Atelier VIF : Visualisation d'informations, Interaction, et Fouille de données - EGC 2018

Organisateurs : : Pierrick Bruneau (LIST), Fabien Picarougne (LINA)

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.

Pierrick Bruneau
LIST

Fabien Picarougne
LINA

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

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 (Fischer et al., 2015; Liu et al., 2016), 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 1a). 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

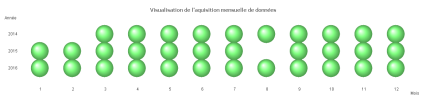
a)  b) 

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 1b 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.

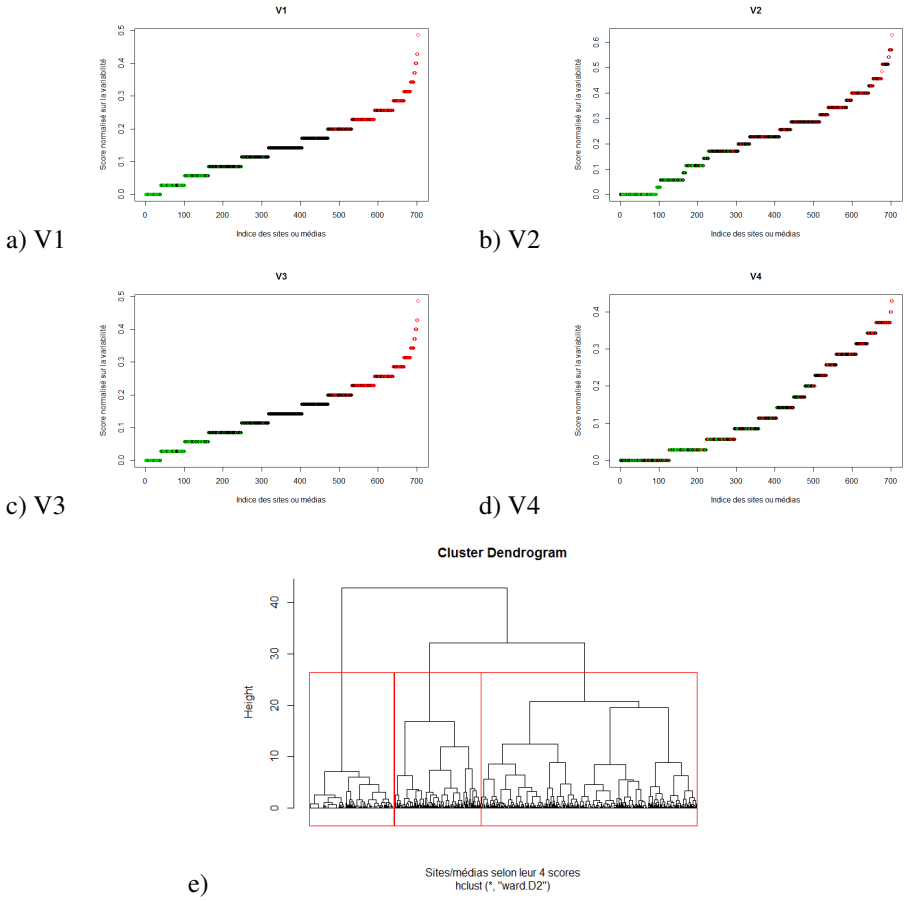
Ainsi les figures 2a-d mettent en évidence des possibles problèmes liés à la récolte de ces sites particuliers et non des tendances globales de la récolte. En effet, nous pouvons remarquer que les problèmes de récoltes sont courants car une majorité des données récoltées a, au regard des autres, des 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 2e.

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 (Louhi et al., 2016).

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.



a) V1 b) V2 c) V3 d) V4 e)

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*

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 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.

Visualisation et parcours de ressources lexicales constituées automatiquement par Word2Vec sur des comptes rendus de maintenance

Meryl Bothua*, Laurent Pierre*

*EDF Lab Paris-Saclay
7 bd Gaspard Monge
91120 PALAISEAU, France
meryl.bothua@edf.fr, laurent.pierre@edf.fr

1 Introduction

Dans le contexte de la transition numérique d'EDF et dans sa volonté d'exploiter l'ensemble de ses données, il est aujourd'hui nécessaire de tester des méthodes de fouille de texte. Une chaîne de traitement a été mise en place pour extraire et analyser des informations à partir de rapports de maintenance. Des tests ont permis de mettre en évidence l'apport de Word2Vec (Mikolov et al. (2013)), pour l'aide à la constitution de ressources lexicales. L'automatisation du processus met en exergue des éléments auparavant noyés dans la masse des données.

2 Exploitation des résultats

Le gain est double : une économie de temps est réalisée grâce à la proposition de termes candidats au peuplement de lexiques ; nous produisons actuellement une sortie RDF avec une ontologie associée et proposons une visualisation sous forme de graphe avec l'outil SemVue. Pour chaque terme du corpus, des candidats sont donnés après prétraitements. Il est possible de filtrer sur la catégorie morphosyntaxique du terme choisi, ainsi que sur son lemme (Schmid (1994)). Le gain est également qualitatif : des synonymes, des abréviations, des possibles fautes d'orthographe et des phénomènes de multilinguisme sont retournés par le système.

