Classifications non supervisées de données évolutives : application au Web Usage Mining

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

*Projet AxIS, INRIA Rocquencourt
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. Il est important pour les opérateurs de sites Web d'analyser l'utilisation de leurs sites afin de mieux connaître le comportement de leurs visiteurs. Il est aussi important de tenir compte de l'aspect temporel dans ces analyses. En effet, la manière dont une visite est réalisée 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 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 de comportements à l'aide des résumés obtenus par une approche évolutive des méthodes de classification. Pour ce faire, nous effectuons une partition de la période disponible de temps en sous-périodes de temps plus significatives. Nous comparons les résultats obtenus par cette méthode avec ceux obtenus par une analyse globale.

1 Introduction

Le Web est un des exemples les plus pertinents de source de données volumineuses et dynamiques grâce à l'augmentation colossale du nombre de documents mis en ligne et des nouvelles informations ajoutées chaque jour. Les profils d'accès au Web ont une nature dynamique, dûe au propre dynamisme du contenu et de la structure d'un site Web ou bien dûe au changement d'intérêt de ses utilisateurs au cours du temps. 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 (vacances d'été, d'hiver, Noël, etc.), des événements externes dans le monde (épidémies, guerres, crises économiques, coupe du monde de *football*), etc.

Une méthode d'analyse dirigée par l'usage consiste à étudier les traces laissées par les utilisateurs d'un site Web lors de leurs passages, plus précisément les informations enregis-

trées dans des fichiers logs du serveur Web. Dans ce cadre, la plupart des méthodes consacrées à la fouille de données d'usage du Web 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. Ceci permet, par exemple, d'étudier d'éventuels changements des centres intérêts des utilisateurs d'un site Web sur ces sous-périodes de temps. On peut ainsi étudier l'évolution temporelle des profils des utilisateurs en les caractérisant par des descriptions capables d'intégrer l'aspect temporel. Le volume des données considérées étant très élevé, il est en outre important de recourir à des résumés pour représenter les profils considérés.

Toutes ces considérations ont motivé d'importants efforts dans l'analyse de données, notamment pour adapter des méthodes de la fouille de données statiques aux données qui évoluent pendant le temps. Le présent article se place dans ce courant de recherche et 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 2 contient un bref état de l'art concernant la problématique abordée, dans la section 3 nous présentons les données d'usage et le site Web qui nous sert de référence. La section 4 décrit l'approche proposée pour l'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 de référence. La dernière section présente les conclusions et les travaux futurs envisagés.

2 Etat de l'art

La fouille du Web (Web Mining, WM) (Kosala et Blockeel, 2000) s'est développée à la fin des années 90 et consiste à utiliser l'ensemble des techniques de la fouille de données afin de développer des outils permettant l'extraction d'informations pertinentes à partir de données du Web (documents, traces d'interactions, structure des pages, structure des liens, etc.). Plus précisément, la fouille de données d'usage du Web (Web Usage Mining, WUM) désigne l'ensemble de techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web (Cooley et al., 1999) (Spiliopoulou, 1999) et c'est précisément sur cette partie que le présent article se concentre.

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 de 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 séquentiels. Ils évoquent également des techniques concernant l'analyse statistique de telles approches.

Cependant, la plupart des méthodes d'analyse de l'usage sont appliquées sur la totalité de données disponibles. Ainsi, les changements intéressants qui pourraient se produire dans la

période considérée ne sont pas pris en compte. Par exemple, quand les données proviennent d'une accumulation portant sur une période de temps potentiellement longue (comme dans le cas des fichiers log Web), on s'attend à ce que les comportements évoluent avec le temps. Pour traiter cela, les modèles d'usage découverts doivent être mis à jour continuellement (avec des algorithmes efficaces) afin de suivre le changement de comportement des visiteurs. Ceci exige de la méthode une surveillance continue des modèles découverts, pré-requis d'importance essentielle pour les applications centrées sur la dimension temporelle. Une solution possible pour traiter ce problème est proposée dans cet article : effectuer le partitionnement du temps avant l'application de la méthode et intégrer un suivi temporel des profils de comportement dans l'algorithme d'extraction de ces derniers.

