La visualisation d'information au service de la veille concurrentielle, de la fouille d'information et de la supervision de systèmes complexes

Guy Melançon, Christophe Douy

La visualisation d'information au service de la veille concurrentielle, de la fouille d'information et de la supervision de systèmes complexes

Christophe Douy
PIKKO
cdouy@pikko-software.com

Guy Melançon
INRIA Futurs & CNRS UMR 5800 LaBRI
Guy.Melancon@labri.fr

Introduction

« Crises, accélération des mutations économiques et sociales, globalisation. Le monde est de plus en plus complexe. Le dirigeant d'entreprise rencontre des difficultés croissantes pour garder le cap. Ses traditionnels instruments de navigation deviennent obsolètes. Entre tempête et brouillard, la maîtrise de l'information devient un enjeu stratégique pour éviter les écueils. »
C'est en ces termes que [Corman and Ingargiola 2005] décrivent le contexte dans lequel se trouvent aujourd'hui toutes les organisations. Dans un contexte de surcharge d'information au quotidien, les acteurs confrontés à un tissu informationnel complexe sont à la recherche de leviers leur permettant de les aider dans leurs prises de décision. En 2001, le projet « How much information » de l'université de Berkeley quantifiait cette surcharge et évaluait l'information produite cette année-là à un exabyte (1 million de terabytes) de données dont 99,997% sont disponibles sous forme électronique seulement (voir [Keim 2001]). En 2003, la quantité d'information rapportée à la population mondiale correspondait à 800 Mo de nouvelles données produites par personne en un an [Peter and Varian 2003].
Ces préoccupations ne sont pas nouvelles : l'intelligence économique est une activité ancienne qui a été très longtemps à la frontière du monde du renseignement industriel et de l'analyse marketing concurrentielle [Carrez and Carayon 2003; Brute de Rémur 2006]. Le développement important des réseaux informatiques, et en tout premier lieu d'Internet, a induit une nouvelle stratégie de communication et la recherche d'une plus grande transparence de la part des organisations, des entreprises et des institutions. De fait, la volumétrie d'informations « en ligne » des organisations est devenue beaucoup plus importante que celle des documents et rapports qu'elles publiaient il y a quelques années.
Le terme de « gisement d'informations » à propos d'Internet prend donc tout son sens et a pour conséquence le développement d'une nouvelle activité consistant à collecter, analyser les informations disponibles afin de fournir de nouveaux outils d'aide à la décision aux acteurs des organisations. De plus, la rapidité de diffusion liée aux médias électroniques a développé un besoin spécifique d'alertes : les responsables concernés veulent être prévenus dès qu'un phénomène nouveau apparaît. Cette activité de collecte et d'analyse n'a que très peu de points communs avec la recherche d'informations telle que la pratique un documentaliste : la surveillance et la collecte s'effectue sur des domaines peu ou pas connus et des outils sont mis en place pour détecter des informations nouvelles et non attendues. Il est rare qu'une organisation prévienne ses concurrents de ses changements d'orientation ou de sa stratégie !
Dans ce contexte où le temps est compté, les organisations misent sur le traitement automatisé de l'information, et souhaite surtout ne pas passer à côté d'une information clé ou d'une opportunité. Cette attente cache une complexité que la technologie doit aider à résoudre : une information, même marginale à un instant donné, peut évoluer en un élément clé du système étudié. Le concept de signaux faibles introduit par Ansoff [Ansoff 1976] capture cette réalité. Le signal faible est une information à caractère anticipatif ; sa présence se confirme en portant attention aux éventuelles discontinuités et ruptures pouvant se produire dans l'environnement d'une organisation. Le signal faible s'observe si on suit la trace des informations se rapportant à un évènement, par exemple.

Etat de l'art

Face à ces exigences, les organisations sont en quête de solutions de gestion de l'information dont le retour sur investissement devient un facteur clé de compétitivité, de réactivité et d'efficacité.

Les moteurs de recherche : un point de départ pour l'analyste

Confortée par la part grandissante du web dans la masse d'informations disponibles, la réponse actuelle à ces besoins de traitement privilégie l'utilisation de moteurs de recherche. Apparus il y a plus de dix ans, avec Google en tête [Brin and Page 1998], ils fournissent un déluge de réponses « pertinentes » aux requêtes de l'utilisateur avec une rapidité stupéfiante. Si ce type d'interaction peut convenir dans un grand nombre de cas, il faut néanmoins constater que certaines problématiques ressortant de l'analyse de l'information disponible restent non résolues :

- Il n'y a pas de présentation synthétique de l'information fournie par le système.
- Les rapprochements entre informations doivent être faits par l'utilisateur.
- Les signaux faibles sont généralement noyés dans la masse.
- Enfin, même filtrée, la quantité d'information à traiter reste importante et prive l'utilisateur de l'efficacité du moteur de recherche.

