Atelier VIF : Visualisation d'informations, Interaction, et Fouille de données - EGC 2017

Organisateurs : : Fabien Picarougne (LINA), Pierrick Bruneau (LIST), Hanene Azzag (LIPN), David Bihanic (Université Paris 1)

PRÉFACE

Désormais bien établi à EGC, l'atelier VIF émane du groupe de travail *Visualisation d'Information, Interaction et Fouille de données*, fruit de la collaboration entre les associations EGC et AFIHM. Celui-ci se propose de faire le point sur l'actualité en visualisation interactive d'informations, tant du point de vue fondamental qu'applicatif. Á la confluence des communautés EGC et VIS et à la croisée des disciplines (Informatique, Géographie, Ergonomie, Design, etc.), les méthodes de visualisation interactive et de fouille visuelle des données sont au cœur des préoccupations de cet atelier. Aussi, il aura pour vocation de favoriser l'échange sur l'évolution récente des axes de recherche dans ces thématiques, et sur l'application des méthodes de visualisation à des problématiques industrielles. Le traitement de données massives (*Big Data*) et des flux de données fera l'objet d'une attention particulière.

Fabien Picarougne	Pierrick Bruneau	Hanene Azzag	David Bihanic
LINA	LIST	LIPN	Uni. Paris 1

Membres du comité de lecture

Le Comité de Lecture est constitué de:

Michaël Aupetit (Qatar Computing Research Institute)
Hanene Azzag (LIPN, Université de Paris 13 Sorbonne)
David Bihanic (Université Paris 1 Panthéon-Sorbonne)
Fatma Bouali (LI Tours et Université de Lille2)
Pierrick Bruneau (Luxembourg Institute of Science and Technology)
Mohammad Ghoniem (Luxembourg Institute of Science and Technology)
Fabrice Guillet (LINA, Université de Nantes)
Patrik Hitzelberger (Luxembourg Institute of Science and Technology)
Pascale Kuntz (LINA, Université de Nantes)
Nicolas Labroche (LI, Université François Rabelais de Tours)
Guy Mélançon (LABRI, Université de Bordeaux)
Monique Noirhomme (Institut d'Informatique, FUNDP, Namur, Belgique)
Benoit Otjacques (Luxembourg Institute of Science and Technology)
Fabien Picarougne (LINA, Université de Nantes)
Bruno Pinaud (LABRI, Université de Bordeaux)
Julien Velcin (Université de Lyon 2)
Gilles Venturini (LI, Université François Rabelais de Tours)

TABLE DES MATIÈRES

Analyse interactive post-hoc pour l'identification dans les contenus multimédia

Pierrick Bruneau

LIST, 5 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette
prenom.nom@list.lu,
http://www.list.lu

Résumé. Les algorithmes d'annotation automatique de contenus multimédia produisent inévitablement des erreurs. Dans le contexte spécifique de l'identification de personnes, nous décrivons comment des technologies visuelles et interactives aident à une meilleure compréhension de ces erreurs, fournissant des clés aux équipes de recherche en analyse de documents multimédia pour améliorer leurs algorithmes.

1 Description de la présentation

L'annotation automatique de contenus multimédia consiste notamment à identifier les locuteurs dans ces contenus, i.e. qui parle quand au cours d'un journal télévisé par exemple. L'initiative MediaEval consiste à proposer des challenges auxquels des équipes de recherche peuvent inscrire leurs algorithmes. Une de ces tâches a justement consisté à identifier les locuteurs de manière non-supervisée, c-a-d en n'utilisant que les clés d'identification disponibles dans le contenu multimédia (caractères incrustés, piste audio) (Poignant et al., 2015).

A posteriori (d'où le qualificatif *post-hoc*), les organisateurs du challenge ainsi que les équipes participantes s'intéressent à l'analyse de leurs résultats, et des erreurs commises par leurs algorithmes. En particulier, une partie des plans concernés a été annotée manuellement avec des propriétés qualitatives, telles que la présence de caractères incrustés sur le plan, la présence de plusieurs personnes, ou le fait qu'elles parlent en même temps.

La collaboration avec les organisateur du challenge MediaEval, des chercheurs spécialisés en analyse de documents multimédia, a engendré un outil d'analyse visuelle et interactive mettant en relation les propriétés décrites ci-dessus et les prédictions réalisées par les algorithmes participant au challenge. L'interface repose sur des classifieurs, qui tentent de prédire le succès d'une détection de locuteur dans un plan donné à partir des propriétés des plans.

Au cours de la présentation, outre la description de l'outil interactif lui-même (voir Figures 1 et 2), nous exposons les motivations qui nous ont amené à abandonner une analyse *hors-ligne* des résultats, ainsi que les observations qui ont émergé de l'utilisation de cet outil.

Références

Poignant, J., H. Bredin, et C. Barras (2015). Multimodal person discovery in broadcast TV at MediaEval 2015. In *Proceedings of MediaEval*.

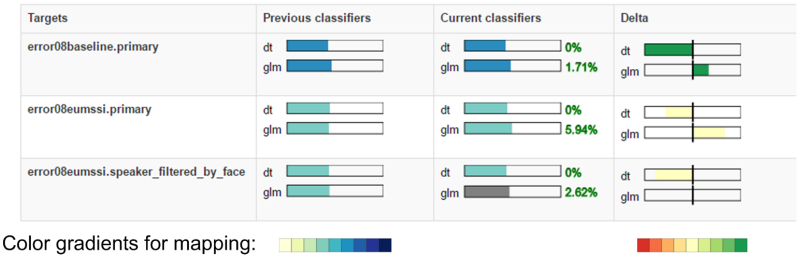

FIG. 1 – *Une vue est focalisée sur les* cibles *à prédire, c-a-d le succès ou l'échec d'algorithmes à détecter les locuteurs d'un plan donné. Chaque cible est associée à un ensemble de classifieurs (e.g. arbres de décision, régression logistique), entrainés pour inférer cette variable de succès. La vue permet de comparer les classifieurs entre eux.*

