

Algerian Democratic and Popular Republic
Ministry of Higher Education and Scientific Research
Higher School of Computer Science
- 08 May 1945 - Sidi Bel Abbès



THESIS

To obtain the **Engineering's Degree**

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science**

Theme:

**Trapezoidal Rare-Sentence Pattern Mining with LLM-Guided
Explanations for Streaming Provenance and Event Data.**

Presented by:

BENGHENIMA Hafsa

In front of the jury composed of:

Dr. ELOUALI Nadia

President

Dr. DIF Nassima

Examiner

Pr. BENSLIMANE Sidi Mohammed

Supervisor

Dr. BENABDERRAHMANE Sid Ahmed

Supervisor

Année universitaire **2025/2026**

Abstract

Modern event streams generate massive, heterogeneous data in which true anomalies are rare, continuously evolving, and difficult to interpret. We propose a two-stage framework for real-time anomaly detection and explanation that combines streaming rarity estimation with multi-signal anomaly fusion. The first stage introduces **Trapezoidal Rare-Sentence Pattern Mining (TRSPM)**, a streaming framework that models rare behavioral patterns and produces a rarity signal using fuzzy rare-pattern mining. The second stage performs deeper anomaly analysis by embedding the shortlisted candidates and combining complementary anomaly signals through **Reciprocal Rank Fusion (RRF)** to improve anomaly ranking. An active-learning mechanism further refines the ranking, while an **LLM-guided Retrieval-Augmented Generation (RAG)** component generates concise, evidence-grounded explanations for the highest-ranked anomalies. The proposed framework is evaluated on cybersecurity, biomedical, business and other datasets, demonstrating strong anomaly detection, ranking performance, and interpretable explanations. Ablation studies further quantify the contribution of streaming rarity estimation, multi-signal fusion, and active learning to the overall effectiveness of the framework.

Keywords: Anomaly Detection, Streaming Data, Rare Pattern Mining, Multi-Signal Fusion, Active Learning, Retrieval-Augmented Generation.

RÉSUMÉ

Les flux d'événements modernes génèrent des volumes massifs de données hétérogènes, dans lesquels les véritables anomalies sont rares, évolutives et difficiles à interpréter. Nous proposons un cadre en deux étapes pour la détection et l'explication d'anomalies en temps réel, combinant une estimation de la rareté en environnement de flux avec une fusion multi-sigaux des anomalies. La première étape introduit **Trapezoidal Rare-Sentence Pattern Mining (TRSPM)**, un cadre de traitement en flux qui modélise les comportements rares et produit un signal de rareté à l'aide d'une approche de fouille de motifs rares fondée sur la logique floue. La seconde étape réalise une analyse plus approfondie des anomalies en représentant les candidats présélectionnés sous forme de plongements sémantiques (embeddings) et en combinant des signaux d'anomalie complémentaires à l'aide de la **Reciprocal Rank Fusion (RRF)** afin d'améliorer le classement des anomalies. Un mécanisme d'apprentissage actif affine ensuite ce classement, tandis qu'un composant de **Retrieval-Augmented Generation (RAG)** guidé par un grand modèle de langage (LLM) génère des explications concises et fondées sur des preuves pour les anomalies les mieux classées. Le cadre proposé est évalué sur des jeux de données issus de la cybersécurité, du domaine biomédical, du domaine commercial ainsi que d'autres domaines, démontrant de solides performances en matière de détection d'anomalies, de qualité du classement et d'interprétabilité des explications. Des études d'ablation quantifient en outre la contribution de l'estimation de la rareté en flux, de la fusion multi-sigaux et de l'apprentissage actif à l'efficacité globale du cadre proposé.

Mots-clés : Détection d'anomalies, Données en flux, Fouille de motifs rares, Fusion multi-sigaux, Apprentissage actif, Génération augmentée par récupération.

ملخص

تُولد تدفقات الأحداث الحديثة كميات هائلة من البيانات غير المتجانسة، حيث تكون الشذوذات الحقيقية نادرة، ومتغيرة باستمرار، ويصعب تفسيرها. في هذه الأطروحة، نقترح إطار عمل مكوّنًا من مرحلتين للكشف عن الشذوذ وشرحه في الزمن الحقيقي، يجمع بين تقدير الندرة في البيانات المتدفقة ودمج إشارات الشذوذ متعددة المصادر. تقدم المرحلة الأولى إطار **Mining Pattern Rare-Sentence Trapezoidal (TRSPM)**، وهو إطار لمعالجة البيانات المتدفقة يقوم بنمذجة الأنماط السلوكية النادرة وإنتاج إشارة للندرة باستخدام تنقيب الأنماط النادرة القائم على المنطق الضبابي. أما المرحلة الثانية فتُجري تحليلاً أكثر عمقاً للشذوذات من خلال تمثيل المرشحين المختارين باستخدام التضمينات الدلالية (Embeddings) ثم دمج إشارات الشذوذ التكميلية باستخدام تقنية **Fusion Rank Reciprocal (RRF)** لتحسين ترتيب الشذوذات. كما تُستخدم آلية للتعلّم النشط لتحسين هذا الترتيب تدريجياً، في حين يتولى مكوّن **Generation Retrieval-Augmented (RAG)** الموجه بواسطة نموذج لغوي كبير (LLM) توليد تفسيرات موجزة ومدعومة بالأدلة للشذوذات الأعلى ترتيباً. وقد تم تقييم الإطار المقترح باستخدام مجموعات بيانات من مجالات الأمن السيبراني، والطب الحيوي، والأعمال، وغيرها من المجالات، حيث أظهر أداءً قوياً في كشف الشذوذ، وتحسين جودة الترتيب، وتوفير تفسيرات قابلة للفهم. كما أظهرت دراسات الإزالة **Studies Ablation** مساهمة كل من تقدير الندرة في البيانات المتدفقة، ودمج الإشارات متعددة المصادر، والتعلّم النشط في تعزيز الكفاءة العامة للإطار المقترح.

الكلمات المفتاحية: كشف الشذوذ، البيانات المتدفقة، تنقيب الأنماط النادرة، دمج الإشارات متعددة المصادر، التعلّم النشط، التوليد المعزّز بالاسترجاع.