Comment exploiter une masse de données pour la comparer à des traces qualitatives de l'usage d'un environnement numérique de formation ?

Philippe Daubias, Simon Flandin, Valérie Fontanieu

Philippe DAUBIAS,
Ingénieur de recherche
DUNES - IFÉ - ENS de Lyon (France)

Simon FLANDIN,
Collaborateur scientifique
CRAFT, Université de Genève (Suisse)

Valérie FONTANIEU
Ingénieure d'études
DUNES - IFÉ - ENS de Lyon (France)

RESUME

Contribution à la recherche en « e-formation » des adultes

Notre recherche s'intéresse aux usages d'un environnement numérique de formation[1] (ENF) en ligne par ses utilisateurs cibles (enseignants débutants). Elle s'inscrit du point de vue théorique dans un paradigme enactif de l'activité et de la cognition, considérée comme située et incarnée. Dans cette approche, les usages numériques sont conçus comme des cours d'action (Theureau & Jeffroy, 1994) (i) traçables (moyennant des méthodes de production et/ou de recueil de traces ad hoc) et (ii) donnant lieu à expérience conscientisable et exprimable (moyennant des techniques de remise en situation ad hoc). Dans la contribution résumée ici, nous nous focalisons sur la conceptualisation et l'exploitation du « traçage » des usages, en privilégiant l'angle méthodologique. Notre ambition est de contribuer au développement de méthodologies articulant, sur la base d'hypothèses théoriques partagées sur l'activité et l'apprentissage, des méthodes de recueil et de traitement de données quantitatives et qualitatives. Nous sommes convaincus qu'articuler la compréhension des usages formatifs au niveau local (analyse qualitative à grain fin) et au niveau global (analyse quantitative d'une masse de données) des utilisateurs-cibles d'un ENF (en l'occurrence des professionnels de l'enseignement débutants) est utile pour alimenter l'état des connaissances en « e-formation » des adultes et pour renseigner la conception d'ENF (amélioration de l'ENF existant ou conception de nouveaux ENF).

Une approche de learning analytics en « e-formation » des adultes

Notre recherche s'inscrit dans le domaine des learning analytics (LAK), en ce que nous établissons un scénario d'analyse de données massives, avant de l'instancier. Plus précisément, les études phénoménologiques qualitatives rassemblées dans la thèse de Flandin (2015), menée auprès de six enseignants débutants, ont notamment montré que des usages-types de l'ENF étaient identifiables et modélisables (Flandin, Auby & Ria, 2016a). L'un de ces usages-types en particulier est lié aux principes fondamentaux de conception de l'ENF et favorise les apprentissages visés par les concepteurs (Flandin, Auby & Ria, 2016b). Pour tester la validité globale de cette hypothèse locale, nous nous proposons d'observer les comportements sur un très large (voire exhaustif) échantillon d'utilisateurs de même profil, sur la même période que celle étudiée par les études mentionnées (août 2010 à juillet 2015). Les résultats obtenus quantitativement seront confrontés aux résultats qualitatifs antérieurs.

Méthode de traitement des données

Pour ce faire, nous effectuons une analyse quantitative sur des masses de données décrivant l'usage de l'ENF au travers des navigations sur la plateforme. Plus précisément, toutes les requêtes des utilisateurs, c'est-à-dire les demandes d'accès aux pages de l'ENF sont enregistrées par le serveur qui l'héberge, dans des fichiers de log d'accès. Ces logs contiennent ainsi toutes les requêtes de 257 414 utilisateurs sur la période de 5 ans considérée, c'est-à-dire la liste horodatée de tous les éléments (pages et

1 Plateforme de vidéoformation NéoPass@ction (Ria & Leblanc, 2011). En ligne : neo.ens-lyon.fr

leurs éléments constitutifs) demandés par les navigateurs de ces utilisateurs. Ces fichiers contiennent les adresses IP des utilisateurs, ce qui rend possible la reconstruction de leurs navigations. À quelques rares exceptions (Agosti et al., 2012, Vishwakarma et al., 2014), les logs d'accès aux sites web sont principalement destinés à des utilisations techniques (Silva, 2007) de métrologie pour établir des volumes de consultation globaux, ou de sécurité informatique pour analyser une tentative d'attaque ou une intrusion via une faille de sécurité sur un serveur. Les logs ont toutefois déjà été envisagés comme moyen d'identifier ou de modéliser des comportements d'utilisateurs (Suneetha & Krishnamoorthi, 2009 ; Kumar et al., 2017), mais sans possibilité de comparer les éventuels résultats obtenus à ceux d'une étude qualitative précise sur le même environnement. Dans notre cas, nous disposons d'une telle étude et moyennant les précautions méthodologiques exposées dans cette communication, il suffit de sélectionner le sous-ensemble de données correspondant à l'étude, pour envisager ce type de comparaison. Nous réordonnons par utilisateurs et filtrons les logs pour éliminer les données non utilisables ou sans intérêt pour l'étude. Nous reconstruisons ensuite les navigations en identifiant les accès d'un même utilisateur. De plus, nous utilisons des informations de la base de données de l'ENF (i) pour catégoriser les utilisateurs et extraire la sous-population visée et (ii) pour enrichir les informations contenues dans les logs d'accès. Cet « enrichissement de logs système » que l'on peut considérer comme une information technique « matérielle » ou de bas niveau, peut aussi s'appliquer à d'autres situations (Daubias, Fontanieu, & Khaneboubi, 2018) où des informations de plus haut niveau (descriptions, métadonnées) permettent d'éclairer les logs techniques, pour permettre une analyse avec l'apport conjoint des deux types d'informations.

