Analyse d'information relationnelle par des graphes interactifs de grandes tailles

Saïd KAROUACH, Bernard DOUSSET

IRIT-Université Paul Sabatier (Toulouse III), Equipe SIG
118, route de Narbonne.
31062 Toulouse cedex 4
tél.: (33) 05 61 55 67 81. fax:. (33) 05 61 55 62 58
{karouach,dousset}@irit.fr
http://atlas.irit.fr

Resume : La découverte de connaissances à partir d'importantes masses de données hétérogènes débouche le plus souvent sur l'analyse relationnelle. La recherche d'informations stratégiques s'appuie en effet sur les liens fonctionnels et sémantiques entre documents, acteurs, terminologie et concepts d'un domaine sans oublier le paramètre temps. De nombreuses méthodes sont proposées pour identifier, analyser et visualiser les mécanismes mis à jour: analyse relationnelle, classifications supervisées et non supervisées, analyse factorielle, analyse sémantique, cartes, dendogrammes, … Mais ces approches demandent souvent une expertise non négligeable pour être comprises et ne s'adressent donc pas aux non initiés. Par contre, la vue d'un graphe mettant en relation une ou deux classes d'éléments interdépendants est directement assimilable par tout le monde. Nous proposons donc un ensemble de visualisations interactives de graphes dont la manipulation doit permettre une découverte de connaissances intuitive et basée sur un langage graphique naturel. Nous illustrons notre propos de nombreux exemples tirés de cas réels d'analyses stratégiques qui ont permis d'évaluer cette approche sur un panel très large de données.

Summary : Knowledge discovery in large heterogeneous data sets often leads up to relational analysis. Strategic information research relies on functional and semantic links between documents, actors, terminology, concepts and time. Several methods are suggested to identify, analyse and visualise discovered mechanisms: relational analysis, supervised or non-supervised classifications, factorial analysis, semantic analysis, maps, dendrograms, ... But these approaches often request an expert's report to be understood and are not intended for uninitiated. On the other hand, a graph's view which shows relation between one or two classes of dependant elements is easily understandable. We suggest then interactive visualization sets of graphs whose manipulation enables intuitive knowledge discovery based on natural graphical language. We illustrate our work with several examples extracted from real cases of strategic analyses which enabled to evaluate this approach with many data.

1 Introduction

Notre cadre d'étude se situe dans le domaine de l'extraction de connaissances à partir de collections de données textuelles semi-structurées. Dans un premier temps, l'analyse de ce type de corpus consiste en un découpage de l'information utile en unités (terminologie, individus, organismes, temps,...) et à l'extraction des plus significatives d'entre elles en fonction des objectifs visés. La seconde étape consiste à appliquer des méthodes statistiques sur ces unités soit pour en déduire leur organisation dans le corpus pris dans sa totalité, soit pour identifier des regroupements locaux (réseaux sémantiques, signaux faibles, collaborations, concurrences, …). Dans ce cadre, les méthodes de traitement automatique de l'information textuelle prennent souvent pour point de départ une représentation de l'information élaborée sous une forme matricielle. Ces matrices se décomposent généralement en deux classes. D'un coté, les matrices représentant les relations entre entités issues du même type de données comme les auteurs, les mots clés ou des concepts, des sites Web, ... De l'autre côté (cas plus complexe), les matrices représentent les relations entre deux entités différentes. Il peut s'agir alors de connexions documents-termes, auteurs-termes, auteurs-affiliations, ... L'analyse relationnelle de ce type de données permet ensuite de détecter des éléments remarquables du domaine étudié comme des classes, des connecteurs, des sous réseaux qui ont une fonction structurante. Notre objectif est de représenter ce type d'information sous forme ergonomique pour une analyse visuelle et exploratoire, dans une optique de découverte de relations cachées (ou implicites) qui permettent de mieux appréhender un domaine. Pour atteindre ce but, notamment dans le cas délicat des grands volumes d'information, nous nous appuyons sur plusieurs techniques issues de différents domaines : théorie des graphes, partitionnement, analyse relationnelle, optimisation du tracé des graphes et visualisation d'information.