3 Visualisation

L'outil SemVue propose une interface de navigation dans un graphe RDF extrêmement facile à mettre en oeuvre à chaque étape du développement d'un entrepot de données. Le coeur de l'interface est défini à l'aide d'une ontologie associée à la modélisation métier proprement dite dans un entrepot de données (Motik et al. (2014)) et permet ainsi à l'interface web de lancer des requêtes SPARQL. La stratégie d'affichage de SemVue est définie à l'aide d'axiomes OWL, ceux-ci configurant dynamiquement le sous-graphe à exposer à l'utilisateur.

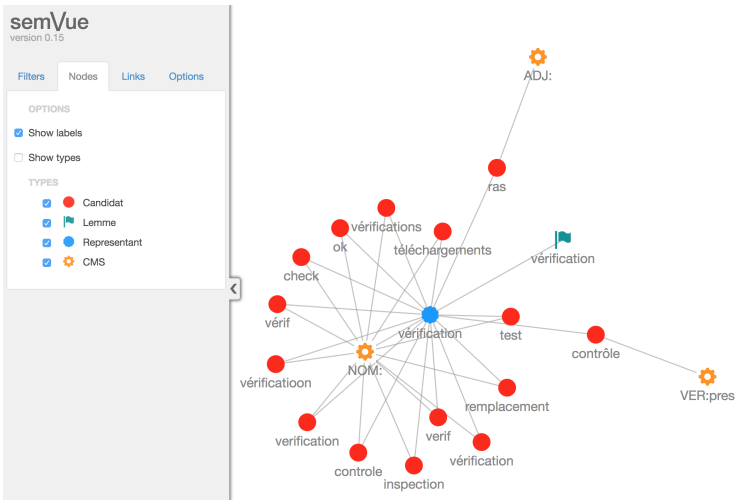


FIG. 1 – *Présentation des résultats avec SemVue.*

4 Travaux futurs

Face à ces résultats prometteurs, nous avons pour perspective de faciliter le processus d'automatisation. Pour l'heure, la méthode que nous avons établie mobilise encore fortement l'utilisateur. Nous souhaiterions développer un outil avec un niveau d'automatisation supérieur pour améliorer le gain de temps et l'expérience utilisateur.

Références

Mikolov, T., K. Chen, G. Corrado, et J. Dean (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*.

Motik, B., Y. Nenov, R. Piro, I. Horrocks, et D. Olteanu (2014). Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems. In C. E. Brodley et P. Stone (Eds.), *Proc. of the 28th AAAI Conf. on Artificial Intelligence (AAAI 2014)*, Québec City, Québec, Canada, pp. 129–137. AAAI Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees.

Summary

EDF started developing a process dedicated to extract and analyze information from maintenance reports. Experiments with Word2Vec were conducted for creating lexical resources from these sets of documents. We populated ontologies from the Word2Vec outputs in order to visualize and check the accuracy of the results, thanks to our ontology driven RDF graph UI. Future work is scheduled for improving automatization of the processing line.

Visualisation dynamique de connaissances : application aux interactions entre facteurs de risque des maladies cardiovasculaires

Rabia Azzi*, Sylvie Despres*, Jérôme Nobécourt*

*Université Paris 13, Sorbonne Paris Cité, LIMICS, INSERM, (UMRS 1142),
Sorbonne Universités, UPMC Université Paris 06, F-93017, Bobigny, France
prenom.nom@univ-paris13.fr,
http://www-limics.smbh.univ-paris13.fr/membres/

La dernière décennie a vu le nombre de décès imputables aux maladies cardiovasculaires (MCV) augmenter considérablement (WHO, 2017). Si les principaux facteurs de risques cardiovasculaires sont aujourd'hui bien connus, leur évaluation à tendance à être réalisée sans considérer leurs interactions.

L'idée guidant ce travail est de concevoir une approche de visualisation dynamique des interactions entre les facteurs de risques cardiovasculaires afin d'aider à la compréhension du déclenchement des effets en cascade produits par l'intervention sur un des facteur (par exemple, agir sur le facteur fumeur pour un patient dépressif peut provoquer une dégradation de son état). La plupart du temps, ces interactions sont représentées par des modèles statistiques synthétisés sous forme de tableaux et représentés en utilisant des graphes. Cette représentation permet de procéder à des parcours du graphe guidé par les interactions existant entre ce mêmes facteurs.

Nous proposons dans ce papier de présenter :

1. brièvement le passage d'un modèle statistique issue de l'étude de (Meneton et al., 2016) à un modèle conceptuel ;
2. la démarche permettant la visualisation dynamique des connaissances représentées dans ce modèle ;
3. le prototype MCVGraphViz [1](cf. figure 1) développé en utilisant la bibliothèque JavaScript D3.js [2] ;
4. les perspectives d'exploitation offerte de ce travail.

