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 Marcelot*

*

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.