Vers une nouvelle interface visuelle dédiée à l'analyse des récoltes multisources de données

Zied Ben Othmane, Damien Bodénès, Amine Aït-Younes, Cyril de Runz

Vers une nouvelle interface visuelle dédiée à l'analyse des récoltes multisources de données

Zied Ben Othmane [*,**], Damien Bodénès[**]
Amine Aït-Younes[*], Cyril de Runz[*]

[*]CReSTIC/MODECO, Université de Reims Champagne-Ardenne, 51687 Reims Cedex 2
zied.ben-othmane@etudiant.univ-reims.fr, {amine.ait-younes,cyril.de-runz}@univ-reims.fr,
[**]Kantar Media, Rue Francis Pedron, 78240, Chambourcy
{Zied.Benothmane,damien.bodenes}@kantarmedia.com

Dans l'objectif d'étudier les investissements publicitaires sur internet, la société Kantar Media a mis en place un ensemble d'outils de récolte de données (*crawlers*) pour récupérer différentes données sur les publicités affichées sur un ensemble de sites. Ces outils fournissent des données largement imparfaites du fait de la non exhaustivité possible de la récolte, de la stratégie d'affichage des publicités par les sites, etc.. Cela amène à une première question : l'information stockée est-elle légitime ? Il y a donc, à ce jour, au minimum, un besoin de modèles d'estimation de la véracité de cette information.

Dans ce travail nous nous questionnons principalement sur la qualité en essayant de fournir un outil d'analyse visuelle des récoltes effectuées guidées par les données récoltées. L'objectif est d'aider à déterminer les biais possibles dans les récoltes. La visualisation ayant montré son intérêt pour l'analyse des grands volumes de données (**??**), nous nous positionnons dans le cadre d'une démarche de visualisation guidée par les données.

Nous ne nous positionnons pas dans ce premier travail sur l'évaluation de la qualité de données via un outil de visualisation basée sur une analyse robuste des données. Pour cela nous différencions dans un premier temps deux cas :

— l'absence de données récoltées sur les publicités sur un site qui peuvent être dues à plusieurs facteurs : arrêt volontaire de la récolte interne, changement de stratégie du site vis à vis des crawlers, arrêt par un des fournisseurs de données de la récolte sur ce site, etc. ;
— la présence des données qui sont elles mêmes soumises à plusieurs facteurs réduisant leur qualité : impossibilité de récoltes permanentes, changement des stratégies de récoltes, etc.

Afin de mettre en évidence ces deux cas, nous avons développé un premier outil exploratoire permettant une visualisation booléenne (présence/absence de données) des récoltes par site (cf. figure **??**a). Ce premier travail a permis à la société de prendre conscience de certains biais dans la récolte interne des données ; e.g. présence de discontinuité à l'échelle du mois voire du trimestre alors que l'on pensait le flux continu à ces échelles. Nos données de récoltes sont agrégées et définies sur quatre variables volumiques. Nous proposons d'analyser les flux de données à l'échelle du mois non pas par leur valeur intrinsèque mais par leur valeur vis-à-vis des autres. Pour cela, nous affectons chaque valeur à son quartile (statistique robuste) évalué

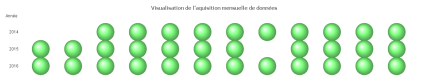
a)  b)

FIG. 1 – *Visualisation par mois de la récolte de données pour un site pour une variable : a) vision booléeene de la présence ou non, b) vision par quartiles calculés par rapport aux valeurs récoltées pour l'ensemble des sites et par mois*

sur les valeurs collectées par mois et par variable : -1 pour l'absence d'information, 1 pour les données inférieures au premier quartile, 2 pour entre le premier quartile et la médiane, 3 entre la médiane et le troisième quartile et enfin 4 pour les données supérieures au troisième quartile. La figure **??**b présente la représentation graphique des 36 mois de récolte pour un site selon les quartiles pour une variable données.

Nous obtenons dès lors pour un mois et un site particulier des données entières ordonnées. L'idée ici est de mettre en évidence non pas des fluctuations globales mais des fluctuations relatives. Nous proposons dès lors d'analyser les variations de quartile pour catégoriser les sites selon la variabilité de la récolte de données les concernant vis-à-vis des autres sites en leur affectant pour chaque variable un score. Ce score correspond à la somme des variations importantes entre deux mois consécutifs (nombre de différences supérieures à 2) normalisée par le nombre d'inter-mois.

