VISUALISATION DE RELATIONS PAR DES GRAPHES INTERACTIFS DE GRANDE TAILLE

KAROUACH Saïd,
karouach@irit.fr

DOUSSET Bernard
dousset@irit.fr
IRIT-SIG, Université Paul Sabatier, 118, route de Narbonne
31062 Toulouse cedex 04
Tél : 05.61.55.67.81

Résumé : Notre plate-forme Tétralogie dédiée à la découverte de connaissances implicites dans de grand corpus hétérogènes d'information électronique est essentiellement basée sur l'analyse relationnelle : entre les acteurs du domaine, les éléments terminologiques et le temps. Afin de mieux communiquer les résultats globaux ou partiels des ces analyses, nous développons actuellement des méthodes de visualisation interactive de graphes basées sur la notion d'attraction et de répulsion. Elle permettent à l'utilisateur de naviguer dans les données, en proposant des vues partielles ou simplifiées qui permettent de comprendre assez vite certaines organisations locales remarquables. Le couplage avec des méthodes de tri et de classification déjà présentes sur la plate-forme permet un pré-placement optimal des sommets ce qui conduit immédiatement à une meilleure lisibilité et à une découverte intuitive des relations les plus significatives.

Abstract : Our Tétralogie platform is dedicated to knowledge discovery in great heterogeneous electronic information corpus. It is primarily based on relational analysis: between the actors' field, terminological elements and time. In order to show total or partial results of these analysis, we develop some interactive graphs visualization techniques, based on attraction and repulsion concept. The user can navigate in data, by proposing partial or simplified views which help him to understand rather quickly some local organizations. The coupling with sort and/or classification methods already present on the platform, allows vertexes optimal preplacement which immediately leads to better legibility and intuitive discovery of the most significant relations.

Mots Clés : Veille scientifique et technique, Analyse relationnelle, Graphes, Visualisations interactives, Java, Internet, Intranet.

Keywords : Science and technology watch, Relational analysis, Graphs, Interactive vizualization, Internet, Intranet

VISUALISATION DE RELATIONS PAR DES GRAPHES INTERACTIFS DE GRANDE TAILLE

INTRODUCTION

Le graphe est largement utilisé comme moyen de représentation et de visualisation de données. Cette méthode de restitution est très appréciée par les utilisateurs, car elle ne nécessite pas de connaissances mathématiques particulières. Le tout est de trouver un graphe à la fois fidèle à la réalité et suffisamment lisible pour être exploité. Nous partons, soit d'une matrice relationnelle calculée au niveau de la plate-forme, soit d'un extrait de matrice récupéré, via un portail, depuis une table de base de données relationnelle (compilation d'une matrice issue de l'analyse initiale).

La matrice analysée appartient à un des types suivants :

- Présence absence (existence d'un lien dans au moins un document)
- Contingence (entre deux variables à item unique : journal, date, 1° auteur, langue)
- Co-occurrence (une des variables peut contenir plusieurs items : auteurs, mots-clés)

Cette matrice peut avoir subi un ou plusieurs traitements préalables :

- Tri par classes de simple connexité
- Tri par blocs diagonaux en mode absolu
- Tri par blocs diagonaux en mode relatif
- Normalisations de tous ordres
- Extraction de sous-graphes
- Seuillages
- Fusion d'éléments
- Elimination d'éléments ne générant pas de connexion

Le but étant toujours de trouver un modèle qui reflète le mieux possible la réalité.

1 - VISUALISATION DYNAMIQUE DE GRAPHES

1.1 - Principe général

A l'origine, Eades **[EADE 84]** comparait les liaisons dans les graphes à des ressorts en analogie avec la loi physique de Hook. Il associait les sommets à des masses et les arêtes à des ressorts reliant celles-ci. Un tel système engendre des forces entre les sommets ce qui entraîne des déplacements respectifs. Après une phase de transition, le système finira par se stabiliser. Eades supposait que le placement final des sommets pourrait correspondre à une configuration satisfaisante du graphe. L'algorithme d'Eades ajoute la notion de force de répulsion entre les sommets et une attraction matérialisée par les arêtes. La condition d'arrêt est simplement un nombre maximum d'itérations. L'analogie à la mécanique serait complète si la condition d'arrêt était fonction de l'énergie du système. L'équilibre est atteint pour un optimum énergétique du graphe. Cette démarche a fait l'objet de plusieurs développements **[KAMA 89], [FRUC 91], [FRIC 94]** conduisant à différents modèles dynamiques de type FDP (Force Directed Placement) dont voici l'algorithme général.

1.2 - Algorithme

Initialisation : Positionnement aléatoire des sommets
Tant que *k < Max_iter* **faire**
    Pour tout sommet *u* **faire**
        Pour tout sommet *v != u* **faire**
        Si *distance(u ; v) < seuil* **alors**
            Calcul de la force de répulsion entre u et v
    Pour toute arête *(u, v)* **faire**
        Calcul de la force d'attraction entre u et v
    Pour tout sommet *u* **faire**
        Cumul des forces
        Déplacement de u en fonction de la température globale
        Diminuer la température globale

Fin

1.3 - Forces d'attraction et de répulsion

La notion de la température globale du système a été introduite pour limiter le déplacement des sommets. Cette température diminue au cours du déroulement de l'algorithme ce qui implique que plus on approchera de l'équilibre , moins un sommet pourra se déplacer. En effet, le déplacement de tous les sommets est inversement proportionnel à la température globale.