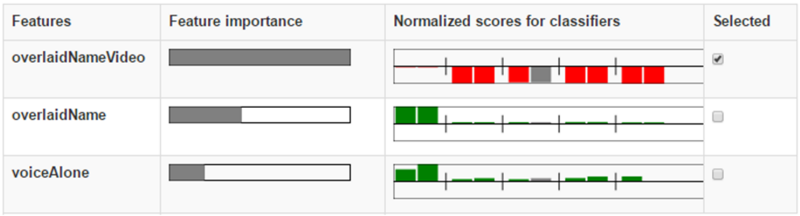

FIG. 2 – *Les propriétés des plans associés sont au coeur d'une autre vue. L'importance des propriétés est déterminée par la valeur absolue maximale de ses scores vis à vis de chacun des classifieurs. Un histogramme révèle le détail de ces scores. Le score est positif ou négatif selon son impact à l'ajout ou au retrait de la propriété associée.*

Coloriage de données multidimensionnelles dans des visualisations guidées par les données à base de profils couleurs extraits d'une base d'images couleurs

Cyril de Runz*, Idriss Oumar Adam*

*CReSTIC, Département Informatique, MODECO,
Université de Reims Champagne-Ardenne
Chemin des Rouliers CS30012 51687 REIMS CEDEX 2, France
cyril.de-runz@univ-reims.fr,
http://http://www.univ-reims.fr/crestic

1 Introduction

L'idée guidant ce travail est de mettre en place des approches de visualisation de données multidimensionnelles (non images) pour lesquelles la couleur et la configuration spatiale des données sont entièrement guidées par les données. Les données guident donc intégralement la visualisation, et les regroupements observés sont extraits de celle-ci. Ces regroupements sont à la fois spatiaux et colorimétriques. Nous cherchons, dans cet article, à avoir une colorisation qui soit guidée par les profils d'une base d'images.

Pour cela, dans un premier temps, en considérant un ensemble d'images couleurs, nous nous inspirons des travaux de Otha et al. (1980) qui identifie une transformation entre couleurs dans une image et ses scores issues d'une analyse en composantes principales (ACP). Cette démarche a inspirée différent travaux de visualisation (Blanchard et al., 2005; de Runz et al., 2012) qui exploite sa transformée inverse afin de colorier des jeux de données multidimensionnelles quantitatives réduits par ACP.

L'objectif de cet article est de travailler sur le lien entre les images couleurs et les ACP construites dessus individuellement afin d'identifier des informations pouvant caractériser ces transformations que nous appelons profils couleurs, ainsi que des regroupements (clusters) de profils et des représentants de ces regroupements à l'aide d'approches de clustering (c.f. figure 1).

Dans un second temps, il s'agit d'exploiter ces profils couleurs représentatifs afin de proposer des coloriages dans les visualisations de données quantitatives multidimensionnelles guidés par les données. Ainsi, contrairement aux travaux de (Blanchard et al., 2005; de Runz et al., 2012) utilisant la transformée inverse d'Otha, nous considérons celle-ci uniquement comme un profil couleur possible parmi d'autres extraits de la précédente étude d'une base d'images couleurs.

La figure 2 présente le résultat obtenu de la colorisation à l'aide des trois profils couleurs représentatifs selon les centroides des trois clusters. Dans les visualisations, les données sont ordonnées selon les classes présentes dans IRIS. Les 5 premières colonnes correspondent à la

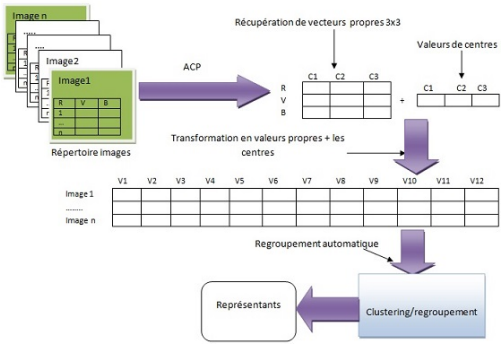

FIG. 1 – *Construction des profils couleurs et obtention des profils représentatifs*

FIG. 2 – *Visualisations des données IRIS coloriées à l'aide des trois profils couleurs représentatifs*

classe Setosa, les cinq suivantes à la classe Versicolor et les 5 dernières à la classe Virginica. Pour tous les profils, et conformément à la littérature, nous pouvons aisément séparer la classe Setosa des autres. Selon nous, sur ces données, nous pouvons distinguer plus nettement, à l'aide du centroïde du deuxième cluster, les classes Versicolor et Virginica qu'avec les autres profils représentatifs.

Références

Blanchard, F., M. Herbin, et L. Lucas (2005). A New Pixel-Oriented Visualization Technique Through Color Image. *Information Visualization 4*(4), 257–265.

de Runz, C., M. Herbin, et É. Desjardin (2012). Unsupervised visual data mining using self-organizing maps and a data driven color mapping. In *16th International Conference on Information Visualisation*, Montpellier, France, pp. 241–245. IEEE Computer Society.

Otha, Y., T. Kanade, et T. Sakai (1980). Color information for region segmentation. *Computer Graphics and Image Processing 13*, 222–241.

Summary

Plateforme web pour l'exploration interactive et multi-niveau de larges collections d'images

Frédéric Rayar*, Sabine Barrat*
Fatma Bouali*,** Gilles Venturini *

*Université François-Rabelais de Tours, Laboratoire d'Informatique
64 avenue Jean Portalis, 37200 Tours, France,
frederic.rayar@univ-tours.fr, sabine.barrat@univ-tours.fr, gilles.venturini@univ-tours.fr
**Université de Lille2, IUT, Dpt STID
25-27 Rue du Maréchal Foch, 59100 Roubaix, France,
fatma.bouali@univ-lille2.fr

1 Introduction

La dernière décennie a vu la réduction du coût des appareils photos, des webcams et des scanners, mais aussi celui des supports de stockage. De fait, la quantité d'images capturées par tout un chacun a explosé, qu'elles soient générées dans le cadre privé, commercial ou dans celui de projets de numérisation (humanités numériques). De plus, l'avènement d'Internet a accentué le fait que le nombre d'images mises en ligne croît de manière exponentielle, notamment avec les sites d'instituts ou encore les réseaux sociaux (Domo, 2015).

