Généralisation de motifs séquentiels pour la fouille de données multi-sources

Julie BU DAHER

Mots-clés – Niveaux et Public concernés

Mots-clés : analyse de données, fouille de motifs séquentiels, données multi-sources.

Niveaux : collège–lycée

Public : élèves

À quelles questions cette thèse tente-t-elle de répondre ?

L'objectif général de ce travail de recherche est d'améliorer l'apprentissage scolaire. D'une part, il vise à aider les élèves à comprendre, évaluer leurs performances scolaires et suivre leurs progrès. D'autre part, il vise à aider les élèves à améliorer leurs acquis en fournissant à chacun des recommandations personnalisées. Afin de répondre à nos objectifs, notre travail de recherche s'intéresse à une phase amont : l'analyse et à la compréhension des données d'activité des apprenants ainsi qu'à l'extraction d'informations de comportement-type.

Notre question de recherche principale était : les sources multiples d'informations sur les apprenants, leur comportement et les ressources pédagogiques permettent-elles d'identifier des comportements-type riches, dans un contexte où la quantité de traces et d'informations est limitée ?

Pourquoi ces questions sont-elles pertinentes ?

L'objectif est d'aider les apprenants à évaluer leurs performances scolaires et à suivre leurs progrès, mais aussi aider les apprenants à améliorer leurs niveaux scolaires. Il est en effet de la plus haute importance de pouvoir fournir une information aux apprenants, a fortiori lorsqu'ils travaillent en ligne, et que l'enseignant n'est pas présent. Pour fournir cette information, il est important de comprendre les traces des activités qu'ils réalisent. Les données exploitées sont les traces numériques des activités sur leur environnement numérique de travail (ENT). L'extraction d'informations à partir des données des apprenants peut concerner l'étude de leur rythme, fréquence de travail, résultats, etc. et permet donc de leur fournir des informations compréhensibles qui leur permettent d'évaluer leurs performances

scolaires et de suivre leurs progrès. Les données disponibles peuvent également concerner les apprenants eux-mêmes et les ressources pédagogiques. Les informations extraites peuvent également être utilisées pour aider les apprenants à améliorer leur niveau scolaire en leur fournissant des recommandations personnalisées.

Quelle méthodologie de recherche a-t-on utilisée ?

L'analyse de données (*Data Analytics*) consiste à examiner des données brutes, d'en extraire de l'information, de tirer des conclusions et de prendre des décisions. En éducation, comprendre les données des apprenants est une première étape fondamentale pour identifier leurs acquis, estimer leurs résultats possibles, recommander des ressources pédagogiques à consulter par les apprenants sur leur ENT ou encore leur recommander des ressources pédagogiques en fonction de leurs besoins et de leurs préférences.

Les données des apprenants sont multiples. Elles incluent à la fois des données sur leur comportement et sur leurs traces d'activité sur l'ENT. Par exemple, ces données représentent les séquences de ressources pédagogiques consultées associées au moment de cette consultation, des données descriptives des apprenants (données démographiques, résultats scolaires, etc.), ou encore des données descriptives sur les ressources pédagogiques consultées.

Dans ce travail de recherche, l'objectif est de permettre l'exploitation de données réelles : traces d'activité réelles d'apprenants sur leur environnement numérique de travail en conditions réelles d'activité, données d'apprenants et de ressources réelles, etc. Nous avons conçu un algorithme qui permet d'extraire des motifs récurrents d'apprentissage, à partir des données recueillies, par techniques de fouille de données (Bu Daher et Brun, 2020 ; Bu Daher *et al.*, 2019 ; Bu Daher *et al.*, 2020) pour mieux comprendre les apprenants. Un exemple de motif est donné ci-dessous.

Le processus de fouille de motifs ainsi que les motifs fréquents générés par ce processus sont interprétables, ce qui permet, par exemple, de proposer des recommandations. Pour permettre cette extraction de motifs, une quantité significative de traces est nécessaire, sur des apprenants d'un même niveau (5e par exemple), pouvant provenir de plusieurs classes, voire de plusieurs établissements. Cependant, sur des données réelles, les quantités de données ne sont pas infinies, et le processus de fouille fait rapidement face à un manque de données. L'algorithme conçu permet de pallier ce manque de données en tirant profit de la multitude de sources des données complémentaires mises à disposition. Ainsi, l'ensemble des données de cette étude contient différents types de données provenant d'une seule source ou de plusieurs sources de données. Malgré sa complexité et son hétérogénéité, cet ensemble de données est riche et contient une grande quantité d'informations qui pour certaines sont pertinentes et exploitables. C'est cette complexité et cette richesse qui constituent le défi scientifique de la fouille associée.

En e-éducation, les données de traces sont séquentielles et peuvent représenter les séquences de ressources pédagogiques (examens, exercices, etc.) que les apprenants consultent sur leur ENT dont voici un exemple : élève-143 : $\langle R_3\ R_8\ R_{13}\ R_{27}\ R_{29}\rangle$, où élève-143 représente l'identifiant de l'élève et R_n représente l'identifiant d'une ressource pédagogique. La séquence signifie que l'étudiant a consulté les ressources R_3 , puis R_8 , puis R_{13} , puis R_{27} et enfin R_{29}.

Quels résultats a-t-on obtenus ?

