Panorama de quelques approches récentes pour la classification non supervisée de graphes

Pascale Kuntz, Fabien Picarougne

Equipe COD - Laboratoire d'Informatique de Nantes Atlantique
Ecole Polytechnique de l'Université de Nantes
rue Christian Pauc, BP 50609
44306 Nantes Cedex3
{pascale.kuntz ; fabien.picarougne}@univ-nantes.fr,
http://www.polytech.univ-nantes.fr/COD/

Résumé. Les avancées technologiques récentes ont permis d'acquérir dans de nombreux domaines des corpus de graphes. Une problématique en plein essor consiste à classer ces données complexes pour établir des typologies. Différentes approches développées en fouille de données sont présentées dans cet article: la visualisation de graphes dans une perspective exploratoire, la caractérisation des graphes par des descripteurs structurels et fonctionnels, par des sous-structures et par des décompositions spectrales, et les méthodes à noyaux.

1 Introduction

Pour les pionniers de l'utilisation intensive de l'informatique dans l'analyse des données il y a une trentaine d'années, la tâche ne manquait pas ; que ce soit en Sciences Humaines ou en Sciences de la Vie par exemple, les données s'offraient à l'analyse sous un angle différent avec des technologies prometteuses. Au début des années quatre-vingt, I.C. Lerman notait : « on peut à présent prétendre dégager, derrière une grande masse d'informations, des structures d'organisation entre ces dernières, et cela au grand bonheur d'investigation dans les nombreuses disciplines (...) où la part due à l'observation et donc à la collecte des données est si importante » Lerman (1981). Si les corpus examinés à cette époque avaient été savamment, et souvent laborieusement, constitués par des experts de divers champs disciplinaires, les technologies d'aujourd'hui permettent de recueillir non seulement des volumes de données très importants sur des échelles de temps de plus en plus réduites, mais également des données de plus en plus complexes qui ne se réduisent plus simplement à des tableaux $Individus \times Variables$.

Les graphes, qui modélisent des systèmes de relations, représentent une illustration paradigmatique de l'accessibilité de ces nouvelles données. A titre d'exemple, T. Abello et al.

Cet article est la transcription d'une communication invitée présentée aux $3^{èmes}$ journées thématiques Apprentissage Artificielle et Fouille de Données. Il n'est donc pas un état de l'art exhaustif mais une introduction à une problématique en plein essor dans de nombreux domaines.

Abello et al. (1999) de chez AT&T considèrent des graphes d'appels quotidiens -dont les sommets sont des numéros de téléphones et les arêtes représentent les communications effectuées- de plus de 50 millions de sommets et 170 millions d'arêtes. Pour estimer des paramètres sur la Toile, Barabasi et al. Albert et al. (1999) travaillent sur des graphes échantillons de 300000 documents et 1500000 liens (environ 0.03% de la taille du web indexable à cette époque). Les graphes de contacts ou de co-citations des réseaux sociaux (« les amis de mes amis de mes amis ... ») portent maintenant sur des centaines de milliers de sommets Tjaden et Wasson (2000), etc... L'objectif est ici d'analyser les propriétés structurelles et fonctionnelles de ces grands réseaux de relations.

En parallèle s'est développée une nouvelle problématique : l'analyse de bases de graphes. Une des premières applications phare est certainement la classification de composants chimiques Wünstel (2003) pour les bases de données pharmacologiques. Et les avancées technologiques ont permis de recueillir des corpus de graphes variés dans des disciplines différentes.

Ces différentes problématiques ont donné lieu cette dernière décennie à une prolifération de travaux aux confins de plusieurs disciplines telles que notamment la fouille de données (« link analysis », « graph mining »), la physique statistique, la métrologie et la sociologie (réseaux sociaux). En fouille de données, ce sujet absent à la fin des années 90 connaît un véritable essor visible notamment à travers la prolifération de sessions spécialisées (KDD workshop on Link Analysis and Group Detection, KDD workshop on Multi-relational Data Mining, European Workshop on Mining Graph, Trees and Sequences).

Dans cet article, nous nous focalisons sur la question de la classification non supervisée d'une base de graphes. Nos données sont un ensemble fini de graphes, et nous cherchons à classer ces graphes sans modèle préalable. En paraphrasant le comte de Buffon (Histoire Naturelle, 1749), il s'agit donc de mettre ensemble les graphes qui se ressemblent et de séparer ceux qui diffèrent les uns des autres.

Si ce problème générique de la classification est déjà bien souvent délicat pour des données plus classiques de type $Individus \times Variables$, il est à notre connaissance encore très ouvert pour le cas où les individus sont des graphes. Cette problématique fait elle même référence à d'autres problématiques (difficiles) abondamment traitées en optimisation combinatoire, notamment la recherche d'isomorphismes de graphes, et le calcul de distances entre graphes de type distance d'édition. Nous ne les abordons pas explicitement ici. Nous tentons seulement de dresser un panorama introductif de quelques approches récemment développées en fouille de données qui contribuent à défricher le problème.

Pour illustrer la variété des « données graphes » qui peuvent être rencontrées dans différents champs disciplinaires, nous commençons cet article par une série d'exemples réels (paragraphe 2). Les graphes peuvent être utilisés comme des modèles théoriques puissants en tant qu'objets combinatoires, mais ils peuvent également se visualiser. Dans le paragraphe 3, nous introduisons rapidement l'intérêt des outils de visualisation de graphes dans un objectif de fouille exploratoire. Puis nous poursuivons dans les chapitres suivants par une présentation d'approches plus automatiques. Pour exploiter l'analogie avec le cas classique $Individus \times Variables$ pour lequel une multitude d'algorithmes ont été proposés, une démarche légitime consiste à décrire les graphes de la base par un ensemble fini de descripteurs. Nous présentons différentes voies : dans le chapitre 4 les descripteurs sont des indicateurs structurels et fonctionnels issus de la physique statistique, dans le chapitre 5 les descripteurs indiquent la présence/absence de sous-graphes, dans le chapitre 6 les descripteurs sont déduits d'une analyse spectrale. Enfin,

nous finissons par une approche en plein essor : la construction de noyaux entre graphes. Notons qu'afin de faciliter l'utilisation des nombreuses entrées bibliographiques présentées, nous avons choisi de les présenter par chapitre.

2 Quelques exemples

Dans la suite, nous considérons un ensemble fini de graphes $\Gamma = \{G_1, G_2, ..., G_N\}$ où chaque graphe $G_i = (V_i, E_i)$ est défini par un ensemble V_i de n_i sommets et un ensemble E_i de m_i arêtes ou arcs si le graphe est orienté. On rappelle que n_i désigne l'ordre du graphe G_i et m_i sa taille. Dans certains cas on considèrera un étiquetage sur les sommets et/ou sur les arêtes. On note X_i la matrice $n \times n$ d'adjacence du graphe G_i.