Ainsi, les collections d'images devenant conséquentes, il est apparu nécessaire de proposer des paradigmes pour les explorer. Dans le cadre de travaux de thèse, nous avons proposé une solution où l'étude de la construction et de la visualisation d'une structure pour explorer une très grande base d'images a été menée de manière conjointe (Rayar et al., 2016). De plus, nous avons réalisé et mis en ligne des plateformes web permettant une exploration interactive et multi-niveau de larges collections d'images.

Nous proposons dans cette présentation de :

1. brièvement présenter la structure sur laquelle s'appuient nos plateformes,
2. présenter et justifier les choix d'implémentation de ces plateformes,
3. présenter quelques cas d'utilisation qui ont permis de valoriser nos travaux,
4. et enfin présenter les résultats d'une première évaluation utilisateur.

2 Plateforme

Les plateformes ont été réalisé à l'aide de technologies web (HTML5/CSS3) et s'appuient sur des librairies et plugins JavaScript pour dessiner les graphes, afficher les images et interagir avec la structure. Un démonstrateur est accessible en ligne : `http://frederic.rayar.free.fr/ice/` et le code source sera ouvert et mis en ligne sous peu.

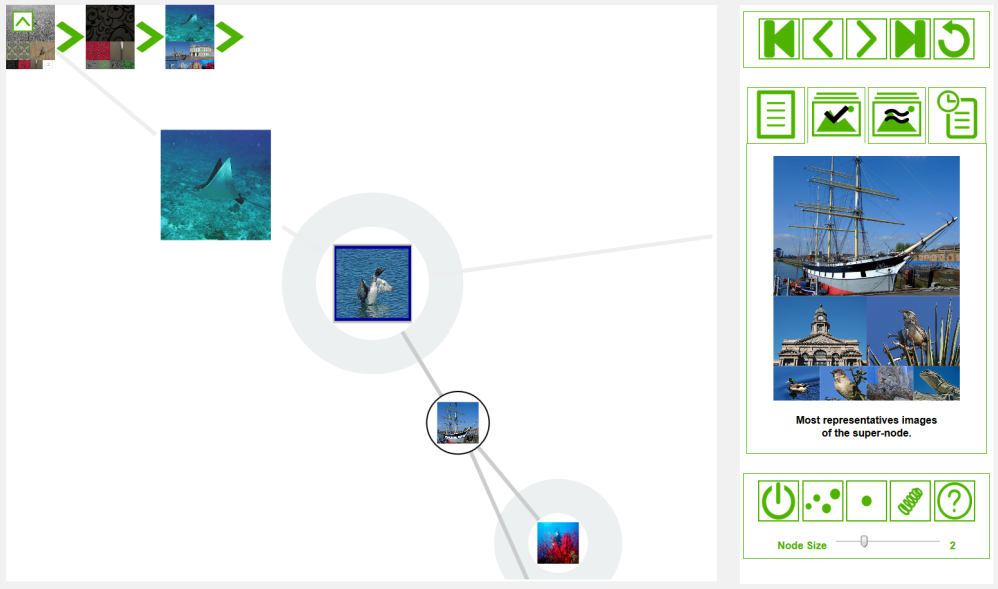

FIG. 1 – *Interface de la plateforme ICE (Image Collection Explorer).*

Références

Domo (2015). Data never sleeps 3.0. https ://www.domo.com/learn/data-never-sleeps-3-0.

Rayar, F., S. Barrat, F. Bouali, et G. Venturini (2016). Incremental hierarchical indexing and visualisation of large image collections. In *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*, pp. 659–664.

Summary

In our research, we address the interactive exploration of large image collections. We propose to present here the platforms that have been developed in the scope these works. More specifically, we will *(i)* briefly describe the hierarchical and graph-based hybrid structure we have proposed, *(ii)* describe and justify the implementation choice of our platforms, *(iii)* illustrates the relevance of our contribution with some use cases and *(iv)* present the results of a user evaluation that has been conducted.

Exploration Visuelle de Données Spatio-Temporelles Brutes

Romain Vuillemot

Université de Lyon, LIRIS
romain.vuillemot@gmail.com,
http://romain.vuillemot.net/

Résumé. Les données spatio-temporelles sont omniprésentes, et pour les comprendre dès leur collecte, il est nécessaire de développer des outils agnostiques vis à vis de leur type, volume et distribution. Nous proposons l'utilisation de graphiques standards (histogrammes, scatterplots, etc.) coordonnés pour l'analyse visuelle de ces données à ce stade très en amont sans hypothèse initiale. Ces graphiques ont la particularité de ne pas être dépendant des caractéristiques du jeu de données et passent à l'échelle facilement. Nous montrons un exemple d'implémentation ce type d'interface appliqué à des centaines de milliers de trajectoires aériennes de manière dynamique et dans le navigateur web.

Contexte. Les activités humaines telles que les déplacements en vélo et en avion, mais aussi les mouvements de pointeurs sur un écran d'ordinateur, génèrent de grands volumes de données qui ont une position (x, y) évoluant au fil du temps (t). L'analyse de ces données permet de comprendre la dynamique temporelle d'événements dans un certain espace, et le comportement d'objets qui se déplacent dans plusieurs lieux (physiques ou virtuels). Le temps et l'espace constituant un index fort de ces données, qu'il est aussi difficile de dissocier, notamment lors de leur représentation : autrement dit les données spatio-temporelles doivent être représentées en points ou trajectoires dans un plan cartésiens. Celles-ci peuvent être contextualisées avec des métadonnées géographiques (ex : plan d'un bâtiment), mais il est difficile de combiner des dimensions supplémentaires ou dérivées (vitesse, accélération, etc.). Ainsi les possibilités de mapping visuel sont limitées à une ou deux dimensions supplémentaires (en utilisant la couleur, des formes et l'animation). Il est donc nécessaire d'utiliser l'interaction utilisateur pour explorer les données et les représentations. Peuquet (2002) et Andrienko et al. (2011) ont proposé des frameworks complets permettant de systématiser l'exploration et la représentation de ces données. Peuquet propose la décomposition en *what+why=when*, *what+when=why* et *when+why=what* afin de décomposer et recombiner les données pour répondre à certaines questions. Adrienko suggère une taxonomie de tâches élémentaires et de recherche de relation, aussi bien directes qu'indirectes.

