Interprétation Interactive de connaissances à partir de traces

Amélie Cordier[1], Béatrice Fuchs[2].

[1] Université de Lyon, Claude Bernard, LIRIS, F-69 100, Villeurbanne, France, amelie.cordier@liris.cnrs.fr

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

Résumé : Dans le processus d'extraction de connaissances à partir de données (ECD), l'analyste est au centre des opérations car c'est lui qui possède les connaissances pour interpréter les résultats. L'interactivité est alors déterminante lors de l'interprétation des motifs issus de la fouille pour choisir ceux qui deviendront des connaissances. Nous proposons une démarche d'interprétation interactive lors du processus d'extraction de connaissances à partir de données (ECD), dans le cadre de la recherche d'épisodes séquentiels à partir de traces. Elle s'appuie sur une visualisation des épisodes séquentiels obtenus par la fouille, sur lesquels l'analyste peut interagir. Il peut trier les résultats à l'aide de mesures de qualité, visualiser les occurrences de motifs sur la trace, et et un mécanisme de révision automatique permet de filtrer les motifs au voisinage d'un motif choisi. Des expérimentations montrent l'intérêt de cette approche qui est appliquée au domaine de l'analyse mélodique.

Mots-clés : extraction de connaissances, Visualisation, interactions, traces

1 Introduction

L'extraction de connaissances à partir de données (ECD) vise à découvrir des connaissances nouvelles dans de gros volumes de données à l'aide de méthodes non triviales, dans un processus *interactif* et *itératif* (Frawley *et al.*, 1992). La nature itérative du processus est due à la complexité du phénomène étudié dont chaque itération améliore graduellement la compréhension. L'interactivité quant à elle est due au travail d'un analyste, expert du domaine, qui joue un rôle central dans le processus d'ECD : il dirige le travail d'analyse en guidant les différentes étapes et en décidant quelles sont les connaissances pertinentes, celles qui font sens dans le domaine étudié. Sa présence est d'autant plus importante que les connaissances du domaine ne sont généralement pas disponibles dans le système (Mathern, 2012).

Les travaux dans le domaine de l'ECD se sont longtemps focalisés sur l'étape de fouille car elle est au centre du processus et pose des problématiques de calcul complexes. En effet, la fouille assure le traitement automatique de gros volumes de données pour y mettre en évidence des régularités. Mais pour passer des régularités aux connaissances, une expertise humaine est indispensable et le travail se heurte à plusieurs difficultés. D'une part le paramétrage de la fouille est loin d'être aisé et il faut en général s'y prendre à plusieurs fois pour trouver un paramétrage « convenable ». D'autre part l'étape d'interprétation est très délicate car des milliers de résultats doivent être traités manuellement. Cette dernière étape est pourtant déterminante, la forme des résultats doit être adéquate pour qu'ils soient compréhensibles, et leur présentation doit faciliter le travail de l'analyste afin de lui permettre de mobiliser ses connaissances. L'expérience a montré que l'utilisation de mesures d'intérêt indépendantes du domaines, des mesures d'intérêt *objectives*, telles que le support ou la longueur (Béchet *et al.*, 2014) sont insuffisantes lorsqu'il y a beaucoup de motifs et de redondance combinatoire (van Leeuwen, 2014). Par ailleurs, il y

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

a un réel besoin de suivre l'avancement du travail de l'analyste, de mémoriser le travail réalisé à chaque étape et de le capitaliser dans une perspective de réutilisation lors de sessions de découverte ultérieures liées à l'itérativité du processus.

On a donc rapidement pris conscience de l'attention à porter à toutes les étapes du processus d'ECD et plus particulièrement à leur caractère interactif qui est déterminant pour mener à bien le processus complet (Holzinger, 2013). Pour susciter la mobilisation cognitive de l'analyste, des outils sont nécessaires pour lui faciliter l'interprétation des résultats. Ces constats expliquent l'intérêt des travaux sur l'interactivité et la visualisation pour assister toutes les étapes du processus d'ECD (Kuntz *et al.*, 2006).

Nous proposons dans cet article une approche visant à assister le travail de l'analyste lors de l'interprétation en lui facilitant le suivi de l'avancement de son travail par une démarche itérative et interactive. A chaque itération, l'analyste peut visualiser et interagir sur les résultats de la fouille, et ses actions sont prises en compte pour gérer l'avancement de son travail et ainsi lui permettre de se focaliser plus rapidement sur des motifs d'intérêt. Dans la suite de l'article, nous situons ce travail dans le domaine de recherche, puis nous présentons le processus d'ECD et la démarche d'interprétation proposée associée aux définitions sous-jacentes, illustrées dans le domaine de l'analyse mélodique. Les premières expérimentations sont présentées ensuite afin étudier l'efficacité du processus, suivies d'une discussion. Pour finir, nous concluons sur l'état actuel du développement et ses perspectives.

