Réticulation virtuelle de réseaux effectifs

Pr Luc QUONIAM (*), Charles-Victor BOUTET (*)
mail@quoniam.info, mnem00@gmail.com

Thèmes : réseaux /Data Mining/visualisation de l'information/cartographie/

Résumé : Les réseaux (d'établissements, d'individus…) constitués dans le monde tangible ont leur pendant numérique. On évoque aujourd'hui les nombres informatiques là où hier encore leurs cousins astronomiques symbolisaient la quantité pléthorique, infinie, cognitivement hors de portée. Comment établir une représentation humainement exploitable, une modélisation de l'intelligence numérique de réseaux de sites web nichés dans la matrice de surcharge informationnelle internet ? Nous proposons, à travers nos travaux, l'élaboration d'un outil de *crawl* qui permettra, *in fine*, d'obtenir une représentation, que nous souhaitons la plus cognitivement confortable, de la structure réticulaire d'un ensemble de sites web appartenant à des membres d'un réseau du monde matériel.

MOTS-CLE : fouille de données, visualisation d'information, web-crawling, réseaux sociaux

KEYWORDS : data mining, infovis, web-crawling, social networks

1 Introduction

"Historiquement, une grande quantité d'information a toujours été une bonne chose : l'information a rendu possible la dissémination des cultures, le développement du commerce et des technologies" (Carlson, 2003). "*Actuellement l'accroissement constant des informations [...] au plan international [...] pose problème de la manière dont ces informations vont être construites, associées et traitées*" (Dou *et al.,* 2003) : la surcharge informationnelle fait problème ainsi que l'organisation de l'information, a fortiori sur internet. Nous souhaitons parvenir à une modélisation cognitivement exploitable d'un ensemble de sites web appartenant aux membres d'un réseau d'entités liées dans l'univers réel afin d'obtenir un outil d'analyse pertinent quant aux relations numériques existant entre lesdits sites web. Dans un premier temps, il nous faudra réaliser un outil capable de parcourir les sites web ciblés : un web-crawler, puisque, à notre connaissance, les outils déjà existant tel que « Soscibot » ne satisfaisaient pas nos objectifs et notamment la faculté de parcourir toutes les pages d'un site web, d'hyperlien en hyperlien, pour isoler les hyperliens d'un membre du réseau vers un autre membre, et de réitérer l'opération sur tous les sites protagonistes du réseau. De facto, préalablement à la phase de modélisation se présentait une phase de data mining.

2 Etude de cas : RMEI

Le Réseau Méditerranéen des Ecoles d'Ingénieurs a été créé en juin 1997 à l'initiative du Groupe Ecole Supérieure d'Ingénieurs de Marseille (ESIM), établissement de la CCI Marseille-Provence. Depuis sa création, le RMEI veut valoriser les atouts spécifiques de la Méditerranéen misant sur une efficacité accrue, par le maillage des Ecoles d'Ingénieurs, des Universités et autres partenaires du pourtour méditerranéen : Le RMEI agit comme interface entre le monde de l'entreprise et les grandes écoles et universités techniques. Il doit faciliter et intensifier la relation Universités – Grandes Écoles - Laboratoires de Recherche - Entreprises de la Méditerranée au service de l'innovation mais aussi du recrutement d'ingénieurs et de scientifiques compétents, enjeux des grands groupes industriels mais aussi des PME/TPE de la Méditerranée. Aujourd'hui, le RMEI c'est plus de 100 000 étudiants - ingénieurs qui sont concernés par les 53 institutions membres issues de treize pays.

2.1 Problématique

Bien que ce réseau d'établissements ait été formé en 1997, son site Internet n'existe que depuis deux ans. Aussi, nous nous proposons de cartographier la structure réticulaire naissante, formée par les 53 institutions susdites ainsi que le site du RMEI lui-même, dans le but d'avoir une vision globale, d'appréhender les relations existantes ou non entre les différents membres de ce réseau de sites Internet afin d'obtenir une connaissance simple *: **le réseau virtuel de RMEI est-il vraiment existant en terme de collaboration ou non*** **?** En d'autres termes : la collaboration effective à travers leurs sites est elle en adéquation avec leur volonté et leur communication sur le sujet.

