Explorer les traces d'apprenants avec l'extraction d'épisodes séquentiels

Béatrice Fuchs
Université de Lyon, UJML3, Laboratoire LIRIS, IAE
1C avenue des Frères Lumière - CS 78242 69372 LYON CEDEX 08
beatrice.fuchs@liris.cnrs.fr

1 Introduction

Des volumes importants de traces sont accumulés par les apprenants dans leur environnement numérique d'apprentissage. Les traces numériques se présentent sous la forme de séquences d'actions contextualisées et temporellement situées. La collecte et l'analyse des ces traces qui témoignent de leur activité est importante afin de mieux comprendre leurs difficultés en vue de leur proposer une assistance adaptée en conséquence ou d'améliorer les outils existants. Une façon d'étudier ces traces est l'extraction de connaissances à partir de données (ECD), qui vise à extraire des connaissances à partir de données dans un processus interactif et itératif. Une des méthodes adaptée à l'exploration des traces est l'extraction d'épisodes séquentiels qui prend en compte les dimensions événementielle et temporelle des traces pour mettre en évidence des séquences typiques d'actions réalisées par des utilisateurs afin de caractériser leur parcours ou leur comportement. L'objectif est de répondre à des questions diverses telles que la catégorisation des utilisateurs, la comparaison de leurs parcours, la réussite ou l'échec en fonction des activités réalisées, *etc*.
Nous proposons une approche qui s'appuie sur l'ECD, la visualisation et les interactions avec l'utilisateur afin d'une part assister son travail et d'autre part tirer parti de son expertise pour mettre en évidence des connaissances du domaine étudié et les exprimer sous la forme de contraintes liées au domaine.
Pour cela, notre démarche s'appuie sur trois éléments principaux : un système dédié à la gestion de base de traces (SBT), un processus d'ECD mis en œuvre dans le prototype DISKIT et une approche visuelle et interactive pour assister le processus, en particulier l'interprétation (TRANSMUTE).

2 L'exploration des traces

Il s'agit de mettre en place une sorte de *laboratoire* d'analyse des traces en combinant un ensemble d'outils pour répondre à des problématiques diverses exploitant les traces numériques.

2.1 Système de gestion de base de traces

Un système à base de traces (SBT) est dédié à la collecte, la modélisation, le stockage et la manipulation de traces. Une trace numérique décrit un ensemble d'actions réalisée par un ou des utilisateurs. Ces actions sont typées, contextualisées et temporellement situées et décrites dans un modèle de trace. Un SBT met à disposition un ensemble de traces ainsi que des opérations génériques de *transformation* pour les manipuler. Parmi les transformations, la *réécriture* permet de mémoriser l'interprétation de sous-séquences de la trace en créant une trace d'un plus haut niveau d'abstraction et augmentant ainsi progressivement son niveau de compréhension. Le système à base de traces sert donc non seulement de *container* de traces, mais il permet également de mémoriser des interprétations réalisées sur des traces en vue d'étudier des phénomènes complexes et en particulier l'apprentissage humain.

2.2 DISKIT

DISKIT met en œuvre le processus d'ECD qui, à partir d'une ou plusieurs traces, extrait un ensemble d'épisodes séquentiels fréquents. DISKIT procède en trois étapes principales : pré-traitement, fouille, puis post-traitement. Le pré-traitement collecte une ou plusieurs traces sur lesquelles d'éventuelles transformations sont réalisées et les met en forme pour la fouille. Les traces sont collectées soit sous la forme d'un fichier texte soit à partir d'un SBT.
La fouille produit un ensemble d'épisodes séquentiels qui sont ensuite mis en forme durant le post-traitement afin d'être compréhensibles, puis restitués. DISKIT est un outil destiné à être complété par d'autres méthodes en amont et/ou en aval afin d'étudier des problématiques plus précises autour des traces.

2.3 Épisodes séquentiels

L'étape de fouille est assurée par un algorithme qui extrait des épisodes séquentiels et des règles séquentielles à un conséquent à partir d'une ou plusieurs séquences d'événements [1]. Un épisode séquentiel est une séquence de types d'actions $a_1, a_2, ..., a_n$ qui se répète de façon récurrente dans les traces, avec une fréquence supérieure à un certain seuil. Si $a_1, a_2, ..., a_n$ est un épisode séquentiel, une règle séquentielle à un conséquent est de la forme : $a_1, a_2, ..., a_{n-1} \rightarrow a_n$ avec un degré de confiance indiquant la proportion des épisodes $a_1, a_2, ..., a_{n-1}$ qui sont suivis de a_n.
DMT4SP prend en charge des contraintes pour limiter le nombre d'épisodes séquentiels générés :

1. DMT4SP : Data Mining Techniques For Sequence Processing, http://liris.cnrs.fr/~crigotti/dmt4sp.html

– le nombre d'occurrences et/ou de séquences minimum,
– la longueur min / max,
– la confiance pour les règles séquentielles,
– l'intervalle de temps min / max entre événements,
– l'intervalle de temps maximal entre les premier et dernier événements des épisodes séquentiels,
– une contrainte de préfixe sous la forme d'une suite de types d'actions par lesquelles les épisodes doivent débuter ainsi qu'une contrainte de suffixe spécifiant un type d'action terminal de l'épisode séquentiel.