2 Visualisation et interactions dans le processus d'ECD

Les représentation visuelles d'informations sont utilisées depuis longtemps en statistiques et en analyse de données et une communauté de chercheurs s'est constituée dans le domaine de la visualisation d'information. Avec l'émergence de la fouille de données et des algorithmes performants pour trouver des régularités dans de grandes quantités de données, on a rapidement pris conscience de l'enjeu à tirer parti des travaux sur la visualisation pour intégrer l'humain dans le processus de découverte de connaissances (Shneiderman, 2002), ce qui a abouti à la fouille visuelle des données qui s'est développée ces dernières années (Bertini & Lalanne, 2009). Plus généralement l'analyse visuelle (Keim *et al.*, 2010) 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 (van Leeuwen, 2014). Ceci a donné lieu à plusieurs travaux dans ce sens avec une approche visuelle (Bothorel, 2014) et/ou interactive (Blanchard *et al.*, 2007a) pour les règles d'association.

Un des premiers problèmes de la fouille est la surabondance des résultats qui rend difficile leur exploration visuelle. Les travaux qui se sont intéressés à ce problème ont d'abord étudié des mesures d'intérêt objectives afin de caractériser la qualité des résultats de la fouille (Guillet & Hamilton, 2007), principalement les règles d'association. 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 subjectives (Marinica *et al.*, 2008; Brisson & Collard, 2008). La plupart des approches se sont intéressées aux règles d'association, ou aux règles temporelles (Blanchard *et al.*, 2007b, 2008), 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 dans ces systèmes visent davantage à

changer de point de vue sur l'affichage les résultats, mais l'assistance à la construction d'un modèle est encore peu abordée dans la littérature (van Leeuwen, 2014).

Nous proposons un cadre pour la découverte d'épisodes séquentiels qui intègre un ensemble d'outils utiles pour l'analyse :

- un algorithme de fouille de séquences qui recherche des régularités dans des traces,
- une interface permettant des interactions avec l'analyste afin de visualiser une trace et des occurrences de motifs dans la trace,
- des mesures indépendantes du domaine pour caractériser la redondance combinatoire afin d'aider à identifier les motifs les plus prometteurs,
- un processus de révision qui consiste à filtrer les motifs au fur et à mesure des sélections de motifs afin d'assister le travail de l'analyste,
- un système à base de traces afin de mémoriser le résultat des analyses et capitaliser ainsi le travail réalisé pour des sessions de découverte ultérieures.

3 Extraction de connaissances à partir de traces

Les traces sont étudiées dans le cadre d'un processus classique d'ECD mis en œuvre dans un cycle composé des étapes principales de pré-traitement (sélection de trace, transformation), fouille, puis post-traitement (visualisation, interprétation). Bien que nous nous plaçons dans le cadre de l'étude des *traces*, les concepts présentés ici puissent s'appliquer à des données temporellement situées quelconques. Nous utilisons le domaine de l'analyse mélodique comme application « jouet » pour évaluer nos propositions. Il s'agit d'analyser une partition musicale décrite par une séquence de notes associées à une durée pour y détecter des motifs mélodiques récurrents.

3.1 Traces et système à base de traces

Notre approche s'appuie sur un système dédié à la gestion de traces qui permet d'une part de collecter des traces et les stocker, mais également à les manipuler à l'aide d'opérations génériques. 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. Dans le domaine de l'analyse mélodique par exemple, les obsels décrivent les notes d'une partition musicale et sont caractérisés par un nom et une durée. 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 celles-ci, 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 non nécessairement contigüs de t_1 par de nouveaux types d'obsels résumant chaque motif. Par exemple, la transformation $\bigstar$ est définie de la façon suivante : $\bigstar : \square\blacktriangle\triangledown\diamondsuit \longrightarrow \square\bigstar\triangledown$ avec $\bigstar = \blacktriangle\diamondsuit$. Un système à base de traces modélisées est un système permettant de collecter, de traiter et de visualiser des traces.

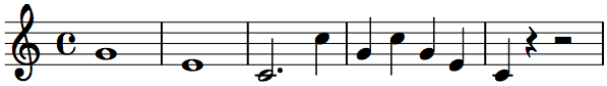
Le framework *kTBS* (*kernel for Trace Based System* [1]), (Champin *et al.*, 2013) réifie cette notion de système à base de traces. La réécriture de traces se situe au cœur du dispositif interactif mis en œuvre dans notre démarche d'interprétation.

3.2 Pré-traitement, fouille

Dans l'étape de pré-traitement, une trace est choisie par l'analyste dans une base de traces pour construire une séquence.

Définition 1 (Séquence)
Une séquence S est un ensemble d'événements typés et datés. Une occurrence d'événement est un couple (e_i, t_i) avec $e_i \in E$, où e_i est un type d'événement, E est l'ensemble des types d'événements et $t_i \in \mathbb{N}$ est l'estampille associée à e_i.

