Incident Diagnosing and Reporting System Based on Retrieval Augmented Large Language Model

Peng Yuan[1], Lu-An Tang[1], Yanchi Liu[1], Kobayashi Yuji[2], Moto Sato[1], Haifeng Chen[1*]

[1]NEC Labs America

[2]NEC Central Research Lab

{pyuan, ltang, yanchi, moto, Haifeng}@nec-labs.com, y_koba@nec.com

Abstract

The Internet of Things (IoT) is widely used in many applications such as smart city, transportation, healthcare, and environment monitoring. A key task of IoT maintenance is to analyze the abnormal sensor records and generate incident report. Traditionally, domain experts engage in such labor intensive tasks. Recent advances in Large Language Model (LLM) have sparked interests in developing AI based systems to automate these labor intensive processes. However, two critical problems hinder the effective application of LLM in IoTs: (1) LLM lacks background knowledge of deployed IoTs; and (2) the incidents are complex events involving many sensors and components. LLM needs to understand the sensor relationships for accurate diagnosis. In this study, we propose a Retrieval Augmented language model based Incident Diagnosing and Reporting system (RAIDR) for IoT applications. RAIDR retrieves related system documents based on the incident features and leverages LLM to analyze anomalies, identify root causes, and automatically generate incident reports. The automated incident reporting process streamlines end users' decision making for system maintenance and troubleshooting.

Introduction

The IoT integrates a large number of devices and sensors to monitor system behaviors and environment changes (Tang et al. 2019). As a key task of IoT applications, managing the detected incidents presents significant challenges. These incidents involve hundreds of sensors and thousands of abnormal records. Traditionally, domain experts undertake this labor-intensive process, dedicate weeks to manually diagnose issues, identify potential root causes, and compile final incident reports. In most cases, the response time is too long for efficient resolution and system repair. With recent advances of LLM, it is appealing to develop an AI based assistant to automate this task (Su et al. 2024). However, existing LLMs have limitations when directly applied to such tasks (Sarda et al. 2023; Ahmed et al. 2023), mainly due to the following two challenges: **Background Knowledge Gap**: Most LLMs lack intrinsic understanding of IoT specific contexts, including the sensor types, environmental conditions, and system intricacies. Without essential background information, the LLM cannot provide meaningful analysis for the detected incidents. **Complexity of IoT Incidents**: An IoT incident is a complex event that involves hundreds of sensors and components. The anomaly typically originates from a single sensor but propagates and affects more sensors over time. The LLM must understand the sensor relationships and dependencies for accurate diagnosis.

In response to these challenges, we propose a novel solution of Retrieval Augmented language model based Incident Diagnosing and Reporting system (RAIDR) for IoT applications. RAIDR first collects the abnormal records from deployed detectors and constructs a graph model to capture the sensor interactions and dependencies. Then, RAIDR generates the features to represent the incident timeline and influences. Such features are served as the key to query and retrieve relevant documents, such as sensor description, operation manual, incident tickets and historical reports. RAIDR generates the prompts based on the retrieved documents and anomaly features. By incorporating context-aware prompts, RAIDR enhances LLM's ability to understand IoT specific environments. Finally, the incident reports are generated with suggested actions to help the end user's trouble shooting.

System Description

RAIDR includes three major modules: (1) incident analysis; (2) document retrieval and (3) report generation.