Les évolutions actuelles

Des logiciels spécialisés dans la veille concurrentielle et l'intelligence économique sont apparus ces dernières années. Tous s'appuient, pour ce qui relève de la fouille de texte, sur des moteurs de recherche. Pour réduire la charge cognitive qui pèse sur l'utilisateur, la plupart des applications proposent des fonctions d'analyse de l'information collectée. Ces fonctions permettent, entre autres, d'effectuer des analyses sémantiques de l'information (recherche de noms propres, par exemple), et de regrouper automatiquement des documents ou des informations proches (on parle de classification ou de clustering).

La visualisation comme support à l'analyse de l'information

Cette restitution sous forme graphique « en bout de chaîne » est encore aujourd'hui à sens unique : elle n'est pas utilisée comme support de l'analyse. Or l'activité de veille se fait en fouillant les informations enfouies dans une masse de données complexes. Cette complexité doit être gérée en fragmentant l'espace à fouiller, et en l'organisant par niveaux de détails. Les approches visuelles permettant d'explorer et d'exploiter ces informations connaissent un succès grandissant. Elles misent sur la capacité de l'œil humain à traiter efficacement des informations graphiques en y identifiant des motifs structuraux complexes. Selon Ware [Ware 2000], 40% des activités de notre cortex sont en effet dédiées au stimuli visuels. Il s'agit donc de tirer profit de ce potentiel d'analyse visuel en offrant à l'utilisateur la possibilité de « naviguer » dans ces espaces fragmentés et de pouvoir obtenir des informations détaillées « à la demande ».
La visualisation d'information s'inspire d'idées ancrées dans des traditions d'origines diverses, dont la statistique graphique, la cartographie, le graphisme par ordinateur, l'interaction homme-machine, la psychologie cognitive, la sémiotique, le design graphique et l'art graphique [Tufte 1983; Tufte 1990; Tufte 1997], [Card, Mackinlay *et al.* 1999; Herman, Marshall *et al.* 2000; Spence 2001]. Lorsqu'on lance une requête sur Google Scholar avec les mots-clés « large information space », c'est sans surprise que l'on trouve parmi les liens les plus pertinents (au sens de Google) les références aux travaux actuels en visualisation d'information.
La complexité dont hérite ainsi l'utilisateur doit en revanche s'accompagner d'outils à l'aide desquelles il interagit sur les représentations graphiques mises à sa disposition. Il lui faut en effet

pouvoir délimiter un périmètre de fouille, contribuer à le structurer et annoter les éléments de données. Le travail de veille s'insère ainsi dans une boucle où, en visualisant l'espace, un modèle évolue et se construit de manière itérative, comme le suggère la figure 1.

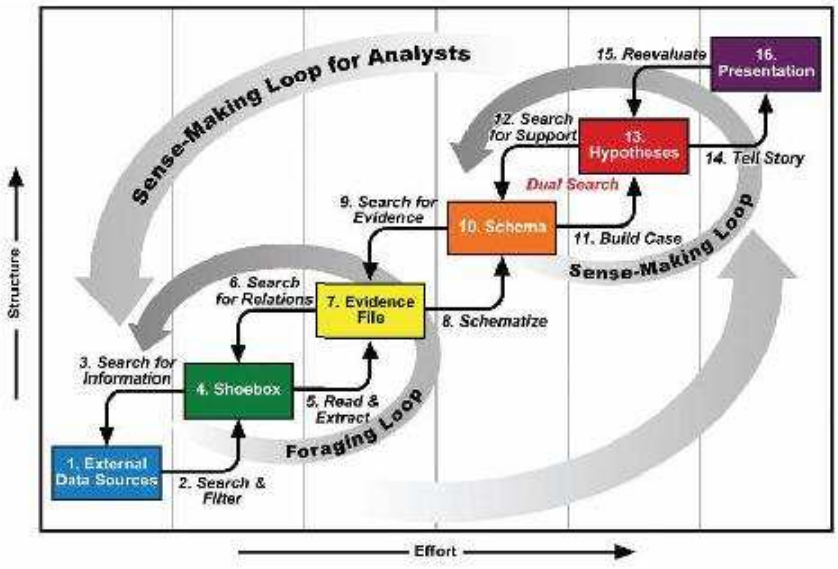


Figure 3. le "sense-making loop" selon [Thomas and Cook 2006].