Dans notre prototype, nous avons adopté la même démarche que celle décrite ci-dessus. Notre modèle diffère de celui-ci dans la définition des forces d'attraction et de répulsion. La force d'attraction entre deux sommets v_i et v_j est définie par :

$$f_a(v_i,v_j)=\beta_{ij}\, d_{ij}^{\alpha_a}/k$$

β_{ij} est le poids de l'arête (v_i, v_j), k est calculé en fonction de la surface de la fenêtre et du nombre de sommets du graphe, d_{ij} est la distance entre v_i et v_j sur le dessin. Si les sommets v_i et v_j ne sont pas reliées par une arête, alors $f_a = 0$.
La force de répulsion est donnée par :

$$f_r(v_i,v_j)=-k^2/d_{ij}^{\alpha_r}$$

la variable α_a (respectivement α_r) est une constante qui sert à définir le degré d'attraction (respectivement de répulsion) entre deux sommets.

Ce type d'algorithme de dessin de graphe donne de bons résultats pour des graphes relativement petits (100 sommets). Son utilisation devient très lourde pour des graphes de grande taille. Une solution consiste à transformer le graphe initial en une structure équivalente de taille moyenne. L'idée est alors de décomposer le graphe en sous-graphes (groupes), et d'appliquer ensuite l'algorithme de dessin sur le graphe des groupes. Ceci nécessite une technique de partitionnement de graphe efficace tenant compte de la taille du graphe initial.

1.4 - Partitionnement de graphe

La limitation du nombre de sommets à afficher améliore la clarté et augmente simultanément la rapidité d'exécution des tâches chargées du placement de ces sommets. Stijn van Dongen **[DONG 00a]** a introduit une technique de partitionnement de graphes de grande taille. Son algorithme MCL (Markov Cluster algorithm) est basé sur la simulation de l'écoulement stochastique dans un graphe. L'idée est de simuler plusieurs écoulements aléatoires dans le graphe, puis de renforcer l'écoulement là où il est déjà fort, et de l'affaiblir là où il est faible. Mathématiquement, l'écoulement est simulé par des opérations algébriques (**expansion** et **inflation**).sur la matrice stochastique (de Markov) associée au graphe. L'expansion de matrice correspond au calcul des probabilités des chemins aléatoires de plus grandes longueurs. Plus spécifiquement, l'expansion augmente l'écoulement par le calcul des puissances de la matrice stochastique, elle permet à l'écoulement de relier différentes parties du graphes, mais elle ne montrera pas la structure fondamentale des clusters. L'inflation favorise et rétrograde les probabilités des chemins dans le graphe. Autrement dit , cet opérateur sert à renforcer ou affaiblir l'écoulement là où c'est nécessaire.

Les étapes d'expansion et d'inflation s'effectuent alternativement sur la matrice de Markov jusqu'à ce qu'aucun changement ne puisse être détecté. La matrice de Markov est alors interprétée en tant que matrice résultat d'un regroupement. Pour plus de détails, nous conseillons de consulter les autres travaux de Stijn van Dongen **[DONG 00b, DONG 00c]**.

2 - RECHERCHE DE CLUSTERS ET DE CONNECTEURS

2.1 - Visualisation du graphe sans simplification

Avant d'avoir recours à l'arsenal des simplifications proposées en mode interactif, le mieux est de visualiser le graphe dans son intégralité. Très souvent, il est suffisamment lisible pour être interprété , mais un écran haute résolution est bien entendu recommandé. Certaines manipulations élémentaires permettent ensuite de découvrir les caractéristiques principales de la structure relationnelle :

- Rechercher les éléments isolés
- Tirer sur un sommet pour évaluer ses liaisons
- Déplacer vers les bords de l'écran les structures homogènes
- Laisser agir le système d'attraction répulsion pour placer les sommets

Dans l'exemple ci-dessous, nous proposons une vue globale qui manque de lisibilité, mais qui avec un peu d'habitude peut déjà être exploitée en l'état. Les leaders du domaine sont plus colorés (matrices quantitatives), la densité des faisceaux dévoile les réseaux importants, la force des liens entre réseaux peut être évaluée par la lecture des valeurs des arêtes.

