Extraction d'épisodes séquentiels à partir de Traces : application au jeu Tamagocours.

Béatrice Fuchs[1]

Université de Lyon, Jean Moulin, IAE, LIRIS, F-69 008, Lyon, France,
beatrice.fuchs@liris.cnrs.fr

Abstract. Nous proposons une démarche de découverte de connaissances à partir de traces afin d'analyser les traces d'activités laissées par des utilisateurs. Elle est mise en œuvre dans un processus d'extraction de connaissances à partir de données (ECD) qui génère des épisodes séquentiels ou des règles séquentielles. Ce processus a été mis en œuvre dans DISKIT et se décline sous deux formes. Dans une forme interactive, l'analyste peut étudier une trace individuellement à l'aide d'une interface graphique interactive. Dans une forme autonome, plusieurs traces peuvent être analysées simultanément sans interaction mais à l'aide de contraintes sémantiques. Cette démarche est mise en œuvre sous ces deux formes pour l'analyse des traces du jeu Tamagocours.

1 Introduction

Nous proposons d'étudier l'analyse des traces dans le cadre d'une approche interactive de l'extraction de connaissances à partir de données (ECD) qui vise à assister le travail de l'analyste lors de l'interprétation. Le processus a été mis en œuvre dans DISKIT et s'appuie sur une étape de fouille de données assurée par le prototype DMT4SP[1] qui extrait des motifs séquentiels sous la forme d'*épisodes séquentiels* ou de *règles séquentielles* à un conséquent. Un épisode séquentiel est une séquence d'actions $a_1, a_2, ..., a_n$ qui se répète de façon récurrente dans les traces. Une règle séquentielle est de la forme $a_1, a_2, ..., a_{n-1} \rightarrow a_n$, où $a_1, a_2, ..., a_{n-1}$ est un épisode séquentiel et a_{n-1} une action. Elle est associée à une mesure de confiance calculée comme le rapport du nombre d'occurrences de $a_1, a_2, ..., a_{n-1}$ avec le nombre d'occurrences de $a_1, a_2, ..., a_n$. Les régularités détectées par la fouille sont des hypothèses possibles de connaissances. Mais le passage des régularités aux connaissances requiert une expertise humaine qui se heurte à plusieurs difficultés, notamment la profusion des résultats, et la difficulté de leur interprétation. Pour aborder ces difficultés nous avons étudié deux approches pour assister le processus : une approche visuelle et interactive, et une approche autonome. Les deux approches ont été appliquées au jeu sérieux Tamagocours pour l'apprentissage de règles de diffusion de ressources numériques.

2 Positionnement

Avec l'émergence de la fouille de données et d'algorithmes performants pour analyser et trouver des régularités dans de grandes quantités de données, on s'est rapidement

[1] Data Mining Technique For Sequence Processing

rendu compte qu'il était important d'aider l'analyste en suscitant sa mobilisation cognitive afin de l'assister lors de l'interprétation des résultats. Ainsi des travaux se sont intéressés à tirer parti des travaux sur la visualisation pour intégrer l'humain dans le processus de découverte de connaissances, ce qui a abouti à la fouille visuelle des données qui s'est développée ces dernières années [1]. L'analyse visuelle [2] vise à faire émerger des connaissances en combinant la puissance de traitement, la visualisation et l'expertise humaine, résumé par *"Analyse first, show the important, zoom, filter and analyse further, details on demand"*. Il s'agit donc de donner un rôle central et actif à l'humain dans le processus de découverte de connaissances [3]. Dans cette mouvance, les travaux se sont focalisés à l'analyse visuelle, par exemple [4] et/ou interactive [5] pour les règles d'association. Un des premiers problèmes de la fouille est la surabondance des résultats qui rend difficile leur exploration. Les travaux qui se sont intéressés à ce problème ont d'abord étudié des mesures d'intérêt indépendantes du domaine afin de caractériser la qualité des résultats de la fouille [6]. Puis les travaux ont visé à intégrer des connaissances du domaine dans le processus d'ECD, sous forme de bases de connaissances, d'ontologies ou de mesures dépendantes du domaine [7]. La plupart des approches se sont intéressées aux règles d'association, ou aux règles temporelles [8], [9], mais à notre connaissance, peu de travaux se sont intéressés aux épisodes séquentiels, aussi bien du point de vue des mesures d'intérêt que de la fouille visuelle. Par ailleurs, les interactions étudiées dans la littérature s'intéressent davantage à l'affichage les résultats, mais l'assistance à la construction d'un modèle est encore peu abordée dans la littérature [3].

