Analyse de données pour des graphes étiquetés

Thibault Laurent, Nathalie Villa-Vialaneix

Analyse de données pour des graphes étiquetés

Thibault Laurent[1] & Nathalie Villa-Vialaneix[2]

[1] *Toulouse School of Economics, 21 allées de Brienne, 31000 Toulouse, France -* *thibault.laurent@univ-tlse1.fr*

[2] *SAMM, Université Paris 1, 90 rue de Tolbiac, 75634 Paris cedex 13, France -* *nathalie.villa@univ-paris1.fr*

Résumé. Nous proposons une méthode de fouille de données pour un graphe dont les sommets sont étiquetés. Deux approches sont décrites et illustrées sur un jeu de données réelles : elles permettent une représentation du graphe qui combine les informations sur sa structure et sur la valeur de ses étiquettes. Cette visualisation peut être utilisée à des fins d'interprétation pour apporter des informations plus nuancées sur la caractérisation des sommets du graphe. [1]

Mots-clés. réseau, graphe, graphe étiqueté, noyau, ACP, visualisation

Abstract. This article presents a data mining method dedicated to graphs whose nodes are labelled. Two approaches are described and illustrated on a real-life data set. A representation of the graph is provided that combines information on its structure as well as on its labels. This representation is intended to be meaningful and to provide a more subtle description of the nodes that their actual labelling.

Keywords. network, graph, labelled graph, kernel, PCA, visualization

1 Introduction

Le cadre de cette communication est la fouille de données de graphes dont les sommets sont étiquetés. Les graphes, aussi souvent appelés réseaux, sont des outils naturels de modélisation de données relationnelles, c'est-à-dire, de données dans lesquelles les individus sont décrits par leurs relations les uns aux autres et non par une série d'attributs qualitatifs ou numériques. L'intérêt pour ce type de modèle a considérablement crût ces dernières années, avec le développement des études sur les réseaux. Les graphes étudiés peuvent contenir un grand nombre de sommets, plusieurs centaines, milliers voire plusieurs dizaines ou centaines de milliers [2] et des informations supplémentaires sur chacun des sommets sont parfois disponibles. Dans ce dernier cas, on parle de *graphes étiquetés*.

La compréhension intuitive de ces grands jeux de données est difficile sans l'appui d'outils de fouille de données. Parmi celles-ci, on trouve les méthodes de visualisation de

1. Ce travail est financé par l'ANR ModULand (Programme blanc SHS 1 2011 0).

2. Voir, par exemple, `http://snap.stanford.edu/data` pour une collection de très grands graphes.

graphes, qui visent à proposer à l'utilisateur une vision globale ou simplifiée du graphe, dans laquelle des informations supplémentaires peuvent être ajoutées aux sommets par le biais d'étiquettes, de couleurs, de forme ou de taille des sommets : de nombreux logiciels de visualisation de graphes ont été développés [3]. Dans la plupart de ces approches, la visualisation est construite de manière indépendante de la valeur des étiquettes du graphe, dans un but purement esthétique, souvent à l'aide de modèle de forces du type de celui initialement proposé par Fruchterman et Reingold [8]. Ainsi, la vision du graphe fournie n'est pas forcément bien adaptée à un objectif de compréhension et d'interprétation, particulièrement dans le cadre où les sommets sont étiquetés. Une approche combinant structure du graphe et valeur des étiquettes permet d'avoir une vision globale de l'intégralité de l'information disponible. Toutefois, certains travaux abordent ce problème : les approches connues sous le nom de *clustered graph visualization* [4, 6, 7, 11], qui s'appuient en général sur des algorithmes de forces modifiés, obtenus en intégrant les contraintes de la classification dans des modèles de forces classiques, ou, [5] qui développe un modèle probabiliste de classification supervisée de projection dans un espace latent.

Dans l'article présenté ici, nous proposons une approche dont l'objectif est légèrement différent : en caractérisant chaque sommet par la distribution des étiquettes de ses voisins, et non seulement par sa propre étiquette, l'idée est de représenter à proximité les sommets dont la distribution des étiquettes des voisins est similaire. Deux approches sont décrites (section 2) : la première basée sur la simple détermination de la distribution et la seconde basée sur un noyau, permettant une caractérisation plus nuancée de la distribution des étiquettes voisines. Les deux approches sont illustrées sur un jeu de données réelles (section 3).

