Détection de changement dans les profils en ligne d'utilisateurs

Parisa Rastin, Basarab Matei, Guénaël Cabanes

LIPN-CNRS, UMR 7030, Université Paris 13
rastin@lipn.univ-paris13.fr

Résumé. L'analyse des données dynamique est difficile. En effet, la structure de telles données évolue dans le temps, potentiellement à une vitesse très rapide. De plus, les objets dans ces ensembles de données sont souvent complexes. Dans cet article, notre motivation pratique est d'analyser l'évolution des profils en ligne d'utilisateurs, c'est-à-dire de suivre la localisation géographique des utilisateurs ainsi que leurs traces de navigation en ligne afin de détecter des changements dans leurs habitudes et leurs intérêts. Nous proposons un nouveau cadre dans lequel nous créons d'abord, pour chaque utilisateur, des signaux de l'évolution de ses intérêts et des localisations physiques enregistrées au cours de sa navigation. Ensuite, nous détectons automatiquement les changements d'intérêt ou de lieu grâce à un nouvel algorithme de détection de sauts dans les signaux.

1 Introduction

L'un des défis principaux de l'exploration de données est la détection de changement dans les ensembles de données dynamiques. Ce phénomène est connu sous le nom de "dérive de concept" (Gama, 2010; Silva et al., 2013). Une des applications directes, qui constitue notre intérêt pratique dans le présent document, est la détection de changement dans l'intérêt des utilisateurs sur la base des données enregistrées au cours de leur navigation en ligne. Cette tâche, appelée "profilage de l'utilisateur", revêt une grande importance économique pour les entreprises du secteur de la publicité en ligne. Les tâches de profilage visent à reconnaître les "états d'esprit" des utilisateurs à partir de leurs navigations sur différent sites Web ou leurs interactions avec des "points de contact" numériques (différentes façons dont une marque interagit et affiche des informations aux utilisateurs). Il est très important de pouvoir détecter les changements d'intérêt d'un utilisateur ou son déménagement dans une autre ville ou un autre pays afin d'ajuster la stratégie publicitaire le concernant. Ces profils sont calculés à partir d'une très grande base de données de navigation sur Internet, qui liste les séquences d'URL ou les points de contact visités par un grand nombre de personnes. Chaque URL d'un "point de contact" est caractérisée par des informations contextuelles et sémantiques. Dans ce contexte, chaque utilisateur est décrit comme une série temporelle de catégories d'URL et d'emplacements physiques. Les catégories d'URL sont calculées à l'aide d'une approche de classification adaptée aux données complexes (Rastin et al., 2016; Rastin et Matei, 2018). Les emplacements sont enregistrés à l'aide des informations de géolocalisation collectées lors de la navigation de l'utilisateur, mais sont limités à une série de codes postaux. La détection de changements dans les

séries temporelles implique l'extraction de périodes "stables", séparées par une période de variation généralement courte. Il y a donc deux stratégies principales : soit l'algorithme cherche à détecter les différentes périodes de stabilité dans la série chronologique, soit il détecte la période de variation (Last, 2002; Aggarwal et al., 2003; Han, 2005; Cao et al., 2006). La détection de la stabilité ou de l'homogénéité est liée à la tâche de classification des flux de données. Dans cet article, nous considérons une fenêtre temporelle glissante avec un pas d'une journée, afin d'obtenir pour chaque fenêtre une distribution de lieux ou d'intérêts. Nous proposons dans cet article une nouvelle approche basée sur le traitement de signaux, décrite dans la section 2, adaptée à la tâche de profilage. Nous avons ensuite testé l'algorithme sur des données simulées pour valider sa qualité par rapport aux approches traditionnelles ; les résultats sont présentés à la section 3. Enfin, nous avons appliqué le cadre proposé à un jeu de données industrielles réelles, comme indiqué à la section 4. Une conclusion est donnée à la section 5.

2 Proposed approach

Algorithm 1 Détection de changements dans un signal de profil utilisateur

Entrée : Vecteur de signal v de longueur N.
sortie : Liste des changements détectés.