Les acteurs de la visualisation

En tout état de cause, les Etats-Unis regroupent une masse critique importante de chercheurs dans le domaine de la visualisation d'information et de la visualisation scientifique. C'est aussi sur le continent américain que l'on compte le plus de sociétés développant et proposant des produits qui intègrent les technologies issues de la visualisation d'information (TOWSAWYER, INXIGHT, SMARTMONEY, HIVEGROUP, etc.).

Dans ces outils de première génération, une métaphore visuelle unique était développée le plus souvent. De plus, l'accent n'était pas mis sur la portabilité et l'intégration du composant de visualisation à un système d'information existant en entreprise, mais reposait le plus souvent sur une technologie dédiée. Ces sociétés ont néanmoins été précurseurs dans leur domaine.

Avec l'engouement actuel pour des outils de cartographie traditionnelle ("terrestre") en ligne (portail IGN en France, ou GoogleMaps) ou embarqués (guidage GPS), le concept de visualisation interactive bénéficie aujourd'hui d'un intérêt plus "grand public". Ainsi, le domaine de la visualisation sort peu à peu de la confidentialité.

Aujourd'hui le marché est mûr pour une nouvelle génération d'outils visuels, qui simplifient l'exploration ou l'analyse de données nombreuses, hétérogènes et/ou changeantes. De nouveaux acteurs adoptent aujourd'hui un positionnement précurseur, en proposant la mise en place de plates-formes exploitant à fond la métaphore visuelle.

Les réponses techniques

La notion fondamentale autour de laquelle les solutions techniques s'articulent est la « proximité » :

- proximité sémantique entre deux fragments gisant dans un univers d'information ;
- proximité entre deux acteurs dans un réseau.

C'est bien cette notion qui guide les calculs des moteurs de recherche. Les astucieux algorithmes et heuristiques qu'ils mettent en œuvre juge de la pertinence d'un document face à une requête : il s'agit bien d'évaluer la proximité entre le contenu d'un document et les mots-clés ou les concepts utilisés pour construire la requête.

L'exploration visuelle de données mise sur la découverte de motifs structuraux pour guider l'analyse : regroupement d'éléments de données en zones distinctes, effets de couleur ou de taille, etc. En questionnant le motif, l'analyste est à même de découvrir les ingrédients propres aux données qui expliquent la présence du motif. Cette vision est simpliste, puisque le processus ne se déroule pas en une seule étape : la plupart du temps, le motif donnera des indications sur la manière dont l'interprétation des données ou des attributs doit être rectifiée, amenant un ajustement des motifs, eux-mêmes sources de nouveaux questionnements. Cette démarche itérative (Figures 1 et 3) convergera vers une analyse argumentée, où la visualisation sera *in fine* utilisée à des fins de publications et de diffusion de la découverte. La visualisation est donc présente à toutes les étapes du processus d'analyse, depuis la découverte « en aveugle », jusqu'à sa diffusion, contribuant tout au long du parcours à sa construction.

Afin de construire une vue où pourront apparaître ces motifs, les données d'entrée sont comparées, leur proximité est calculée (on parle aussi de *similarité*, parfois même de *distance* dans une acception très large). Une première approche consiste à ne prendre en compte que les proximités les plus significatives, faisant alors émerger un réseau de liens entre certaines informations. L'objet mathématique correspondant est un graphe : des sommets correspondant aux données extraites de l'espace d'information, et des arêtes liant ces sommets auxquelles peuvent être associés des attributs reflétant la nature ou l'intensité de la relation. On peut alors mobiliser tout l'attirail de la théorie des graphes et du dessin de graphes [Battista, Eades et al. 1998; Kaufmann and Wagner 2001] pour construire une vue enrichie d'attributs graphiques.

