

Assistance à la réutilisation de processus d'analyse de traces d'apprentissage via une approche narrative et sémantique

Alexis Lebis

► **To cite this version:**

Alexis Lebis. Assistance à la réutilisation de processus d'analyse de traces d'apprentissage via une approche narrative et sémantique. Septièmes Rencontres Jeunes Chercheurs en EIAH (RJC EIAH 2018) , Apr 2018, Besançon, France. Actes des 7ièmes RJC-EIAH 2018, 2018. <hal-01769541>

HAL Id: hal-01769541

<https://hal.archives-ouvertes.fr/hal-01769541>

Submitted on 18 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Assistance à la réutilisation de processus d'analyse de traces d'apprentissage *via* une approche narrative et sémantique

Alexis Lebis^{1,2}

¹Sorbonne Universités, UPMC, CNRS, LIP6 UMR 7606, F-75005 Paris, France

²Univ Lyon, Université Lyon1, CNRS, LIRIS UMR 5205, F-6922 Villeurbanne, France
alexis.lebis@lip6.fr

Résumé. Le partage et la réutilisation de processus d'analyse de traces d'apprentissage, ainsi que leur adaptation à des contextes différents, sont devenus des enjeux importants. L'objectif est double : impliquer et soutenir la communauté dans le cycle du processus d'analyse, de son élaboration à son utilisation, et lui donner les outils nécessaires pour promouvoir une co-construction de ces analyses. Dès lors, il apparaît important de fournir aux différents acteurs des outils pour les assister dans leur tâche respective, au sein de l'analyse. Dans cet article, nous présentons trois types d'assistance envisageables grâce à une représentation sémantique des processus d'analyse de traces.

Keywords: processus d'analyse de traces, narration, sémantique, ontologie, inférence, raisonnement, assistance.

1 Introduction

Dans le domaine des *Learning Analytics*, l'analyse des traces des apprenants en contexte pédagogique a pour but d'extraire des informations pertinentes en vue de comprendre et d'améliorer l'apprentissage. D'un point de vue informatique, ces analyses sont concrétisées par des processus d'analyse de traces. Ces processus d'analyse sont composés d'une succession ordonnée d'opérations, implémentées dans un outil d'analyse, appliquées sur des traces d'apprentissage [1]. Ces processus d'analyse sont soumis à des contraintes qui sont soit liées aux contextes d'apprentissage, soit aux spécificités techniques des données et des outils d'analyses. Ces contraintes rendent le partage, la réutilisation et l'adaptation des processus complexes voire peu pertinents [2].

Le constat que nous faisons est que le paradigme actuel des processus d'analyse, entièrement dédié à résoudre un besoin computationnel, ne permet pas d'inclure efficacement les différentes informations propres à l'analyse [3]. En effet, les outils d'analyse actuels ne permettent pas de tenir compte correctement du contexte d'apprentissage des processus d'analyse, des dépendances qui y sont associées et des choix effectués lors de l'analyse. De plus, l'information est rarement structurée pour une réutilisation automatique par la machine. C'est pourquoi nous utilisons un framework ontologique qui nous permet de raisonner sur les processus d'analyse afin de produire une

assistance concernant leur consultation, leur adaptation et leur réutilisation. Cette assistance est destinée aux différents acteurs de l'analyse (*i.e.* analyste, décideur...). Dans la suite de cet article, nous présentons, en section 2, un état de l'art sur les processus d'analyse de traces et sur le raisonnement sémantique. En section 3, nous décrivons brièvement notre approche narrative pour décrire les processus d'analyse de traces grâce à un framework ontologique. L'objectif de cette approche est de s'émanciper des contraintes techniques et de tenir compte des contraintes contextuelles tout en structurant l'information. La section 4 présente des possibilités d'assistance concernant la consultation, l'adaptation et la réutilisation des processus d'analyse.