Pour ce faire, nous devions :

- Etre en mesure de récolter les données (de *crawler* l'ensemble des sites web)
- Parvenir à les synthétiser (En l'occurrence, construire un modèle de représentation visuelle cognitivement avantageux/interprétable mettant en exergue les propriétés remarquables où leur absence – ici, les liens entres membres du réseau)
- Trouver un moyen d'afficher le modèle en question afin que l'humain puisse le percevoir, et ce sans dégrader la qualité de la représentation construite à l'étape précédente

2.1.1 Construction de la chaine de traitement

Devant l'inexistence d'un outil capable d'effectuer de bout en bout les tâches susmentionnées, du *crawling* à l'affichage, Nous s devions constituer notre propre boîte à outils soit une chaîne de traitement dont la flexibilité autoriserait son application à des cas tout à fait hétérogènes au-delà du cas RMEI. Afin de faire aboutir notre entreprise, nous devions en premier lieu disposer d'un outil de crawl qui parcourrait l'ensemble des sites web du RMEI, d'hyperlien en hyperlien, en effectuant du TDM (*Text Data Mining*) afin de raffiner la récolte des données et aussi accélérer le processus de collecte : un seul de ces sites web pouvait comporter plusieurs millions de pages et la récupération de données à travers Internet n'est pas instantanée. Des langages de programmation parfaitement indiqués pour cette tâche tels que Perl ou Python disposaient déjà de librairies de fonctions spécifiquement dédiées.

2.1.1.1 Web crawling

Après de nombreux tests, notre choix s'est porté sur la librairie URLNet programmée en langage Python : celle–ci dispose d'un jeu de fonctions, d'une orientation logique axée sur le concept d'arbre (arborescence d'un site web) et de forêts (un à N jeux de N arbres) dotées des fonctions de filtrage et d'une grande flexibilité manquant au logiciel Socscibot développé par le cybermetrics group de l'université de Wolverhampton. De plus, l'acte de crawling lui-même, point critique dans ce genre de démarche car chronophage, était l'un des plus rapides parmi les bibliothèques testées par nos soins.

2.1.1.1.1 Méthode exploratoire et construction logicielle du module de crawling

La méthode exploratoire utilisée est simple :

On dispose de n URLs de sites web que l'on numérote de 1 à n. Pour chaque URL, on envoie un agent logiciel récupérer son contenu html, particulièrement les hyperliens. Chaque hyperlien est analysé pour être :

Stocké dans un tableau d'urls parcourus s'il fait partie de l'arborescence du site à partir duquel il a été récolté (trivial pour les URLs de départ)

Expurgé s'il est exclu des arborescences des n racines/URLs de base (cas d'un hyperlien vers un site externe au réseau)

S'il fait partie de l'arborescence de la racine/URL où on l'a récupéré, on le stocke dans un tableau d'URLs à parcourir si l'URL en question n'a pas déjà été parcouru.

S'il fait partie de l'arborescence de la racine/URL où on l'a récupéré, on l'expurge si l'URL en question a déjà été parcouru.

S'il fait partie de l'arborescence d'une racine du réseau différente de la racine de l'URL où on l'a récupéré, c'est donc un lien d'un membre du réseau vers un autre membre. Ce lien est stocké dans un tableau à deux entrées : Numéro du site originaire du lien et numéro du site destinataire du lien (dans un graphe à eux dimensions, il s'agirait de nœuds)

On réitère l'opération à l'aide d'un plus grand nombre d'agents jusqu'à ce que le tableau d'URLs à parcourir soit vide, signe que les n sites web auront été explorés. De là, nous disposons de tous les nœuds pour tracer un graphe représentatif des relations établies entre chacun des sites web du réseau. Les liens/URLs récoltés, en nombre et en profondeur, pourront éventuellement faire office de critères utiles à la symbolisation de l'importance de l'interconnexion entre plusieurs membres.

2.1.1.1.2 De l'Extraction de Connaissances de Données

Selon le Gartner Group, le *Data Mining* est un procédé de découverte de corrélations significatives, de règles de tendances en parcourant de grands volumes de données stockées dans des référentiels en utilisant des technologies de reconnaissance de forme mais également des techniques statistiques et mathématiques.

Il convient cependant de distinguer les diverses déclinaisons de ce concept global. Quand une nouvelle discipline émerge cela prend habituellement quelque temps et un bon nombre de discussions avant que les concepts et les limites soient normalisés et c'est justement le cas du *text mining* (Kroeze *et al.,* 2003) : Dans un papier d' inauguration, "*Untangling text data mining*", (Hearst, 1999) a abordé le problème de clarification des concepts et la terminologie de "*text-mining*" comme il est important de distinguer le *"text data mining"* (TDM) et l'accès à l'information (*text retrieval*).

Les bases de données bibliographiques peuvent contenir les champs clairement structurés, tels que l'auteur, le titre, la date et l'éditeur, aussi bien que du texte. En ce sens, l'extraction de connaissances de base de données (ECBD, en Anglais *KDD : knowledge discovery in databases*) contenant des champs textuels relève du *Text Mining*. L'exploitation du Web est un champ plus large que le *Text Mining* parce qu'il contient des éléments hétérogènes (*e.g.* multimedia) et non nécessairement structurés qui nécessitent des traitements particuliers. On parlera d'ECD (extraction de connaissances de données) pour désigner les méthodes automatiques ou semi-automatiques d'analyse de l'information, issues des mathématiques, qui permettent de faire émerger d'une masse de données volumineuse des informations cachées.