Comme nous l'avons noté précédemment les illustrations les plus nombreuses sont en bioinformatique. Nous présentons brièvement d'autres illustrations provenant de disciplines différentes (urbanisme, histoire, entomologie). Au-delà des problématiques applicatives spécifiques aux différents champs que nous ne détaillons pas ici, ces exemples illustrent la variété des familles de graphes rencontrés : graphes étiquetés par des alphabets standardisés restreints ou des alphabets très grands, multi-graphes orientés, graphes plongés dans des espaces géométriques bi-dimensionnels ou tri-dimensionnels. Les approches proposées pour la classification de ces corpus sont généralement applicables à certaines familles. Mais à notre connaissance, il n'existe pas à l'heure actuelle de démarche générique, ni même comparative, offrant une vue synthétique sur l'adéquation méthodes-caractéristiques des graphes considérés.

2.1 Corpus de graphes de déplacements

La compréhension de la propension à se déplacer est une question clé en aménagement du territoire urbain. Pour les ménages, l'accroissement de la mobilité n'est plus à caractériser uniquement par l'accroissement des sorties du domicile mais plutôt par l'accroissement du nombre de kilomètres parcourus et par la complexification croissante des stratégies de déplacements. Le graphe de la figure 1 est extrait d'une enquête menée dans la communauté urbaine de Brest Wiel et al. (1996) : il représente les déplacements d'un ménage associés à des motifs pré-déterminés (école, travail, achat).

2.2 Corpus de réseaux sociaux

De par l'accessibilité des bases de données scientométriques, les réseaux de co-citations sont devenus un terrain d'étude privilégié. Lorsque l'on se focalise sur une analyse selon un point de vue donné (e.g. temporel, géographique, communautaire, ...) on peut construire un corpus de réseaux (e.g. des réseaux de différentes époques ou associés à différentes localisations). Au-delà de l'exemple cité, les technologies actuelles facilitent la collecte de tels corpus dans différents champs des Sciences Humaines.

Par exemple, la figure 2 représente des graphes de sociabilité entre des paysans du Moyen-Age Boulet et al. (2007). Ces graphes ont été constitués à partir d'actes manuscrits (figure 3) de transactions agraires recueillis sur trois siècles dans une zone géographique restreinte extrêmement bien documentée du Sud-Ouest de la France GraphComp. Des ces actes différents

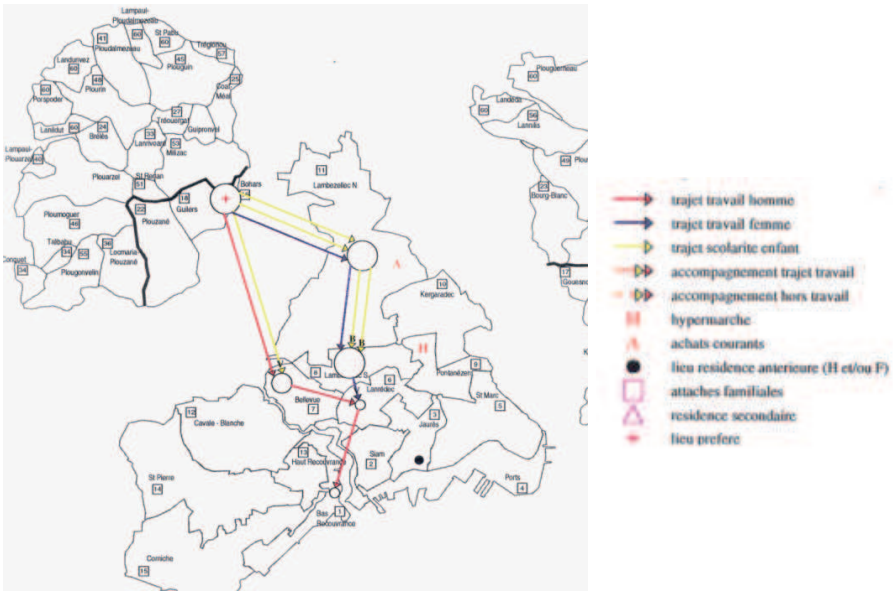

FIG. 1 – *Comportements de mobilité et évolution de l'organisation urbaine de Brest et de sa communauté urbaine.*

réseaux peuvent être construits selon le point de vue d'intérêt ; la figure 2 modélise les relations entre les noms d'individus apparaissant dans une transaction : les sommets du graphe sont donc les individus nommés dans la transaction, et il existe une arête entre deux sommets si les noms associés sont cités dans une même transaction. Pour les médiévistes, ces réseaux de sociabilité permettent de porter un autre regard sur le monde paysan, qui bien que constituant plus de 90% de la population de l'époque, a laissé très peu de traces écrites comparativement aux classes dominantes que furent la noblesse et le clergé.

Les graphes sont ici étiquetés par un alphabet très large (plusieurs dizaines de milliers de noms) et leur ordre peut être assez grand (les sources documentaires contiennent près de 6000 actes relativement bien répartis selon les intervalles de temps considérés).

2.3 Corpus de réseaux de galeries chez les insectes sociaux

Les structures en réseaux abondent dans les sociétés animales. Ceux qui matérialisent le déplacement des individus sur leur territoire résultent d'une activité collective et permettent d'exprimer des échanges sociaux. L'étude des modes de production et d'utilisation collective de ces réseaux est un axe de recherche majeur chez les spécialistes des insectes sociaux (e.g. Bonabeau et al. (1999)). Nous présentons ici deux exemples : des réseaux de galeries chez les fourmis et chez les termites.

