Groupement de données évolutives dans la fouille d'usage du Web [1]

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

∗ *INRIA, Domaine de Voluceau - Rocquencourt, B.P. 105, 78153 Le Chesnay, France*

∗∗ *CIn - UFPE, Caixa Postal 7851, CEP 50732-970, Recife (PE), Brésil*

RÉSUMÉ
Le changement d'intérêt des utilisateurs et les modifications liées à la structure et au contenu d'un site Web peuvent influencer la manière dont une visite y est réalisée. Ainsi, les modèles associés à ces comportements dans la fouille d'usage du Web doivent être mis à jour continuellement. Une solution est de mettre à jour ces modèles à l'aide des résumés obtenus par une approche évolutive des méthodes de classification.

SUMMARY
The way in which a Web site is visited can indeed evolve due to modifications of the structure and the content of the site, or due to changes in the behaviour of certain users. Thus, the models associated with these behaviours in the Web Usage Mining domain must be updated continuously. A solution to this problem is to update these models using summaries obtained by an evolutionary approach of the classification methods.

MOTS CLÉS: Analyse de données, Fouille de données d'usage du Web

1 Introduction

Les profils d'accès à un site Web peuvent être influencés par certains paramètres de nature temporelle (l'heure et le jour de la semaine, des événements saisonniers, etc.). Cependant, 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 ainsi ceux qui prédominent sur la totalité de la période. En conséquence, certains types de comportements, qui ont lieu pendant de courtes sous-périodes ne sont pas pris en compte par les méthodes classiques. Il est pourtant important d'étudier ces comportements et de réaliser une analyse portant sur des sous-périodes significatives. Le présent article propose de suivre le changement de comportement à l'aide des résumés obtenus par une approche évolutive de la classification.

[1]Les auteurs tiennent à remercier le projet INRIA/FACEPE et la CAPES-Brésil.

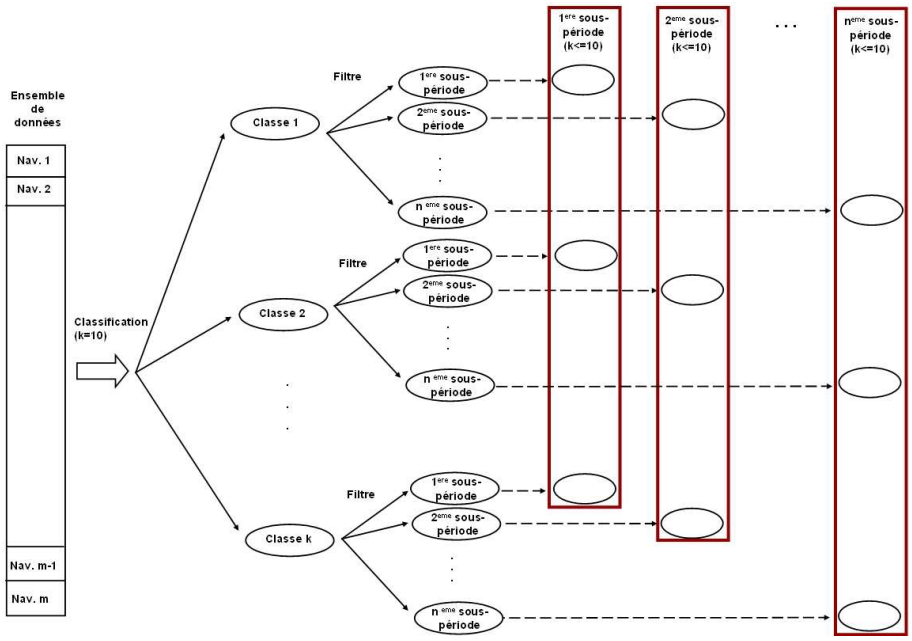

Figure 1: Schéma de la classification globale

2 Approche de classification par sous-périodes de temps

L'approche proposée dans cet article consiste dans un premier temps à diviser la période analysée en sous-périodes plus significatives (mois de l'année). Ensuite, une classification est réalisée sur les données de chaque sous-période, aussi bien que sur la période complète. Les résultats fournis sont donc comparés les uns avec les autres. Nous avons proposé quatre classifications de la manière suivante : **(1)** *Classification globale* (cf figure 1) : cette classification est obtenue sur la totalité des individus; **(2)** *Classification locale indépendante* (cf figure 2): pour chaque zone temporelle *a priori*, on réalise une classification de l'ensemble d'individus concernés. Comme chaque zone est distincte, chaque classification est donc indépendante des autres; **(3)** *Classification locale "précédente"* (cf figure 3) : ici, on utilise la structure classificatoire de la sous-période temporelle précédente pour obtenir une partition de la sous-période courante; et **(4)** *Classification locale dépendante* (cf figure 4) : ici, on initialise l'algorithme pour une sous-période temporelle avec les résultats de cet algorithme appliqué sur la sous-période précédente.

