Constructing Large-Scale Real-World Benchmark Datasets for AIOps

Zeyan Li\textsuperscript{1}, Nengwen Zhao\textsuperscript{1}, Shenglin Zhang\textsuperscript{2}, Yongqian Sun\textsuperscript{2}
Pengfei Chen\textsuperscript{3}, Xidao Wen\textsuperscript{1}, Minghua Ma\textsuperscript{4}, Dan Pei\textsuperscript{1}
\textsuperscript{1}Tsinghua University, \textsuperscript{2}Nanjing University, \textsuperscript{3}Sun Yat-sen University, \textsuperscript{4}Microsoft Research

\section*{ABSTRACT}
Recently, AIOps (Artificial Intelligence for IT Operations) has been well studied in academia and industry to enable automated and effective software service management. Plenty of efforts have been dedicated to AIOps, including anomaly detection, root cause localization, incident management, etc. However, most existing works are evaluated on private datasets, so their generality and real performance cannot be guaranteed. The lack of public large-scale real-world datasets has prevented researchers and engineers from enhancing the development of AIOps. To tackle this dilemma, in this work, we introduce three public real-world large-scale datasets about AIOps, mainly aiming at KPI anomaly detection, root cause localization on multi-dimensional data, and failure discovery and diagnosis. More importantly, we held three competitions in 2018/2019/2020 based on these datasets, attracting thousands of teams to participate. In the future, we will continue to publish more datasets and hold competitions to promote the development of AIOps further.

\section{1 INTRODUCTION}
In recent years, traditional manual operations and maintenance have gradually become time-consuming and error-prone due to the increasing complexity of modern IT service architectures. To tackle this dilemma, Artificial Intelligence for IT Operations (AIOps) has been proposed to relieve some manual intervention required. Gartner [35] defines AIOps as the application of machine learning (ML) and data mining to IT operations. AIOps combine big data and ML functionality to enhance and partially replace traditional manual efforts, including performance monitoring, anomaly detection, event correlation and analysis, incident management, etc.

In the literature, tremendous efforts have been devoted to the research of AIOps, including anomaly detection [7, 29, 34, 43], root cause localization [5, 6, 18, 19, 24, 25, 28, 33, 37, 40, 41, 47], failure prediction [11, 27, 39, 44, 46, 49], incident management [9, 10, 12, 31], resource management [14, 15] and so on. However, due to the lack of public large-scale real-world datasets, it is difficult to compare these approaches fairly. Although these works achieve promising performance on their respective private datasets, these approaches could perform unstably in the real world. It is also difficult to transfer an approach to real generic application due to the various nature of different target systems for AIOps. Thus, AIOps is at the crossing point between Artificial Intelligence research and operation engineers, which usually requires the cooperation of academic researchers and industry to obtain real data and scenarios and carry out research on this basis. However, it is challenging for every researcher to find industrial partners since the monitoring data used for AIOps are highly confidential. Therefore, the lack of public large-scale real-world datasets is one significant limitation of AIOps research and application.

Although there exist several public datasets, these datasets suffer from several limitations. First, the majority of existing datasets covers only one AIOps scenario, e.g., Yahoo Benchmark [21] for KPI anomaly detection and [51] for faulty microservice localization. However, AIOps involves diverse scenarios, for example, KPI anomaly detection, log anomaly detection, capacity planning, root cause localization, etc. It is unsatisfactory to evaluate comprehensive AIOps approaches separately on such scenario-specific datasets. Second, the scale of the dataset is limited. For example, Yahoo Benchmark for KPI anomaly detection only includes hundreds of data points. In general, real-world service systems tend to support numerous users and perform complex architecture. Thus, the dataset should be large and diverse to be close to the real world. Third, some datasets are synthetic, not from the real world. Such datasets cannot reflect the real system behavior, which could incur risk for real applications. As a result, rich, large-scale, and real-world datasets to assist the research of AIOps are highly desired.

