Flux de Données contre Séries Temporelles

Pierrick Bruneau
CRP-GL
Département ISC
41, rue du Brill
L-4422, Belvaux
bruneau@lippmann.lu

Olivier Parisot
CRP-GL
Département ISC
41, rue du Brill
L-4422, Belvaux
parisot@lippmann.lu

Benoit Otjacques
CRP-GL
Département ISC
41, rue du Brill
L-4422, Belvaux
otjacque@lippmann.lu

RÉSUMÉ
Les usages informatiques actuels génèrent des séquences de données complexes et volumineuses (e.g. capteurs, réseaux sociaux). Pour les traiter, deux paradigmes s'imposent naturellement : considérer ces données comme des séries temporelles, et/ou comme des flux de données (i.e. *data streams* dans la littérature). Bien que semblables, ces deux principes diffèrent de manière assez sensible. Nous pensons que leur combinaison avec des techniques de visualisation et d'interaction fait émerger de nouvelles problématiques, qui sont autant de pistes de contributions intéressantes.

Mots Clés
data streams; séries temporelles; data mining; visualisation

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous. Voir `http://www.acm.org/about/class/1998/` pour la liste complète des catégories ACM.

PRÉSENTATION
Au sein du département "Informatique, Systèmes, Collaboration" (ISC) du Centre de Recherche Public - Gabriel Lippmann (CRP-GL), notre équipe dédie sa recherche au développement d'approches *visual analytics*, i.e. de combinaisons pertinentes entre méthodes de visualisation d'information et de data mining. Ces deux communautés sont historiquement assez décorrélées, et notre objectif est de proposer des travaux prenant en compte l'état de l'art joint de ces deux perspectives.

Cet axe est développé dans le contexte de partenariats avec des organismes publics ou privés, voire de la participation à des projets de recherche nationaux et internationaux (e.g. FNR, FP7). En parallèle, le CRP-GL dispose de fonds publics pour son développement stratégique, dont la prospection dans notre axe de recherche fait partie intégrante.

Le CRP-GL a vocation à la recherche appliquée, aussi nous sommes toujours à l'écoute de cas d'utilisation potentiels pour notre thématique de recherche ; toutefois une partie de notre effort est également consacré à une réflexion plus fondamentale.

Notre équipe travaille sur des contributions aux domaines des *visual analytics* et du *data mining* visuel, avec en ligne de mire l'adaptation au traitement de flux de données. En parallèle d'une réflexion fondamentale sur ces sujets, nous sommes impliqués dans divers projets collaboratifs, dont un projet européen CHIST-ERA, visant à la conception d'une plateforme web collaborative de traitement et de visualisation d'annotations multimédia (Camomile), ainsi qu'une collaboration avec le département "Environnement et Agro-Biotechnologies" du CRP-GL, pour l'analyse de données réelles issues de la création d'énergie renouvelable.

PROBLÉMATIQUES D'INTÉRÊT EN DATA MINING
Les flux de données engendrés par certaines plateformes web (e.g. Twitter, métriques Google), voire par des capteurs individuels désormais peu onéreux soulèvent de nombreuses problématiques scientifiques. En particulier, un des objectifs stratégiques de notre département est de faire avancer l'état de l'art sur le traitement de ces flux dans un contexte *visual analytics*.

L'analyse des séries temporelles est utilisée classiquement dans de nombreux domaines (e.g. traitement du signal, économétrie, météorologie). Ici, une série de valeurs (ou de vecteurs) va être utilisée "hors ligne" pour apprendre un modèle (e.g. de prédiction, ou de détection de motifs dans des cours de bourse) [4, 3]. Une fois appris, le modèle est figé, et ne peut pas tirer parti des nouvelles données arrivant dans le flux sans un réapprentissage complet. Pire, la complexité d'apprentissage de ces modèles dépend souvent de la taille du flux, au mieux linéairement. Le réapprentissage serait donc de plus en plus couteux au fil du temps. Notons que des travaux considèrent le problème "dual", i.e. compte-tenu d'un ensemble de séries temporelles en flux, caractériser des similarités entre elles, e.g. rechercher des motifs temporels partagés avec l'algorithme *Dynamic Time Warping* (DTW) [1], ou encore estimer les similarités deux-à-deux de séries temporelles à plusieurs niveaux de résolution [8]. Cependant les problèmes de complexité évoqués plus haut sont patents dans ce cas également.

