MMS Explore : un outil de visualisation interactive pour l'analyse qualité de flux données temporelles

Zied Ben Othmane*, Cyril De Runz**
Amine Ait Younes**, Vincent Mercelot*

*KantarMedia, 2 Rue Francis Pédron, 78241 Chambourcy - France
{zied.benothmane, vincent.mercelot}@kantarmedia.com,
http://www.kantarmedia.com
**CReSTIC, MODECO, Université de Reims Champagne-Ardenne
Chemin des Rouliers CS30012 51687 REIMS CEDEX 2
{cyril.de-runz, amine.ait-younes}@univ-reims.fr
http://crestic.univ-reims.fr/

Résumé. Les flux données web récoltés par les robots doivent avoir un haut niveau de véracité pour pouvoir déterminer des connaissances précises. Aussi, analyser leur qualité est primordial à la vue notamment des imperfections intrinsèques aux données. Dans cet article, nous présentons un outil de visualisation interactive permettant d'analyser la qualité de ces flux temporels.

1 Introduction

Les données récoltées par les capteurs sont de plus en plus mises en question au vue de la difficulté de vérifier leurs sens. Cette dernière est due à la grande masse de données récoltées mais aussi à cause d'autres facteurs comme la temporalité, l'incertitude (Ba et al., 2016), l'incomplétude, etc. L'information traitée à partir de ces données peut alors être mise en cause. Ainsi la détermination de connaissances précises et fiables nécessite donc un système vérifiant leur véracité (Lukoianova et Rubin, 2014).

Nous sommes dans un contexte de récolte des données web. Nous nous focalisons sur l'analyse des fichiers de log spécifiques à la récolte des bannières publicitaires. Le système de recueil enregistre a priori toutes les informations qui tournent autour de la publicité sur une page web (e.g. le nom du site, l'url, la catégorie, la méta-catégorie, etc.). Cet axe de travail est au départ d'une chaîne de traitements non abordée dans cet article. L'information finale dépend fortement des données d'entrées récoltées par les capteurs.

Dans ce contexte, il est important d'avoir un système vérifiant leur qualité. Nous souhaitons développer un prototype pour l'analyse de la qualité de ces données en tenant compte de leurs spécificités. Nos données sont imprécises. En effet, le nombre des bannières récoltées dans une période du temps peut varier sans forcément que le nombre réel de bannières affichées varie. Cela peut être dû à des bogues de développement, à l'emploi de technologies visant à gêner la récolte des données, etc. De plus, nous avons observé des discontinuités remarquables dans la récolte en fonction des sites.

Notre modèle d'analyse doit donc prendre en considération toute la complexité des données. Afin d'explorer les informations fournies par notre modèle, nous proposons un système visuel interactif. Dans cet article, nous présentons dans un premier temps les approches utilisées pour la définition de notre modèle. Nous présentons ensuite notre outil, appelé MMS Explore, qui met en œuvre les indicateurs nécessaires pour évaluer la qualité des flux temporels.

2 Approches intégrées

Afin d'analyser la qualité du recueil, notre outil exploite les approches et notions suivantes (pour les définitions formelles, cf. Z. b. Othmane (2018)) :

— Quantiles : Nous proposons de projeter les données dans leur quantile par rapport à un ensemble étudié. C'est une approche qui apporte un nouveau mode de comparaison des données en les passant du volume brute au volume relatif . Le volume relatif (quantiles) donne naissance à des chronologies de positionnement temporel interne et externe du recueil (Utkin et al., 2014). Le positionnement interne d'un média permet une comparaison de ses données à un instant vis à vis de ses données connues lors de l'analyse. On positionne la donnée récoltée à un instant t par un capteur donné dans l'ensemble des données récoltées connues par ledit capteur. Le positionnement externe correspond à la position d'une donnée à un temps t par rapport à toutes les autres données d'une sélection définie. On positionne donc la donnée par rapport à l'ensemble des autres données récoltées par les capteurs sélectionnés (Destercke et al., 2015). Ces deux informations nous permettent de construire des indicateurs de cohérence et de variabilité du flux analysé. Pour l'absence d'une donnée, l'approche lui affecte une position distinct dite Q_O. Cette nouvelle position reflète l'ignorance totale sur la valeur de la récolte (absence/non-existence). Cette approche permet de couvrir de l'imprécision des données et de gérer l'incertitude résultante en les représentant dans un domaine approprié (Cappiello, 2015). Le choix des paramètres de ce domaine tel que le nombre de quantile, l'échelle, etc. sont paramétrables dans notre outil (Z. b. Othmane, 2018).

— Variabilité : la variabilité est définie selon deux versions permettant de calculer un score par rapport au mouvement du recueil d'une sélection. Ces définitions permettent d'avoir des scores relatifs au mouvement interne et externe, i.e. sur la volatilité d'un mouvement vis à vis du reste mais aussi sur sa cohérence intrinsèque. Les scores engendrés nous ont permis de détecter de possibles problèmes au niveau du recueil. Ces derniers sont représentés dans notre outil par des points singuliers pouvant mettre en évidence l'existence de possibles dysfonctionnements (anomalies) au niveau des capteurs.