3 Données d'usage

Les données d'usage d'un site Web proviennent essentiellement des fichiers log des serveurs concernés (dans une approche centrée serveur). Chaque ligne du fichier log décrit une requête reçue par le serveur associé : elle indique ainsi le document demandé, la provenance de la requête, la date de la demande, etc. Diverses techniques de pré-traitement permettent d'extraire des logs des *navigations*, comme par exemple celles de Tanasa et Trousse (2004), utilisées dans le présent article. Une navigation est une suite de requêtes provenant d'un même utilisateur et séparées au plus de 30 minutes. Elle constitue donc la trajectoire d'un utilisateur sur le site.

Pour l'application de notre approche, nous utilisons comme site de référence celui du Centre d'Informatique de Recife-Brésil [1], le laboratoire d'un des auteurs. Ce site est constitué d'un ensemble de pages statiques (pages personnelles des professeurs, pages de support de cours, etc.) et dynamiques, ces dernières étant gérées par *servlets* programmées en Java. La partie dynamique du site consiste en 91 pages très bien organisées sous la forme d'un arbre sémantiquement structuré (cf Rossi et al. (2006a) Rossi et al. (2006b) Da Silva et al. (2006a) Da Silva et al. (2006b) pour une analyse de cette partie du site). Nous avons étudié les accès au site du 1 juillet 2002 au 31 mai 2003 (le fichier de logs contient environ 2Go de données brutes).

Afin d'analyser les traces d'usage plus représentatif, nous avons sélectionné les navigations longues (contenant au moins 10 requêtes et avec une durée totale d'au moins 60 secondes) et supposées humaines (le ratio durée sur requêtes doit être supérieur à 4, c'est-à-dire, correspondre à moins de 15 requêtes par minute). Ces efforts ont comme but d'extraire les navigations humaines et d'exclure celles provenant des robots. L'élimination des navigations courtes est motivée par la recherche de patrons d'utilisation du site, à l'exclusion des accès simples (par l'intermédiaire d'un moteur de recherche) qui n'engendre pas un parcours sur le site. Après le filtrage et élimination des cas aberrants nous avons obtenu un total de 138 536 navigations, 184 275 pages (dont 91 dynamiques), 56 314 utilisateurs et 1,19 minutes comme durée moyenne de visualisation de pages.

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

4 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 (Chi et al., 2002) (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 (mois) dans le but de rechercher l'apparition et/ou l'évolution des comportements qu'on manquerait par une analyse globale de toute la période. Ensuite, une classification est réalisée sur les données de chaque sous-période, aussi bien que sur la période complète. Les résultats fournis pour chaque classification sont donc comparés les uns avec les autres. Notons que le partitionnement temporel porte sur l'ensemble des navigations qui est ainsi découpé en sous-groupes. Chaque navigation est donc affectée à un sous-groupe, ce qui est assez différent des analyses statistiques élémentaires dans lesquelles on compte par exemple les accès à une page en fonction de l'heure d'accès, sans tenir compte des navigations.

Dans notre approche, le partitionnement du temps est formulé en fonction du mois, cependant d'autres possibilités de partitionnement du temps (intervalles de 15 jours, jours fériés et non fériés, périodes dans la journée : matin, après midi, soir, etc.) sont aussi envisageables.