3 Démarche d'analyse

Nous proposons une démarche d'analyse des traces d'activité qui s'appuie d'une part sur la visualisation et l'interactivité et d'autre part l'introduction de connaissances du domaine dans le processus. Notre approche s'articule autour de trois éléments principaux : un processus d'ECD mis en œuvre dans DISKIT, un système à base de traces (SBT) et une interface graphique interactive pour assister l'interprétation (TRANSMUTE). DISKIT assure le processus d'ECD qui, à partir d'une ou plusieurs traces, extrait un ensemble de motifs 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 en fonction des souhaits de l'analyste, et les met en forme pour la fouille. La fouille est assurée par un prototype d'extraction d'épisodes séquentiels qui produit un ensemble de motifs séquentiels. Ces motifs sont mis en forme durant le post-traitement afin d'être compréhensibles, puis restitués. Un système à base de traces (SBT) est dédié à la collecte, le stockage, et à la manipulation des traces. Il met à disposition un ensemble de traces ainsi que des opérations génériques de *transformation*. L'approche visuelle et interactive est assurée par TRANSMUTE dans laquelle l'analyste joue un rôle moteur. Il peut interagir avec la trace analysée ainsi qu'avec les motifs proposés par la fouille. Une interface graphique personnalisable propose des modalités d'interactions afin d'assister le travail d'analyse et repérer les motifs potentiellement intéressants : tri des motifs selon plusieurs critères, filtrage dynamique en fonction des sélections opérées par l'analyste.

4 Le cadre du jeu Tamagocours

Dans le cadre du jeu Tamagocours, deux types d'analyse ont été réalisées. La première est visuelle et interactive exploite le SBT et TRANSMUTE et ne permet l'analyse simultanée que d'une seule trace. La deuxième est autonome et exploite plusieurs traces simultanément. Lors de l'analyse interactive, TRANSMUTE a été paramétré de façon à ce que les actions du jeu apparaissent sous une forme visuelle facilement interprétable (figure 1). TRANSMUTE permet de visualiser sur une même ligne temporelle la trace en cours d'analyse et les motifs sélectionnés par l'analyste. Les motifs candidats issus de la fouille peuvent être ordonnés selon une ou plusieurs mesures d'intérêt. La sélection d'un motif dans cette liste s'accompagne d'un filtrage automatique des motifs restants en fonction des actions en commun avec le motif sélectionné, ce qui élimine de nombreux motifs redondants, favorisant ainsi la focalisation de l'analyse sur d'autres motifs. L'analyse en mode autonome permet l'analyse de plusieurs traces simultané-

