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.

Références

Benedek, A. et B. Trousse (2003). Adaptation of self-organizing maps for case indexing. *In 27th Annual Conference of the Gesellschaft fur Klassifikation, Germany*, 31–45.

Charrad, M. (2005). *Techniques d'extraction des connaissances appliquées aux données du Web*. Mémoire de mastère, Ecole Nationale des Sciences de l'Informatique de Tunis.

Charrad, M., Y. Lechevallier, M. B. Ahmed, et G. Saporta (2009). Block clustering for web pages categorization. *Lecture Notes in Computer Science Series Springer 5788/2009*, 260–267.

Charrad, M., Y. Lechevallier, G. Saporta, et M. B. Ahmed (2008). Web content data mining : la classification croisée pour l'analyse textuelle d'un site web. *Revue des Nouvelles Technologies de l'Information (Cépaduès) 1*, 43–54.

Diday, E. (1971). Une nouvelle méthode de classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de Statistique Appliquée 19 2*, 19–33.

Fu, Y., K. Sandhu, et M. Shih (2000). A generalization-based approach to clustering of web usage sessions. *In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, Springer*, 21–38.

Govaert, G. (1983). *Classification croisée*. Thèse de doctorat, Université Paris 6.

Kimball, R. et R. Merz (2000). Le data webhouse : Analyser des comportements clients sur le web. *Editions Eyrolles, Paris*.

Landauer, T. et S. Dumais (1997). A solution to platoŠs problem : the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review 104*, 211Ű240.

Lechevallier, Y., D. Tonasa, B. Trousse, et R. Verde (2003). Classification automatique : Applications au web mining. *In Yadolah Dodge and Giuseppe Melfi, editor, Méthodes et Perspectives en Classification, Presse Académiques Neuchâtel*, 157–160.

Marascu, A. et F. Masseglia (2006). Extraction de motifs séquentiels dans les flots de données d'usage du web. *Extraction et Gestion des Connaissances (EGCŠ06), Lille*, 627–638.

Srivastava, J., R. Cooley, M. Deshpande, et P.-N. Tan (2000). Web usage mining : Discovery and applications of usage patterns from web data. *SIGKKD Explorationsy*, 12–23.

Tanasa, D. (2005). *Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support*. Thèse de doctorat, Université de Nice Sophia Antipolis.

Summary

The Web Content and Usage based (WCUM) Approach proposed in this paper deals with tha analysis of both content and usage of the web site to better understand the behaviour of web site users. Our main contribution consists in associating the content analysis with usage analysis and adaptating block clustering algorithms, traditionally used in bioinformatics, to web mining problems in order to discover homogeneous blocs of instances and attributes.