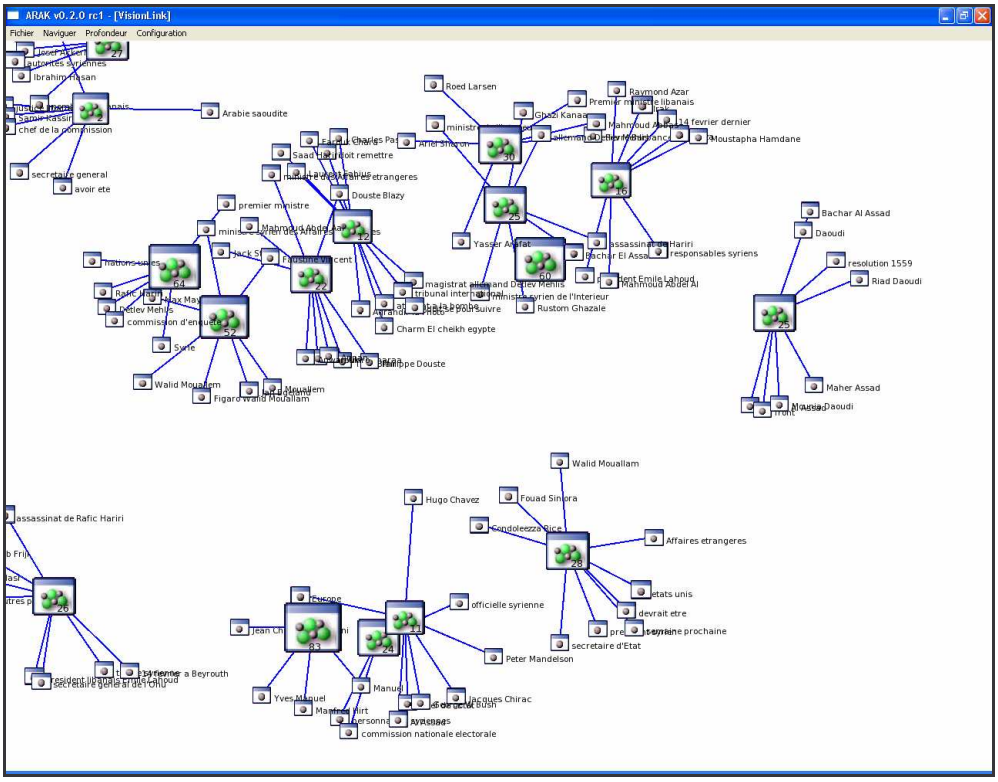


Figure 4. VISIONLINK : un composant de visualisation associative développé par PIKKO.

C'est ce que montre la figure 2. Bien qu'on puisse juger de la proximité entre toutes paires d'éléments d'information, seuls les liens entre données jugées assez proches sont retenus. L'exemple regroupe des fragments d'information extraits du web concernant l'actualité (« Rafiq Hariri »). Le processus d'extraction a fait émerger des entités nommées (personnes, lieu, date) et des « concepts ». Ces fragments sont alors regroupés selon leur proximité en termes de co-occurrences dans les documents analysés et peuvent faire apparaître des liens attendus ou non-attendus entre éléments de l'actualité. L'algorithme de dessin tient compte de ces proximités et regroupe automatiquement les éléments formant des groupes thématiques. Les groupes eux-mêmes peuvent ensuite devenir les sommets d'un graphe-résumé donnant une vue globale sur l'organisation des fragments.

Dans certains cas, le découpage des informations en catégories thématiques est donné. Ce découpage se fait le plus souvent selon un schéma arborescent hiérarchisant les catégories et sous-catégories. Les cartes interactives (« treemaps ») tirent profit de cette organisation arborescente (voir la figure 4 plus loin).

La proximité entre éléments de données peut aussi faire l'objet d'un traitement en amont de la visualisation. Les éléments peuvent être agrégés selon des procédures diverses (voir [Nakache and Confais 2005]). Ces opérations sont encore aujourd'hui l'objet de travaux de recherche ; de nombreux travaux se penchent sur l'analyse de réseaux, cherchant à mettre à profit la théorie des réseaux [Bornholdt and Schuster 2003; Dorogovtsev and Mendes 2003; Newman 2003; Brandes and Erlebach 2005]. La difficulté tient souvent à concilier ce qui peut être déduit de la structure du graphe ou de la distribution statistique des proximités, avec la sémantique du domaine étudié.