Visualisations simples et génériques. Notre motivation de design pour explorer les données (x, y, t) repose sur la nécessiter de disposer d'outils d'analyse ne faisant pas d'hypothèse sur les données collectées. Le but étant de pouvoir identifier d'éventuelles anomalies, mais aussi tendances comme des corrélations. Ces étapes sont préalablès à toute phase de traitement et modélisation de données, comme la classification ou la prédiction. La Figure 1 illustre le type d'approche que nous avons déjà implémenté, et les détails de l'interface sont les suivants :

— ① **Vue centrale.** Affichage de la position (x, y) et les trajectoires sous forme d'arcs. Cette vue sert essentiellement d'aperçu sans nécessairement être interactive. D'autres

représentations que l'espace cartésien peut être utilisés, comme les Space-Time cubes avec une vue 3D de l'espace de données.

— ② ③ **Vues de distributions.** Au moyen d'histogrammes, il s'agit d'afficher les propriétés de chacune des dimensions de manière indépendante. L'utilisation de ces graphiques simples permet aussi de nettoyer les données Kandel et al. (2011). Des scatterplots peuvet aussi être utilisés, mais leur nature bi-variable nécessiteraient d'afficher toutes les permutations comme sous forme de matrice Elmqvist et al. (2008).

— ④ **Détails des données.** Il s'agit d'afficher sous forme de tables les données brutes et les valeurs exactes pour toutes les dimensions.

— **Coordination des vues.** Les vues précédentes doivent être rendues interactives afin de permettre le filtrage et la composition de requêtes complexes. Ces requêtes doivent être réalisées avec un temps de réponse quasiment immédiat (< 100ms) afin de garder l'attention de l'humain. Cette coordination dynamique a certes été proposée il y a plus de 25 ans avec HomeFinder Williamson et Shneiderman (1992), mais de nouveaux challenges se posent désormais avec de grands volumes de données, à la fois en termes de rapidité d'affichage et de progressivité du rendu si il n'est pas immédiat.

Nous avons testé cette interface illustrée figure 1 sur différents jeux de données (trajectoires d'avions et de vélos, motion capture 3D d'un humain, position curseur sur un écran, position d'individus dans un bâtiment) qui ont des propriétés suffisamment diverses pour mettre de valider la généricité de l'approche. Nos perspectives de travaux visent donc à étudier comment ce type d'interfaces peuvent être améliorés afin de supporter de manière complète les tâches d'exploration de Andrienko et al. (2011) et Peuquet (2002). En particulier nous sommes intéressés par les problèmes inverses en utilisant l'interaction avec la sortie de l'interface pour explorer les paramètres de configuration d'entrée similaire à Vuillemot et Perin (2015).

Références

Andrienko, G., N. Andrienko, P. Bak, D. Keim, S. Kisilevich, et S. Wrobel (2011). A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages & Computing 22*(3), 213–232.

Elmqvist, N., P. Dragicevic, et J. Fekete (2008). Rolling the Dice : Multidimensional Visual Exploration using Scatterplot Matrix Navigation. *IEEE Transactions on Visualization and Computer Graphics 14*(6), 1539–1148.

Kandel, S., A. Paepcke, J. Hellerstein, et J. Heer (2011). Wrangler : Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, New York, NY, USA, pp. 3363–3372. ACM.

Peuquet, D. J. (2002). *Representations of Space and Time*. Guilford Press.

Vuillemot, R. et C. Perin (2015). Investigating the Direct Manipulation of Ranking Tables for Time Navigation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, New York, NY, USA, pp. 2703–2706. ACM.

Williamson, C. et B. Shneiderman (1992). The Dynamic HomeFinder : Evaluating Dynamic Queries in a Real-estate Information Exploration System. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, New York, NY, USA, pp. 338–346. ACM.

Annexe

FIG. 1 – *Aperçu de l'interface générique d'exploration de données spatio-temporelles que nous proposons, appliquée avec un jeu de données de trajectoires d'avions, affichées sous forme de points et d'arcs dans la vue principale ①. Des vues latérales descriptives d'attributs permettent le filtrage ② et peuvent être combinées entre elles ③. Une table affiche les détails de chaque dimension ④. Après quelques cliques, l'utilisateur découvre que la plupart des vols en retard le sont le soir. L'interface utilise les jeux de données de ASA Data Expo et la bibliothèque DC.JS (*`https://dc-js.github.io/dc.js/`*) autour de Crossfilter (*`http://square.github.io/crossfilter/`*) Le jeu de données affiche 231 038 vols américains au premier trimestre 2001.*

Représentativité, généricité et singularité : augmentation de données pour l'exploration de dossiers médicaux

Joris Falip*, Amine Aït Younes*, Frédéric Blanchard* Michel Herbin*

*CReSTIC, Université de Reims Champagne Ardenne
joris.falip@univ-reims.fr,
http://crestic.univ-reims.fr

1 Introduction

Les données médicales, issues entre autres de dossiers médicaux électroniques, sont de plus en plus abondantes et proposent des défis qui leur sont propres. Pour exploiter ces données, les experts ont besoin d'outils d'exploration et de visualisation favorisant l'émergence de nouvelles hypothèses cliniques. Il est ainsi important, dans le domaine de la santé, de tenir compte des cas atypiques et inhabituels tout en évitant la généralisation (Nourizadeh et al., 2013). Chaque cas est en effet un cas particulier, et tenter de générer des individus typiques issus de moyennes statistiques nous éloigne de la réalité en proposant aux professionnels de santé des modèles abstraits ne correspondant finalement vraiment à aucun des cas concrets qu'ils cherchent à explorer. Le projet *CoSyRES* vise à proposer des algorithmes adaptés à ces contraintes (Blanchard et al., 2010) pour la visualisation et l'exploration de ces données. Nous avons choisi un paradigme basé-instance, très proche de celui utilisé par les professionnels de santé se basant sur leur expérience. Dans cet article, nous proposons une méthode exploratoire permettant de faire émerger des associations de patients similaires et d'observer des regroupements de patients autour des cas jugés comme typiques. L'approche mise en avant ici est adaptée aux difficultés inhérentes aux bases de données médicales : informations absentes ou erronées, données hétérogènes et haute dimensionnalité (Blanchard et al., 2015).