1. http ://www-limics.smbh.univ-paris13.fr/MCVGraphViz/
2. https ://d3js.org/

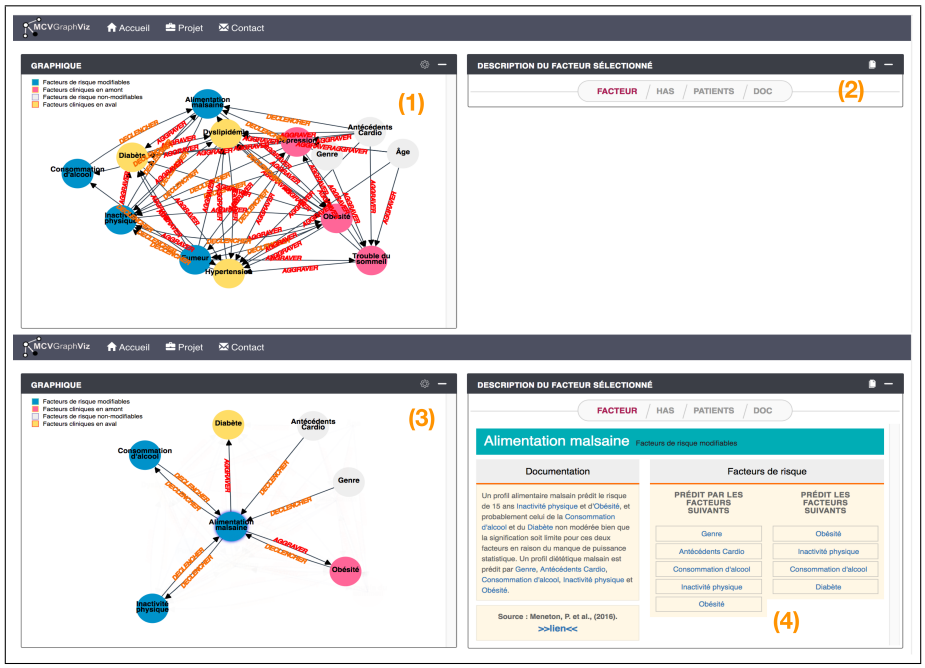


FIG. 1 – *(1) aperçu de l'interface de visualisation des interactions entre les facteurs de risque cardiovasculaires, affichées sous forme de noeuds et d'arcs dans la vue initiale. (2) zone de description des noeuds. (3) vue focalisée obtenue via la selection d'un noeud. (4) description du noeud sélectionné.*

Références

Meneton, P., C. Lemogne, E. Herquelot, S. Bonenfant, M. G. Larson, R. S. Vasan, J. Ménard, M. Goldberg, et M. Zins (2016). A global view of the relationships between the main behavioural and clinical cardiovascular risk factors in the gazel prospective cohort. *PLOS ONE 11*(9), 1–20.

WHO (2017). *World Health Statistics 2017 :Monitoring Health for the SDGs Sustainable Development Goals*. World Health Statistics Annual. World Health Organization.

Summary

In this paper we propose a methodology to obtain a dynamic visualization of interactions between cardiovascular risk factors. In this presentation, we : *(i)* propose a solution to make a transition from a statistical model to a conceptual model, *(ii)* build a dynamic visualization of knowledge represented in this model, *(iii)* introduce our prototype MCVGraphViz, *(iv)* present the prospects for exploitation of this work.

Supervision et respects de la vie privée, l'enjeu ethique des interfaces de visualisation

Roberto Marroquin, Julien Dubois, Christophe Nicolle

Univ. Bourgogne Franche-Comté, Laboratoire Le2i-FRE 2005, Dijon, France
{roberto-enrique.marroquin-cortez, julien.dubois, cnicolle}@u-bourgogne.fr

Les vidéos captées par un réseau de caméras sont habituellement analysées en temps réel par un opérateur humain dans une salle de supervision à l'aide de dizaines d'écrans. Avec l'accroissement de la taille des réseaux et la duplication du nombre de capteurs, il devient difficile, voire impossible, pour un être humain d'identifier des événements rapidement. De plus, les informations visuelles nécessitent une connaissance contextuelle, qui dans un bâtiment est un facteur essentiel à la prise de décision. L'interprétation de l'action perçue dépend souvent du statut courant ou passé de l'utilisateur, des lieux, des objets présents ou d'événements antérieurs. Un bâtiment dit intelligent ne doit plus se limiter à relier de l'information de capteurs, mais il doit aussi être capable de présenter un contexte d'analyse, d'interpréter les multiples paramètres de manière cohérente et d'adapter automatiquement les services fournis au changement de contexte tout en préservant la vie privée des usagers du bâtiment.

L'utilisation d'un système de vidéosurveillance n'est pas triviale. Il peut être considéré comme un système de génération continue de données hétérogènes (flux vidéo, activités, trajectoires...) dont l'analyse, toujours par un être humain, pose de nombreux problèmes : identification des personnes et des activités, association cognitive d'activités produites par plusieurs scènes simultanées, détection visuelle d'occultation, décision d'une action. Dans un centre commercial, dont la fréquentation peut atteindre 20 000 personnes par jour, le système de supervision peut être considéré comme une source de données massives, en perpétuelle transformation et fortement hétérogènes. Pour traiter ces données de manière automatique, outils issues de la recherche dans le domaine du BigData doivent être utilisés. De plus, il est nécessaire de qualifier dans ce traitement la vérité et la valeur des données identifiée (Emani et al., 2015) pour restituer dans une interface de visualisation une information reconstruite juste et pertinente pour un déclenchement d'actions appropriées, sans risque de surcharge cognitive pour l'opérateur humain.

Enfin, l'utilisation de ces systèmes pose toujours le problème de l'atteinte à la vie privée des personnes. Comment garantir que la vidéo brute ne va pas être récupérée et diffusée par des tiers, sans rapport avec les raisons initiales du contexte d'acquisition ? L'exploitation des données personnelles, en France, est un enjeu législatif important. Néanmoins, dans le reste du monde, l'écart entre la protection personnelle et les avancées techniques se creusent(Winkler et Rinner, 2014). Pour combler ce fossé, nous avons développé une approche basée sur un réseau de caméras intelligentes, ou chaque caméra comprend ce qu'elle voit et transmet de la connaissance et non de la vidéo sur le réseau. Toutes les caméras envoient leur connaissance à un serveur central qui va les combiner et, selon des règles métiers, va effectuer des raisonnements pour reconstruire l'activité des usagers dans un bâtiment.

