ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Improving IT Service Delivery with AI-based Incident Management

Aakash Aluwala

USA

ABSTRACT

This paper discusses how machine learning, predictive analytics, and process automation can be incorporated in improving the IT incident management practice. While the IT structures are becoming vast, there is a requirement for the use of more automated processes such as ticket categorization and the diagnosis and solution of such tickets. In this article, the author presents a literature review of this topic along with examples of AI sub-processes at different stages of the incident management. It also looks at some of the crucial factors that organizations need to bear in mind when deploying AI-based solutions on factors like data handling, model creation, performance tracking and compatibility with present IT service management frameworks.

***Corresponding author**

Aakash Aluwala, USA.

Received: October 14, 2023; **Accepted:** October 21, 2023; **Published:** October 28, 2023

Keywords: Incident Management, Artificial Intelligence, Machine Learning, Automation, It Service Delivery, AIOps

Introduction

IT incident management is becoming a crucial area for companies that are rapidly evolving to depend on IT systems. However, complex systems are no longer effectively managed by conventional methods which posit risks to the users. As these incidents greatly influence businesses that lack updated resources, there is a need to transform the whole management into a seamless and efficient system. Hence, AI seems promising to address the problem as it can increase service availability while decreasing resolution times. Futhermore, AI can ensure the service delivery with less disruptions and resources.

Some of the most recent research shows that AI is gradually finding its way into IT operations particularly in the automation of core activities. Analysis explains how AI and machine learning can be applied to enhance various activities of the ITIL framework such as the IT service incident management [1]. They also describe present day AI applications that center on automating routine job to release the support teams for higher value work. Similarly, other researchers in their systematic review of AI solutions adopted in IT management processes observe that the current solutions have a bias towards automating IT operation functions such as incident management [2].

AI can be applied to various sub-processes of the incident management life cycle. AI technologies such as Natural Language Processing (NLP) can be applied to automatically categorize incident by ticket description [3]. This can assist in directing the incident to the right support team within a short span of time. With the help of machine learning, pattern recognition abilities help identify causes of similar events from ticket history [4]. It can also predict the time to resolve a new incident by using time series analysis of the time taken to resolve incidents in the past and the availability of support resources [5]. Such estimates are particularly important for managing user expectations.

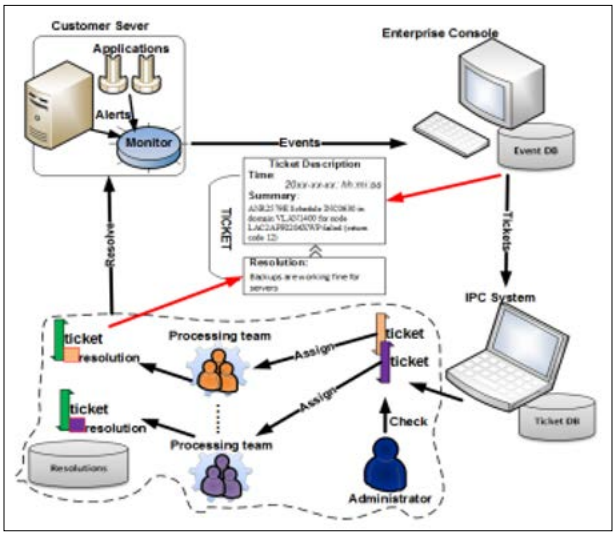

Figure 1: ITIL Service Management System [19].

Literature Review

Some authors analysed the applicability of supervised machine learning for the automatic classification of analysis by the characteristics and content from mobile brand reviews [6]. To achieve this, the researchers used Naive Bayes, Decision Tree, Random Forest and Support Vector Machine (SVM) classifiers to determine the model that obtained the highest precision, accuracy and recall. They discovered that the SVM classifier had the highest accuracy of the models in analyzing all the four measures concerning dissimilar brands.

