Application of Change Point Detection to System Logs for Fault Detection

Yu Komohara, Satoru Kobayashi, Hideya Ochiai, Hiroshi Esaki
The University of Tokyo
Tokyo, 113-8656, Japan
{como, sat}@hongo.wide.ad.jp, ochiai@elab.ic.i.u-tokyo.ac.jp, hiroshi@wide.ad.jp

ABSTRACT

Once a trouble occurs in network systems, system operators have to detect it rapidly. System-fault investigation largely depends on system logs generated by machines in network. Log messages contain a lot of useful data such as hosts, timestamp, and how programs run. However, due to the enormous volume of system logs, it is impossible to read all system logs. Thus, operators cannot find useful information from them. Automated system log analysis can improve such a hard situation for system operators. In this research, we propose a method to evaluate the change of system automatically by analyzing hidden information, change points. We applied a change point detection method for system log. In addition, we set a threshold for change point score and we evaluated the change of system by the number of logs the change point score of which exceeded the threshold. We made an experiment using system logs of servers in a research institution, and we succeeded to find large change points related to system behavior.

1. INTRODUCTION

Due to the expansion of Internet, the number and complexity of machines which consists of network has increased rapidly. In this situation, the amount logs generated by network machines are very large. When a trouble occurs in network systems, operators solve the problem reading System logs. System log is generated by programs running in network systems. The log messages have much useful information for trouble-shooting such as hosts, timestamps, and the detail of programs behavior. However, the amount of system logs is too large for operators to read. Thus, many researches are trying to avoid the situation by automated analysis for system logs. When something changes in network systems, features of log messages also changes. Thus, operators can find what happens in network systems by monitoring features of log messages. There are a lot of features in log messages. In this study, we focus on the frequency of log messages. Although, the frequency of log messages is hidden information of system logs. It is difficult for operators to notice the change of it. Therefore, we apply change point detection to system logs and we find change points of network systems. Such information enables us to detect troubles rapidly.

2. RELATED WORK

One of the approaches of automated system log analysis is improving readability of log messages. Qui et al. [1] analyzed syslog messages of routers. Router configs contain the most of the location information in router syslog messages. For this reason, router syslog messages are analyzed with router configs. And the authors developed a system that groups enormous volume of syslog messages into a small number of meaningful network events.

Other approach is to find useful information for trouble-shooting from log messages. To extract useful information, Change point detection is used in some researches. Change point detection is the method to find essential changes in the long-term data series. Barry et al. [2] developed Bayesian Change Point detection. Bayesian Change Point detection find an underlying sequence of parameters partitioned into contiguous blocks of equal parameter values and the beginning of each block is defined as a change point. Chen et al. [3] implemented the system which infer the causes of performance problems. They applied BCP to tcp request latency data to strengthen effectiveness.

Yamanishi et al. [4] developed ChangeFinder, that distinguish change points from outliers using two-stage learning of time series models. They analyzed the frequency of connection requirements by ChangeFinder to detect MS.Blast worm. However, change point detection can be used to analyze general time-series data. For this reason, much useful information can be gained by using change point detection for the whole log data in network system.

3. METHODOLOGY

3.1 Change Point Detection

We use change point detection to find hidden information in system logs. As mentioned in Section 2, there are some change point detection method. In this study, we use ChangeFinder [4]. ChangeFinder distinguishes significant change points from outliers. In addition, it is an online machine learning method. Thus, we are able to implement real time analysis system.

However, Simple usage of ChangeFinder fails in some cases. There are logs that have cyclic ups and downs at a prescribed period. ChangeFinder may detect many change points from such log messages. For this reason, before applying the change point detection for system logs, we have to adjust ChangeFinder to system logs.

Appropriate parameters depend on the kind of time series data. In this study, we tuned parameters and ChangeFinder does not find change points which appear every day.

