USTEP: Structuration des logs en flux grâce à un arbre de recherche évolutif

Arthur Vervaet*, Raja Chiky**, Mar Callau-Zori*

*3DS Outscale, Saint-Cloud, France
**ISEP - Institut Supérieur d'Electronique de Paris, Issy-les-Moulineaux, France
arthur.vervaet@outscale.com, raja.chiky@isep.fr

Résumé. Les registres de *logs* enregistrent en temps réel des informations relatives à l'exécution d'un système informatique. Ceux-ci sont régulièrement consultés à des fins de développement ou de surveillance. Afin d'être utilisables plus facilement pour des tâches d'exploration automatique, les messages des *logs* doivent être structurés. Cette étape se déroulant en amont des opérations d'analyses, elle peut devenir un goulot d'étranglement temporel et influencer l'efficacité des méthodes en aval. Dans ce papier, nous présentons USTEP, une méthode de structuration des *logs* en flux. Basée sur un arbre de recherche évolutif, USTEP est capable de découvrir et d'encoder efficacement de nouvelles règles de structuration. Nous présentons ici une évaluation des performances de USTEP sur un panel de 13 jeux de données issus de systèmes différents. Nos résultats mettent en valeur la supériorité de notre approche en matière d'efficacité et de robustesse par rapport à d'autres méthodes de structuration en ligne.

1 Introduction

Pour les services en ligne de grande envergure, une seule anomalie peut impacter des millions d'utilisateurs (Liang et al., 2021). Pouvoir détecter de tels événements en temps réel permet aux équipes de surveillance de limiter leur impact. Les journaux de *logs* enregistrent une vaste gamme d'événements au sein d'un système informatique et sont communément utilisés pour détecter les origines d'une panne, analyser l'activité ou renforcer la sécurité d'une plateforme(Moh et al., 2016). Les *logs* ont prouvé leur valeur pour la détection automatisée d'anomalies dans plusieurs scénarios (Du et al., 2017; Zhang et al., 2019; Meng et al., 2019). Afin de ne pas avoir à traiter des données textes brutes, les méthodes précédemment citées ont recours à une étape de structuration des messages *logs*. Les *logs* et leur message étant produits par des lignes de code spécifiques (*e.g., print(), logger.log()*), le processus de structuration vise à retrouver le motif sous-jacent à chaque entrée (Figure 1).

La qualité de cette étape de structuration a un impact sur l'efficacité des méthodes en aval. La Figure 2 illustre l'influence de la qualité de structuration des logs (*Parsing Accuracy (PA)*) sur la précision (*Anomaly Detection (AD) accuracy*) de Deeplog (Du et al. (2017)) une méthode de détection d'anomalies. Ici, une perte de 0,2 en *PA* réduit de 0,75 la précision *AD* maximale (0,97 à 0,22) et de 0,59 la précision *AD* moyenne (0,78 à 0,19). Le graphique présenté est une distribution des résultats de 10 instances de Deeplog entraînées sur différentes versions