2 État de l'art

Les travaux portant sur l'étude des processus d'analyse, de leur cycle de vie et de leur nomenclature ont permis de mettre en avant leur structure complexe, avec l'identification, entre autres, des phases de prétraitement, d'analyse et de post-traitement [4].

Ces processus d'analyse sont composés d'opérations manipulant les traces de manière prédéfinie [1]. De plus, ces processus (ou des parties de processus) peuvent aussi être utilisés comme opérateurs pour une autre analyse [4]. Ces processus permettent d'extraire des informations pertinentes, à des fins variées : personnalisation, diagnostic des connaissances, etc. Pouvoir exploiter des processus d'analyse déjà implémentés afin de répondre à d'autres besoins constitue un intérêt majeur pour les acteurs du domaine. Ces acteurs pourraient alors utiliser des techniques déjà éprouvées, ou bien encore s'assurer de la pertinence des résultats obtenus. Cette nécessité de réutilisation peut être observée au vu des efforts concernant le partage des processus d'analyse [1, 11]. Cependant, réutiliser un processus d'analyse est une tâche complexe [12], à la fois parce qu'il existe des dépendances techniques liées aux outils d'analyse, mais aussi des dépendances contextuelles (*e.g.* contexte pédagogique) [13] et que l'information portant sur ces processus est rare et peu, voire pas, structurée.

Dans le domaine des *Learning Analytics*, il est toutefois possible de remarquer l'utilisation ponctuelle d'éléments sémantiques pour des opérations de fouille de données (notamment le *clustering*) [6] pour tenter d'apporter de l'information supplémentaire. Mais c'est surtout dans le domaine des *workflows* que l'on peut noter les tentatives les plus fructueuses pour structurer l'information et intégrer aux processus des éléments sémantiques, comme avec wf4ever [7]. Il s'agit d'ajouter aux workflows des ressources supplémentaires pour les documenter, en accord avec une ontologie.

En prenant en compte ces travaux, nous avons proposé une approche narrative des processus d'analyse pour structurer et formaliser à la fois les processus d'analyse et l'information associée [10] (cf. section 3), *via* une ontologie. Ce réseau sémantique nous permet d'envisager un raisonnement automatisé, comme le montre Sowa [5], afin de proposer des assistances pertinentes aux différents acteurs de l'analyse dans leur tâche de réutilisation et d'adaptation d'un processus d'analyse existant. L'objectif est en effet de combiner notre approche narrative des processus d'analyse avec des travaux nous permettant de raisonner dans un tissu d'information sémantique, où l'imprécision relative de l'information existe [8,9].

3 Notre approche narrative des processus d'analyse

Afin de s'émanciper des contraintes liées aux spécificités techniques des processus d'analyse, et de prendre en compte leurs contextes, nous avons proposé un framework ontologique pour leur narration [10] permettant une description des processus d'analyse dans le but de les rendre capitalisables, *i.e.* compréhensibles, partageables et réutilisables dans des contextes plus ou moins similaires à leur contexte initial.

Cette approche narrative s'émancipe des prérequis computationnels. Pour ce faire, un processus d'analyse décrit dans notre framework est représenté à l'aide de divers éléments. Tout d'abord, les **opérateurs narrés** représentent les concepts d'opérations à appliquer, et qui sont le plus petit dénominateur commun entre les opérations ayant le même objectif qui existent dans des outils d'analyse différents. En guise d'exemple, considérons un opérateur *Filtre Temporel*. Son implémentation, ses configurations, son utilisation et son comportement sont dépendants d'un outil d'analyse, mais l'intention est identique pour chaque implémentation d'un tel filtre : filtrer des données temporelles. C'est cette intention qu'un opérateur narré représente.