La séquence est construite à partir des obsels de la trace sélectionnée qui comportent une date de début et de fin et sont associés à un type d'obsel. Dans le cas de la partition musicale, les obsels sont des notes associées à une durée qui permettent de calculer leurs date de début et de fin, mais dans le cas général d'une trace, la durée n'est pas toujours disponible. Soit l'exemple de partition musicale suivant :



Les notes de cette partition peuvent être décrites par la trace suivante [2] :

Types d'obsels	G	E	C	C	G	C	G	E	C
durées	4	4	3	1	1	1	1	1	1
date de début	0	4	8	11	12	13	14	15	16

À partir de cette trace, la séquence suivante est construite avec $E = \{C, E, G\}$:
$S = \{(G, 0), (E, 4), (C, 8), (C, 11), (G, 12), (C, 13), (G, 14), (E, 15), (C, 16)\}$

À l'étape suivante, l'analyste fournit les paramètres de fouille qui servent à contraindre la fouille. L'étape de fouille utilise DMT4SP [3] (Nanni & Rigotti, 2007), un prototype d'extraction d'épisodes ou de règles séquentiels à partir d'une ou plusieurs séquences d'événements, conformément à la sémantique d'occurrence minimale (Mannila *et al.*, 1997). DMT4SP produit un ensemble de motifs fréquents satisfaisant les contraintes spécifiées dans le paramétrage.

Définition 2 (Motif, occurrence)
Un motif $m = (e_1, e_2, \ldots, e_n), e_i \in E$ est une séquence de types d'événements de longueur $l_m = n$. Une occurrence o_m^j du motif m est un ensemble d'estampilles $\{t_i\}_{i=1,n}$ tq $(e_i, t_i)_{i=1,n} \in S$. $O_m = \{o_m^j\}_j$ est l'ensemble des occurrences du motif m dans S.

Définition 3 (Fréquence, support)
On appelle fréquence ou support d'un motif m le nombre d'occurrence de ce motif dans S. Elle est notée $\sigma(m) = |O_m|$. La fouille retourne un ensemble M de motifs fréquents tels que $M = \{m_i\}, \forall i\ \sigma(m_i) \geqslant \sigma_{min}$, où σ_{min} est le support minimum choisi par l'analyste.

1. `http://tbs-platform.org/tbs/doku.php`
2. Ici les trois valeurs utilisées – la ronde, blanche pointée et noire ont pour durées respectives 4, 3 et 1 temps.
3. `http://liris.cnrs.fr/~crigotti/dmt4sp.html`

Dans l'exemple de partition musicale précédent, si $\sigma_{min} = 2$, les occurrences du motif (G, E, C) sont $o_1 = \{0, 4, 8\}$ et $o_2 = \{14, 15, 16\}$. $\sigma((G, E, C)) = 2 \geqslant \sigma_{min}$.

D'autres contraintes peuvent être spécifiées par l'analyste afin de limiter les résultats, comme la fenêtre temporelle, définie comme l'intervalle de temps maximal séparant le premier et le dernier événement des motifs [4].

3.3 Post-traitement et Interprétation interactive

Les motifs produits par la fouille sont mis en forme lors du post-traitement à l'aide des informations présentes dans la trace afin de les rendre intelligibles pour l'interprétation, puis ils sont affichés dans l'application TRANSMUTE. L'analyste peut trier les motifs selon plusieurs critères, les sélectionner, voir leurs occurrences dans la trace et choisir ceux qu'il estime les plus pertinents. L'interprétation consiste à construire une nouvelle trace transformée à partir de la trace analysée dans laquelle l'analyste mémorise les motifs qu'il a sélectionnés en créant de nouveaux types d'obsels qui se substituent aux occurrences des motifs sélectionnés dans la trace. Lorsqu'un motif est choisi, les occurrences des autres motifs ayant au moins un obsel en commun avec les occurrences du motif sélectionné sont éliminées, leur support est recalculé, et ceux dont le support est insuffisant sont éliminés : c'est l'opération de révision. La révision a pour effet de diminuer graduellement le nombre de résultats et facilite ainsi les prochains choix de l'analyste qui peut se focaliser sur d'autres motifs. Lorsque l'analyste a sélectionné tous les motifs et leur a associé les types d'obsels les remplaçant, il peut déclencher la réécriture qui procède à la création d'une nouvelle trace transformée. Ces opérations ainsi que le prototype TRANSMUTE sont présentés ci-après.

3.3.1 Le prototype Transmute

TRANSMUTE est un outil de génération de transformation de traces à partir des interactions de l'analyste qui permet d'afficher et d'interagir avec une trace et les motifs issus de la fouille. L'architecture de TRANSMUTE repose sur le module DISKIT qui met en œuvre le cycle d'ECD et SAMOTRACES, un framework Javascript pour la visualisation et les interactions (Barazzutti *et al.*, 2016). L'interface (figure 1) comporte dans la partie supérieure la trace en cours d'analyse associée à son modèle où les occurrences des motifs sélectionnés par l'analyste sont affichés et dans la partie inférieure, à gauche les paramètres de la fouille saisis par l'analyste et à droite les motifs obtenus par la fouille. Suite à la sélection de deux motifs, on observe dans la partie inférieure les motifs filtrés par la révision qui apparaîssent estompés dans l'interface.

