Une méthode de visualisation de traces d'activités sur la plateforme open edX : le cas du Mooc « Enseigner et former avec le numérique »

www.adjectif.net/spip/spip.php

Pour citer cet article :

Boelaert, Julien et Khaneboubi, Mehdi (2015). Une méthode de visualisation de traces d'activités sur la plateforme open edX : le cas du Mooc « Enseigner et former avec le numérique ». *Adjectif.net* [En ligne]. Mis en ligne le Vendredi 19 juin 2015. URL : http://www.adjectif.net/spip/spip.php?article350

Résumé :

Ce texte propose une analyse de données susceptible d'éclairer les pratiques d'apprentissages en ligne dans le cadre de Moocs. Après avoir codé des activités pour chaque semaine du cours (répondre au quiz, regarder des vidéos, etc.), un classement automatique permet d'identifier 4 types de parcours.

Mots clés :

France

Introduction

Les *massive open online courses* (Moocs) génèrent des traces d'activités en ligne que l'on peut considérer comme des données d'apprentissage. Lors de la deuxième édition du Mooc *enseigner et former avec le numérique* (eFAN) [1], on a compté plus de 4800 inscrits et presque 500 dépôts d'activités [2]. Les actions en ligne des participants ont généré des traces de connexions sur les serveurs du Mooc (journal de connexion ou *logs)*. Le journal de connexions à un serveur web est par exemple à la base des services web comme *Xiti* ou *Google analytics*.

Les logs constituent des informations non déclaratives qui concernent toutes celles et ceux qui se connectent sur un site web. Dans notre contexte, il s'agit de données potentiellement intéressantes pour par exemple identifier des trajectoires d'apprenants, repérer des comportements typiques ou simplement effectuer des mesures d'audience.

En se basant sur des logs de connexion, Kizilcec et al. (2013) ont mesuré les flux de participants de semaine en semaine et construit différents profils comportementaux dans trois Moocs d'informatique et de mathématiques sur la plateforme *Coursera* [3]. La modélisation statistique qu'ils ont réalisé n'est pas sans défauts. D'abord, les choix de codages opérés insistent sur le fait d'avoir suivi les vidéos et quiz *à temps*, c'est-à-dire la semaine de leur mise en ligne conduit à une vision relativement scolaire des pratiques d'apprentissage en ligne, qui n'est pas forcément pertinente dans l'étude des *Moocs*, dont l'organisation interne est souvent modulaire.

Ensuite, la représentation graphique qu'ils produisent, en bulles et en flux, donne une illusion d'abstraction des pratiques en ligne tout en masquant la complexité des données. Enfin, leur méthode de classification automatique (clustering) des *types* des participants ne prend pas en compte l'aspect temporel des semaines de

cours.

Autrement dit, dans leur modèle statistique, l'ordre dans lequel un participant a consulté les contenus de cours n'est pas pris en compte. Ainsi, le résultat de leur classification serait exactement le même si l'on modifiait l'ordre des semaines dans la base de données. Nous proposons ici une autre méthode de visualisation et de classification des données de connexions, appuyée sur les outils de l'analyse de séquences.

Données de connexion

La plateforme *open edx* [4], administrée en France sous le nom de *France université numérique* (FUN) [5], sur laquelle s'est tenu le Mooc eFAN2, a produit un fichier répertoriant environ 790 000 connexions pour lesquelles figurent notamment :

— la date et l'heure de la connexion,

— l'URL de la page consultée,

— le type d'action réalisé avec un objet vidéo,

— le type de navigateur, de système d'exploitation, etc.

Comme souvent avec ce type de données, le fichier qui nous a été communiqué n'était pas exhaustif et comporte des défauts et des manques. On compte par exemple environ 3 700 inscrits dans les logs alors que la plateforme en recense un peu plus de 4 600. En revanche, les données dont on dispose sont suffisamment cohérentes et consistantes pour conduire une réflexion méthodologique. Comment résumer graphiquement des comportements complexes sans pour autant produire des visualisations trop abstraites ? Quelle méthode de classification automatique est susceptible de produire des profils de participants prenant en compte les contraintes temporelles inhérentes à ce mode d'enseignement ?