Pour dépasser les limites liées aux formats de traces ainsi qu'à leur granularité, nous proposons de représenter les traces sous la forme de **graphes de variables**. Ces graphes représentent les variables contenues dans les traces, tout en faisant apparaître les relations qui existent entre elles. De cette manière, la sémantique initiale des traces est préservée, voire améliorée. Par exemple, considérons une trace issue d'un MOOC contenant les variables *étudiant*, *certification*, *nombre de cours lus* (v_1), *de vidéos vues* (v_2) et *messages postés* (v_3). Un graphe de variables G associé serait alors un nœud *étudiant* relié aux nœuds v_1 , v_2 et v_3 , respectivement par les relations *a participé*, *a regardé* et *a communiqué*. Les nœuds *certification*, *étudiant* et *MOOC* seraient respectivement liés par les relations *décerné à* et *décerné par*.

Ces opérateurs narrés et ces graphes de variables sont utilisés au sein d'étapes. Une **étape** matérialise l'application d'un opérateur sur un jeu de données particulier. Elle est donc constituée d'un opérateur narré, éventuellement configuré, et d'un graphe de variables en entrée. Bien que notre approche n'ait pas pour objectif d'effectuer des calculs avec les données, l'application d'un opérateur narré sur un graphe de variables va produire en sortie un nouveau graphe de variables. Cette production représente l'effet attendu d'un opérateur implémenté sur des données concrètes. Au final, un **processus d'analyse narré** est constitué d'étapes. Chaque étape a en entrée un graphe de variables, un opérateur narré configuré, et en sortie un graphe de variables. En guise d'exemple, prenons un processus d'analyse visant à prédire la certification d'étudiants dans un MOOC. Une étape de ce processus serait d'entraîner un modèle prédictif spécifique (disons une régression logistique) sur un ensemble E de variables. Ces *variables* sont sélectionnées dans le graphe de variables G précédent (ici, $E = \langle v_1, v_2, v_3, certification \rangle$) et mis en correspondance avec l'entrée de l'opérateur narré O qui représente la régression. En sortie, un nœud *Modèle* est créé, en fonction de E et O .

De plus, pour remédier au problème du peu d'information et de son manque de structure, notre framework prévoit plusieurs solutions. Premièrement, le recours à des **éléments narratifs** pour enrichir la **description** des processus : il s'agit de représenter des sous-ensembles précis d'information de manière explicite. Par exemple, être ca-

pable d'identifier que certaines informations d'une étape d'un processus sont une *Hypothèse* ou de décrire les connaissances produites en sortie d'un processus d'analyse et leur cadre d'utilisation. Deuxièmement, ces éléments narratifs sont mis en relation avec les éléments qu'ils décrivent (*e.g.* l'hypothèse d'une étape, le contexte d'une analyse) *via* des arcs. En résulte alors un réseau descriptif organisé. Enfin, nous utilisons un **vocabulaire contrôlé** pour désambiguïser l'information, où chaque terme peut partager des relations avec d'autres termes, l'objectif étant de pouvoir raisonner efficacement avec les processus d'analyse narrés. Par exemple, un *Étudiant de MOOC* peut être sémantiquement défini comme un *Étudiant* (lui-même défini comme un *foaf:Agent* d'après l'ontologie Friend Of A Friend) *impliqué* dans un *MOOC*. Ainsi, lors de la définition d'un graphe de variables, d'un paramétrage d'un opérateur narré ou bien encore d'une description *via* un élément narratif, il est possible d'utiliser des termes qui sont sémantiquement définis dans le vocabulaire contrôlé, plutôt que du texte difficile à interpréter.

4 Assistance à la réutilisation des processus d'analyse

Dans cette section, nous présentons trois types d'assistance qui pourraient être proposées en ayant accès à un ensemble de processus d'analyse narrés, au sein de notre framework [3,10]. Nous montrons aussi que le fait d'adopter une démarche narrative des processus offre des perspectives d'assistances destinées non seulement aux analystes, mais également à tous les acteurs du cycle de l'analyse. Enfin, nous expliquons comment l'assistance est susceptible d'évoluer *via* des retours utilisateurs.