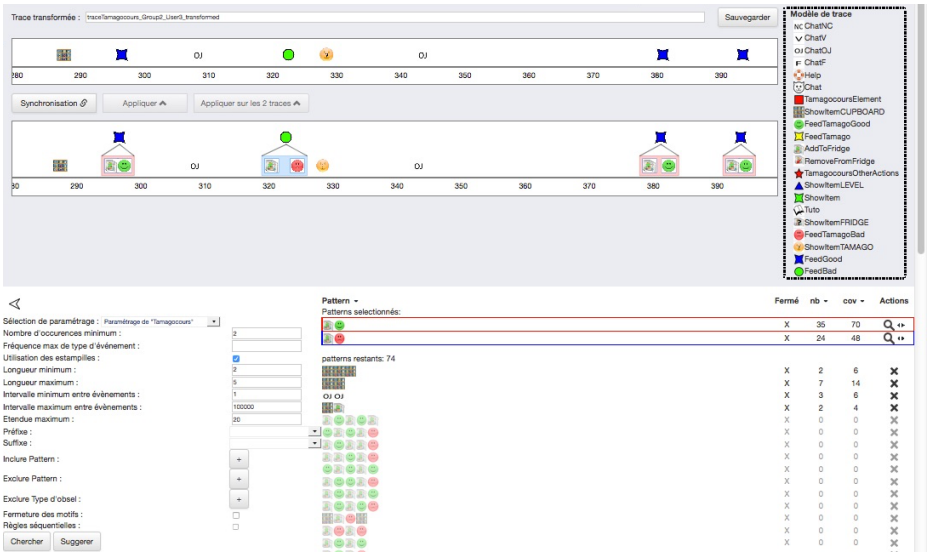

Fig. 1. L'interface d'analyse de Transmute.

ment, mais pas de façon interactive car l'interface de TRANSMUTE n'est pas conçue pour l'affichage de plusieurs traces.

Afin de limiter le nombre de motifs, de nombreux types de contraintes sont paramétrables. Les contraintes de pré-traitement permettent de limiter les types d'action à rendre en compte pour l'analyse. L'étape de fouille permet de préciser le type d'analyse : épisodes séquentiels ou règles séquentielles. Les contraintes de support permettent de spécifier le nombre minimum d'occurrences d'un motif et/ou de traces contenant le motif, et pour les règles la confiance des règles. Les contraintes syntaxiques imposent la longueur minimum et/ou maximum des motifs, un type d'action terminal ou un préfixe.

Les contraintes temporelles limitent l'intervalle de temps entre la première et la dernière action des motifs ou entre deux actions consécutives. Les contraintes de post-traitement complètent l'éventail des possibilités de la fouille avec des contraintes plus ciblées. La fermeture des motifs permet d'obtenir une représentation plus compacte des motifs en assurant qu'aucun motif n'est inclus dans un autre motif ayant le même support. Des restrictions sur la présence ou l'absence d'une ou plusieurs patterns d'actions dans les motifs résultats peuvent être également données. Enfin il est possible de préciser un ensemble d'attributs pour lesquels la valeur est imposée constante à l'intérieur de chaque occurrence de motif. Cette contrainte indépendante du domaine a montré une élimination drastique de motifs non pertinents et s'est avérée très utile pour filtrer un très grand nombre d'occurrences de motifs redondants, et par conséquent, de motifs redondants.

5 Conclusion

Nous proposons une approche de l'ECD centrée sur l'analyste afin d'appréhender la découverte de connaissances à partir de traces. Elle s'appuie sur la visualisation, les interactions et des contraintes permettant de transcrire simplement des connaissances du domaine. L'analyse autonome permet une étude globale des traces tandis que l'analyse interactive vise à examiner plus finement et précisément les traces individuellement.

References

1. Enrico Bertini and Denis Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, pages 12–20. ACM, 2009.
2. Daniel A Keim, Jörn Kohlhammer, Geoffrey Ellis, and Florian Mansmann. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.
3. Matthijs van Leeuwen. Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, pages 169–182. Springer, 2014.
4. Gwenael Bothorel. *Algorithmes automatiques pour la fouille visuelle de données et la visualisation de règles d'association: application aux données aéronautiques*. PhD thesis, 2014.
5. Julien Blanchard, Fabrice Guillet, and Henri Briand. Interactive visual exploration of association rules with rule-focusing methodology. *Knowledge and Information Systems*, 13(1):43–75, 2007.
6. Fabrice Guillet and Howard J Hamilton. *Quality measures in data mining*, volume 43. Springer, 2007.
7. Claudia Marinica, Fabrice Guillet, and Henri Briand. Post-processing of discovered association rules using ontologies. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pages 126–133. IEEE, 2008.
8. Julien Blanchard, Fabrice Guillet, and Régis Gras. On the discovery of significant temporal rules. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, pages 443–450. IEEE, 2007.
9. Julien Blanchard, Fabrice Guillet, and Régis Gras. Assessing the interestingness of temporal rules with sequential implication intensity. In *Statistical Implicative Analysis*, pages 55–71. Springer, 2008.