— Stabilité : La stabilité est la mise en relation de la variabilité externe avec la variabilité interne (dite cohérence). Son objectif est de quantifier la stabilité totale, i.e. une mesure globale de la qualité du fonctionnement d'un capteur. Ce score peut se présenter comme une indicateur statique déterminé suite à une mesure sur un temps précis ou sur une catégorie précise ou bien pendant une période de temps. Ce quantifieur est défini pour un média (un capteur), une variable étudiée, et une période.

— La discontinuité de la récolte : Nous sommes dans un contexte des données temporelles volumineuses, la détection des arrêts et reprises de la récolte d'un ensemble des capteurs est délicat (Cappiello et al., 2018). Notre outil intègre des graphiques aidant

à juger la continuité des enregistrements (présence/absence par exemple). Un système d'alerte allant sur 4 positions complète le dispositif. Il indique le niveau d'urgence, i.e. le degré de la nécessité d'une intervention rapide.

3 Présentation de l'outil

Notre outil de visualisation interactive exploite les données de logs organisées par variables d'étude. La partie *Back-end* s'occupe de la préparation des données et indicateurs nécessaires à l'interface visuelle interactive contenant les tableaux de bords appropriés (partie *Front-end*). Ces deux parties interagissent ensemble selon les choix interactifs de l'utilisateur. Le logiciel dans ses deux parties peut appeler des scripts externes pour affiner les analyses.

L'outil propose divers *dashboards*, complémentaires et interagissant entre eux, qui intègrent les approches d'analyses cités ci-dessus pour valider la qualité d'un recueil. L'interactivité assurée par notre outil permet de mieux comprendre la qualité de la récolte en permettant une exploration profonde.

Afin de permettre à l'utilisateur d'avoir une idée globale sur la qualité des flux temporels, notre outil propose des :

— Visualisations en graphiques ordinaires : des graphiques permettant une navigation dans divers axes d'analyse.
— Visualisations des mesures statistiques : des mesures qui se changent automatiquement à la suite d'une sélection ou un axe de travail.
— Visualisations binaires : un mode de représentation d'existence/absence de la donnée selon un axe de temps ou une catégorisation appropriée.
— Visualisations en méta-plan : un type de visualisation informant sur les positionnements des données.
— Visualisations analytiques : appels externes à des scripts de fouille de données offrant des visualisations analytiques.

Notre outil propose principalement deux types d'indicateurs : les indicateurs statiques visant à fournir une information précise par un score bien défini (voir Fig. 2), et les indicateurs dynamiques qui apparaissent dans des flux d'informations progressifs, e.g. indicateur de stabilité de n médias sur une période de temps (voir Fig. 1).

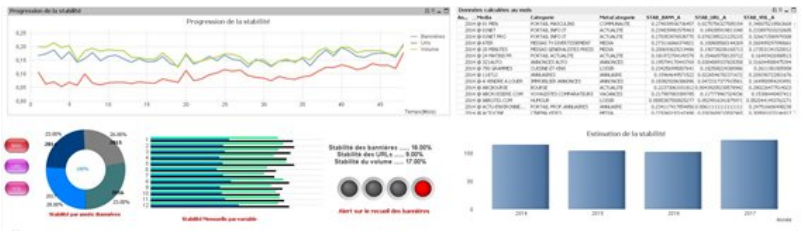

FIG. 1 – *Information sur la stabilité de la récolte*

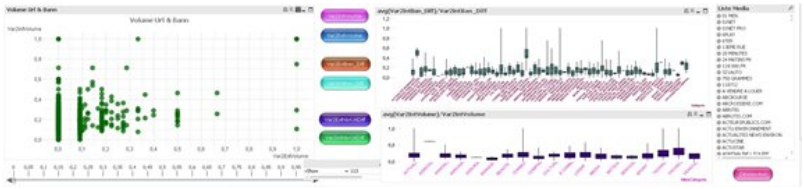

FIG. 2 – *Information sur la variabilité de la récolte*

Références

Ba, M. L., L. Berti-Equille, K. Shah, et H. M. Hammady (2016). Vera : A platform for veracity estimation over web data. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, Republic and Canton of Geneva, Switzerland, pp. 159–162. International World Wide Web Conferences Steering Committee.

Cappiello, C. (2015). On the role of data quality in improving web information value. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA, pp. 1433–1433. ACM.

Cappiello, C., W. Samá, et M. Vitali (2018). Quality awareness for a successful big data exploitation. In *Proceedings of the 22Nd International Database Engineering & Applications Symposium*, IDEAS 2018, New York, NY, USA, pp. 37–44. ACM.

Destercke, S., M.-H. Masson, et M. Poss (2015). Cautious label ranking with label-wise decomposition. *European Journal of Operational Research 246*(3), 927 – 935.

Lukoianova, T. et V. Rubin (2014). Veracity roadmap : Is big data objective, truthful and credible ? *Advances in Classification Research Online 24*(1), 4–15.

Utkin, L. V., Y. A. Zhuk, et I. A. Selikhovkin (2014). An imprecise model of combining expert judgments about quantiles. *European Journal of Technology and Design* (1), 49–60.

Z. b. Othmane, C. d. Runz, A. a. Y. e. D. B. (2018). A multi-sensor visualization tool for harvested web information: Insights on data quality. *iV2018 - 22st International Conference on Information Visualisation 22*, 10–13.

Summary

The web data streams collected by the robots must have a high level of veracity to be able to determine precise knowledge. Also, to analyze their quality is essential, especially in view of imperfections intrinsic to the data. In this article, we present an interactive visualization tool to analyze the quality of these temporal flows.