Afin d'analyser notre démarche, la suite de cet article développe différents aspects liés à la visualisation des graphes. Dans la section 2, nous présentons nos motivations dans le choix du concept de graphe comme modèle de représentation. Dans la section 3, nous discutons des moyens visuels et graphiques qu'on peut utiliser pour rendre une représentation de graphes plus pertinente et plus riche. Nous abordons, dans la quatrième partie, l'utilisation des techniques de partitionnement de graphes dans un objectif de détections de groupes homogènes et de structuration des données pour réduire la complexité visuelle des informations affichées. Nous décrivons notre démarche basée sur l'utilisation de la technique **MCL**. Ensuite, nous montrons, par des exemples, le long de la section 6 comment le mixage de ces différentes approches peut contribuer à visualiser, manipuler, explorer et naviguer dans l'espace informationnel, dans un environnement interactif et dynamique.

2 Représentation de données : concept de graphe

Le concept graphe est généralement utilisé comme modèle de représentation dès que les données sont intrinsèquement liées. Ce type de données peut être vu comme un graphe, associé à une matrice, dont les arêtes représentent les relations entre les données. La représentation de données relationnelles par des graphes est largement utilisé dans différents domaines qui traitent de l'information textuelle. L'analyse de réseaux a pour objectif de concevoir des représentations synthétiques qui puissent exprimer l'interaction entre les différentes entités représentées. En effet, la lecture des réseaux doit faire surgir l'information

(endogène) en permettant d'identifier visuellement la morphologie structurelle de l'information analysée. Toutefois, la représentation du réseau doit être lisible pour être bien interprétée par son utilisateur. La lisibilité d'un réseau n'est pas en soi un concept facile à déchiffrer. La communauté du dessin de graphes a établi des critères esthétiques [Di Battista *et al.*, 1999) qui permettent de déterminer la qualité d'un dessin donné. Toutefois, le dessin d'un graphe sera lisible s'il reflète les structures macroscopiques qui existent en son sein.

Le but recherché par la visualisation des graphes est d'en faciliter la lecture. Une des techniques la plus utilisée pour dessiner un graphe est celle basée sur la notion d'attraction et de répulsion [Eades, 1984]. Ce type d'algorithme de dessin de graphe donne de bons résultats pour des graphes relativement petits (quelques centaines de 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.

3 Détection d'éléments importants

Dans un graphe de relations sur un ensemble de sommets, tous les éléments n'ont pas la même importance ou le même rôle dans la structure locale ou globale du graphe. En visualisation d'information, la couleur (ou intensité de couleur) est une variable visuelle très utilisée pour mettre en valeur les caractéristiques des entités affichées. La couleur peut refléter soit l'importance d'une sommet soit son appartenance à un groupe de sommets ayant certaines caractéristiques communes. Quel que soit son objectif, l'utilisation de la couleur est basée une valeur numérique, appelée valeur métrique ou nœud métrique, attribuée à chaque sommet du graphe. Il existe deux types de métriques : une métrique basée sur la structure (elle utilise seulement des informations relatives à la structure du graphe : le degré d'un sommet), une métrique basée sur le contenu (elle utilise des données associées au sommet). L'avantage d'une métrique structurale est qu'aucune connaissance de domaine n'est exigée. Nous utilisons la notion de métrique pour détecter visuellement les éléments importants du graphe, mais elle peut aussi être utilisée pour réaliser des opérations de filtrage et de regroupement.

3.1 Définition de métrique

Soit A= (a_{ij}) la matrice d'adjacence d'un graphe valué G quelconque de n sommets. La fonction qui associe à un sommet v_i, la quantité $m_i = \sum_j a_{ij}$, étant la somme des poids de ses arêtes incidentes, définit une métrique [Melançon *et al.*, 1999]. Cette métrique est structurale puisqu'elle tient compte de la structure du graphe. Elle donne une indication sur *l'importance* d'un sommet dans le graphe en fonction des **intensités des liens** qu'il entretient au sein de la structure globale. L'analyse d'un graphe basée sur ce type de métrique permet de repérer les groupes de sommets qui sont souvent liés ensembles, les liens indirects qu'ils peuvent nouer, le rôle clé de certains sommets situés à l'interface entre plusieurs groupes, ... Par exemple, dans un graphe qui représente des co-publications, cette métrique définit un indicateur sur la contribution personnelle d'un auteur au sein d'une équipe. Naturellement, le même concept peut être appliqué aux arêtes. Etant donné que chaque paire de sommets

(adjacents) est liée par une arête, le poids (pondération) représente l'intensité du lien. Il est alors possible de définir ce poids comme la métrique associée à l'arête.