2.1 Algorithme et critères d'évaluation

Pour le groupement de données, nous utilisons un algorithme de type nuées dynamiques applicable sur un tableau de données (voir tableau 1). Pour toutes les procédures de classification, nous avons demandé 10 classes avec un nombre d'initialisations aléatoires

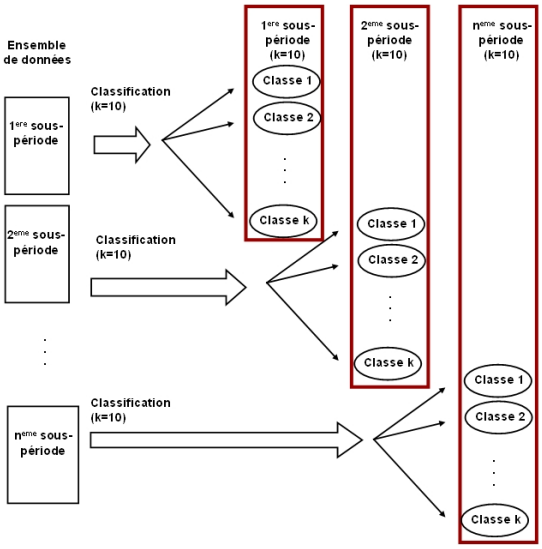

Figure 2: Schéma de la classification locale indépendante

No	Champ	Signification
1	IDNavigation	Code de la navigation
2	NbRequests_OK	Nombre de requêtes réussies (statut = 200) dans la navigation
3	NbRequests_bad	Nombre de requêtes échouées (statut $\neq$ 200) dans la navigation
4	PRequests_OK	Pourcentage de requêtes réussies (= NbRequests_OK/ NbRequests)
5	NbRepetitions	Pourcentage de requêtes répétées dans la navigation
6	PRepetitions	Pourcentage de repetitions (= NbRepetitions / NbRequests)
7	DureeTotale	Durée totale de la navigation (en secondes)
8	MDuree	Moyenne de la durée des requêtes (= DureeTotale / NbRequests)
9	MDuree_OK	Moyenne de la durée des requêtes réussies (= DureeTotale_OK / NbRequests_OK)
10	NbRequests_Sem	Nombre de requêtes rapportées aux pages dynamiques qui forment la structure sémantique du site
11	PRequests_Sem	Pourcentage des requêtes sémantiques (=NbRequests_Sem/ NbRequests) dans la navigation
12	TotalSize	Somme d'octets transférés dans la navigation
13	MSize	Moyenne d'octets transférés (= TotalSize / NbRequests_OK)
14	DureeMax_OK	Durée maximale parmi les requêtes réussies

Table 1: Attributs descriptifs des navigations.

égal à 100 (sauf dans les cas des classifications locales précédente et dépendante).

Pour analyser les résultats, nous utilisons la F-mesure de van Rijsbergen (1979) pour une analyse classe par classe et l'indice de Rand corrigé d'Hubert et Arabie (1985) pour une analyse globale. Pour les deux indices, la valeur 0 correspond à une absence totale de liaison entre les partitions considérées, alors que la valeur 1 indique une liaison parfaite.

2.2 Application et résultats

Nous utilisons comme site de référence celui du CIn de Recife-Brésil [2]. Ce site est constitué d'un ensemble de pages statiques et pages dynamiques (cf Rossi et al. (2006) et Da Silva

[2]Le site analysé est actuellement accessible à l'adresse : `http://www.cin.ufpe.br/`

