A Data Mining Friendly Anonymization Scheme for System Logs using Distance Mapping

Gabriela Limonta and Yoan Miche[0000−0001−8864−2312]

Nokia Bell Labs Finland
firstname.lastname@nokia-bell-labs.com

Abstract. In this document, we investigate the use of Distance Mapping ideas and what they enable for system logs. In the field of telecommunication networks, log files are used for service quality assurance, and are coming from various devices, back end systems, and general usage of the network. Typically, these log files are not allowed to be monetized or shared with third parties, thanks to legal restrictions on privacy issues. While there are some existing early solutions to this, such as Differential Privacy or Homomorphic Encryption, we propose here to look at Distance Mapping to transform the raw data (the system log files) into highly usable but anonymized data. The resulting data can be used directly by Machine Learning algorithms, visualization algorithms, or be considered for re-embedding. While this approach transforms the data format significantly and limits its usage only for distance-based data mining and machine learning tools, it is an elegant and computationally feasible methodology for such applications.

1 Motivation

Logs are generated by applications to record important information and events during the runtime of a system. Automated log analysis has been an active area of research [5,8] focusing on how to use logs effectively. These event logs can be used in combination with machine learning techniques for different purposes, including forensic analysis [2] and anomaly detection [1].

5G telecommunications enable increased connectivity and the deployment of diverse Internet of Things (IoT) applications. By combining logs from different sources of a distributed system, we can build an overview of its state, which provides situational awareness when investigating an incident. Logs play a crucial role in analyzing the state of both 5G infrastructure [6] and the IoT applications built on top of it [4].

Log files can also be used for data mining. For example, telecommunications operators hold large amounts of log data, which often includes information related to the traffic, location and movement of connected mobile equipment. This data can be monetized by mining it internally to build better services or by selling it to external parties for commercial purposes. However, the latter is often not possible due to regulations on privacy and data protection rules and the risks of re-identification of users or systems.

2 Problem and Proposed Solution

To use the contents of log files in machine learning, data is transformed typically to a format that can be used by an algorithm, i.e. numerical data, (often arbitrarily mapping values to integers). This raises issues, as detailed in [3], and falls short with text fields as in log files. We propose to use the technique in [3], named Distance Mapping, followed by neighbor re-embedding [7].

The idea in Distance Mapping is to create a function that maps probabilistically the distances between elements in the original metric space $(\mathbb{X}, d)$ (with $\mathbb{X}$ a set of values, and d a distance function over that set) to distances in a well-known space, such as the canonical Euclidean $(\mathbb{R}, d_{\texttt{EUC}})$: Distances between log entries (across all fields: timestamp, message,...) are mapped to distances between points in a Euclidean space that are as likely (probabilistically) to be at that distance. The goal is to preserve the structure of the data, specifically distance-based, such as density, separation of clusters and classes,...

With pairwise distances between log entries mapped, the mapped distances can be used by distance based machine learning, e.g. K-Means. Some techniques do not use directly distances, so we perform re-embedding of the data into another space [3]: We convert pairwise distances between points to a set of points that respects those pairwise distances. For outlier detection, visualization, and machine learning, re-embedding the mapped distances into $\mathbb{R}^d$ is ideal. For privacy, and having data of the same format as the original, we can decide to re-embed the mapped distances into the original space. We obtain a data set of the same format as the original, with log entries that relate to each other (in terms of pairwise distances) as in the original data, and yet none of the entries are the same as in the original data.

We theorize that this last approach allows for generating synthetic data that preserves (statistically) pairwise distances between original data points, and the structure of the data (understood here as based on distances).

References

1. Du, M., Li, F., Zheng, G., Srikumar, V.: DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning. In: Proc. of the 2017 ACM SIGSAC Conference. pp. 1285–1298. ACM, Dallas Texas USA (Oct 2017)
2. Kahles, J., Törrönen, J., Huuhtanen, T., Jung, A.: Automating Root Cause Analysis via Machine Learning in Agile Software Testing Environments. In: 12th IEEE Conf. on Software Testing, Validation and Verification (ICST). pp. 379–390 (Apr 2019)
3. Miche, Y., Ren, W., Oliver, I., Holtmanns, S., Lendasse, A.: A framework for privacy quantification: measuring the impact of privacy techniques through mutual information, distance mapping, and machine learning. Cognitive Computation **11**(2), 241–261 (2019)
4. Noura, H.N., Salman, O., Chehab, A., Couturier, R.: DistLog: A distributed logging scheme for IoT forensics. Ad Hoc Networks **98**, 102061 (Mar 2020)
5. Oliner, A., Stearley, J.: What Supercomputers Say: A Study of Five System Logs. In: 37th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN'07). pp. 575–584 (Jun 2007)

6. Sundqvist, T., Bhuyan, M.H., Forsman, J., Elmroth, E.: Boosted Ensemble Learning for Anomaly Detection in 5G RAN. In: Artificial Intelligence Applications and Innovations. pp. 15–30. Springer, Cham (2020)
7. Yang, Z., Peltonen, J., Kaski, S.: Scalable Optimization of Neighbor Embedding for Visualization. In: International Conference on Machine Learning. pp. 127–135. PMLR (May 2013)
8. Zhu, J., He, S., Liu, J., He, P., Xie, Q., Zheng, Z., Lyu, M.R.: Tools and Benchmarks for Automated Log Parsing. In: IEEE/ACM 41st Int. Conf. on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 121–130 (May 2019)