Comme dans un très grand nombre d'expérimentations, les données massives collectées ou « traces » (Lund & Mille, 2009) ne sont qu'un aspect de la réalité car elles ne reflètent que ce qui se passe à l'interface du système informatique. Dans notre cas, les logs d'accès sont enregistrés au niveau du serveur et cette restriction s'amplifie, car on ne voit que les interactions du client avec le serveur, pas ce qui se produit sur le navigateur du client. Notre analyse ne permettra pas par exemple de savoir si l'utilisateur a visionné une vidéo, mais pourra juste donner l'indication de son téléchargement et de l'absence d'autres interactions avec le serveur pendant un laps de temps donné. Pour les parties de site où l'utilisateur peut naviguer sans que cela n'implique le téléchargement d'éléments (en cas d'utilisation d'AJAX/Javascript), il n'y a pas de trace de l'activité au niveau du serveur.
En plus des questions d'éthique que soulèvent les traitements de masses de données, nous sommes attentifs à la préservation de l'anonymat des utilisateurs de la plateforme (Barbaro & Zeller, 2006 ; Mivule, 2017). Les logs d'accès contiennent les adresses IP des clients, ce qui pourrait permettre d'identifier les utilisateurs réels (Reffay et al., 2012), c'est pourquoi, afin d'assurer l'anonymat des données sans perdre d'informations utiles à l'analyse (Reffay & Teutsch, 2007), nous avons (i) étiqueté de façon automatique chaque navigation, c'est-à-dire les requêtes successives ayant la même adresse IP, avec l'identifiant utilisateur_x où x est le numéro d'ordre d'apparition de l'adresse IP sans lien avec l'adresse IP elle-même, (ii) utilisé une base de géolocalisation, associant des plages d'adresse IP à des localisations géographiques pour caractériser les différents utilisateurs afin de maintenir la possibilité de faire émerger en cours d'analyse des spécificités géographiques, pouvant être le résultat d'une politique locale de l'institution par exemple, et (iii) supprimé les adresses IP, ce qui coupe totalement le lien à l'utilisateur réel. Le corpus de données anonymes obtenu se compose ainsi de 5 années de logs triés par navigation et enrichis d'informations sur la nature des ressources consultées.

Analyse statistique des données

Notre objectif est d'analyser les données quantitatives massives pour comparer les résultats obtenus avec ceux de l'étude de Flandin (2015). Nous ne guidons pas l'analyse de données massives en fonction des résultats précédemment obtenus pour laisser la possibilité au modèle statistique d'apporter un éclairage différent sur les données. Pour cela, nous utilisons des méthodes de classification automatique (ou clustering) avec différentes variables descriptives des navigations des utilisateurs, construites à partir des données enrichies, pour affiner l'analyse. Ceci permet de questionner les résultats obtenus qualitativement en validant, invalidant ou précisant les observations préalables.

Perspectives

Le but du projet est d'une part (i) de déterminer si les modèles produits majoritairement de manière qualitative, via six études de cas, ont (ou non) une validité plus générale à l'échelle de tous les utilisateurs de l'ENF de même profil, et d'autre part (ii) d'évaluer les principes technologiques mobilisés pour la conception amont et de contribuer à leur amélioration. Les perspectives ouvertes par l'avancée méthodologique que nous présentons dans cette contribution peuvent être déclinées selon trois phases successives :

1. la mise en œuvre du protocole automatique permettant de reconstruire des parcours de navigation exploitables à partir des données (massives) recueillies en continu sur l'utilisation de l'ENF ;
2. l'élaboration et la mise en œuvre d'un protocole de traitement statistique des résultats permettant d'obtenir des modèles "quantitatifs" de l'utilisation de l'ENF ;
3. la comparaison des modèles qualitatifs et quantitatifs obtenus, la validation/invalidation des principes de conception de l'ENF et la dérivation éventuelle de nouveaux principes de conception en e-formation des adultes.