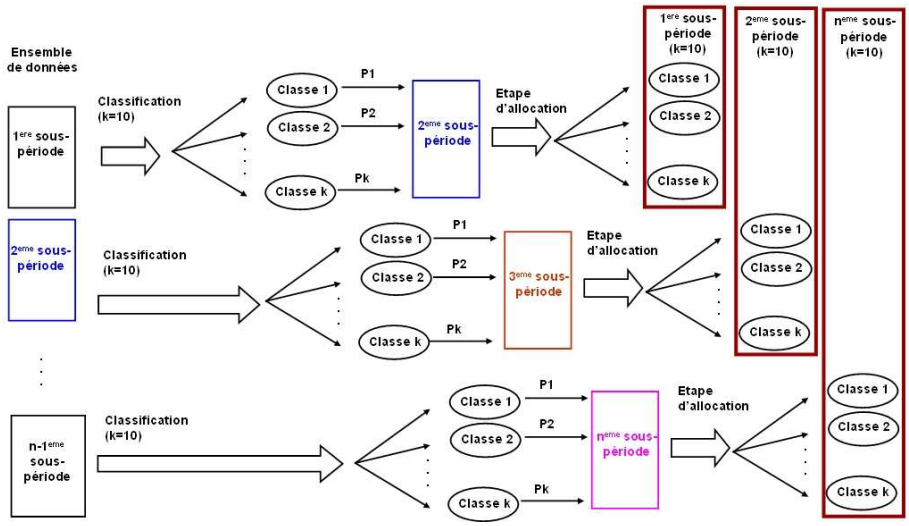

Figure 3: Schéma de la classification locale précédente

et al. (2006) pour une analyse de cette partie du site). Nous avons étudié les accès au site du 1 juillet 2002 au 31 mai 2003.

Pour le prétraitement des données, nous avons utilisé la méthodologie proposée par Tanasa et Trousse (2004) dans laquelle une *navigation* est définie comme une suite de requêtes provenant d'un même utilisateur et séparées au plus de 30 minutes. Après le prétraitement et l'élimination des cas aberrants nous avons obtenu 138 536 navigations.

A partir des valeurs de l'indice de Rand corrigé (cf figure 5), dans le cas de confrontation des classifications indépendante versus globale, il y a presque systématiquement des valeurs faibles, c'est-à-dire que certaines classes de la classification indépendante ne sont pas retrouvées dans la classification globale. On voit aussi que la classification "précédente" ne donne pas des résultats très différents de ceux obtenus par la classification dépendante. Ces différences sont confirmées par la F-mesure (cf figure 6). On note que les classes sont très stables dans le temps si on utilise la méthode de classification dépendante. En fait, aucun indice ne descend au dessous de 0.877, ce qui représente une très bonne valeur. Par contre, dans le cas de la classification indépendante, on obtient des classes très différentes de celles obtenues globalement (valeurs inférieures à 0.5).

3 Conclusions et perspectives futures

A travers nos expérimentations, nous avons abordé la problématique du traitement des données dynamiques d'usage du Web. La méthode de classification locale dépendante montre que les classifications obtenues ne changent pas ou peu au cours du temps, alors que la méthode de classification locale indépendante est plus sensible aux changements

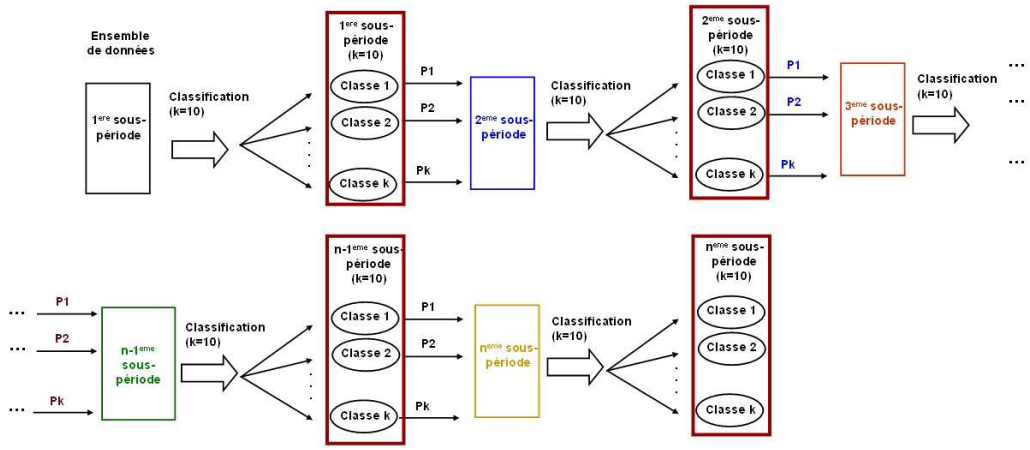

Figure 4: Schéma de la classification locale dépendante

qui peuvent se passer d'une sous-période à l'autre. Comme futurs travaux, nous pouvons signaler l'application d'autres méthodes de classification permettant la découverte automatique du nombre de classes ainsi que les fusion et les scissions.

Bibliographie

[1] Cooley, R., Mobasher, B. et Srivastava, J. (1999). Data Preparation for Mining WWW Browsing Patterns. *Journal of Knowledge and Information Systems*, **1**(1) 5–32.

[2] Da Silva, A., De Carvalho, F., Lechevallier, Y. et Trousse, B. (2006). Mining Web Usage Data for Discovering Navigation Clusters. *11th IEEE Symposium on Computers and Communications (ISCC 2006)*, 910–915.

[3] Hubert, L. et Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, **2** 193–218.