3.2 Analysis Flow

Fig.1 shows the analysis flow. First, we estimate message templates of system logs. We use the incremental learning method proposed by Mizutani [5]. Second, we count the frequency of log messages for each templates. When an event happens in network system, the frequency of log messages related to the event become high. This is because we focus on the frequency of log messages. Third, we apply change point detection to the frequency of log templates by ChangeFinder. Yamanishi [4] analyzed a time series of frequency of connection requirements to specific port by ChangeFinder. However, by using ChangeFinder to all frequency data of log templates, we can gain much useful information about what happens in network system. Finally, we evaluate the change of network systems by the number of groups the change point score of which exceed a threshold.

User-defined parameters are as follows: 1) A period to total frequency is user-defined parameter 2) parameters of ChangeFinder 3) Threshold of change point score.

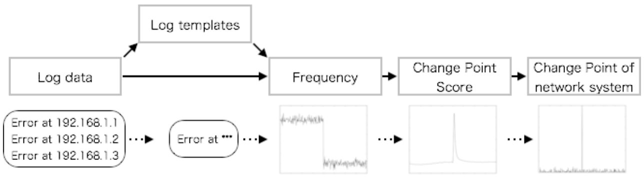

Figure 1: Analysis flow

4. EVALUATION

We apply our implementation for system logs of 15 servers in a research institution. The data was collected for 1 week. The log messages consists of 933,240 lines and 1,302 message templates. We generate a sequence of time-series data of the frequency of log templates in every 10 minutes for the change point detection input. The parameters of ChangeFinder are as follows. Discounting parameter value r is 0.01 and a fixed length of the window T is 5. Fig.2 show the result, where the horizontal axis shows time and the vertical axis shows the number of log templates which have high change point scores.

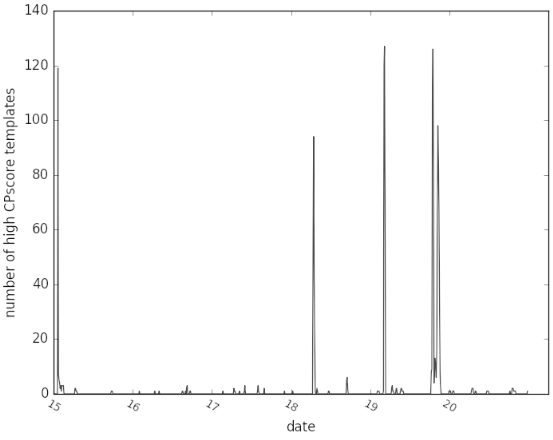

Figure 2: The number of log templates which have high change point score

There are many small change points and 3 large change points. Before these large change points, some events related to servers has occurred.

Machine reboot happened before the first change point and Automatic Updates of OS happened before the second change point. With regard to the third change point, unexpected shutdown happened before the change point.

5. CONCLUSION

In this paper, we proposed the method to evaluate the change of network systems. To investigate the change, we used hidden information of system logs, the frequency of log templates. We applied our implementation to the system logs of servers in a research institution. As a result, we found the large change points from system logs and investigated what happened at the change points. For the future works, we will consider source machines that generated system logs. In addition, it is also useful to examine which log templates show high change point score. Based on semantics of log messages, we will make a further investigation into the change points. Additionally, we implement real time analysis system.

6. REFERENCES

[1] T. Qui, ”What Happened in my Network ? Mining Network Events from Router Syslogs” IMC ’10 Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 472-484 , 2010.

[2] D. Barry and J. A. Hartigan, “A bayesian analysis for change point problems,” Journal of the American Statistical Association, vol. 88, no. 421, pp. 309 – 319, 1993.

[3] Pengfei Chen, Yong Qi, Pengfei Zheng, Di Hou, ”CauseInfer:Automatic and Distributed Performance Diagnosis with Hierarchical Causality Graph in Large Distributed Systems” IEEE INFOCOM 2014, pages 1887-1895, 2014.

[4] Jun-ichi Takeuchi, Kenji Yamanishi, ”A Unifying Framework for Detecting Outliers and Change Points from Time Series”, IEEE Transactions on Knowledge & Data Engineering, vol.18, no. 4, pp. 482-492, April 2006.

[5] M. Mizutani, “Incremental Mining of System Log Format” Services Computing (SCC), 2013 IEEE International Conference, pages 595-602, 2013.