Ainsi les figures **??**a-d mettent en évidence des possibles problèmes liés à la récolte de ces sites particuliers et non des tendances globales de la récolte. En effet, nous pouvons remarquer que les problèmes de récoltes sont courants car une majorité des données récoltées a, au regard des autres, des variations importantes un inter-mois sur sept. La classification non supervisée des données construites grâce aux 4 scores obtenus pour chaque site permet de définir des groupes de stratégies de récoltes selon leur variabilité **??**e.

Dans ce premier travail, nous avons cherché à proposer une approche de visualisation de récolte de données issus de différents capteurs (crawlers internes à la société) en cherchant à mettre en évidence des comportements locaux vis à vis des autres plus que des tendances globales ayant des répercussions locales dans une démarche d'analyse de la qualité des capteurs et de la véracité de l'information exploitée. Comme perspective, nous envisageons : i) de continuer à construire des indicateurs et des interfaces de visualisation basés sur des approches robustes, et 2) d'étudier la combinaison de la démarche avec des approches d'analyse visuelle de flux de données sous forme de voisinage (**?**).

Références

Fischer, F., J. Fuchs, F. Mansmann, et D. A. Keim (2015). Banksafe : Visual analytics for big data in large-scale computer networks. *Information Visualization 14*(1), 51–61.

Liu, T., F. Bouali, et G. Venturini (2016). On visualizing large multidimensional datasets with a multi-threaded radial approach. *Distributed and Parallel Databases 34*(3), 321–345.

Louhi, I., L. Boudjeloud-Assala, et T. Tamisier (2016). Approche de clustering de flux basée sur les graphes de voisinage. *Revue des Nouvelles Technologies de l'Information Extraction*

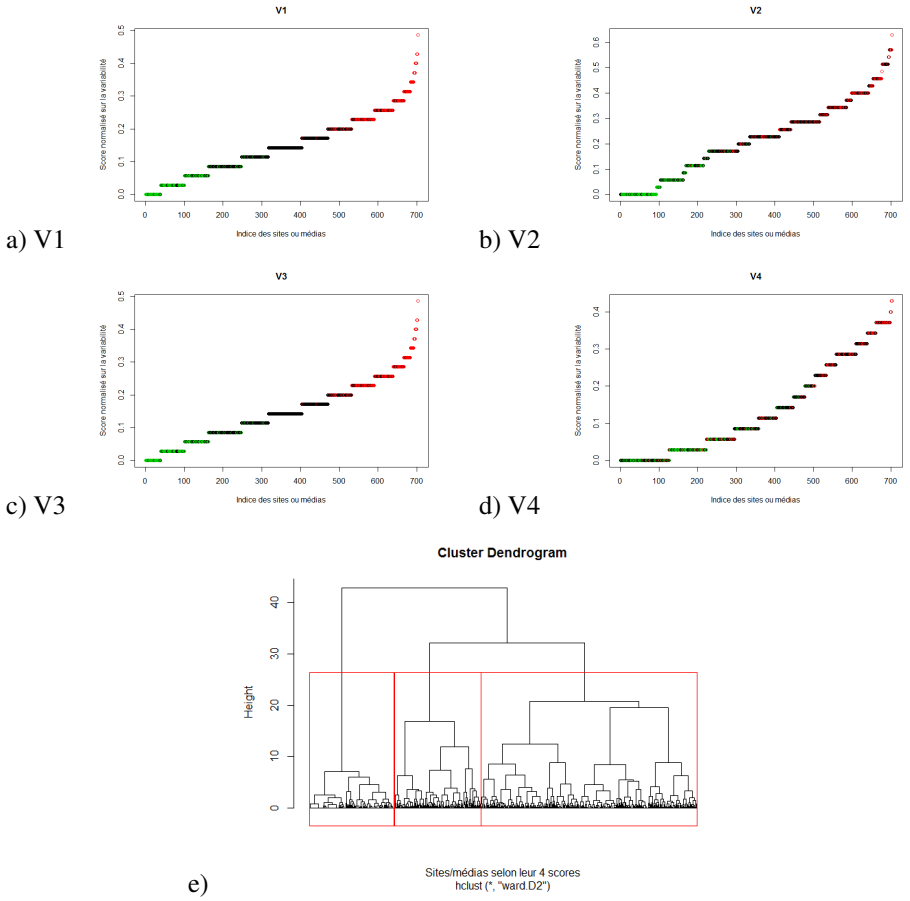


FIG. 2 – *Scores de variabilité par sites et résultat de classification ascendante hierarchique (CAH) : a-d) Visualisation par variable, e) dendogramme de la CAH avec Ward*

et Gestion des Connaissances, RNTI-E-30, 533–534.

Summary

Kantar Media wants to study the digital ad campaigns through uncertain data harvested by crawlers. Therefore, by studying volume data through their quantiles, we develop visualizations that inform on the veracity of the harvesting data.