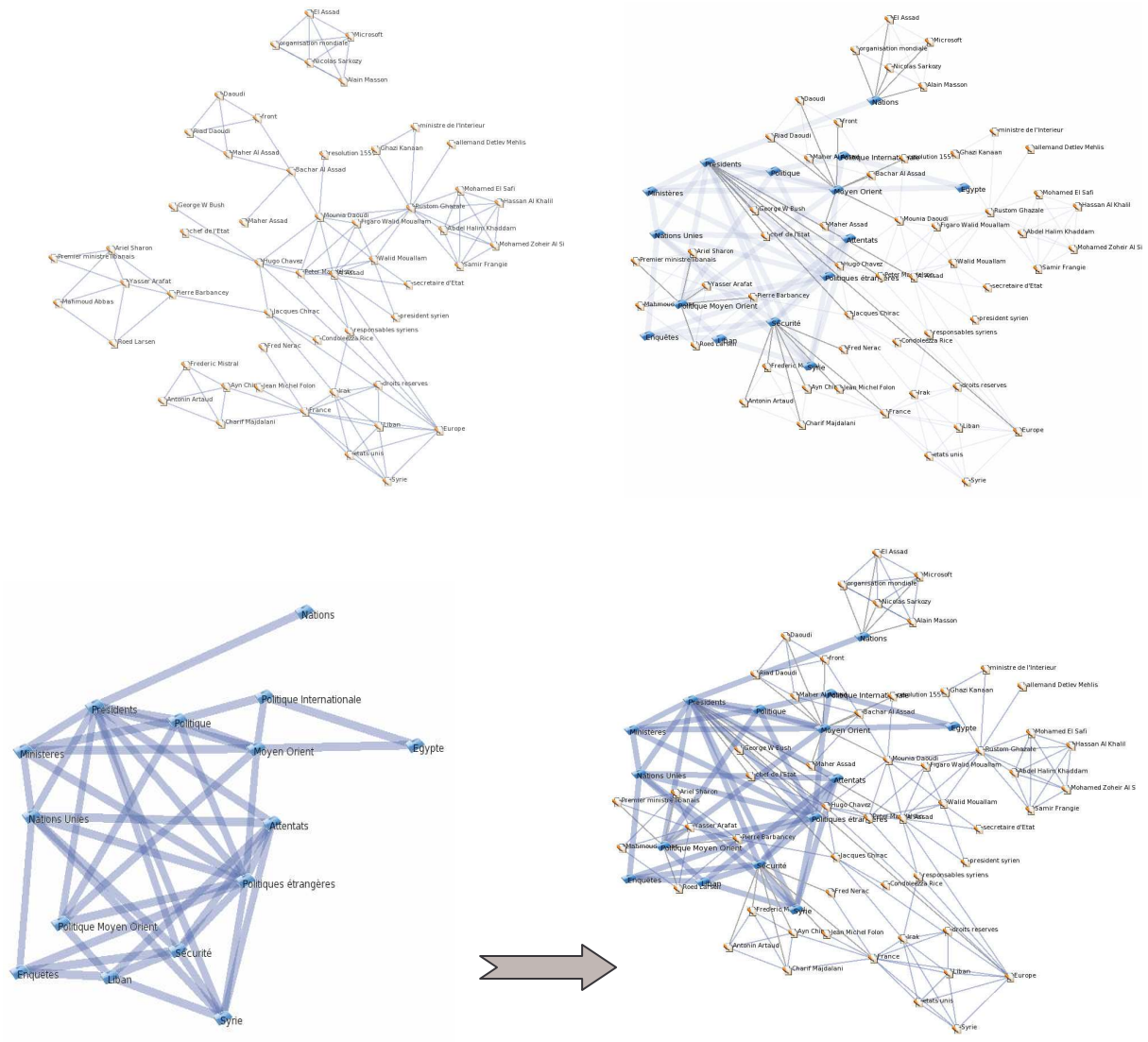


Figure 5. VISIONLINK : du détail au concept

Indice de cohésion et densité des voisinages

Un exemple plus précis illustrera notre propos. Supposons donné un graphe qui résulte du seuillage des valeurs de proximités : seuls restent les liens entre données jugées suffisamment proches. Notons que cette opération n'a pas pour effet de diminuer le nombre d'éléments d'information considérés. Un

second processus de filtrage peut alors être mis en œuvre pour tirer partie de la topologie du graphe, visant à substituer ce graphe de départ par un graphe formé de moins de sommets, dans le but de gagner en lisibilité. Les régions les plus denses du graphe pourront alors être représentées par un seul et unique sommet dont la taille est proportionnelle au nombre de sommets sous-jacents, par exemple.
A cette fin, on calcule pour chaque arête du graphe une statistique permettant de juger de son rôle dans le réseau :

- soit elle contribue à la cohésion du voisinage,
- soit elle est un chemin de passage entre deux régions plus densément connectées.

Les régions plus densément connectées sont susceptibles de rassembler des ressources qui renseignent un même thème, voire qui sont redondantes. L'identification des arêtes-transition permet à l'inverse de donner une lecture de l'articulation entre les thématiques qui émergent des voisinages denses du graphe.