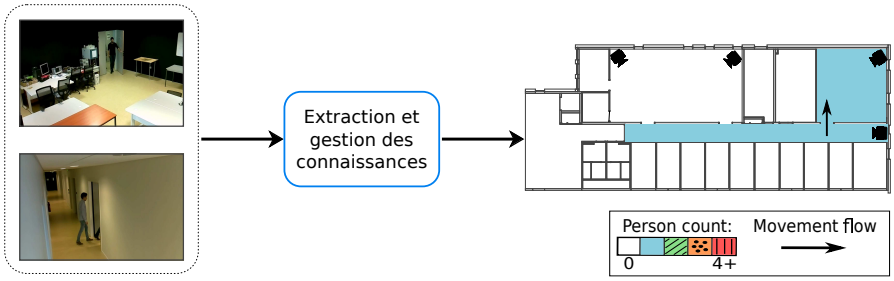


FIG. 1 – *La HeatMap (sur la partie droite) montre le nombre de personnes dans chaque espace et leur déplacement à chaque instant. Cette interface est obtenue en temps réel par un raisonnement réalisé sur une combinaison d'informations fournies par les caméras intelligences. Dans cet exemple, il y a deux personnes à l'étage, l'un est dans le couloir, l'autre a quitte le couloir pour entrer dans un espace (affiché dynamiquement par une flèche).*

À partir de ce système, nous pouvons obtenir des informations, importantes en matière de surveillance, tout en préservant la vie privée des usagers d'un bâtiment :

— Combien de personnes sont actuellement dans le bâtiment et où se trouvent-elles ?
— Quels sont les événements qui se sont déroulés les 5 dernières minutes ? Où sont maintenant les personnes qui ont été associées à ces événements ? (Ce n'est pas le nom de la personne qui est retourné, mais un identifiant anonyme donné par le système).
— Quel est la trajectoire suivie par une personne pendant un certain délai ?

Notre système fournit un ensemble de services de compréhension de l'activité humaine dans un bâtiment dont les réponses sont affichées dans des interfaces de visualisation innovante pour le domaine de la télésurveillance. Par exemple, la figure 1 présente une *HeatMap* qui affiche en temps réel le taux d'occupation de chaque pièce et les trajectoires des personnes, tout en préservant le secret de leur identité. Cette interface peut ainsi remplacer un affichage multiple de vidéos sur un écran dans un poste de télésurveillance. Une démonstration du prototype est accessible sur le site `http://wisenet.checksem.fr/#/heatmap`.

Pour construire un tel système, nous avons dû construire un vocabulaire d'expression des connaissances pour apprendre aux caméras à exprimer la connaissance contenue dans les images. Pour le raisonnement, nous avons utilisé des règles logiques associant ce vocabulaire à une description contextuelle du bâtiment (par exemple les relations topologiques entre les espaces). Ce vocabulaire, enrichi de connaissance contextuelle, nous a permis de concevoir un système de raisonnement pour comprendre l'activité des usagers et déclencher des décisions programmées.

Références

Emani, C. K., N. Cullot, et C. Nicolle (2015). Understandable big data : A survey. *Computer science review 17*, 70–81.

Winkler, T. et B. Rinner (2014). Security and privacy protection in visual sensor networks : A survey. *ACM Computing Surveys (CSUR) 47*(1), 2.

Quelles fonctionnalités pour un outil de visualisation d'ontologie ?

Sylvie Despres*, Jérôme Nobécourt*

*Université Paris 13, Sorbonne Paris Cité, LIMICS, INSERM, (UMRS 1142),
Sorbonne Universités, UPMC Université Paris 06, F-93017, Bobigny, France
prenom.nom@univ-paris13.fr,
http://www-limics.smbh.univ-paris13.fr/membres/

1 Introduction

Nous présentons une analyse succincte des fonctionnalités des principaux outils (*GraphViz*, *VOWL/WebOWL* et *CropCircles*) de représentation graphique d'ontologie mis à disposition d'un ontologue lors de la tâche de construction d'une ontologie. Nous définissons alors les fonctionnalités qui sont implémentées dans Protupos, service web que nous avons implémenté. Il est dédié à la visualisation d'ontologies par différents acteurs intervenant dans la construction de cette ressource.

2 Outils existants pour la visualisation des connaissances pour un ontologue

2.1 Le cas de *Graphviz*

Graphviz est un plugin de Protégé qui s'exécute dans son propre onglet. Il sert principalement à la visualisation graphique d'ontologie grâce à l'utilisation d'un graphe de nœuds et d'arcs qu'il est possible d'étendre et de réduire dynamiquement *via* des actions sur des boutons. Sur le principe, *Graphviz* est très utile pour parcourir la hiérarchie et la manipuler (expansion, respectivement réduction, des propriétés relatives à un nœud). Il permet également la visualisation de la hiérarchie des classes inférées. Les primitives permettant de manipuler cette hiérarchie ne sont pas très intuitives, et dans la plupart des cas les actions effectuées avec ces primitives ne se combinent pas. Le fait que *Graphviz* permette uniquement la visualisation des propriétés hiérarchiques constitue sa principale limite.

