

Approche connexionniste pour l'analyse des données issues d'usage d'Internet : Classification et Visualisation

Khalid Benabdeslem* et Younès Bennani**

*Université Lyon1/LIESP, 8 Boulevard Niels Bohr, 69622 Villeurbanne, France
kbenabde@bat710.univ-lyon1.fr

**Université Paris13/LIPN, 99 Avenue J-B. Clément, 93430 Villetaneuse, France
younes.bennani@lipn.univ-paris13.fr

Résumé. Dans ce papier, nous présentons une chaîne complète de fouille de données comportementales issues des navigations des clients de sites web commerciaux. Nous présentons plus particulièrement, le développement d'une approche d'apprentissage non supervisé pour ce type de données stockées sous forme de traces de navigation dans des fichiers Log.

La première partie de cette étude concerne le problème du codage des données qui seront ensuite utilisées pour l'analyse. En effet, actuellement, les sites web sont dynamiques et les pages ne peuvent pas être caractérisées par des variables fixes comme : la hiérarchie des adresses URL, le contenu, etc. Elles sont représentées par des identificateurs numériques n'ayant aucun « sens » et qui servent comme adresses de récupération des informations dans des bases de données pour remplir les contenus des pages. Pour cette raison, nous proposons une nouvelle méthode de codage de sessions à partir du fichier Log. Cette technique consiste à caractériser une page donnée par un vecteur de poids d'importance de passage, i.e. par ses poids de précédence et de successions relatives à toutes les autres pages qui apparaissent dans le fichier Log.

Dans la deuxième partie de ce travail, nous analysons les propriétés des cartes topologiques de Kohonen et nous proposons une version adaptée aux données comportementales. Cette étape nous permet (1) de construire une cartographie du site web tel qu'il est aperçu par les clients (2) de regrouper les pages pour un objectif de codage de sessions (3) et de projeter les interactions des clients du fichier Log sur la cartographie sous forme de trajectoires symbolisant leurs comportements.

1 Introduction

Les sites Web représentent actuellement une véritable source de production de grands volumes d'informations. Cependant, ce gisement d'information ne représente pas une évidence de compréhension des utilisateurs qui se retrouvent généralement perdus devant de telles quantités d'informations, Zeboulon et al. (2001). Une technique de e-Mining est donc nécessaire pour comprendre les interactions des utilisateurs pour répondre aux mieux à leurs besoins prioritaires. Le e-Mining est une chaîne complète de fouille de données qui permet d'analyser des formes basées sur des interactions pour extraire des connaissances sur les comportements des utilisateurs dans les sites Web, Cadez et al. (2000). Dans ce contexte, nous définissons une session d'utilisateur comme une séquence temporelle des pages qui l'ont intéressé durant son parcours dans le site.