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

Alzennyr Da Silva*, Yves Lechavellier*,
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 résument 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

auteurs discutent en quelques lignes des méthodes pour découvrir les modèles séquentiels, les motifs fréquents et les modèles périodiques partiels dans les flux de données.

Le présent article propose de suivre le changement de comportement à l'aide des résumés obtenus par une approche évolutive de la classification appliquée sur des sous-périodes de temps. L'article est organisé comme suit : la section suivante présente l'approche d'analyse de l'usage basée en sous-périodes temporelles. Nous présentons aussi dans cette section les expériences réalisées, en analysant les résultats et en les comparant à ceux des méthodes classiques. La dernière section présente les conclusions et les travaux futurs envisagés.

2 Approche de classification par sous-périodes de temps

La caractérisation de groupes d'utilisateurs consiste à identifier des traits d'usage partagés par un nombre suffisant d'utilisateurs d'un site Web et ainsi fournir des indices permettant d'inférer le profil de chaque groupe (Da Silva et al., 2006a,b). 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.

Dans ce contexte, nous avons réalisé quatre classifications de la manière suivante :

- **Classification globale** : cette classification est obtenue sur la totalité des individus ;
- **Classification locale indépendante** : pour chaque zone temporelle *a priori*, on réalise une classification de l'ensemble des navigations concernées. Comme chaque zone est distincte, chaque classification est donc indépendante des autres ;
- **Classification locale "précédente"** : ici, on utilise la structure classificatoire de la période temporelle précédente pour obtenir une partition de la période courante ;
- **Classification locale dépendante** : ici, on initialise l'algorithme pour une période temporelle avec les résultats de cet algorithme appliqué sur la période précédente.

2.1 Algorithme et critères d'évaluation

Pour la classification des navigations, nous utilisons un algorithme de type nuées dynamiques (cf Celeux et al. (1989)) applicable sur un tableau de données (voir tableau 1). L'algorithme doit en particulier : (1) pouvoir affecter de nouvelles observations à une classification existante, et (2) pouvoir initialiser l'algorithme avec les résultats d'une autre réalisation de lui-même. Pour toutes les procédures de classification, nous avons demandé 10 classes avec un nombre d'initialisations aléatoires égal à 100, sauf dans le cas de la classification locale dépendante.

Pour analyser les résultats, nous utilisons deux critères. Pour une analyse classe par classe, nous considérons la F-mesure de van Rijsbergen (1979). Pour une analyse plus globale, nous utilisons l'indice de Rand corrigé (cf Hubert et Arabie (1985)). 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.

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 <> 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

TAB. 1 – *Attributs descriptifs des navigations.*

2.2 Application et résultats

Les données d'usage d'un site Web proviennent essentiellement des fichiers log des serveurs concernés. Diverses techniques de pré-traitement permettent d'extraire des *navigations* à partir de ces fichiers, comme par exemple celles de Tanasa et Trousse (2004), utilisées dans cet article. Une navigation est une suite de requêtes provenant d'un même utilisateur et séparées au plus de 30 minutes.

Nous utilisons comme site de référence celui du Centre d'Informatique de Recife-Brésil [1]. Ce site est constitué d'un ensemble de pages statiques et pages dynamiques, ces dernières étant gérées par *servlets* programmées en Java (cf Rossi et al. (2006a,b) pour une analyse de cette partie du site). Nous avons étudié les accès au site du 1 juillet 2002 au 31 mai 2003. Après le filtrage et l'élimination des cas aberrants nous avons obtenu un total de 138 536 navigations.

Nous avons réalisé un suivi des prototypes des classes (mois par mois) pour les classifications locales indépendante et dépendante, puis nous avons projeté ces prototypes dans le plan factoriel (voir figure 1). Sur cette représentation, chaque cercle représente un prototype. Dans la classification dépendante, les dix classes sont représentées par des couleurs différentes. On note une certaine stabilité malgré la diversité de mois analysés. Dans le cas de la classification indépendante, la trajectoire temporelle est simplement matérialisée par les lignes qui joignent un prototype à son plus proche voisin dans la période temporelle précédente. Cela ne donne pas des trajectoires parfaitement identifiées car certains prototypes partagent à un moment donné le même prédécesseur. On note en fait que seules quatre classes sont parfaitement identifiées et stables, les autres subissant des fusions et séparations au cours du temps. Par l'analyse de la variance intra-classe, nous pouvons constater que les classes obtenues par la classification locale indépendante présentent plus de cohésion au sens de ce critère (voir figure 2).