Caractérisation de trajectoires

Les outils d'analyse des trajectoires [6] semblent parfaitement indiqués pour l'analyse des usages de Moocs. Ces outils de visualisation et d'analyse statistique, dont certains algorithmes ont initialement été développés pour l'étude des séquences d'ADN, permettent de prendre en compte explicitement la dimension temporelle des phénomènes observés. Ils reposent sur la définition de « séquences », c'est-à-dire des données dans lesquelles on suit un certain nombre d'individus au cours de plusieurs périodes (Gabadinho et al., 2011). À chaque période, un individu est considéré comme dans un certain « état ». Ces états sont définis initialement dans un *dictionnaire*.

Dans le cas des Moocs, le découpage en semaines de cours donne assez naturellement lieu à un découpage en périodes : chaque vidéo, chaque activité appartiennent à une certaine semaine de cours, et l'on peut donc pour chacune de ces semaines définir un état pour chaque utilisateur, selon son activité. Par exemple, un participant a-t-il regardé les vidéos de la semaine 3 ? A-t-il passé l'interrogation de cette semaine ? Notons que cette « périodisation » est un choix parmi d'autres possibles, et qu'elle ne rend pas compte de toute la complexité des comportements ; elle ne différencie pas un utilisateur qui aurait suivi toutes les vidéos dans l'ordre, regardant chacune le jour de sa sortie, d'un autre utilisateur qui les aurait toutes regardées en une seule fois à la fin du cours. Elle semble néanmoins intéressante, en ce qu'elle permet d'analyser l'activité sur le cours dans l'ordre dans lequel il a été pensé par ses concepteurs.

Les logs du Mooc eFAN2 nous permettent de définir cinq états pour chaque couple utilisateur-semaine :

- absent : l'utilisateur ne s'est jamais connecté aux ressources correspondant à cette semaine ;
- présent : l'utilisateur s'est connecté aux pages de la semaine, mais n'a pas regardé de vidéos, ni répondu aux quiz (il a pu télécharger les vidéos, ou lire les textes, ou simplement consulter les pages pour suivre les liens vers d'autres sites) ;
- vidéo : l'utilisateur a regardé au moins une vidéo de la semaine, mais n'a pas répondu au quiz

(malheureusement, les logs edX ne permettent pas de savoir si l'utilisateur a téléchargé une vidéo, on sait seulement s'il l'a regardée directement sur le site du cours) ;

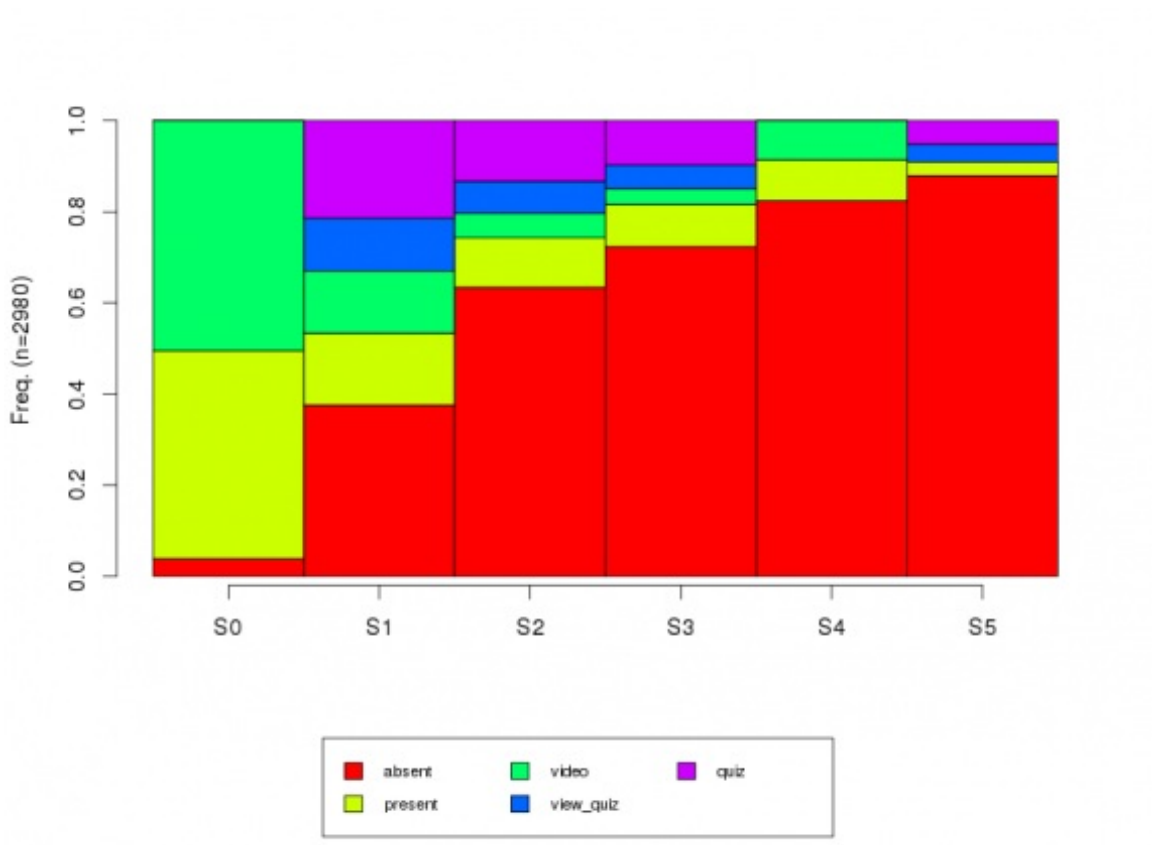
- view quiz : l'utilisateur a consulté le quiz de la semaine, mais n'y a pas répondu ;
- quiz : l'utilisateur a répondu à au moins une question du quiz de la semaine.