3.3.2 Mesures d'intérêt

La fouille produit généralement un grand nombre de motifs caractérisés par une forte redondance combinatoire : un épisode apparaît sous la forme d'un grand nombre de variantes qui ne peuvent être éliminées faute de connaissances suffisantes, car elles satisfont toutes les contraintes qui ont été spécifiées. Le travail de l'analyste qui doit examiner ces résultats est fastidieux et le choix des « bons » épisodes est une tâche difficile. Afin d'aider l'analyste, des

4. Tous ces types de contraintes ne sont pas détaillés dans cet article.

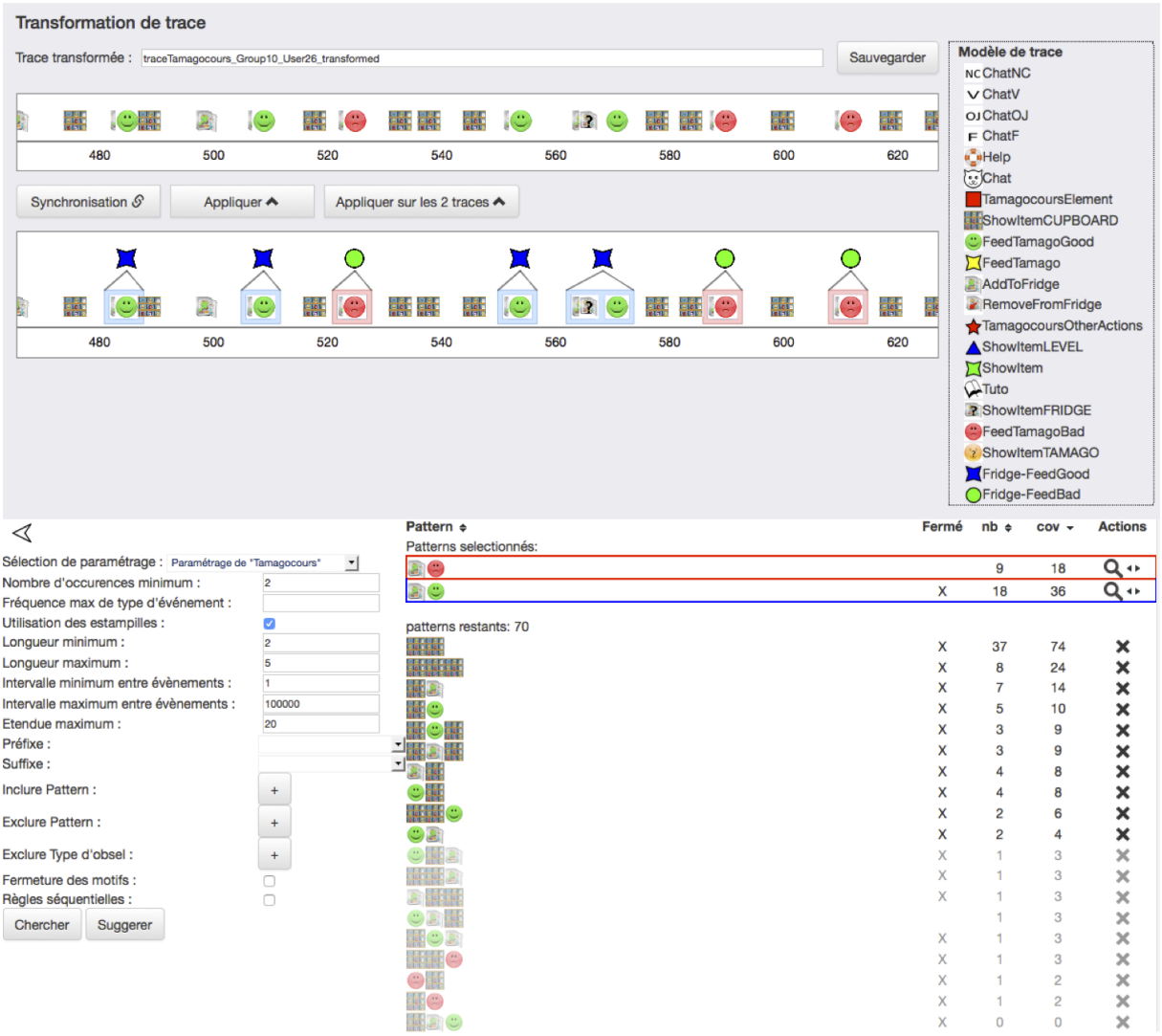


FIGURE 1 – L'interface de TRANSMUTE.

mesures sont utiles pour évaluer l'intérêt des motifs obtenus. Elles peuvent être spécifiques au domaine (subjectives), mais des mesures objectives peuvent également apporter une aide. En particulier nous nous sommes intéressés à des mesures objectives pour caractériser la redondance combinatoire des motifs.

Tout d'abord, la propriété de fermeture des motifs est importante à prendre en considération car elle fournit une représentation plus compacte des motifs et limite très fortement le nombre de motifs générés. Pour cette raison, de nombreux algorithmes de fouille la prennent en compte. Un épisode est fermé s'il n'est pas inclus dans un épisode de longueur supérieure ayant le même support [5]. Nous nous sommes limités à une seule séquence à l'aide de la fréquence telle que

5. Dans le cas des épisodes séquentiels, elle peut être définie de plusieurs façons lorsque l'on traite simultanément plusieurs séquences, car il existe plusieurs définitions du support : en terme de nombre d'occurrences d'un motif ou en terme de nombre de séquences où un motif est présent.

définie précédemment.