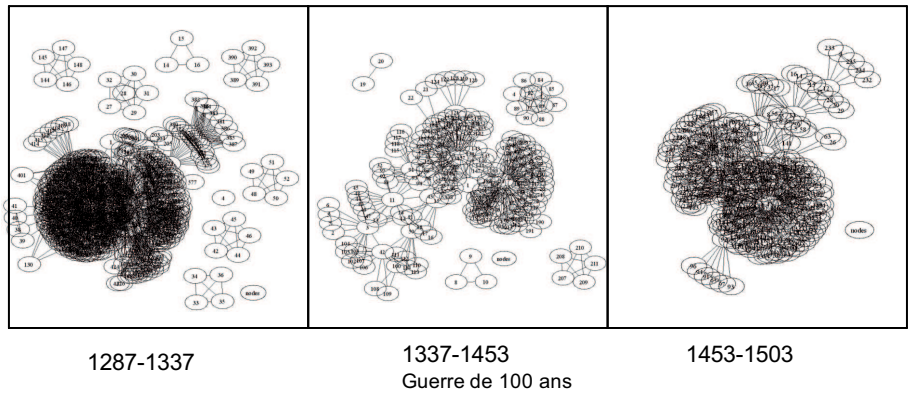

FIG. 2 – *Relations sociales dans la société paysanne médiévale. [projet ANR GRAPH-COMP]*

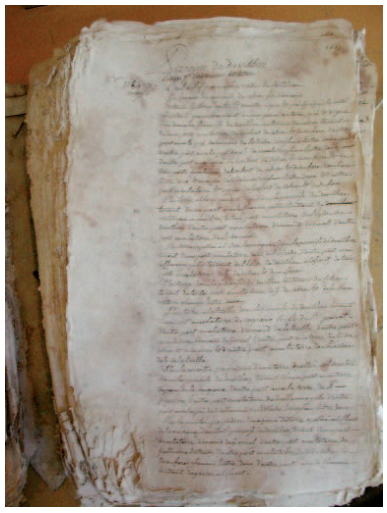

FIG. 3 – *Manuscrit d'actes notariés. (source Archives Départementales du Lot)*

2.3.1 Réseaux bi-dimensionnels

La forme la plus répandue de nid chez les fourmis est un réseau de galeries composé de chambres (cavités) et de galeries reliant les chambres entre elles, et dans la nature, le nid à la surface du sol. Souterraines et extrêmement fragiles, les architectures de ces réseaux sont difficilement accessibles dans leur espace naturel. Par conséquent, des expérimentations en laboratoire ont été menées pour caractériser leurs propriétés géométriques et combinatoires et leurs dynamiques de croissance Buhl et al. (2004). Le dispositif expérimental consistait à déposer des fourmis (100 et 200) sur le pourtour d'un disque de sable de faible épaisseur et à les laisser creuser pendant 72 heures (au-delà pour des raisons encore mal connues le creusement se modifie). Un exemple de réseau réel obtenu est illustré sur la figure 4, et un modèle de graphe associé est présenté sur la figure 5.

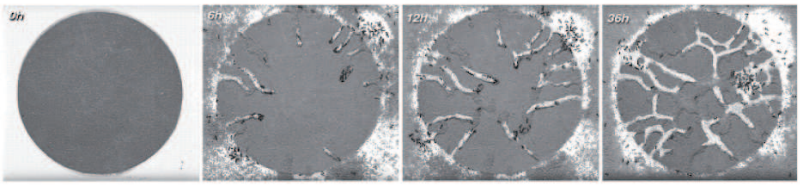

FIG. 4 – *Réseau de galeries creusées par des fourmis.*

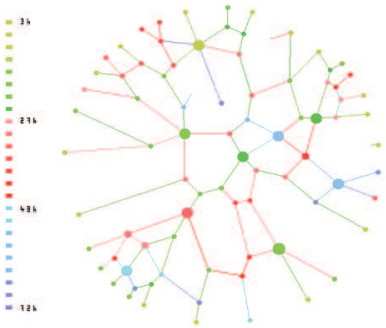

FIG. 5 – *Graphe associé à un réseau de galeries creusées par des fourmis.*

Les graphes peuvent être ici considérés par construction comme des graphes bi-dimensionnels planaires (les sommets correspondent aux chambres où s'intersectent les galeries). Ils sont d'ordre restreint (moins d'une centaine de sommets).

2.3.2 Réseaux tri-dimensionels

Les termites produisent des architectures de nids variées tant par la forme que par la taille des structures. La caractérisation des structurations en réseaux internes est cependant mal connue. Pour contribuer à établir une typologie de ces réseaux, des nids issus d'une collection du Muséum d'Histoire Naturelle ont été scannés (figure 6) et les réseaux de galeries ont été extraits à partir d'une reconstition tri-dimensionnelle des coupes scannées Perna et al. (2008).

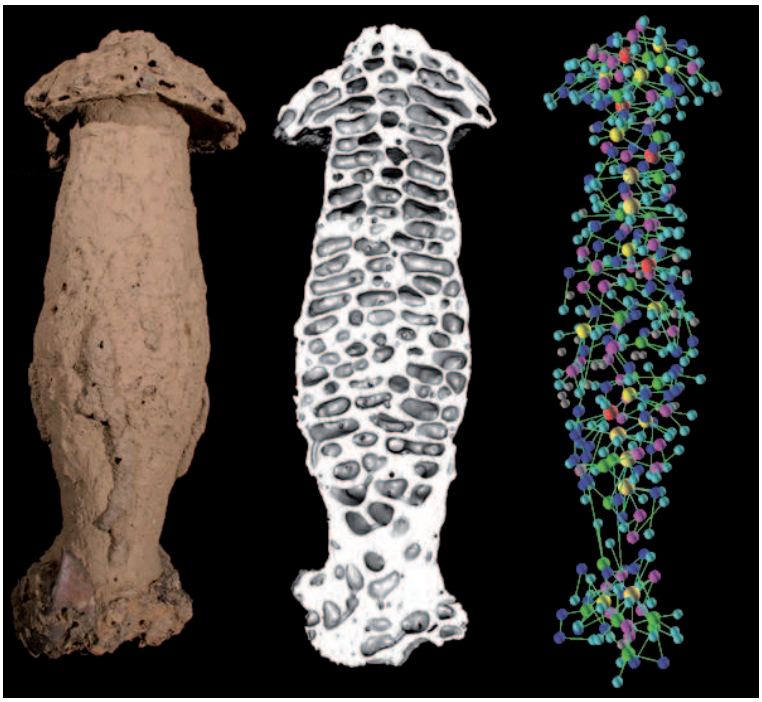

FIG. 6 – *Nid de termite tomographié et graphe associé reconstitué. [projet ANR MESOMORPH]*

Les graphes sont donc ici plongés spatialement dans un espace tri-dimensionnel. Leur ordre est de quelques centaines de sommets.

Références

Abello, J., P. Pardalos, et M. Resende (1999). On maximum clique problems in very large graphs. In J. Abello et J. Vitter (Eds.), *External memory algorithms and visualization*, Volume 50 of *DIMACS Series on Discrete Mathematics and Theoretical Computer Science*, pp. 119–130. American Mathematical Society.

Albert, R., H. Jeong, et A.-L. Barabasi (1999). The diameter of the world-wide web. *Nature 401*, 130–131.