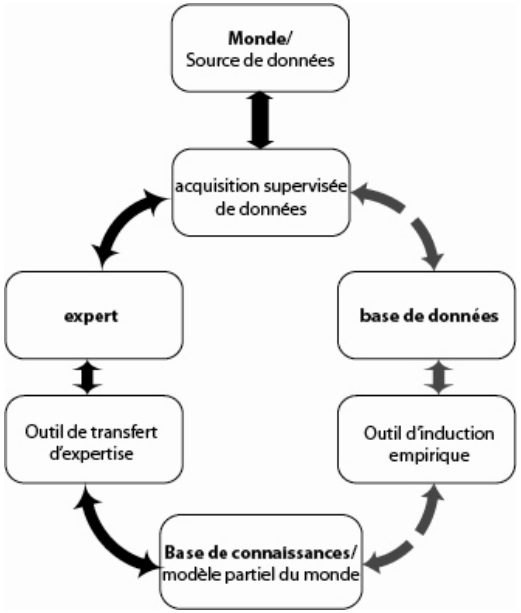

Figure 1 : Transfert d'expertise en balance avec l'induction empirique, deux processus d'Extraction de Connaissances de Données

2.1.1.1.3 De l'extraction de connaissance de données à l'automation de modèles visuels.

Dans son analyse " *the trade-off between knowledge and data in knowledge acquisition*" (Piatesky-Shapiro – Frawley, 1992), Gaines détaille deux approches de l'Extraction de Connaissances de Données : d'un côté l'apprentissage par la machine à créer des connaissances depuis des données (à droite sur la figure un), de l'autre un processus similaire dans lequel la création de connaissances à partir de données est assistée par un expert (à gauche sur la même figure) au lieu d'être à charge d'un outil d'induction empirique. Gaines (Ibid.) remarque que l'expertise humaine peut être source d'erreurs et qu'elle ne peut totalement remplacer l'induction empirique mais peut servir à la guider et à l'expédier. C'est dans cette optique et à cause de la problématique suivante que nous options pour un fonctionnement de pre-réglages de chacun de nos modules par expertise humaine préalablement au processus de construction tel qu'illustré en figure 2, et ce, tout au long du processus de prototypage.

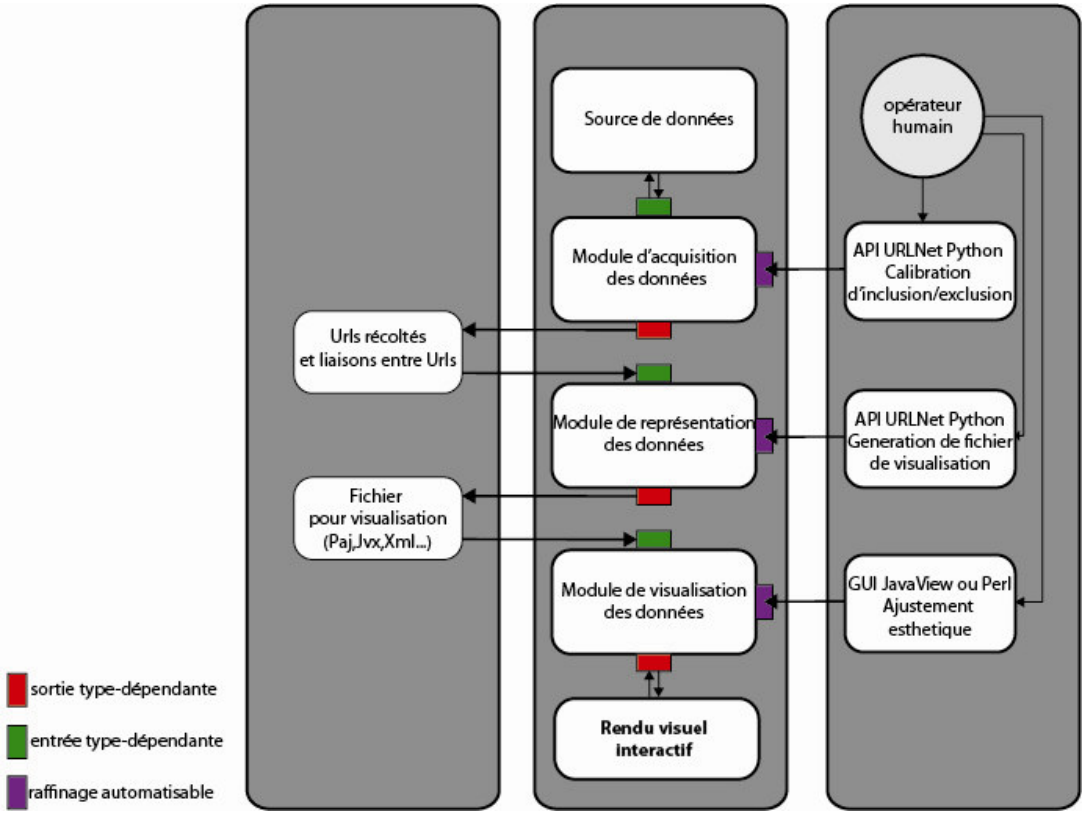

Figure 2 : Chaine de traitement - depuis le crawl jusqu'à l'affichage, l'expertise humaine peut intervenir pour améliorer l'induction empirique a priori