A partir des valeurs de l'indice de Rand corrigé (cf figure 3), 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, ce qui confirme l'intuition acquise par l'observation des prototypes dans le plan factoriel : ces derniers bougent "peu" au cours du temps.

Ces différences sont confirmées par la F-mesure (cf figure 4). Ce qui apparaît nettement, c'est que les classes sont très stables dans le temps si on utilise la méthode de classification

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

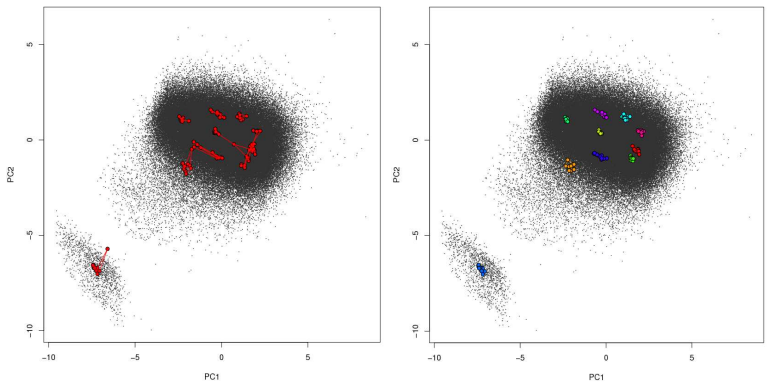

FIG. 1 – *Classifications locales : indépendante (gauche) et dépendante (droite).*

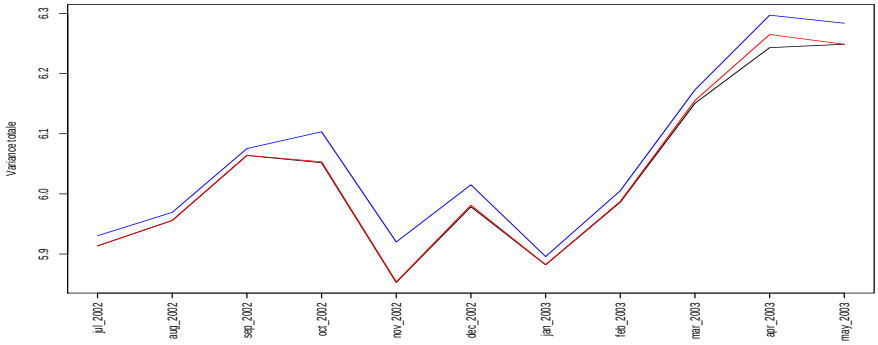

FIG. 2 – *Variance intra-classe des classifications : indépendante (trait noir), dépendante (trait rouge) et globale (trait bleu).*

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 au contraire des classes très différentes de celles obtenues globalement (avec des valeurs inférieures à 0.5).

3 Conclusions et perspectives futures

Dans cet article, nous avons abordé la problématique du traitement des données dynamiques dans le contexte de l'analyse de l'usage du Web. A travers nos expérimentations, nous pouvons dire que 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 qui peuvent se passer d'une sous-période à l'autre. Dans un plan secondaire, l'approche de classification locale indépendante permet

