Web Content Data Mining : la classification croisée pour l'analyse textuelle d'un site Web

Malika Charrad*, Yves Lechevallier**
Gilbert Saporta***, Mohamed Ben Ahmed*
*Laboratoire RIADI, Ecole Nationale des Sciences de l'Informatique, Tunis
malika.charrad@riadi.rnu.tn
mohamed.benahmed@riadi.rnu.tn
**INRIA-Rocquencourt, 78153 Le Chesnay cedex
yves.lechevallier@inria.fr
***CNAM, 292 rue Saint-Martin, 75141 Paris cedex 03
saporta@cnam.fr

Résumé. Notre objectif dans cet article est l'analyse textuelle d'un site Web indépendamment de son usage. Notre approche se déroule en trois étapes. La première étape consiste au typage des pages afin de distinguer les pages de navigation ou pages « auxiliaires » des pages de contenu. La deuxième étape consiste au prétraitement du contenu des pages de contenu afin de représenter chaque page par un vecteur de descripteurs. La dernière étape consiste au block clustering ou la classification simultanée des lignes et des colonnes de la matrice croisant les pages aux descripteurs de pages afin de découvrir des biclasses de pages et de descripteurs. L'application de cette approche au site de tourisme de Metz prouve son efficacité et son applicabilité. L'ensemble de classes de pages groupés en thèmes facilitera l'analyse ultérieure de l'usage du site.

1 Introduction

Le Web représente aujourd'hui la principale source d'information. Ce gisement contenant une grande quantité de données non-structurées, distribuées et multi-medias a besoin d'être maintenu, filtré et organisé pour permettre un usage efficace. Cette tâche s'avère difficile à réaliser avec la large distribution, l'ouverture et la forte dynamicité du Web. Par conséquent, plusieurs travaux de recherche ont tenté d'analyser le contenu des sites Web et comprendre le comportement des utilisateurs de ces sites. L'approche que nous proposons dans cet article se situe dans ce cadre. Notre objectif est d'analyser un site Web en se basant sur le contenu et indépendamment de l'usage. En d'autres termes, nous cherchons à réduire la quantité d'information contenue dans le site Web en un groupe de thèmes qui pourraient susciter l'intérêt des internautes. Il sera par la suite possible d'analyser le comportement des utilisateurs vis-à-vis de ces thèmes.