3.2 Codage de la métrique

Quel que soit l'objectif d'une métrique, son codage par une couleur consiste à définir une fonction qui attribue une intensité de couleur à chaque valeur métrique. La définition d'une telle fonction dépend de la distribution de la métrique. Considérons la normalisation des valeurs m_i par leur maximum $\max(m_j)$ définie par :

$$M_i = \frac{m_i}{\max_j(m_j)}$$

A partir des valeurs M_i, nous définissons un spectre de nuances adapté à la distribution de ces valeurs. Pour coder l'intensité de la couleur à partir des valeurs métriques relatives à chaque sommet, nous utilisons le modèle défini par la famille de fonctions non linéaires de type :

$$f_n(x) = \frac{(n+1).x}{n.x+1}$$

Le codage de la métrique en utilisant ce type de fonctions permet d'attribuer une intensité d'une couleur à chacun des sommets en fonction de sa valeur métrique. Le paramètre n joue le rôle d'amplificateur de l'intensité dans le cas de petite valeur métrique. Les mêmes fonctions peuvent être utilisées pour la coloration des arêtes afin d'identifier visuellement les liens forts et faibles dans le graphe. Elle peut être utilisée à la fois pour définir l'épaisseur et l'intensité de couleur des arêtes. Il faut noter que d'autres variables visuelles, comme la taille, peuvent être utilisées pour le codage d'une métrique. Dans ce cas, un sommet du graphe est représenté par une icône (circulaire ou rectangulaire) ayant une taille proportionnelle à sa valeur métrique (cf. Fig 1).

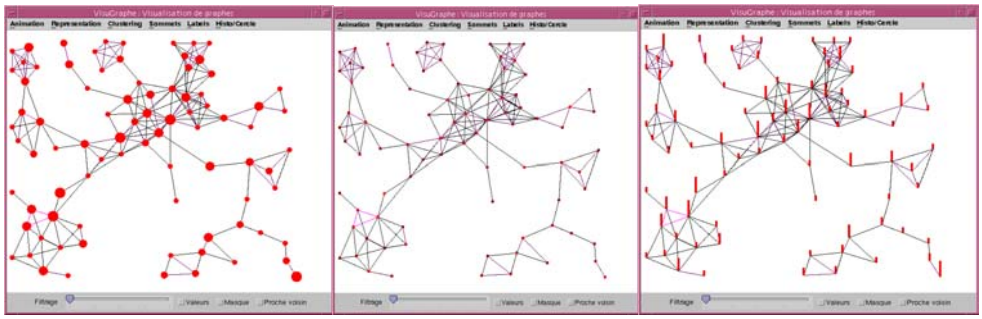

Fig 1 - *Différents codages de la métrique*

4 Détection de groupes : partitionnement de graphes

En visualisation de graphe, la taille constitue l'obstacle majeur pour les algorithmes de dessin. Mise à part l'utilisation d'attributs visuels comme la couleur, le partitionnement est un moyen efficace pour contourner un tel obstacle, en générant un autre graphe de niveau supérieur donnant une idée globale sur la structure des données sous-jacentes, plus facile à visualiser et dans lequel il est possible de naviguer. Dans un contexte de visualisation

interactive de graphes, la méthode de partitionnement utilisée doit être efficace (en temps de calcul) pour garantir la manipulation de la structure visualisée dans un environnement interactif. Il existe plusieurs techniques de partitionnement de graphes de grande taille, et les plus utilisés se basent sur des approches spectrales [Alpert et Kahng, 1995] [Kuntz et Henaux, 2000] [Jouve *et al.*, 2001] ou de partitionnement multiniveaux notamment les algorithmes de la famille METIS [Karypis et Kumar, 1998].