2 Représentativité et émergence d'associations

La méthode exposée ici a pour but de faire émerger des associations entre les individus étudiés, permettant ainsi d'effectuer des regroupements autour d'individus emblématiques. Pour cela, chaque individu va voter en attribuant, selon chaque dimension, un score aux individus qui lui sont semblables. En prenant l'exemple d'un ensemble de données composé de **n** individus décrits chacun par **f** variables, l'émergence de cette structure a lieu en cinq étapes successives :

1. Etablissement, sur chacune des **f** dimensions, d'une matrice de dissimilarité entre les individus.
2. Chaque individu, pour chaque dimension, classe ses voisins par proximité. On obtient donc **f** matrices de rangs, comprenant chacune **n** classements.

3. Transformation des rangs en scores, chaque individu se voyant attribuer **f** scores par chacun des **n** autres individus. Ces scores peuvent être vus comme l'expression de "préférences" individuelles : un score élevé est attribué en cas de proximité forte, un score faible ou nul se voit attribué si les deux individus sont éloignés.
4. Emergence des préférences via l'agrégation des scores obtenus à l'étape précédente. Le score final d'un individu est obtenu en agrégeant les scores qui lui ont été attribués. La valeur obtenue est proportionnelle à la *représentativité* de l'individu, sa capacité à représenter un sous-groupe de la population étudiée.
5. Choix des représentants. En choisissant un facteur **k** modélisant le niveau de granularité souhaité, chaque individu va choisir parmi ses **k** plus proches voisins celui qui a le plus haut score comme étant l'instance la plus représentative de ses caractéristiques.

Des expériences menées sur des jeux de données synthétiques et contrôlés confirment la cohérence des résultats fournis par l'algorithme présenté.

3 Conclusion

L'algorithme proposé permet la visualisation et l'exploration de données de santé à l'aide d'un graphe orienté et valué. Cette approche orientée cas permet de conserver le côté personnel et individualisé de la donnée : un facteur très important dans les applications médicales. La structure de graphe et les notions de représentativité, généricité et singularité permettent d'augmenter les données. La visualisation est ainsi facilitée et l'exploration guidée. L'objectif est de faire émerger de nouvelles hypothèses cliniques, à partir des données patients.

Références

Blanchard, F., A. Aït Younes, et M. Herbin (2015). Linking data according to their degree of representativeness (dor). *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems 15*(4).

Blanchard, F., P. Vautrot, H. Akdag, et M. Herbin (2010). Data representativeness based on fuzzy set theory. *Journal of Uncertain Systems 4*(3), 216–228.

Nourizadeh, A., F. Blanchard, A. Aït Younes, B. Delemer, et M. Herbin (2013). Exploratory data analysis of insulin therapy in the elderly type 2 diabetic patients. *Studia Informatica Universalis 11*(3), 32–49.

Summary

In order to classify individuals according to exemplars that represent them accurately, data visualisation applied to the medical field need to avoid overgeneralization : each case must be treated as a particular instance. This paper presents an algorithm allowing each individual in a dataset to rank other individuals and vote for those that match their important features. Aggregating all these votes give us a way to visualize data according to typical individuals representing subsets of closely-related patients.

Screenographie - de l'autre côté de l'écran

1 Introduction

Les données numériques sont plus que jamais au cœur de nos quotidiens et de nos préoccupations. Qu'il s'agisse de *big data*, de données personnelles ou issues du *quantified self*, chacun envisage et questionne un futur où le nombre de ces données sera croissant. Entre incompréhension et indifférence, les enjeux soulevés par cette déferlante paraissent se jouer loin de nous, pourtant nous sommes tous, à notre manière, auteurs de cet amoncellement. Chaque pas, chaque *like* ou chaque *clic* peut ainsi se trouver enregistré et jour après jour écrire nos histoires individuelles et collectives. Depuis leur création jusqu'à leur (re)présentation, ou restitution, ces données deviennent une matière à penser, à voir le monde et à créer de nouvelles connaissances. Dans ce contexte, dépasser la représentation visuelle et schématique des données pour aller vers des formes sensibles est un moyen à explorer afin de les rendre accessibles et de proposer des modes de présentation plus immédiats pour le grand public. Le design des données, employé dans ce projet, permet de créer des objets - représentation visuelle, matérielle, sonore, etc. - afin que chacun puisse percevoir et penser les données issues de l'activité numérique.

2 Le projet Screenographie

Pour collecter ces traces de nos quotidiens, l'un des capteurs le plus complet et le plus discret n'est autre que notre téléphone mobile, qui nous colle à la peau et nous accompagne partout. En effet il embarque un nombre croissant de capteurs qui permettent de tracer des portraits de plus en plus fidèles et personnels de nos préférences ou de nos mobilités.

Objectifs de recherche Le projet *Screenographie* propose donc de travailler sur la représentation des usages - et les usages eux-mêmes - des terminaux mobiles, à partir d'un nouveau type de données issues du *finger tracking*. Cette technique qui enregistre les pixels touchés par nos doigts sur un écran de smartphone dans un intervalle de temps, permet de créer un nouveau jeu de données. L'un des enjeux du projet sera alors d'étudier la transformation – de la collecte à la restitution en passant par le traitement – de cet ensemble, à la fois personnel et universel.

Le second sera d'investiguer ce type de données pour étudier la façon dont nous usons nos téléphones, bien que cela soit imperceptible, et ainsi de matérialiser l'usure invisible de nos écrans de téléphones.

