Approche pour le suivi de l'évolution des données d'usage du Web : application sur un jeu de données en *marketing*

Alzennyr Da Silva

Projet AxIS, INRIA Paris-Rocquencourt, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay – France
Ceremade, Université Paris-Dauphine, Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16 – France
{*Alzennyr.Da_Silva@inria.fr*}

RÉSUMÉ. *Dans la fouille des flux des données d'usage du Web, la dimension temporelle joue un rôle très important car les comportements des internautes peuvent changer au cours du temps. Dans cet article, nous présentons l'application d'une approche de classification automatique basée sur des fenêtres sautantes pour la détection et le suivi de changements sur un jeu de données en marketing. Cette approche combine les cartes auto organisatrices de Kohonen et la méthode de Ward pour la découverte automatique du nombre de clusters de comportement et deux indices de validation basés sur l'extension pour la détection des changements au cours du temps.*

MOTS-CLÉS : *Classification automatique, Données évolutives, Fouille d'usage du Web*

1. Introduction

La fouille de données d'usage du Web (*Web Usage Mining*, en anglais) désigne l'ensemble des techniques basées sur la fouille de données pour analyser le comportement des utilisateurs d'un site Web [COO 99, SPI 99]. Dans la conférence SLDS 2009 [1], un jeu de données concernant le suivi des achats d'un panel de consommateurs a été diffusé pour analyse dans le cadre d'un concours ouvert aux jeunes chercheurs. L'objectif de cet article est de décrire l'analyse des résultats obtenus à partir de l'application de notre approche de détection et de suivi des changements [DAS 08],[DAS 09] sur ce jeu de données.

2. Description du jeu de données

Le jeu de données en question concerne le suivi des achats de 10 068 clients pendant 14 mois (du 09 juillet 2007 jusqu'au 08 septembre 2008) sur 2 marchés de biens de consommation. Chaque marché commercialise 3 marques de produits. Les données ont été fournies dans un fichier de 3 745 296 lignes contenant les champs décrits dans le tableau 2. Dans ce fichier, tous les croisements *date x marché x marque* ont été présentés, même en cas d'absence d'achat où la valeur est nulle. Pour l'application de notre méthode, nous avons défini un tableau croisé *client x marque achetée* ordonné selon la date et l'identification client (cf. tableau 2). Ce tableau contient un total de 262 215 lignes.

1. Symposium Apprentissage et Science des Données 2009, `www.ceremade.dauphine.fr/SLDS2009`

TABLE 1. *Champs descriptifs des données*

ident	identification client
date	premier jour de la semaine d'achat concernée
marché	marché concerné par les achats (marche_1, marche_2)
marque	marque concernée par les achats (A, B, C, D, E, F)
valeur	valeur des achats

TABLE 2. *Extrait des premières lignes du tableau croisé*

ident	A	B	C	D	E	F	date
2	0	0	0	0	1.65	0	09/07/2007
6	0	0	0	0	1.76	0	09/07/2007
8	7.7	0	0	0	0	0	09/07/2007
10	3.52	0	0.88	0	0	0	09/07/2007
11	1.87	0	0	0	0	0	09/07/2007
12	0	0	0	1.65	0	0	09/07/2007
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

3. Approche de classification automatique pour la détection et le suivi de changements sur des données évolutives

Notre approche de classification automatique pour la détection et le suivi de changements sur des données évolutives consiste dans un premier temps à partitionner le flux de données d'entrée en fenêtres sautantes de taille fixe. Soit W_1 la première fenêtre du flux de données, l'idée est d'appliquer une méthode quelconque de classification non supervisée sur les données de W_1. Soit $G = (g_1, \ldots, g_c, \ldots, g_k)$ l'ensemble de prototypes des clusters découverts par cette classification. Ensuite, on repère la fenêtre suivante dans le temps, nommée W_2. Cette fenêtre ne se chevauche pas avec la première. Le pas suivant consiste à affecter les données de W_2 aux prototypes de G, ce qui nous fournit une première partition P_1. Par la suite, on applique la même méthode de classification non supervisée sur les données de W_2, ce qui définit une nouvelle partition P_2. La détection de changement entre les deux fenêtres sera donc mesurée par la comparaison des deux partitions P_1 et P_2 à l'aide de deux critères d'évaluation.