Récemment, Stijn van Dongen 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 (**puissances matricielles** et **normalisations**) appliquées sur la matrice stochastique (de Markov) associée au graphe. Pour plus de détails, nous conseillons de consulter les travaux [Van Dongen, 2000]. L'évaluations de la méthode MCL a montré la rapidité et la qualité de ses résultats dans divers domaines [Enright *et al*., 2002].

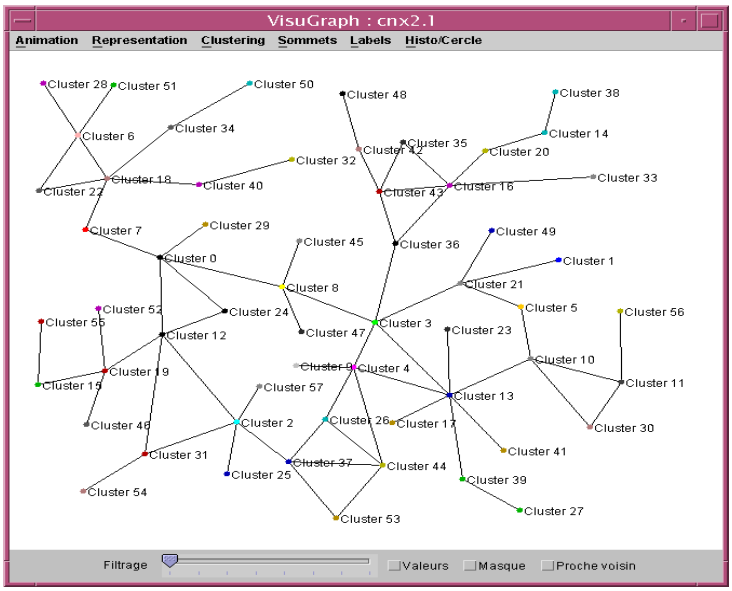

Fig 2 *: Graphe de classes obtenu par MCL*

Après avoir découvert une partition du graphe étudié , nous pouvons réduire le nombre d'éléments à afficher en limitant notre vue aux groupes eux-mêmes (cf. Fig. 2). Cette représentation du graphe de classes (ou graphe quotient) diminue considérablement la complexité du dessin à base de forces. Si nous revenons au graphe initial (cf. Fig. 3), l'appartenance de chaque sommet à une classe est signalée par une couleur identique. Il est alors facile de remarquer que les liens inter-classes sont plus faibles que les liens intra-classes et que la répartition des classes reste identique sur les deux vues.

