Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces

Pierre-Loup Barazzutti, Amélie Cordier, Béatrice Fuchs

Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces

Pierre-Loup Barazzutti * Amélie Cordier *, Béatrice Fuchs **.

* LIRIS CNRS, UMR 5205, Université Lyon 1
**LIRIS CNRS, UMR 5205, Université Lyon 3
prenom.nom@liris.cnrs.fr

Résumé. Alors que l'extraction de connaissances à partir de données (ECD) est un processus qualifié d'interactif et d'itératif, l'interactivité des outils est souvent limitée et son étude est relativement récente. Elle est pourtant déterminante lors de l'interprétation pour choisir les motifs qui deviendront des connaissances. Nous proposons TRANSMUTE, un outil d'assistance à l'interprétation dans le processus d'ECD, dans le cadre de la recherche d'épisodes séquentiels à partir de traces. La phase d'interprétation est itérative et à chaque itération les résultats de la fouille sont mis à jour dynamiquement en fonction des interactions avec l'analyste. Des outils de visualisation et des mesures de qualité indépendantes du domaine permettent de caractériser l'intérêt des motifs à interpréter pour faciliter leur choix et accompagner le travail de l'analyste afin de l'aider à se focaliser plus rapidement sur les motifs potentiellement intéressants.

1 Introduction

L'extraction de connaissances à partir de données (ECD) vise à analyser des données dans un processus itératif et interactif composé de plusieurs étapes. Durant ce processus, de nombreuses interactions ont lieu avec un utilisateur humain expert du domaine ou analyste qui y joue un rôle moteur. Pour aider l'analyste dans sa tâche, de nombreux outils permettent toutes sortes de visualisations réalistes des résultats, mais leurs capacités d'interaction restent le plus souvent limitées à des manipulations graphiques. Nous avons développé dans TRANSMUTE des fonctionnalités illustrant l'interactivité dans le processus d'ECD en vue d'apporter une assistance à toutes les étapes. Nous nous focalisons dans cet article sur l'étape d'interprétation en proposant un scénario dans lequel l'analyste peut observer dynamiquement l'effet de ses actions et de ses choix au fur et à mesure de l'avancement de son travail. Dans la suite de l'article, nous rappelons d'abord le cadre d'application de ce travail. Puis nous présentons le processus d'ECD et la démarche d'interprétation mis en place, puis leur mise en œuvre dans TRANSMUTE. Pour finir, nous concluons sur l'état actuel du développement et les perspectives.

Ce travail a été réalisé dans le cadre du projet OFS - Open Food System, programme investissements d'avenir.

2 Traces et système à base de traces

Nous nous plaçons dans le cadre de l'étude des *traces*, bien que les concepts présentés ici puissent s'appliquer à n'importe quelles données temporellement situées, comme par exemple la description d'une partition musicale qui est utilisée comme application jouet dans la démonstration. Une trace est constituée d'une séquence d'éléments observés temporellement situés appelés des *obsels*. Elle est associée à un *modèle de trace* décrivant les types d'obsels, leurs attributs et leurs relations avec d'autres types d'obsels. Le modèle de trace permet d'interpréter les informations de la trace pour faciliter son exploitation ultérieure. Les traces sont manipulées par un ensemble d'opérations élémentaires appelées *transformations* qui sont de différents types : filtrage d'obsels, fusion de traces, etc. Parmi ces transformations, la *réécriture* crée une nouvelle trace appelée *trace transformée* qui vise à augmenter progressivement le niveau de compréhension et d'abstraction de la trace initiale. La réécriture consiste à construire une nouvelle trace t_2 à partir d'une trace primaire t_1 en remplaçant dans t_2 des motifs, c'est-à-dire des séquences d'obsels, de t_1 par de nouveaux types d'obsels résumant chaque motif. Un système à base de traces modélisées est un système permettant de collecter, de traiter et de visualiser des traces. Le framework *kTBS* (*kernel for Trace Based System*) [1], (Champin et al., 2013) réifie cette notion de système à base de traces de façon générique. La réécriture de traces se situe au cœur du dispositif mis en place dans Transmute.