Descriptif du projet Les écrans des téléphones mobiles sont d'une matière qui ne s'use pas. Si ils se cassent et se fissurent, leur surface inerte ne se patine pas, ne se déforme pas, et ce malgré le fait que nous la sollicitons régulièrement. Ils semblent absorber la caresse de nos doigts et garder secrets nos usages. En rendant visible le trajet de nos doigts sur notre écran, *Screenographie* propose des radiographies de ces usages. Le dispositif composé d'un retroprojecteur et de plaques de plexiglas gravées, retranscrit en grand format l'écran et ce qui se passe à travers - et au delà de - lui. Cette projection en transparence et en lumière permet de créer des portraits de nos usages en fonction des applications enregistrées. Après avoir testé différents mode de capture - continue/discontinue, longues/courtes, une ou plusieurs personnes - nous avons décidés de faire des captures courtes pour une application donnée, décidée en amont. Ainsi contrairement à l'approche répandue "users centric" nous proposons une approche "uses centric", qui permet de raconter autrement l'histoire de nos usages mobiles, en dessinant un motif propre à chaque application du téléphone. Afin d'étudier les corrélations entre les types d'applications et les trajets dessinés par les doigts de l'enquêté, différents type d'applications ont été enregistrées ; réseaux sociaux, divertissement, communication, etc… Les patterns qui se dessinent, indiquent clairement que certaines applications induisent des gestes particuliers. Ainsi les réseaux sociaux nous poussent à *scroller*, là où les jeux invitent à accumuler rapidement des micro-gestes. Ces représentations mettent alors en lumière la gestuelle associée aux différentes applications que nous utilisons quotidiennement et nous questionne sur le rapport charnel à nos téléphones.

3 Conclusions provisoires

Ce travail, encore en cours, a permis dans un premier temps de faire apparaître des patterns, qui peuvent permettre de comparer l'ergonomie des différentes applications. Un deuxième volet, nous permettra d'enquêter sur la dimension universelle de ces traces, en comparant les enregistrements d'une même application chez différents enquêtés. Le mode de captation actuel, contraignant - installation d'une application tierce, documentation manuelle des enregistrements - devra être repensé et permettre une réflexion sur la mise en forme des données dès leur création. Il sera l'occasion d'étudier comment notre usage du téléphone, très personnel, peut aussi raconter une histoire collective au travers des gestes identiques que nous reproduisons.

Summary

Cell phones' screens are made of a material that doesn't wear out. This inert surface doesn't skate or deform, and this despite the fact that we regularly touch it to consult websites, send text or watch multimedia content ... By making the path of our fingers visible on our screen, Screenography offers X-rays of our uses. These representations highlight the gesture associated with the various applications that we use daily and questions us about the physical and sensitive relationship to our phones.

Une analyse de données textuelles des archives numériques de Libération, Le Monde, Le Figaro (1997-2015) pour explorer le traitement médiatique de l'islam en France

Pierre Bellon, Gauthier Bravais, Lucas Piessat

Agence Skoli
9 rue de la Martinière - 69001 LYON
gauthier@skoli.fr
http://www.skoli.fr

1 Avant-propos : accompagner la transformation numérique de la recherche

Skoli est une agence indépendante et spécialisée qui accompagne les organisations de recherche en territoire numérique pour :
- L'usage des données numériques (recueil, traitement, analyse, datavisualisation)
- La création d'interfaces web innovantes de restitution de travaux de recherche

Entre autres phénomènes, l'environnement numérique oblige les organisations de recherche à repenser et augmenter les standards de la production intellectuelle. En effet, il s'affirme à la fois comme le canal privilégié du partage des savoirs et comme un immense réservoir de connaissances disponibles, continuellement alimenté par nos traces numériques. Pour la recherche, la capacité d'explorer ce déluge de données numériques est l'opportunité d'appréhender des sujets jusqu'alors inaccessibles, à condition d'en maîtriser les technologies. En matière de médiation scientifique, elle relance aussi le champ de la "datavisualisation" (considérablement dynamisé grâce au numérique).

C'est sur ce type d'enjeux de transformation digitale de la recherche qu'intervient notre agence, en nous mettant à la disposition de chercheurs le temps d'un projet.

2 "*L'islam, objet médiatique*" (2016), analyse de données textuelles et création d'une interface web de restitution pour le grand public

2.1 Le cadre de l'étude

En France, le rapport des médias à l'islam est devenu l'objet de controverses. Face à ce constat, l'objectif de notre étude était d'apporter un décryptage rigoureux et chiffré. Formant un tandem original à ce niveau, nous sommes associés avec Moussa Bourekba (chercheur au CIDOB, spécialiste du monde arabo-musulman) afin de réunir deux types de

compétences : la méthode et la production scientifique d'une part , un savoir-faire en matière de programmation informatique, de data intelligence, d'écriture numérique, de web design et d'architecture d'information de l'autre.

2.2 La méthode

Nous avons analysé le traitement médiatique de l'islam dans la presse française à travers l'exploration d'un corpus constitué de milliers d'articles issus des archives numériques de 3 quotidiens : Le Monde, Le Figaro et Libération. Au final, notre étude recense les occurrences des termes "islam", "musulman", ainsi que les mots, les adjectifs ou les événements qui leur sont associés. Elle permet notamment d'observer à quel point le 11 septembre 2001 constitue l'acte de naissance d'un sujet médiatique désormais récurrent, mais longtemps resté marginal. Elle permet aussi, dans une logique plus qualitative, d'analyser le vocabulaire employé par les quotidiens sur le sujet, ainsi que son évolution dans le temps.

En toute transparence, notre avons publié le détail de notre méthodologie sur l'interface de restitution elle même.

2.3 L'interface web de restitution

L'interface de restitution que nous avons développée adresse le grand public et constitue de notre point de vue un format original de transmission et de publication des savoirs. Elle mêle analyses et datavisualisations dynamiques/animées et propose un mode de navigation original, basé entièrement sur le *scroll*. Vous pouvez accéder à l'application web ici : http://islam-objet-mediatique.fr/

Preuve de l'intérêt de ce type de coproduction et d'un format de médiation proche du datajournalisme, l'étude a bénéficié depuis sa publication d'une audience exceptionnelle pour le jeune chercheur et la jeune agence que nous sommes : effet de viralité certain sur Twitter, reprise dans Mediapart, émission dédiée sur France Culture en 2017, etc.