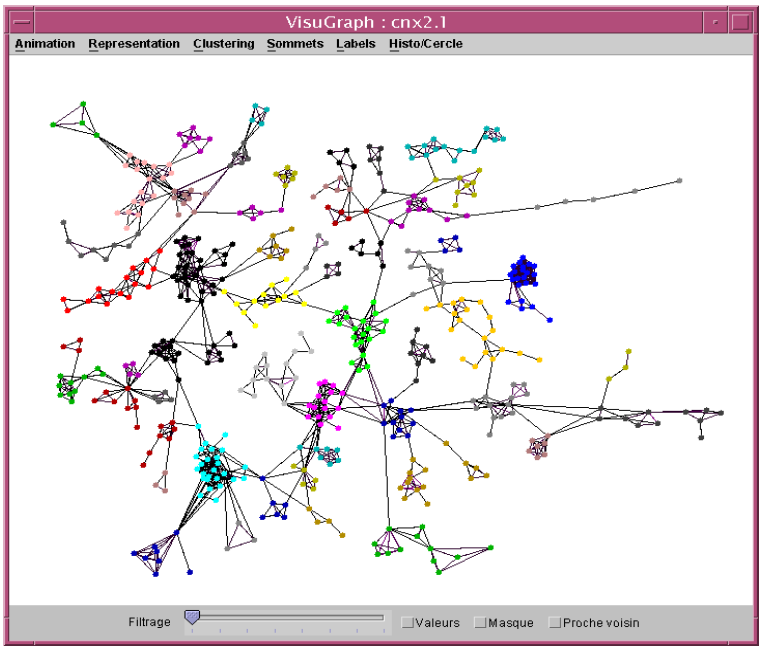

Fig 3 *:Graphe partitionné*

Nous venons de voir qu'il est possible de réduire la complexité visuelle liée à la taille du graphe initial, en procédant à son partitionnement. Cependant, il est important de noter que la visualisation proprement dite dépend fortement de la qualité de la partition obtenue, donc de la méthode de partitionnement utilisée et des moyens interactifs mis en œuvre pour manipuler et naviguer dans une telle partition. Autrement dit, le problème de visualisation d'un graphe de grande taille doit prendre en compte trois critères essentiels : le choix de la *méthode de partitionnement*, le(s) *mode*(s) de *représentation* du graphe partitionné et les *moyens de navigation et d'exploration*.

5 Modes d'exploration et de navigation

Le processus de découverte de connaissances basé sur une technique de visualisation est un processus interactif qui commence habituellement par des essais de visualisation des données dans leur globalité, suivi de la définition d'une stratègie d'exploration. Le mantra de Schneiderman [Shneiderman, 1996] résume bien cette idée : « Overview first, zoom and filter, then details on demand ».

Lors de l'analyse d'un ensemble de données issus d'un corpus quelconque, l'analyse peut avoir comme objectif l'exploration de ces données à partir d'un centre d'intérêt particulier. Qu'il soit partitionné ou non, l'exploration d'un graphe peut être réalisée à partir d'un sommet particulier qu'on appellera *focus*. Afin de réaliser une navigation locale dans le graphe initial souvent trop complexe, il est possible de travailler sur un sous-graphe. Pour cela, nous partons d'un focus particulier choisi dans une liste alphabétique et nous étendons

progressivement le graphe, depuis ce sommet, par transitivité. Cette technique permet de nous concentrer sur un extrait pertinent issu d'une information ciblée (acteur, mot-clé, concept).

5.1 Identification du focus

Compte tenu de la taille du graphe initial, la recherche d'un tel focus peut s'avérer longue. Pour cela, nous utilisons un moyen interactif d'identification basé sur la notion d'*oeil de poisson* sur la liste des étiquettes des sommets et non pas sur la représentation géométrique du graphe. Ainsi l'utilisateur peut accéder rapidement au focus via son étiquette (cf. Fig. 4).