Bonabeau, E., M. Dorigo, et G. Theraulaz (1999). *Swarm Intelligence : From Natural to Artificial Systems*. Oxford.

Boulet, R., F. Hautefeuille, B. Jouve, P. Kuntz, B. L. Goffic, F. Picarougne, et N. Villa (2007). Sur l'analyse de réseaux de sociabilité dans la société paysanne médiévale. In *MASHS 2007 : Computational methods for modelling and learning in social and human sciences*, Brest, France.

Buhl, J., J. Gautrais, R. Solé, P. Kuntz, S. Valverde, J. L. Deneubourg, et G. Théraulaz (2004). Efficiency and robustness in ant networks of galleries. *European Physical Journal B 42*, 123–129.

GraphComp. Site de saisie de données issues de manuscrits d'actes notariés de 1200 à 1700, http ://graphe.dyndns.org.

Lerman, I. (1981). *Classification et analyse ordinale des données*. Dunod.

Perna, A., C. Jost, S. Valverde, J. Gautrais, G. Théraulaz, et P. Kuntz (2008). The topological fortress of termites. In S. LNCS (Ed.), *Proc. of Biowire*. to appear.

Tjaden, B. et G. Wasson (2000). The oracle of bacon. Technical report, University of Virginia.

Wiel, M., A. Morvan, S. Tauty, R.-P. Desse, P. L. Guirriec, J.-P. Barthélemy, P. Kuntz, et E. Dilasseur-Lesaux (1996). *Comportements de mobilité et évolution de l'organisation urbaine, Tomes I, II, III*. Agence d'Urbanisme de la Communauté Urbaine de Brest.

Wünstel, M. (2003). Software development. In J. Gasteiger et T. Engel (Eds.), *Chemoinformatics - A Textbook*. Wiley-VCH Weinheim.

3 Fouille visuelle de graphes

« Quand on étudie la théorie des graphes, eh bien, on pourrait très bien parler de graphes en termes de fonctions 0 et 1 (...), mais non, on les traite en forme de figures parce qu'on veut visualiser l'objet, mettre des points pour représenter les sommets ; les arêtes ce sont des lignes continues qu'on dessine sur le plan et ce sont des propriétés d'un type graphique et visuel qu'on étudie (...) » (Berge). Un des intérêts majeurs des graphes est effectivement de pouvoir caractériser les propriétés d'un système de relations via à un arsenal combinatoire sophistiqué tout en facilitant l'accès à ces structures complexes via notamment les représentations visuelles adaptées.

Berge proposait de voir le dessin dans sa tête ! Mais heureusement depuis les premiers travaux de Knuth Knuth (1963) dans les années 60, la représentation visuelle des graphes sur des supports plus partageables est devenue un domaine de recherche à part entière animé par les communautés « Graph Drawing » (le symposium international Int. Symp. on Graph Drawing est organisé annuellement avec des actes publiés sous la forme de Lecture Notes in Computer Science chez Springer) et « Infoviz » (e.g. Herman et al. (2000)).

Le problème générique du tracé consiste à dessiner un graphe G sous une forme intelligible sur un support standard bi-dimensionnel. La qualité du dessin est évidemment décisive pour l'appropriation de la représentation par l'utilisateur Purchase (2000). Pour préciser cette notion délicate, qui reste in fine subjective, on retient généralement quatre concepts de base Battista et al. (1999) : la convention de tracé, les contraintes physiques, les critères esthétiques, et les contraintes sémantiques. La *convention de tracé* spécifie les règles géométriques de lecture

du tracé qui sont souvent inhérente aux pratiques en vigueur dans le domaine d'application (représentation polygonale où chaque arête est représentée par une ligne polygonale, représentation rectiligne où chaque arête est représentée par un segment de droite, représentation orthogonale etc ...). Les *contraintes physiques* sont inhérentes au support de représentation et à l'oeil humain ; elles imposent notamment des écarts minimums à respecter entre les entités géométriques (points, boîtes, ...) représentant les sommets et les courbes représentant les arêtes. Les *critères esthétiques* tentent de formaliser les propriétés à satisfaire pour faciliter la lisibilité d'un tracé. Ces critères sont définis par des contraintes combinatoires : minimisation du nombre de croisements d'arêtes, minimisation de la somme des longueurs des arêtes ou de la longueur de l'arête maximale, minimisation des coudes, ... Il n'existe pas d'ordonnancement générique de ces critères ; l'interprétation de chaque tracé dépendant de sa propre sémantique. Cependant, des travaux en psychologie cognitive ont montré que la réduction des croisements est le critère prépondérant pour la lisibilité et la mémorisation Purchase (2000). Les *contraintes sémantiques* sont associées à l'interprétation des composantes du graphe ; par exemple, des proximités sémantiques doivent être respectées dans le positionnement des sommets.

De par la variété des combinaisons de ces différentes contraintes, de nombreux algorithmes ont été proposés : nous renvoyons pour une présentation détaillée aux ouvrages de Battista et al. (1999) et de Kaufmann et Wagner (2001). Au-delà de ce problème générique, de nouvelles approches ont été récemment proposées pour aborder le problème spécifique de la représentation des grands graphes où les critères esthétiques sont de natures différentes et les contraintes de temps de calcul et de gestion de la mémoire peuvent devenir critiques. Des logiciels offrent maintenant des fonctionnalités opérationnelles pour représenter une grande variété de graphes. Nous renvoyons à l'ouvrage de Jünger et Mutzel Junger et Mutzel (2003), et au site Web GVSR -Graph Visualization Software References Pinaud et al. (2006).

Dans le cas où aucun modèle de référence n'est donné, ces outils permettent une première analyse des graphes, selon une démarche similaire à celle plus classique de l'analyse de données exploratoire. Cette démarche est actuellement en plein essor en « fouille visuelle de données » (e.g. Poulet et Kuntz (2006); F. Poulet (2008). Cependant, elle n'en est qu'à ses débuts et différents problèmes méthodologiques émergent des premiers retours d'expérience. La complexité des problèmes de tracé, dont beaucoup sont NP-difficiles, a conduit au développement d'heuristiques de résolution très variées où les approches stochastiques jouent un grand rôle. De plus, la nécessité de traiter des graphes de plus en plus complexes a conduit ces dernières années à une forte diversification des modèles de représentation associées aux conventions de tracé Kitchin et Dodge (2002). Ainsi, les restitutions visuelles d'un même graphe avec différents logiciels peuvent être fort différentes, ce qui peut entrainer de nombreux biais d'interprétation.