The employment of auto-classification at the first stage of ticket processing can lead to a decrease in the time required for the

actual examination of tickets by the service desk agents [2]. It enables direct first level routing and effective sorting of the incidents according to their severity levels. However, the author points out that it may not be enough to only automate the initial classification and consider AI for helping with root cause analysis and resolution as well.

Authors elaborated that through the use of machine learning algorithms, specifically clustering algorithms, one can analyze patterns and relationships within the incident data to identify the root causes of related incidents [7]. They learned how Anthropic, an AI safety startup, applies unsupervised machine learning from support cases of Anthropic's own systems [8].

Incidents were clustered using similar attributes such as the stamp of occurrence, services involved, error messages etc [9]. It uncovered some dependencies between particular kinds of failures and allowed for identifying the underlying technical problems. According to the authors, this root cause analysis capability of machine learning can be helpful to reduce a lot of time that support teams spend in identifying correlations [10]. It also helps to address residual risks before they result in more accidents or blackouts.

The AI engines are able to look at all current and historical incident data in Servicenow [11]. They can parse the ticket body, perform basic actions independently using scripts, and offer recommendations to the agents that involve human intervention when necessary [12]. Similarly, Research pays much attention to the use of mature ITSM platforms that can serve as knowledge repositories for AI models to obtain carefully filtered training and evaluation data [13].

Monitoring Tools and Strategies

Supervision is vital to ensure that AI systems in incident management functions are performing optimally. One such requirement is the ability to monitor the key metrics of the model over time in order to look for signs of decline. There are logging platforms such as TensorBoard, Neptune and MLFlow that can be used to log metrics during the training and inference process [14]. Some of the measures include Accuracy, Precision, Recall for the classification models and Clustering quality measures for other models. With an eye on accuracy in a validation set, it is easier to understand when retraining might be due. Tools interact with model registries to associate monitored metrics with version information.

Figure 2: Tools in Machine Learning Libraries [14].

Logged model performance metrics can be analyzed using tools with unsupervised algorithms that can spot statistically significant anomalies without human intervention. For instance, if accuracy deviates from the baseline range it usually has, then an anomaly alert is triggered. Other methods such as time series decomposition and clustering help in such contextual anomaly detection [15]. It acts as an indicator of upcoming events.

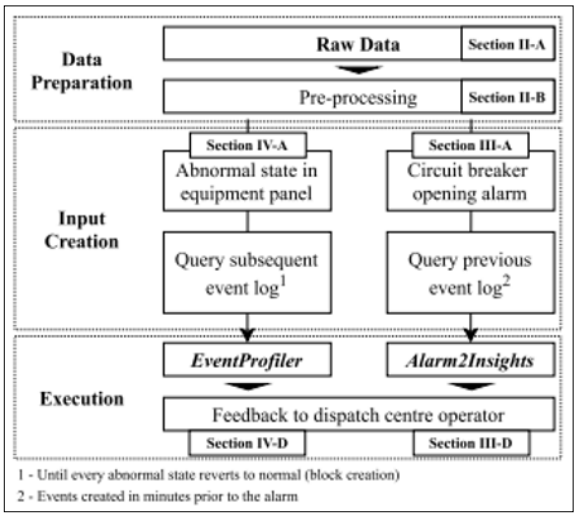

Figure 3: Model Chain for the Data-Driven Analysis of SCADA Alarms [9].

This neutral approach that involves gradually introducing the new models reduces risks of new defects. Some of the platforms such as Kubeflow Pipelines and Argo are equipped with A/B testing pipelines for evaluating multiple model versions [14]. A small percentage of production traffic can be directed to a "canary" version to check its stability before spreading to all users. If there are differences in any of the indicators, they are easily identified. This strategy helps catch bugs earlier in safer deployment phases.

Evaluating inputs given to the models and the outputs/predictions acquired assist in validating the models. It checks raw inputs for certain deviations from their expected distribution or features because such abnormalities may lead to reduced or erroneous output quality.

