Cas d'étude : TAMAGO

Scénario : S4 - S4V1 et S4V2
Auteurs de l'analyse : Béatrice Fuchs, avec l'assistance de Caroline Emin pour la recherche de patterns
Période de l'analyse : Juillet 2015, sur des données de mars-avril 2015

La problématique posée pour l'analyse :

Tester l'utilisabilité du prototype Transmute dans la recherche de motifs séquentiels récurrents dans le cadre de l'étude de Tamagocours. Utilisation d'algorithmes de découverte de connaissances non supervisée. Quelques fonctionnalités de découverte supervisée.

Description du stockage des données:

Les données ont été collectées à partir d'un fichier CSV comprenant la totalité des traces pour la période, tous groupes et utilisateurs confondus.

Plateformes/outils utilisés :

Transmute et l'ensemble des modules qui constituent son architecture :

- DisKit : pré et post-traitement (processus d'extraction de connaissances à partir de traces)
- DisKit inclut un module de fouille de séquences DMT4SP (Data Mining Techniques For Sequence Processing) développé par C. Rigotti du laboratoire LIRIS à Lyon
- KTBS : noyau de stockage et de gestion de base de traces
- Samotraces : module d'affichage de traces et de gestion des accès au KTBS

Points forts	**Points faibles**
Transmute : Facilité d'utilisation dans une application et adaptabilité (outil générique). Possibilités d'interaction avec la trace et les motifs découverts pour juger de leur pertinence : – Visualisation des occurrences de motifs trouvés sur la trace analysée – Mémorisation des motifs intéressants et transformation de la trace pour les y faire apparaître – Tri selon plusieurs critères : fréquence du motif, couverture de la trace. Les mesures permettent très rapidement d'identifier des motifs intéressants – Elimination automatique de motifs « redondants » lors d'une sélection.	Principalement liés au fait qu'il s'agit de la première version d'un prototype : – La préparation du travail avant utilisation nécessite l'intervention d'un informaticien. – Pas de « préférences utilisateur » pour la personnalisation de l'outil – Quelques bugs résiduels. – L'interface n'est pas toujours très pratique. – Un seul algorithme de fouille est disponible pour le moment. – Problèmes de performances au delà de 2000 événements dans une trace ou motifs issus de la fouille. – Licence DMT4SP
DisKit : Permet de trouver des épisodes séquentiels (motifs séquentiels) sous contraintes (découverte non supervisée). De nombreux types de contraintes peuvent être introduits pour limiter les résultats et des recherches très fines peuvent être réalisées. Post-traitement efficace et complète les contraintes qui manquent dans la fouille. On peut générer des règles séquentielles à	Manque de fonctionnalités de pré - traitement : – transformer les traces en vue d'analyses spécifiques (par ex concaténation de types d'action avec des attributs ou généralisations de types d'actions). – Requête élaborées pour sélectionner une trace à partir de la base de traces En post-traitement :

un seul conséquent et contraindre la valeur du conséquent pour rechercher tous les motifs séquentiels qui y mènent avec une confiance minimum.	– Rajouter des contraintes sur attributs. Ces fonctionnalités devraient pouvoir être améliorées avec le KTBS. Pas de possibilité de traiter simultanément plusieurs traces (des problèmes théoriques restent à étudier). Un bug de Transmute n'a pas permis de tester les règles avec un conséquent donné.
DMT4SP : algorithme très performant qui propose de nombreuses possibilités de contraintes.	Problème de licence, pas open-source, l'auteur délivre l'exécutable de façon ponctuelle. Remarque : le programme existe en version Linux ou Unix uniquement.
Samotraces : Visualisation très réaliste et personnalisable et en fonction du domaine d'application. La petite taille des traces de Tamagocours se prêtait bien à l'analyse.	Problèmes de performances pour des traces de grande taille ou pour des résultats de fouille très grands (> 2000).

Description des pré-traitements:

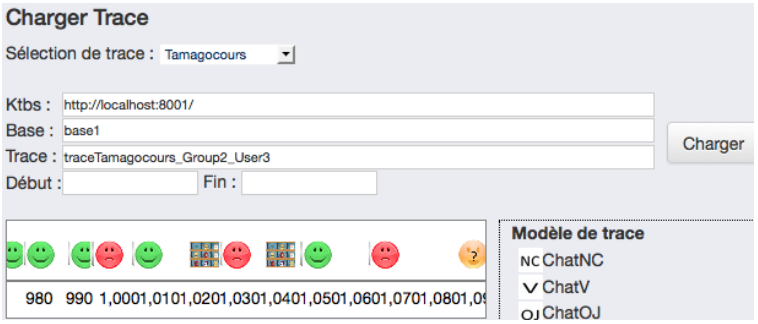
Plateformes/outils utilisés:
Un petit programme python a été écrit pour générer les traces à partir du fichier CSV dans un format reconnu par le KTBS. Quelques transformations ont été réalisées (concaténations d'attributs notamment).