Initialisez $j = [\log_2(N)]$
Initialisez la liste globale des sauts $L_g = \emptyset$
while $j > [\log_2(N)] - 4$ **do**
  Lissage :
  for $i \leftarrow 1, length(v^j)$ **do**
    $v_k^{j-1} = \frac{v_{2k-1}^j + v_{2k}^j}{2}$
  Initialiser la liste locale des sauts $L_e = \emptyset$
  Calcul de la fonction de coût en fonction des différences finies du premier ordre :
  for $k \leftarrow 1, length(v^{j-1})$ **do**
    $dv_k^{j-1} = |\Delta v_k^{j-1}| + |\Delta v_{k+1}^{j-1}|$
  Calcul des maxima locaux de la fonction de coût :
  for $k \leftarrow 1, length(v_k^{j-1})$ **do**
    if $dv_k^{j-1} > \max(dv_{k-2}^{j-1}, dv_{k-1}^{j-1}, dv_{k+1}^{j-1}, dv_{k+2}^{j-1})$ **then**
      $L_j \leftarrow L_j + \{k\}$
  $j \leftarrow j - 1$
Définir L_g comme l'intersection de tous les L_j : $L_g = \cap_{j=[\log_2(N)]-4}^{[\log_2(N)]} L_j$

Afin de détecter les changements de profil des utilisateurs, nous avons appliqué l'algorithme de détection de changement décrit ci-dessous. Cet algorithme détecte des "sauts" inhabituels dans un signal caractérisant des variations de profil d'un utilisateur. Pour construire un tel signal, nous avons défini comme profil de référence la distribution d'étiquettes ou de codes postaux dans les premières fenêtres temporelles. Ensuite, la fenêtre est décalée d'un jour à la fois, afin de produire une série de distribution. La similarité entre deux distributions de

probabilité (fenêtre de référence et fenêtres décalées) est calculée par la divergence de Jensen-Shannon (JS) (Manning et Schütze, 1999; Dagan et al., 1997), une version symétrisée et lissée de la divergence de Kullback-Leibler $D(P \parallel Q)$ entre deux distributions discrètes. Notez que toutes les probabilités égales à P ou Q sont ignorées dans le calcul, ce qui signifie que deux distributions totalement différentes auront une valeur JS de 1. L'approche proposée a été testée sur des ensembles de données artificielles pour validation, puis appliquée sur les ensembles de données réels pour analyser les changements de profil et d'état d'esprit des utilisateurs. L'algorithme 1 décrit l'approche de détection de changement multi-échelles. L'idée est la suivante : un processus de lissage itératif élimine les fluctuations aléatoires du signal (lignes 5 et 6), puis détecte des variations anormalement élevées (lignes 12 à 14). Les signaux sont des fonctions continues fragmentées présentant des discontinuités à certains emplacements x_i, c.-à-d. $v(x_i^+) \neq v(x_i^-)$. Nous considérons ici que v_k^j sont les moyennes d'une fonction v discrétisée sur les intervalles $I_{j,k} = 2^{-j}[k, k+1[$. Dans une approche multi-échelle basée sur des coefficients, une stratégie pour détecter les singularités au niveau j est basée sur un critère qui utilise les différences de premier ou de second ordre de v^j, la détection des singularités de saut est effectuée à chaque niveau indépendamment. Nous calculons ensuite le nombre N_j de singularités au niveau j et définissons $j_{\max}$ comme le plus grand niveau j tel que $N_{j-1} = N_j$. Nous définissons également le niveau $j_{\min}$ comme le plus petit j tel que $N_j = N_{j_{\max}}$. Une singularité détectée dans $I_{j,k}$ pour $j_{\min} < j < J$ est dite recevable s'il existe une singularité dans $I_{j+1,2k}$ ou $I_{j+1,2k+1}$.