L'idée principale derrière cette approche consite à ce que la partition P_1 reflète la segmentation des individus de la fenêtre courante selon l'organisation des clusters dans la fenêtre précédente. La partition P_1 contient donc une information d'organisation ancienne dans le temps. De l'autre côté, la partition P_2 n'est pas influencée par la classification précédente et reflète donc la segmentation actuelle des individus, classés indépendamment du passé.

Sur un jeu de données stable, les deux partitions P_1 et P_2 seront très similaires, si ceci n'est pas le cas, alors une rupture entre les deux fenêtres est vérifiée. Dans ce cas, une analyse plus approfondie entre les clusters de ces deux partitions s'avère nécessaire afin de détecter la nature des changements vérifiés.

Pour cette comparaison, nous appliquons deux indices de validation basés sur l'extension (individus) : la F-mesure [RIJ 79] et l'indice corrigé de Rand [HUB 85]. Le premier indice assume des valeurs contenues dans l'intervalle $[0, +1]$ et le deuxième indice dans l'intervalle $[-1, +1]$. Dans les deux cas, les valeurs proches de 1 correspondent à des partitions très semblables, alors que les valeurs proches de 0 correspondent à des partitions dissimilaires.

Pour la méthode de classification, nous utilisons les cartes auto organisatrices de Kohonen [KOH 95] initialisées à partir d'une ACP (Analyse en Composantes Principales) sur des données d'entrée [ELE 99]. La couche de compétition est initialisée avec une centaine de neurones disposés sur une grille rectangulaire. Après la conver-

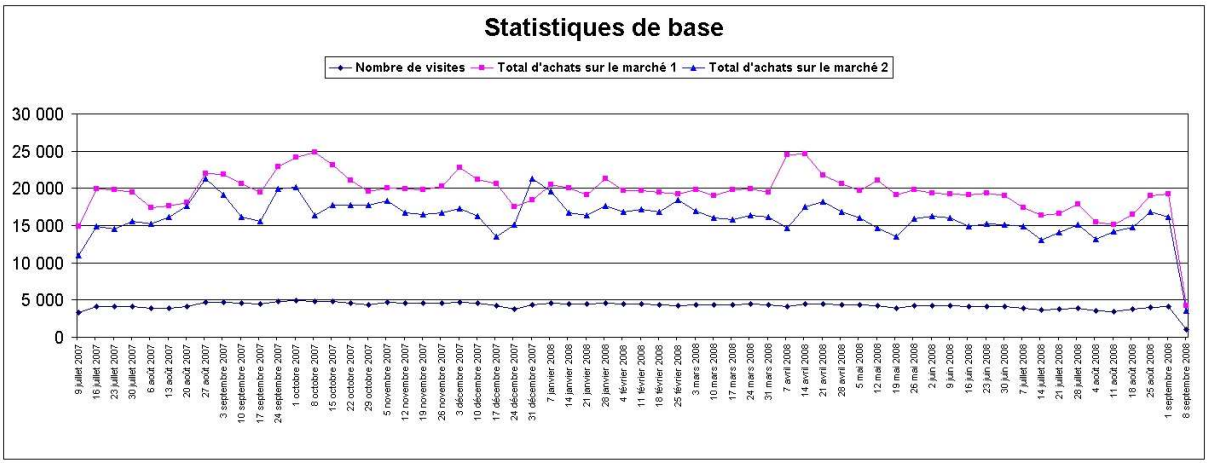

FIGURE 1. *Statistiques de base sur les données analysées.*

gence, une CAH (Classification Ascendante Hiérarchique) utilisant le critère de Ward est appliquée sur les prototypes correspondant aux neurones de la grille. La coupure du dendrogramme résultant est effectuée selon le critère du gain d'inertie intra-classe. Les neurones dans un même groupe sont donc fusionnés. De cette manière, le nombre de clusters présents dans les données est automatiquement découvert.