[4] Rossi, F., De Carvalho, F., Lechevallier, Y. et Da Silva, A. (2006b). Dissimilarities for Web Usage Mining. *Data Science and Classification (IFCS 2006)*, 39–46.

[5] Tanasa, D. et Trousse, B. (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. *IEEE Intelligent Systems*, **19**(2) 59–65.

[6] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London.

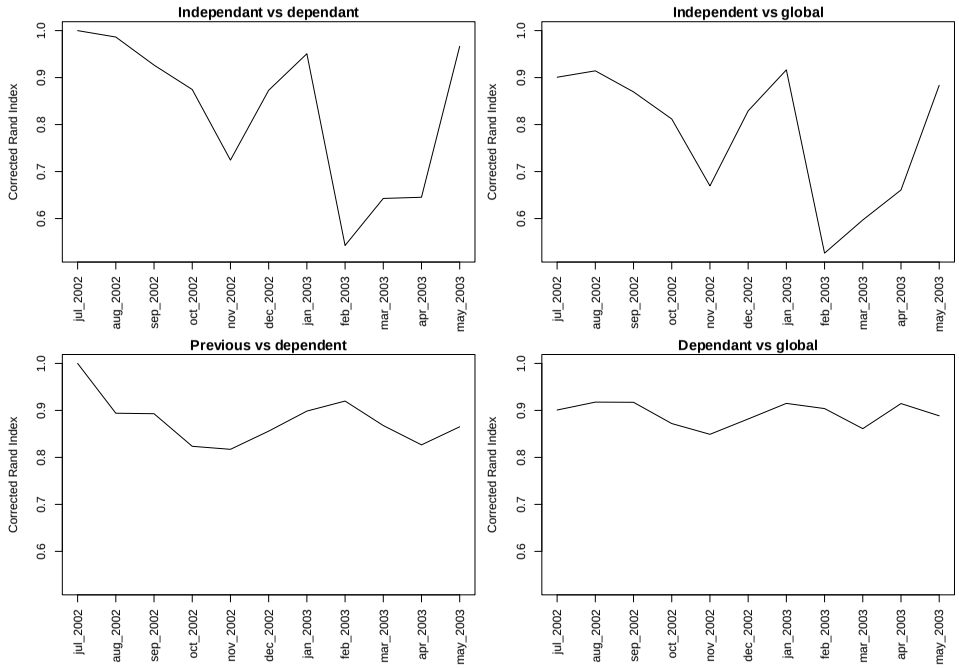

Figure 5: Indice de Rand corrigé partition par partition.

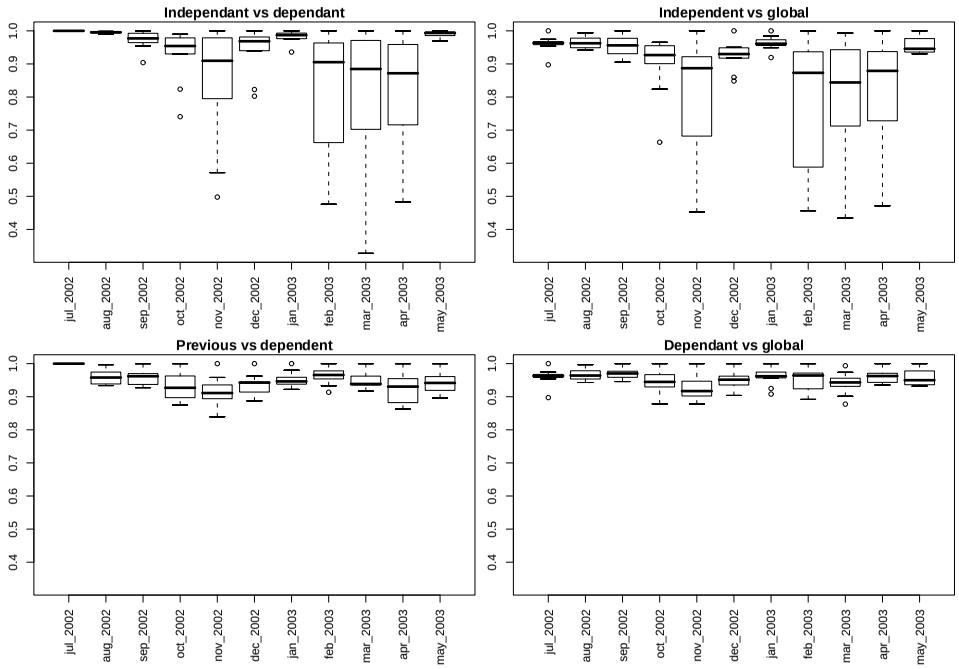

Figure 6: Boxplots correspondant aux F-mesures classe par classe.