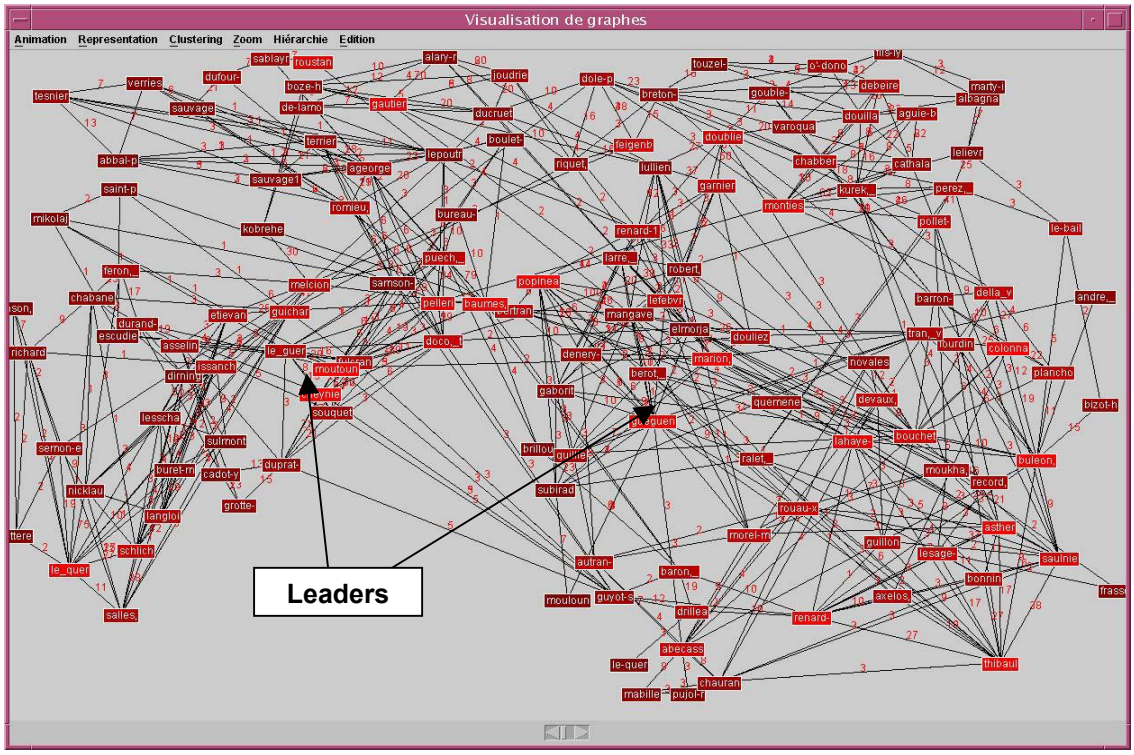


Figure 1. : Graphe de liens sans simplification : manque de lisibilité.

2.2 - Aide à l'analyse par simplifications interactives

Ci-dessous, une simplification par seuillage du graphe précédent qui permet de détecter les groupes isolés, ceux qui sont connectés et les individus qui servent d'interfaces. La complexité du graphe étant réglable, le processus de découverte peut se baser sur cette fonction pour faire apparaître progressivement les détails par abaissement du seuil une fois que les signaux fort ont été repérés et correctement répartis sur l'écran.

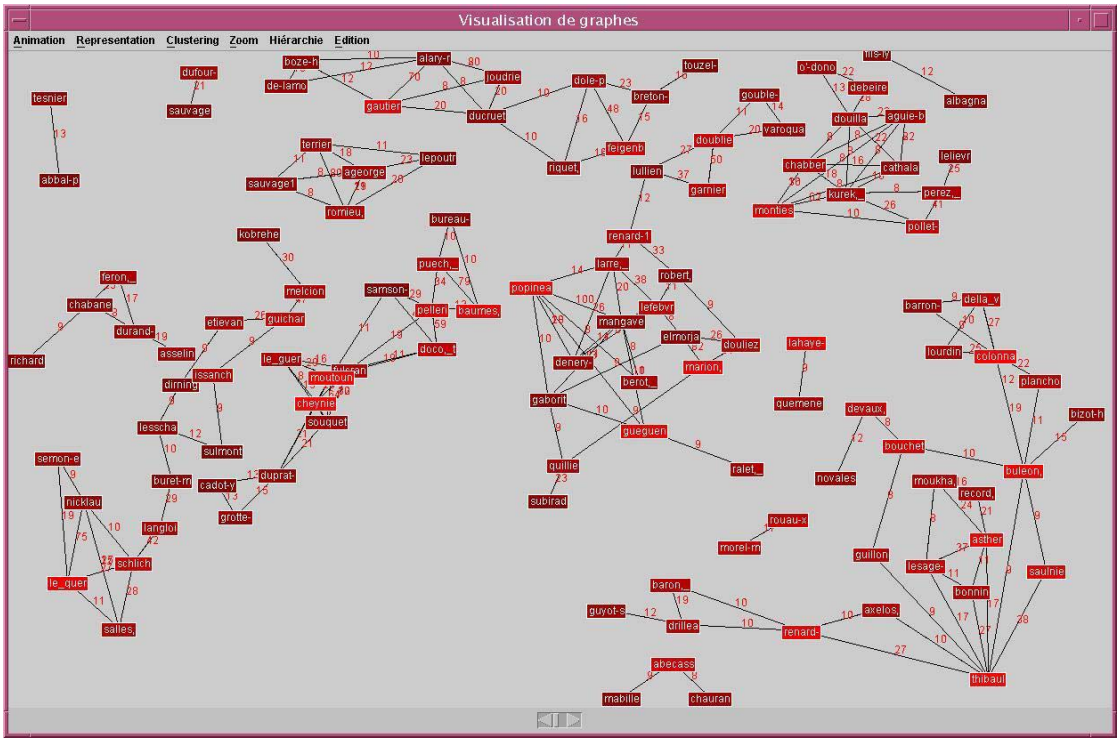


Figure 2. : Graphe obtenu après simplification : les éléments clés sont identifiés.

3 - CLASSES CONNEXES ET SOUS GRAPHES

3.1 - Tri par classes de connexité

Une matrice qu'elle soit carrée ou rectangulaire peut être triée en fonction de ses classes de simple connexité. Un zoom arrière en 2D fait alors clairement apparaître sa structure comme ci-dessous. Chaque classe peut alors être étudiée séparément afin d'en expliquer la structure et d'en découvrir les éléments clés : ici les connecteurs entre équipes de recherche.

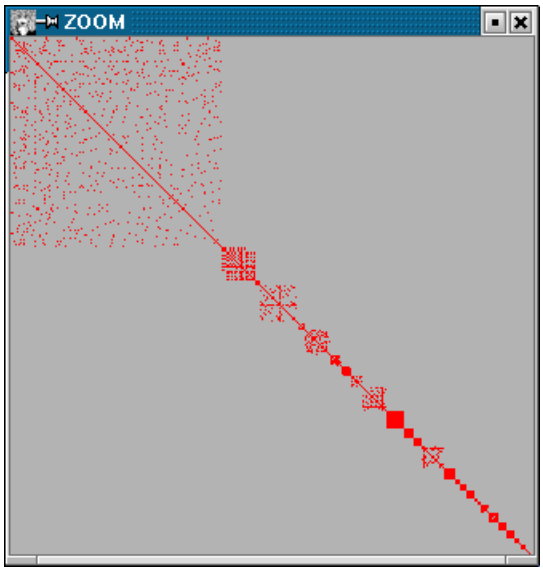


Figure 3. : zoom arrière 2D d'une matrice carrée Auteurs² triée par connexité.