3 Extraction de connaissances à partir de traces

Les traces sont étudiées dans le cadre d'un processus d'ECD mis en oeuvre dans DisKit [2], dans un cycle composé des étapes principales 1) prétraitement (sélection de trace, transformation), 2) fouille, et 3) post-traitement (visualisation, interprétation). L'étape de fouille de DisKit utilise DMT4SP [3], un prototype d'extraction de motifs et de règles à partir d'une ou plusieurs séquences d'événements, conformément à la sémantique d'occurrence minimale définie dans (Mannila et al., 1997). Durant le post-traitement, les résultats de la fouille sont présentés à l'analyste qui doit choisir les motifs qu'il estime les plus pertinents compte tenu de ses connaissances du domaine. La fouille produit très souvent un nombre important de motifs caractérisés par une forte redondance combinatoire. Il n'est pas facile pour l'analyste de s'y retrouver et faire un choix pertinent. Nous proposons une démarche itérative et interactive pour l'aider à appréhender méthodiquement cette étape (fig. 1). La liste des motifs est complétée par des indicateurs d'intérêt permettant à l'analyste de les trier et ainsi mettre en avant les motifs potentiellement les plus intéressants. Actuellement ces indicateurs sont la fréquence, la longueur et la *couverture* permettant de caractériser le nombre d'obsels distincts couverts par le motif. Lors de la sélection d'un motif, l'analyste peut voir l'impact de son choix à la fois sur la trace et sur la liste des motifs. Sur la trace,

1. `http://tbs-platform.org/tbs/doku.php`
2. DIScovering Knowledge from Interaction Traces
3. Data Mining Techniques For Sequence Processing, `http://liris.cnrs.fr/~crigotti/dmt4sp.html`

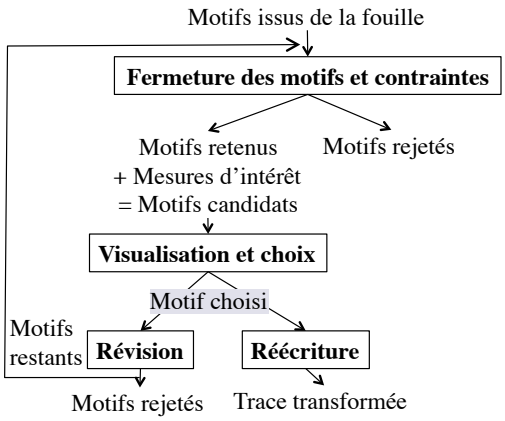


FIG. 1 – *Le scénario d'interprétation de* TRANSMUTE

il peut directement voir et accéder à toutes les occurrences du motif sélectionné. Pour un motif sélectionné donné, il peut alors créer un nouveau type d'obsel qui se substituera à chaque occurrence du motif sélectionné dans la trace transformée. Lorsqu'un motif est sélectionné, tous les motifs dont les occurrences ont au moins un obsel en commun avec les occurrences du motif sélectionné sont éliminés. Les indicateurs des motifs restants sont alors recalculés et la liste est triée en conséquence. Les motifs ne répondant plus aux contraintes fixées initialement par l'analyste (support) sont automatiquement éliminés, ce qui diminue graduellement le nombre de résultats et facilite ainsi le prochain choix de l'analyste qui peut se focaliser sur d'autres motifs. Lorsque l'analyste a sélectionné tous les motifs et défini les types d'obsels les remplaçant, il peut confirmer les changements et déclencher la création d'une nouvelle trace transformée. Ceci a pour conséquence de supprimer définitivement tous les motifs précédemment estompés de l'interface. L'analyste a la possibilité de continuer l'interprétation soit à partir des résultats restants, soit en réitérant le processus avec d'autres contraintes sur la même trace ou sur la trace nouvellement créée.

