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

id. Les épisodes sont construits incrémentalement pendant une troisième phase en concaténant des id. EMMA définit deux notions : borne et borne projetée. Une borne d'un épisode P est un intervalle $[t_s, t_e]$ dans lequel P apparaît. Une borne projetée d'un épisode P représente l'intervalle de taille au maximum w dans lequel l'algorithme cherche les id pour étendre P. EMMA est un algorithme rapide et facile à adapter.

La communauté d'extraction de motifs fréquents admet que dans la plupart des applications, il est suffisant d'extraire un ensemble d'épisodes condensé et significatif. Par conséquent, les travaux récents se focalisent sur la détection d'épisodes comportant certaines caractéristiques et contraintes : épisodes maximaux (Gan et Dai, 2011), ou approximativement fermés et non dérivables (Gan et Dai, 2012).

Tout comme il est possible d'extraire des règles d'association à partir de motifs, des règles d'épisodes peuvent être extraites à partir d'épisodes. Ces règles d'épisodes peuvent aussi être utilisées dans un objectif de prédiction (Daurel, 2003). La majorité des règles d'épisodes construites à partir d'épisodes fermés ou maximaux ont la caractéristique d'avoir un antécédent long (composé de nombreux items). De notre point de vue, elles ne sont pas adaptées à la prédiction au plus tôt d'informations, car pour prédire une conséquence, il faut attendre l'apparition de la totalité des items de l'antécédent. Dans ce cas, pour détecter rapidement des événements, une règle d'épisodes avec un antécédent plus petit est plus pertinente qu'une règle avec un antécédent plus long. Par conséquent, notre premier objectif est d'extraire des règles d'épisodes composées d'un antécédent de taille minimale, que nous appellerons "règles d'épisodes minimales".

Dans la tâche d'identification au plut tôt d'informations, nous faisons l'hypothèse qu'il est inutile de chercher à former des règles d'épisodes très complexes ou très précises, mais des règles d'épisodes représentant les premiers signaux déclencheurs d'un événement. Nous trouvons donc qu'il est inutile d'extraire les règles d'épisodes contenant plusieurs fois le même item. Il est vrai qu'une apparition multiple d'un item porte plus d'informations, mais l'unicité des items permettra de diminuer la complexité de l'algorithme. Nous appelons "épisode sans répétition" un épisode composé d'items uniques. Notre second objectif est d'extraire des épisodes sans répétition.

Nous souhaitons pouvoir anticiper/prédire des informations "lointaines". Notre troisième objectif est donc d'extraire des règles d'épisodes ayant une conséquence éloignée temporellement de l'antécédent.

La notion "minimale" n'est pas nouvelle dans l'état de l'art. Elle a été proposée dans (Rahal et al., 2004) dans le cadre d'extraction de règles d'association fiables, peu fréquentes ayant l'antécédent le plus petit et la conséquence fixée par l'utilisateur à partir de données transactionnelles dans le but de prédire au plus tôt la conséquence. Cette approche est différente de notre travail. En effet, nous voulons extraire les règles d'épisodes (et non pas des règles d'association), sans répétition et avec une conséquence temporellement éloignée de l'antécédent.

Nous présentons maintenant l'approche proposé pour atteindre nos objectifs.

2 Notre approche : extraction de règles d'épisodes minimales

Rappelons que notre objectif est de pouvoir anticiper au plus tôt des événements de façon fiable. Pour cela, nous proposons un algorithme d'extraction de règles d'épisodes qui, en plus

d'être fréquentes et fiables comme celles extraites par les algorithmes de l'état de l'art, sont sans répétition, minimales et avec une conséquence éloignée temporellement de l'antécédent.

Nous utilisons les mêmes étapes d'initialisation et d'encodage comme dans l'algorithme EMMA, mais nous proposons une autre approche pour obtenir des règles d'épisodes avec les caractéristiques souhaitées. Le principe de notre algorithme est le suivant : chaque règle d'épisodes est construite en fixant le préfixe de la future règle (1-uplet épisode), puis en fixant la conséquence, qui est la plus éloignée possible de ce préfixe, et enfin en complétant l'antécédent avec des 1-uplet épisodes les plus proches possibles du préfixe. Nous détaillons ci-dessous cet algorithme.

2.1 Déroulement de l'algorithme

Identification du préfixe : Chaque id obtenu dans la deuxième étape de EMMA représente un préfixe d'un épisode potentiel.

Soit id_i un préfixe, nous construisons $P_{roj}ID_{fin}(id_i)$ la liste des id apparaissant loin de id_i (dans les dernières positions des bornes projetées de id_i). Nous construisons également $P_{roj}ID_{deb}(id_i)$, la liste des id apparaissant à proximité de id_i (présents au début des bornes projetées de id_i). La taille des intervalles "début" et "fin" est une proportion de w.