Définition 4 (Inclusion stricte de motifs)
Un motif $m_1 = (e_1, e_2, \dots e_p)$ *est inclus dans le motif* $m_2 = (e'_1, e'_2, \dots e'_n)$ *noté* $m_1 \prec m_2$ *si* $m_1 \neq m_2$ *et* $\exists \{j_i \in \mathbb{N}\}_{i=1\dots p}$ *tels que* $1 \leqslant j_i < j_{i+1} \leqslant n$ *et* $\forall i\ e_i = e'_{j_i}$

Définition 5 (Motif fermé)
Un motif m *est fermé si* $\forall\ m' \in M$ *tel que* $m \prec m'$ *alors* $\sigma(m) > \sigma(m')$

Dans l'exemple de la séquence
$S = \{(G,0),(E,4),(C,8),(C,11),(G,12),(C,13),(G,14),(E,15),(C,16)\}$,
le motif $m_1 = (G,E) \prec m_2 = (G,E,C)$ avec $\sigma(m_1) = \sigma(m_2)$, donc m_1 n'est pas fermé. En revanche, le motif $m_3 = (G,C) \prec m_2$ avec $\sigma((G,C)) = 3$, m_3 est fermé car il n'est inclus dans aucun motif ayant le même support.

En complement à la fermeture des motifs, nous avons défini les mesures ci-dessous pour évaluer la redondance combinatoire.

Définition 6 (Couverture événementielle)
La couverture événementielle CE_m *est l'ensemble des estampilles distinctes des occurrences d'un motif* m. $CE_m = \bigcup_{i=1}^{\sigma(m)} o_m^i$. *L'indicateur de couverture événementielle* IC_m *d'un motif est le nombre d'estampilles distinctes de ses occurrences :* $IC_m = |CE_m|$.

Définition 7 (Étalement événementiel)
L'étalement EE_m *d'un motif* m *de longueur* n *est le nombre d'événements de* S *dont l'estampille est incluse dans les intervalles des occurrences de* m. $EE_m = |\{t_k, \forall o_m^i \exists I_m^i, t_1^i \leqslant t_k \leqslant t_2^i\}|$

Définition 8 (Bruit)
Le bruit B_m *d'un motif est le nombre d'événements ne faisant partie des occurrences d'un motif donné et insérés dans l'intervalle temporel d'un motif :* $B_m = EE_m - CE_m$

Dans certains cas, il est possible de calculer des indicateurs temporels du même type que les indicateurs événementiels précédents mais qui portent sur la durée des événements, lorsque la durée des événements est disponible dans la trace. C'est le cas dans le domaine de l'analyse mélodique où les notes sont associées à une durée.

Définition 9 (Couverture temporelle)
La couverture temporelle CT_m *d'un motif* $m \in M$ *est la durée totale des événements de la couverture événementielle. On note* $d((e,t))$ *la durée associée à l'événement* (e,t). *Soit* m *un motif de couverture événementielle* $CE_m, \forall (e_i,t_i) \in CE_m, CT_m = \Sigma_i d((e_i,t_i))$

Dans l'exemple précédent, considérons les motifs $m_1 = (G,C)$ et $m_2 = (G,E,C)$.
$CE_{m_1} = \{0,8,12,13,14,16\}$, $IC_{m_1} = 6$, $EE_{m_1} = 8$, $B_{m_1} = 8-6 = 2$ et $CT_{m_1} = 11$.
$CE_{m_2} = \{0,4,8,14,15,16\}, IC_{m_2} = 6, EE_{m_2} = 6$, $B_{m_2} = 6-6 = 0$ et $CT_{m_2} = 14$.

Ces définitions permettent de décrire les principes de la révision interactive.

3.3.3 Révision interactive

Au fur et à mesure que des motifs sont choisis, il devient inutile de continuer à présenter tous les motifs qui sont, parfois à peu de choses près, des variantes des motifs choisis. Dans l'exemple de la partition précédente, si le motif (G, E, C) est choisi, il est inutile de considérer d'autres motifs tels que (G, E) ou (E, C) par exemple, car leur occurrences comportent des événements présents dans les occurrences de (G, E, C). Le choix d'un motif donné rend caduque l'examen de toutes ses variantes que l'analyste ne devrait pas avoir à considérer par la suite. Dans ce but, le principe de la révision repose sur le filtrage de tous les motifs redondants avec le motif choisi, afin de favoriser la focalisation sur d'autres motifs - ou régions de l'espace de recherche non encore explorés. L'étape d'interprétation est itérative et à chaque itération, le choix d'un motif est suivi d'une révision qui a pour conséquence un filtrage des motifs restant à considérer. La révision joue un rôle important dans l'assistance fournie à l'analyste en phase d'interprétation car elle agit comme un filtrage de l'espace de recherche autour d'un motif choisi par l'analyste. Elle se base sur la couverture événementielle pour rechercher les motifs à supprimer.

Définition 10 (Révision)
Soit m_c un motif choisi par l'expert de couverture événementielle CE_c.
Soit $m_i \in M, m_i \neq m_c$ un motif, l'ensemble des occurrences de m_i invalidées par le choix de m_c est : $O(m_i|m_c) = \{\forall o_i \in O(m_i), o_i \cap CE_c \neq \emptyset\}$.
On note $M_{m_c} \subset M$ l'ensemble des motifs de M invalidés par le choix de m_c par l'analyste : $M_{m_c} = \{m_i \in M, m_i \neq m_c,$ tq $\sigma(O(m_i|m_c)) < \sigma_{min}\}$
Lorsque le motif m_c a été choisi, l'ensemble des motifs restants à examiner par l'expert à l'itération suivante, est : $M \setminus M_{m_c}$.