4 Transmute

TRANSMUTE était initialement conçu pour visualiser et transformer des traces de façon interactive et fournit pour cela un ensemble de fonctionnalités pour créer des obsels, manipuler des traces, visualiser et stocker les résultats de transformations. La motivation sous-jacente est de fournir à l'analyste des outils interactifs pour assister l'interprétation des traces. Dans ce travail, nous avons enrichi TRANSMUTE avec des fonctionnalités d'ECD qui permettent à présent de paramétrer et lancer un algorithme de fouille et d'afficher les résultats afin de construire des traces transformées, et d'assister l'étape de post-traitement du processus d'ECD ainsi que la manipulation des résultats. L'architecture de TRANSMUTE s'articule autour de plusieurs modules (fig. 2). Il

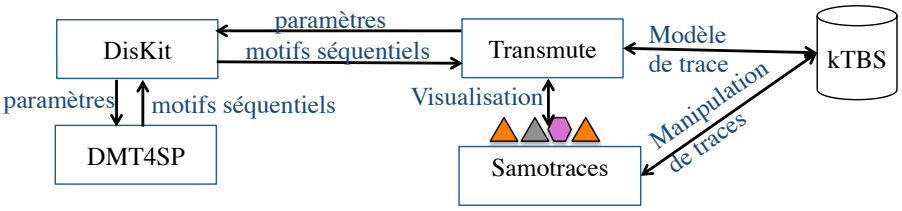
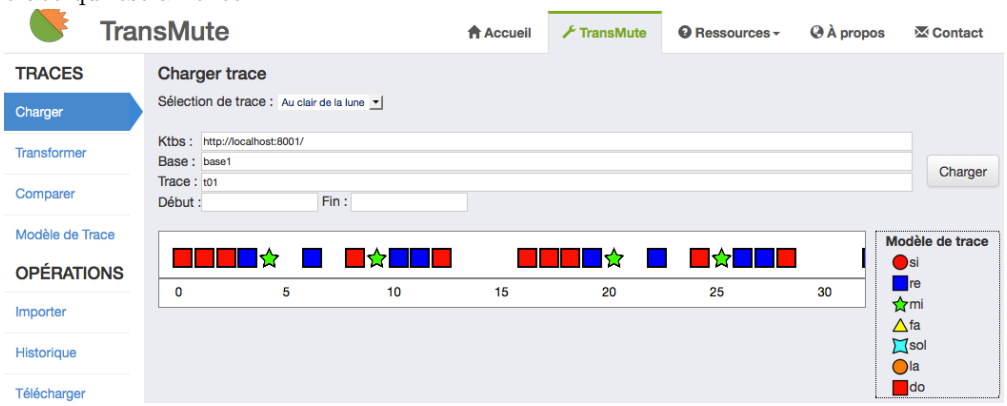


FIG. 2 – TRANSMUTE *: architecture*

s'appuie d'abord sur *Samotraces*[4], un framework Javascript permettant de construire des visualisations de traces personnalisables en paramétrant l'affichage des éléments observés en fonction de leur type, leurs attributs, *etc.* Samotraces permet également la communication avec le gestionnaire de traces appelé kTBS fournissant toutes les manipulations de base sur les traces : collecte de traces, traitement, transformations, exportations, etc.

DISKIT met en oeuvre les étapes du processus d'ECD. L'étape de fouille utilise DMT4SP pour trouver les épisodes séquentiels fréquents à partir d'une ou plusieurs séquences d'événements. DMT4SP prend en compte de nombreuses contraintes telles que le support, des contraintes temporelles, (fenêtre temporelle maximale, gap min / max), des contraintes syntaxiques (longueur min / max, préfixe, suffixe), etc. Ces contraintes sont complétées par DISKIT en post-traitement avec d'autres types de contraintes telles que la recherche de motifs fermés (tels que définis dans Tatti et Cule (2010)), des contraintes de présence ou d'absence de motifs spécifiques, afin de réduire le nombre de motifs produits. Les résultats de la fouille sont mis en forme et restitués. Actuellement, DISKIT ne traite qu'une seule trace à la fois.

