Rapport de thèse

1ère année - Alexis Lebis

Contexte

Cette thèse se situe dans le cadre du projet ANR HUBBLE [1] dont l'objectif est de créer un observatoire national pour la construction et le partage de processus d'analyse de données massives, issues des traces laissées dans des environnement de type e-learning. HUBBLE devrait permettre d'analyser et d'expliquer des phénomènes d'enseignement et d'apprentissage avec ces environnements.

Cette thèse est dirigée par Vanda Luengo de l'équipe de recherche MOCAH du LIP6 et co-encadrée par Marie Lefevre et Nathalie Guin de l'équipe TWEAK du laboratoire LIRIS. La thèse est rattachée à l'UPMC Université Paris 6.

Dans cette thèse, nous nous intéressons plus précisément à la capitalisation des processus d'analyse de traces dans le contexte du e-learning. Ce type d'apprentissage est défini par l'utilisation d'outils informatiques par des apprenants, en présence ou non de tuteurs. Au sein de ces outils, divers types de ressources pédagogiques sont disponibles (e.g. vidéos, quiz). Ces dispositifs génèrent des données qui relatent l'interaction des utilisateurs entre eux (e.g. messages privés, forum), ainsi que les activités faites sur les ressources des dispositifs. Nous parlons dès lors de traces d'interactions, qui sont considérées comme réservoirs de connaissances ; connaissances pouvant être découvertes par l'analyse de ces traces.

Une analyse est représentée par un processus d'analyse, ici de traces. Il s'agit de l'application de techniques, d'opérations et de méthodes, appelées opérateurs [2], pour produire des connaissances comme des modèles ou des indicateurs. Les acteurs de cette analyse peuvent être des statisticiens, des analystes ou encore des chercheurs. Concevoir un processus d'analyse de traces est une tâche s'avèrant souvent complexe et fastidieuse : le besoin doit être correctement cerné et les données, ainsi que les connaissances du domaine, correctement communiquées. La publication des résultats est également cruciale afin que ceux-ci puissent être utilisés convenablement. Actuellement, les processus d'analyse sont dépendants des outils d'analyses utilisés, des opérateurs qui y sont disponibles, ainsi que de la représentation des données contenues dans les traces. Cela fait disparaître l'objectif des opérateurs utilisés, élude le sens des données manipulées et altère la démarche intellectuelle. De fait, en pratique, pour un même besoin, un processus d'analyse peut être décliné sur différents outils d'analyse, en fonction du format des données concernées : la capitalisation est inexistante [3].

Lorsque nous parlons de capitalisation, nous entendons la faculté de disposer de l'existant pour en tirer profit. Dans notre cas, cela concerne la reproductibilité, la réutilisation, le

partage et l'adaptabilité des processus pour le spectre des outils d'analyse disponibles. Dans la littérature, il est possible de distinguer deux styles d'approches allant dans le sens de la capitalisation.

L'une est axée sur les processus d'analyses en eux-mêmes, où à plus forte raison d'une certaine partie des opérateurs le constituant. Actuellement, il est possible de dégager une tendance qui est de proposer un outil d'analyse, ouvert en terme d'accessibilité et de consultation, où les processus et opérateurs sont partageables. Cependant, de tels processus ne sont pas réutilisables trivialement en dehors de leurs outils, ni de leurs contextes initiaux, sans compter les lacunes descriptives qui peuvent exister.
Il existe aussi des travaux proposant de partager vers d'autres outils d'analyses. Cependant, ils se heurtent à des problèmes complexes, comme l'interopérabilité technique, une perte de sémantique des opérateurs utilisés ou encore l'évolution des outils d'analyses [4].

Le second axe se concentre sur les données, où l'objectif est d'en proposer la capitalisation, toute proportion éthique gardée. L'un des besoins majeurs étant de proposer des représentations avec pour objectifs d'assurer la fiabilité des données utilisées, mais aussi de permettre l'alignement des concepts identiques et permettre de rendre les données compréhensibles à la fois d'un point de vue humain et machine. Des travaux proposent d'effectuer un *mapping* vers des formalismes plus génériques, proposant ainsi des cadres de travail contrôlés pour ré-exploiter facilement les données. Cependant, augmenter la généricité des données n'exempte pas les opérateurs qui les utilisent d'être contraints par les outils gérant le *mapping* : cela ne résout pas le problème de capitalisation des processus d'analyse de traces [5].