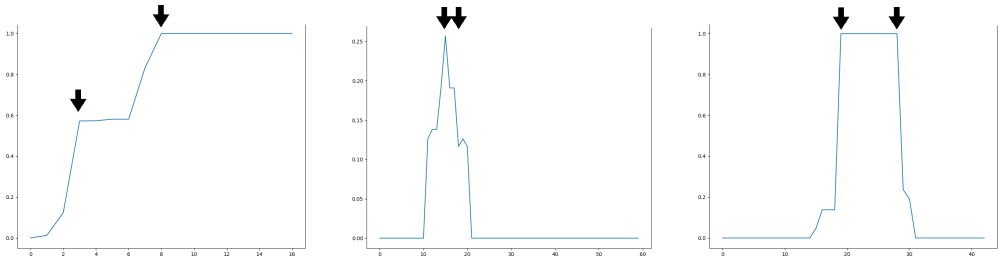
3 Validation expérimentale

Pour valider la qualité de notre algorithme dans un environnement contrôlé, nous l'avons testé sur des ensembles de données artificielles. Pour générer ces données, nous avons considéré trois catégories de variations de profils : soit le profil de l'utilisateur change avec le temps en un profil totalement nouveau, soit il devient partiellement différent, soit il reste stable. Nous avons généré 10000 signaux pour chacune de ces catégories. Pour construire un signal, nous avons d'abord généré deux ensembles de 1 à 5 étiquettes aléatoires, chacun représentant des profils possibles avant et après le changement. Un seul ensemble est créé pour simuler l'absence de changement. Pour simuler un changement partiel, nous avons forcé les deux ensembles à partager 1 ou 2 étiquettes. Nous avons simulé une période de deux mois. Cent "time-stamps" aléatoires ont été générés au cours de cette période, chacun associé à une étiquette du premier ou du second ensemble, selon une date de changement choisie au hasard. Pour démontrer l'efficacité de l'approche proposée, nous avons évalué ses performances en termes de temps de calcul et avons calculé les moyennes des différences absolues entre la date de modification détectée et prévue et les avons comparées à un ensemble d'algorithmes de l'état de l'art : Jump penalty, PWC bilateral, Robust jump penalty, Soft mean-shift, Total variation, Robust TVD et Medfiltit (Little et Jones, 2011). Le tableau 3 présente les résultats de la comparaison. Les bonnes performances de l'algorithme proposés sont dûs à l'utilisation d'une fonction de lissage bien adaptée à la fonction de coût et surtout aux applications successives de ce lissage de façon itérative, permettant de détecter des variations stables dans le signal. Le processus est par ailleurs peu complexe, ce qui explique les faibles temps de calculs observés.

Algorithmes	Profil stable		Changement complet		Changement partiel	
	Temps (s)	Erreur	Temps(s)	Erreur	Temps(s)	Erreur
Proposé	0.85	2.78	0.94	1.67	0.87	2.07
Jump penalty	29.4	14.33	25.26	4.11	31.93	4.47
PWC bilateral	83.87	14.31	12.83	3.77	18.71	4.19
Robust penalty	8.43	14.26	93.48	4.02	90.63	4.57
Soft mean-shift	8.57	16.57	21.99	3.6	21.5	4.56
Total variation	45.55	13.12	103.44	3.27	116.8	4.02
Robust TVD	7016.69	14.82	4405.04	3.8	4390.13	4.61
Medfiltit	1.32	13.92	0.98	2.95	1.29	3.93

TAB. 1 – *Résultats expérimentaux*

4 Application

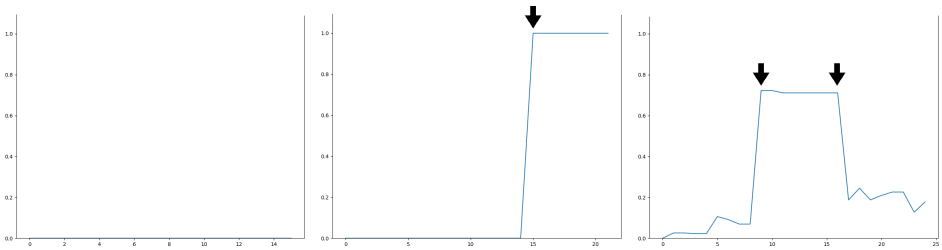


(a) Déménagement (b) Séjour court (c) Séjour long

FIG. 1 – *Exemple de signaux obtenus lors du déménagement d'un utilisateur ou lors de déplacements temporaires. Les flèches indiquent les changements détectés*

Afin de suivre l'évolution des habitudes de déplacement des utilisateurs, la géolocalisation (codes postaux) associées à des horodatages sur une période de 74 jours pour 598 utilisateurs ont été utilisés. L'objectif de ces données est de pouvoir détecter le moment où un utilisateur déménage dans un endroit différent ou passe du temps en dehors de sa zone habituelle. Lors de la création du signal, nous avons utilisé une fenêtre de 10 jours. Dans Figure 1.a, la dissimilarité de Jensen-Shannon augmente fortement pendant deux jours, reste stable pendant trois jours, puis augmente à nouveau. Deux changements sont détectés, le premier étant un changement partiel. Ce type de signal peut être interprété comme un mouvement en deux étapes, avec une période pendant laquelle l'utilisateur passe du temps aux deux endroits avant de se déplacer définitivement. Un autre cas intéressant est celui où l'utilisateur part en vacances ou pour son travail quelque temps, avant de retourner à son lieu de résidence habituel. Les figures 1.b et 1.c montrent deux exemples pour ce cas.

Pour suivre les changements réels d'intérêt individuel, nous avons utilisé un ensemble de données du journal de navigation de 142794 utilisateurs, fournissant à chaque utilisateur une liste des "time-stamps" associés à l'URL visitée à ce moment, sur une période de 30 jours.



