الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbes

Thesis

To obtain the diploma of **Master's Degree**

Field of Study: **Computer Science**

Specialization: **Artificial Intelligence and Data Science[AIDS]**

Theme

Log-Based Anomaly Detection: A Literature Review, Comparative Evaluation, and Future Directions

---

Presented by
Khodja Yousra
Meski Melissa

Defended on: **September, 2025**
In front of the jury composed of

Dr. KHALDI Belkacem	President of the Jury
Dr. BEKKOUCHE Mohammed	Thesis Supervisor
Ms. BENALI Celine	Co-Supervisor
Dr. BEDJAOUI Mohammed	Examiner

Academic Year: 2024/2025

Abstract

In modern large-scale computing environments, system logs constitute a primary source of information for understanding system behavior and diagnosing faults. However, the ever-increasing volume and complexity of logs render manual inspection impractical, creating the need for automated anomaly detection methods. This thesis addresses this challenge by providing a comprehensive study of log-based anomaly detection, covering the fundamentals of log systems, preprocessing techniques such as parsing and log representation, and a systematic exploration of detection methods ranging from traditional approaches to advanced deep learning and transformer-based models.

Through an extensive review and evaluation, we observe that statistical and classical machine learning methods are often effective for smaller datasets or logs with limited structural diversity, but they struggle to maintain performance when applied to larger and more complex datasets with diverse templates. Deep learning approaches, particularly recurrent and convolutional architectures, demonstrate improved ability to capture sequential patterns in logs, although they require significant training data and computational resources, and can face difficulties in generalizing across different system environments. Recent transformer-based methods show state-of-the-art performance, as they leverage self-attention to model long-range dependencies in log sequences, achieving superior results in terms of precision, recall, and F1-score across benchmark datasets such as HDFS, BGL, and Thunderbird. Overall, the results confirm that while traditional methods remain useful in certain constrained scenarios, the future of log anomaly detection lies in advanced sequence modeling approaches, with transformer-based architectures emerging as the most promising direction.

Keywords– System logs, Log anomaly detection, Log parsing, Log representation, Machine learning, Deep learning, Transformer models, Sequence modeling.

الملخص

في بيئات الحوسبة الحديثة واسعة النطاق، تُعَدّ سجلات الأنظمة مصدرًا أساسيًا للمعلومات لفهم سلوك الأنظمة وتشخيص الأعطال. ومع ذلك، فإن الزيادة المستمرة في حجم السجلات وتعقيدها تجعل الفحص اليدوي أمرًا غير عملي، مما يخلق حاجة ملحّة إلى أساليب آلية لاكتشاف الانحرافات. يتناول هذا البحث هذا التحدي من خلال دراسة شاملة لاكتشاف الانحرافات القائم على السجلات، حيث يغطي أساسيات أنظمة السجلات، وتقنيات المعالجة المسبقة مثل التحليل الصرفي والتمثيل البنيوي للسجلات، بالإضافة إلى استكشاف منهجي لطرق الكشف بدءًا من الأساليب التقليدية وصولًا إلى النماذج المتقدمة للتعلم العميق والأساليب المعتمدة على المحولات.

ومن خلال مراجعة وتقييم موسعين، لاحظنا أن الطرق الإحصائية وأساليب التعلم الآلي الكلاسيكية غالبًا ما تكون فعّالة مع مجموعات البيانات الصغيرة أو السجلات ذات البنية البسيطة، لكنها تفشل في الحفاظ على أدائها عند تطبيقها على مجموعات أكبر وأكثر تعقيدًا تحتوي على قوالب متنوعة. أما تقنيات التعلم العميق، وخاصة الشبكات العصبية المتكررة والالتفافية، فقد أظهرت قدرة أعلى على التقاط الأنماط التسلسلية في السجلات، لكنها تتطلب كميات ضخمة من بيانات التدريب وموارد حسابية كبيرة، كما أنها تواجه صعوبات في التعميم عبر بيئات أنظمة مختلفة. أما الأساليب الحديثة المعتمدة على المحولات فقد حققت نتائج متقدمة على مستوى الأداء، إذ تستفيد من آلية الانتباه الذاتي لنمذجة الاعتمادات طويلة المدى في تسلسل السجلات، مما أتاح لها تحقيق نتائج متفوقة من حيث الدقة والاسترجاع ودرجة F1 على مجموعات بيانات مرجعية مثل HDFS و BGL و Thunderbird.

بشكل عام، تؤكد النتائج أن الأساليب التقليدية ما زالت مفيدة في بعض السيناريوهات المحدودة، إلا أن مستقبل اكتشاف الانحرافات في السجلات يتجه نحو أساليب متقدمة لنمذجة التسلسل، حيث تبرز البُنى المعتمدة على المحولات باعتبارها الاتجاه الأكثر وعدًا.

الكلمات المفتاحية -- سجلات النظام، كشف الانحرافات في السجلات، تحليل السجلات، تمثيل السجلات، التعلم الآلي، التعلم العميق، نماذج المحوّل ،(Transformers) نمذجة التسلسل.

Résumé

Dans les environnements informatiques modernes à grande échelle, les journaux système constituent une source d'information essentielle pour comprendre le comportement des systèmes et diagnostiquer les pannes. Cependant, l'augmentation constante du volume et de la complexité des journaux rend l'inspection manuelle impraticable, ce qui crée un besoin urgent de méthodes automatisées de détection d'anomalies. Ce mémoire répond à ce défi en proposant une étude complète de la détection d'anomalies basée sur les journaux, couvrant les fondements des systèmes de logs, les techniques de prétraitement telles que l'analyse syntaxique et la représentation des journaux, ainsi qu'une exploration systématique des méthodes de détection allant des approches traditionnelles aux modèles avancés d'apprentissage profond et basés sur les transformateurs.

À travers une revue et une évaluation approfondies, nous observons que les méthodes statistiques et d'apprentissage automatique classique s'avèrent souvent efficaces pour les petits jeux de données ou pour des journaux présentant une faible diversité structurelle, mais elles peinent à maintenir leurs performances lorsqu'elles sont appliquées à des ensembles plus vastes et plus complexes avec des modèles de logs variés. Les approches d'apprentissage profond, en particulier les architectures récurrentes et convolutionnelles, démontrent une meilleure capacité à capturer les dépendances séquentielles dans les journaux, bien qu'elles nécessitent des volumes de données d'entraînement importants ainsi que des ressources computationnelles élevées, et qu'elles rencontrent des difficultés à généraliser entre différents environnements systèmes. Les méthodes récentes basées sur les transformateurs affichent des performances à l'état de l'art, car elles exploitent le mécanisme d'auto-attention pour modéliser les dépendances à long terme dans les séquences de logs, obtenant des résultats supérieurs en termes de précision, rappel et F1-score sur des jeux de données de référence tels que HDFS, BGL et Thunderbird.

Dans l'ensemble, les résultats confirment que, bien que les méthodes traditionnelles conservent une utilité dans certains scénarios contraints, l'avenir de la détection d'anomalies dans les journaux repose sur des approches avancées de modélisation séquentielle, les architectures basées sur les transformateurs apparaissant comme la direction la plus prometteuse.

Mots-clés– Journaux système, Détection d'anomalies dans les journaux, Analyse syntaxique des journaux, Représentation des journaux, Apprentissage automatique, Apprentissage profond, Modèles Transformers, Modélisation de séquences.