2.1.2 Information Overload , Automation et flexibilité

La surcharge informationnelle, dans notre problématique, constitue une masse qu'il faut gérer. L'opérateur, en tant qu'individu humain -cognitivement limité- dispose d'un atout : la possibilité de l'automation. La nature matérielle de l'outil informatique posant une première contrainte stricte : organiser l'information potentiellement pléthorique dans un espace de représentation limité, les premières intentions de recherche dans le domaine de la visualisation d'information ont été d'automatiser cette optimisation spatiale. En effet, la majorité des outils propose une série de modèles classiques (treemaps, cartes auto-organisées, algorithmes basés sur la force…).

Dans le cas particulier des algorithmes basés sur la force, les atomes de l'ensemble représenté sont uniformément répartis dans l'espace (2D ou 3D). Nos séries de prototypages nous ont amenés à constater que, malgré l'utilisation conjointe de procédés interactifs (navigation, zoom…) censés permettre une meilleure appréhension d'un le lot de données pertinentes au sein d'une masse conséquente, de couleurs permettant la mise en exergue desdites données pertinentes, on assiste un phénomène d'occlusion visuelle. Afin de passer outre ce phénomène et permettre la possibilité d'une perception, d'une préhension des données intéressantes, nous profitons de la

flexibilité offerte par les fichiers de données XML et du langage Perl afin d'automatiser des modifications des données de représentation spatiale des atomes. La figure N illustre la création d'un fort contraste spatial par translation des données pertinentes par opposition aux autres données qui restent groupées de façon homogène suivant l'algorithme *Früchterman-Rheingold* initialement utilisé. Pour sa grande flexibilité et la parfait adéquation avec nos attentes, nous avons opté pour le module de visualisation « JavaView »

2.1.3 Fondements pour une bonne représentation visuelle des données

2.1.3.1 De la cartographie, des lieux et de l'imagination

Lorsque l'on conçoit, que l'on apprend à concevoir une carte, il faut toujours garder à l'esprit une des caractéristiques fondamentales de l'outil cartographique : celui-ci utilise un langage visuel dont les principes, les règles, les qualités, les limites résultent tous les exigences physiologiques de l'œil humain (Poidevin, 1999).

Quel écolier ne s'est pas retrouvé perdu, noyé par un flot d'informations durant l'étude d'une discipline au cours d'une année scolaire ? Et quelle solution immédiate peut-on lui fournir ? C'est là la problématique plus globale de l'individu face à un flot –voire un torrent, *a fortiori* de nos jours- d'informations, problématique que nous faisons notre dans cette étude de cas aux proportions informatiques. Perdu, voilà le mot juste. Pourtant, depuis les antiques *Ars memoriae* de l'époque de Cicéron jusqu'aux récentes cartes heuristiques de Tony Buzan, l'humain s'efforce d'organiser spatialement le savoir (Yates, 1987) (Cicéron – Yon, 2003). Tony Buzan formulera récemment les principes de crochets ou liaisons qui incitent naturellement l'individu à créer un lien de toute nouvelle information assimilée avec les informations déjà en mémoire (ou sur support) (Buzan, 2004), et si Poidevin évoque l'œil qui perçoit des images, nous souhaitons revenir sur les solutions historiques trouvées pour palier à la surcharge informationnelle, laquelle existait déjà avant l'invention de l'imprimerie.

Les orateurs et bien d'autres se sont penchés sur l'assimilation de l'information par l'être humain. Nous retenons un principe fondamental des *Ars Memoriae: « La mémoire artificielle***1** *est fondée sur des lieux et des images2. »* (Yates, opt. Cit.). Ainsi, avant même de songer aux exigences physiologiques de l'œil, nous devons nous pencher sur les exigences cognitives de l'humain à traiter les images et les lieux.

Des générations d'élèves et de professionnels assimilent encore la géographie et indirectement la cartographie à des disciplines d'inventaires dont le seul but serait de situer les lieux, les faits et les phénomènes. Cette vision limitée et fortement stéréotypé vient du fait que l'école et l'enseignement n'ont généralement pas été préparés à transmettre l'utilité opérationnelle de la géographie et de la cartographie. Il en résulte qu'en tant que support pédagogique, les cartes novice trop souvent qu'à répondre à la question « où ? » (Poidevin, 1999).