Identification de la conséquence : id_i est étendu avec les éléments $id_j \in P_{roj}ID_{fin}(id_i)$ qui représentent une conséquence potentielle d'un épisode dont le premier élément est id_i, formant ainsi un épisode candidat $< id_i, id_j >$. La fréquence de cet épisode candidat est calculée en utilisant le nombre de bornes dans sa liste de bornes. Si $< id_i, id_j >$ n'est pas fréquent, on arrête cette itération et on considère que id_j ne peut pas être une conséquence de id_i. On itère de nouveau pour étendre id_i avec d'autres $id \in P_{roj}ID_{fin}(id_i)$. Si $< id_i, id_j >$ est fréquent, alors la règle $id_i \rightarrow id_j$ est construite. Si elle est fiable, elle est considérée comme minimale. Elle n'est donc plus étendue et elle est ajoutée à la liste finale des règles d'épisodes. Si la règle est fréquente mais pas fiable, alors le 2-uplet épisode $< id_i, id_j >$ est étendu pour compléter l'antécédent.

Complétion de l'antécédent : Rappelons que nous cherchons à compéter l'antécédent de façon à obtenir non seulement une règle fiable mais aussi une règle avec un antécédent minimal. Les itérations de complétion de l'antécédent s'arrêtent donc dès que la règle d'épisodes est $fiable$, ou s'il n'y a plus d'épisode candidat.

À partir de $< id_i, id_j >$, id_i est étendu itérativement avec les id dans sa liste $P_{roj}ID_{deb}(id_i)$. Soit $id_s \in P_{roj}ID_{deb}(id_i)$, si l'épisode candidat $< id_i, id_s, id_j >$ n'est pas fréquent, alors le préfixe de l'antécédent id_i est étendu avec un autre $id \in P_{roj}ID_{deb}(id_i)$, sinon la confiance de la règle $id_i, id_s \rightarrow id_j$ est calculée. Si la règle n'est pas fiable, alors l'épisode $< id_i, id_s >$ est étendu avec d'autres $id \in P_{roj}ID_{deb}(id_i, id_s)$, sinon elle est acceptable, et donc minimale, son antécédent n'est donc plus étendu et elle est ajoutée à la liste finale de règles d'épisodes.

Épisode sans répétition : Comme nous l'avons mentionné, notre algorithme a pour but d'éviter l'apparition multiple d'un item dans les règles d'épisodes. À chaque fois qu'un épisode doit être étendu, une vérification est faite pour éviter de l'étendre avec des id ayant des items en commun, ce qui aura également l'avantage de réduire la complexité de l'algorithme.

3 Conclusion et perspectives

Dans ce travail préliminaire, nous proposons un algorithme de fouille de séquence de données dans le but de prédire au plus tôt et de façon fiable, des événements. Notre objectif à court terme est de fouiller des messages issus des réseaux sociaux et de prédire l'apparition d'informations. Nous avons mis en évidence que l'extraction de règles d'épisodes est adaptée à nos données représentées sous la forme d'une séquence complexe ordonnée selon le temps. Pour atteindre nos objectifs, nous avons déterminé plusieurs caractéristiques des règles d'épisodes à extraire : elles doivent être sans répétition, fréquentes, fiables, ayant l'antécédent le plus petit possible et la conséquence éloignée temporellement de l'antécédent.

L'algorithme, tel que nous l'avons défini, ne permet d'extraire que des conséquences de taille 1 (1-uplet épisodes). Il est évident qu'une conséquence plus longue sera plus porteuse d'informations sur les futurs événements. Une de nos perspectives est donc d'adapter cet algorithme de façon à ce qu'il puisse extraire des conséquences plus longues.

Références

Daurel, T. (2003). *Représentations Condensées d'Ensembles de Règles d'Association*. Ph. D. thesis, L'Institut National des Sciences Appliquées de Lyon.

Gan, M. et H. Dai (2011). Fast mining of non-derivable episode rules in complex sequences. In *Modeling Decision for Artificial Intelligence*, pp. 67–78. Springer.

Gan, M. et H. Dai (2012). Mining condensed sets of frequent episodes with more accurate frequencies from complex sequences. *Int. J. of Innovation Computing, Information and Control 8*(1), 453–470.

Huang, K.-Y. et C.-H. Chang (2008). Efficient mining of frequent episodes from complex sequences. *Information Systems 33*(1), 96–114.

Mannila, H., H. Toivonen, et A. I. Verkamo (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery 1*(3), 259–289.

Rahal, I., D. Ren, W. Wu, et W. Perrizo (2004). Mining confident minimal rules with fixed-consequents. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pp. 6–13. IEEE.

Summary

The messages posted daily on social networks are valuable source of information. Their mining can be used to predict the emergence of information. Our goal in this article is to propose an épisode rules mining algorithm with the objective of early and confident information prediction.