Références

Battista, G. D., P. Eades, R. Tamassia, et I. Tollis (1999). *Graph Drawing : algorithms for the visualization of graphs*. Prentice Hall.

F. Poulet, B. Le Grand, T. D. (Ed.) (2008). *Actes de l'atelier Visualisation et Extraction de connaissances, 8èmes journées francophones Extraction et Gestion des Connaissances*, Sophia-antipolis.

Herman, I., G. Melançon, et M. S. Marshall (2000). Graph visualization and navigation in information visualization : A survey. *IEEE Trans. Vis. Comput. Graph. 6*(1), 24–43.

Junger, M. et P. Mutzel (2003). *Graph Drawing Software*. Secaucus, NJ, USA : Springer-Verlag New York, Inc.

Kaufmann, M. et D. Wagner (Eds.) (2001). *Drawing Graphs, Methods and Models (the book grow out of a Dagstuhl Seminar, April 1999)*, Volume 2025 of *Lecture Notes in Computer Science*. Springer.

Kitchin, R. et M. Dodge (2002). *Atlas of Cyberspace*. Harlow, England : Pearson Education.

Knuth, D. E. (1963). Computer-drawn flowcharts. *Commun. ACM 6*(9), 555–563.

Pinaud, B., P. Kuntz, et F. Picarougne (2006). The website for graph visualization software references (gvsr). In M. Kaufmann et D. Wagner (Eds.), *Graph Drawing*, Volume 4372 of *Lecture Notes in Computer Science*, pp. 440–441. Springer.

Poulet, F. et P. Kuntz (2006). *Visualisation en extraction de connaissances, numéro spécial Revue Nationale des Technologies de l'Information (RNTI), E1*. Cépaduès-Editions.

Purchase, H. C. (2000). Effective information visualisation : a study of graph drawing aesthetics and algorithms. *Interacting with Computers 13*(2), 147–162.

4 Descripteurs structurels et fonctionnels

En physique statistique et en analyse de réseaux sociaux, la construction de mesures, fonctions de caractéristiques combinatoires et/ou géométriques, du réseau permet de caractériser d'un point de vue macroscopique certaines topologies spécifiques, les plus connues étant les « réseaux invariants d'échelle » Albert et Barabasi (2002), et les « petits mondes » Watts et Strogatz (1998). Certaines mesures peuvent avoir une interprétation fonctionnelle. Ces mesures peuvent être utilisées comme descripteurs pour la classification de graphes. Dans ce cas, se pose la question délicate et peu étudiée à notre connaissance de l'existence de corrélations « intrinsèques » entre certaines mesures.

Nous ne citons ci-dessous pour un graphe $G = (V, E)$ d'ordre n quelconque que les principales mesures et renvoyons, par exemple, à Dorogovtsev et Mendes (2003) pour les détails.

- la *distribution des degrés* : $f(k)$ est la fréquence du degré k et le degré moyen $\overline{k}$ du graphe G. L'exposant γ de la distribution des degrés peut être considéré lorsque celle-ci suit une loi caractéristique telle qu'une loi puissance $f(k) \sim k^{-\gamma}$
- l'*efficacité globale* est définie par la moyenne de l'inverse des plus court chemins entre les paires de sommets. Notons d_{ij} la distance du plus court chemin entre les sommets quelconques v_i et v_j de V. L'efficacité globale est définie par

$$E(G) = \frac{1}{n(n-1)} \sum_{v_i \neq v_j \in V} \frac{1}{d_{ij}}$$

  Le calcul de la distance d_{ij} peut tenir compte uniquement de la combinatoire du réseau (nombre d'arêtes minimum ou somme des poids minimum lorsqu'il y a une pondération) ou de sa géométrie (somme des longueurs des arêtes minimum) ;

– le *coût* rend compte du nombre -éventuellement pondéré- des connexions dans le graphe. Notons w_{ij} le poids de l'arête $\{v_i, v_j\}$; $w_{ij} = 1$ pour tout v_i, v_j si le graphe n'est pas pondéré. Le côut est défini par

$$Cost\,(G) = \frac{\sum_{v_i \neq v_j \in V} x_{ij} w_{ij}}{\sum_{v_i \neq v_j \in V} w_{ij}}$$

– le *coefficient de « clustering »* caractérise la densité de connexions dans le voisinage de chaque sommet. Soit v_i un sommet de G. Etant donnée une définition de voisinage (par exemple les sommets adjacents de v_i), on note $\delta\,(v_i)$ le degré de v_i. Si tous les voisins étaient connectés entre eux alors il y aurait $\delta\,(v_i)\,.\,(\delta\,(v_i)+1)\,/2$ arêtes. Le coefficient de clustering associé à v_i est défini par

$$C\,(v_i) = \frac{2\delta\,(v_i)}{\delta\,(v_i)\,.\,(\delta\,(v_i)+1)}$$

Le coefficient de clustering $\overline{C}$ est la moyenne des $C\,(v_i)$ sur l'ensemble des sommets v_i de G.

– la *robustesse* caractérise la facilité/difficulté avec laquel le graphe se fragmente en plusieurs composantes connexes lorsque l'on détruit des sommets et/ou des arêtes. Considérons une expérience dans laquelle on détruit un pourcentage de r sommets (resp. arêtes) dans G. Notons s la taille relative de la plus grande composante connexe du graphe résultant. La destruction des sommets peut s'effectuer de façon aléatoire ou préférentielle en détruisant de préférence les sommets ayant des degrés forts. L'expérience est réitérée un nombre fixé de fois. La robustesse est la valeur moyenne de r pour laquelle $s = 0.5$.

Références

Albert, R. et A.-L. Barabasi (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics 74*, 47–97.

Dorogovtsev, S. N. et J. F. F. Mendes (2003). *Evolution of Networks : From Biological Nets to the Internet and WWW (Physics)*. New York, NY, USA : Oxford University Press, Inc.

Watts, D. J. et S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature 393*(6684), 440–442.

5 Caractérisation par sous-structures

Pour comparer des graphes, on peut décrire chacun d'eux par un vecteur binaire indiquant la présence/absence de sous-structures caractéristiques. Cette description permet ainsi l'application des nombreuses dissimilarités sur tableau de présence/absence bien connues de la littérature taxonomique Hubálek (1982). Pour certaines applications, la classification de composants chimiques en particulier, ces structures caractéristiques sont prédéterminées à l'avance. Le problème se pose alors comme un problème difficile d'isomorphisme de sous-graphes. Dans d'autres cas, les structures ne sont pas connues initialement. Différents travaux en fouille des données consistent à rechercher des sous-structures fréquentes.