Les applications et leurs débouchés commerciaux

De nombreuses applications liées aux problématiques de sécurité globale, dans un cadre étendu de "renseignement" peuvent profiter des techniques de visualisation. Ainsi, les technologies de cartographies décisionnelles (p.ex. la technologie ARAK de l'éditeur PIKKO) permettent :

- la synthèse des ressources observées, grâce à la création de cartes de l'environnement synthétiques et interactives, sur le modèle des cartes géographiques,
- le couplage de ce type d'informations non géographiques aux traditionnels SIG, en passe eux de devenir une commodité dans de nombreuses organisations.
- l'exploration de réseaux volumineux ou complexes (réseaux d'individus ou d'entreprises)

Application à l'intelligence territoriale

L'intelligence territoriale se propose de relier la veille et l'action publique au service du développement économique et industriel d'un territoire. Les collectivités territoriales (les chambres de commerce, d'industrie ou d'artisanat ; services de développement économique au niveau région ou département p.ex) ont un besoin grandissant d'outils d'analyse de plus en plus en plus opérationnels pour :

- mieux gérer leurs territoires et mieux maîtriser leurs ressources ;
- développer une capacité de réaction rapide face à des situations de crise ;
- maîtriser leurs infrastructures (routes, réseaux..) ;
- recenser les bonnes pratiques d'autres territoires à des fins de « benchmarking » ;
- posséder la connaissance des savoir-faire et des produits du territoire pour réaliser un marketing territorial.

Des outils de visualisation d'information comme le composant EasyKube de la société PIKKO proposent une solution cohérente avec les objectifs et les besoins de la veille territoriale. La catégorisation des informations est restituée au travers d'une vue tabulaire multi-niveau exploitant un paradigme de visualisation, le *treemap*, ayant fait ses preuves [Johnson and Schneiderman 1991; Shneiderman 1992; van Wijk and van de Wetering 1999; Wattenberg 1999; Bruls, Huizing et al. 2000]. L'analyste qui emploie cet outil dispose d'une interface de visualisation d'information multidimensionnelle, lui permettant véritablement de plonger dans l'information pour l'explorer de façon interactive sous toutes ses facettes.

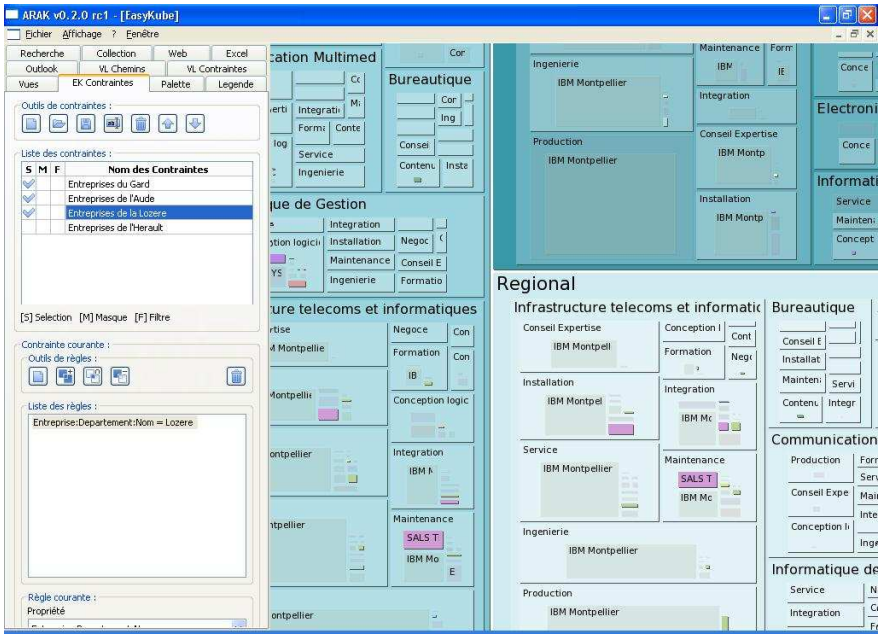


Figure 6. EASYKUBE, développé et commercialisé par PIKKO, propose une visualisation intégrée des informations hiérarchisées, et facilite l'accès aux informations à des fins de consultation ou d'annotation.

Application aux activités de renseignement

Dans le cadre de la gestion d'enquêtes et au sens large l'administration de bases dites de 'renseignements', dans des domaines variés, que ce soit celui de la lutte contre le blanchiment d'argent, du contre-espionnage industriel ou de la riposte contre-terroriste, il est nécessaire de construire intuitivement des bases d'analyse, manipuler facilement d'importants volumes d'informations et concevoir sa propre matrice d'analyse.

Solutions

La recherche de « grappes de connectivité » c'est à dire la détermination automatique de la présence de réseaux au sein d'un graphe complexe (graphe au sens d'interactions entre individus, et personnes morales) peut largement être facilitée par des outils de visualisation.

De même la recherche des liaisons les plus directes ou les plus plausibles entre deux entités (en fonction de critères paramétrables) peut apparaître immédiatement dans un logiciel mettant à profit la visualisation d'information (cf. figure 5).