Dans l'exemple, soient les motifs $m_1 = (G, E, C)$, $m_2 = (G, E)$ et $m_3 = (G, C)$.
$CE_{m_1} = \{0, 4, 8, 13, 15, 16\}$, les occurrences de m_2 sont $o^1_{m_2} = \{0, 4\}$ et $o^2_{m_2} = \{14, 15\}$, et pour m_3, $o^1_{m_3} = \{0, 8\}$,$o^2_{m_3} = \{12, 13\}$,$o^3_{m_3} = \{14, 16\}$. Si l'expert choisit le motif m_1, $o^1_{m_2}$ et $o^2_{m_2}$ sont supprimées ainsi que $o^1_{m_3}$ et $o^3_{m_3}$. Les supports de m_2 et m_3 deviennent respectivement 0 et $1 < \sigma_{min}$. La révision suite au choix de m_1 a donc pour effet d'éliminer m_2 et m_3.

La révision commence initialement avec l'ensemble M de tous les motifs issus de la fouille. A chaque itération, l'ensemble M est progressivement épuré de la redondance combinatoire autour d'un motif choisi en éliminant les motifs voisins dans l'espace de recherche.

Les mesures objectives ainsi que la révision interactive sont évalués dans le paragraphe suivant à l'aide d'une expérimentation dans le domaine de l'analyse musicale.

4 Expérimentations

Dans le domaine de l'analyse mélodique, trois partitions ont été étudiées pour lesquelles les motifs intéressants à retrouver dans les résultats de la fouille sont fournis par l'expert. Nous les appellerons dans la suite les *motifs experts*. Nous proposons de mesurer l'efficacité du processus par l'effort requis par l'analyste pour trouver les motifs intéressants en terme de nombre de motifs à examiner pour trouver tous les motifs experts, ce qui correspond au rang du dernier motif expert trouvé dans les résultats de fouille. Ainsi, plus ce rang est faible, plus l'effort requis par l'analyste est faible et plus la stratégie mise en place est efficace. L'effort de l'expert

est mesuré en triant les motifs selon plusieurs critères : l'ordre de sortie de la fouille (sans tri), la fréquence, l'indice de couverture événementielle, la couverture temporelle. Dans un deuxième temps le bruit est utilisé comme premier critère de tri et les quatre critères précédents comme deuxième critère. Ensuite, le rang le plus élevé des motifs experts est observé pour chaque critère de tri et indique l'effort requis pour trouver tous les motifs experts. Nous reportons les résultats dans deux tables ci-dessous : dans la première table, l'effort est d'abord mesuré sans révision, et dans la deuxième table, il est mesuré avec révision à chaque fois qu'un motif expert est trouvé. Intuitivement, la suppression de motifs à chaque fois qu'un motif expert est sélectionné devrait mener à une diminution du nombre de motifs et aura un effet sur les rangs des motifs experts restants qui devraient nécessairement diminuer. Néanmoins, des motifs experts sont susceptibles d'être éliminés à la suite de la révision, et il convient alors d'observer leur taux de rappel. Le paramétrage utilisé pour la fouille a été choisi de façon à assurer la présence de tous les motifs experts ($\sigma_{min} = 2$).

Sans révision, le rang du dernier motif expert est résumé dans le tableau 1 pour les trois pièces, selon le critère de tri utilisé ($\nearrow$ indique un tri croissant et $\searrow$ un tri décroissant).

pièce	motifs fouille	motifs experts	sans tri	$\sigma \searrow$	$IC \searrow$	$CT \searrow$	Bruit $\nearrow$ puis $\sigma \searrow$	Bruit $\nearrow$ puis $IC \searrow$	Bruit $\nearrow$ puis $CT \searrow$
1	3 853	11	3 838	3 838	3 797	2 735	240	369	**233**
2	12 947	20	12 818	12 829	12 591	**9 516**	12 667	12 672	12 668
3	59 786	29	53 805	56 061	31 309	30 151	**22 364**	25 317	25 420

TABLE 1 – Rang du dernier motif expert sans révision.

Sans révision, les meilleurs résultats sont obtenus avec un tri par bruit croissant et couverture temporelle décroissante pour la pièce 1, par couverture temporelle décroissante pour la pièce 2 et par bruit croissant puis fréquence décroissante pour la pièce 3, et un taux de rappel des motifs experts de 100% car aucun motif n'a été supprimé. La diminution de l'effort de l'expert est respectivement de 94%, 26% et 58% par rapport à un traitement sans tri. Dans tous les cas, l'utilisation des mesures a permis une diminution sensible de l'effort de l'expert.