Recherche de processus en fonction du besoin. Actuellement, la recherche d'une analyse répondant à un besoin donné est complexe. En effet, à notre connaissance, lorsque les outils d'analyse permettent une telle recherche, il s'agit généralement d'effectuer une requête en ne se basant que sur la correspondance de mots-clés fournis par l'utilisateur dans un champ de recherche. On peut aussi noter une pratique qui consiste à rechercher par opérateur. Cependant, une telle recherche est non contextualisée et ne convient pas pour répondre à un besoin particulier.

Nous proposons d'exploiter l'approche narrative pour permettre une recherche directement basée sur un besoin d'analyse. L'objectif ici est d'être capable de comprendre le besoin et de raisonner sur l'ensemble des processus narrés, pour identifier lesquels peuvent être pertinents. Cela permettrait d'assister intelligemment les différents acteurs de l'analyse dans l'exploration de processus pouvant potentiellement résoudre ledit besoin. Cela permettrait également de fournir à un analyste des alternatives issues d'autres processus pour, au final, lui fournir de nouvelles idées et méthodologies pour concevoir et améliorer ses analyses.

Pour ce faire, un besoin est décrit comme un ensemble de dimensions finies. Ces dimensions représentent les différentes propriétés du besoin, comme sa description, son contexte, les variables initiales à analyser ou encore d'éventuelles hypothèses. Tous les termes utilisés par les acteurs pour décrire ces dimensions sont issus de l'ontologie et du vocabulaire contrôlé de notre approche. De plus, ces termes peuvent

être mis en relation entre eux *via* des propriétés sémantiques, elles aussi issues du vocabulaire contrôlé, pour former des graphes de concepts à rechercher. Par exemple, un enseignant pourrait définir la dimension de contexte pédagogique D_C de son besoin B comme étant un *Serious Game* sur les *maths*, et une autre dimension D_B faisant référence au besoin lui-même comme *connaître* si un *étudiant* va *terminer* ou non le *jeu*. Ainsi $B = D_C \cup D_B$.

Nous considérons qu'un processus d'analyse narré, ou une partie de processus (*i.e.* un ensemble d'étapes), répond à un besoin si, après évaluation de toutes les dimensions du besoin, le résultat est supérieur à un seuil d'acceptation défini par l'utilisateur.

Pour procéder à cette évaluation, chaque dimension est projetée dans notre ontologie des processus d'analyse narrés. Cette projection définit quels éléments de l'ontologie doivent être requêtés, ainsi que leur importance, d'après des heuristiques. Ainsi, les termes décrivant chaque dimension sont recherchés dans les éléments jugés les plus pertinents de l'ontologie. Pour chaque terme recherché, le degré de correspondance est évalué avec les termes contenus dans les éléments réifiant l'ontologie. En exploitant les propriétés du web sémantique, notamment les équivalences entre les termes, la pertinence des recherches se trouvera renforcée. De plus, en s'inspirant des travaux d'approximation comme Corese [8], il est possible de définir la notion de termes *similaires*. Cela permet d'envisager une propagation de la recherche à des solutions n'étant plus strictement égales, mais similaires (avec un certain degré), à un besoin donné. Qui plus est, quand un terme initial est substitué par un terme similaire, ses propriétés sémantiques sont aussi exploitables. En reprenant la description du besoin B , il est par exemple possible d'assimiler un *Serious Game* comme une *situation pédagogique* en utilisant le vocabulaire contrôlé, et ainsi explorer dans ce réseau sémantique les termes subsumés par ce dernier terme. L'analyse prédictive sur les MOOC présentée en section 3, qui est aussi une situation pédagogique, peut devenir un potentiel candidat. Notamment si, dans le vocabulaire, la notion *terminer le jeu* est définie comme équivalente (selon un certain degré) avec la *certification*. Les concepts non alignés, comme la discipline *maths*, diminuent alors le score de correspondance.

De plus, pour traiter l'imprécision engendrée par l'utilisation des différents degrés de similarité, les ensembles flous semblent pertinents. De cette manière, il est possible d'estimer si un processus narré (ou une partie de processus) répond au besoin de l'utilisateur, en accord avec la similarité attendue.