Debuggers allow executing all inferred methods on sample inputs for pinpointing the root of a defect [15]. Aids such as ModelDB and CloudWatch Insights help debug deep neural network layers by enabling visualization of decision trees, connectivity graphs, and activation pattern. Any deviation from perceived normal behavior identifies issues for directing debugging to specific areas.

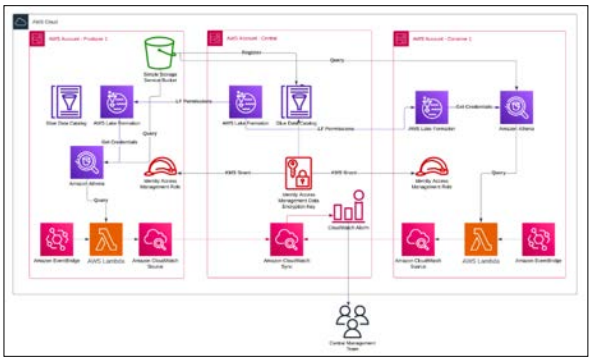

Figure 4: Example of CloudWatch Model used by JPMorgan [16].

Tasks

AI adoption to improve the processes related with incident management will require the identification of specific tasks to be performed. The first set of tasks includes data collection and data preprocessing to obtain the labeled data that are necessary for

building the supervised machine learning models. IT teams must identify the appropriate historical data sources covering various systems and services. They must then identify key features that need to be derived from raw data for use in models. This data must be extracted at scale from multiple repositories by custom scripts and tools and normalized for analysis. Another important step is the utilization of the data annotation process where it involves human workers and AI for analyzing incidents and tagging them with the correct categories, causes, solutions, and other tags.

After obtaining the effective quality of training data, the model is ready to be trained and evaluated. Specific machine learning models have to be selected depending on the goal of applying it to incident management – for classification in routing of incidents, clustering for identifying root cause and others for automation etc. The annotated dataset should then be then split into training, validation and test sets for model iteration over iterations. Cross-validation and tuning involve running inferential tests with the purpose of identifying the best model configuration and hyperparameters. Validation metrics describe absolute levels of quantified models before it is put into a production environment it expects to deliver certain accuracy levels. Testing on real unlabeled data also affirms the efficiency of the model in a way that also affirms extra testing on real data adds validity to the model.

Integrating trained AI models within current available ItSM tools using APIs or interfaces is another set of imperative tasks [17]. This enables the constant checking of degradations in the performance of AI systems, and the subsequent retraining of the AI when such situations occur. The ability to version a model and log when a certain version is used is also another aspect of the regulation and governance process. The use of new production data to continue recalibration of models in recurrent training cycles helps in updating models.

For example, Researcher have developed an AI method in which workflow of ticket response can be automated from an analysis of historical tickets [18]. This included five different ticket attributes, including status, competency, subjects, timestamps, and two types of resolutions; plus relationship representations or edges between the nodes. A recursive neural network trained on this graph learnt multistep resolution procedures of recurring incident type [19].

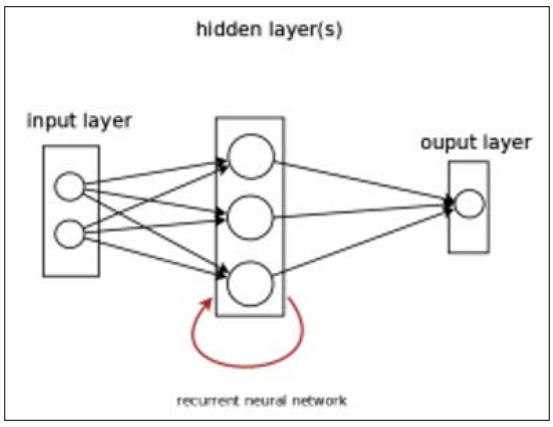

Figure 5: Recurrent Neiural Network with a Single Layer [19].

The model then looks for similar historical tickets and assigns the response playbook that is specific to the automated steps of resolution. This makes it possible for systems to resolve ordinary occurrences without the necessity for human-interference provided occurrences are in accordance to expected standard sequences. The authors assert that it enhances MTTR (Mean Time To Resolve) by minimizing repetitive manual actions and lightening the load of MTAs at the expense of time devoted to more complicated occurrences [20].