Lorsque la révision est introduite, le rang du dernier motif expert trouvé est résumé dans le tableau 2. Les meilleurs résultats sont obtenus avec un tri par bruit croissant pour la pièce 1 et par couverture temporelle décroissante pour les pièces 2 et 3.

pièce	motifs fouille	motifs experts	sans tri	$\sigma \searrow$	$IC \searrow$	$CT \searrow$	Bruit $\nearrow$ puis $\sigma \searrow$	Bruit $\nearrow$ puis $IC \searrow$	Bruit $\nearrow$ puis $CT \searrow$
1	3 853	11	502	52	49	24	13	13	**12**
2	12 947	20	801	222	95	**71**	204	204	204
3	59 786	29	13490	5103	537	**527**	1533	1533	1533

TABLE 2 – Rang du dernier motif expert avec révision.

La diminution de l'effort est significative par rapport à un traitement sans tri et sans révision (tableau 3), et montre l'efficacité la révision conjointement aux mesures. Cependant, le rappel

des motifs experts est de 82% pour la pièce 1 et 100% pour la pièce 3 et pour la pièce 2, il est de 33% lorsque le tri n'utilise pas le bruit et de 62% lorsque le bruit est introduit. L'introduction du bruit a favorisé les motifs constitués de notes plus proches et amélioré le rappel, qui reste néanmoins encore décevant.

					Bruit ↗ puis		
pièce		$\sigma \searrow$	$IC \searrow$	$CT \searrow$	$\sigma \searrow$	$IC \searrow$	$CT \searrow$
1	Diminution	90%	90%	95%	97%	97%	98%
	Rappel	82%			82%		
2	Diminution	72%	88%	91%	75%		
	Rappel	33%			62%		
3	Diminution	62%	96%	96%	89%		
	Rappel	100%			100%		

TABLE 3 – Synthèse de la diminution de l'effort par rapport à un traitement des motifs de la fouille sans révision et sans tri et du taux de rappel des motifs experts.

L'ordre dans lequel les motifs sont choisis a une incidence sur les motifs éliminés par la révision, d'où l'importance du choix des critères de tri. Ces résultats montrent qu'il est important de disposer de plusieurs mesures pour tenir compte des caractéristiques de chaque pièce qu'il convient par conséquent d'observer avant l'analyse pour prendre en compte leurs particularités. Des mesures subjectives telle que celle présentée dans (Fuchs, 2011), n'ont pas été introduites dans ce travail et sont indispensables pour compléter ces mesures objectives. On peut également mentionner que TRANSMUTE permet d'annuler une sélection de motif, et que cette souplesse d'utilisation n'a pas été prise en compte dans cette expérimentation.

5 Discussion

Le prototype Transmute qui met en oeuvre cette approche possède des limitations principalement liées à l'interface, car il n'a pour l'instant bénéficié d'aucune optimisation. Il est utilisable sur de petites traces, et ne permet pas de traiter un nombre de motifs trop important (quelques milliers d'obsels et de motifs), et il reste du travail à réaliser pour lever ce verrou. Le module DISKIT en revanche peut traiter des traces et un nombre de motifs beaucoup plus importants. Une validation qualitative a montré que TRANSMUTE a pu être pris en main aisément par les utilisateurs (Barazzutti, 2015). TRANSMUTE s'est également avéré utilisable pour l'analyse de traces d'un jeu sérieux collaboratif pour l'apprentissage de règles de diffusion de ressources numériques (TAMAGOCOURS) montré à la figure 1. Une base de trace a été construite à partir des sessions de jeu de 244 étudiants répartis en 86 groupes représentant au total environ 26 000 obsels. Les groupes et utilisateurs ont été analysés séparément, représentant des traces de quelques centaines d'obsels.

Actuellement, les mesures permettant de trier les motifs sont prédéfinies dans TRANSMUTE. Il serait souhaitable que l'analyste puisse choisir lui-même des mesures. Des mesures subjectives peuvent être conçues à l'aide du kTBS, car les motifs issus de la fouille sont mis en relation avec les obsels de la trace analysée. Leurs attributs et relations rendent possible la réalisation de

calculs plus complexes qui ne peuvent bien-entendu être implémentés que par un informaticien et mis à disposition d'un analyste.

6 Conclusion et perspectives

Nous avons présenté une démarche d'interprétation itérative et interactive dans un processus d'ECD à partir de traces. Elle s'appuie sur l'utilisation de mesures d'intérêt pour trier les motifs, une visualisation des motifs issus de la fouille où l'analyste peut interagir pour voir l'impact de ses actions sur la trace. Un filtrage dynamique au voisinage des motifs sélectionnés favorise une meilleure focalisation sur de nouvelles régions de l'espace de recherche. La création d'une trace transformée et la réécriture permettent de mémoriser le travail de l'analyste afin de le prendre en compte lors de sessions de travail ultérieures. Les premières expérimentations réalisées sont encourageantes. La plateforme TRANSMUTE réifie cette approche.