Nous proposons une approche qui gère les données multi-sources en un seul processus de fouille. Dans ce cadre, nous proposons de limiter la complexité du processus de fouille en considérant une source comme étant principale et les sources de données supplémentaires sont fouillées de manière sélective, c'est-à-dire uniquement lorsque cela est nécessaire pendant le processus.

Deux résultats principaux peuvent être associés à ce travail de recherche. Le premier résultat concerne l'identification de plusieurs types de relations entre sources de données, et la nature des motifs fouillés associés. Un premier type de relations existe entre la source de données séquentielles

et la source de données descriptive des apprenants. La source de données descriptive des apprenants fournit des informations spécifiques à chacun. Lorsque ces données sont rapprochées de la source de données séquentielles, elles permettent d'obtenir des motifs séquentiels fréquents plus précis que ceux générés par l'exploration traditionnelle de données séquentielles uniquement. Ainsi, ce type de relations permet d'extraire des informations associant le comportement numérique des apprenants et la source de données descriptive de ceux-ci.

Un deuxième type de relations existe entre la source de données séquentielle et la source de données descriptives des ressources pédagogiques. La source de données séquentielle contient des séquences des ressources pédagogiques, et la source de données descriptives des éléments fournit des données supplémentaires sur chaque élément des séquences. Lorsque les données descriptives des ressources sont fournies aux données séquentielles, chaque ressource est décrite par des attributs descriptifs supplémentaires. Ces attributs représentent des informations plus générales que les identifiants des ressources, donnant donc plus de généralités aux motifs fouillés.

Le second résultat concerne l'algorithme de fouille qui gère les deux types de relations. L'algorithme conçu tire l'avantage de la source de données descriptives des ressources pédagogiques afin de générer des motifs séquentiels généraux. Afin de gérer le problème de la similarité des données et de générer des motifs plus fréquents, nous définissons deux mesures de similarité : la similarité de motifs qui compare différents motifs et la similarité d'éléments qui compare différents éléments dans les motifs où un élément représente une ressource pédagogique. Ensuite, nous formons des motifs généraux à partir de motifs similaires, c'est-à-dire contenant des éléments similaires. Un motif général est un motif qui contient des informations plus générales que les motifs fréquents traditionnels. Enfin, nous proposons une nouvelle méthode pour détecter des motifs généraux fréquents. Notre algorithme permet de résoudre le problème de la similarité des données et de la faible couverture des données en générant des motifs plus fréquents ; de plus, les motifs généraux fréquents sont riches car ils contiennent divers types d'informations.
Les motifs fréquents extraits de cet algorithme sont exploités pour former des règles d'association, et ces règles sont utilisées pour fournir des recommandations personnalisées aux apprenants. Les informations descriptives fréquentes des apprenants extraites par l'algorithme permettent d'identifier les apprenants ayant des profils similaires, et les informations séquentielles fréquentes associées sont utilisées 1/ pour identifier le comportement passé des apprenants, et, 2/ pour leur fournir des recommandations personnalisées basées sur leurs besoins académiques. Ces recommandations sont représentées sous la forme d'une ressource pédagogique ou d'une séquence de ressources pédagogiques à consulter par les étudiants sur leur ENT pour les aider à améliorer leur niveau académique. Les motifs généraux fréquents extraits de cet algorithme permettent d'obtenir plus de possibilités de recommandations.

Les données du projet sont encore en cours d'acquisition. L'algorithme conçu, bien que visant un contexte éducatif, est volontairement générique. La validation de cet algorithme a été effectuée sur des données de structure et caractéristiques proches. Ce sont des données musicales qui ont été exploitées. Trois sources de données sont disponibles : des données d'écoute de musique par des utilisateurs, des données descriptives des ces utilisateurs et des données de description des musiques. Le corpus de données est composé des écoutes de 2 000 utilisateurs sur 34 000 musiques, où chaque utilisateur a en moyenne 48 consultations. L'évaluation de l'algorithme a porté à la fois sur sa capacité à s'exécuter en un temps limité, avec une évaluation sur le temps d'exécution et la mémoire utilisée. L'évaluation relative aux motifs extraits a permis de valider non seulement le fait que certains motifs ne nécessitent pas de généralisation, car présents en quantité suffisante, mais également que le processus de généralisation permet effectivement d'extraire des motifs généraux, palliant ainsi au manque de données.

Que dois-je retenir de cette étude pour ma pratique ?

En l'état actuel, l'outil n'est pas utilisable par les élèves, en particulier en raison d'un souci d'accès aux données de comportement numérique. Cependant, dans le futur il est envisagé de l'intégrer dans un

outil d'auto-évaluation pour les apprenants afin de suivre leurs progrès, comprendre leurs performances scolaires dès le début du processus d'apprentissage. De plus, les informations extraites des données des apprenants via l'algorithme que nous avons développé pourraient également être utilisées pour fournir des recommandations personnalisées aux apprenants afin de les aider à améliorer leur niveau scolaire, quelle que soit la discipline.

Références

Bu Daher, J. et Brun, A. (2020). Handling Item Similarity in Behavioral Patterns through General Pattern Mining. Dans *The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'20)*.

Bu Daher, J., Brun, A. et Boyer, A. (2019). Multi-source Relations for Contextual Data Mining in Learning Analytics. *arXiv preprint arXiv :1907.04643*.

Bu Daher, J., Brun, A. et Boyer, A. (2020). Multi-source data mining for e-learning. *arXiv preprint arXiv :2009.08791*.