De plus, il n'est pas non plus possible de trivialement combiner ces deux axes, principalement à cause de l'intrication des dépendances -entre autres structurelles- qui existent entre opérateurs et données. Et quand bien même, de telles approches n'apporterait aucune solution pour l'assistance aux utilisateurs dans la compréhension et la réutilisation directe des analyses, ainsi que pour l'adaptabilité des processus d'analyses. Mais, malgré cela, ces travaux constituent une preuve en soit que le besoin d'échange et de partage est bien réel.

Objectif de la thèse

L'objectif de cette thèse est de proposer aux différents acteurs concernés par les processus d'analyse de traces de e-learning des modèles et des méthodes capables de les abstraire, ***a posteriori,*** des différentes plates-formes d'analyses. Cela afin de rendre ces processus d'analyses indépendants des contraintes techniques et représentatifs de leur contexte d'utilisation, pour être reproduits, réutilisés, modifiés, partagés et enrichis par toute la communauté, permettant de ce fait une capitalisation efficace.

Bilan de la 1ère année

État de l'art

L'objectif principal de l'étude bibliographique a été de cerner les différentes disciplines liées à l'analyse des traces dans le domaine du e-learning pour prendre en compte les avantages et dégager les limitations existantes.

Un pan de l'analyse est lié à l'Educational Data Mining (*EDM*). L'EDM est directement issue du Knowledge Discovery in Database [6], discipline qui vise à pallier les lacunes du processus de data mining en l'intégrant dans un cycle plus important, permettant d'obtenir des données raffinées plus sensées. L'EDM est principalement une approche très peu (voire non) supervisée, bien qu'itérative, qui vise à prendre en compte les spécificités du monde éducatif, comme une granularité plus fine des données ou encore l'identification des utilisateurs. Il faut cependant noter que "la plupart des outils de data mining pour l'EDM sont trop complexes d'utilisation pour les éducateurs, et les possibilités vont bien au delà de leurs besoins" [7].

Les Learning Analytics (*LA*) concernent la collection, l'étude, l'analyse des données provenant des étudiants et de leurs contextes, dans l'objectif de comprendre et d'optimiser l'apprentissage et les environnements de lequel cet apprentissage prend part [8]. Il se situe à la croisée de plusieurs autres disciplines, dont l'EDM. Il s'agit d'une approche itérative et interactive, qui implique fortement les différents acteurs, dont ceux pédagogiques, afin de tirer pleinement partie des spécificités du domaine du e-learning [9].

Ces différents acteurs occupent des rôles précis dans le cycle analytique que l'on peut extraire de la littérature. On identifie évidemment l'analyste, mais aussi l'expert du système éducatif. En plus de cela, on extrait de la littérature le preneur de décision, par exemple un directeur d'école, et aussi le client des données, comme les enseignants, et les sujets des données qui sont la source des données. Notre étude des rôles a pour but de mettre en avant l'ensemble hétéroclyte des compétences disponibles et la nécessité de les considérer dans nos travaux, ceci pour des apports pertinents [10].

De plus, un champ de travail et de recherche qui a attiré notre attention est celui des *workflows*, dû à sa grande diversité d'utilisation (e.g. génétique, entreprise, astronomie). Des travaux intéressants existent pour représenter des analyses de manière séquentielle et visuelle, inscrit dans une optique de partage scientifique. Prendre ces travaux en compte, en plus de leurs récents apports sur l'aspect descriptif et sémantique, constitue un élément important pour une solution de capitalisation efficace [11].

Proposition

Formalisme

Pour tendre vers une capitalisation, nous avons proposé qu'un processus d'analyse puisse être considéré comme indépendant des outils. Pour cela il est nécessaire que chacun des éléments le constituant le soit aussi (paramètres, données, opérateurs...). Nous avons donc travaillé sur un formalisme permettant de représenter un opérateur de manière indépendante des outils d'analyse, ainsi que des données qu'il va exploiter.

