search methods like PELT, Binary Segmentation, Bottom-Up, and Window-based change point detection [8]. Some of these methods require to set the number of changepoints to detect beforehand, some require a threshold value. We explore how to determine those parameters and when to use which method. The last step is to analyze the detected segments and, e.g., if two segments may be a part of the same context but were interrupted by another segment or event, we can cluster them together.

Evaluation

For evaluation, we compare detected segments and anomalies with existing labeled datasets. The results of those tests are compared and the goal is to determine which methods work best on which type of time series data and why.

References

[1] Zhen Ming Jiang and Ahmed E. Hassan. A Survey on Load Testing of Large-Scale Software Systems. IEEE Transactions on Software Engineering, vol. 41, no. 11, pp. 1091-1118. 2015.

[2] David Arthur and Sergei Vassilvitskii. K-Means++: The Advantages of Careful Seeding. In Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2007). pp. 1027–1035. 2007.

[3] Liron Hakim Bobrov. The Numbers Are In! Black Friday and Cyber Monday. https://www.similarweb.com/corp/blog/the-numbers-are-in-2019-black-friday-cyber-monday/. 2019.

[4] Henning Schulz, Tobias Angerstein, and André van Hoorn. Towards Automating Representative Load Testing in Continuous Software Engineering. In Companion of the 9th ACM/SPEC International Conference on PerformanceEngineering (ICPE). pp. 123–126. 2018.

[5] Aminikhanghahi, S., Cook, D.J. A Survey of Methods for Time Series Change Point Detection. Knowledge and Information Systems, vol. 51. pp. 339–367. 2017.

[6] Eamonn Keogh, Jessica Lin, and Ada Fu. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In Proceedings of the Fifth IEEE

International Conference on Data Mining (ICDM '05). pp. 226–233. 2005.
[7] Saeed Aghabozorgi, Ali Seyed Shirkhorshidi, and Teh Ying Wah. Time-series Clustering - A Decade Review. Information Systems, vol. 53, no. C. pp. 16–38. 2015.
[8] Charles Truong, Laurent Oudre, Nicolas Vayatis. Selective Review of Offline Change Point Detection Methods. Signal Processing. pp. 167:107-299. 2020.
[9] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Zachary Zimmerman, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Mining and Knowledge Discovery, vol. 32, no. 1. pp. 83–123. 2018.