Extraction de règles d'épisodes minimales dans des séquences complexes

Lina Fahed, Armelle Brun, Anne Boyer

Université de Lorraine - LORIA - Équipe KIWI
Campus scientifique BP 239 54506 Vandoeuvre-lès-Nancy Cedex
{Lina.Fahed, Armelle.Brun, Anne.Boyer}@loria.fr

Résumé. Les messages déposés quotidiennement sur les réseaux sociaux et les blogs sont très nombreux et constituent une source d'informations précieuse. Leur fouille peut être utilisée dans un but de prédiction d'informations. Notre objectif dans cet article est de proposer un algorithme permettant la prédiction d'informations au plus tôt et de façon fiable, par le biais de l'identification de règles d'épisodes.

1 Introduction

Avec l'émergence du web 2.0, les internautes ne sont plus de simples consommateurs, ils sont également acteurs par le biais des messages qu'ils peuvent déposer, des commentaires qu'ils peuvent laisser et de toute action qu'ils peuvent effectuer. Dans ce cadre, les messages laissés dans les réseaux sociaux représentent une source précieuse d'informations, que de nombreuses recherches cherchent à analyser dans le but d'en comprendre le contenu, d'en extraire les relations cachées, mais aussi de prédire de l'information. Le flux de messages peut être considéré comme une séquence ordonnée par la date de création des messages. On appellera "item" un élément représentant un message (mot du message, opinion ou sujet extrait, etc.). À un temps t, plusieurs items apparaissent donc dans cette séquence : l'ensemble des items du message créé au temps t. Ce type de séquence est appelée "séquence complexe".

Dans le cas où les données sont formées d'une unique et longue séquence, l'extraction d'épisodes est une tâche essentielle. Un épisode est un motif temporel composé d'items "relativement proches", qui apparaît souvent tout au long de la séquence ou sur une partie de cette séquence (Mannila et al., 1997). Mannila (Mannila et al., 1997) a proposé les premiers algorithmes d'extraction d'épisodes : $Winepi$ et $Minepi$ qui seront la base de nombreux autres algorithmes proposés par la suite. Ces deux algorithmes extraient dans un premier temps les épisodes les plus petits, et forment incrémentalement des épisodes plus grands en se basant sur leur fréquence. Ces méthodes ont la caractéristique d'extraire un ensemble complet d'épisodes.

L'extraction d'épisodes dans des séquences complexes est une problématique récente qui nécessite un algorithme adapté pour prendre en compte l'existence de plusieurs items à chaque temps. Huang et Chang (Huang et Chang, 2008) proposent un algorithme appelé EMMA qui extrait un ensemble complet d'épisodes à partir d'une séquence complexe. Dans les deux premières phases, EMMA extrait un ensemble de motifs fréquents représentant des 1-uplet épisodes, associe un identifiant id à chaque 1-uplet épisode, puis encode la séquence avec ces