To promote the development of AIOps, we have published three large-scale and real-world datasets, which are based on production systems of our industrial partner (e.g., Sougo, eBay, Tencent, Suning, and China Mobile Zhejiang). These datasets are for KPI anomaly detection, root cause localization for multi-dimension data, and failure discovery and diagnosis, which is a comprehensive scenario, respectively. There have been many research works [8, 13, 19, 23, 24, 38, 45, 50] using our datasets. Furthermore, based on these datasets, we have held an AIOps algorithm competition in 2018/2019/2020, attracting 125/141/141 teams and hundreds of participants\textsuperscript{1}. The solutions of top teams are also shared publicly [4]. We believe such datasets would benefit the AIOps research and practice, as ImageNet [16] did to the fields of computer vision and deep learning. We would continue constructing and publishing new datasets and holding AIOps algorithm competitions once a year.

The main contributions of our work are summarized as follows.

\begin{itemize}
\item Up to now, we have published three large-scale real-world datasets \cite{1–3}, including KPI anomaly detection, root cause localization for multi-dimensional data, and failure discovery and diagnosis. These datasets have been used by many research works.
\item We hold the yearly AIOps algorithm challenge, which involves thousands of researchers and engineers and significantly promotes this community’s development.
\end{itemize}

\section{2 OUR WORK}
Following existing work \cite{36}, we categorize the research of AIOps into two macro-areas, failure management and resource management. Failure management aims to reduce the negative impact incurred by failures so as to ensure service availability and user

\textsuperscript{1}\url{https://competition.aiops-challenge.com/home/competition}
experience. Resource management aims to allocate and save various resources (e.g., power, VMs) in IT service management. We currently mainly focus on failure management, and resource management could be our future work.

As presented in Fig. 1, the goals of failure discovery, failure diagnosis, and failure mitigation are to reduce the mean time to detect (MTTD), to reduce the mean time to recover (MTTR), and to increase the mean time between failures (MTBF), respectively. In our work, to tackle the bottleneck of AIoPS research, motivated by ImageNet [16], we aim to construct large-scale real-world benchmark datasets for AIoPS. We first focus on failure management and have published three related datasets, including KPI anomaly detection (Section 3), root-cause localization for multi-dimensional data (Section 4), and failure discovery and diagnosis (Section 5). We will introduce each dataset and related background in detail in the following three sections.

3 KPI ANOMALY DETECTION

In this section, we first introduce the background of KPI anomaly detection. Then we present the details of our published dataset to enhance the research of KPI anomaly detection.

3.1 Background

To closely monitor service running status and identify system failures, engineers tend to continuously collect many key performance indicators (KPIs) and identify anomalies from KPIs. Accurate KPI anomaly detection can trigger prompt troubleshooting, help to avoid economic loss, and maintain user experience [29, 38, 43, 48].

In recent years, tremendous efforts have been dedicated to KPI anomaly detection. Existing works could be classified into three categories, i.e., supervised methods (e.g., Opprentice [29] and EDAGS [21]), unsupervised methods (e.g., Donut [43]), and semi-supervised methods (e.g., ADS [7]). However, there exist several significant challenges, restricting the research of KPI anomaly detection [38].

- Various KPI patterns. Based on our experience and observation, KPI patterns could be classified into three categories: seasonal (e.g., transaction volume), stable (e.g., success rate), and unstable (e.g., CPU utilization). Therefore, a successful approach should be generalized enough for different patterns.
- Diverse KPI anomaly patterns. The definition of KPI anomalies is the behavior violating normal patterns, like spikes, dips, and jitters. Besides, different scenarios and components have different levels of tolerance to anomalies. Thus, the designed approaches should have the ability to detect diverse anomaly patterns.
- Lack of sufficient labeled data. On the one hand, the number of KPIs is huge in large-scale systems. On the other hand, labeling anomalies requires experienced engineers with domain knowledge [50]. Thus, it is challenging and time-consuming to obtain sufficient labeled data to construct a supervised model and evaluate performance.