2.2 Le cas *VOWL/WebOWL*

VOWL est un plugin de Protégé qui s'exécute dans son propre onglet et qui ne prend pas en compte les actions effectuées par ailleurs. Par exemple, il n'est pas possible de se centrer sur un concept et de naviguer autour de ce dernier dans l'onglet VOWL. Sur le principe, VOWL est une spécification de notations graphiques pour les ontologies écrites en OWL

`http://purl.org/vowl/spec/`. L'outil est clairement tourné vers l'ontologue. Le problème majeur de son utilisation avec Protégé réside dans ses faibles capacités de paramétrage. Ainsi, *VOWL* affiche toute l'ontologie en utilisant la syntaxe graphique sans permettre d'action utilisateur.

Il existe un service web, *WebOWL* (`http://visualdataweb.de/webvowl/`) implémentant VOWL indépendamment de Protégé. *A priori*, l'interface de cet outil semble aussi compliqué que VOWL. Cependant sa grande force est que les actions peuvent être effectués grâce aux boutons et aux menus. Un inconvénient majeur de *WebOWL* est que comme pour tout service web, l'ontologie doit être téléchargée.

Dans les deux cas, la prise en main de l'outil nécessite une phase d'apprentissage et d'appropriation importante.

2.3 Le cas de *CropCircles*

Cet outil intégré à SWOOP permet de visualiser la hiérarchie des concepts. Un concept est représenté par un cercle, les fils de ce concept sont représentés à l'intérieur de ce cercle et la hiérarchie est visualisé selon un représentation en coupe. Il est possible de zoomer sur une couche particulière de la hiérarchie en cliquant dans le cercle et de naviguer dans la hiérarchie en cliquant à l'extérieur du cercle. La vision en coupe par niveau hiérarchique permet notamment de bien appréhender la densité d'un sous-arbre par rapport à un autre. En se plaçant en 2D, la représentation permet de visualiser l'ontologie en coupe de la racine jusqu'aux feuilles. *CropCircles* est construit exclusivement pour parcourir ces coupes dans un plan 2D. Les seules actions possibles sont zoomer/dézoomer ce qui simplifie son utilisation et la rend très agréable et pratique.

3 Fonctionnalités pour l'outil de visualisation Protupos

Un outil n'est réellement utilisable que si son paramétrage prend en compte les besoins de l'utilisateur. Par conséquent, il est préférable de disposer de modes de visualisation présentant les entités utiles à l'utilisateur dans son contexte d'usage.

Après l'analyse des principaux outils étudiés, nous avons retenu les fonctionnalités suivantes pour le service web Protupos :

- visualisation intéractive de la hiérarchie : utilisation d'une représentation circulaire zoomable par actions de l'utilisateur ;
- visualisation de la hiérarchie par coupe de niveau
- visualisation de chaîne de propriétés centrée sur un nœud avec la possibilité de développer dynamiquement la construction de la chaîne ;
- visualisation sous la forme de réseaux d'un type de propriété ;
- visualisation centrée sur un concept en utilisant une fonction "élastique" (permettant de modifier le focus) ;
- nuage de tags associés à un concept.

La présentation de ces fonctionnalités sera effectuée en utilisant l'ontologie OOGO (Ontologie des Outils de Gestion d'Ontologie, `https://www6.inra.fr/reseau-in-ovive/Actions-du-reseau/Seminaires/Seminaire-du-24-novembre-2017`).

Raisonnement à partir de cas visuel : méthodes et application au traitement du cancer du sein

Jean-Baptiste Lamy*, Boomadevi Sekar**, Gilles Guezennec*
Jacques Bouaud*,***, Brigitte Séroussi*,****

*LIMICS, Université Paris 13, Sorbonne Paris Cité, 93017 Bobigny,
INSERM UMRS 1142, UPMC Université Paris 6, Sorbonne Universités, Paris
jean-baptiste.lamy@univ-paris13.fr
**School of Computing and Mathematics, Ulster University, United Kingdom
***AP-HP, DRCI, Paris ****AP-HP, Hôpital Tenon, Département de Santé Publique, Paris

Le raisonnement à partir de cas (RAPC) (Choudhury et Begum, 2016) est un raisonnement par analogie dans lequel la solution à un nouveau cas est déterminée en s'inspirant de cas anciens similaires, dont la solution est connue. Le RAPC comprend plusieurs étapes : la *recherche* des cas similaires dans une base de cas, l'*adaptation* des solutions des cas similaires au nouveau cas, et l'*apprentissage* du nouveau cas. L'étape d'adaptation est délicate et repose souvent sur une expertise humaine. Dans ce travail, nous proposons de faciliter cette étape avec une interface visuelle interactive combinant une approche quantitative utilisant un nuage de points et une approche qualitative utilisant des boîtes arc-en-ciel, une technique récente pour visualiser des ensembles (Lamy et al., 2017).

Ce travail s'inscrit dans le projet européen DESIREE (http://www.desiree-project.eu, financé par le programme Horizon 2020 de l'Union Européenne, *grant agreement* No. 690238). Le projet vise à combiner plusieurs approches d'aide à la décision pour faciliter la prise en charge du cancer du sein. Le RAPC peut être appliqué à l'aide à la décision clinique, par exemple pour déterminer la classe de traitement appropriée pour un patient atteint de cancer du sein, parmi quatre classes possibles : endocrinothérapie, chimiothérapie, radiothérapie, chirurgie.