Chaque classe est en fait un sous graphe totalement indépendant qui peut être visualisé et analysé à part. Ci-dessous les structures de deux équipes de la partie basse du zoom.

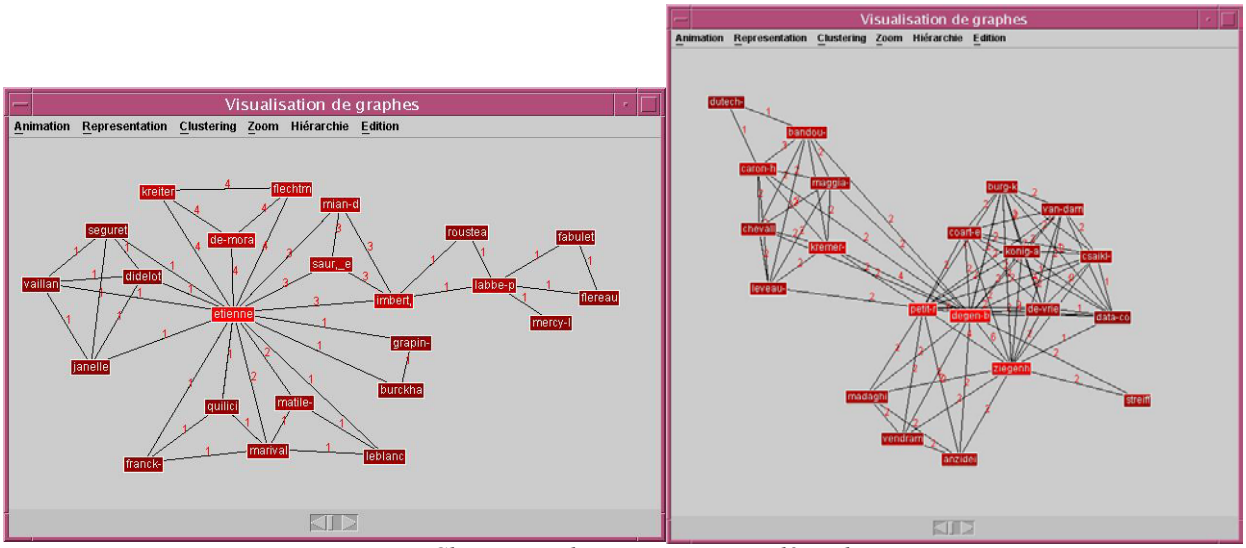


Figure 4. : Classes simplement connexes indépendantes.

3.2 - Extraction de sous graphes

La partie haute du zoom précédent représente un collaboratoire beaucoup plus important, qui est formé de plusieurs équipes cohérentes. Les collaborations entre ces équipes s'effectuent au travers de connecteurs qui sont parfaitement identifiables dans les graphe qui suit. Mais l'étude du fonctionnement interne de chaque équipe n'est pas facilitée par la densité de ce graphe. Le mieux est de couper les liens qui unissent une équipe à ses voisins sans perdre les connecteurs, donc extraire un sous graphe et l'étudier à part.

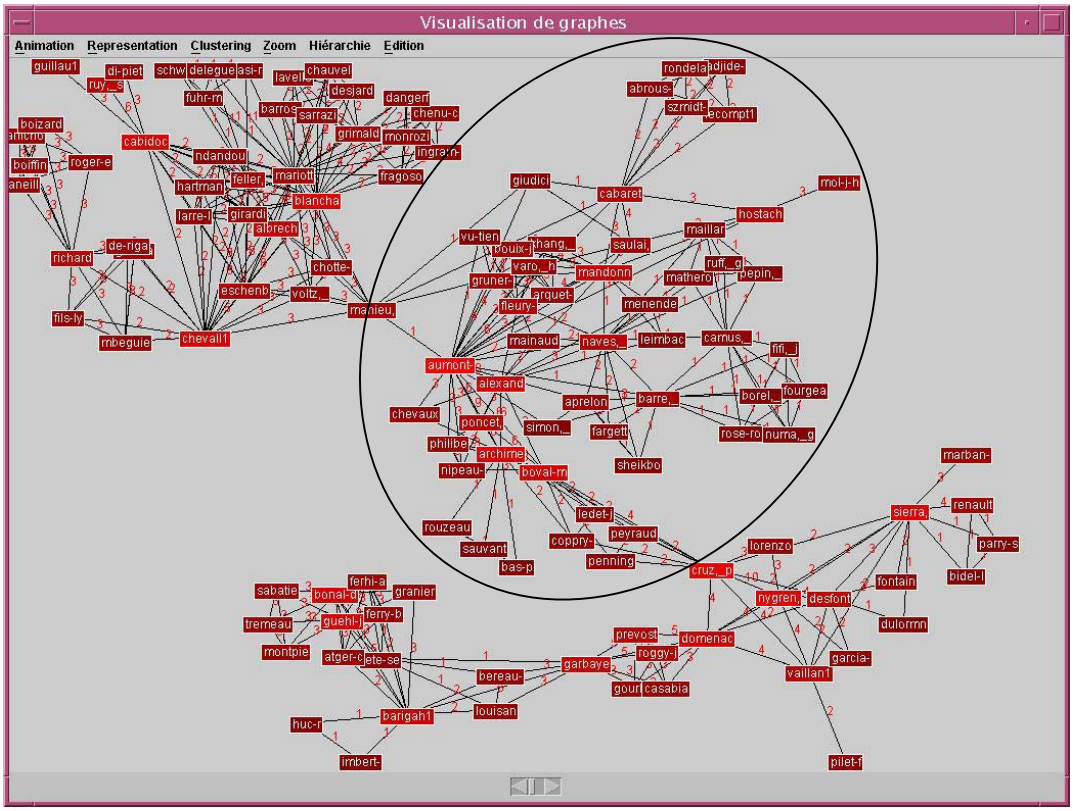


Figure 5. : Sous-graphes ou classes connexes indépendantes.