Although there exist rich KPI anomaly detection works, these approaches are mainly evaluated on their private datasets. Besides, there are no large-scale real-world KPI anomaly detection datasets currently for model training and evaluation. Thus, despite the promising performance of their papers, the performance of real deployment in the industry is far from satisfying. Although there exist several public datasets about KPI anomaly detection like Yahoo Benchmark [21] and Numenta Anomaly Benchmark [22], they only contain KPIs with limited data points and synthetic anomalies. In comparison, the number and type of KPIs are large and diverse. Consequently, a large-scale and diverse KPI anomaly dataset is highly desired in the community of KPI anomaly detection.

3.2 Dataset

To tackle the above challenges, we constructed a large-scale real-world KPI anomaly detection dataset (named A) with sufficient labels, covering various KPI patterns and anomaly patterns. This dataset is collected from five large Internet companies (Sougo, eBay, Baidu, Tencent, and Ali). Specifically, dataset A includes 27 KPIs in total. For each KPI, experienced engineers in these companies manually labeled anomalies carefully. Table 1 shows some statistical information about this dataset, including the total number of data points, anomaly ratio and time span. We could observe that the time span of each KPI is from two months to seven months, which is large-scale compared with existing datasets like Yahoo [21]. Figure 2 presents five KPIs within a time span of one week.

![Figure 1: Pipeline and key measurements in failure management](image)

Table 1: Overview of KPI anomaly detection dataset A

| KPI ID | #Points | Anomaly ratio | Time span (#days) |
|-------|---------|---------------|------------------|
| ...   | 294019  | 0.19%         | 207              |
| ...   | 295361  | 0.19%         | 207              |
| ...   | 295258  | 0.36%         | 207              |
| ...   | 241148  | 3.17%         | 183              |
| ...   | 294048  | 1.04%         | 207              |

According to our experience and observation, the KPI patterns in this dataset are indeed various.

Based on this dataset, we held the first AIoPS algorithm challenge in 2018\(^2\), attracting 125 teams and over 300 participants from both academia and industry. Besides, motivated by the novel evaluation

\(^2\)https://competition.aiops-challenge.com/home/competition/148445227200032281
strategy of KPI anomaly detection proposed in [43], we import alarm delay restriction and design a novel evaluation method used in the competition\(^3\). This dataset and evaluation method have been widely used in many research works [13, 38, 50]. The best F1-score achieved in this competition is 0.8216. The solutions proposed by the top teams are shared in public [4].

4 ROOT CAUSE LOCALIZATION FOR MULTI-DIMENSIONAL DATA

In this section, we first introduce the background of root cause localization for multi-dimensional data. Then we introduce our published large-scale dataset for this problem.

4.1 Background

To closely monitor the events occurring in a system, various types of structured logs are generated, carrying critical information for failure diagnosis. The structured logs record events with many fields describing the events. For example, an online shopping platform could generate structured logs recording orders. As shown in Table 2, the field Timestamp describes when the order occurs, Dollar Amount describes the dollar amount of the order, Province describes the location of the corresponding customer, and ISP describes the network that the customer used. Engineers would like to collect the overall value of some fields, such as the total dollar amount, to monitor the status of the whole system. We call such fields measures [6, 24]. Other fields characterize the events and we call them attributes [6, 18, 24, 28]. By grouping original structured logs by some attributes and aggregating values of some fields, multi-dimensional data [24] are generated. For example, to generate the multi-dimensional data in Table 3, we group structured logs like those in Table 2 by Timestamp, Province, and ISP, and aggregate values of Dollar Amount. Note that in Table 3, we only present the top teams are shared in public [4].

