Web Usage Mining : extraction de périodes denses à partir des logs

Florent Masseglia *, Pascal Poncelet**, Maguelonne Teisseire***, Alice Marascu *

* INRIA Sophia Antipolis, 2004 route des Lucioles - BP 93, 06902 Sophia Antipolis, France
{Alice.Marascu,Florent.Masseglia}@sophia.inria.fr
**EMA-LGI2P/Site EERIE, Parc Scientifique Georges Besse, 30035 Nîmes Cedex, France
{Pascal.Poncelet}@ema.fr
***LIRMM UMR CNRS 5506, 161 Rue Ada, 34392 Montpellier cedex 5 - France
{teisseire}@lirmm.fr

Résumé. Les techniques de Web Usage Mining existantes sont actuellement basées sur un découpage des données arbitraire (*e.g.* "un log par mois") ou guidé par des résultats supposés (*e.g.* "quels sont les comportements des clients pour la période des achats de Noël ? "). Ces approches souffrent des deux problèmes suivants. D'une part, elles dépendent de cette organisation arbitraire des données au cours du temps. D'autre part elles ne peuvent pas extraire automatiquement des "pics saisonniers" dans les données stockées. Nous proposons d'exploiter les données pour découvrir de manière automatique des périodes "denses" de comportements. Une période sera considérée comme "dense" si elle contient au moins un motif séquentiel fréquent pour l'ensemble des utilisateurs qui étaient connectés sur le site à cette période.

1 Introduction

L'analyse du comportement des utilisateurs d'un site Web, également connue sous le nom de Web Usage Mining, est un domaine de recherche qui consiste à adapter des techniques de fouille de données sur les enregistrements contenus dans les fichiers logs d'accès Web (ou fichiers "access log") afin d'en extraire des relations entre les différentes données stockées Cooley et al. (1999), Masseglia et al. (2003), Mobasher et al. (2002), Spiliopoulou et al. (1999). Ces derniers regroupent des informations sur l'adresse IP de la machine, l'URL demandée, la date, et d'autres renseignements concernant la navigation de l'utilisateur. Parmi les méthodes développées, celles qui consistent à extraire des motifs séquentiels Agrawal et Srikant (1995) s'adaptent particulièrement bien au cas des logs mais dépendent du découpage qui est fait des données. Ce découpage provient soit d'une décision arbitraire de produire un log tous les *x* jours (*e.g.* un log par mois), soit d'un désir de trouver des comportements particuliers (*e.g.* les comportements des internautes du 15 novembre au 23 décembre lors des achats de Noël). Pour comprendre l'enjeu de ces travaux, prenons l'exemple d'étudiants connectés lors d'une séance de TP. Imaginons que ces étudiants soient répartis en 2 groupes. Le groupe 1 était en TP le lundi 31 janvier. Le groupe 2 en revanche était en TP le mardi 1^{er} février. Chacun de ces