Les résultats obtenus sur des données à grande échelle permettront d'alimenter le processus de conception continuée par l'étude des usages de l'ENF (Béguin & Rabardel, 2000 ; Flandin & Gaudin, 2014 ; Leblanc, 2012) de façon bien plus systématique que les résultats actuellement disponibles (obtenus via six études de cas).

Ils permettront plus largement de contribuer à la compréhension des usages autonomes et en ligne d'environnements dédiés à la formation professionnelle et à la stabilisation de principes de conception performants.

Mots clés : IHM ; User eXperience ; Learner eXperience ; Teacher eXperience ; EIAH ; Instructional Design ; Learning Analytics (LAK) ; Phénoménologie ; Formation professionnelle.

BIBLIOGRAPHIE

Agosti, M., Crivellari, F. & Di Nunzio, G. M. (2012). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery* 24.3 (2012): 663 - 696.

Barbaro, M. & Zeller Jr., T. (2006). A Face Is Exposed for AOL Searcher, No.4417749. *The New York Times*, 9 Aug 2006, p. C4.

Béguin, P. & Rabardel, P. (2000). Concevoir pour les activités instrumentées. *Revue d'Intelligence Artificielle, 14,* 1-2, 35-54.

Daubias P., Fontanieu V. & Khaneboubi M. (2018). Log is in the R : Une méthode pour analyser l'audience d'un site pédagogique. *Actes du colloque les usages du numérique en éducation : regards critiques (RUNED2018)*, Lyon, France, mars 2018.

Flandin, S. (2015). *Analyse de l'activité d'enseignants stagiaires du second degré en situation de vidéoformation autonome : Contribution à un programme de recherche technologique en formation.* Thèse de doctorat, Université Blaise Pascal, Clermont-Ferrand.

Flandin, S., Auby, M. & Ria, L. (2016a). Étude de l'utilisation d'un environnement numérique de formation: méthode de remise en situation à l'aide de traces numériques de l'activité. *Activités*, *13*(13-2).

Flandin, S., Auby, M. & Ria, L. (2016b). À quoi s'intéressent les enseignants dans les exemples en formation ? Étude de l'utilisation par des stagiaires de ressources basées sur la vidéo. *Recherches en Éducation, 27,* 118-133.

Flandin, S., & Gaudin, C. (2014). Conception continuée dans l'usage en vidéoformation des enseignants. *Actes du 3ème Colloque International de Didactique Professionnelle,* Caen, 28-29 octobre.

Kumar A., Ahirwar V. & Singh, R. K. (2017). A Study on Prediction of User Behavior Based on Web Server Log Files in Web Usage Mining. *International Journal of Engineering and Computer Science*, *6*(2), pp. 20233-20236.

Leblanc, S. (2012). *Conception d'environnements vidéo numériques de formation. Développement d'un programme de recherche technologique centré sur l'activité dans le domaine de l'éducation.* Note de synthèse pour l'Habilitation à Diriger des Recherches non publiée. Université de Montpellier 3, Montpellier.

Lund K. & Mille A. (2009). Traces, traces d'interactions, traces d'apprentissages, définitions, modèles informatiques, structurations, traitements et usages. In Marty J.-C. et Mille A. (Eds.), *Analyse de traces et Personnalisation des EIAH, Traité Informatique et Systèmes d'Information* (pp. 21-56). Lavoisier-Hermes.

Mivule, K. (2017). Web Search Query Privacy, an End-User Perspective. *Journal of Information Security*, 8, 56-74

Reffay, C. & Teutsch, P. (2007). Anonymisation de corpus réutilisables : masquer l'identité sans altérer l'analyse des interactions. *Actes de la conférence EIAH2007*, Lausanne, Suisse, juin 2007.

Reffay C., Blondel F.M. & Giguet E. (2012). Stratégies pour l'anonymisation systématique d'un corpus d'interactions plurilingues. *Actes de la conférence IC2012*, Grenoble, France, juin 2012.

Ria, L. & Leblanc, S. (2011). Conception de la plateforme de formation Néopass@ction à partir d'un observatoire de l'activité des enseignants débutants : enjeux et processus. *Activités*, *8*(8-2).

Silva, S. (2007). *Web Server Administration*, ISBN-13: 978-1-4239-0323-9

Suneetha, K.R & Krishnamoorthi, R. (2009). Identifying User Behavior by Analyzing Web Server Access Log File, *IJCSNS International Journal of Computer Science and Network Security*, VOL. 9 No.4, April 2009

Theureau, J. & Jeffroy, F. (1994). *Ergonomie des situations informatisées. La conception centrée sur les cours d'action des utilisateurs*. Toulouse : Octarès.

Vishwakarma, Amit & Singh, Kedar Nath (2014) A Survey on Web Log Mining Pattern Discovery, *(IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 5 (6), 2014, 7022-7031