Il est important de remarquer que notre approche est totalement indépendante de la méthode de classification non supervisée appliquée, la seule restriction est que celle-ci doit fournir un prototype pour chaque cluster découvert.

4. Analyse des résultats

Dans notre approche, nous avons utilisé une fenêtre de temps de taille égale à une semaine. C'est-à-dire, notre méthode de classification non supervisée est appliquée sur l'ensemble d'individus ayant réalisé des achats dans une même semaine, ceci afin de segmenter les clients en fonction de leurs habitudes d'achat dans le temps. La figure 1 présente quelques statistiques de base, comme le nombre de visites de clients et le total d'achat par marché et par semaine. Comme nous pouvons remarquer, le nombre de visites est assez stable pendant toute la période analysée. Ce nombre avoisine 5 mille visites par semaine. De plus, la valeur totale d'achats sur le marché 2 est presque toujours supérieure à celle du marché 1 pour une même période de ventes.

Le nombre total de clusters de préférences d'achat découverts par notre méthode varie entre 2 et 8 (cf. figure 2). L'axe des abscisses du graphique dans la figure 2 répresente la date de début de la semaine analysée. Le nombre de clusters le plus stable étant 7 pour la majorité des semaines. Un cluster représente un ensemble de clients ayant des préférences d'achat similaires pour une même semaine. Remarquons que pendant les trois premières semaines du mois de novembre de 2007, le nombre total de clusters de préférences d'achat a été stabilisé et égal à 2. Ceci peut être une réponse à une stratégie de vente mise en ligne pendant cette période de temps.

Les valeurs obtenues par les deux indices de comparaison de partition cités dans la section 3 sont montrés dans les figures 3 et 4. La F-mesure effectue une analyse cluster par cluster en cherchant la meilleure représentation (match) d'un cluster dans la première partition par un cluster correspondant dans la deuxième partition. La F-mesure obtient donc autant de valeurs qu'il y a de clusters dans la première partition. On trace une *boxplot* à partir de ces valeurs pour chaque semaine analysée. Les périodes les plus stables sont celles qui présentent les valeurs les plus élevées de la F-mesure. Sur le jeu de données analysé, ces périodes correspondent aux semaines débutées le 20 août 2007, 17 septembre 2007, 10 décembre 2007, 25 février 2008 et 02 juin 2008 (cf. figure 3). Ces mêmes