Nous considérons ces cinq états comme représentatifs de degrés d'investissement dans le *Mooc*, listés ci-dessus dans un ordre croissant ; un utilisateur qui a à la fois regardé des vidéos et consulté les questions du quiz sera codé dans l'état *view quiz*, qui indique un niveau d'investissement plus élevé.

D'autres choix auraient été possibles par exemple, construire plus d'états différents (vidéo uniquement, quiz uniquement, vidéo et quiz…), mais la nomenclature choisie présente le double avantage d'être simple et de représenter un gradient d'investissement cohérent.

Visualisation des trajectoires

Le graphique 1 est un « chronogramme » : il représente, pour chaque semaine de cours, la part des participants qui étaient dans chacun des cinq états [7] présentés à la section précédente.

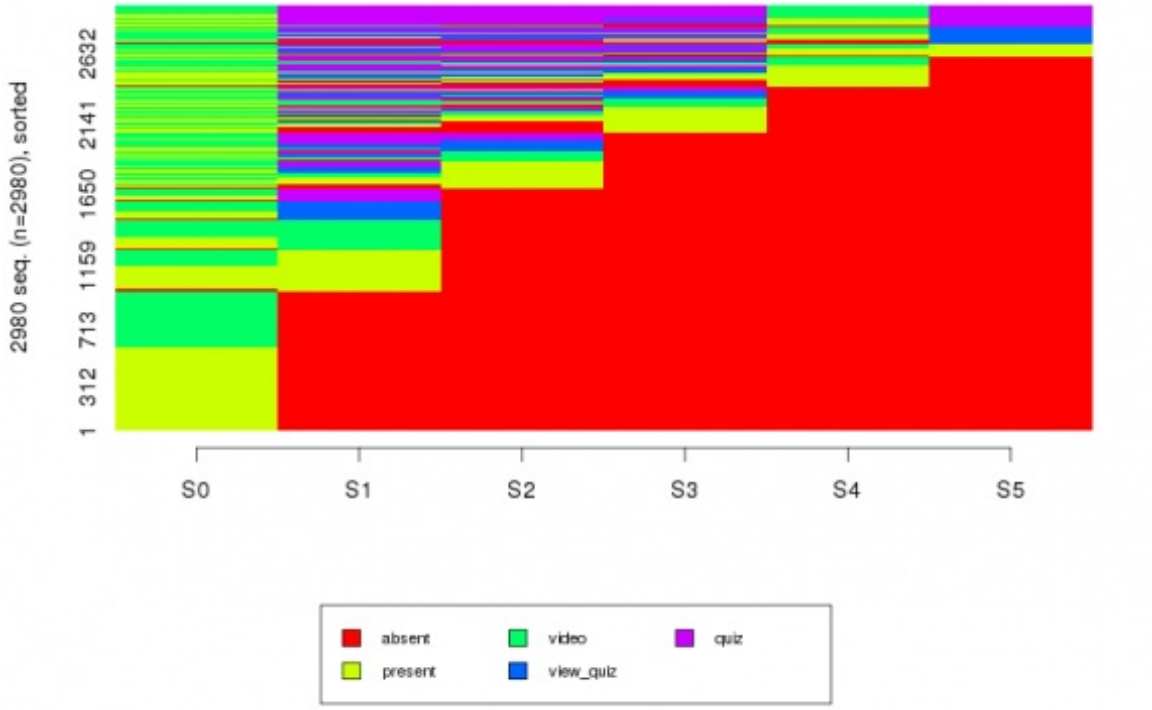


Graphique 1 : fréquence des différents états, semaine par semaine.

On remarque d'emblée que, de semaine en semaine, l'« absentéisme » (en rouge) augmente, jusqu'à concerner près de 90 % des utilisateurs en dernière semaine. De plus, de moins en moins de participants répondent aux quiz (21 % en semaine S1, 5 % en semaine S5). Par ailleurs, chaque semaine, une part non négligeable des actifs ne s'intéresse pas aux quiz, et se contente de regarder les vidéos ou de lire les textes (états *vidéo* et *présent*). À noter que les semaines S0, S4 et S5 sont un peu particulières : il n'y avait pas de quiz lors des semaines S0 et S4, et pas de vidéos lors de la semaine S5.

Le graphique 2 est construit à partir des mêmes données, mais les représente sous forme de « tapis » : les

trajectoires individuelles de tous les inscrits. Les trajectoires sont ordonnées selon leur état de fin de trajectoire.