It indicates a failure occurring in the system when the overall value of a measure becomes abnormal (e.g., the total dollar amount decreases compared with the expected normal value). However, in practice, a failure usually only causes the measure values under specific attribute combinations [18] abnormal [24]. An attribute combination refers to a slice of the whole multi-dimensional data by selecting records with specified values of some attributes, and it has been formally defined in existing literature [18]. For example, in Table 3, only the measure values under \{Province=Beijing\} are significantly abnormal. Such attribute combinations show the scope of the failure, and thus, are substantial clues for failure diagnosis. In Table 3, for example, by localizing \{Province=Beijing\}, engineers get to know that the failure is related to the network or servers for users in Beijing, which directs further diagnosis. Therefore, we call such attribute combinations as root causes of multi-dimensional data.

| Table 2: Example structured logs |
|----------------------------------|
| Order ID | Timestamp | Dollar Amount | Province | ISP         |
|----------|-----------|---------------|----------|-------------|
| T001     | 2021.11.11 10:00:01 | $16 | Beijing | China Mobile |
| T002     | 2021.11.11 10:00:05 | $21 | Beijing | China Unicom |

| Table 3: Example multi-dimensional data |
|------------------------------------------|
| Province | ISP | Dollar Amount | Expected Dollar Amount |
|----------|-----|---------------|------------------------|
| Beijing  | China Mobile | 5   | 10         |
| Beijing  | China Unicom | 10  | 20         |
| Shanghai | China Unicom | 30  | 31         |
| Guangdong| China Mobile | 10  | 9.8        |
| Zhejiang | China Unicom | 2   | 2          |
| Guangdong| China Unicom | 200 | 210        |
| Shanxi   | China Unicom | 20  | 22         |
| Jiangsu  | China Unicom | 200 | 203        |
| Tianjin  | China Mobile | 41  | 43         |
| Total    |    | 518           | 550.8                  |

4.2 Dataset

Most existing works [5, 6, 18, 19, 25, 28, 37, 40, 41, 47] on root cause localization for multi-dimensional data use their private datasets for evaluation. Therefore, to enhance the research of root cause localization for multi-dimensional data, we construct and publish a dataset [2], named B, based on the real-world large online shopping platform of Suning. Ltd.

In dataset B, there are five attributes, which are named as i.e., p and 1 for confidential reasons. The measure of B is the average number of order per minute. In total, the dataset B contains 400 synthetic failures with ground-truth root cause labels, spanning eight weeks. We use synthetic failures rather than real-world failures because it is hard to collect hundreds of production failures with ground-truth root causes. The synthesis process of failures contains six steps.

1. We first select a time point for anomaly injection where the overall measure values are smooth.
2. We randomly choose \( m \in \{1, 2, 3, 4\} \) different attributes out of the five attributes.
3. Then we randomly choose \( n \in \{1, 2, 3, 4\} \) different attribute combinations corresponding to the \( m \) selected attributes. These \( n \) selected attribute combinations are the ground-truth root cause where we are going to inject anomalies.

\(^3\)https://github.com/iopsai/iops/tree/master/evaluation
(4) For each selected attribute combination, we modify the measure values of it, i.e., the measure values of those rows at the select time point satisfying the attribute combination. The modified values are calculated by generalized ripple effect (GRE) [24] with a random magnitude.

(5) We add extra noises, which are generated by Gaussian distribution, on both modified rows and other rows to simulate the cases where GRE does not work perfectly and there are forecast residuals, respectively.

(6) We check the synthetic failure with the following conditions. When anyone of them are triggered, we go back to step 2 to regenerate a failure at the time point.
   - There is otherwise attribute combinations sharing almost the same rows with the selected ground-truth attribute combinations.
   - The injected noises is so large that the overall measure value of unmodified records (in step 3) is also abnormal.

Overall, among the 400 synthetic faults, 144/131/89/36 faults contains 1/2/3/4 root-cause attribute combinations, and 96/136/108/60 faults involve 1/2/3/4 attributes.

Based on the dataset B, we held the second AIOps challenge in 2019. There are 141 teams and 547 participants from both academia (39%) and industries (61%) in this competition. The best performance achieved by them in this competition is 0.9593 with respect to F1-score. The solutions proposed by the top teams are published in [4].