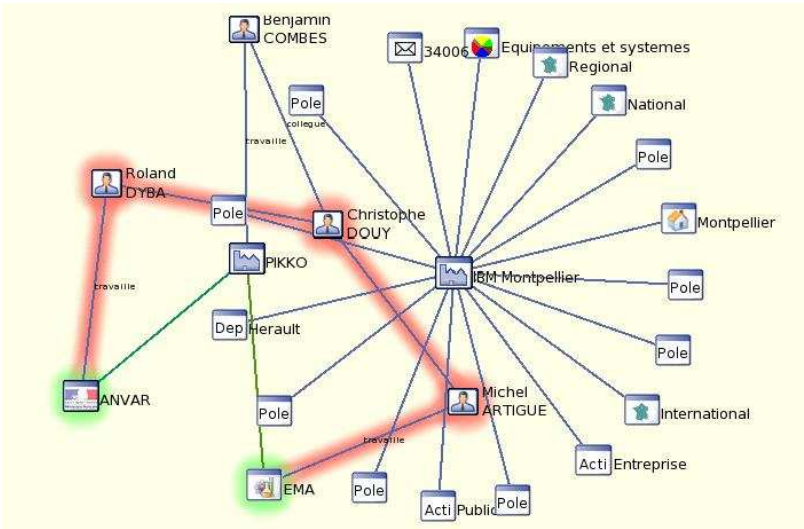


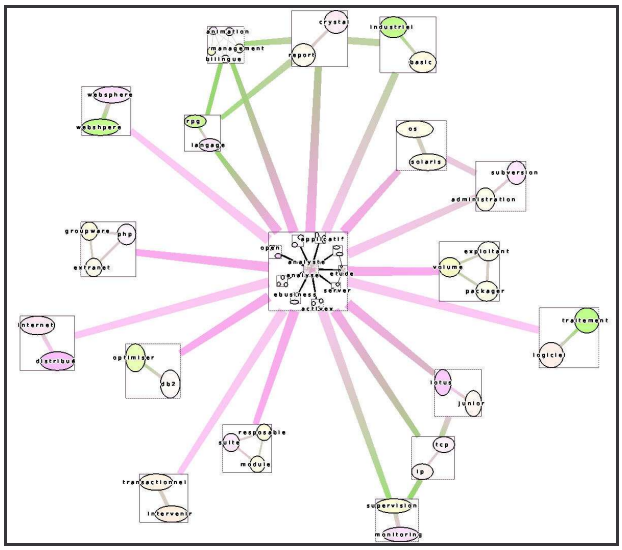
Figure 7. VisionLINK, recherche de chemins dans un réseau de personnes

Conclusion

La visualisation a indéniablement une place à se faire dans le contexte de l'exploration, de l'analyse et de supervision de masses d'information hétérogènes, et s'inscrire d'emblée parmi les outils et concepts de la sécurité globale. Typiquement, les activités de veille et de renseignement rassemblent des informations de sources diverses afin de voir poindre le germe d'une hypothèse enfouie dans la masse. La capacité à explorer visuellement, à trier l'information à même les métaphores graphiques et à organiser interactivement l'information répond à un réel besoin de l'analyste, et pendant toutes les étapes de son travail : des premiers pas de l'exploration, en passant par l'annotation, et jusqu'à la publication des hypothèses et des éléments confirmatoires.
Un axe de recherche nouveau consiste à envisager les informations à analyser comme un système complexe dynamique évoluant et s'organisant à différents niveaux d'échelle, comme c'est le cas dans de nombreux domaines.

- Notamment les réseaux multi-niveaux apparaissent utiles lorsqu'il s'agit de décrire les grands réseaux spatiaux en géographie quantitative [Amiel, Melançon et al. 2005] : réseaux de transports, réseaux de filiations d'entreprises, migrations alternantes.
- C'est le cas aussi dans les réseaux sociaux qui, bien que définis à partir d'interaction entre individus, contiennent et décrivent implicitement une dynamique inter-groupes.

La figure ci-dessous illustre le type de métaphores visuelles produites à l'heure actuelle avec le logiciel TULIP[120]. Le réseau sémantique sous-jacent à cette image a été éclaté selon des méthodes d'analyse des réseaux et de classification. L'image permet de décoder instantanément l'organisation de certaines composantes en périphérie du « cœur » de réseau. La navigation permet alors d' « entrer » dans chacune des composantes en suivant les échelles. L'exploration donne lieu à interprétation, des données, mais aussi de leurs mises en relation.
C'est ce type de navigation, des mécanismes sous-jacents algorithmiques, calculatoires mais aussi des interactions qu'il semble très prometteur d'appliquer aux problématiques de renseignement.