Graphique 2 : « tapis » des trajectoires individuelles des participants.

Cette visualisation se lit de la façon suivante : chaque ligne est un utilisateur, que l'on suit de gauche à droite de la semaine S0 à la semaine S5. Ainsi, les lignes tout en bas du graphique, qui sont d'abord jaunes puis rouges, représentent des participants qui étaient présents lors de la semaine S0 (mais n'ont pas regardé les vidéos ni les quiz sur le site), et n'ont pas consulté les pages des semaines suivantes. Sur ce graphique nous avons ordonné les individus selon leur état à la semaine S5 ; ainsi, les trajectoires représentées tout en haut du graphique sont celles des participants qui ont répondu au quiz lors de la dernière semaine.

On remarque, ici encore, l'importance croissante des « absents », mais le fait d'observer les trajectoires individuelles permet cette fois de parler d'une forme de « décrochage » : un grand nombre d'utilisateurs ne reviennent jamais après leur première absence. Cependant, les profils des « décrocheurs » semblent assez divers : certains d'entre eux ont par exemple répondu aux quiz des semaines précédentes.

Les utilisateurs ayant passé le test final (en haut sur le graphique, les lignes dont la dernière case est violette) semblent au contraire présenter des trajectoires plus homogènes : la plupart d'entre eux a répondu aux quiz de toutes les semaines précédentes.

Une logique de l'investissement semble se dégager de ce graphique : le tri des trajectoires selon l'état de fin semble pertinent, et l'investissement semble suivre un processus cumulatif. En effet, les décrochages sont le plus souvent définitifs, et plus un apprenant participe activement au *Mooc*, plus il a de chances de le suivre jusqu'au bout.