2.4 Contraintes

Un des problèmes de la fouille de données en général et plus particulièrement de la fouille d'épisodes séquentiels est la surabondance de résultats produits caractérisés par une forte redondance combinatoire. Pour pallier à cette difficulté, DISKIT prend en charge plusieurs options et contraintes supplémentaires.
Lors du pré-traitement, il est possible de filtrer des types d'obsels pour éliminer par exemple ceux qui «bruitent» trop la fouille et de de découper les traces selon un ou plusieurs attributs constituant un *contexte* afin d'imposer à la fouille de ne rechercher que des épisodes qui ont du sens. Ce découpage des traces présente l'avantage d'explorer des traces plus petites ce qui accélère le temps de traitement et limite de façon très importante le nombre d'épisodes inutiles générés.
DISKIT applique également des contraintes lors du post-traitement. La propriété de fermeture des épisodes permet d'obtenir une représentation «compacte» des épisodes séquentiels. La contrainte de «pattern» consiste à ne sélectionner que les épisodes qui contiennent un motif défini en terme de séquence de types d'événements. Ceci permet de focaliser l'exploration sur un motif particulier afin d'étudier les autres actions qui l'accompagnent lors du parcours. De la même façon il est possible de spécifier un motif qui ne doit pas être présent et ainsi de filtrer les résultats de la fouille. Enfin une contrainte d'appariement permet de contraindre les valeurs d'un ou plusieurs attributs caractéristiques des actions de façon à ce que leurs valeurs soient identiques dans les différentes occurrences d'épisodes.

2.5 Visualisations et interactions

Afin de permettre à l'utilisateur d'intervenir pendant le processus d'ECD, TRANSMUTE a été conçu pour visualiser une trace et les épisodes issus de la fouille. Il permet de paramétrer l'affichage des actions en fonction de leurs caractéristiques, (type, attributs, *etc.*), et d'assister la phase d'interprétation. Les épisodes séquentiels sont caractérisés par des mesures d'intérêt afin de les trier et ainsi mettre en avant ceux qui sont potentiellement les plus «intéressants». L'utilisateur peut sélectionner des épisodes et observer l'impact de son choix sur la trace ainsi que sur la liste des épisodes restants. Il peut voir et accéder à toutes les occurrences de l'épisode sélectionné dans la trace afin d'estimer la pertinence de son choix. Pour un épisode sélectionné donné, l'analyste peut décider de créer une observation qui constitue son interprétation de l'épisode, et la mémoriser dans une trace transformée. À l'issue d'une sélection, la liste des épisodes est mise à jour dynamiquement : toutes les occurrences d'épisodes ayant au moins une action en commun avec les occurrences de l'épisode sélectionné sont identifiés et éliminés. Les mesures associées aux épisodes sont alors recalculées (notamment le support) et les motifs ne satisfaisant plus les contraintes sont éliminés à leur tour et la liste des épisodes restants est triée en conséquence. Ce processus permet de diminuer graduellement le nombre de résultats et ainsi de faciliter le prochain choix de l'analyste en l'aidant à se focaliser sur d'autres épisodes.

3 Application aux learning analytics

Des expérimentations des propositions ont été réalisées sur les traces de Tamagocours, un jeu sérieux collaboratif pour l'apprentissage de règles de diffusion de ressources numériques. Il s'agit d'apporter des outils pour l'exploration des traces afin d' apporter des réponses à des questions qui peuvent être générales sur l'apprentissage humain ou plus spécifiques.
Quelques pistes de travail sont évoquées ci-dessous, certaines sont encore exploratoires.

Catégorisation des utilisateurs. Les épisodes séquentiels associés aux utilisateurs peuvent être exploités pour catégoriser les utilisateurs en fonction de la fréquence observée des différents épisodes. Des outils tels que l'analyse de concepts formels sont notamment en cours d'étude car ils s'avèrent intéressants pour l'acquisition de connaissances. Les catégories obtenues peuvent ensuite contribuer à la personnalisation de l'enseignement, ainsi qu'à l'évolution des outils.

Caractériser l'échec des apprenants. Il s'agit d'explorer les traces afin de mettre en évidence les séquences d'actions susceptibles d'expliquer la réussite ou l'échec des apprenants. Il s'agit de d'étudier les épisodes ou les règles séquentielles menant aux actions caractéristiques d'un échec, ce qui permet d'émettre différentes hypothèses sur les raisons de l'échec et de proposer une assistance adaptée pour remédier aux difficultés des apprenants.

Corréler l'absence d'une activité avec l'échec. Il s'agit d'étudier comment influe la réalisation ou non d'une activité donnée sur la réussite ou l'échec final de l'utilisateur. Ce type de requête implique de trouver des séquences d'action où un type d'action donné est *présent* ou *absent*. L'absence d'une action dans un épisode est difficile à détecter et nécessite des mécanismes de filtrage particulier proches de ceux utilisés dans TRANSMUTE.

Les perspectives s'orientent principalement sur l'étude d'applications afin de mettre en évidence des besoins récurrents et de rechercher des outils génériques pour les traiter. Des besoins spécifiques sont également intéressants à étudier.