Pour un graphe biparti issu d'une matrice non symétrique, le problème est similaire : on peut étudier à part chaque composante simplement connexe, ou extraire des sous graphes en supprimant les liens qui les unissent au reste du graphe.

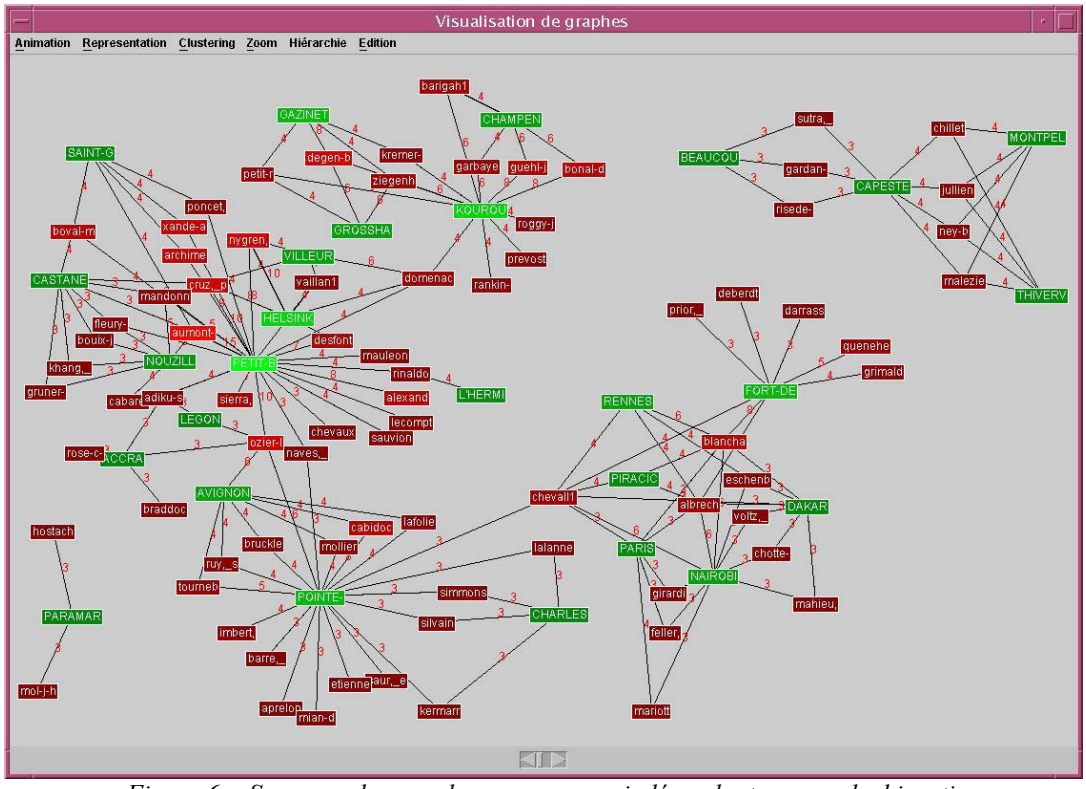
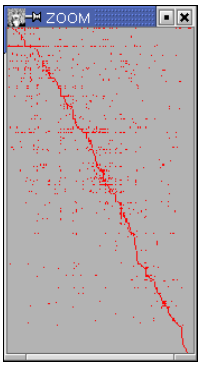


Figure 6. : Sous-graphes ou classes connexes indépendantes : graphe biparti.

Ci-dessus, un extrait d'une matrice Auteurs – Villes où des liens entre sous graphes sont alternativement dus à certains auteurs ou à certaines villes. Il est assez facile d'isoler quatre groupes en ne coupant qu'un ou deux liens. Les éléments isolés peuvent alors être listés afin de servir de filtre pour une analyse stratégique du groupe ainsi formé (par croisement avec toutes les autres informations disponibles : acteurs et thèmes)

4 - PLACEMENT DIRIGE DES SOMMETS

4.1 - Tri par blocs diagonaux



Afin de rendre plus lisible la visualisation de graphes complexes et connexes, nous avons tenu compte, dans le placement des sommets, de la sérialisation des items obtenue à l'issue d'un tri par blocs diagonaux. Les sommets sont alors distribués de façon circulaire, leurs principales connexions s'effectuent avec leurs voisins et seules les liaisons inter-clusters attirent maintenant l'attention. Ce principe est applicable aussi bien pour les matrices symétriques qu'asymétriques. Ci-contre, le zoom arrière 2D d'une matrice asymétrique ainsi réorganisée par permutation des lignes et des colonnes.

4.2 - Classification simultanée des lignes et des colonnes

Une autre méthode consiste à trier simultanément les lignes et les colonnes de la matrice. C'est par exemple possible après avoir réalisé une analyse factorielle des correspondances qui normalise les lignes et les colonnes et les plonge dans le même espace. Il suffit alors de placer les sommets du graphe dans l'ordre obtenu sur l'arbre planaire de classification hiérarchique en mettent les lignes à l'extérieur (plus nombreuses) et les colonnes à l'intérieur.

