Approche pour le suivi de l'évolution des données d'usage du Web : application sur un jeu de données en *marketing*

Alzennyr Da Silva, Yves Lechevallier

Projet AxIS, INRIA Paris-Rocquencourt
Domaine de Voluceau, B.P. 105, 78153 Le Chesnay – France
{Alzennyr.Da_Silva@inria.fr,Yves.Lechevallier@inria.fr}

Résumé Dans la fouille des flux des données d'usage du Web, la dimension temporelle joue un rôle très important car les comportements des internautes peuvent changer au cours du temps. Dans cet article, nous présentons l'application d'une approche de classification automatique basée sur des fenêtres sautantes pour la détection et le suivi de changements sur un jeu de données en *marketing*. Cette approche combine les cartes auto organisatrices de Kohonen et la méthode de Ward pour la découverte automatique du nombre de clusters de comportement ainsi que deux indices de validation basés sur l'extension pour la détection des changements au cours du temps.
Keywords : Classification automatique, données évolutives, fouille d'usage du Web (Web Usage Mining, WUM).

1 Introduction

La fouille de données d'usage du Web (*Web Usage Mining*, en anglais) désigne l'ensemble des techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web [1, 8]. Dans la conférence SLDS 2009 [1], un jeu de données concernant le suivi des achats d'un panel de consommateurs a été diffusé dans le cadre d'un concours ouvert aux jeunes chercheurs. L'objectif de cet article est de décrire l'analyse des résultats obtenus à partir de l'application de notre approche de détection et de suivi des changements [2, 3] sur ce jeu de données.

2 Description du jeu de données

Le jeu de données en question concerne le suivi des achats de 10 068 clients pendant 14 mois (du 09 juillet 2007 jusqu'au 08 septembre 2008) sur 2 marchés de biens de consommation. Chaque marché commercialise 3 marques de produits. Les données ont été fournies dans un fichier de 3 745 296 lignes contenant les champs décrits dans le tableau 2. Dans ce fichier, tous les combinaisons *date* x *marché* x *marque* ont été présentés, même en cas d'absence d'achat. Pour l'application de notre méthode, nous avons défini un tableau croisé *client* x *marque* ordonné selon la date et l'identification client (cf. tableau 3). Ce tableau contient un total de 262 215 lignes.

[1] Symposium Apprentissage et Science des Données 2009, `www.ceremade.dauphine.fr/SLDS2009`