Une des premières publications en ce sens Inokuchi et al. (2000) propose une adaptation de l'algorithme A Priori initialement appliqué à la fouille de règles d'association. Cet algorithme recherche des sous-structures fréquentes dans une base de données en utilisant la propriété d'anti-monotonie : la fréquence d'une sous-structure est toujours plus petite ou égale à celle de ses sous-structures. Appliquée aux règles d'association qui modélisent une tendance implicative entre des conjonctions d'attributs (ex : $ab \rightarrow c$ où a, b et c sont des attributs décrivant des objets dans une base) cette propriété permet de limiter l'exploration de l'espace de recherche des règles potentiellement intéressantes : si abc et bcd sont des conjonctions d'attributs fréquentes de taille 3 alors l'algorithme va considérer comme candidate à l'évaluation la conjonction $abcd$ de taille 4.

Appliquée aux graphes, cette propriéte consiste à engendrer un sous-graphe candidat de taille $k+1$ à partir de sous-graphes fréquents de taille k. Différentes approches ont été proposées. Les approches « bottom-up » consistent à engendrer des candidats ayant un sommet, une arête ou un chemin supplémentaire par concaténation Inokuchi et al. (2000); Kuramochi et Karypis (2001); Vanetik et al. (2002). Par exemple, dans le cas le plus simple, on construit des sous-graphes candidats de taille $k+1$ à partir de deux sous-graphes candidats de taille k qui ne différent que d'une arête. Considérons deux sous-graphes de taille k, $G(X_k)$ et $G(Y_k)$ de matrices d'adjacence respectives X_k et Y_k. Ces matrices peuvent se décomposer sous la forme suivante :

$$X_k = \begin{pmatrix} X_{k-1} & x_1 \\ x_2 & 0 \end{pmatrix} \text{ et } Y_k = \begin{pmatrix} X_{k-1} & y_1 \\ y_2 & 0 \end{pmatrix}$$

La matrice d'adjacence du graphe candidat $G(Z_{k+1})$ de taille $k+1$ peut donc s'écrire sous la forme

$$Z_{k+1} = \begin{pmatrix} X_{k-1} & x_1 & y_1 \\ x_2 & 0 & z_{k,k+1} \\ y_2 & z_{k+1,k} & 0 \end{pmatrix}$$

Une illustration est donnée sur la figure 7.

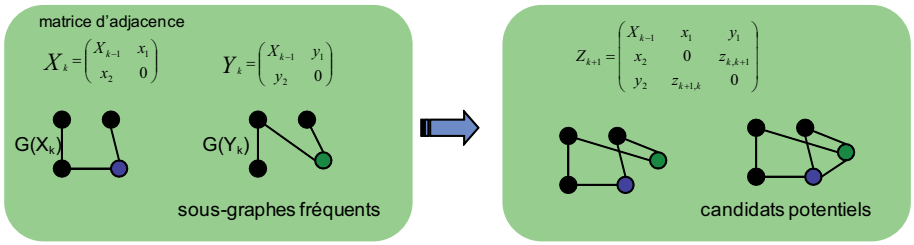

FIG. 7 – *Construction de sous-graphes candidats de taille $k+1$ à partir de deux sous-graphes candidats de taille k.*

Pour tenter d'améliorer ce processus, des approches « par croissance » basées sur une extension récursive d'une sous-structure ont été récemment proposées Yan et Han (2002); Borgelt et Berthold (2002); Huan et al. (2003, 2004).

Ces approches ont été essentiellement appliquées à des bases de graphes modélisant des molécules. Les bases peuvent être de très grande taille mais les graphes considérés ont des

degrés relativement faibles (inférieurs à 5), un étiquetage des sommets et des arêtes qui restreint le champ des combinaisons possibles dans l'exploration de l'espace de recherche et des tailles restreintes (moins d'une trentaine de sommets).

Cependant, la complexité de ces approches reste très élevée et comme l'ont récemment souligné Kuramochi et Karypis (2007) *« it is unclear whether those algorithms can operate efficiently on other types of graph datasets as they do on chemical graphs »*.

Références

Borgelt, C. et M. R. Berthold (2002). Mining molecular fragments : Finding relevant substructures of molecules. In *ICDM '02 : Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, Washington, DC, USA, pp. 51. IEEE Computer Society.

Huan, J., W. Wang, et J. Prins (2003). Efficient mining of frequent subgraphs in the presence of isomorphism. *icdm 00*, 549.

Huan, J., W. Wang, J. Prins, et J. Yang (2004). Spin : mining maximal frequent subgraphs from graph databases. In W. Kim, R. Kohavi, J. Gehrke, et W. DuMouchel (Eds.), *KDD*, pp. 581–586. ACM.

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (present-absence) data : an evaluation. *Biological Review 57*, 669–689.

Inokuchi, A., T. Washio, et H. Motoda (2000). An apriori-based algorithm for mining frequent substructures from graph data. In *PKDD '00 : Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, London, UK, pp. 13–23. Springer-Verlag.

Kuramochi, M. et G. Karypis (2001). Frequent subgraph discovery. In *ICDM '01 : Proceedings of the 2001 IEEE International Conference on Data Mining*, Washington, DC, USA, pp. 313–320. IEEE Computer Society.

Kuramochi, M. et G. Karypis (2007). Graph matching – finding topological frequent patterns from graph datasets. In D. J. Cook et L. B. Holder (Eds.), *Mining Graph Data*, Chapter 6, pp. 117–158. Wiley.

Vanetik, N., E. Gudes, et S. E. Shimony (2002). Computing frequent graph patterns from semistructured data. In *ICDM '02 : Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, Washington, DC, USA, pp. 458. IEEE Computer Society.

Yan, X. et J. Han (2002). gspan : Graph-based substructure pattern mining. In *ICDM '02 : Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, Washington, DC, USA, pp. 721. IEEE Computer Society.

6 Descripteurs basés sur une décomposition spectrale

Depuis la précédente décennie l'analyse spectrale des graphes Mohar (1992); Chung (1997); Cvetkovic et al. (1997) connait un essor important qui peut s'expliquer en partie par l'accroissement des capacités de calcul qui rendent accessibles des calculs matriciels à grande échelle.