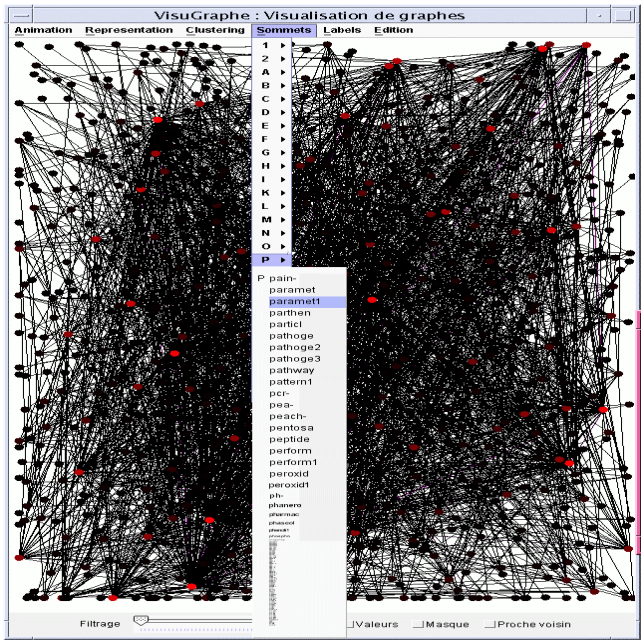

Fig 4 *: Choix d'un focus*

5.2 Exploration pas à pas : vues locales

A partir d'un focus et en fonction du besoin de l'analyse, plusieurs opérations peuvent être effectuées sur le graphe initial. Généralement, ces opérations ont un objectif commun : l'extraction d'un sous graphe particulier ayant une sémantique spécifique. Cette sémantique est fortement liée à la nature des données analysées. Par exemple, si le graphe représente un réseaux de liens entre concepts issus d'un corpus, le sous graphe en question peut être le réseau de liens (ou réseau sémantique) relatif au focus ou le réseau de concepts ayant une même particularité que le focus choisi, ... Dans le cas où le graphe représente les co-publications entre chercheurs, il peut s'agir d'extraire le réseau de personnes (ou collaboratoire) qui ont co-signées un certain nombre de publications. Quel que soit l'objectif de ces opérations, il s'agit d'extraire une ``**structure réduite**'' (SR) du graphe initial, en fonction d'un besoin ou ``d'un point de vue'' particulier de l'utilisateur.

D'un point de vue visualisation, il est inconcevable de produire un affichage de ce type de structure ayant un niveau de profondeur très élevé. Par contre, il est tout à fait possible d'afficher une SR et particulièrement celle qui correspond à quelques niveaux de transitivité à partir du focus (cf. Fig. 5). Toutefois, dans le cas de réseaux très denses, la SR peut être complexe même pour deux niveaux. Il est possible de contrôler la taille de la SR, à l'aide d'un curseur qui représente les niveaux possibles, en ajustant le niveau de profondeur. Ainsi, l'utilisateur peut se rendre compte de la complexité des résultats tout le long de son exploration. Une démarche complémentaire à celle-ci consiste à permettre à l'utilisateur, dans un premier temps, de ne visualiser que les sommets adjacents (de niveau 1) au focus, tout en lui autorisant de continuer son exploration à partir d'un des sommets affichés (pseudo-focus). Dans ce cas, la SR de niveau 1 est enrichie, d'une part, par les voisins du pseudo-focus qui viennent de s'agréger à la SR du premier niveaux et, d'autre part, par la génération des nouveaux liens que peuvent avoir les voisins du pseudo-focus avec ceux de la SR relative au focus initial. Ceci fournit une vue locale sur les données, car la SR n'est qu'une partie du graphe global.

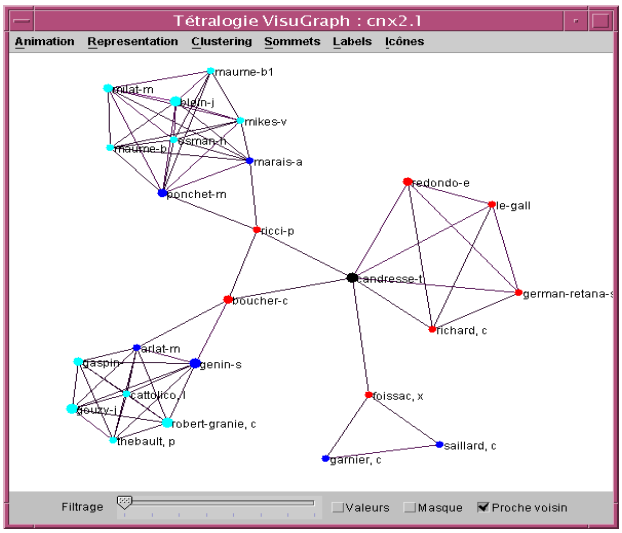

Fig 5 *: Structure réduite relative à un focus (en noir)*

Pour ne pas surcharger l'espace d'affichage et rendre la SR résultante moins complexe, l'utilisateur peut à tout moment réduire le nombre d'éléments affichés soit par filtrage interactif soit en éliminant certains éléments affichés qu'il juge inutile pour la suite de l'exploration. L'intérêt majeur d'une telle démarche est de permettre à l'utilisateur de personnaliser son mode d'exploration en fonction de ses objectifs, ce qui favorise les associations d'idées.