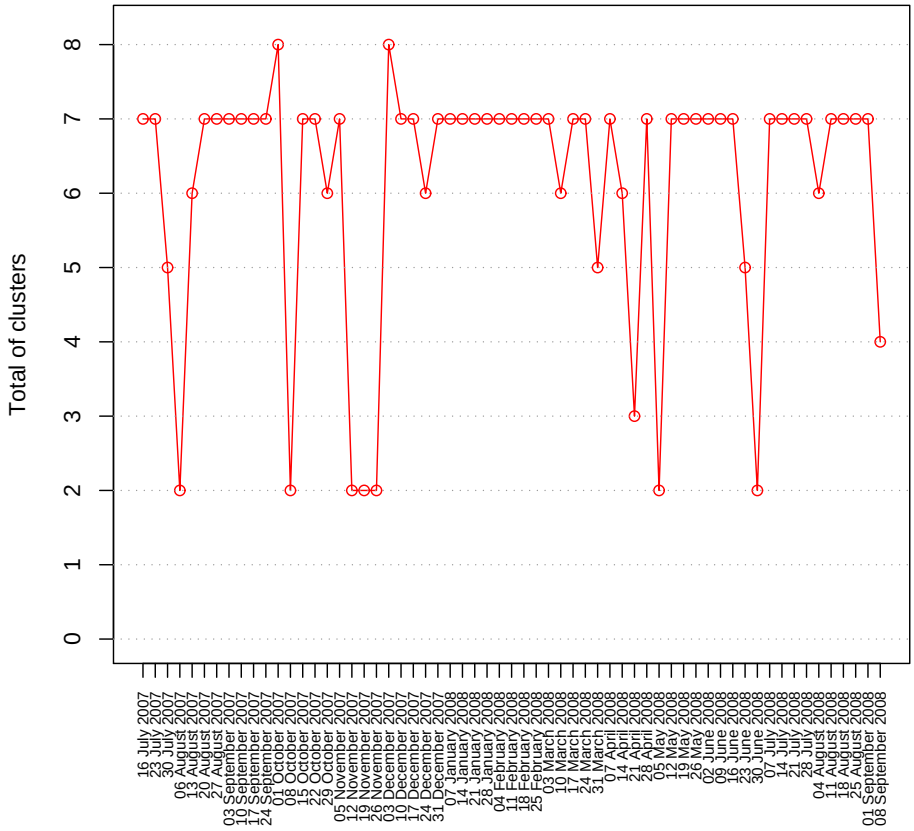

FIGURE 2. *Nombre de clusters de comportement par fênetre de temps (semaine).*

périodes ont été également repérées par l'indice de Rand corrigé (cf. figure 4). Cet index fournit une mesure globale basée sur tout l'ensemble de clusters dans les deux partitions. Ceci montre l'accord entre ces deux indices et leur aptitude à résoudre ce type de problème. Pour les autres semaines analysées, ce jeu de données reste assez instable. Les clusters de comportement subissant un nombre important de changement au cours du temps.

Afin de réaliser un suivi des changements subis par les clusters, notre méthode implémente des fonctionnalités additionnelles capables de repérer les transformations subies par un cluster de comportement au cours du temps. Ces transformations peuvent être de différentes natures, à savoir :

- **Disparation d'un cluster de comportement** : repérée quand un cluster de comportement existant dans une fenêtre de temps précédente n'est plus présent dans la fenêtre de temps suivante.
- **Apparition d'un cluster de comportement** : repérée quand un cluster de comportement inexistant dans une fenêtre de temps précédente est aperçu dans la fenêtre de temps suivante.
- **Scission d'un cluster de comportement** : repérée quand un certain nombre de clients appartenant à un cluster de comportement migre vers d'autres clusters. Ceci peut indiquer un changement de préférence pour une nouvelle marque de produit.
- **Fusion de deux ou plusieurs clusters de comportement** : repérée quand les clients appartenant à différents clusters de comportement migrent vers un même cluster de comportement. Ceci indique quand deux clients