Enfin, analyser l'effet des opérations sur les variables représente un atout majeur par rapport aux solutions traditionnelles. En regardant l'évolution des graphes de variables, ainsi que les variables concernées par une opération, il est possible d'extraire des informations implicites qui peuvent ensuite être exploitées lors de la recherche, pour répondre au besoin. Par exemple, des coefficients de corrélation entre 0 et 1 de variables d'apprentissage sont obtenus *via* l'opérateur associé. Puis, par filtrage, ne sont gardés que ceux supérieurs à 0,9. Le système peut alors inférer qu'il s'agit de variables fortement corrélées. Le résultat de cette recherche est un ensemble de solutions, constituées de processus d'analyse narrés et d'ensembles d'étapes. Ces solutions sont identifiées en comparant si le degré de similarité obtenu est supérieur à celui attendu par l'utilisateur. Chaque élément contribuant au score est, grâce à l'ontologie, expliqué à l'utilisateur, lui permettant ainsi de vérifier le raisonnement du système.

Découverte d'informations critiques. Une autre assistance envisagée consiste en la découverte automatique d'informations critiques pour certains éléments de l'analyse. Une information est dite critique si, lorsque le contexte d'un processus d'analyse change, elle n'est plus pertinente et occasionne des erreurs (*e.g.* d'interprétation). Les paramétrages des opérateurs sont notamment concernés. Pour remédier à ce problème, nous prévoyons de nous appuyer sur les éléments narratifs utilisés lors de la description du processus d'analyse et sur des retours utilisateurs. Si l'on reprend les variables E utilisées pour entraîner le modèle de la section 3, des éléments narratifs de type contraintes et hypothèses peuvent être associés au processus, afin d'exprimer dans quelle situation il est pertinent de les utiliser. Par exemple, une hypothèse pourrait être qu'entraîner le modèle seulement avec les nombres d'actions comme c'est le cas ici, ne convient que dans le cadre d'un MOOC. Dès lors, le système d'assistance pourra détecter les éléments (*e.g.* étapes) ne convenant pas pour la réutilisation de l'analyse répondant à un nouveau besoin.

Ainsi, l'objectif est de mettre en correspondance, pour un processus d'analyse narré, ses particularités (*e.g.* paramétrage d'opérateurs) avec les informations narratives fournies. Ensuite, *via* la prise en compte des retours d'expérience utilisateurs dans d'autres contextes, cette correspondance particularités-informations sera modulée *via* un apprentissage par renforcement, de type *Q-learning*.

Grâce à cette correspondance évolutive, lors de la réutilisation d'un processus d'analyse narré, il sera alors possible d'assister l'analyste en lui indiquant les éléments auxquels il doit porter attention. De plus, en combinant cette correspondance avec la recherche par besoin (*cf.* 1^{ère} assistance), il est possible d'indiquer directement les éléments des solutions qui ne conviennent pas dans le contexte du besoin recherché.

Adaptation automatique. Enfin, la troisième assistance envisagée concerne l'adaptation automatique de processus d'analyse narrés en fonction d'un besoin ou d'un contexte exprimé par un utilisateur. Pour être capable d'adapter les processus automatiquement, il faut que la machine puisse identifier les éléments à modifier et trouver des substituts qui conviennent. Pour réaliser une telle assistance, nous prévoyons de nous appuyer sur les deux assistances préalablement présentées.

En effet, il semble pertinent de ne chercher à adapter que les processus qui correspondent le plus au besoin recherché par l'utilisateur. *Via* les mécanismes de raisonnement de la 1^{ère} assistance, le système est capable de récupérer une liste des processus d'analyse narrés jugés pertinents. À partir de cette liste de solutions, utiliser les mécanismes de découverte d'informations critiques de la 2^{ème} assistance nous permet de vérifier si chaque processus peut s'appliquer convenablement au besoin utilisateur. Pour chaque élément jugé critique, une alternative est recherchée dans la banque des processus d'analyse narrés, avec la méthodologie de la 1^{ère} assistance.

