Découverte de motifs graduels partiellement ordonnés : application aux données d'expériences scientifiques

Simon Ser*, Fatiha Saïs*, Maguelonne Teisseire**

*LRI, Université Paris Sud, Bât. 650-Ada Lovelace,
91405 Orsay Cedex, France
simon.ser@emersion.fr, Fatiha.Sais@lri.fr
http://www.lri.fr/~sais
** TETIS Irstea Université de Montpellier,
500, rue J. F. Breton 34093, Montpellier Cedex 5, France
maguelonne.teisseire@irstea.fr
http://textmining.biz/Staff/Teisseire/

Résumé. Les données séquentielles sont aujourd'hui omniprésentes et concernent divers domaines d'application. La fouille de données de séquences permet d'extraire des informations et des connaissances pouvant être à forte valeur ajoutée. Cependant, lorsque les données de séquences sont riches en données numériques, des méthodes de fouille de données plus fines sont nécessaires pour extraire des connaissances plus expressives représentant la variabilité des valeurs numériques ainsi que leur éventuelle interdépendance. Dans cet article, nous présentons une nouvelle méthode de découverte de séquences graduelles fréquentes représentées par des graphes à partir d'une source de données de séquences en RDF (Resource Description Framework [1]). Ces dernières sont transformées en graphes graduels partiellement ordonnés, *gpo*. Nous proposons un algorithme permettant de découvrir les sous-graphes *gpo* fréquents. Une expérimentation sur deux jeux de données réelles ont montré la faisabilité et la pertinence de notre approche.

1 Introduction

En raison du développement du marché des objets connectés et des techniques de géolocalisation, les données séquentielles sont omniprésentes et produites en quantité de plus en plus importante. Elles concernent une multitude de domaines d'application, allant de la médecine jusqu'aux télécommunications en passant par l'éducation (Kumar et al. (2011)). Les données séquentielles peuvent être produites sous la forme de suites d'informations ordonnées (e.g., parcours de patients, séquences de génomes, expériences scientifiques) ou sous la forme de séries temporelles (e.g., analyse du signal, données météorologiques, économétrie).

La fouille de motifs séquentiels, qui consiste à découvrir des sous-séquences fréquentes à partir d'une base de données séquentielles, a suscité un grand intérêt dans une multitude de

1. https://www.w3.org/RDF/