5. Solutions and Implementation

An ideal approach would be implementing an AI platform that includes essential tooling for data, modelling, monitoring, and interpretability. The choice of implementation strategies takes into account the organizational requirements and infrastructure of IT.

Historical and real-time incident data need to be collected concerning an incident and placed into a scalable data lake to create an authoritative source [21]. Such product includes AWS Glue, Databricks, and Hadoop for data warrior, semantic web, and access management. A golden record schema reduces redundancies and variation, and is governed and curated. ETL pipelines are always processing new data from various sources at or close to real-time.

The unified factory deploys, trains and fine-tunes Machine Learning pipelines in terms of classification, clustering and supervised/predictive tasks. Some of these are modular reusable components for example TensorFlow Extended and PyTorch. Notebooks facilitate experimentation [22]. Pipelines are connected with the data hub and can use the value of scalable distributed computing through Spark. A model registry is versioned, depends on the others and has an ability to be automated.

Service agents are enhanced via ML models with the aid of xAI in SSA frameworks within ITSM solutions such as ServiceNow by providing recommendations [7]. Natural queries are queries that allow users to find cases matching their query with their resolution at the top. The usage of interconnected dashboards concerns the most frequently posed questions and the indicators for leaders. Agents rating their recommendations to ensure that models are expanding and becoming better. The tasks that involve the repetitive user queries are managed by bots on their own. It is safely rolled out by a governance body [23,24].

Results

The above research findings indicate that the use of AI techniques provides superior outcomes when implemented with gathering and managing the incident. It might be noted that in preliminary runs on several large organizations, the automated classification based on machine learning algorithms was found to work with reasonable degree of accuracy and reliability when tested with new unlabeled tickets. This was instrumental in cutting the time taken by tier 1 agents to review and route the tickets.

The use of chatbots to handle simple requests significantly reduced the load on the service desks and brought the number of repetitive requests down to a third of the initial value in the first month only. This significantly relieved the burden of workload pressure on the agents. At the same time, resolution playbooks derived from previously closed tickets and performed via workflow engines indicated the possibility of addressing a significant number of new incidents without consulting with people.

However, while accuracy emerged high on structured data, NLP models were unable to perform well in embedded natural language text at times. This remains a key area of concern to address. In addition, bias auditing also found minor underperformance on the distorted types of incidents that were not adequately represented in past data sets. More so, expansion of the training data continues

to be leveraged to address such long-tail issues.

Conclusion

In conclusion, this paper shows that AI applied selectively offers tremendous capacity to improve IT incident management steps beyond what would be achievable using manual work alone. Promising approaches like machine learning, predictive analytics and process automation present initial yet undeniable positive results to automate redundant sub-tasks in the context of managing incidents. This can be translated to significant gains such as increased service desk productivity, faster problem solving, less down time and most importantly; enhanced user experiences.

However, there is still more work to be done, particularly on providing human-interpretable explanations for the decision-making process of the algorithms, learning from scenarios where dataset bias poses a challenge, and ensuring that AI is an enhancement to human intelligence in handling of things like anomalies. More innovation with greater emphasis on addressing responsible challenges can further enhance the capabilities for delivering unmatched services through artificial intelligence in IT operations management.

