

WCUM pour l'analyse d'un site web

Malika Charrad*,*** Yves Lechevallier**
Mohamed Ben Ahmed*, Gilbert Saporta***

*Ecole Nationale des Sciences de l'Informatique
malika.charrad@riadi.rnu.tn,
**INRIA-Rocquencourt
yves.lechevallier@inria.fr
***Conservatoire National des Arts et Métiers
gilbert.saporta@cnam.fr

Résumé. Dans ce papier, nous proposons une approche WCUM (Web Content and Usage based Approach) permettant de relier l'analyse du contenu d'un site Web à l'analyse de l'usage afin de mieux comprendre les comportements de navigation sur le site. L'apport de ce travail réside d'une part dans la proposition d'une approche reliant l'analyse du contenu à l'analyse de l'usage et d'autre part dans l'extension de l'application des méthodes de block clustering, appliquées généralement en bioinformatique, au contexte Web mining afin de profiter de leur pouvoir classificatoire dans la découverte de biclasses homogènes à partir d'une partition des instances et une partition des attributs recherchées simultanément.

1 Introduction

La caractérisation des internautes fréquentant un site Web est un problème incontournable pour assister l'internaute et prédire son comportement. Ces considérations ont motivé d'importants efforts dans l'analyse des traces des internautes sur les sites Web. D'autres efforts ont été concentrés sur l'analyse du contenu des pages Web. Sachant que le comportement des utilisateurs sur un site web dépend fortement du contenu des pages du site et inversement le contenu du site devrait répondre aux attentes des usagers du site, nous proposons de faire la liaison entre le contenu et l'usage d'un site web. Notre idée est d'exploiter les différentes informations relatives au contenu d'un site Web et de son usage en vue de l'analyser. Le point de départ de cette approche est le contenu textuel du site et les fichiers logs contenant les traces des utilisateurs. L'approche WCUM relie l'analyse du contenu à l'analyse de l'usage d'un site Web. Elle se déroule en deux principales étapes. La première consiste à l'analyse textuelle d'un site Web en appliquant un algorithme de block clustering à la matrice croisant les descripteurs aux pages afin de résumer le contenu du site en un ensemble de thèmes. La seconde étape consiste à introduire ces thèmes dans l'analyse de l'usage du site. L'application de cette approche nécessite d'une part l'aspiration du site afin de transformer ses pages en fichiers texte, et d'autre part la collecte des fichiers Logs contenant la trace des utilisateurs sur le site.