De deux choses l'une : d'une part, l'inclination naturelle susmentionnée à relier les informations nous montre que la cartographie a un rôle majeur à jouer quant à l'organisation mentale de l'information puisque l'être humain, historiquement, associe l'information à des lieux mentaux (*Loci*). D'autre part, pour en revenir à cet écolier perdu, nous pensons que la carte, dans ce contexte, doit nécessairement répondre à la question *« où ? » -où sui-je, où vais-je raccrocher ce nouveau savoir par rapport à mon propre savoir, ma propre logique ?-* Une fois la réponse apportée, les capacités cognitives naturelles peuvent se mettre en branle et la nouvelle information trouver sa place naturelle au sein de l'espace mental que l'individu s'est constitué depuis longtemps. Dans le cas contraire, l'individu est confronté à une problématique d'utopie (pas de lieu).

Conformément à ces principes, nous considèrerons toute représentation visuelle que nous construirions comme un lieu possiblement susceptible de trouver sa place au sein de représentations pré-existantes

En ce qui concerne le formalisme issu de la cartographie, nous retiendrons dans un certaine mesure les figurés ponctuels de Bertin (Bertin et al., 2005), qui sont des constructions graphiques qui ont soit un contour géométrique (cercle, carré, rectangle, triangle, losanges...), soit un contour expressif s'ils évoquent la forme réelle de la donnée représentée (un avion pour un aéroport, une croix pour une église...) qui ont prouvé leur efficacité à travers les cartes géographiques utilisent des figurés ponctuels, linéaires ou zonaux qui reposent sur trois formes élémentaires : le point, la ligne et l'aire qui sont les combinaisons de six variables visuelles, appelée également variable rétinienne, mesurée et étudiée par Bertin (Ibid.) tel qu'illustré en figure 1.

1 Artificielle : issue de l'art (de mémoriser) par opposition à la mémoire naturelle

2 Image : image construite par faculté d'imagination tant et si bien que « *L'art de mémoire est comme une écriture intérieure* » (Yates, Op. Cit.), concept rejoignant celui de Franck Herbert qui évoque un « œil intérieur » (Herbert, 2005)

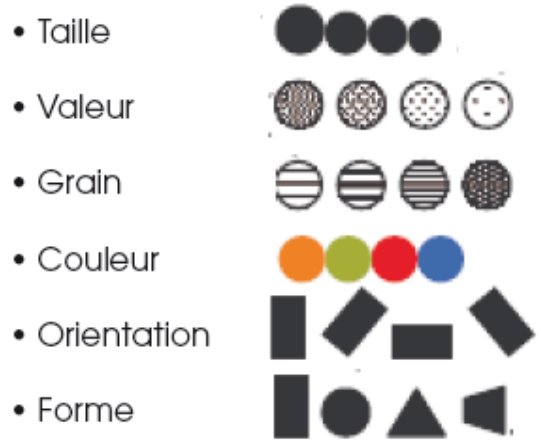

Figure 3 : les six variables rétiniennes selon Bertin

2.1.3.2 De la cognition et de l'infovis

Comme notre représentation présente les limites de l'outil informatique, nous devons prendre en compte les règles propres à l'infovis. Ainsi, la dimension Z, dans un espace tridimensionnel, au sens figuré, désigne la composante de qualification permettant de délimiter les objets géographiques à travers leurs relations : proportionnalité, ordre ou différence/association (id.). Comme nous souhaitons justement mettre en exergue les relations entre les différents sites web membres du réseau, nous avons opté pour une représentation comprenant ad minima trois dimensions spatiales, ce qui exclut de fait les représentations surfaciques classiques à deux dimensions (treemap pourtant propices pour des représentations de grands nombres d'éléments, carte auto-organisée de Kohonen…).

2.1.3.2.1 Interaction

Selon l'approche écologique de la perception due au psychologue J.J. Gibson (1979), la perception est indissociable de l'action : il faut agir pour percevoir et il faut percevoir pour agir (Hascoët et al., 2001). Il apparait donc souhaitable que l'utilisateur puisse interagir avec les données modélisées.

2.1.3.2.2 Limitations

Dès 1956, George Miller pose, dans son célèbre « *The Magical Number Seven, Plus or Minus Two* » (Miller, 1956) que la capacité cognitive humaine est classiquement de l'ordre de sept, plus ou moins deux items, capacité nommée empan mnésique de nos jours. Nous savons par ailleurs que cette limitation de sept, plus ou moins deux prend un sens plus global selon Bandler et Grinder (Bandler – Grinder, 2005) Nous avons donc essayé de restreindre notre système de modélisation à un périmètre de dimensions de cet ordre. Les trois dimention d'un espace tridimentionnel, une quatrième résidant dans des possibilités d'interaction ainsi que la colorisation nous ont semblé évidemment les premiers choix pour lesquels opter lors de notre phase de prototypage.