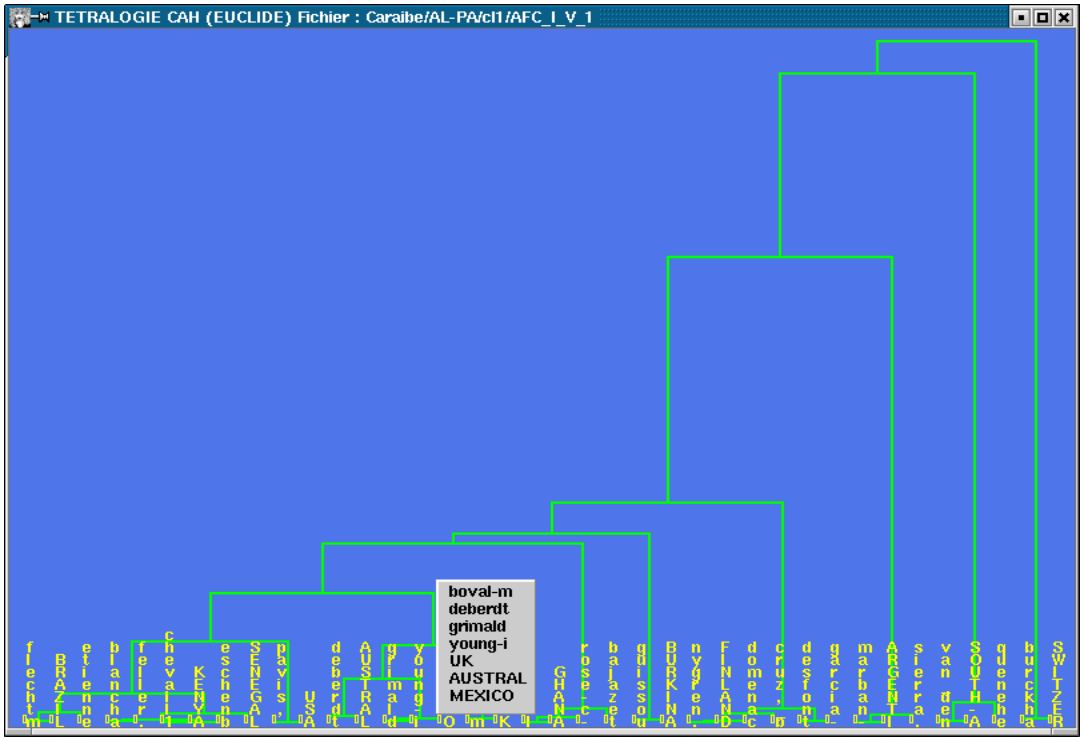


Figure 7. : Classification ascendante hiérarchique après AFC sur Auteurs Pays.

4.3 - Visualisation d'une matrice triée

Dans l'exemple suivant, où nous avons croisé des Auteurs et des Pays, les Auteurs (lignes de la matrice) sont disposés à la périphérie, les pays moins nombreux à l'intérieur. Nous pouvons immédiatement remarquer :

- Les pays totalement isolés,
- Les auteurs isolés dans un pays isolé,
- Les groupes d'auteurs qui collaborent à l'international,
- Les acteurs de ces collaborations à l'intérieur des groupes,
- Les pays liés au travers des auteurs,
- La force de ces liaisons,
- Les éléments clés (connecteurs uniques, plaques tournantes),
- Le leader de chaque groupe (par la coloration)
- La taille de chaque cluster qu'il soit ou non isolé.

En fait, les éléments remarquables (liens importants) sont ceux qui traversent la figure, ils représentent, ici, des collaborations internationales.