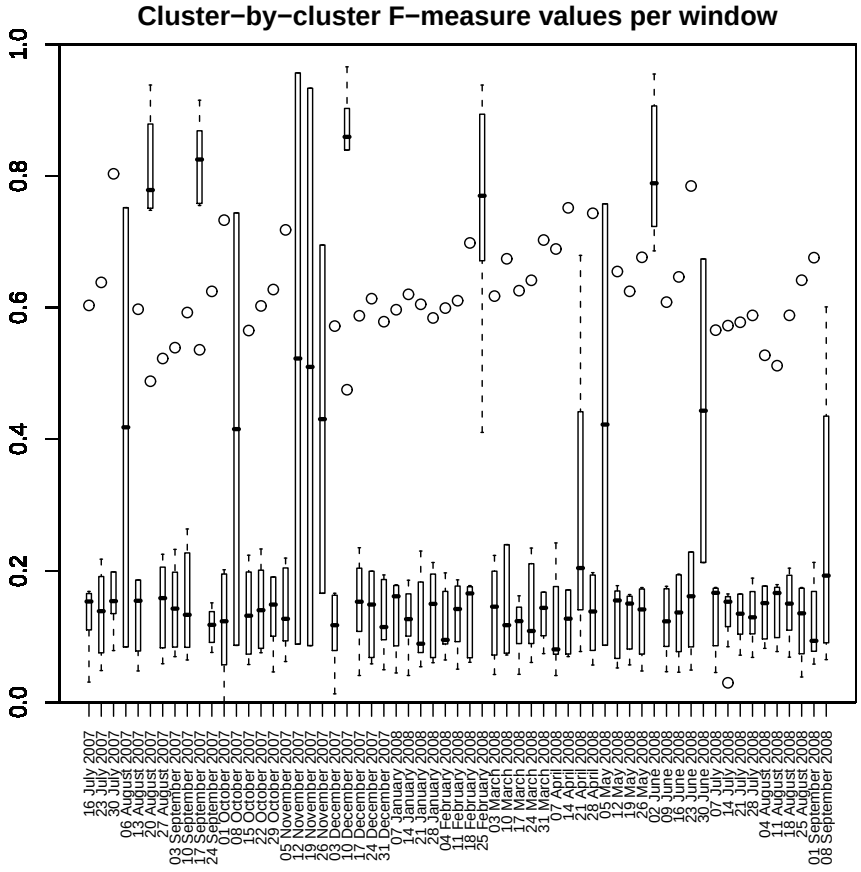

FIGURE 3. *Valeurs obtenus par la F-mesure pour chaque semaine analysée.*

ayant des habitudes d'achat distinctes passent à partager les mêmes préférences d'achat. Un exemple typique concerne le phénomène vérifié lors de la mise en place des promotions sur produits qui attirent différents clients autres que le public cible.

– **Pas de changement** : dans ce dernier cas, le cluster de comportement est considéré comme survivant dans la fenêtre de temps suivante. Il peut cependant subir des changements liés au nombre d'effectifs (rétrécissement ou grossissement). Ces changements peuvent indiquer le changement de popularité de certaines marques de produit.

Ces genres de changements vérifiés sur les données analysées sont décrits dans la figure 5. Pendant nos expérimentations, nous n'avons remarqué aucune scission de clusters de comportement.

Toujours en cohérence avec les résultats obtenus par la F-mesure et l'indice de Rand corrigé, notre méthode a détecté le plus grand nombre de clusters survivants (cf. graphique *Total of survivals* de la figure 5) pendant les mêmes semaines de stabilité trouvées par les deux indices de comparaison de partition. Pendant ces semaines, la quasi totalité des clusters de préférences d'achat a survécu.

En analysant le graphique *Total of appearances* nous remarquons, pour la majorité des semaines analysées, un grand nombre de nouveaux clusters (autour de 6). Le nombre total de clusters par fenêtre étant assez stable et égal