References

1. Glintschert M (2020) AI-Driven IT and its Potentials – A State-of-the-Art approach. SSRN Electronic Journal 48.
2. Mandal A, Agarwal S, Malhotra N, Sridhara G, Ray A, et al. (2019) Improving IT Support by Enhancing Incident Management Process with Multi-modal Analysis. in Lecture notes in computer science 431-446.
3. Cristian M, Christian S, Dumitru-Tudor T (2019) A Study in the Automation of Service Ticket Recognition using Natural Language Processing. IEEE.
4. Zhang K, Xu J, Min MR, Jiang G, Pelechrinis K, et al. (2016) Automated IT system failure prediction: A deep learning approach. IEEE.
5. Lee J, Ni J, Singh J, Jiang B, Azamfar M, et al. Intelligent maintenance systems and predictive manufacturing. Journal of Manufacturing Science and Engineering 142.
6. Guia M, Silva R, Bernardino J (2019) Comparison of Naïve Bayes, Support Vector Machine, Decision Trees and Random Forest on Sentiment Analysis. 11th International Conference on Knowledge Discovery and Information Retrieval.
7. Glintschert M (2020) AI-Driven IT and its Potentials – A State-of-the-Art approach. SSRN Electronic Journal.
8. Hendrix J, Morozoff D (2022) Media forensics in the age of disinformation. in Advances in computer vision and pattern recognition 7-40.
9. Andrade JR, Rocha C, Silva R, Viana JP, Ricardo J Bessa, et al. (2022) Data-Driven Anomaly Detection and event log profiling of SCADA alarms. IEEE Access 10: 73758-73773.
10. Ma Q, Li H, Thorstenson A. (2021) A big data-driven root cause analysis system: Application of Machine Learning in quality problem solving. Computers & Industrial Engineering 160: 107580.
11. Patel M, Smita Patil, Katerina Tzanavara, Manish Chauhan, Yuhao Wang, et al. (2019) Service Now: CMDB Research. School of Professional Studies
12. Lammers M (2019) A QA-pair generation system for the incident tickets of a public ICT Shared Service Center. Available: https://essay.utwente.nl/77562/.
13. Heinsuo L (2020) SMART TICKETING Continuous learning system for document classification. 2020. Available: https://trepo.tuni.fi/bitstream/handle/10024/122011/HeinsuoLeo.pdf?sequence=2.
14. Dott R, Enrico G, Correlatore D, Alessandro B, Candidato E, et al. (2021) MLOps -Standardizing the Machine Learning Workflow Thesis on Big Data. Available: https://amslaurea.unibo.it/23645/1/tesi_enrico_salvucci.pdf.
15. Andrade JR, Rocha C, Silva R, Viana JP, Ricardo J Bessa, et al. (2022) Data-Driven Anomaly Detection and event log profiling of SCADA alarms. IEEE Access 10: 73758-73773.
16. Xu Z, Ma S, Zhang X, Zhu S, Xu B (2018) Debugging with intelligence via probabilistic inference. ICSE.
17. Amazon (2022) How CloudWatch cross-account observability helps JPMorgan Chase improve Federated Data Lake Monitoring | AWS Cloud Operations & Migrations Blog. aws.amazon.com, https://aws.amazon.com/blogs/mt/how-cloudwatch-cross-account-observability-helps-jpmorgan-chase-improve-federated-data-lake-monitoring.
18. Hintsch J, Khan A, Siegling A, Turowski K (2017) Application software in Cloud-Ready Data Centers: a survey. Service science: research and innovations in the service economy 261–288.
19. Zhou W, Wei Xue, Ramesh Baral, Qing Wang, Chunqiu Zeng, et al. STAR. KDD.
20. Jacob S, Qiao Y, Jacob P, Lee B (2020) Using Recurrent Neural Networks to Predict Future Events in a Case with Application to Cyber Security. BUSTECH 2020 : The Tenth International Conference on Business Intelligence and Technology 13-19.
21. Sabharwal N, Bhardwaj G (2022) Hands-on AIOps.
22. BOUMA T (2018) Current Big data needs and trends. Muni.cz, 2018. https://is.muni.cz/th/j7ctf/Diplomka___16__Archive.pdf.
23. Singh P, Manure A (2019) Learn TensorFlow 2.0: Implement Machine Learning and Deep Learning Models with Python. 2019. Available: http://link.springer.com/content/pdf/10.1007/978-1-4842-5558-2.pdf.
24. IBRAHIM A (2019) The Cyber Frontier: AI and ML in Next-Gen Threat Detection. 2019.