De plus, une trace du raisonnement peut également être construite grâce à l'utilisation conjointe des deux autres mécanismes d'assistance. Cela permet d'expliquer à l'analyste les inférences effectuées par la machine. Mais surtout, cela permet de demander audit analyste si les propositions d'adaptation automatique semblent pertinentes. Ainsi, grâce aux réponses de l'analyste, la détection d'informations critiques et les mécanismes d'adaptation pourront être affinés.

5 Conclusion

Dans cet article, nous avons présenté des assistances, rendues possibles par une ontologie dédiée à la narration des processus d'analyse. Ces assistances permettront de rechercher des processus d'analyse narrés répondant au besoin des acteurs, *via* plusieurs dimensions, tout en identifiant les spécificités dudit besoin pour adapter les processus par rapport au besoin initial. L'objectif est de renforcer les efforts liés à la capitalisation en y intégrant des solutions d'assistance jusqu'alors peu ou pas existantes dans les outils d'analyse traditionnels. Nous espérons ainsi tendre vers la définition d'un écosystème impliquant à la fois l'analyste et les différents acteurs intervenant lors de l'élaboration d'une analyse (*e.g.* décideur) afin de donner un sens à la co-construction des analyses. De plus, ces assistances constituent des pistes importantes pour une démarche qualité, notamment en limitant les biais potentiels lors de la réutilisation de processus d'analyse de traces d'apprentissage.

Références

1. Mandran, N., Ortega, M., Luengo, V., Bouhineau, D. : Dop8 : merging both data and analysis operator life cycles for TEL. In : Proceedings of LAK'15, ACM (2015) 213-217.
2. Clow, D. : An overview of learning analytics. *Teaching in Higher Education* 18(6) (2013) 683-695.
3. Lebis, A., Lefevre, M., Luengo, V., Guin, N. : Towards a capitalization of processes analyzing learning interaction traces. In : Proc. of the EC-TEL'16, Springer 397-403, (2016).
4. Baker, R.S., Yacef, K. : The state of educational data mining in 2009 : A review and future visions. *JEDM* 1(1) (2009) 3-17.
5. Sowa, J.F. : *Conceptual structures : information processing in mind and machine*. Addison-Wesley Pub., Reading, MA, (1983).
6. Alves de Medeiros, A.K., van der Aalst W.M.P. : Process Mining towards Semantics. In : *Advances in Web Semantics I*, LNCS 4891. Springer, Berlin, Heidelberg (2008) 35-80.
7. Page, K. et al. : From workflows to research objects : an architecture for preserving the semantics of science. Proc. of the 2nd International Workshop on Linked Science (2012).
8. Corby, O., Dieng-Kuntz, R., Faron-Zucker, C. Gandon, F. : *Ontology-based approximate query processing for searching the semantic web with corese*. Technical report (2009).
9. Widyantoro, D.H., Yen, J. : A fuzzy ontology-based abstract search engine and its user studies. In : Proc. of the 10th IEEE Int. Conference on Fuzzy Systems (2001) 1291-1294.
10. Lebis, A., Lefevre, M., Luengo, V., Guin, N. : Capitalisation of analysis processes : Enabling reproducibility, openness and adaptability thanks to narration. In : Proceedings of LAK'18, ACM (2018) 245-254.
11. Siemens, G. et al. : *Open Learning Analytics: an integrated & modularized platform*. Technical report (2011). Society for Learning Analytics Research.
12. Belhajjame, K et al. : Why Workflows Break – Understanding and Combating Decay in Taverna Workflows. In : Proc. of the 8th International Conference on E-Science (2012) 1-9
13. Chatti, M.A., Dyckhoff, A.L., Schroeder, U. et Trüs, H. : A reference model for learning analytics. *International Journal of Technology Enhanced Learning* 4, 5-6 (2012), 318-331