La Figure 1 montre l'interface que nous proposons. La figure présente un nouveau cas (*query*, en blanc) et 12 cas similaires extraits d'une base de données (en jaune ou en rouge selon la classe du traitement qui leur a été prescrit, respectivement endocrinothérapie ou chirurgie). Le nuage de points (à gauche) présente chaque cas par un point ; la distance entre deux points représentant la similarité entre les deux cas (deux points proches correspondant à deux cas similaires). Nous pouvons voir sur la figure que le cas le plus similaire a été traité par chirurgie (couleur rouge), mais que la majorité des cas similaires a reçu une endocrinothérapie (couleur jaune). Le nuage de point est construit à partir d'une matrice de distances et d'une technique de réduction de dimension (*Multidimensional Scaling*) en coordonnées polaires. Celle-ci préserve intégralement les distances entre le nouveau cas et un autre, au détriment des autres distances.

Les boîtes arc-en-ciel (à droite) présentent chaque cas dans une colonne, en se limitant aux deux classes majoritaires. Les cas similaires sont groupés par classe, avec le nouveau cas au centre. En dessous, des boîtes présentent les caractéristiques communes à différents sous-groupes de cas. Les boîtes sont sélectionnées par Information Mutuelle (MI), de sorte à garder

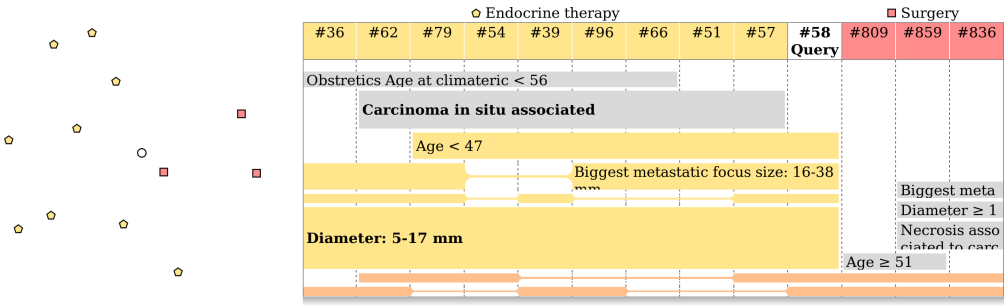


FIG. 1 – *Copie d'écran de l'interface proposée.*

celles qui sont le plus à même d'aider à déterminer la classe du nouveau cas. Ensuite, les boîtes sont colorées en fonction de la proportion de cas de chaque classe qu'elles comprennent. Par exemple la boîte « Age < 47 » occupe toutes les colonnes des patients âgés de moins de 47 ans. Nous voyons que cela comprend 7 cas similaires traités par endocrinothérapie, plus le nouveau cas. L'âge est donc un argument pour classer le nouveau cas en endocrinothérapie, de même que les autres boîtes jaunes.

Une couche interactive permet d'une part d'afficher des détails à la demande, d'autre part de relier les deux parties de la visualisation : par exemple lorsque le curseur de la souris est placé sur une boîte, un message *popup* affiche davantage d'information, et les points correspondant aux cas inclus dans la boîte sont mis en évidence.

Les premiers tests avec les cliniciens ont donné des résultats prometteurs. Les cliniciens ont trouvé intéressante l'approche qualitative, car elle permet de relier les recommandations du système avec des caractéristiques des patients qui leur parlent. En revanche, ils ont parfois été dubitatifs, lorsque le nuage de points et les boîtes arc-en-ciel amènent à des conclusions différentes (par exemple si le cas le plus similaire sur le nuage de points ne correspond pas au traitement majoritaire sur les boîtes arc-en-ciel).

Références

Choudhury, N. et S. A. Begum (2016). A survey on case-based reasoning in medicine. *International Journal of Advanced Computer Science and Applications (IJACSA) 7*(8), 136–144.

Lamy, J. B., H. Berthelot, C. Capron, et M. Favre (2017). Rainbow boxes : a new technique for overlapping set visualization and two applications in the biomedical domain. *Journal of Visual Language and Computing 43*, 71–82.

Summary

In Case-Based Reasoning, adaptating the solutions of old cases to new cases is challenging. We propose a visual interface to facilitate this process, combining a scatter plot with rainbow boxes. We describe a preliminary application to breast cancer therapy.

Exploration de résumés personnalisés de données

Grégory SMITS et Olivier PIVERT

Univ Rennes, CNRS, IRISA - UMR 6074
{gregory.smits | olivier.pivert}@irisa.fr

Résumé. Ce document présente une approche d'exploration et d'extraction de connaissances à partir d'un résumé linguistique et personnalisé des données.

Résumé personnalisé de données

Qu'il s'agisse d'un cadre professionnel ou personnel, le volume ainsi que l'hétérogénéité des jeux de données disponibles ne cessent d'augmenter. La valeur d'un jeu de données dépend essentiellement des connaissances que l'utilisateur peut en extraire. Un des enjeux cruciaux auxquels la communauté scientifique de gestion des données et des connaissances doit répondre concerne le développement de méthodes et d'outils permettant à un utilisateur de traduire rapidement des données brutes en connaissances interprétables et exploitables. Une manière d'accélérer ce processus de transformation des données en connaissances repose sur la génération de résumés des données, résumés qui pourront ensuite être restitués graphiquement à l'utilisateur afin qu'il dispose d'un aperçu des données. Un second axe permettant d'accélérer l'appropriation de données par un utilisateur concerne la personnalisation des représentations et des explications générées.