3 Propositions d'intervention

Nous vous proposons d'orienter notre intervention sur le retour d'expérience de l'étude "*L'islam, objet médiatique*", et d'aborder un ou plusieurs axes parmi les suivants :
- techniques de *scraping* des archives numériques
- techniques de constitution, "nettoyage" et traitement du corpus
- techniques de création de datavisualisations dynamiques
- librairies et technologies numériques utilisées
- plus largement, notre regard sur l'intérêt (et les réserves), pour la recherche, d'utiliser des données numériques et de créer des formats web innovants de restitution

PORGY : a Visual Analytics Platform for System Modelling and Analysis Based on Graph Rewriting

Bruno Pinaud*, Oana Andrei**, Maribel Fernández***
Hélène Kirchner****, Guy Melançon*, Jason Vallet*

*Université de Bordeaux, LaBRI, France, {prénom.nom}@u-bordeaux.fr
**School of Computing Science, University of Glasgow, UK
***King's College London, UK
****Inria, France, helene.kircher@inria.fr

We propose PORGY [1] a visual modelling framework (Fig. 1) designed for specifying, simulating, and analysing complex systems. PORGY is built on top of the open-source visualisation framework TULIP [2]. PORGY is based on the use of *port graphs with attributes* to represent system states. In a port graph, edges connect to nodes at specific points, called ports. Nodes, ports and edges describe the system components and their relationships, while attributes encapsulate the data values associated with entity. We use graph transformations based on port graph rewrite rules to describe the evolution of the system.

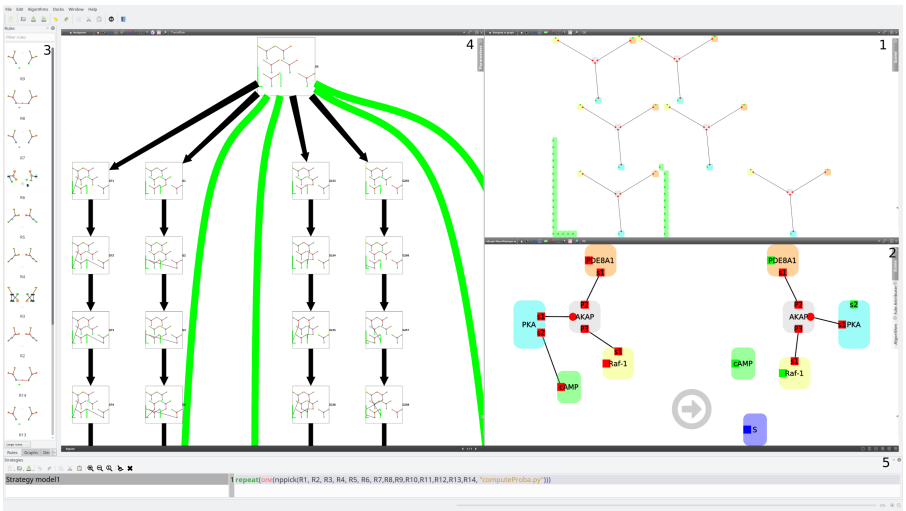

FIG. 1 – *Overview of* PORGY*: (1) editing a graph; (2) editing a rule; (3) all available rules; (4) the derivation tree, a complete trace of the computing history; (5) editing a strategy.*

1. http://porgy.labri.fr
2. http://tulip.labri.fr

Port graph rewrite rules are graphical representations of transformations in the system, thus they provide a direct, visual mechanism to observe the system's behaviour. In addition to port graphs and rewrite rules our modelling approach includes *strategy expressions* to steer rule applications. Strategies allow to use operators to combine graph rewriting rules, as well as operators to define the location where rules should, or should not, apply. Often more than one transformation is possible at a given state, in which case instead of a single transformation step, we may have several alternatives to choose from, and in turn, generate several different sequences of transformations. The various transformation sequences are organised as a tree structure, which we call *Derivation Tree* (DT).

In order to support the tasks involved in the study of a graph rewriting system, PORGY provides facilities to view each component at the same time (rules, strategy, any state of the rewritten graph, DT), to perform on-demand rewriting (strategy-based or rule-based) with drag-and-drop mechanisms, to synchronise the different views to track evolution of system properties, to explore a DT with all possible derivations at different scales, to track the rewriting process throughout the whole DT, or to plot the evolution of a chosen parameter.

Generally speaking, PORGY has been designed with the Visual information-seeking mantra of Shneiderman (1996) in mind: *Overview first, zoom and filter, then details on demand*.

Overview First. To understand the behaviour of non-deterministic systems, it is often useful to execute several times the same rewrite program on the same input to look for potential variations. These can be seen as branches in the DT. Although it is often a large data structure, it provides an indexed representation of the system evolution where each node represents one system state, and an edge is the application of a rule or a strategy. PORGY allows to analyse the derivation tree and work with it at different levels. For instance, Small Multiples (SMs) allow to see consecutive graph states like a comic-strip.

Zoom and Filter. One may be interested in plotting the evolution of a parameter computed out of each intermediate state. An interactive scatter plot can be built. Moreover, thanks to the TULIP backend, all graphical views are synchronised. For instance, if some interesting points are selected inside the scatter plot, they are also immediately selected inside the corresponding branch of the derivation tree. The synchronisation is also valid for graph elements.

Details on Demand. We can investigate further the selected nodes by zooming in and seeing distinctly the graphs. Hovering the mouse pointer over an edge allows to see which elements were changed by the application of the rule. The modified elements are emphasised in the picture, to clearly display which ones have evolved.

In this short abstract, we have introduced some key features of PORGY, an open-source general-purpose modelling and analysis environment based on graph rewriting. Domain-specific versions of PORGY can be easily implemented by extending or refining the features presented here thanks to the TULIP plugin system.

References

Fernández, M., H. Kirchner, and B. Pinaud (2016). Strategic Port Graph Rewriting: An Interactive Modelling and Analysis Framework. Research Report, Inria.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. of the IEEE Symp. on Visual Languages*, pp. 336–343.