2.1.3.2.3 Optimisation spatiale et esthétique en trois dimensions

Depuis plus de deux décennies, les algorithmes basés sur la force tel que l'algorithme Fruchterman-Reingold sont connus pour être propices à une modélisation de graphes en deux ou trois dimensions, esthétique et simple à mettre en œuvre : le principe de tels algorithmes est de considérer les nœuds comme des ressorts exerçant des forces les uns sur les autres, le tout résultant en graphes dont les arcs prennent des longueurs homogènes et s'entrecroisent au minimum tout en optimisant l'organisation spatiale tel qu'illustré en figure 2. La problématique de cette organisation

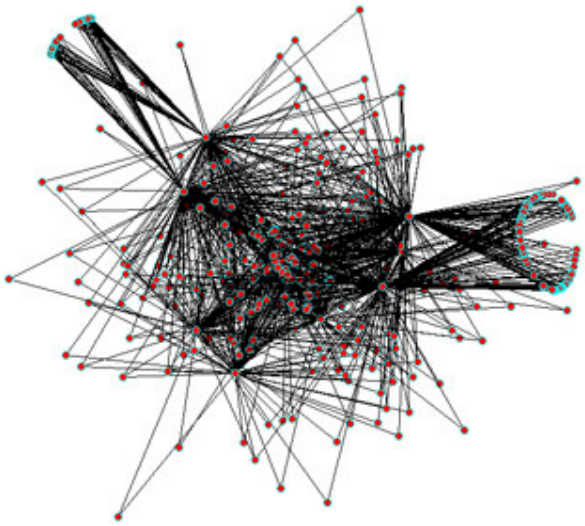

Figure 4 : la blogosphère francopohone de l'i.e en 2008- Algorithme Fruchterman-Reingold 3D

3 Résultats prototypés

3.1 Mesure réticulaire directe de niveau 4

Après que notre module de *crawling* a récolté toutes les pages jusqu'à un niveau de profondeur 4 des 53 sites web du réseau RMEI tel que http://www.domaine.com/niveau1/niveau2/niveau3/niveau 4, notre module de représentation des données a retenu l'ensemble des 53 sites ainsi que l'ensemble des liens existant d'un de ces sites vers un autre. Il a également calculé les coordonnées en 3 dimensions de l'ensemble obtenu selon l'algorithme Früchtermann-Rheingold, puis le module d'affichage a permis d'obtenir la figure 5. On s'aperçoit que la structure réticulaire de l'ensemble est inexistante.

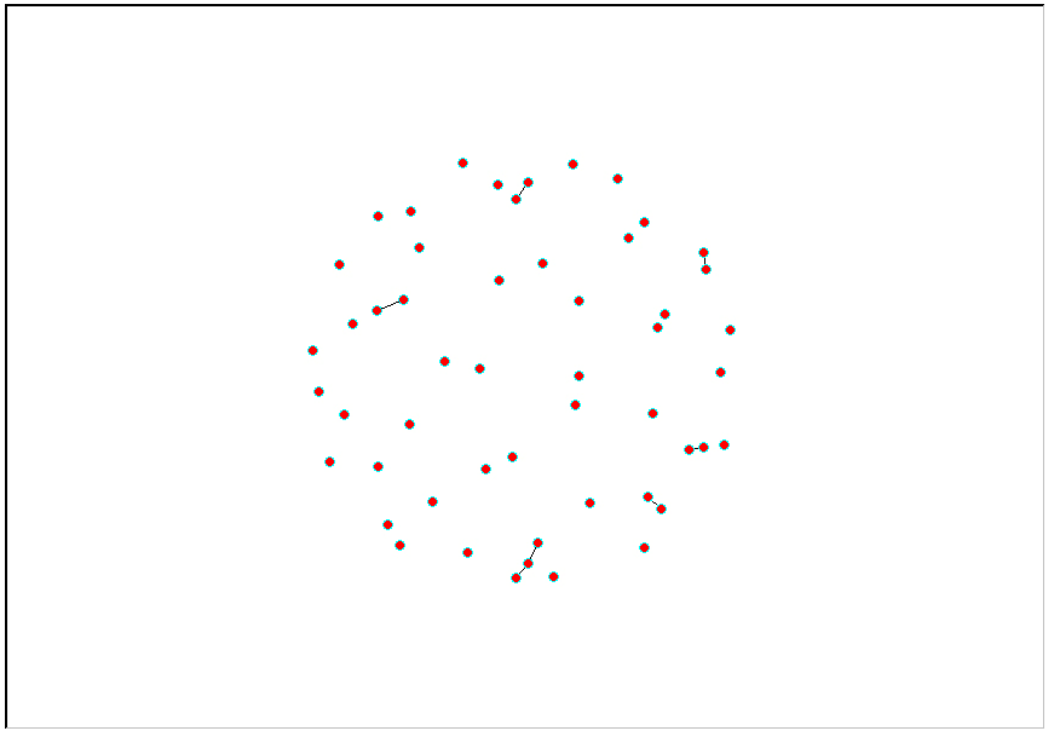

Figure 5 : structure réticulaire virtuelle des 53 sites du RMEI – algorithme Früchtermann-Rheinglod – Module de visualisation Javaview dans sa déclinaison applet