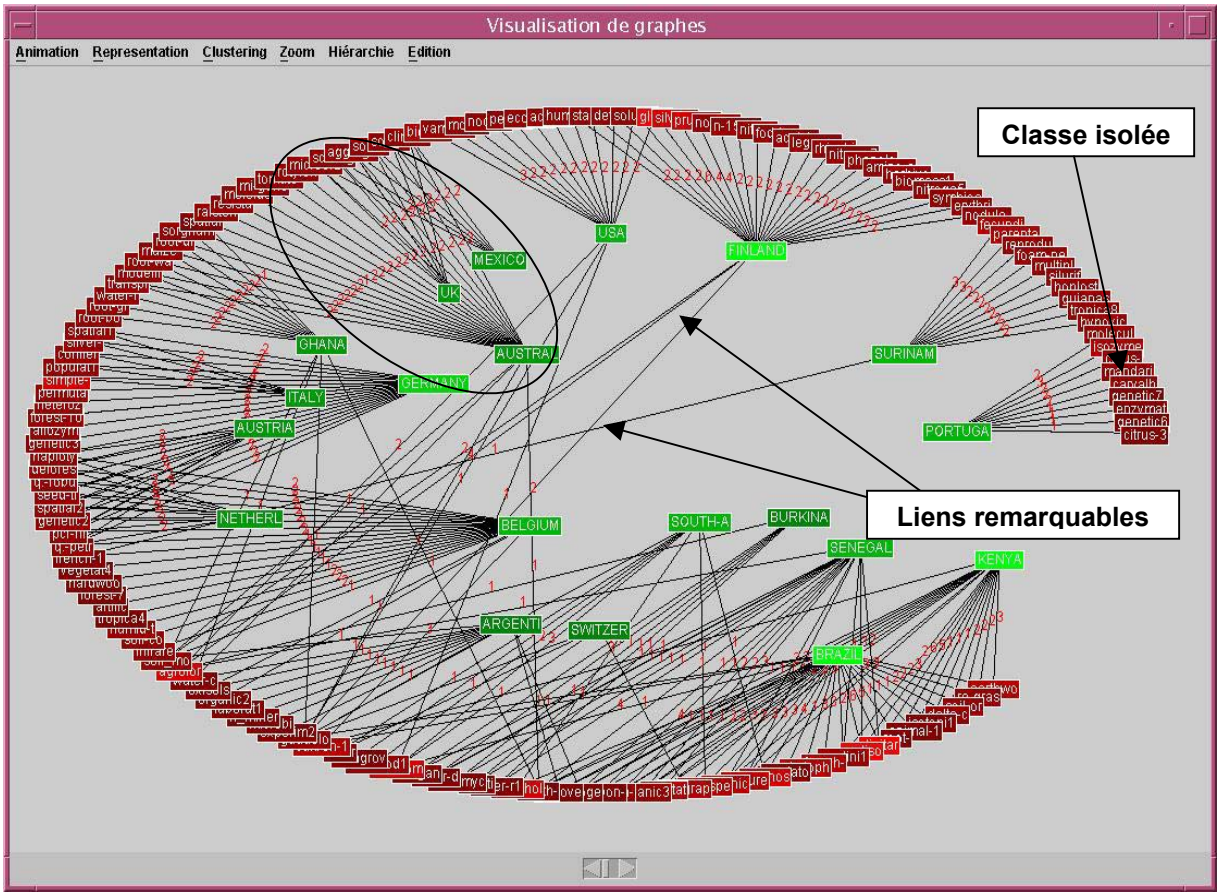


Figure 8. : Placement des items en fonction d'un tri par blocs diagonaux.

4.4 - Autre possibilité de partitionnement

D'autres méthodes de placement des sommets peuvent être issues des classifications proposées par Tétralogie, en particulier la classification par la méthodes des centre mobiles. Cette méthode propose en effet une partition en *n* classes effectuée à partir d'un représentant de chaque classe. Il est possible de choisir ces représentants directement sur le graphe comme on le fait déjà sur une carte factorielle en 3D. Une fois la partition réalisée les sommets sont placé de façon judicieuse en fonction de la taille de la fenêtre. Une autre utilité de ce partitionnement est l'extraction automatique de sous graphes plus faciles à analyser, cette possibilité prend toute son importance dès que le nombre de sommets dépasse plusieurs centaines.

5 - CAS PARTICULIER DES ARBRES PARTIELS

Une méthode radicale de simplification d'un graphe connexe est de ne conserver qu'un de ses arbres partiels extremum (ici maximum). Pour cela, il suffit, pour chaque sommet, de garder sa connexion la plus forte (première arête de chaque cocycle). Le graphe obtenu permet de savoir quels sont les éléments clés du domaine (connecteurs privilégiés). De plus, le placement des sommets obtenu à partir d'un de ses arbres planaires peut servir de base pour la reconstruction du graphe complet ou de l'une de ses simplifications moins poussées. En voici une illustration qui permet de juger de la lisibilité de cette méthode dans le cas d'une matrice asymétrique (graphe biparti).

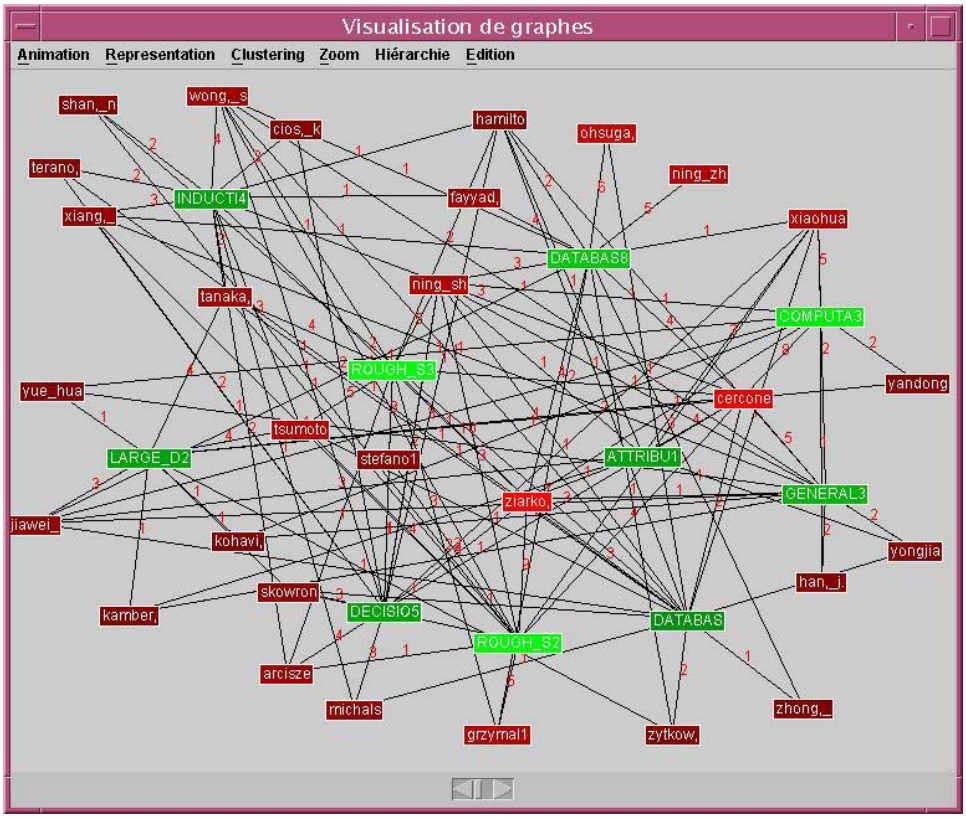


Figure 9. : Graphe simplement connexe de départ.