Lors du processus d'exploration, l'utilisateur peut se sentir désorienter. L'utilisation de la couleur des sommets de chaque niveau permet de réduire ce sentiment, par le fait qu'il est facile de distinguer les sommets par leurs couleurs respectives. Par exemple, le focus (niveau 0) aura par défaut une couleur noire, ses voisins (niveau 1) auront une couleur rouge et ceux de niveau 2 bleu et ainsi de suite. Toutefois, il est nécessaire d'accompagner l'exploration

avec un mécanisme d'historique d'opérations effectuées. Cet historique enregistre la séquence du déroulement de l'exploration, en affichant parallèlement les sommets empruntés depuis le début de la navigation. Cette séquence, mise à jour au fur et à mesure des opérations d'augmentation ou réduction de la SR, permet à l'utilisateur d'effectuer des opérations de type "*Annuler*" et "*Retour en arrière*". L'intérêt d'un tel historique est primordiale pour éviter les problèmes de localisation et de désorientation (D'où vient-on ? Quels liens ont été suivis ?).

Notons que lors de cette exploration, la construction des structures réduites est dynamique et que son affichage est géré par un dessin basé sur l'algorithme de forces. L'affichage du sous graphe s'adapte automatiquement aux opérations d'augmentation ou de réduction. Comme la représentation d'un graphe peut induire différentes interprétations, l'utilisateur peut lui-même placer dynamiquement certains sommets à des positions qu'il juge pertinentes. Il peut immobiliser des sommets sur des positions particulières en fonction de son point de vue.

5.3 Extraction de groupes

Dans le cas d'un graphe partitionné, d'autres types de structures peuvent être sélectivement extraites et visualisées en fonction de différents critères. En effet, l'utilisateur peut extraire, par exemple, le groupe auquel appartient le focus pour découvrir son rôle par rapport aux autres sommets du graphe. Le centre d'intérêt peut être une classe à part entière. La classe extraite peut faire l'objet d'une analyse « locale » plus fine. Dans le cas de graphe de grande taille, l'extraction et l'analyse de ses sous-parties permet de diviser un problème complexe en sous problèmes de tailles praticables.

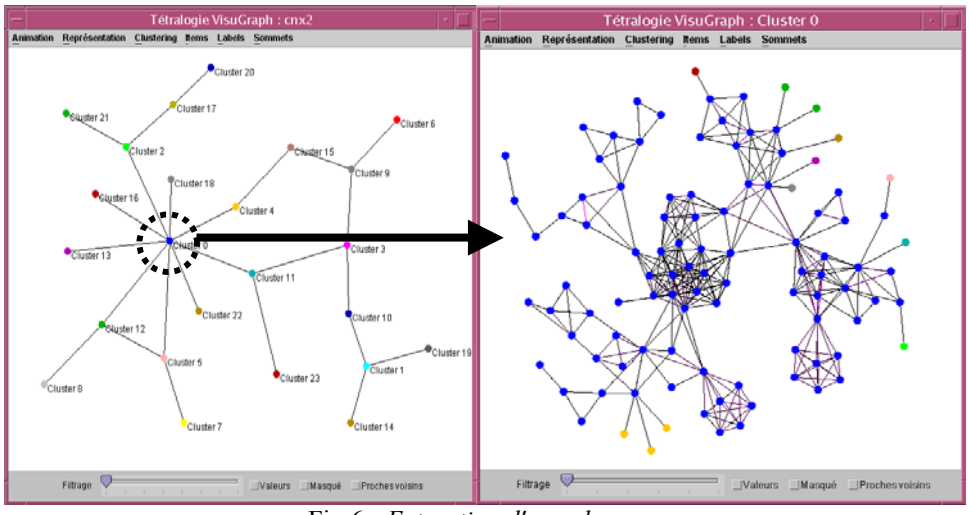

Fig 6 *: Extraction d'une classe*

Par exemple, la figure 6 représente la structure interne de la classe 0, à laquelle nous ajoutons les éventuels sommets qui jouent le rôle de connecteurs entre celle-ci et les autres classes.