3.2 Mesure réticulaire indirecte de niveau 4

Afin d'approfondir notre connaissance quant à la structure réticulaire du réseau RMEI, nous parcourons le site de l'association RMEI http://rmei.info jusqu'à un niveau 4 et faisons de même pour tous les noms de domaines rencontrés lors de ce parcours. Finalement, notre module de crawl a parcouru 3758 sites web sur une profondeur de quatre niveaux. Les informations récoltées par notre crawler (domaines et liens entre domaines) sont interprétées par notre module de présentation des données qui calcule la position en trois dimensions selon l'algorithme Früchterman-Rheingold de ces domaines ainsi que leurs liens.

3.2.1 Extraction visuelle des données

Malgré l'affichage en trois dimensions, la possibilité d'interaction (zoom, rotation, décalage…) et des points de couleur distincte (en violet sur la figure 6) des 53 sites du RMEI, la masse obtenue ne permet pas de distinguer les relations entre les 53 points qui nous intéressent. L'algorithme Früchterman-Rheingold homogenise les données dans l'espace. Aussi, nous décidons de nous servir de la flexibilité du format de fichier généré (Format Jvx qui est en fait du XML) en créant un script Perl qui modifie les coordonnées des points du lot intéressant en augmentant simplement leur valeur, ce qui revient à une opération de translation qui sera désormais automatisée en cas de forte charge informationnelle : l'induction empirique de notre module de représentation devient adaptée à la masse. Nous avons créé une différence de densité spatiale qui se traduit par un contraste évident tel qu'illustré en figure 6. Nous remarquons que la structure réticulaire est toujours inexistante.

4 Conclusion

L'élaboration de notre propre chaîne de traitement nous a permis, grâce à la flexibilité des technologies utilisées et conjointement à la connaissance de la cognition ainsi qu'au savoir-faire en matière de cartographie et d'*infovis*, d'obtenir un outil capable de crawler rapidement et de façon calibrée ainsi qu'un module de modélisation et un autre de visualisation en mesure de fournir des représentations convenables, nous permettant de tirer une connaissance immédiate de cette masse de données et *in fine* de répondre à la question initialement posée : *le réseau virtuel RMEI est-il vraiment existant en terme de collaboration ou non ?* – A la lumière de notre analyse, nous pouvons en déduire que ce n'est pas le cas. L'outil ainsi créé fait toujours l'objet de prototypages. Vu la grande latitude permise, notamment du module de visualisation, nous pensons être en mesure de l'améliorer encore et, idéalement, de provoquer les mécanismes propres aux « lieux » et aux « images »