(a) Intérêt stable (b) Changement d'intérêt (c) Changement temporaire

FIG. 2 – *Exemples d'intérêt stable, de changement d'intérêt et de changement temporaire d'intérêt des utilisateurs, en fonction de leurs journaux de navigation.*

Chaque URL a été associée à une classe prédéfinie et la navigation de l'utilisateur peut être exprimée dans une distribution de classes visitées variant dans le temps. Les figures 2.a à 2.c illustrent différents scénarios de changement d'état d'esprit de l'utilisateur. Figures 2.a est un utilisateur qui ne change pas d'intérêt pendant un mois. La figure 2.b est un exemple de résultat pour la détection d'un changement d'intérêt individuel, l'utilisateur modifiant son intérêt au cours du temps. Une troisième catégorie d'état d'esprit observé est un groupe d'utilisateurs qui changent d'intérêt pour une période limitée puis reviennent à leur intérêt initial. La figure 2 .c illustre ce type d'utilisateurs. Comme vous le voyez, ces signaux montent et restent stables sur une période de temps puis diminuent. Cela signifie que la différence entre la fenêtre de référence et les fenêtres décalées augmente pendant un certain temps, mais qu'à la fin de la période enregistrée, la distribution des catégories d'URL visitées revient à une distribution similaire à la distribution de référence.

5 Conclusion

Dans cet article, nous avons proposé un nouvel algorithme multi-échelles de détection de changement pour analyser les variations de profils individuels des utilisateurs en fonction de leurs données de navigation et de géolocalisation. Nous avons d'abord créé, pour chaque utilisateur, un signal de l'évolution de la répartition de l'intérêt des utilisateurs en ligne et un autre signal basé sur la distribution des emplacements physiques enregistrés au cours de leur navigation. Ensuite, nous avons proposé un algorithme de détection de sauts capable de détecter automatiquement les changements. Nous avons détecté différents scénarios : au cours de la période analysée, certains utilisateurs ont conservé le même profil, certains ont eu un changement net de profil et d'autres n'ont montré qu'un changement provisoire. Les tests expérimentaux effectués sur des signaux simulés ont montré que l'approche proposée est plus rapide et fait moins d'erreurs pour cette tâche que les algorithmes de l'état de l'art.

Références

Aggarwal, C. C., J. Han, J. Wang, et P. S. Yu (2003). A framework for clustering evolving data streams. In *Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29*, VLDB '03, pp. 81–92. VLDB Endowment.

Cao, F., M. Estert, W. Qian, et A. Zhou (2006). *Density-Based Clustering over an Evolving Data Stream with Noise*, pp. 328–339. Society for industrial and applied mathematics.

Dagan, I., L. Lee, et F. Pereira (1997). Similarity-based methods for word sense disambiguation. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, EACL '97, pp. 56–63.

Gama, J. (2010). *Knowledge Discovery from Data Streams* (1st ed.). Chapman & Hall/CRC.

Han, J. (2005). *Data Mining : Concepts and Techniques*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Last, M. (2002). Online classification of nonstationary data streams. *Intell. Data Anal. 6*(2), 129–147.

Little, M. A. et N. S. Jones (2011). Generalized methods and solvers for noise removal from piecewise constant signals. i. background theory. *Proc. R. Soc. A 467*(2135), 3088–3114.

Manning, C. D. et H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA : MIT Press.

Rastin, P. et B. Matei (2018). Prototype-based Clustering for Relational Data using Barycentric Coordinates. In *Proceeding of the International Joint Conference on Neural Networks (IJCNN)*, IJCNN'18.

Rastin, P., T. Zhang, et G. Cabanes (2016). *A New Clustering Algorithm for Dynamic Data*, pp. 175–182. Cham : Springer International Publishing.

Silva, J. A., E. R. Faria, R. C. Barros, E. R. Hruschka, A. C. P. L. F. de Carvalho, et J. Gama (2013). Data Stream Clustering : A Survey. *ACM Comput. Surv. 46*(1), 13–31.

Summary

The analysis of dynamic data is challenging. Indeed, the structure of such data changes over time, potentially in a very fast speed. In addition, the objects in such data-sets are often complex. In this paper, our practical motivation is to perform users profiling, i.e. to follow users' geographic location and navigation logs to detect changes in their habits and their interests. We propose a new framework in which we first create, for each user, a signal of the evolution in the distribution of their interest and another signal based on the distribution of physical locations recorded during their navigation. Then, we detect automatically the changes in interest or locations thanks a new jump-detection algorithm. We compared the proposed approach with a set of existing signal-based algorithms on a set of artificial data-sets and we showed that our approach is faster and produces less errors for this kind of task. We then applied the proposed framework on a real data-set and we detected different categories of behavior among the users, from users with very stable interest and locations to users with clear changes in their behaviors, either in interest, location or both.