# Clustering de séquences d'évènements temporels

Romain Guigourès**, Dominique Gay*, Marc Boullé*, Fabrice Clérot*

*Orange Labs
prenom.nom@orange.com,
**Zalando
prenom.nom@zalando.de

**Résumé.** Nous proposons une nouvelle méthode de clustering et d'analyse de séquences temporelles basée sur les modèles en grille à trois dimensions. Les séquences sont partitionnées en clusters, la dimension temporelle est discrétisée en intervalles et la dimension évènement est partitionnée en groupes. La grille de cellules 3D forme ainsi un estimateur non-paramétrique constant par morceaux de densité jointe des séquences et des dimensions des évènements temporels. Les séquences d'un cluster sont ainsi groupés car elles suivent une distribution similaire d'évènements au cours du temps. Nous proposons aussi une méthode d'exploitation du clustering par simplification de la grille ainsi que des indicateurs permettant d'interpréter les clusters et de caractériser les séquences qui les composent. Les expériences sur des données artificielles ainsi que sur des données réelles issues de DBLP démontrent le bien-fondé de notre approche.

## 1 Introduction

Les données contenant une information temporelle constituent un défi pour le processus de découverte de connaissances (Yang et Wu, 2006). Les données temporelles sont complexes dans le sens où un objet de la base est décrit par une ou plusieurs séquences d'éléments ordonnés dans le temps. Selon la nature des éléments temporels (catégoriels ou numériques, ponctuels ou continus dans le temps), il existe une grande diversité de méthodes d'extraction de connaissances (Mörchen, 2007). Ici, nous nous intéressons aux données de séquences d'évènements catégoriels et ponctuels, où chaque évènement d'une séquence est associé à un temps $t$, et que nous appelons simplement séquences d'évènements temporels. La fouille de séquences d'évènements temporels trouve des applications dans de nombreux domaines : e.g., dans le domaine médical, Patnaik et al. (2011) explore des bases de dossiers médicaux électroniques de patients à la recherche de motifs d'évènements temporels fréquents ; dans le domaine du Web, Masseglia et al. (2008) et Saleh et Masseglia (2011) extraient des comportements fréquents d'utilisateurs par période de temps ; en sciences sociales, Studer et al. (2010) cherche à grouper des individus selon leur parcours de vie. La majeure partie des efforts de recherche s'est focalisée sur l'extraction de motifs fréquents dans les données de séquences d'évènements temporels (ou TAS pour "Temporally-Annotated Sequences", voir e.g., (Giannotti et al., 2006)). Dans cet article, nous nous intéressons au problème de clustering de séquences : le but est de créer des groupes de séquences qui partagent des caractéristiques similaires. Dans la