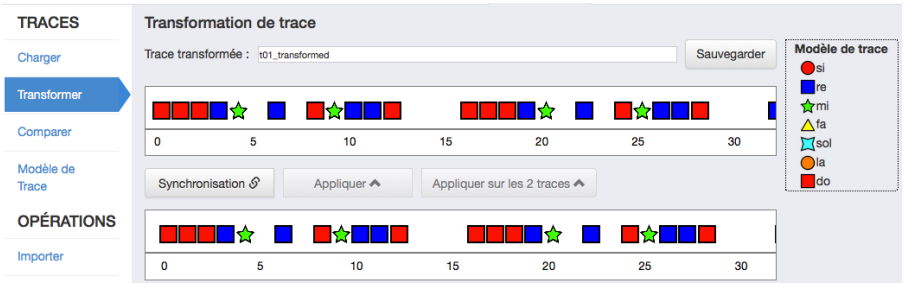
Une session typique d'utilisation de TRANSMUTE est la suivante. Tout d'abord l'utilisateur choisit, parmi les traces disponibles stockées dans la base de traces, une trace qui est affichée :



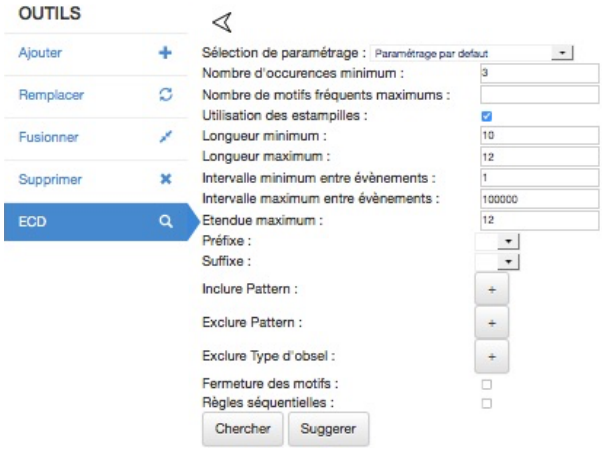
Sur cette trace, l'utilisateur choisit ensuite le type de transformation appliquer à cette trace, en l'occurrence il s'agit dans cette démonstration d'un processus d'ECD. La

4. `https://github.com/bmathern/samotraces.js`

trace en cours de traitement et la trace transformée sont affichées avec leur modèle :



L'utilisateur est alors invité à choisir les paramètres à utiliser puis à lancer la fouille :



Le miner est lancé et les résultats sont mis en forme et affichés. L'interface de d'interprétation interactive de Transmute comporte trois parties principales (fig. 3) : (1) La première partie montre en haut la trace en cours d'analyse et en bas la trace en cours de transformation où les motifs sont affichés. (2) Sur la droite, un résumé du modèle de trace permet de faciliter la compréhension de la trace et des motifs. (3) La troisième partie montre la liste des motifs issus de la fouille.

5 Conclusion et perspectives

Nous avons présenté une démarche d'interprétation interactive dans un processus d'ECD mise en oeuvre dans Transmute. L'analyste peut directement voir l'impact de ses actions sur la trace et sur les résultats de la fouille restant à examiner. Il reste de nombreuses fonctionnalités à mettre en oeuvre notamment pour assister le pré-traitement. Un « bon » paramétrage de la fouille n'est pas une tâche facile et nous pensons poursuivre notre travail sur l'interactivité en améliorant Transmute dans ce sens. Nous souhaitons également permettre à l'utilisateur de choisir les indicateurs qu'ils souhaite pour le tri des résultats de la fouille, et lui proposer aussi bien des indicateurs dépendants que indépendants du domaine.

FIG. 3 – *L'interface principale de* TRANSMUTE

Références

Champin, P.-A., A. Mille, et Y. Prié (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica* (59), 171–204.

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

Tatti, N. et B. Cule (2010). Mining closed strict episodes. In *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pp. 501–510.

Summary

Analysis of traces is an increasingly active issue in knowledge discovery (KD). Unfortunately, to date, few tools are designed to help interactively the analyst during the knowledge discovery process. In this paper, we present TRANSMUTE, a tool to assist the analyst throughout an interactive KD process in traces. TRANSMUTE displays the results of the mining together with the source trace so that it is easier for the analyst to understand them and thus to refine the parameters of the algorithm accordingly. TRANSMUTE is developed as a web platform and relies on a Python implementation of the KD process.