Une des utilisations possibles de la théorie des sous-ensembles flous consiste à transformer les domaines de définition, généralement de nature numérique ou catégorielle, des attributs qui décrivent les données en variables linguistiques. Comme l'illustre la figure 1, une variable linguistique associée à une partition floue des données permet à la fois de discrétiser un domaine de définition et également d'associer une étiquette linguistique à chaque regroupement de données. De plus, de par la possibilité de représenter des transitions graduelles entre les différents éléments de la partition, les ensembles flous constituent un cadre théorique idéal pour représenter le caractère subjectif et imprécis des termes que nous utilisons pour décrire des phénomènes observables (e.g. *'prix élevé'*, *'accélération forte'*, *'métier à risque'*, etc). Les partitions floues définies par l'utilisateur sur les attributs qui l'intéressent forment un vocabulaire utilisateur.

Nous avons conçu un algorithme distribuable dans une architecture de calcul pour quantifier dans quelle mesure chaque terme du vocabulaire utilisateur couvre un jeu de données. Le résultat de ce processus de réécriture des données forme un vecteur composé de termes linguistiques subjectifs. L'ensemble des termes linguistiques du vocabulaire qui couvrent un tant soit peu les données constitue un résumé que nous pouvons par exemple représenter graphiquement comme un nuage de mots ou une vue *ad-hoc*. La figure 1 (droite) décrit linguistiquement 127

FIG. 1 – *Partition floue décrivant la durée d'un vol (gauche) et résumé des données (droite)*

millions de vols domestiques aux Etats-Unis sur six dimensions [1]. Une mesure d'information dédiée à cette représentation a également été proposée pour identifier les termes les plus informatifs d'un résumé. La représentation graphique des données est ensuite ajustée pour mettre en exergue ces termes importants.

Exploration interactive et extraction de connaissances à partir des résumés

Outre le fait de fournir une vue synthétique de grands volumes de données, la représentation graphique d'un vecteur de réécriture constitue également une interface graphique d'exploration des données. En cliquant sur un terme apparaissant dans le résumé, l'utilisateur déclenche la sélection du sous-ensemble des données satisfaisant ce terme. Ce sous-ensemble est lui-même résumé, puis représenté graphiquement afin de permettre une exploration interactive des données par combinaison conjonctive de termes linguistiques subjectifs. Nous avons construit une structure d'indexation dédiée aux vecteurs de réécriture pour permettre une navigation fluide dans les données. Nous avons également montré dans Smits et al. (2016) qu'il était possible d'extraire très efficacement des connaissances utiles (règles d'association, termes atypiques, etc.) à partir des vecteurs de réécriture construits lors du résumé des données.

Notre objectif actuel est de collaborer avec des experts en interface graphique pour construire des représentations dédiées à nos résumés linguistiques et aux connaissances que nous extrayons.

Références

Smits, G., O. Pivert, et R. R. Yager (2016). A soft computing approach to agile business intelligence. In *Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on*, pp. 1850–1857. IEEE.

Summary

This short document presents a data exploration and knowledge extraction approach based on a linguistic and subjective summary of the data.

1. http ://stat-computing.org/dataexpo/2009/the-data.html

Visualisation Immersive de Graphes en 3D pour explorer des graphes de communautés

Laurent Brisson*, Thierry Duval*, Rémi Sahl

*IMT Atlantique, Département LUSSI, Technopole Brest Iroise, 29238 Brest Cedex 3
Lab-STICC (UMR CNRS 6285)
{laurent.brisson, thierry.duval}@imt-atlantique.fr

1 Introduction

Nous voulons étudier l'apport de la visualisation 3D immersive comparativement à la visualisation 2D classique pour assister l'exploration et l'analyse des communautés dans les réseaux représentés sous forme de grands graphes. La finalité de ce travail est d'être en mesure de proposer des solutions de visualisations valorisant et s'appuyant sur les travaux effectuées dans le domaine de l'analyse des réseaux sociaux et en particulier les algorithmes de détection de communauté, d'analyse d'influence ou de propagation d'information.

2 Visualisation de Graphes et Réalité Virtuelle Immersive

Le domaine de la visualisation d'informations commence à découvrir le potentiel de la 3D immersive pour la visualisation de données. Les premiers travaux sur le sujet de Ware et al. (1996) ont montré que la RV immersive peut apporter beaucoup pour la visualisation de réseaux en 3D. Plus récemment Tory et al. (2006) ont montré qu'ajouter une troisième dimension combinée avec des effets visuels appropriés pouvait également améliorer les performances des utilisateurs lors de l'analyse visuelle de données. Dernièrement Kwon et al. (2016) ont montré que la visualisation 3D immersive pouvait être plus efficace que des visualisations sur des écrans 2D classiques. Il faut cependant rester prudent sur l'usage de la 3D de façon à ne pas fatiguer inutilement les utilisateurs comme l'ont constaté McIntire et Liggett (2014). On retiendra néanmoins que les travaux de Hand (1997) rappellent que l'usage du "motion parallax" via un tracking de la tête de l'utilisateur améliore grandement la perception du relief et améliore le confort de l'utilisateur de technologies immersives.

3 Layouts 3D pour la visualisation 3D immersive de graphes

En nous basant sur des premières propositions de Greffard et al. (2012) pour visualiser des graphes à l'aide d'algorithmes "force-directed", et sur les travaux de Kwon et al. (2015) et Kwon et al. (2016) proposant un nouveau layout 3D projetant des graphes classiques 2D sur une sphère, nous avons proposé une nouvelle visualisation 3D de graphes de communautés.