Dans un contexte non supervisé, nous avons réalisé quatre classifications :

- **Classification globale** : on réalise une classification sur tout l'ensemble d'individus (navigations), sans tenir compte des sous-périodes de temps. Pour des raisons d'analyse des résultats, on réalise en suite une intersection entre les classes obtenues et la partition temporelle. La classification globale engendre donc une classification chacun des 11 groupes temporels ;
- **Classification locale indépendante** : pour chaque mois analysé, on réalise une classification sur l'ensemble de navigations appartenant au mois en question. Chaque classification est donc indépendante des autres classifications réalisées sur les autres mois ;
- **Classification locale "précédente"** : cette classification peut être obtenue à partir d'une autre classification quand l'algorithme utilisé est capable d'affecter de nouveaux individus aux groupes obtenus (cf la section suivante). On utilise donc la structure classificatoire de la période temporelle précédente pour obtenir une partition sur la période suivante, ce qui correspond dans notre cas à la première étape de l'algorithme de nuées dynamiques ;
- **Classification locale dépendante** : cette classification peut être obtenue avec des algorithmes itératifs de type nuées dynamiques (cf la section suivante) qu'on peut initialiser de façon adaptée. Ici, on initialise la classification d'une période temporelle avec les prototypes obtenus de la classification de la période temporelle immédiatement précédente.

4.1 Algorithme et critères d'évaluation

Pour la classification des navigations dans notre méthode, nous utilisons un algorithme de type nuées dynamiques (cf Celeux et al. (1989)) applicable sur un tableau de données caractérisé par des attributs descriptifs des navigations (voir tableau 1). D'autres algorithmes de partitionnement sont bien entendu envisageables, mais ils doivent être compatibles avec les techniques de classification décrites dans la section précédente. Il faut en particulier :

1. pouvoir affecter de nouvelles observations à une classification existante ;

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 où l'algorithme était initialisé avec les résultats obtenus de la sous-période temporelle précédente.

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

TAB. 1 – *Attributs descriptifs des navigations.*

Pour analyser les résultats, nous utilisons deux critères. Pour une analyse classe par classe, nous considérons la F-mesure de van Rijsbergen (1979). Pour comparer deux partitions, nous cherchons la meilleure représentation de la classe a de la première partition par une classe b dans la seconde partition, au sens de la F-mesure (ce qui nous donne autant de valeurs numériques qu'il y a de classes dans la première partition). Cette mesure permet une analyse fine, mais elle ne tient pas compte des effectifs relatifs des classes. Pour une analyse plus globale, nous utilisons l'indice de Rand corrigé (cf Hubert et Arabie (1985)) qui permet de comparer directement deux partitions. 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.

4.2 Résultats et discussion

Pour mieux comprendre l'évolution des classes par rapport aux sous-périodes de temps analysées, nous avons réalisé un suivi des prototypes des classes (mois par mois) pour la classification locale indépendante et la classification locale dépendante, puis nous avons projeté ces prototypes dans le plan factoriel obtenu sur la population totale (voir figure 1). Sur cette représentation, chaque disque représente un prototype. Dans la classification dépendante (à droite), les dix classes sont représentées par des couleurs différentes et les traits représentent la trajectoire des prototypes. On note une certaine stabilité malgré la diversité de mois analysés. Dans le cas de la classification indépendante (à gauche), 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.