Certains de ces algorithmes sont incrémentaux, donc capables en théorie de s'adapter à un flux de taille

croissante en utilisant un minimum de ressources calculatoires supplémentaires [8, 12]. Toutefois, comme cela a été constaté dès la dérivation de la version incrémentale de l'algorithme classique k-means [9], l'ajout perpétuel de nouveaux éléments fait que l'importance des derniers arrivants décroit graduellement, jusqu'à devenir négligeable : on aurait alors un modèle progressivement figé *de facto*. Des auteurs ont proposé de restreindre le calcul de statistiques exhaustives dans une fenêtre glissante dans le contexte d'un algorithme EM [11], mais la question du paramétrage de la taille de cette fenêtre n'a pas été abordée, et semble peu évidente. Des études théoriques sur les facteurs d'oubli ont été également proposées [10], mais des contributions restent à faire pour en évaluer la valeur pratique.

Le paragraphe précédent a effleuré un dilemme qu'on peut caractériser comme suit : quel est le meilleur compromis entre la stabilité d'un modèle, et sa capacité à s'adapter ? Certains travaux récents traitent de cette question, par exemple [2], en tentant de caractériser et détecter le changement de manière non-paramétrique, s'affranchissant ainsi notamment du problème de taille de fenêtre glissante évoqué précédemment.

TRANSCRIPTION EN TERMES DE VISUALISATION ET D'INTERACTION

Les questions présentées dans la section précédente, plutôt relatives au *data mining*, ont des transcriptions diverses dans le contexte de la visualisation d'information, et de l'interaction avec cette dernière.

Le dilemme classique *biais-variance* [6] peut être transposé à la visualisation des séries temporelles numériques : celles-ci peuvent présenter des profils de courbe très bruités, rendant leur visualisation difficile, à petite ou à grande échelle. A contrario, un lissage trop important (e.g. via une fenêtre de Parzen) risque de détruire l'information intrinsèque d'une courbe temporelle.

Un flux de données est potentiellement infini, et il faut pourtant parvenir à le représenter en utilisant une bande passante visuelle finie : des moyens de résumer visuellement les données "périmées" du flux doivent être trouvés, faisant ainsi écho au facteur d'oubli de la section précédente. Des travaux traitent de la visualisation efficace de séries temporelles [5], mais en considérant rarement un tel facteur d'oubli.

Maintenant imaginons qu'il existe un modèle capable de s'adapter de manière satisfaisante à la nature locale du flux. Comment le visualiser, et interagir avec lui, sachant qu'à l'instar du flux de données, il évoluera également dans le temps ? Ces problèmes se rapprochent également du domaine du *storytelling*, pour lequel il existe un compromis à réaliser entre la liberté d'interaction et une restitution efficace de l'information [7].

BIBLIOGRAPHIE

1. Berndt, D., and Clifford, J. Using dynamic time warping to find patterns in time series. *AIII Workshop on KDD* (1994), 359–370.
2. Bondu, A., and Boullé, M. A supervised approach for change detection in data streams. *International Joint Conference on Neural Networks* (2011), 519–526.
3. Durbin, J., and Koopman, S. J. *Time Series Analysis by State Space Methods*. Oxford University Press, 2001.
4. Gardner, G., Harvey, A. C., and Phillips, G. D. A. Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Applied Statistics* (1980), 311–322.
5. Heer, J., Kong, N., and Agrawala, M. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009), 1303–1312.
6. James, G. Variance and bias for general loss functions. *Machine Learning 51* (2003), 115–135.
7. Kosara, R., and Mckinlay, J. Storytelling: The next step for visualization. *IEEE Computer 46*, 5 (2013), 44–50.
8. Lin, J., Vlachos, M., Keogh, E., and Gunopulos, D. Iterative incremental clustering of time series. *Advances in Database Technology (EDBT 2004). LNCS 2992* (2004), 106–122.
9. McQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), 281–297.
10. Sato, M. Online model selection based on the variational bayes. *Neural Computation 13*, 7 (Jul 2001), 1649–1681.
11. Smidl, V., and Quinn, A. *The variational Bayes method in signal processing*. Springer, 2006.
12. Toyoda, M., Sakurai, Y., and Ishikawa, Y. Pattern discovery in data streams under the time warping distance. *The VLDB Journal 22*, 3 (2013), 295–318.