There are several existing research works [19, 24] utilizing the data provided by AIOps Challenge 2019. Note that the datasets \( B_0 \sim B_4 \) used and published by Squeeze [24] are based on exactly the same original data as the dataset \( B \) described above and generated by almost the same synthesis process.

5 FAILURE DISCOVERY AND DIAGNOSIS

In Section 3 and Section 4, we introduce our datasets on KPI anomaly detection and root cause localization for multi-dimensional data. These datasets are useful to benchmark failure discovery and failure diagnosis algorithms respectively. However, on the one hand, engineers require automatic failure discovery and diagnosis in an end-to-end manner. On the other hand, root cause localization for multi-dimensional data provides only clues to the exact root causes, and it still requires further diagnosis to obtain fine-grained root causes given such clues. Thus, we construct a dataset for benchmarking both failure discovery and fine-grained failure diagnosis approaches. In this section, we first introduce the background of the problem. Then we introduce the details of our dataset.

5.1 Background

A distributed system is composed of many cooperating components, such as load balancers, web servers, application servers, and databases. For example, in Fig. 3, we present the architecture of a production distributed system (named \( \mathcal{S} \)) at a major Internet service provider, China Mobile Zhejiang. In Fig. 3, each vertex is a component. These components are of five classes, including OSB (Oracle Service Bus), service, DB, Docker, and OS. To monitor the status of the components, engineers collect various KPIs of these components, such as queries per minute, success rate, and average response time.

By KPI anomaly detection on these KPIs, we can discover failures in the distributed systems and figure out which components are abnormal. However, due to the complex dependencies, a failure occurring at a root-cause component can propagate to other components and cause their KPIs abnormal as well [23]. Therefore, to diagnose the failure, it is necessary to distinguish the root-cause components from other abnormal components. Furthermore, to localize the fine-grained root causes of a failure, it is required to identify root-cause metrics and the root-cause components further. For each component, there are plenty of metrics (e.g., CPU utilization and network throughput) collected. When we identified the root cause metrics, for example, CPU usage, it is clear that the failure is caused by CPU exhaustion on the component. In summary, we aim to discover failures by KPI anomaly detection and then identify the root-cause components and the root-cause metrics.

As introduced in Section 3, KPI anomaly detection has been extensively studied. The existing approaches of identifying root-cause components [17, 20, 23, 26, 32, 45, 51] can be classified into two categories, i.e., trace-based [17, 23, 26, 45, 51] and dependency graph-based [20, 32]. There are also some works [30, 42] further identifying root-cause metrics. MicroDiag [42] models the dependencies among metrics by causality discovery and ranks metrics by PageRank. FluxRank [30] ranks metrics by learning-to-rank.

5.2 Dataset

Most existing works on failure diagnosis use private datasets for evaluation. Some existing works [23, 51] published their datasets, but these datasets contain only traces and thus do not support fine-grained failure diagnosis. Furthermore, though there are some public datasets (e.g., Azure\(^2\) and Alibaba\(^3\)) including fine-grained metrics, there are no ground-truth failure time points and root causes. Therefore, to provide a comprehensive benchmark for failure discovery and fine-grained diagnosis, we construct and publish a dataset [3], named \( \mathcal{C} \). Dataset \( \mathcal{C} \) has already been used by existing research works [8, 23, 45].

The dataset \( \mathcal{C} \) is based on \( \text{sysA} \), whose architecture is shown in Fig. 3. There are three types of monitoring data in the dataset \( \mathcal{C} \), i.e., traces, KPIs, and metrics. User requests are sent to one of the OSB components and then invoke other components, including Service and DB. The whole execution process of a user request is called a trace. We record each trace by its spans on each component, including the attributes of each span and the causal relationships among spans. A span is a unit of work done in a component of a

---

\(^2\)https://github.com/Azure/AzurePublicDataset
\(^3\)https://github.com/alibaba/clusterdata
The attributes of the spans on each component can be aggregated into KPIs to reflect the overall status of each component. In Table 4, we present an example trace, where each row is a span. Each span has a unique ID (the field id), the ID of its parent span (the span that invokes this span) (pid), and the unique ID of the whole trace (traceId). With these three fields we can identify the spans of a trace and the causal relationships among the spans. The other fields characterize a span. For example, the span of the first row occurs at 1590249600016 (Unix timestamp in milliseconds), costs 274 milliseconds, and has a successful response.