6 Conclusion

Dans cet article, nous avons présenté une approche de visualisation interactive comme moyen de représentation et d'analyse d'information issue d'un processus de découverte de connaissances (ECD). Cette approche se distingue de celles utilisées en analyse de données classiques. En effet, les démarches basées sur l'analyse de données permettaient déjà d'accéder à ce type de découverte, mais leur mode de représentation graphique est assez mal adapté pour une restitution grand public. L'utilisation des graphes comme un outil de représentation de données est une technique intuitive et leur visualisation a bénéficié ces dernières années des progrès significatifs réalisés dans plusieurs domaines : théorie des graphes, partitionnement de graphes et visualisation d'information. En plus, la lecture d'un graphe ne nécessite pas de connaissances particulières comme celles exigées pour interpréter une carte factorielle ou un arbre de classification. [Karouach et Dousset, 2002). Le tout est de trouver un graphe à la fois fidèle à la réalité et suffisamment lisible. L'approche basée sur le concept graphe permet une représentation automatique de l'information. L'utilisation des techniques de partitionnement comme MCL a deux objectifs différents mais complémentaires. D'une part, c'est un moyen efficace pour contourner la complexité des structures de grande taille lors de leur visualisation. D'autre part, elles permettent d'extraire les structures macroscopiques contenues dans les données. Associée à des techniques de partitionnement et des moyens interactifs, notre approche peut aider l'utilisateur à réaliser des analyses plus fines, tout en lui permettant une manipulation et une exploration dynamique de son espace informationnel. La navigation dans cet espace peut être réalisée soit sur le graphe de classes soit sur le graphe initial via des sommets représentatifs choisis par l'utilisateur.

Références :

[Alpert et Kahng, 1995] C.J. Alpert, A.B. Kahng. *Recent developments in netlit partitioning : A survey*. Integration : The VLSI journal, vol. 19, pp.1-18, 1995.

[Di Battista et al., 1999] Di Battista G., P. Eades, R. Tamassia et I. Tollis, *Graph Drawing : Algorithm for the visualisation of graphs*, Prentice Hall, 1999.

[Eades, 1984] P. Eades. *A heuristic for Graph Drawing.* Congressus Numerantium, vol. 42, pp. 149-160, 1984.

[Enright et al., 2002] A.J. Enright, S. Van Dongen et C.A Ouzounis. An efficient algorithm for lagre-scale detection of protein families,Nucleic Acids Research, vol. 30, pp. 1575-1584, 2002.

[Jouve et al., 2001] B. Jouve, P. Kuntz et F. Velin. *Extraction de structures macroscopiques dans des grands graphes par une approche spectrale*. ECA, Hermès Science publication édition, vol. 1, pp. 173-184, 2001.

[Karouach et Dousset, 2002] S. Karouach, B. Dousset. *Visualisation de relations par des graphes interactifs de grande taille*. 9ème journées de sur les systèmes d'information élaborée : Bibliométrie - Information stratégique - Veille technologique, CD-ROOM, Ile Rousse (Corse), octobre 2002.

[Karypis et Kumar, 1998] G. Karypis, V. Kumar. *Multilevel k-way partitioning scheme for irregular graphs*. Journal of Parrallel and distributed Computing, vol. 48, pp.96-129, 1998.

[Kuntz et Henaux, 2000] P. Kuntz et F. Henaux. *Numerical comparaison of two spectral decomposition for vertex clustering*. Data Analysis, Classification and Related Methods, Proc. Of IFCS'2000, Springer Verlag, pp.581-586, 2000.

[Melançon et al, 1999] G. Melançon, I. Herman et M. Delest. *Indices visuels et métriques combinatoires pour la visualisation de données hiérarchiques*, Proc. of the IHM'99 Workshop (Onzièmes journées sur l'ingénierie de l'Interaction Homme-Machine), 166-173, Montpellier, 1999.

[Shneiderman,1996] B. Shneiderman. *The eyes have it : A Task by Data Taxonomy for Information Vsualizations*, Proc. Visual Language'96, CO, CS-TR-3665, pp. 336-343, septembre 1996.

[Van Dongen, 2000] S. Van Dongen; *Graph Clustering by Flow Simulation*. Thèse de l'université d'Ultrecht, Allemagne, May 2000.