Rappelons que le Laplacien discret peut se voir dans une première approche comme une adaptation à un graphe du Laplacien $\nabla.\nabla f$ (où $\nabla f = (\partial f/\partial x, \partial f/\partial y...)$ représente le gradient de f) qui permet d'évaluer une différence entre une fonction f en un point fixé et la moyenne de f dans une région autour de ce point. A titre illustratif considérons le graphe G de la figure 8. Le gradient ∇G peut s'écrire sous la forme

$$\nabla G = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

FIG. 8 – *Graphe exemple.*

où les lignes sont associées aux arêtes (a, b, c, d) et les colonnes aux sommets (1, 2, ...5). Le signe est ici arbitraire et pourrait être inversé. Le Laplacien $L = \nabla.\nabla G$ peut alors s'écrire comme le produit de la transposée de ∇G par lui-même :

$$L = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Cette expression se retrouve directement à partir de la matrice d'adjacence X de G. Notons D la matrice $n \times n$ diagonale des degrés des sommets de G. Alors,

$$L = D - X$$

La matrice L est une matrice semi-définie positive, et ses valeurs propres sont par conséquent positives. Notons $\lambda_1, \lambda_2, ..., \lambda_n$ ses valeurs propres et $\mu_1, \mu_2, ..., \mu_n$ les vecteurs propres associés. On a

$$L = MM^t$$

où $M = \left(\sqrt{\lambda_1}\mu_1, \sqrt{\lambda_2}\mu_2, ..., \sqrt{\lambda_n}\mu_n\right)$.

Le spectre du Laplacien discret permet de rendre compte efficacement de propriétés des graphes. Cependant, son utilisation pour la classification de graphes n'est pas immédiate. En effet, la topologie du graphe est invariante par permutation des étiquettes des arêtes et des sommets, mais ce n'est pas le cas de la matrice spectrale L qui est dépendante d'un réordonnancement des lignes. Par conséquent, à moins d'un étiquetage standardisé lié à une sémantique particulière sur les sommets, on ne peut pas comparer différents graphes en comparant directement leurs matrices spectrales.

Une première alternative basée sur une composition des vecteurs propres s'est restreinte à la comparaison de graphes de même taille (e.g. Umeyama (1988)). Une extension probabiliste

-qui traite les différences comme des valeurs manquantes- a ensuite été proposée pour palier cette restriction Luo et Hancock (2001). D'autres approches utilisent des plongements des graphes dans des espaces de dimension restreinte définis par les premiers vecteurs propres, qui restituent les propriétés structurelles « les plus marquées » du graphe (e.g. Kosinov et Caelli (2002)).

Plus récemment, les polynômes symétriques fonctions des vecteurs propres ont été introduits pour éviter le problème de non invariance par permutation Wilson et al. (2005). Notons M_{ij} les composantes de la matrice M. On remarque que $\lambda_j = \sum_{i=1,n} M_{ij}^2$est un polynôme symétrique pour les composantes du vecteur propre μ_j. D'où l'idée d'utiliser les polynômes symétriques élémentaires (qui forment une « base » pour les polynômes symétriques) pour caractériser un graphe. Rappelons que pour un ensemble de variables $\{x_1, x_2, ..., x_n\}$ ces polynômes élémentaires s'écrivent sous la forme suivante :

$$
\begin{aligned}
S_1\left(x_1, x_2, ..., x_n\right) &= \sum_{i=1,n} x_i \\
S_2\left(x_1, x_2, ..., x_n\right) &= \sum_{i=1,n} \sum_{j=i+1,n} x_i x_j \\
&\cdots \\
S_n\left(x_1, x_2, ..., x_n\right) &= \prod_{i=1,n} x_i
\end{aligned}
$$

Le polynôme $\prod_{i=1,n}(x - x_i)$ a pour racines $x_1, x_2, ..., x_n$, et donc $x^n - S_1 x^{n-1} + S_2 x^{n-2} + ... + (-1)^n S_n = 0$. Ainsi, on peut construire les n polynômes élémentaires $S_1(M_{1j}, M_{2j},..., M_{nj})$, ..., $S_n\left(M_{1j}, M_{2j}, ..., M_{nj}\right)$ pour chaque vecteur $(M_{1j}, M_{2j}, ..., M_{nj})$: ils sont invariants par permutation et permettent de retrouver l'information contenue dans la description spectrale. Cela constitue au total n^2 descripteurs pour chaque graphe. Pour rendre cette approche opérationnelle, les auteurs préconisent une réduction préalable de l'espace de description, par exemple par ACP.

Références

Chung, F. (1997). *Spectral Graph Theory*. Number 92 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Cvetkovic, D., P. Rowlinson, et S. Simic (1997). *Eigenspaces of Graphs*. Encyclopedia of Mathematics and its Applications. Cambridge : Cambridge University Press.

Kosinov, S. et T. Caelli (2002). Inexact multisubgraph matching using graph eigenspace and clustering models. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, London, UK, pp. 133–142. Springer-Verlag.

Luo, B. et E. R. Hancock (2001). Structural graph matching using the em algorithm and singular value decomposition. *IEEE Trans. Pattern Anal. Mach. Intell. 23*(10), 1120–1136.

Mohar, B. (1992). Laplace eigenvalues of graphs—a survey. *Discrete Math. 109*(1-3), 171–183.

Umeyama, S. (1988). An eigendecomposition approach to weighted graph matching problems. *IEEE Trans. Pattern Anal. Mach. Intell. 10*(5), 695–703.

Wilson, R. C., E. R. Hancock, et B. Luo (2005). Pattern vectors from algebraic graph theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*(7), 1112–1124.

7 Méthodes à noyaux

Les méthodes à noyaux connaissent une utilisation extensive en apprentissage automatique (e.g. Schölkopf et Smola (2002)). Elles permettent de plonger les données initiales dans des espaces de grandes tailles pour extraire certaines de leurs propriétés en gardant des coûts de calculs efficaces.

Etant donné un ensemble d'objets O, un noyau peut se voir comme une généralisation d'un produit scalaire. Rappelons qu'un noyau défini positif sur O est une fonction $K : O \times O \rightarrow \mathbb{R}$ qui est symétrique ($K(o_i, o_j) = K(o_j, o_i)$ pour toute paire (o_i, o_j) de O) et vérifie $\sum_{i=1,n} \sum_{j=1,n} \alpha_i \alpha_j K(o_i, o_j) \geq 0$ pour tout $\alpha_i, \alpha_j \in \mathbb{R}$. N'importe quel noyau défini positif peut être représenté comme un produit scalaire dans un espace de Hilbert. Les propriétés mathématiques de ces noyaux ne sont pas récentes (e.g. Schoenberg (1938); Aronszajn (1950)), mais leur utilisation en apprentissage est intimement liée à l' « astuce noyau » : tout algorithme qui ne nécessite que des produits scalaires entre vecteurs peut être effectué implicitement dans un espace de Hilbert.