Quatre modes de participation au Mooc

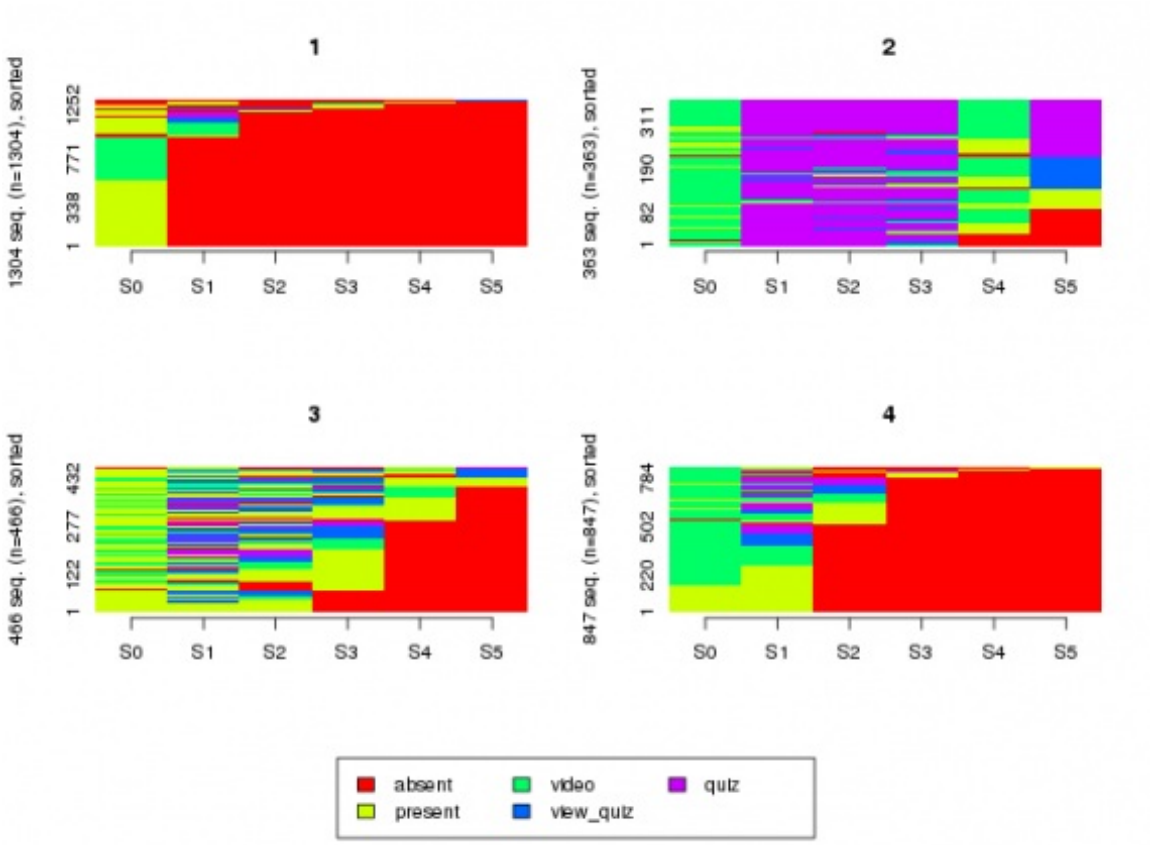
La classification automatique (*clustering*) nous permet ensuite de déterminer de manière statistique les différents groupes présents dans la population. Pour ce faire, il faut d'abord définir une « distance » entre

trajectoires. Sur cette définition, notre méthode s'écarte de celle de Kizilcec et al. (2013). En effet, outre les différences dans le codage des états, la distance qu'ils utilisent est une distance de *Hamming* : elle compte le nombre d'éléments différents entre deux trajectoires. Par exemple, on dira que deux trajectoires A-B-C-D-E et E-A-B-C-D ont entre elles une distance de 5, puisqu'à chacune des cinq périodes les états des deux trajectoires sont différents.

Pour prendre en compte la dimension temporelle des données de trajectoires, il est préférable d'utiliser une distance d'*optimal matching*. Cette distance est définie comme le nombre minimal d'opérations élémentaires (remplacement comme dans la distance de Hamming, mais aussi insertion et suppression d'états) qu'il faut pour rendre égales deux trajectoires. Ainsi, pour passer de la trajectoire E-A-B-C-D à la trajectoire A-B-C-D-E, il suffit de supprimer le E initial, et d'ajouter un E à la fin, soit deux opérations, et une distance de 2 [8]. Cette différence d'approche explique le fait que la distance de Hamming est insensible à une permutation des périodes, alors que la distance d'*optimal matching* ne l'est pas.