Un opérateur indépendant représente un objectif, et regroupe des opérateurs ayant cet objectif mais implémentés dans des outils d'analyses différents et donc ayant des contraintes techniques différentes. Par exemple, réaliser un intervalle de temps peut être fait de différentes manières en fonction des outils, mais un opérateur indépendant *Time Range* représentera le concept de créer un intervalle, sans dépendre d'outils. Ces opérateurs indépendants n'ont pas vocation à calculer directement les données. Ils tiennent lieux d'éléments descriptifs, faisant ainsi des processus d'analyse décrits l'expression d'une méthodologie pour obtenir une information pertinente. L'instanciation de ces processus est envisagée par l'utilisation d'instructions décrite de manière stéréotypée pour chaque opération, dans chacun des outils d'analyses.

De plus, les processus d'analyse sont basés sur l'exploitation des données et sont intrinsèquement liés à elles. Il n'est donc pas possible pour obtenir une description efficace d'un processus de ne pas les considérer. Cependant, ces données sont elles aussi soumises à des contraintes, comme leur formalisme (e.g. CSV ou RDF), posant des problématique d'interopérabilité et de capitalisation. Dès lors, il convient de trouver un point d'équilibre entre l'information contenue, son utilité et les contraintes techniques qu'elle génère. Nous suggérons, comme point d'équilibre pour une capitalisation efficace, de travailler avec les concepts que dégagent les données -appelés types d'éléments traces (TTE, e.g. heure de début), à l'instar des opérateurs indépendants, et non pas directement avec leurs valeurs (e.g. 12h45) : cela permet de s'affranchir du problème de la représentation des données. Le descripteur de l'analyse déclare alors les TTEs requis initialement pour effectuer l'analyse. Ces TTEs seront exploités par les opérateurs indépendants pour générer des informations tout au long de la description du processus (e.g. feedback), comme le montre la figure 1.

Nous avons regroupé ces notions d'opérateur indépendant et de types d'éléments tracés au sein d'un méta-modèle permettant de décrire des processus d'analyse indépendants. Les différents éléments de notre formalisme ont été obtenus par l'étude empirique de différents outils : kTBS, UnderTracks, Knime, Weka, Usage Tracking Language, SPSS.

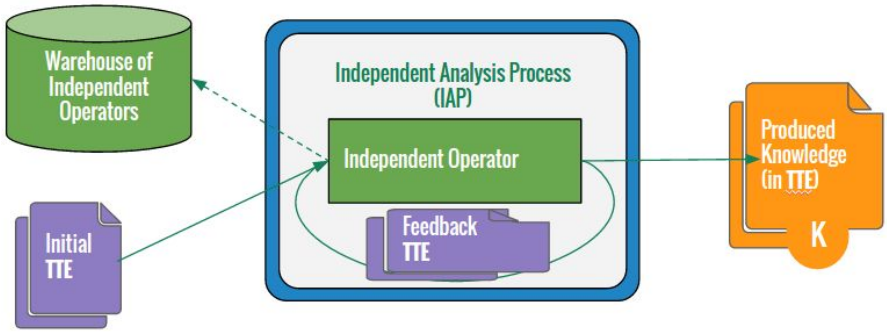

Figure 1 : Représentation des éléments impliqués dans la description d'un processus d'analyse indépendant permettant, sur cet exemple, d'obtenir une connaissance (K).

Implémentation et expérimentation

Les différents éléments de notre formalisme ont été implémenté au sein d'une application web qui a servie de support à six expérimentations impliquant différents acteurs [4] : 2 informaticiens, 3 statisticiens et 1 cogniticien. La démarche expérimentale se composait de trois parties.

La première consistait à matérialiser un besoin d'analyse puis à décrire -textuellement ou graphiquement- le processus d'analyse associé et à évaluer subjectivement sa difficulté. Cette partie permet d'étudier comment un acteur réfléchit et conçoit un processus, exempt de spécificités techniques et de valeurs concrètes : a-t-il besoin des valeurs pour s'exprimer ? à quelle granularité opère-t-il sa description ?