Outre l'approfondissement des mesures pour l'interprétation, une piste d'amélioration consiste à explorer dans leur ensemble les motifs pour sélectionner ceux qui présentent ensemble une meilleure couverture globale de la séquence initiale (Vreeken *et al.*, 2010), ou encore de traiter les motifs par groupes selon une mesure de similarité. D'autres perspectives concernent l'assistance aux autres phases du processus d'ECD et en particulier le pré-traitement. Tout d'abord un « bon » paramétrage de la fouille n'est pas une tâche facile et nous pensons poursuivre le travail sur l'interactivité pour aider l'analyste dans la phase de paramétrage. Pour cela, nous considérons plusieurs pistes : (1) un processus d'ECD entièrement interactif où les interactions avec l'analyste et les résultats de la fouille sont utilisés pour guider le réglage des paramètres, et (2) la recommandation de paramétrages à partir d'expériences antérieures ou à partir des interactions de l'utilisateur sur la trace elle-même. Plus généralement, la préparation des données peut être enrichie avec toutes sortes de requêtes sur les traces afin de varier les dimensions à analyser. De plus, la prise en compte de plusieurs dimensions permettrait une analyse plus fine et précise dans le domaine de l'analyse musicale. Enfin l'analyse simultanée de plusieurs traces et la façon aborder la représentation des motifs sur un grand nombre de traces est un sujet important dès lors qu'il s'agit d'étudier par exemple les traces d'utilisateurs différents ou des pièces musicales à plusieurs voix.

Références

BARAZZUTTI P.-L. (2015). *Transmute : un outil interactif d'assistance à la découverte de connaissances*. Mémoire de master en informatique, Université Claude Bernard Lyon 1.

BARAZZUTTI P.-L., CORDIER A. & FUCHS B. (2016). Transmute : un outil interactif pour assister l'extraction de connaissances à partir de traces. In B. CRÉMILLEUX & C. DE RUNZ, Eds., *Extraction et Gestion des Connaissances - EGC 2016*, volume RNTI-E-30 of *Extraction et Gestion des Connaissances*, p. 463–468, Reims, France : Cyril de Runz RNTI.

BÉCHET N., CELLIER P., CHARNOIS T. & CRÉMILLEUX B. (2014). Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. *Revue d'Intelligence Artificielle*, **28**(2-3), 245–270.

BERTINI E. & LALANNE D. (2009). 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*, p. 12–20 : ACM.

BLANCHARD J., GUILLET F. & BRIAND H. (2007a). Interactive visual exploration of association rules with rule-focusing methodology. *Knowledge and Information Systems*, **13**(1), 43–75.

BLANCHARD J., GUILLET F. & GRAS R. (2007b). On the discovery of significant temporal rules. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*, p. 443–450 : IEEE.

BLANCHARD J., GUILLET F. & GRAS R. (2008). Assessing the interestingness of temporal rules with sequential implication intensity. In *Statistical Implicative Analysis*, p. 55–71. Springer.

BOTHOREL G. (2014). *Algorithmes automatiques pour la fouille visuelle de données et la visualisation de règles d'association : application aux données aéronautiques*. PhD thesis.

BRISSON L. & COLLARD M. (2008). How to semantically enhance a data mining process ?. In *ICEIS*, volume 19, p. 103–116 : Springer.

CHAMPIN P.-A., MILLE A. & PRIÉ Y. (2013). Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. *Intellectica*, (59), 171–204.

FRAWLEY W. J., PIATETSKY-SHAPIRO G. & MATHEUS C. J. (1992). Knowledge discovery in databases : An overview. *AI Magazine*, **13**(3), 57–70.

FUCHS B. (2011). Co-construction interactive de connaissances. Application à l'analyse mélodique. In A. MILLE, Ed., *Ingénierie des connaissances*, p. 705–720.

GUILLET F. & HAMILTON H. J. (2007). *Quality measures in data mining*, volume 43. Springer.

HOLZINGER A. (2013). Human-computer interaction and knowledge discovery (hci-kdd) : What is the benefit of bringing those two fields to work together ? In *Availability, Reliability, and security in Information Systems and HCI*, p. 319–328. Springer.

KEIM D. A., KOHLHAMMER J., ELLIS G. & MANSMANN F. (2010). *Mastering the information age-solving problems with visual analytics*. Florian Mansmann.

KUNTZ P., LEHN R., GUILLET F. & PINAUD B. (2006). Découverte interactive de règles d'association via une interface visuelle. *Visualisation en Extraction des Connaissances*, p. 113–125.

MANNILA H., TOIVONEN H. & VERKAMO A. I. (1997). Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, **1**, 259–289.

MARINICA C., GUILLET F. & BRIAND H. (2008). Post-processing of discovered association rules using ontologies. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, p. 126–133 : IEEE.

MATHERN B. (2012). *Découverte interactive de connaissances à partir de traces d'activité : Synthèse d'automates pour l'analyse et la modélisation de l'activité de conduite automobile*. Thèse de doctorat en informatique, Université Claude Bernard Lyon 1.

NANNI M. & RIGOTTI C. (2007). Extracting trees of quantitative serial episodes. In S. DŽEROSKI & J. STRUYF, Eds., *Knowledge Discovery in Inductive Databases : 5th International Workshop, KDID 2006 Berlin, Germany, September 18, 2006 Revised Selected and Invited Papers*, p. 170–188, Berlin, Heidelberg : Springer Berlin Heidelberg.

SHNEIDERMAN B. (2002). Inventing discovery tools : combining information visualization with data mining. *Information visualization*, **1**(1), 5–12.

VAN LEEUWEN M. (2014). Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, p. 169–182. Springer.

VREEKEN J., LEEUWEN M. & SIEBES A. (2010). Krimp : mining itemsets that compress. *Data Mining and Knowledge Discovery*, **23**(1), 169–214.