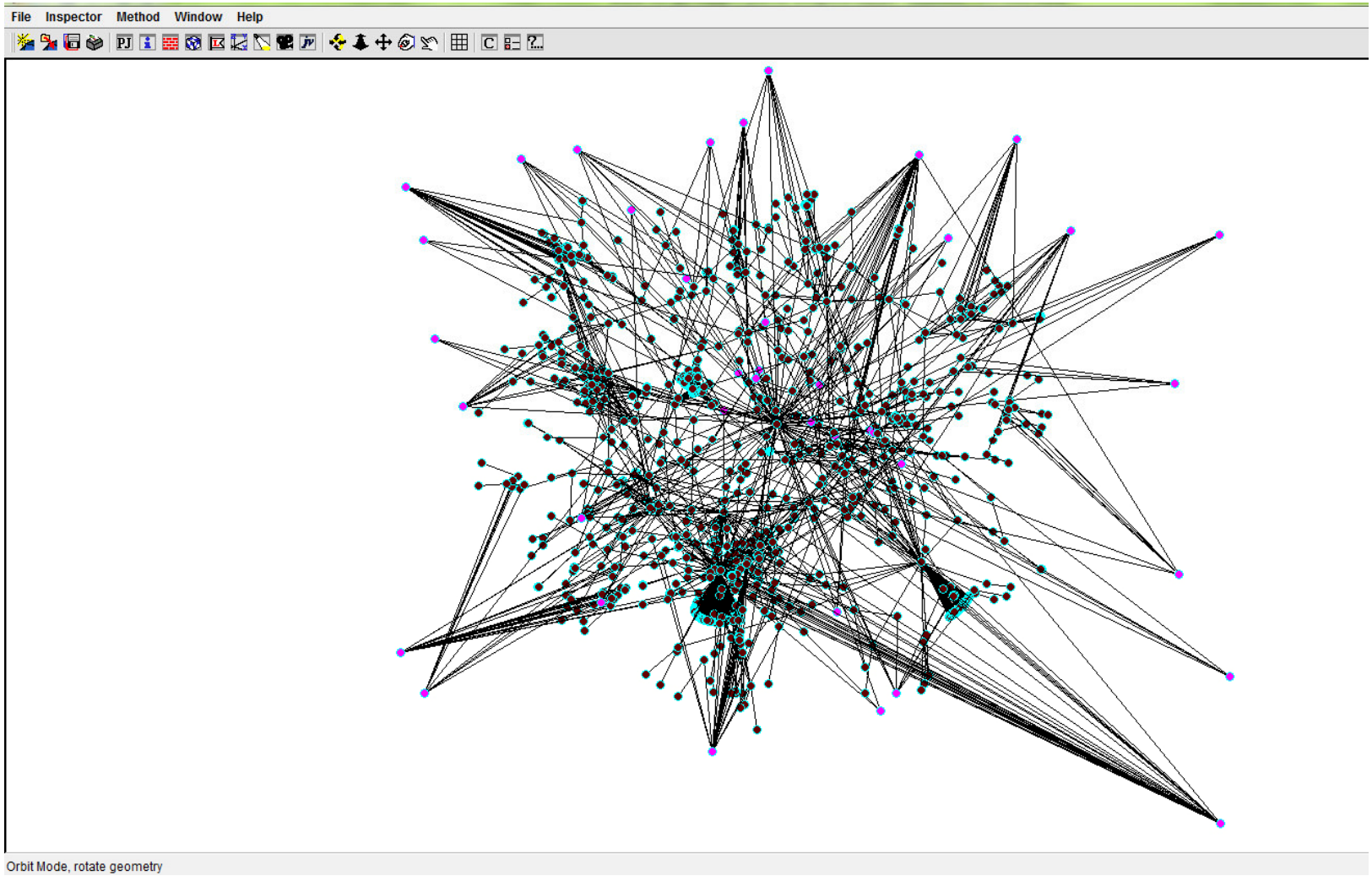

Figure 6 : les membres du RMEI colorisés en mauve sont aisément discernables après translation hors de la masse – Javaview en mode Standalone

5 Bibliographie :

[1] BANDLER R., GRINDER J., et ANDREAS C., *Transe-Formations : Programmation NeuroLinguistique et techniques d'hypnose éricksonienne*, DUNOD, 2005.

[2] BERTIN J. et Collectif, *Sémiologie graphique : Les diagrammes, les réseaux, les cartes*, EHESS, 2005.

[3] BUZAN T., *Tout sur la mémoire*, EDITIONS D'ORGANISATION, 2004.

[4] CARLSON C., *Information Overload, retrieval strategies and internet user empowerment*, RCLIS.ORG, 2003; http://eprints.rclis.org/archive/00002248/01/Information_Overload.pdf

[5] CICERON et YON A., *L'Orateur*, BELLES LETTRES, 2003.

[6] DOU H. et coll., De *la création des bases de données au développement de systèmes d'intelligence pour l'entreprise*, ISDM, Mai. 2003; http://isdm.univ-tln.fr/PDF/isdm8/isdm8a67_penteado.pdf

[7] HASCOËT M. et BEAUDOIN-LAFON M., *Visualisation interactive d'information*, INFORMATION – INTERACTION - INTELLIGENCE, vol. 1, 2001; http://www.revue-i3.org/volume01/numero01/article01_01_03.pdf

[8] HEARST M.A., *Untangling text data mining*, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, College Park, Maryland: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 1999, pp. 3-10; http://portal.acm.org/citation.cfm?id=1034678.1034679&coll=Portal&dl=GUIDE&CFID=67534762&CFTOKEN=33126076

[9] HERBERT F., *Dune - Tome 1*, POCKET, 2005.

[10] KROEZE J.H., MATTHEE M.C., et BOTHMA T.J.D., *Differentiating data- and text-mining terminology*, Proceedings of the 2003 annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology, SOUTH AFRICAN INSTITUE FOR COMPUTER SCIENTISTS AND INFORMATION TECHNOLOGISTS, 2003, pp. 93-101; http://portal.acm.org/citation.cfm?id=954014.954024&coll=Portal&dl=GUIDE&CFID=67534762&CFTOKEN=33126076

[11] MILLER G. A., *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, THE PSYCHOLOGICAL REVIEW, vol. 63, 1956, pp. 81-97

[12] POIDEVIN D., *La carte, moyen d'action : Conception - réalisation*, ELLIPSES MARKETING, 1999.

[13] PIATETSKI-SHAPIRO G. et FRAWLEY W., *Knowledge Discovery in Databases*, MIT PRESS, 1992.

[14] YATES F.A., *L'Art de la mémoire*, GALLIMARD, 1987.