Pour appliquer ce type d'approche à la classification de graphes, la difficulté principale consiste à construire un noyau défini positif $K(G_i, G_j)$ qui ait du « sens » sur les graphes. Notons que ce problème est différent de celui de la construction d'un noyau sur les sommets d'un graphe (noyau de diffusion, ...).

Depuis l'article fondateur de Haussler (1999), différentes propositions ont été développées dans la littérature Kashima et Inokuchi (2002); Kashima et al. (2003); Gärtner (2003); Gaertner et al. (2003); Kashima et al. (2004).

Toutes reposent sur la même idée de base : deux graphes sont similaires si ils ont de nombreux chemins en commun. Notons $C_i(G_1)$ (resp. $C_i(G_2)$) l'ensemble des chemins de longueur i de G_1 (resp. G_2). Une forme générique de noyau peut alors s'écrire de la façon suivante :

$$K(G_1, G_2) = \lim_{n \rightarrow \infty} \sum_{i=1}^{n} \sum_{c_1 \in C_i(G_1)} \sum_{c_2 \in C_i(G_2)} \varpi_i k_\sigma(c_1, c_2)$$

où ϖ_i est un réel pondérant les chemins de longueur i, et $k_\sigma(c_1, c_2)$ est un noyau défini sur les chemins.

Les principales différences entre les noyaux présentés dans la littérature reposent sur trois aspects : la façon de prendre en compte les composantes du graphe (étiquettes sur les noeuds et/ou sur les arcs), la façon de parcourir le graphe et l'affectation des poids.

Pour illustrer la démarche, nous présentons ici un noyau introduit par Gärtner (2003). Considérons un graphe G dont les sommets et les arêtes sont étiquetés. Un chemin c de longueur i dans G peut donc être décrit par une séquence d'étiquettes de la forme $s = s_1 s_2 ... s_{2n+1}$ où s_i représente une étiquette d'arc ou de sommet. On considère l'ensemble de tous les descripteurs (séquences) possibles ; et, pour chaque séquence s on compte le nombre de chemins

de G qui coïncident avec cette séquence. Ainsi, on associe à chaque descripteur s la valeur $\Phi_s(G) = \sqrt{\varpi_i} card\{c \in C_i(G); s_i = et_i(c)\}$ où $et_i(c)$est la ième étiquette du chemin c.

Considérons maintenant deux graphes G_1 et G_2. On note

$$V(G_1 \times G_2) = \{(v_1, v_2) \in V_1 \times V_2; et(v_1) = et(v_2)\}$$
$$E(V_1 \times V_2) = \{(u_1, u_2), (v_1, v_2) \in V^2(G_1 \times G_2); (u_1, v_1) \in E_1, (u_2, v_2) \in E_2, et(u_1, v_1) = et(u_2, v_2)\}$$

On note $E_\otimes$ la matrice d'adjacence du produit direct $E(V_1 \times V_2)$: ses composantes $E_{\otimes ij}$ valent 1 si $(v_i, v_j) \in E(V_1 \times V_2)$ et 0 sinon. Et $V_\otimes$ est l'ensemble des sommets du produit direct $V(G_1 \times G_2)$.

Le noyau $K(G_1, G_2)$ défini, si la limite existe, par

$$K(G_1, G_2) = \sum_{i,j=1}^{|V_\otimes|} \left(\sum_{n=0}^{\infty} \varpi_n E_\otimes^n \right)_{ij}$$

coïncide avec le produit scalaire $\langle \Phi(G_1), \Phi(G_2) \rangle$ défini sur l'espace des descripteurs.

Son calcul opérationnel dépend de la série matricielle. Il peut être calculé efficacement, avec une complexité cubique, pour certains choix des poids ϖ_i.

Références

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society 68*.

Gaertner, T., P. Flach, et S. Wrobel (2003). On graph kernels : Hardness results and efficient alternatives. In *Proceedings of the 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop*, pp. 129–143. Springer-Verlag.

Gärtner, T. (2003). A survey of kernels for structured data. *SIGKDD Explor. Newsl. 5*(1), 49–58.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, University of California at Santa Cruz.

Kashima, H. et A. Inokuchi (2002). Kernels for graph classification. In *ICDM Workshop on Active Mining*.

Kashima, H., K. Tsuda, et A. Inokuchi (2003). Marginalized kernels between labeled graphs. In T. Faucett et N. Mishra (Eds.), *20th International Conference on Machine Learning*, pp. 321–328. AAAI Press.

Kashima, H., K. Tsuda, et A. Inokuchi (2004). Kernels for graphs. In K. T. Schoelkopf, B. et J. Vert (Eds.), *Kernel Methods in Computational Biology*, Cambridge, MA ; USA, pp. 155–170. MIT Press.

Schölkopf, B. et A. Smola (2002). *Learning with Kernels*. Cambridge, MA, USA : MIT Press.

Schoenberg, I. J. (1938). Metric spaces and positive definite functions. *Transactions of the American Mathematical Society 44*(3), 522–536.

8 Conclusion

Nous avons présenté dans cet article différentes propositions qui permettent d'aborder le problème de la classification d'une base de graphes sans modèle préalable. Chacune des pistes évoquées soulève des questions délicates.

Au-delà des problèmes algorithmiques inhérents à la manipulation de graphes, la fouille de données visuelle ne peut faire abstraction des avancées en psychologie cognitive concernant le traitement humain de l'information visuelle. En particulier, un débat est actuellement en cours sur l'apport des restitutions visuelles tri-dimensionnelles. De plus, d'un point de vue opérationnel pour l'analyste, une méthodologie reste à définir pour l'intégration de la visualisation dans un processus de fouille.

Au-delà des problématiques de calculs inhérentes à chacune des méthodes exposées, la description des graphes par des vecteurs de caractéristiques repose, mais avec d'autres mesures, le problème de l'analyse des relations entre variables. Ce problème bien classique peut ici différer de son positionnement courant : la structuration des données peut induire certaines corrélations dans toute la population de graphes dont on observe un échantillon.

Les méthodes à noyaux peuvent permettre de faire abstraction de cette question en considérant comme espace de description un espace de Hilbert. Cependant, la construction opérationnelle de noyaux pertinents sur des graphes de familles variées est encore un problème largement ouvert tant sur le plan théorique (définition du noyau) qu'algorithmique (calcul efficace du noyau).

Summary

In various domains, new graph corpus are now available thanks to the recent technological advances. A problem in full expansion consists in clustering these complex data for defining typologies. Different approaches developed in data mining are presented in this paper: graph visualization for structure exploration, graph characterization by structural and functional descriptors, by sub-structures and by spectral decompositions, and kernel methods.