Exploration visuelle de graphes multi-couches basée sur un degré d'intérêt

Antoine Laumond*, Guy Melançon*, Bruno Pinaud*

*Université de Bordeaux, UMR 5800 LaBRI
{prenom.nom}@u-bordeaux.fr

Le projet BLIZAAR [1] consiste à explorer de manière visuelle et intéractive des graphes multi-couches dynamiques avec une application dans deux domaines bien distincts : des données historiques sur la construction européenne et des données biologiques sur la composition et le fonctionnement de certains types de plantes. Les graphes multi-couches (Kivelä et al., 2014) issus de ces corpus de données sont très denses. Pour permettre une analyse à la fois fine et à grande échelle, nous proposons de revisiter le concept de *Degree Of Interest* (DOI) de Van Ham et Perer (2009) pour le rendre plus exploratoire et plus intéractif.

À partir d'un graphe initial, l'utilisation du DOI permet d'extraire un sous-graphe pertinent. Une valeur est calculée pour chaque sommet x en fonction d'un sommet dit *focus* y, initialement choisi par l'utilisateur comme un sommet intéressant, et d'une requête utilisateur z :

$$DOI(x|y,z) = \alpha.API(x) + \beta.UI(x,z) + \gamma.D(x,y)$$

API est un score calculé en fonction de la topologie du graphe (par exemple degrés, centralités). UI utilise une requête z spécifiée par l'utilisateur pour fournir un score au sommet x (par exemple recherche de tags dans un attribut). Enfin, D représente la distance entre le sommet observé x et le sommet focus y.

Une fois DOI calculé pour chaque sommet, un algorithme glouton extrait les sommets les plus intéressants. Le sommet focus est sélectionné et ses voisins sont marqués. Parmi les sommets marqués, on sélectionne celui avec le plus haut DOI et on marque ses voisins. On réitère la procédure jusqu'à sélectionner le nombre de sommets souhaité et ainsi créer un sous-graphe G_s pertinent.

Cependant, le choix initial du sommet focus et de la requête associée par l'utilisateur est figé. La méthode présentée dans la suite de ce résumé propose de lever ce verrou en prenant notamment en compte un ensemble évolutif de sommets focus afin de transformer le DOI en une méthode d'exploration pour graphe multi-couche.

L'idée est de proposer des itérations pour faire évoluer G_s et ainsi permettre l'exploration du graphe initial. En plus du sommet focus, on ajoute un ensemble de sommets que l'utilisateur sélectionne au fil des itérations. Ainsi, sur le premier sous-graphe obtenu par application du DOI standard, on sélectionne un ou plusieurs sommets qui vont rejoindre le sommet focus. Le calcul de UI et de D ne s'effectue plus alors entre x et y mais entre x et un ensemble Y de

1. Financement ANR (BLIZAAR ANR-15-CE23-0002-01) et FNR (BLIZAAR INTER/ANR/14/9909176). http://blizaar.list.lu

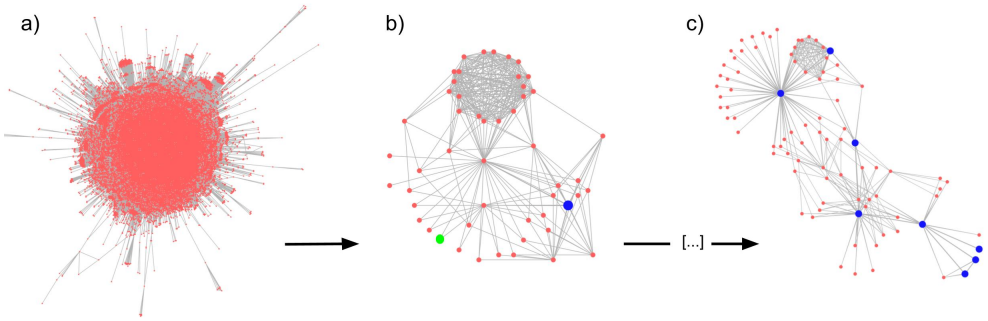

FIG. 1 – *Exemple d'applications successives du DOI sur un graphe. a) correspond au graphe initial. b) est obtenu via un sommet focus (bleu) et permet la sélection d'un ou plusieurs nouveaux sommets (vert) afin de calculer une nouvelle itération. c) correspond au graphe obtenu suite à 7 itérations.*

sommets focus. Le sous-graphe obtenu est alors recalculé en fonction de Y à chaque itération. L'utilisateur peut ensuite sélectionner de nouveaux sommets et ainsi affiner l'intérêt du sous-graphe obtenu en fonction non plus d'une hypothèse de départ fixe mais d'une directive que l'on peut faire évoluer au fur et à mesure de l'exploration. On peut ainsi naviguer dans un large graphe sans être parasité par des informations non pertinentes pour l'utilisateur (Fig.1).

Des améliorations sont en cours de développement notamment au niveau ergonomique en ajoutant une fonctionnalité Bring-And-Go (Moscovich et al., 2009) pour faciliter encore davantage l'exploration de grands graphes ainsi qu'en différenciant le calcul du DOI en fonction des types de sommets dans le cas des graphes multi-couches. Quelques obstacles devront aussi être travaillés notamment au niveau des performances afin que le calcul des nouveaux sous-graphes soit le plus rapide possible. Il serait aussi intéressant de travailler sur le problème de lisibilité. En cas de graphe initial très connecté, le sous-graphe obtenu peut très rapidement perdre en pertinence en étant lui même très dense en lien. Enfin, un utilisateur risque d'être facilement confus entre les différentes itérations des sous-graphes dont le layout est recalculé. Un mécanisme pour conserver la carte mentale de l'utilisateur améliorerait grandement l'efficacité de l'exploration.

Références

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, et M. A. Porter (2014). Multilayer networks. *J. of Complex Networks* (2), 203–271.

Moscovich, T., F. Chevalier, N. Henry, E. Pietriga, et J.-D. Fekete (2009). Topology-aware navigation in large networks. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, pp. 2319–2328. ACM.

Van Ham, F. et A. Perer (2009). "search, show context, expand on demand" : Supporting large graph exploration with degree-of-interest. *IEEE Trans. on Visualization and Computer Graphics 15*(6), 953–960.

Index