Pour la visualisation d'une communauté, nous avons choisi un algorithme permettant de projeter des layouts 2D sur des sphères, associé à un algorithme de edge bundling de Lambert et al. (2010) proposé par le logiciel Tulip utilisé pour la création du graphe 2D initial, et à une projection des arêtes sur la sphère à l'aide de courbes de Béziers (voir figure 1 à gauche).

Nous avons ensuite considéré chaque communauté comme un nœud d'un autre graphe reliant ces communautés réparties le long d'un cercle (voir figure 1 à droite). Pour faciliter la visualisation et regrouper les liens partant des membres d'une communauté et allant vers des membres d'une autre communauté, nous avons fait une première adpatation en 3D de l'algorithme KDEB (Kernel Density Estimation edge Bundling) de Hurter et al. (2012).

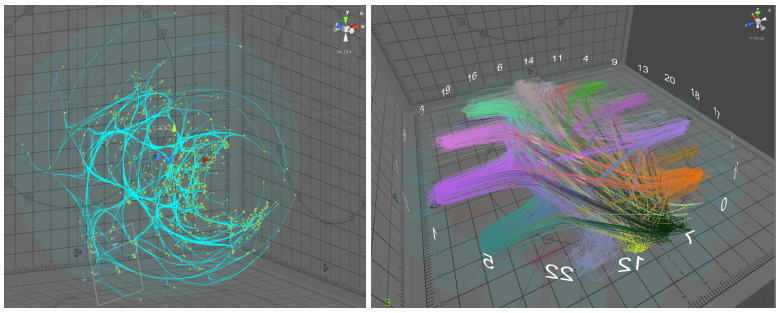


FIG. 1 – *Métaphores de visualisations 3D de liens intra-communauté (à gauche) et inter-communauté (à droite).*

Ces représentations facilitent à la fois les observations de liens inter-communautés et intra-communautés.

4 Conclusion et perspectives

Ces résultats sont une première étape vers la proposition de nouvelles métaphores de représentations spatiales adaptées à la visualisation de graphes de communautés. Il reste à les évaluer en les comparant à des représentations 2D équivalentes.

Pour la représentation intra-communauté, il reste à améliorer l'algorithme de répartition des nœuds qui utiliserait leurs propriétés topologiques pour fournir une visualisation plus intuitive ne se limitant pas à une projection de tous les nœuds sur une même sphère mais plutôt à l'intérieur d'une zone délimitée par 2 sphères concentriques de diamètres différents, en prenant par exemple en compte la connectivité intra-communauté et extra-communauté des nœuds.

En ce qui concerne la représentation inter-communautés, il reste à améliorer l'algorithme de edge bundling et à explorer également d'autres métaphores spatiales pour ne pas se limiter à un cercle mais également projeter les communautés sur un espace 3D.

Ces métaphores devront supporter le passage à l'échelle, car il doit falloir pouvoir traiter des représentations de plusieurs centaines de communautés de plusieurs milliers de membres.

Ces visualisations devront permettre une collaboration (potentiellement asymétrique) entre différents acteurs impliqués dans l'analyse des communautés : data scientist et expert métier). Ces asymétries de collaborations pourront être immersive / non immersive et/ou 2D / 3D.

Références

Greffard, N., F. Picarougne, et P. Kuntz (2012). Immersive Dynamic Visualization of Interactions in a Social Network. In *Challenges at the Interface of Data Analysis, Computer Science, and Optimization, Studies in Classification, Data Analysis, and Knowledge Organization*, pp. 255–262.

Hand, C. (1997). A Survey of 3D Interaction Techniques. *Computer Graphics Forum 16*(5), 269–281.

Hurter, C., O. Ersoy, et A. Telea (2012). Graph bundling by kernel density estimation. *Computer Graphics Forum 31*(3pt1), 865–874.

Kwon, O.-H., C. Muelder, K. , Lee, et K.-L. Ma (2015). Spherical layout and rendering methods for immersive graph visualization. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, Volume d, pp. 63–67. IEEE.

Kwon, O. H., C. Muelder, K. Lee, et K. L. Ma (2016). A study of layout, rendering, and interaction methods for immersive graph visualization. *IEEE Transactions on Visualization and Computer Graphics 22*(7), 1802–1815.

Lambert, A., R. Bourqui, et D. Auber (2010). Winding roads : Routing edges into bundles. *Computer Graphics Forum 29*(3), 853–862.

McIntire, J. P. et K. K. Liggett (2014). The (possible) utility of stereoscopic 3D displays for information visualization : The good, the bad, and the ugly. In *2014 IEEE VIS International Workshop on 3DVis (3DVis)*, pp. 1–9.

Tory, M., A. Kirkpatrick, M. Atkins, et T. Moller (2006). Visualization task performance with 2D, 3D, and combination displays. *IEEE Transactions on Visualization and Computer Graphics 12*(1), 2–13.

Ware, C., D. Hui, et G. Franck (1996). Evaluating stereo and motion cues for visualizing information nets in three dimensions. *ACM Transactions on Graphics (TOG) 15*(2), 121–140.

Summary

In this paper we present some new metaphors for immersive 3D visualization of communities networks: one spherical projection for visualization of intra-community links and one circular distribution for visualization of inter-community links.

Index