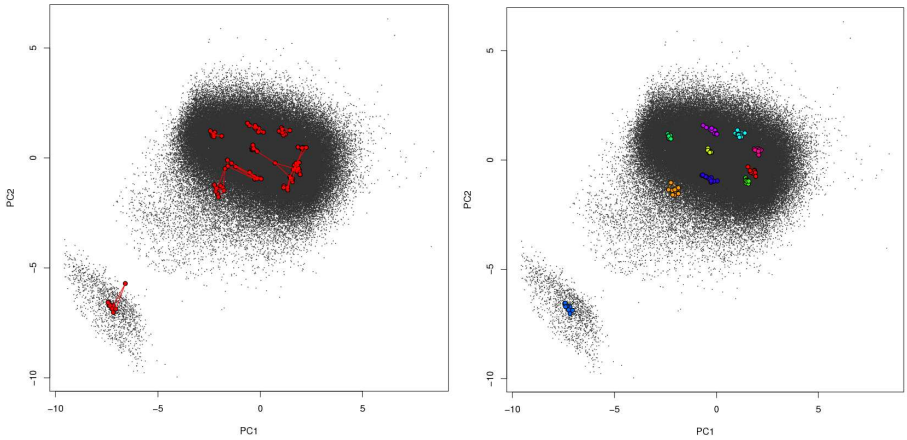

FIG. 1 – *Projection et suivi des prototypes des classes pour les classifications locales : indépendante (gauche) et dépendante (droite).*

Si on projette sur le plan factoriel les prototypes de la classification globale (G1, G2, ..., G10), puis les prototypes obtenus par les classifications locales indépendante et dépendante (voir figure 2), on voit que la classification locale indépendante est capable d'identifier des nouvelles classes qui ne sont pas trouvées par les deux autres classifications.

Nous avons également calculé la variance intra-classe pour les trois classifications (dépendante, indépendante et globale) mois par mois. Comme attendu, le score est meilleur pour la classification locale indépendante, puis classification locale dépendante et finalement pour la classification globale (voir figure 3). A partir de cela, nous pouvons constater que les classes obtenues par la classification locale dépendante présentent plus de cohésion.

Vue la faible différence entre les scores, on s'attendrait donc à ce que les classes soient assez proches. Cependant, la représentation dans le plan factoriel induit un premier doute, car les prototypes dans le cas indépendant semblent parfois assez différents de ceux des cas globaux et dépendants.

En fait, les valeurs de l'indice de Rand ne montrent que la classification indépendante est parfois très différente de la classification globale (figure 4). Ces différences sont confirmées par la F-mesure. Comme on obtient 10 valeurs de F-mesure pour chaque mois (une par classe), on trace une *boxplot* de résumés, ce qui donne la figure 5. On voit que dans le cas de confrontation des classifications indépendante versus globale il y a presque toujours 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.

En confrontant la classification globale via la F-mesure (classe par classe) aux classifications locales dépendante et indépendante, on affine l'analyse ci-dessus. Ce qui apparaît nettement, c'est que les classes sont très stables dans le temps si on utilise la méthode de classifica-

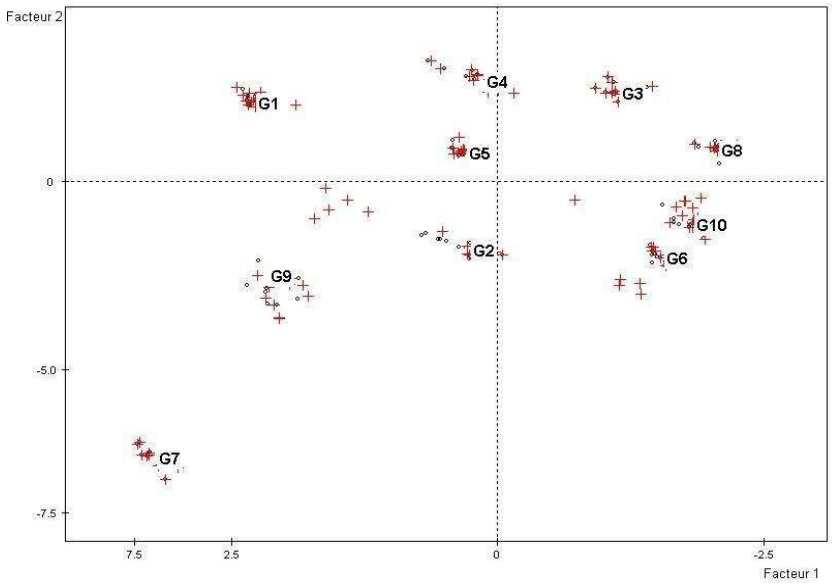

FIG. 2 – *Projection des prototypes des classes pour les classifications : globale (G1, G2, ..., G10), indépendante (+) et dépendante (o).*

tion dépendante. En fait, aucune classe ne descend au dessous de 0.877 pour la F-mesure, 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 faibles de la F-measures, inférieures à 0.5).