Points forts	**Points faibles**
Rapidité	Spécifique au traces étudiées Alimentation manuelle du KTBS (modèle de trace + traces nombreuses)

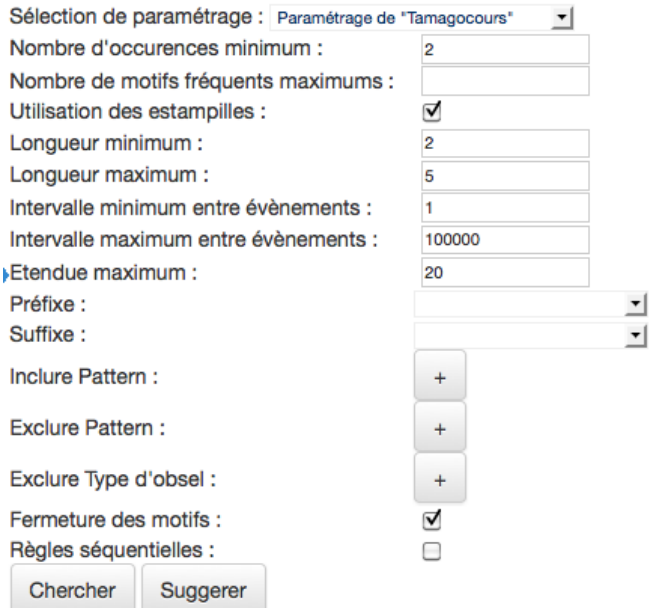
Description des analyses :

Plateforme par plateforme :

Mode opératoire méthodologique :
1. A l'aide de l'interface de Transmute :
   1.1. Choix et chargement de la trace à partir du KTBS
   1.2. Affichage de la trace et du modèle de trace associé



1.3. Choix des paramètres de fouille par l'utilisateur et lancement de la fouille.



1.4. Affichage de la liste des motifs résultant de la fouille associés à des indicateurs (fréquence, couverture : nombre d'actions couvertes par le motif dans la trace).

1.5. Etape d'interprétation interactive et itérative :

1.5.1. Observation et tri des résultats

1.5.2. Sélection d'un motif :

1.5.2.1. Filtrage automatique des motifs redondants en fonction de la sélection

1.5.2.2. Visualisation des différentes occurrences dans la trace et navigation.

1.5.3. Choix d'un nouveau « type d'observé » pour le motif choisi

1.5.4. Réécriture de la trace en remplaçant les occurrences du motif sélectionné par le nouveau type d'observé créé pour le représenter.

Mode opératoire technique, logiciels utilisés :

Préparation du travail en amont :

2. KTBS :
   - 2.1. Modélisation de la trace et création du modèle de trace dans KTBS
   - 2.2. Découpage de la trace par groupe + utilisateur : 243 traces de 10 à 390 actions (en moyenne 102 actions par trace).
   - 2.3. Alimentation de la base de traces du KTBS avec les traces obtenues. En pratique, seules quelques traces analysées ont été stockées dans le KTBS par manque de temps pour élaborer une solution automatique.
3. Paramétrage de Transmute :
   - 3.1. sélection d'un jeu d'icônes réalistes pour chaque type d'action répertorié.

Lors de l'analyse :

4. Par le module DisKit :
   - 4.1. Pré-traitement : lecture de la trace à partir du KTBS, génération d'une séquence pour le miner, éventuellement filtrage de certaines actions.
   - 4.2. Lancement du miner et récupération des résultats au format TSV.
   - 4.3. Post-traitement : mise en forme des résultats et prise en compte de contraintes additionnelles (inclusion ou exclusion de motifs, fermeture des motifs,

Scripts produits pour l'analyse des données

néant

Résultats obtenus:

Les premières analyses ont été réalisées avec une dizaine de traces et on a facilement mis en évidence les motifs récurrents identifiés lors des analyses précédentes (SAF-AF).
Nous n'avons pas pu mener à terme une analyse exhaustive pour rechercher de nouveaux motifs faute de temps. Nous avons néanmoins observé des motifs récurrents caractéristiques du profil de certains groupes d'utilisateurs de Tamagocours.

Points forts des analyses	Points faibles des analyses
– Interprétation facilitée par l'utilisation d'icones réalistes et par l'utilisation de mesures d'intérêt caractéristiques (support, couverture) – Visualisation des motifs dans la trace – Filtrage des motifs en fonction du choix de l'utilisateur	– Des bugs résiduels (affichage, indicateurs) n'ont pas permis de mettre en œuvre les transformations issues de la sélection d'un motif.

Description des itérations:

Actuellement le scénario S4V1 a été concluant quoique non appliqué à la totalité des traces. Le scénario S4V2 doit être poursuivi pour affiner et découvrir de nouvelles stratégies mises en œuvre dans Tamagocours.

Décrire la production des nouvelles données :
- Processus automatique de création de traces à partir d'un ou plusieurs fichiers pour alimenter le KTBS de façon entièrement automatique.

Décrire le changement dans les méthodes d'analyse :
- Pas de changement à court terme (travail important d'ingénierie nécessaire).
- À moyen/long terme, les fonctionnalités doivent être complétées et améliorées.
- Expérimenter avec d'autres algorithmes de mining.

Cette première analyse a mis en lumière l'ensemble des fonctionnalités nécessaires et des travaux qu'il faut continuer à explorer :
- Traitement de plusieurs séquences simultanément : DMT4SP permet de traiter plusieurs traces. Néanmoins ceci a une incidence sur DisKit en particulier pour déterminer la fermeture des motifs. Des aspects théoriques doivent être abordés pour cela.
- Visualisation des motifs dans le cas de plusieurs séquences : Transmute n'a pas été conçu pour la découverte à partir de plusieurs traces. La visualisation graphique peut montrer des limites lorsqu'il y a un grand nombre de traces (surcharge). Il est en revanche possible de générer un fichier CSV que l'utilisateur peut ensuite manipuler à souhait. Néanmoins, les possibilités d'interaction et les fonctions de Transmute ne seront alors plus disponibles. À étudier.
- Pré-traitement : sélection des données (trace + requête sur trace) et transformation des traces (combinaison d'attributs notamment) pour une plus grande variété d'analyses.
- Assistance au paramétrage du miner.
- Post-traitement : contraintes sur attributs, filtrage des occurrences de motifs en recouvrement.
- Itérations sur des cycles d'analyse en utilisant la trace transformée à l'issue d'analyses précédentes.

Points forts des itérations	Points faibles des itérations