*Corresponding author: Haifeng@nec-labs.com

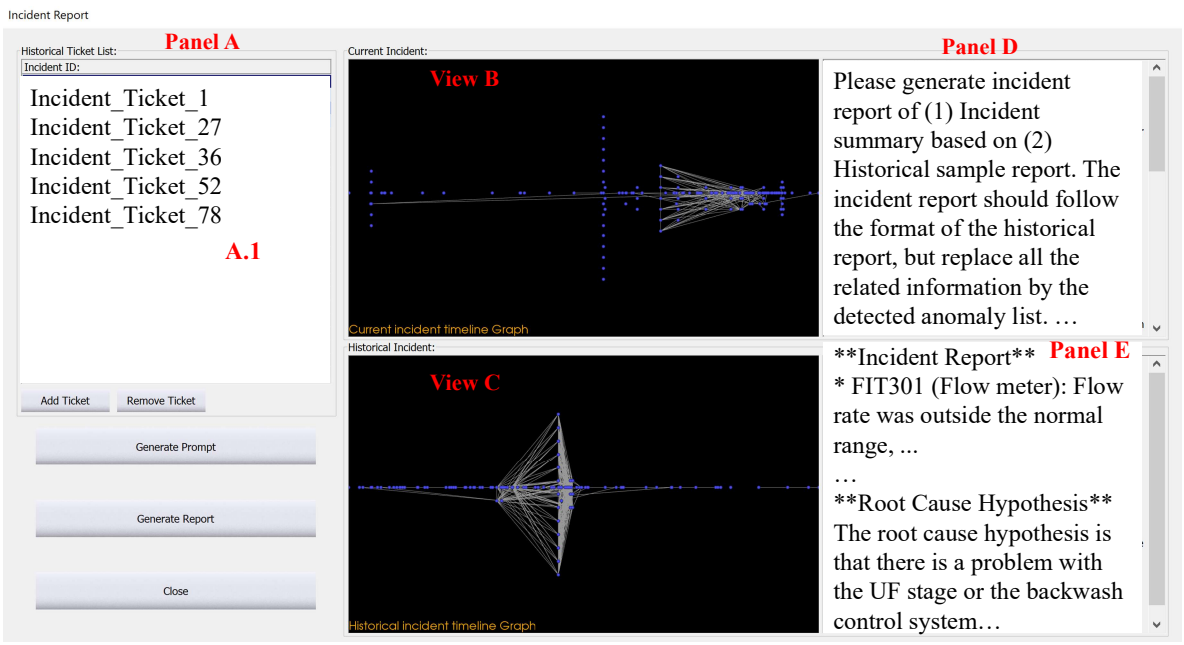

Figure 1: The incident report dashboard of RAIDR

The first module is the incident analysis module. The system takes detected anomalies from deployed IoT detectors as input. These detectors are developed based on different mechanism (Sarda et al. 2023; Yuan et al. 2022). RAIDR also takes the sensor's relationship graph as another input. This graph can be trained based on the correlation of the sensor readings or categorical value switch timestamps (Yuan et al. 2023). In the document retrieval module, RAIDR retrieves the incident signatures. These signatures are key features to identify each incident. Such features are used as the key to query the historical datasets of incident tickets and retrieve the similar ones. In the third module, RAIDR integrates retrieved documents to generate some examples as the in-context learning prompts for LLM. Then RAIDR uses the LLM to generate incident report following the same format of historical tickets. The generated incident report is then output to end users for their trouble shooting tasks.

Demo Scenario

In the demo, we will show an example of using RAIDR to analyze anomalies from a real IoT dataset: At the beginning, the users upload detected anomalies of the system incidents. The incident periods and detailed anomalies are listed. There are many anomalies been detected from dozens of sensors. RAIDR conducts a clustering algorithm to group the anomalies on relevant sensors, and shows the overall anomaly scores of the system over time. RAIDR automatically computes the system level incident threshold by number of abnormal sensors. After that, RAIDR determines the incident periods and highlights them on the timeline. After checking the incident graphs, the user can start the process of incident analysis. The detailed dashboard is shown as Figure 1. RAIDR retrieves the similar tickets in the list of A.1. RAIDR lists out the timeline graph of the new incident (View B) and the historical similar one (View C). After that, the user just clicks the button of "Generate Prompt". RAIDR will integrate all the retrieved documents to prepare the input prompt (Panel D) for LLM. The prompt includes three parts: (1) the sensors' background information, (2) the report examples from historical tickets, and (3) the error messages and detected anomalies of the new incident. Finally, the user clicks the button of "Generate Report". These prompts are input to LLM and the final incident report will be generated.

