

Vers la simulation et la détection des changements des données évolutives d'usage du Web

Alzennyr Da Silva^{*1}, Yves Lechevallier*, Francisco De Carvalho**

* Projet AxIS, INRIA Paris-Rocquencourt

Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay – France

{Alzennyr.Da_Silva, Yves.Lechevallier}@inria.fr

** CIN/UFPE, Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil

fatc@cin.ufpe.br

Résumé. Dans le domaine des flux des données, la prise en compte du temps s'avère nécessaire pour l'analyse de ces données car leur distribution sous-jacente peut changer au cours du temps. Un exemple typique concerne les modèles des profils de navigation des internautes. Notre objectif est d'analyser l'évolution de ces profils, celle-ci peut être liée au changement d'effectifs ou aux déplacement de clusters au cours du temps. Afin d'analyser la validité de notre approche, nous mettons en place une méthodologie pour la simulation des données d'usage à partir de laquelle il est possible de contrôler l'occurrence des changements.

1 Introduction

La fouille de données d'usage du Web (*Web Usage Mining*, WUM) désigne l'ensemble de techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web (Cooley et al. (1999); Spiliopoulou (1999)).

De manière contradictoire à la quantité colossale des données mises en ligne sur Internet, l'une des difficultés la plus importante liées à la fouille d'usage du Web est la pénurie (voir inexistence) de *benchmarks* de données d'usage du Web pour l'application et la comparaison de différentes techniques d'analyse. Ceci est dû au fait que les données d'usage contiennent des informations privées. Pour cela, nous proposons dans cet article une méthodologie pour la génération de données artificielles d'usage du Web sous la forme de tableau de contingence *navigations* \times *catégories de pages*. Notre principale motivation est la possibilité de mesurer l'efficacité de notre approche de détection de changements sur un ensemble de données contenant des changements de comportements pré-établis et sur lesquels nous avons un contrôle total. Notre proposition présente un algorithme de création de données artificielles ainsi que la simulation de changements liés à l'effectif et au déplacement des classes artificielles. Enfin, nous validons notre approche sur trois études de cas de différentes complexités (cf. figure 1).

Les résultats ici présentés sont la suite des travaux déjà exposés dans les deux dernières conférences EGC (cf. Da Silva et Lechevallier (2008) et Da Silva et al. (2007)).

¹L'auteur remercie la CAPES-Brésil pour son soutien à ce travail de recherche.