La deuxième partie permettait l'appropriation, par la pratique, des concepts de notre approche. Le sujet était guidé dans la création d'un processus d'analyse au sein du prototype. Nous étudions principalement si le sujet arrivait à créer une analogie entre sa démarche intellectuelle et l'approche mise en œuvre par le prototype.

La troisième partie consistait à transcrire le processus d'analyse décrit lors de la partie 1 dans le prototype. Cela nous permet d'étudier une éventuelle divergence sémantique entre les deux processus et d'en étudier les raisons. De plus, cette partie permet d'analyser le comportement du sujet vis-à-vis des concepts abstraits qu'il manipule et la manière dont il opère la description de son processus générique.

Tout au long de l'expérimentation, les sujets étaient autonomes. Ils étaient filmés et évalués par un observateur *via* des grilles d'évaluation portant sur le comportement et l'aisance face aux différents concepts de l'approche et du prototype. Les deux dernières parties étaient suivies d'un questionnaire. Tous les documents et productions expérimentales autorisés à être diffusés, ainsi que le prototype complet, sont accessibles en ligne[1].

Les résultats expérimentaux encouragent fortement l'idée qu'il est possible de décrire un processus d'analyse de cette manière, en conservant sa sémantique et ses objectifs, tout en

[1] https://liris.cnrs.fr/~alebis/iogap.html

permettant de connaître les outils d'analyses capable de le réaliser. Nous avons montré que la réification d'un processus d'analyse issu d'une démarche cognitive était possible.

Cycle analytique itératif

L'étude de la littérature axés autour de l'analyse et des différentes personnes impliquées nous a permis d'identifier et de proposer une nomenclature du cycle analytique, se voulant itératif par nature (cf. Figure 2).

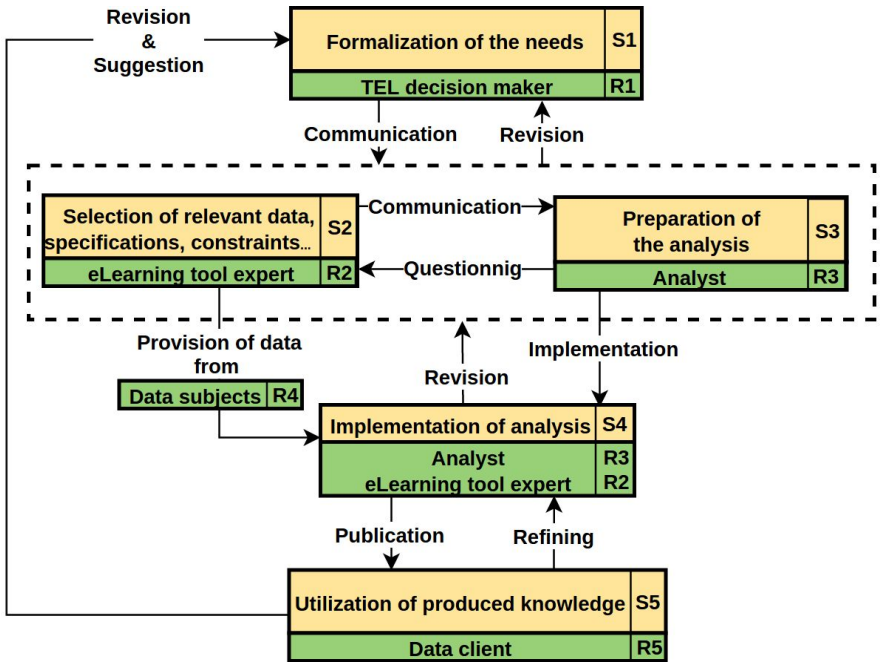

Figure 2 : Représentation du cycle analytique itératif obtenu par l'étude de la littérature. Il fait intervenir le preneur de décision (TEL decision maker), l'expert du système (eLearning tool expert), l'analyste (Analyst), les sujets des données (Data subjects) et les clients des données (Data client).

Ce cycle contextualise les différents rôles au sein de différentes étapes et montre les principales interactions y survenant. L'analyste, qui est chargé de définir les stratégies de l'analyse ainsi que de l'implémenter, se retrouve en étroite collaboration avec l'expert du système, qui a la lourde tâche de fournir le contexte -principalement technique, ainsi que les données, à l'analyste en fonction du ou des besoins formalisés. Ces besoins sont formalisés par le preneur de décision qui a approuvé les besoins émis par le client des données.