D'une certaine façon, il est surprenant de constater que les partitions découvertes par la classification locale dépendante sont très proches de celles obtenues par la classification globale. Nous pourrions ainsi spéculer qu'une analyse effectuée sur des sous-périodes de temps soit capable d'obtenir les mêmes résultats censés être révélés par une analyse globale effectuée sur toute la période disponible de temps. En outre, la méthode de classification locale dépendante peut être considérée comme une approche de type "diviser pour régner" et pour cela, serait capable de régler certaines contraintes liées au temps total de traitement de données et aux limitations physiques des machines (telles que la taille de la mémoire, vitesse du microprocesseur, etc.).

En conclusion, nous pouvons dire que la méthode de classification locale dépendante montre que les classifications obtenues ne changent pas ou peu au cours du temps (ce qui est confirmé par la variance intra-classe), alors que la méthode de classification locale indépendante ne réussit pas à découvrir ce fait, étant plus sensible aux changements qui peuvent se passer d'une sous-période à l'autre.

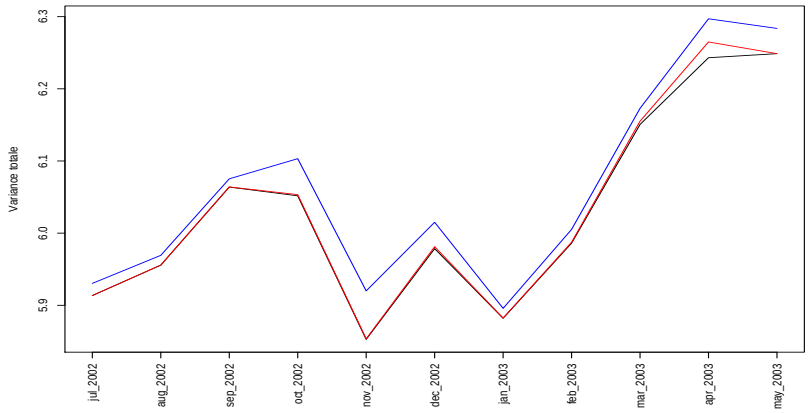

FIG. 3 – *Variance totale pour les classifications : indépendante (trait noir), dépendante (trait rouge) et globale (trait bleu).*

5 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. Les questions traitées ont montré la nécessité de définition et/ou d'adaptation de méthodes capables d'extraire des connaissances et de suivre l'évolution de ce type de données. Bien qu'il existe de nombreuses méthodes performantes d'extraction de connaissances, peu de travaux ont été consacrés à la problématique de données pouvant évoluer avec le temps.

A travers nos expérimentations, nous avons montré que l'analyse des données dynamiques par sous-périodes offre un certain nombre d'avantages. Ainsi, elle permet de rendre la méthode plus efficace dans la découverte des classes et des possibles changements qui peuvent avoir lieu pendant de courts sous-périodes de temps et ne sont pas détectés par une analyse globale. Dans un plan secondaire, notre approche permet 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, ce qui permettrait d'identifier plus clairement les possibles fusionnements et scissions des classes au cours du temps.

