

# Découverte de motifs séquentiels et de règles inattendus

D. H. Li\*, A. Laurent\*\*, P. Poncelet\*

\*LGI2P - EMA, SITE EERIE  
{haoyuan.li,pascal.poncelet}@ema.fr

\*\*LIRMM - CNRS - Université Montpellier II  
laurent@lirmm.fr

**Résumé.** Les travaux autour de l'extraction de motifs séquentiels se sont particulièrement focalisés sur la définition d'approches efficaces pour extraire, en fonction d'une fréquence d'apparition, des corrélations entre des éléments dans des séquences. Même si ce critère de fréquence est déterminant, le décideur est également de plus en plus intéressé par des connaissances qui sont représentatives d'un comportement inattendu dans ces données (erreurs dans les données, fraudes, nouvelles niches, ...). Dans cet article, nous introduisons le problème de la détection de motifs séquentiels inattendus par rapport aux croyances du domaine. Nous proposons l'approche USER dont l'objectif est d'extraire les motifs séquentiels et les règles inattendues dans une base de séquences.

## 1 Introduction

Pour faire face aux besoins des nouvelles applications (médicales, suivi de consommation, suivi des navigations sur un serveur Web, etc), de plus en plus de données sont stockées sous la forme de séquences. Pour traiter ces bases et en extraire des connaissances pertinentes, les motifs séquentiels ont été proposés Agrawal et Srikant (1995). Ils permettent, étant donnée une base de données de séquences, de trouver toutes les séquences maximales fréquentes au sens d'un support minimal défini par l'utilisateur. Si la découverte de corrélations dans les données séquentielles est primordiale pour le décideur, il n'en reste pourtant pas moins que certains problèmes ne peuvent être résolus par la recherche de tendances. De nouveaux motifs intéressent le décideur : les motifs inattendus qui contredisent les croyances acquises sur le domaine pour, par exemple, détecter des attaques sur un réseau.

Rappelons que notre objectif n'est pas de trouver les motifs rares, mais bien les motifs contredisant une connaissance, ce qui n'existe pas dans la littérature. La recherche de connaissance inattendue à partir d'une base de croyance a été introduite dans Silberschatz et Tuzhilin (1995) et Padmanabhan et Tuzhilin (2006) présentent une approche de découverte de règles d'association inattendues. Spiliopoulou (1999) propose un cadre basé sur la connaissance du domaine et des croyances pour trouver des règles séquentielles inattendues à partir de séquences fréquentes. Même si ces travaux considèrent des séquences inattendues, ils sont différents de notre problématique dans la mesure où la notion d'inattendue concerne des séquences fréquentes sur la base afin de trier les résultats obtenus. Notre objectif est d'extraire, à