Perspectives

Bien que les résultats des expérimentations menées dans le cadre de notre approche visant la capitalisation des processus d'analyse soient fortement encourageants, il faut cependant noter plusieurs limites, notamment : (1) le manque de sémantique des différents éléments manipulables dû au caractère abstrait de l'approche, limitant de ce fait la compréhension globale de l'analyse et les facultés descriptives et (2) un manque avéré de feedback pour les utilisateurs, concernant entre autres, l'influence que peut avoir un opérateur indépendant sur

des types d'éléments tracés (TTE) et (3) la non prise en compte du contexte d'implémentation.

Nous travaillons donc actuellement sur la question d'une description plus riche des processus d'analyses indépendants, de leurs objectifs et de leurs contextes applicatifs ; cela afin qu'ils puissent se suffire à eux-même sur un plan sémantique, et pour servir de support d'aide à la compréhension, à la décision, à la réutilisation et à la diffusion scientifique. Nous exploitons pour cela le vocabulaire xAPI au sein d'une ontologie décrivant des processus d'analyses.

Cependant, cette description requiert aussi le besoin de s'adapter aux spécificités propres du domaine de l'e-learning. Il est intéressant de se demander s'il est possible de faire ressortir des concepts récurrents et des habitudes d'analyses propres aux données de e-learning, dans l'idée d'aiguiller l'effort scientifique. Pour ce faire, nous envisageons un écosystème évolutif et relationnel concernant la base de vocabulaire contrôlé offerte par xAPI, mais aussi des opérateurs et processus d'analyses indépendants.

Cette description plus riche soulève également la question de savoir quelles informations utiles peuvent être inférées par une description des processus d'analyse de haut niveau. Nous nous posons aussi cette question afin d'assister l'utilisateur dans son élaboration *via* des mécanismes de recommandation et de lui apporter des feedbacks pertinents et contextuels. À partir de tous ces éléments, il faudra donc voir comment assister efficacement les différents acteurs de l'analyse. Il serait intéressant, par exemple, de rechercher des processus connexes à des besoins, ou encore de s'assurer de la crédibilité scientifique d'une analyse et de la bonne lecture des résultats.

Références

[1] http://hubblelearn.imag.fr/?lang=fr

[2] Mandran, N., Ortega, M., Luengo, V., Bouhineau, D. : Dop8 : merging both data and analysis operators life cycles for technology enhanced learning. In : *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. pp. 213–217. ACM (2015)

[3] Dyckhoff, A.L., Zielke, D., Bültmann, M., Chatti, M.A., Schroeder, U. : *Design and implementation of a learning analytics toolkit for teachers.* vol. 15, pp. 58–76. JSTOR (2012)

[4] PMML 4.1 by DMG. http://dmg.org/pmml/pmml-v4-1.html, accessed : 2016-09-20

[5] Lebis, A. : Vers une capitalisation des processus d'analyse de traces. In : Rencontres Jeunes Chercheurs en EIAH (RJC-EIAH 2016). Montpellier, France (Jun 2016)

[6] Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146. http://doi.org/10.1016/j.eswa.2006.04.005

[7] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases, 17(3), 37. http://doi.org/10.1609/aimag.v17i3.1230

[8] Ferguson, R. (2012). The state of learning analytics in 2012: a review and future challenges. *Technical Report KMI-12-01*, *4*(March), 18. http://doi.org/10.1504/IJTEL.2012.051816

[9] Duval, E. (2011). Attention please!: Learning analytics for visualization and recommendation. LAK '11 Proceedings of the 1st International Conference on Learning Analytics and Knowledge, 9–17. http://doi.org/10.1145/2090116.2090118

[10] Drachsler, H., Greller, W., Greller Wolfgang, & Drachsler Hendrik. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, *15*(3), 42–57. http://doi.org/10.1145/2330601.2330634

[11] Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., … Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric research objects. *Journal of Web Semantics*, *32*, 16–42. http://doi.org/10.1016/j.websem.2015.01.003