Remerciements

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

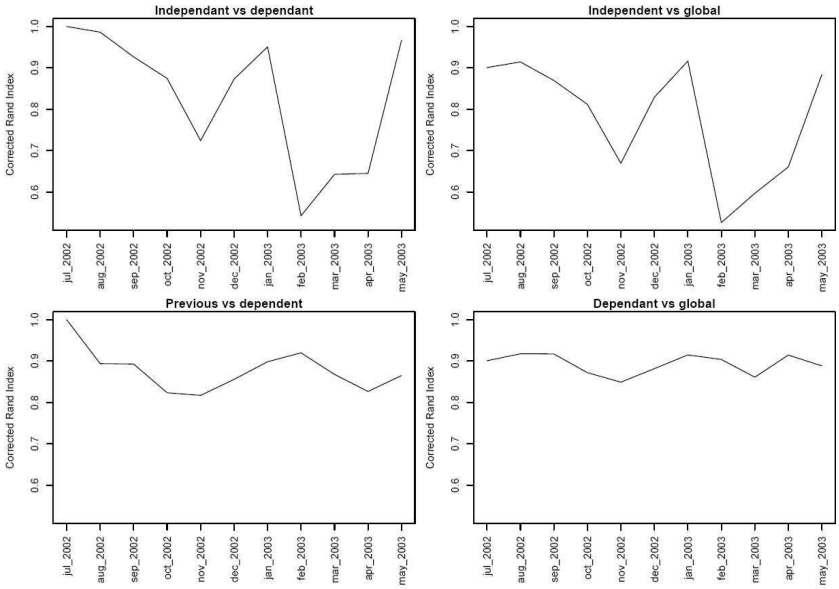

FIG. 4 – *Indice de Rand corrigé classe par classe.*

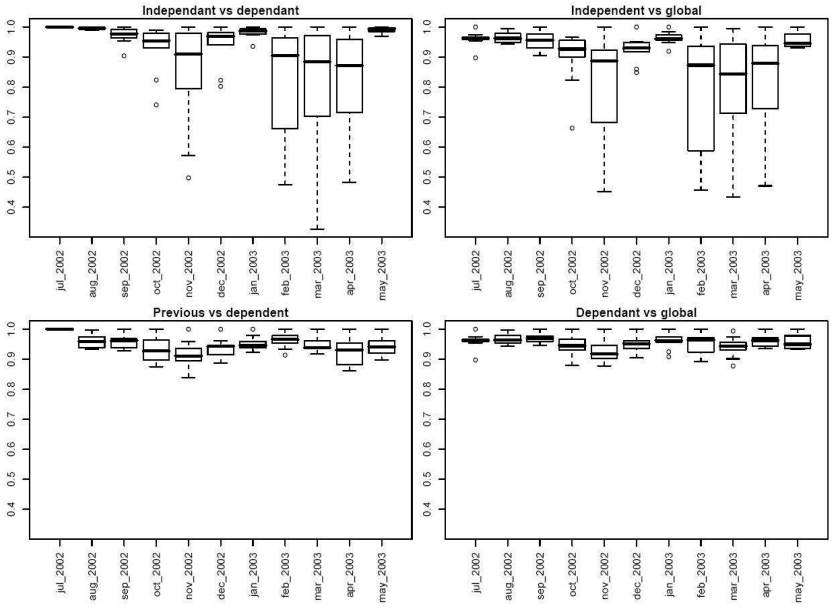

FIG. 5 – *Boxplots correspondant aux F-measures classe par classe.*

Références

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

Chi, E. H., A. Rosien, et J. Heer (2002). Lumberjack: Intelligent discovery and analysis of web user traffic composition. *ACM SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles (WEBKDD)*, 1–16.

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.

Kosala, R. et H. Blockeel (2000). Web mining research: A survey. *ACM SIGKDD Explorations 2*, 1–15.

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.

Spiliopoulou, M. (1999). Data mining for the web. *Workshop on Machine Learning in User Modelling of the ACAI99*, 588–589.

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

Web usage analysis is very important for Web sites' operators as it provides some understanding of behaviour of their visitors. The way in which a site is visited can indeed evolve due to modifications of the structure and/or the contents of the site, or due to changes in the behaviour of certain user groups. Thus, the models associated with these behaviours 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 method. For that, we carry out a split of the time into more significant time sub-periods. We compare the results obtained with this method to those produced by a global analysis.