Table 4: A example trace in the dataset C

| callType   | startTime | elapsedTime | success | traceId     | id            | pid            | cmdId | serviceName |
|------------|-----------|--------------|---------|-------------|---------------|----------------|-------|-------------|
| OSB        | 1590249600016 | 274.0        | True    | 6a171e24568385015da7 | 6a171999799385015da7 | None | os_021   | sub_001       |
| RemoteProcess | 1590249600025 | 257.0        | True    | 6a171e24568385015da7 | 6a171999799385015da7 | 6a171919979385015da7 | None | os_021   | sub_001       |
| CSF        | 1590249600035 | 33.0         | True    | 6a171e24568385015da7 | 6a171999799385015da7 | 6a171919979385015da7 | None | os_021   | sub_001       |
| RemoteProcess | 1590249600037 | 30.0         | True    | 6a171e24568385015da7 | 6a171999799385015da7 | 6a171919979385015da7 | None | os_021   | sub_001       |
| LOCAL      | 1590249600054 | 7.0          | True    | 6a171e24568385015da7 | 6a171999799385015da7 | 6a171919979385015da7 | None | os_021   | sub_001       |
| JDBC       | 1590249600054 | 3.0          | True    | 6a171e24568385015da7 | 6a171999799385015da7 | 6a171919979385015da7 | None | os_021   | sub_001       |

In the dataset C, we provide 169 injected failures with ground-truth information in total. We use these failures rather than real-world failures because it is hard to collect so many real-world failures with ground-truth root causes. On svaA, we injected 7 types of failures, which are summarized in Table 6.

Table 6: Failure injection in the dataset C

| Component   | Injection type | Description                                      |
|-------------|----------------|--------------------------------------------------|
| Database    | close          | Close the listening port of the target instances |
| Database    | session limit  | Decrease the session limit of the target instance |
| Container   | CPU stress     | Stress the CPU of the target container            |
| Container   | network delay $ | Drop packets randomly on the target container     |
| Container   | network delay $ | Delay packets randomly on the target container    |
| Physical node| network delay $ | Drop packets randomly on the target node         |
| Physical node| network delay $ | Delay packets randomly on the target node        |

Based on the dataset C, we held the third AIoPs Challenge in 2020. It attracted 141 teams and 517 participants from both academia and industries. In this competition, the participants are asked to detect and diagnose these failures in an online manner. For each failure, each team is allowed to submit at most two potential root-cause metrics. For network failures on containers, including both loss and delay, as we did not collect network metrics on containers, the participants are asked to localize the root-cause containers only. If the result submitted by a team of a failure achieves a precision score greater than or equal to 0.5, then the result is considered valid. For each failure, the valid results from different teams are ranked by time to diagnose/F-0.5 score, then each of them gets max(10 − i + 1, 0) points (i is the rank). The best score achieved by the participants is 755, given 129 failures in total.

6 CONCLUSION

AIoPs has attracted a great deal of attention from academics and industry. A significant limitation of the research of AIoPs is the lack of public real-world and large-scale datasets. To tackle this problem, we have published three datasets, including KPI anomaly detection, root-cause localization for multi-dimensional data, failure discovery, and diagnosis. More importantly, we held an algorithm competition once a year based on the public datasets, attracting hundreds of teams to take part. Our work is helpful for practitioners and researchers to apply AIoPs to enhance service reliability. In the future, we will continuously publish more datasets involving various AIoPs scenarios and hold the competition. Welcome to pay attention to and actively participate in the competition.