En pratique, la définition des distances (quelle que soit la méthode choisie) peut être enrichie par une pondération des différentes opérations : ainsi, si l'on considère que l'état « présent » est plus différent de l'état « absent » que de l'état « quiz », on peut donner un coût de 2 au remplacement d'un état « présent » par un « absent », et un coût de 1,5 au remplacement de « présent » par « quiz ». La définition de ces poids permet au chercheur d'insuffler à l'algorithme un peu de sa connaissance du phénomène étudié.

Une fois toutes les distances entre trajectoires calculées [9] on fait appel à un algorithme de classification automatique pour regrouper les individus en classes. Comme Kizilcec et al. (2013), nous faisons appel à l'algorithme de classification des *K-means*, dont les résultats nous portent à choisir un découpage en quatre classes. Le résultat de la classification est représenté dans le graphique 3, où l'on peut lire les « tapis » de trajectoires de chacune des classes.



Graphique 3 : « tapis » des trajectoires individuelles, pour les quatre classes d'utilisateurs.

Les deux classes les plus nombreuses sont les classes 1 et 4, qui comptent respectivement 1304 et 847

individus, et comprennent des trajectoires d'apprenants peu investis dans le *Mooc*. La classe 1 (parcours éclair) regroupe les participants qui ont décroché dès la semaine S1, et ont participé au plus à deux semaines de cours. On remarque d'ailleurs qu'environ la moitié d'entre eux n'a pas regardé les vidéos de la semaine S0, ce qui n'est pas souvent le cas dans les autres classes. Les individus de la classe 4 (parcours d'essai) ont, pour leur part, décroché à la semaine S2. On y compte plus de visionnages de vidéos et de réponses aux quiz que dans la classe 1, et presque tous ont eu une activité lors de la semaine S1 ; ils ont ensuite décroché, et soit n'ont consulté aucune des pages des semaines suivantes, soit ont eu une activité ponctuelle au cours d'une seule des semaines S2 à S4.

Les individus de la classe 3 (parcours soutenu, 466 observations) ont eu une activité plus suivie tout au long du *Mooc* : ils ont au maximum trois semaines d'absences, dont très peu d'absences aux semaines S0 à S4. Beaucoup d'entre eux sont codés comme simplement présents (en jaune), c'est-à-dire qu'ils ne regardent ni les vidéos ni les quiz. On manque malheureusement d'informations sur l'activité précise de ces individus : peut-être ont-ils téléchargé les vidéos, peut-être ont-ils simplement lu les textes, peut-être se sont-ils uniquement connectés aux pages pour accéder aux liens vers les activités (qui se déroulaient sur un site tiers)… Toujours est-il que seule une dizaine d'entre eux a répondu au quiz de la dernière semaine, et que l'on peut penser qu'ils n'étaient pas très intéressés par l'évaluation proposée dans le *Mooc*. On pourrait qualifier les apprenants de la classe 3 de « flâneurs », en ce qu'ils ont pioché dans l'offre du *Mooc* les éléments qui pouvaient les intéresser, mais se sont peu investis.

Enfin, la classe 2 (parcours complet) regroupe les 363 apprenants les plus actifs. Leurs trajectoires sont très homogènes : presque tous ont répondu aux quiz des semaines S1 à S3, et 140 d'entre eux ont répondu au quiz final (soit la grande majorité des réponses au quiz final). Le graphique 3 montre bien à quel point la classe 2 est distincte des autres, et fonctionne comme le noyau dur de la participation au *Mooc*.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Classe 1	1	1	2	2.214	3	15
Classe 2	1	8	12	13.37	17	62
Classe 3	1	2	4	4.91	6	22
Classe 4	1	2	3	3.929	5	17

Tableau 1 : répartition des nombres de jours d'activité sur le *Mooc* suivant les 4 clusters

Tableau 1 : répartition des nombres de jours d'activité sur le *Mooc* suivant les 4 clusters

La différence entre la classe 2 et les autres classes est également perceptible dans les statistiques de nombre de jours d'activité des apprenants (comptés comme le nombre de dates distinctes auxquelles un apprenant s'est connecté au *Mooc*), statistiques résumées dans le tableau 1. Les individus du groupe 2 se sont connectés en médiane 12 jours, contre 2 à 4 jours pour les autres classes. Les trois quarts de la classe 2 se sont connectés au moins 8 jours différents, alors que dans les autres classes les trois quarts des participants se sont connecté moins de 6 jours différents. Il est cependant intéressant de noter que les dates de connexions de toutes les classes sont réparties tout le long des 8 semaines de logs, à compter de la semaine S0 d'eFAN 2. Ainsi, nombreux sont les apprenants des classes 2 et 4 à s'être connectés pour la première fois plusieurs semaines après le début officiel du *Mooc*, ce qui illustre bien la diversité des façons de s'investir dans un cours en ligne de ce type.