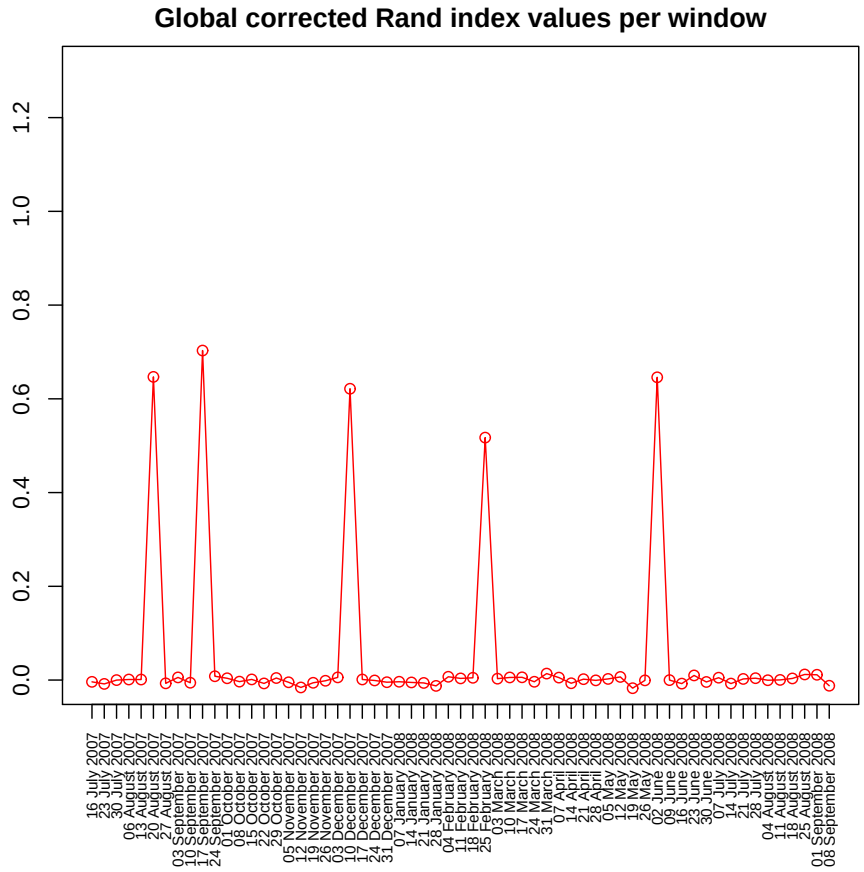

FIGURE 4. *Valeurs obtenus par l'indice de Rand corrigé pour chaque semaine analysée.*

à 7 (cf. figure 2), on pourrait s'attendre à un grand nombre de disparition de clusters. Cependant, en analysant le graphique *Total of disappearances* de la figure 5, nous constatons un faible nombre de disparation de clusters au cours du temps. L'explication de ceci est liée au fait qu'une grande partie des clusters se fondent les uns avec les autres (cf. le graphique *Total of absorptions* de la figure 5) au cours du temps.

Afin de mieux connaître les profils d'achat de ces clients, nous avons tracé sur la figure 6 la distribution des pourcentages d'achat par marque pour les 7 clusters détectés par la méthode pendant les périodes les plus stables, ces clusters représentent les préférences d'achat les plus typiques. Ces profils d'achat peuvent être décrits comme suit :

- **Cluster 1** : clients ayant une préférence d'achat pour la marque B en premier plan et la marque A en deuxième plan.
- **Cluster 2** : profil mixte de clients ayant des fortes préférences pour les marques A et D.
- **Cluster 3** : clients ayant une préférence d'achat majoritaire pour la marque A.
- **Cluster 4** : clients ayant une préférence d'achat pour la marque F en premier plan et les marques A et D en deuxième plan.
- **Cluster 5** : clients ayant une préférence d'achat pour la marque C en premier plan et les marques E et A en deuxième plan.