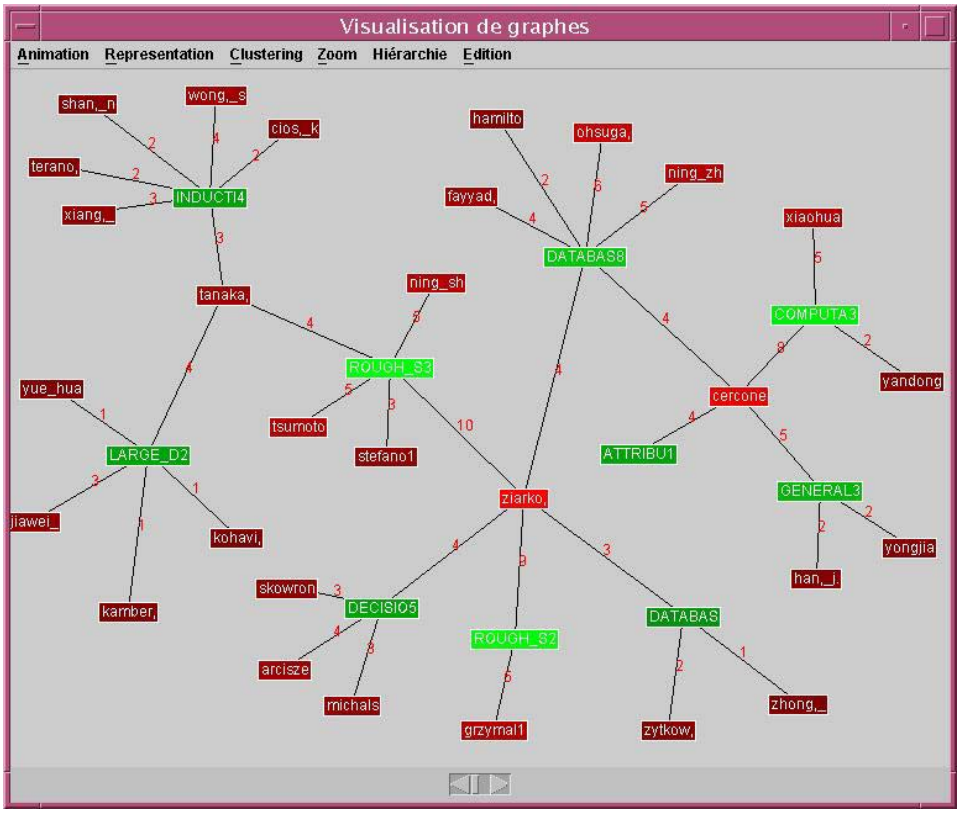


Figure 10. : Arbre partiel maximum extrait du graphe précédent.

CONCLUSION

La visualisation interactive et dynamique de graphes de grande taille apporte à notre plate-forme de veille un outil de restitution accessible à tous. En effet, la lecture d'un graphe ne nécessite pas de connaissances particulières comme celles exigées pour interpréter une carte factorielle, un arbre de classification ou un zoom interactif sur une matrice réorganisée. De plus, cet outil est disponible, grâce à Java, sur la plate-forme d'analyse elle-même, sur la restitution de l'analyse sous forme de site Web implantable sur CD/Rom ainsi que sur un portail ouvert sur l'ensemble des dictionnaires et matrices utilisés dans la macro-analyse et compilés sous forme de base de données relationnelle implantée **[SOSS 01]** en MySql et interfacée en Php. A tous les niveaux de la démarche de veille, ces graphes sont donc accessibles pour aider à la compréhension des phénomènes relationnels inhérents à toute activité humaine. L'étude des réseaux sert à découvrir les stratégies les plus fines qui ne sont pas révélées explicitement dans les écrits analysés mais qui s'y trouvent bel et bien et qu'il suffit de mettre en lumière par de telles méthodes. Les démarches basées sur l'analyse de données à la française nous permettaient déjà d'accéder à ce type d'informations stratégiques, mais leur mode de représentation graphique était assez mal adapté pour une restitution grand public. Le vide est maintenant comblé et ces nouveaux outils viennent compléter les modes de communication graphique et intuitive que nous privilégions actuellement comme les cartes géographiques interactives pour la géostratégie **[HUBE 00], [KARO 01], [DOUS 02].**

BIBLIOGRAPHIE

[EADE 84] P. Eades

A heuristic for graph drawing. Congressus Numeranium, Vol 42, pp. 149-160, 1984.

[KAMA 84] T. Kamada, S. Kawai

An algorithm for drawing general undirected graphs. Information Processing Letters, vol. 31, pp. 7-15, 1989.

[FRUC 91] T. Fruchterman, E. Reingold

Graph Drawing by Force-Directed Placement. Software Pratice an Experience, 1991.

[FRIC 94] A. Frick, A. Ludwig, H. Lehldau

A fast adaptative layout algorithm for undirected graphs. In Proceeding of Graph Drawing'94, vol. 894, pp. 388-403, 1994.

[DONG 00a] S. van Dongen

A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam. May 2000.

[DONG 00b] S. van Dongen

Graph Clustering by Flow Simulation. Thèse , Université d'Ultrecht, Allemagne, May 2000.

[DONG 00c] S. van Dongen

Performance criteria for graph clustering and Markov experiments. Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam. May 2000.

[HUBE 00] G. Hubert, J. Mothe, A. Benammar, T. Dkaki, B. Dousset, S. Karouach

Textual document Mining using graphical interface. International Human Computer Interaction, HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey, pp 918-922 (volume 1), 05-10 août 2001.

[KARO 01] S. Karouach, B. Dousset

Visualisation interactive pour la découverte de connaissances : GeoECD. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 291-300, (Barcelone, Espagne), octobre 2001.

[SOSS 01] D. Sosson, M. Vassard, B. Dousset

Portail pour la navigation en ligne dans les analyses stratégiques. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

[DOUS 02] B. Dousset, S. Karouach

Collaboration interactive entre classifications et cartes thématiques ou géographiques. 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.