

Construction et analyse de résumés de données évolutives : application aux données d'usage du Web

Alzenny Da Silva*, Yves Lechevallier*,
Fabrice Rossi*, Francisco De Carvalho**

* INRIA – Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay cedex – France
{Alzennyr.Da_Silva, Yves.Lechevallier, Fabrice.Rossi}@inria.fr
**Centro de Informatica - CIN/UFPE
Caixa Postal 7851, CEP 50732-970, Recife (PE) – Brésil
fatc@cin.ufpe.br

Résumé. La manière dont une visite est réalisée sur un site Web peut changer en raison de modifications liées à la structure et au contenu du site lui-même, ou bien en raison du changement de comportement de certains groupes d'utilisateurs ou de l'émergence de nouveaux comportements. Ainsi, les modèles associés à ces comportements dans la fouille d'usage du Web doivent être mis à jour continuellement afin de mieux refléter le comportement actuel des internautes. Une solution, proposée dans cet article, est de mettre à jour ces modèles à l'aide des résumés obtenus par une approche évolutive des méthodes de classification.

1 Introduction

Les profils d'accès à un site Web peuvent être influencés par certains paramètres de nature temporelle, comme par exemple : l'heure et le jour de la semaine, des événements saisonniers, des événements externes dans le monde (guerres, crises économiques), etc. Dans ce contexte, la plupart des méthodes consacrées à la fouille de données d'usage du Web (Web Usage Mining) (Cooley et al., 1999) prennent en compte dans leur analyse toute la période qui enregistre les traces d'usage : les résultats obtenus sont donc naturellement ceux qui prédominent sur la totalité de la période. Ainsi, certains types de comportements, qui ont lieu pendant de courtes sous-périodes ne sont pas pris en compte, et restent donc ignorés par les méthodes classiques. Il est pourtant important d'étudier ces comportements et donc de réaliser une analyse portant sur des sous-périodes significatives. Le volume des données considérées étant très élevé, il est en outre important de recourir à des résumés pour représenter les profils considérés.

L'analyse de l'usage a commencé relativement récemment à tenir compte de la dépendance temporelle des profils de comportement. Dans (Roddick et Spiliopoulou, 2002), les auteurs examinent les travaux antérieurs. Ils résumant les solutions proposées et les problèmes en suspens dans l'exploitation de données temporelles, au travers d'une discussion sur les règles temporelles et leur sémantique, mais aussi par l'investigation de la convergence entre la fouille de données et la sémantique temporelle. Tout récemment, dans (Laxman et Sastry, 2006) les