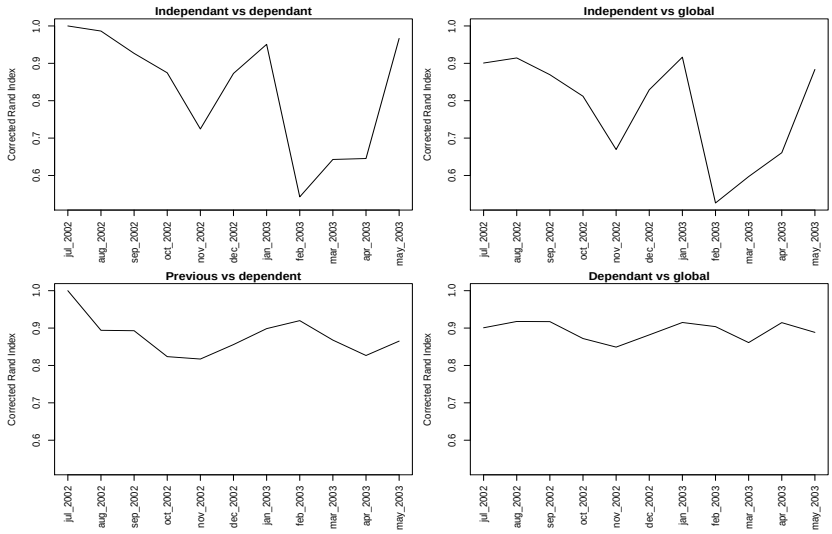

FIG. 3 – *Indice de Rand corrigé classe par classe.*

aussi de s'affranchir des difficultés liées aux limites des machines (telles que la taille de la mémoire, vitesse du processeur, etc.) car nous concentrons l'analyse sur une partie des données disponibles. Comme possibilité de futurs travaux nous pouvons signaler l'application d'autres méthodes de classification et la mise en oeuvre des techniques permettant la découverte automatique du nombre de classes, ainsi que l'introduction du processus de fusion ou de scission des classes.

Remerciements

Nous tenons à remercier le projet INRIA/FACEPE et la CAPES-Brésil pour leur soutien.

Références

Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, et H. Ralambondrainy (1989). *Classification Automatique des Données*. Paris : Bordas.

Cooley, R., B. Mobasher, et J. Srivastava (1999). Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems 1*(1), 5–32.

Da Silva, A., F. D. Carvalho, Y. Lechevallier, et B. Trousse (2006a). Mining web usage data for discovering navigation clusters. *ISCC 2006*, 910–915.

Da Silva, A., F. De Carvalho, Y. Lechevallier, et B. Trousse (2006b). Characterizing visitor groups from web data streams. *GrC 2006*, 389–392.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

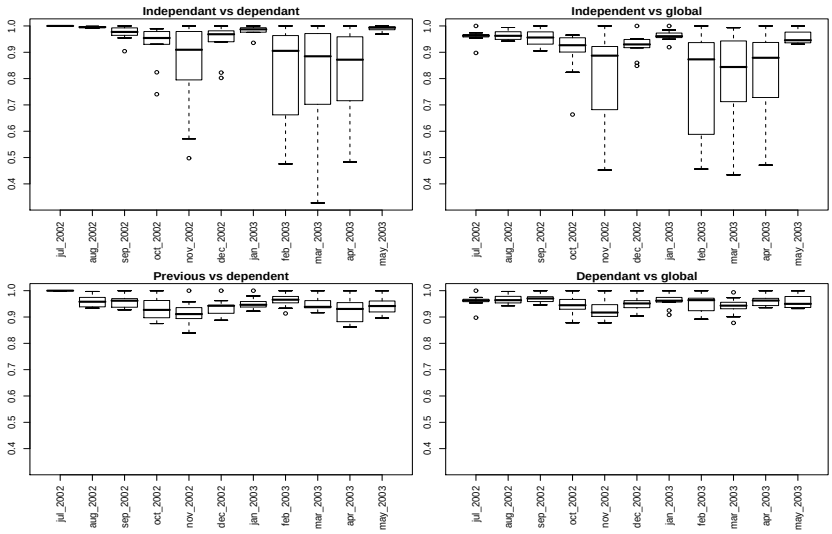

FIG. 4 – *Boxplots correspondant aux F-mesures classe par classe.*

Laxman, S. et P. S. Sastry (2006). A survey of temporal data mining. *SADHANA - Academy Proceedings in Engineering Sciences, Indian Academy of Sciences 31*(2), 173–198.

Roddick, J. F. et M. Spiliopoulou (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on KDE 14*(4), 750–767.

Rossi, F., F. De Carvalho, Y. Lechevallier, et A. Da Silva (2006a). Comparaison de dissimilarités pour l'analyse de l'usage d'un site web. *EGC 2006, RNTI-E-6 II*, 409–414.

Rossi, F., F. De Carvalho, Y. Lechevallier, et A. Da Silva (2006b). Dissimilarities for web usage mining. *IFCS 2006*, 39–46.

Tanasa, D. et B. Trousse (2004). Advanced data preprocessing for intersites web usage mining. *IEEE Intelligent Systems 19*(2), 59–65.

van Rijsbergen, C. J. (1979). *Information Retrieval* (second ed.). London : Butterworths.

Summary

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