Conclusion et perspectives

Nous avons cherché à présenter comment visualiser synthétiquement les flux de participants à un Mooc pour produire une analyse de données susceptible d'éclairer les pratiques d'apprentissage en ligne. Après avoir décrit des activités pour chaque semaine du cours (répondre au quiz, regarder des vidéos, etc.), un classement automatique par la méthode des *K-means*, sur des distances d'*optimal matching*, permet d'identifier 4 profils d'apprenants ayant des comportements cohérents. Nous distinguons ainsi différents modes d'interactions avec le cours que l'on peut interpréter en termes d'investissement des participants.

Parmi les 3 700 inscrits figurant dans les logs [10], on identifie environ 2100 apprenants dans 2 groupes d'apprenants (classe 1 et 4) les moins investis dans le *Mooc* qui sont restés au plus les trois premières semaines. Environ 460 participants (classe 3) ont eu une activité suivie. Beaucoup d'entre eux ont seulement consulté les textes sur la plateforme, mais peut-être ont-ils téléchargé les vidéos ou réalisé les activités qui se déroulaient sur un site tiers. Enfin, 363 apprenants (classe 2) sont les plus actifs et ont des trajectoires très homogènes : presque tous ont répondu aux quiz des premières semaines et regardé les vidéos. Presque 40 % d'entre eux ont répondu au quiz final.

L'originalité de notre étude par rapport à celle de Kizilcec et al. (2013) réside d'abord dans l'intelligibilité de la méthode : d'abord définir les activités réalisées pour chaque semaine de cours, ensuite calculer une distance entre les participants et enfin réaliser une classification automatique. Calculer une distance d'*optimal matching* entre les individus plutôt que celle de *Hamming* comme l'ont fait Kizilcec et al. (2013), permet de conserver dans le modèle la dimension temporelle et ordonnée des activités sur le Mooc. De ce fait, notre classification automatique constitue une synthèse correspondant mieux aux activités des participants que celle de Kizilcec.

La méthode de classification des *K-means* appliquée aux distances d'*optimal matching* permet de dégager des classes de comportements cohérents, et de bien distinguer les différents modes d'utilisation du *Mooc*. Les graphiques de représentations de trajectoires permettent une analyse détaillée des différents types de comportements. Il serait intéressant, pour poursuivre l'analyse, de croiser ce *clustering* avec des variables concernant les individus (âge, sexe, motivations pour suivre le *Mooc*…) ; en outre, l'exploitation d'autres éléments des logs indiquant des actions plus précises (par exemple de savoir qui télécharge les vidéos plutôt que de les regarder sur le site, leurs interventions sur le forum…) permettrait d'éclairer le comportement des apprenants plutôt que de simplement les qualifier de « présents ». Enfin, pour poursuivre ce travail il faudrait réaliser des entretiens avec des participants au Mooc appartenant à chaque classe dans une perspective compréhensive.

Remerciements

La méthode présentée dans ce texte est le produit d'une réflexion réalisée au laboratoire STEF (www.stef.ens-cachan.fr) avec Mattias Mano et Matthieu Cisel, qu'ils en soient ici remerciés.

Références

Gabadinho A., Ritschard G., Müller N. S., Studer M., « *Analyzing and Visualizing State Sequences in R with TraMineR*, *Journal of Statistical Software* », 2011, Volume 40, Issue 4.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). « Deconstructing disengagement : analyzing learner subpopulations in massive open online courses. » *In Proceedings of the Third International Conference on Learning Analytics and Knowledge*. p. 170–179. http ://dl.acm.org/citation.cfm ? id=2460330