الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا لإعلام الآلي - 1945 ماي 08 - بسيدي بلعباس
Ecole Supérieure en Informatique
-08 Mai 1945- Sidi Bel Abbes

THESIS

To obtain the diploma of **Master**
Field: **Computer Science**
Specialty: **Ingénierie des Systèmes Informatiques (ISI)**

Theme

Anomaly Detection Using Logs Data

Presented by :
Aid Abderrahmane

Submission Date : 02/07/2025

In front of the jury composed of:

- Dr. KLOUCHE Badia	President
- Dr. Hanae Naoum	Supervisor
- Pr. Sidi Mohammed BENSLIMANE	Co-Supervisor
- Dr. ELHANNANI Souad	Examiner
- Mrs. Sanna Senouci	Guest

Academic Year : 2024/2025

Abstract

System crashes and security compromises from out-of-ordinary behaviors in computing systems can greatly threaten the operations and the cybersecurity of an organization. Anomaly detection in log data is fundamental to system security and greatly depends on the accuracy and real-time detection. But discovering problems from unstructured high-dimensional log data is relatively hard, with most of them depending on the manual approach or incomplete automation to discard false positives, and suffering from harmful imbalance and varying log formats.

This thesis investigates the problem of detecting anomalies from log data, and applies machine learning and deep learning approaches to overcome limitations of established methods. These difficulties such as redundant runtime information, absence of labeled anomalies, and heterogeneous log structures, have been addressed from the use of sophisticated preprocessing methods such as log parsing and semantic embedding extraction, and the use of strong model such as transformer based architecture and hybrid autoencoder-recurrent neural network. These approaches improve the detection precision and portability in a wide range of log environments, which can promote the proactive monitoring of systems and cyber security.

Keywords: Anomaly Detection, Log Data, Machine Learning, Deep Learning, Log Parsing, Transformer Models, Cybersecurity, System Monitoring.

Résumé

Les pannes système et les compromissions de sécurité dues à des comportements inhabituels dans les systèmes informatiques peuvent gravement menacer les opérations et la cybersécurité d'une organisation. La détection d'anomalies dans les données de journaux est essentielle pour la sécurité des systèmes et repose largement sur la précision et la détection en temps réel. Cependant, identifier les problèmes à partir de données de journaux non structurées et à haute dimension est relativement difficile, la plupart des approches dépendant de méthodes manuelles ou d'une automatisation incomplète pour éliminer les faux positifs, et souffrant de déséquilibres nuisibles et de formats de journaux variés.

Cette thèse examine le problème de la détection d'anomalies dans les données de journaux et applique des approches d'apprentissage automatique et d'apprentissage profond pour surmonter les limites des méthodes établies. Ces défis, tels que les informations d'exécution redondantes, l'absence d'anomalies étiquetées et les structures de journaux hétérogènes, ont été abordés grâce à l'utilisation de méthodes de prétraitement sophistiquées telles que l'analyse syntaxique des journaux et l'extraction d'incorporations sémantiques, ainsi que l'utilisation de modèles puissants comme l'architecture basée sur les transformers et un réseau neuronal récurrent-autoencodeur hybride. Ces approches améliorent la précision de la détection et la portabilité dans une large gamme d'environnements de journaux, favorisant ainsi une surveillance proactive des systèmes et la cybersécurité.

Mots-clés : Détection d'anomalies, Données de journaux, Apprentissage automatique, Apprentissage profond, Analyse syntaxique des journaux, Modèles de transformers, Cybersécurité, Surveillance des systèmes.

الملخص

يمكن أن تشكل الأعطال النظامية والانتهاكات الأمنية الناتجة عن السلوكيات غير العادية في الأنظمة الحاسوبية تهديدًا كبيرًا لعمليات المنظمات وأمنها السيبراني. يُعد الكشف عن الحالات الشاذة في بيانات السجلات أمرًا أساسيًا لأمن النظام، ويعتمد بشكل كبير على الدقة والكشف في الوقت الفعلي. ومع ذلك، يُعتبر اكتشاف المشكلات من بيانات السجلات غير المنظمة وذات الأبعاد العالية أمرًا صعبًا نسبيًا، حيث تعتمد معظم الطرق الحالية على أساليب يدوية أو أتمتة غير كاملة لاستبعاد الإيجابيات الخاطئة، وتعاني من اختلالات ضارة وتنوع كبير في صيغ السجلات. تتناول هذه الأطروحة مشكلة الكشف عن الحالات الشاذة في بيانات السجلات، من خلال تطبيق تقنيات التعلم الآلي والتعلم العميق للتغلب على قيود الأساليب التقليدية. تمت معالجة التحديات المرتبطة مثل وفرة المعلومات غير الضرورية أثناء التشغيل، وغياب الحالات الشاذة الموسومة، والهياكل غير المتجانسة للسجلات، باستخدام طرق معالجة مسبقة متقدمة مثل تحليل السجلات واستخلاص التضمينات الدلالية، بالإضافة إلى اعتماد نماذج قوية كالهندسة القائمة على المحولات والشبكات العصبية التكرارية-الأوتوكودر الهجينة. تسهم هذه النهوج في تحسين دقة الكشف وقابلية النقل عبر مجموعة واسعة من بيئات السجلات، مما يعزز المراقبة الاستباقية للأنظمة ويقوي الأمن السيبراني.

الكلمات المفتاحية: الكشف عن الحالات الشاذة، بيانات السجلات، التعلم الآلي، التعلم العميق، تحليل السجلات، نماذج المحولات، الأمن السيبراني، مراقبة النظام.