**Titre :** Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning

**Auteurs :** Hendrik Drachsler, Toine Bogers, Riina Vuorikari, Katrien Verbertd, Erik Duvald,

Nikos Manouselise, Guenter Behamf, Stephanie Lindstaedtf, Hermann Sternf, Martin Friedrichh, Martin Wolpersh

**Date de publication :** 2010

**Objet** : Discussion du manque de jeux de données standard pour systèmes de recommandation des systèmes d’e-learning, définission de guidelines et proposition d’un standard.

**Contenu**

Constat de la difficulté de comparer et d'évaluer les différentes solutions de personnalisation de contenu appliquées au e-learning en raison de l'absence de véritable jeu de données de référence. La disponibilité de ce type de ressources favoriserait la répétabilité des tests et expériences mis en place pour tester les différentes méthodes de recommandation et permettrait de tirer plus facilement des conclusions quant à la validité des différentes démarche mises en œuvre. La disparité des différents contextes d'apprentissage est mise en cause, notamment entre d'une part les solutions s'adressant aux élèves intégrés à un cursus formel comme une université par exemple, et les élèves apprenant en dehors du cadre d'un tel cursus. Ces derniers ne sont pas contraint à un corpus d'éléments d'apprentissage à assimiler et utilisent d'avantage des sources d'information plus variées comme des blogs, agrégateurs d'articles et plate-formes de partage de documents. Il est noté qu'aucun jeu de données suffisamment important n'a été mis à disposition publiquement pour aucun de ces types d'apprentissage.

Des lignes de conduite sont définies pour aider les chercheurs à sélectionner de bon jeux de données. Celles-ci sont premièrement la nécessité de sélectionner des données représentatives des véritables conditions d'utilisation, en terme de nombre d'utilisateurs de différents groupes au sein d'une même plate-forme et de types d'items considérés. Deuxièmement, il est nécessaire de sélectionner des jeu de données concernant suffisamment d'utilisateurs. Enfin, il est dit que les jeux de données doivent être comparables les uns aux autres, et donc partager une structure similaire.

L'article recommande pour ce faire de varier les sites utilisés comme sources de données afin de toucher des populations plus variées démographiquement.

Il est aussi proposer de tirer des échantillons égaux de données d'utilisateurs parmis des groupes établis en fonction de la taille des comptes de ces utilisateurs en nombre d'abonnés afin d'assurer la diversité des profils considérés. Une liste de sites acceptant de partager leurs données publiquement est aussi proposée. L'article propose également de ne pas prendre du tout en compte les utilisateurs les plus actifs pour éviter de considérer les données liées au spam. Différentes solutions sont ensuite proposées pour éliminer le bruit. Enfin, l'accent est mis sur la nécessité de partager dans des formats standards les données expérimentales ainsi récoltées.

Il est ensuite discuté de protection de la vie privée, et un format standard pour partager les données issus de plate-formes de bookmarking social est proposé.