Log is in the R** : Une méthode pour analyser l'audience d'un site pédagogique

Philippe Daubias*[1], Valérie Fontanieu*[1], and Mehdi Khaneboubi[2]

[1]IFE - ENS de Lyon (IFE) – ENS de Lyon – France
[2]Sciences Techniques Éducation Formation (STEF) – École normale supérieure [ENS] - Cachan, INRP, École normale supérieure (ENS) - Cachan – France

Résumé

** Titre clin d'oeil à (Tabard et al. 2006), qui enregistre l'activité utilisateur au niveau client et non au niveau serveur.
Cette recherche s'inscrit dans le domaine des learning analytics, où l'on analyse des données massives liées à l'éducation. Les données massives ou " traces " (Lund & Mille 2009) collectées automatiquement retracent ce qu'il s'est produit au sein d'un système informatique et présentent un seul aspect d'une activité documentaire plus complexe. Comment cerner l'audience d'un site web à partir de données techniques ? Quelles informations sur l'activité des utilisateurs peut-on établir ? Dans cette communication, à vocation essentiellement méthodologique, nous présenterons de façon détaillée la collecte, le filtrage et l'enrichissement de logs système ainsi que le mode de diffusion d'un questionnaire que nous avons proposé en parallèle aux utilisateurs du site pour caractériser des usages de ressources pédagogiques. Nous évoquerons ce que nos données ne permettent pas d'observer et les biais possibles dans leur production et leur analyse.

Comment caractériser l'activité enseignante dans la conception, la recherche, la sélection, la modification et la recomposition des ressources présentées aux élèves ? Dans le cadre du projet ReVEA (Ressources vivantes pour l'enseignement et l'apprentissage), nous avons effectué une analyse quantitative sur les journaux de connexions du site Planet-Terre. Ces données ont été enrichies par des métadonnées portant sur les documents consultés puis analysées avec le logiciel R (R. Core Team 2014). Le site Planet-Terre, hébergé à l'ENS de Lyon pour la DGESCO, diffuse des documents (articles, images, vidéos, etc.) traitant de géologie, destinés essentiellement à des enseignants et des étudiants.

Tous les accès aux différents éléments du site Planet-Terre (pages web, mais aussi éléments les composants) sont enregistrés, comme pour l'immense majorité des sites web, par un serveur dans des fichiers appelés " logs ". Ces fichiers de log d'accès contiennent la liste horodatée de toutes les pages et leurs éléments constitutifs demandés par les navigateurs des utilisateurs (nommée requête " clients "). Ces données sont principalement destinées à des utilisations techniques, et sont notamment utilisées par la recherche en sécurité informatique. Cependant, des logs d'accès ont déjà servi par exemple à analyser les requêtes sur des bibliothèques en ligne et des moteurs de recherche (Agosti et al. 2012), mais jamais à notre connaissance pour une analyse d'usage d'un site pour l'éducation. Comme c'est le cas pour d'autres sites de ressources destinés aux enseignants, les contenus du site Planet-Terre sont

*Intervenant

décrits conformément au standard LOM-Fr. Le standard international LOM (Learning Object Metadata), décliné pour la France en LOM-Fr sert à décrire des objets d'apprentissage, c'est-à-dire " toute entité numérique ou non qui peut être utilisée, réutilisée ou référencée lors d'une formation dispensée à partir d'un support technologique ". Nous avons utilisé ces métadonnées pour enrichir les informations contenues dans les logs.

Par ailleurs, les logs d'accès contiennent les adresses IP des clients. Ces informations sont " indirectement nominatives ", c'est-à-dire qu'elles ne permettent pas directement d'identifier l'utilisateur réel, mais Reffay et al. (2012) ont montré qu'il ne fallait pas minimiser la possibilité de lever cet anonymat. Pour assurer une bonne anonymisation et garantir une exploitation sans risques de ces informations, nous avons supprimé l'adresse IP des données, sans perte d'informations. L'adresse IP permet d'une part de regrouper les requêtes par utilisateurs (ou tout du moins par points d'accès à internet), mais peut également être utilisée pour déterminer la localisation géographique de l'utilisateur. Ces deux informations étant potentiellement utiles à l'analyse, nous avons étiqueté les requêtes correspondantes à une même adresse IP avec un même identifiant et utilisé une base de géolocalisation, associant des plages d'adresse IP à des localisations géographiques pour caractériser les différents utilisateurs. Il est ainsi impossible de remonter à l'adresse IP initiale sans pour autant perdre les informations de navigation et de localisation géographique de l'utilisateur. Ainsi nous avons produit des lots de logs enrichis, pour des périodes consécutives de trois mois allant de septembre 2015 à août 2017 correspondant au rythme scolaire trimestriel.

Tabard A., Roussel N. & Ledontal C. (2006). *All you need is log.* WWW 2006 Workshop on Logging Traces of Web Activity : The Mechanics of Data Collection, May 2006, Edinburg, United Kingdom.

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

Maristella Agosti, Franco Crivellari & Giorgio Maria Di Nunzio (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), 663-696.

R. Core Team (2014). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
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, juin 2012.

Mots-Clés: Learning Analytics (LAK), Enseignement secondaire, Géologie, Formation professionnelle, User eXperience, Teacher eXperience, Big data in education