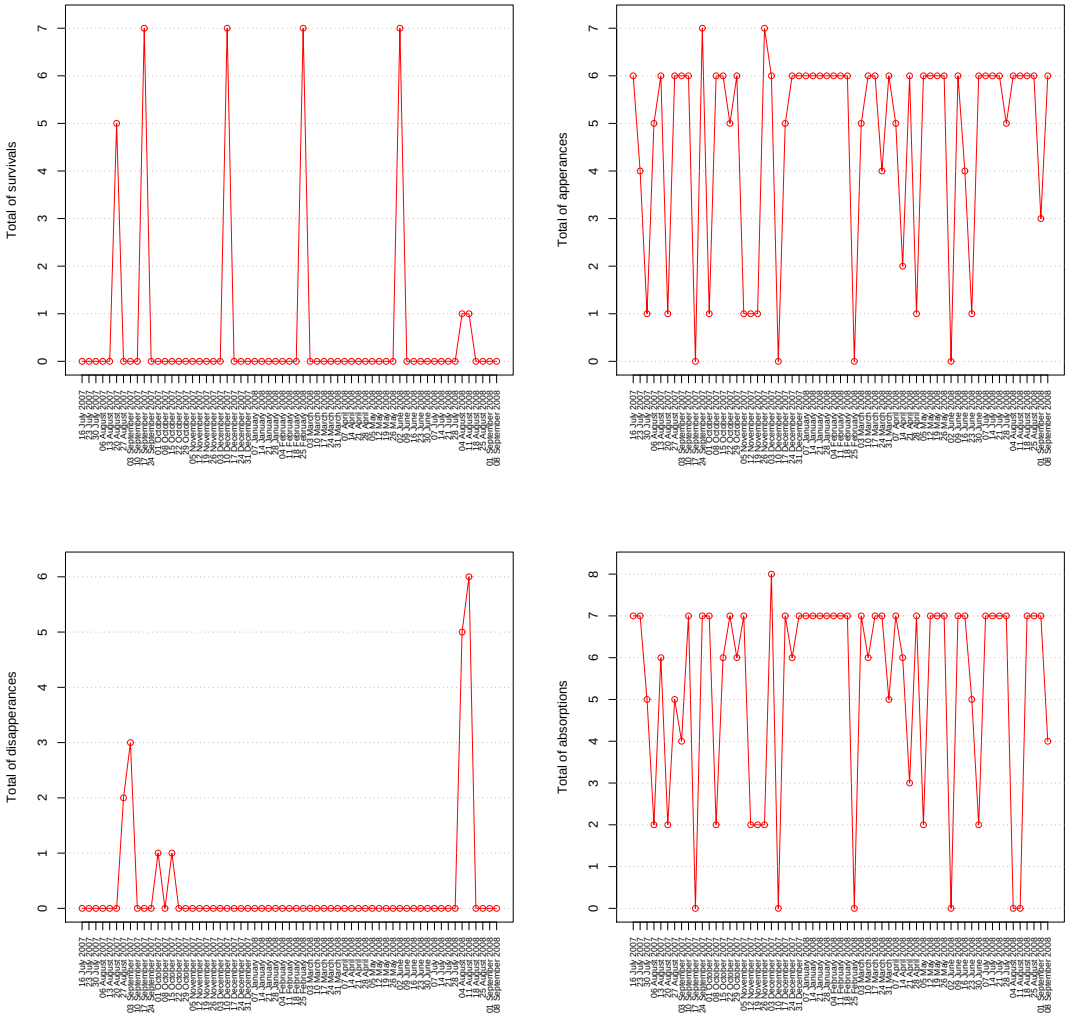

FIGURE 5. *Nombre de changements detectés sur les clusters de comportement au cours du temps.*

– **Cluster 6** : clients ayant une préférence d'achat majoritaire pour la marque D.
– **Cluster 7** : clients ayant une préférence d'achat pour la marque E en premier plan et les marques A et D en deuxième plan.

La marque A a une forte préférence parmi tous les profils d'achat. Les autres marques sont présentes dans des profils ponctuels. En consultant les résultats obtenus, il est possible d'extraire la liste de clients appartenant à un cluster spécifique.

Pour suivre le comportement d'achat d'un client spécifique, il suffit de vérifier si le client en question change de cluster au cours du temps. Si le client reste toujours dans un même cluster au cours du temps, celui si peut être considéré 'fidèle' aux marques concernées par ce cluster. D'un autre côté, si le client change considérablement de cluster au cours du temps, celui si peut être classé tel un 'zappeur' entre les marques de produit.