Conclusion

In this paper, we present a novel solution of Retrieval Augmented language model based Incident Diagnosing and Reporting system (RAIDR) for IoT applications. RAIDR generates the timeline of detected incidents and uses the incident signature to retrieve relevant tickets. RAIDR then generates prompts based on the retrieved documents and uses LLM to generate final incident reports. RAIDR can be used to a wide variety of IoT applications and help troubleshooting for different types of incidents.

References

Shu Tang, Dennis R. Shelden, Charles M. Eastman, Pardis Pishdad-Bozorgi, Xinghua Gao. 2019. A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Automation in Construction* 101:127-139.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, Junhong Lin. 2024. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. arXiv:2402.10350.

Komal Sarda, Zakeya Namrud, Raphael Rouf, et.al. 2023. ADARMA auto-detection and auto-remediation of microservice anomalies by leveraging large language models. In *33rd Annual International Conference on Computer Science and Software Engineering*, 200-205.

Toufique Ahmed, Supriyo Ghosh, Chetan Bansal, et.al. 2023. Recommending Root-Cause and Mitigation Steps for Cloud Incidents using Large Language Models. In *ICSE*, 1737-1749.

Peng Yuan, Lu-An Tang, Haifeng Chen, David S. Chang, Moto Sato, Kevin Woodward. 2023. Temporal Graph Based Incident Analysis System for Internet of Things. In *ECML/PKDD*, 305-309.

Peng Yuan, Lu-An Tang, Haifeng Chen, Moto Sato, Kevin Woodward. 2022. 3D Histogram Based Anomaly Detection for Categorical Sensor Data in Internet of Things. *Open J. Internet Things* 8(1): 32-43.

Peng Yuan, Lu-An Tang, Haifeng Chen, Moto Sato, Kevin Woodward: Explainable Anomaly Detection System for Categorical Sensor Data in Internet of Things. 2022. In *ECML/PKDD*, 594-598.

OpenAI. 2023. ChatGPT. https://chat.openai.com.

NEC. 2023. Cotomi. https://neclab.eu/about-us/press-releases/detail/nec-develops-and-provides-generative-ai-for-business-use.

Hugo Touvron, Louis Martin, Kevin Stone, et.al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Li, Z., Luo, C., Zhao, Y., Sun, Y., Sui, K., Wang, X., Liu, D., Jin, X., Wang, Q. and Pei, D. 2019. Generic and robust localization of multi-dimensional root causes. In *2019 IEEE 30th International Symposium on Software Reliability Engineering(ISSRE),* 47–57.

Liu, P., Chen, Y., Nie, X., Zhu, J., Zhang, S., Sui, K., Zhang, M. and Pei, D. 2019. Fluxrank: A widely-deployable framework to automatically localizing root cause machines for software service failure mitigation. In *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE),* 35–46.

Meng, Y., Zhang, S., Sun, Y., Zhang, R., Hu, Z., Zhang, Y., Jia, C., Wang, Z. and Pei, D. 2020. Localizing failure root causes in a microservice through causality inference. In *2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS),* 1–10.

Ma, M., Yin, Z., Zhang, S., Wang, S., Zheng, C., Jiang, X., Hu, H., Luo, C., Li, Y., Qiu, N. and Li, F. 2020. Diagnosing root causes of intermittent slow queries in cloud databases. In *Proceedings of the VLDB Endowment* 13(8): 1176–1189.

Lv-an Tang, Bin Cui, Hongyan Li, Gaoshan Miao, Dongqing Yang, Xinbiao Zhou. 2007. Effective variation management for pseudo periodical streams. In *2007 ACM SIGMOD international conference on Management of data,* 257-268.