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)).

[1] L'auter remercie la CAPES-Brésil pour sont soutien à ce travail de recherche.

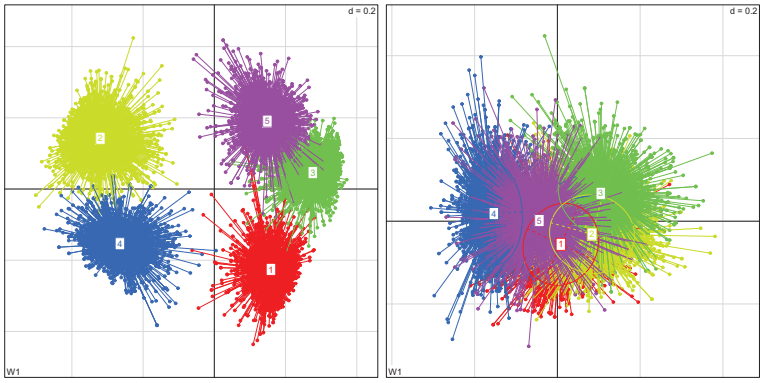

FIG. 1 – *Cinq classes artificielles : bien separées (à gauche) et recouvrantes (à droite).*

Références

Cooley, R., B. Mobasher, et J. Srivastava (1999). Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems 1*(1), 5–32.

Da Silva, A. et Y. Lechevallier (2008). Stratégies de classification non supervisée sur fenêtres superposées : application aux données d'usage du web. In *Actes des 8ème journées Extraction et Gestion des Connaissances (EGC 2008), Revue des Nouvelles Technologies de l'Information (RNTI)*, Volume I, pp. 219–220. cépaduès.

Da Silva, A., Y. Lechevallier, F. Rossi, et F. De Carvalho (2007). Construction et analyse de résumés de données évolutives : application aux données d'usage du web. In *Actes des 8ème journées Extraction et Gestion des Connaissances (EGC 2007), Revue des Nouvelles Technologies de l'Information (RNTI)*, pp. 539–544. cépaduès.

Spiliopoulou, M. (1999). Data mining for the web. *Workshop on Machine Learning in User Modelling of the ACAI99*, 588–589.

Summary

In the data stream domain, taking into account the time factor has become a necessity since the subjacent distribution of the data can evolve over time. A typical example concerns the Web surfer usage profiles. Our aim is to analyze the evolution of these profiles which can be related to the change in the cluster elements or to the displacement of clusters over time. In order to validate the changing indicators, we set up a methodology to simulate usage data which makes it possible to control the occurrence of the changes.