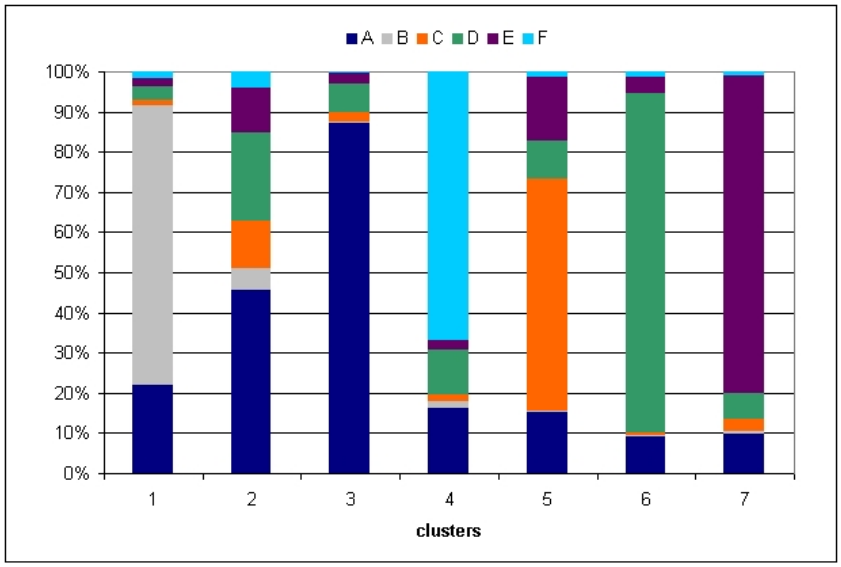

FIGURE 6. *Clusters de préférences d'achat les plus typiques.*

5. Conclusion

Dans cet article nous avons décrit les résultats obtenus par de notre méthode de classification non supervisée pour la détection et suivi des clusters de comportement dans le temps appliquée sur un jeu de données du *marketing* dans le cadre du concours jeunes chercheurs de la SLDS 2009. Nous avons pu suivre l'évolution de clusters de préférences d'achat des clients sur 6 marques de produits de deux marchés différents au long de 14 mois.

Notre méthode a permis la traçabilité des changements subis par les clusters de préférences d'achat des clients ainsi que l'identification de la nature de ces changements au cours du temps (apparition/disparition des groupes de comportement, migration d'individus de et vers un autre groupe de comportements). Notre méthode applique également des tests statiques afin de vérifier si les changements détectés sont statistiquement significatifs.

L'un des enjeux majeurs dans ce genre d'analyse est la nécessité de jongler avec plusieurs changements qui peuvent se passer en simultané sur différents groupes de comportement.

Tous les résultats obtenus par l'application de notre méthode sur le jeu de données en question ont été enregistrés dans une base de données MySQL et peuvent de ce fait faire l'objet a posteriori d'un rapport technique. Il est ainsi possible de fournir à des questions spécifiques des réponses plus détaillées. Par exemple, la segmentation des clients pour une semaine spécifique peut être facilement repérée par une simple requête à la base de données. Il est également possible de mesurer la popularité de certaines marques de produit en fonction du suivi des effectifs d'un même groupe de comportement au cours du temps.

6. Bibliographie

[COO 99] COOLEY R., MOBASHER B., SRIVASTAVA J., Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, vol. 1, n^o 1, 1999, p. 5-32.

[DAS 08] DA SILVA A., LECHEVALLIER Y., Stratégies de classification non supervisée sur fenêtres superposées : application aux données d'usage du Web, *Actes des 8ème journées Extraction et Gestion des Connaissances (EGC 2008), Revue des Nouvelles Technologies de l'Information (RNTI)*, vol. I, cépaduès, 29 January - 1 February 2008, p. 219-220.

[DAS 09] DA SILVA A., LECHEVALLIER Y., DE CARVALHO F., Simulation et détection de l'évolution des données temporelles issues de l'usage du Web, HÉBRAIL G., PONCELET P., QUINIOU R., Eds., *Actes de l'atelier fouille de données temporelles et analyse de flux de données à EGC 2009*, 2009.

[ELE 99] ELEMENTO O., Apport de l'analyse en composantes principales pour l 'nitialisation et la validation de cartes topologiques de Kohonen, *Actes des 7èmes journées de la Société Francophone de Classification (SFC'99)*, Nancy, France, 1999.

[HUB 85] HUBERT L., ARABIE P., Comparing Partitions, *Journal of Classification*, vol. 2, 1985, p. 193–218.

[KOH 95] KOHONEN T., *Self-Organizing Maps*, vol. 30 de *Springer Series in Information Sciences*, Springer, third édition, 1995, Last edition published in 2001.

[RIJ 79] VAN RIJSBERGEN C. J., *Information Retrieval*, Butterworths, London, second édition, 1979.

[SPI 99] SPILIOPOULOU M., Data Mining for the Web, *Workshop on Machine Learning in User Modelling of the ACAI99, ,* 1999, p. 588-589.