Bibliographie

Amiel, M., G. Melançon and C. Rozenblat (2005). "Réseaux multi-niveaux : l'exemple des échanges aériens mondiaux." M@ppemonde 79(3-2005).
Amiel, M., G. Melançon, *et al.* (2005). "Réseaux multi-niveaux : l'exemple des échanges aériens mondiaux." M@ppemonde **79**(3-2005).
Ansoff, I. H. (1976). "Managing strategic surprise by response to weaks signals." California Management Review **XVII**(2): 21-33.

[120] Voir le site www.tulip-software.org

Battista, G. d., P. Eades, *et al.* (1998). Graph Drawing: Algorithms for the Visualisation of Graphs, Prentice Hall.
Bornholdt, S. and G. Schuster, Eds. (2003). Handbook of Graphs and Networks: From the Genome to the Internet, Wiley-VCH.
Brandes, U. and T. Erlebach, Eds. (2005). Network Analysis. Lecture Notes in Computer Science 3418, Springer.
Brin, S. and L. Page (1998). "The anatomy of a large-scale hypertextual web search engine." Computer Networks and ISDN Systems **30**(1-7): 107-117.
Bruls, M., K. Huizing, *et al.* (2000). Squarified Treemaps. Joint Eurographics and IEEE TCVG Symposium on Visualization (Data Visualization '00), Amsterdam, Springer-Verlag.
Brute de Rémur, D. (2006). Ce que intelligence économique veut dire, Organisation Editions.
Card, S. K., J. D. Mackinlay, *et al.* (1999). Readings in Information Visualization. San Francisco, Morgan Kaufmann Publishers.
Carrez, G. and B. Carayon. (2003). "Rapport Carayon." from http://www.assembleenationale.fr/12/budget/plf2004/b1110-36.asp.
Corman, V. and E. Ingargiola (2005). Guide pratique : intelligence économique et PME. MEDEF. Paris.
Dorogovtsev, S. N. and J. F. F. Mendes (2003). Evolution of Networks : From Biological Nets to the Internet and WWW, Oxford University Press.
Herman, I., M. S. Marshall, *et al.* (2000). "Graph Visualisation and Navigation in Information Visualisation: A Survey." IEEE Transactions on Visualization and Computer Graphics **6**(1): 24-43.
Johnson, B. and B. Schneiderman (1991). Tree-maps: a Space-filling Approach to the Visualisation of Hierarchical Information Structures. IEEE Visualisation'91, IEEE CS Press.
Kaufmann, M. and D. Wagner, Eds. (2001). Drawing Graphs, Methods and Models. Lecture Notes in Computer Science, Springer.
Keim, D. A. (2001). "Visual exploration of large data sets." Communications of the ACM **44**(8): 38-44.
Nakache, J.-P. and J. Confais (2005). Approche pragmatique de a classification, Editions TECHNIP.
Newman, M. E. J. (2003). "The structure and function of complex networks." SIAM Review **45**: 167–256.
Peter, L. and H. R. Varian. (2003). "How Much Information." Retrieved March 2006, from http://www.sims.berkeley.edu/how-much-info-2003.
Shneiderman, B. (1992). "Tree visualization with tree-maps: 2-d space-filling approach." ACM Transactions on Graphics **11**(1): 92-99.
Spence, R. (2001). Information Visualization. Harlow, England, ACM Press/Addison-Wesley.
Thomas, J. J. and K. A. Cook (2006). Illuminating the Path: The Research and Development Agenda for Visual Analytics. Illuminating the Path: The Research and Development Agenda for Visual Analytics(eds)**:** 33-68, IEEE Computer Society.
Tufte, E. R. (1983). The Visual Display of Quantitative Information. Cheshire, CT, USA, Graphics Press.
Tufte, E. R. (1990). Envisioning Information. Cheshire, CT, USA, Graphics Press (8th printing, June 2001).
Tufte, E. R. (1997). Visual Explanations. Cheshire, CT, USA, Graphics Press.
van Wijk, J. J. and H. van de Wetering (1999). Cushion Treemaps: visualisation of hierarchical information. IEEE Symposium on Information Visualization (InfoVis '99), IEEE CS Press.
Ware, C. (2000). Information Visualization: Perception for Design. Orlando, FL, Morgan Kaufmann Publishers.
Wattenberg, M. (1999). Visualizing the stock market. CHI '99 extended abstracts on Human factors in computing systems Pittsburgh, Pennsylvania, ACM Press.