2 Description de l'approche proposée

L'approche proposée repose sur la représentation de chaque sommet par un comptage ou une diffusion des étiquettes de ses voisins. Deux méthodes sont décrites : la première repose sur une ACP, dont les pondérations sont adaptées au contexte, des comptages. La seconde repose sur une idée similaire, par diffusion des étiquettes à l'aide d'un noyau. Dans la suite, on notera $\mathcal{G} = (V, W)$ un graphe simple, symétrique et pondéré, de sommets $V = \{x_1, \ldots, x_n\}$ et de matrice d'adjacence W, $\mathcal{C} = \{c_1, \ldots, c_K\}$ les étiquettes possibles des sommets, E la matrice du codage disjonctif des étiquettes des sommets : $E_{ic} = 1$ signifie donc que le sommet x_i porte l'étiquette c. Notons que le cadre d'étude n'exclut pas des étiquettes multiples pour un sommet.

3. Voir, entre autres, les projets libres Gephi [2] `http://gephi.org` ou Tulip [1] `http://tulip.labri.fr`

2.1 ACP de la distribution des étiquettes voisines

La première approche repose sur le calcul simple de WE dont les éléments, n_{ic} correspondent au nombre de voisins du sommet x_i portant l'étiquette c. Si on note $D = \mathrm{Diag}(d_1, \ldots, d_n)$ la matrice diagonale des degrés des sommets, chaque ligne de $P_l = D^{-1}WE$ est donc la distribution des étiquettes chez les voisins du sommet considéré. Notons que, contrairement au cas de l'analyse de tables de contingence classiques, la somme des éléments de $D^{-1}WE$ n'est pas nécessairement égale à 1 puisque les étiquettes multiples sont autorisées. Le choix de la pondération de chaque ligne par $\frac{1}{d_i}$ est dicté par la nécessité de ne pas donner une importance trop grande aux sommets de fort degré.

L'utilisation d'une ACP sur la matrice P_l revient donc à rechercher un plongement du graphe dans un espace de faible dimension dans lequel la proximité entre sommets dans le plan de projection est déterminé par la proximité dans la distribution des étiquettes dans le voisinage. Les colonnes de la matrice sont pondérées par $\frac{n_c}{n}$ où $n_c = \sum_i E_{ic}$ est le nombre de sommets dont l'étiquette est c. Le choix de cette pondération est équivalent à l'utilisation de la métrique $\delta(p_i, p_{i'}) = \sum_c \frac{n}{n_c}\left(\frac{n_{ic}}{d_i} - \frac{n_{i'c}}{d_{i'}}\right)^2$ entre les lignes $p_i = \frac{n_{ic}}{d_i}$ de la matrice P_l. Ce choix a pour but de favoriser les différences dans les distributions des étiquettes faiblement représentées sur le graphe ; dans ce sens, elle peut-être comprise de manière relativement similaire à l'utilisation de la métrique du χ^2 dans une AFC, même si, ici, $\sum_i n_{ic} \neq n_c$.

2.2 Diffusion d'étiquettes par noyau

Une des limites de l'approche précédente est qu'elle ne tient pas compte du tout de la valeur même de l'étiquette propre de chacun des sommets mais seulement de la distribution des étiquettes dans le voisinage. De plus, les étiquettes des seuls voisins directs sont prises en compte. Pour pallier ce problème, nous décrivons ici une approche alternative, basée sur l'utilisation d'un noyau de diffusion sur le graphe. Le noyau de la chaleur du graphe $\mathcal{G}$ est la matrice $n \times n$ définie par $K^\beta = e^{-\beta L}$ où L est le Laplacien du graphe : $L = D - W$. Il peut être interprété comme la limite d'un processus de diffusion de chaleur (d'où son nom) le long des arêtes du graphe [9] et correspond au produit scalaire entre les images des sommets du graphe par un plongement, ϕ dans un espace de Hilbert $\mathcal{K}^\beta$: $K^\beta_{ij} \equiv K^\beta(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{K}^\beta}$. Le fait de connaître le produit scalaire entre les images des sommets par ϕ sans avoir besoin d'expliciter ni ϕ, ni $\mathcal{K}^\beta$ a popularisé les méthodes dites « à noyau » [10, 3].

Utilisant la propriété de reproduction du noyau de la chaleur, la matrice $K^\beta E$ a donc pour éléments $\tilde{f}^\beta_{ic} = \langle \phi(x_i), \sum_{j:c_j=c} \phi(x_j) \rangle_{\mathcal{K}^\beta}$ et peut donc être vue comme une mesure de dissimilarité entre le sommet considéré x_i et un sommet moyen calculé à partir des sommets d'étiquette c. Notons que, si dans l'approche décrite dans la section 2.1, l'étiquette propre du sommet x_i n'intervient pas dans le calcul de n_{ic}, elle est ici ré-introduite de manière naturelle par le biais de la prise en compte de $\phi(x_i)$ dans le calcul de $\tilde{f}^\beta_{ic}$ lorsque

$c_i = c$. Le paramètre β permet de régler l'importance que l'on souhaite donner aux étiquettes des voisins directs dans le calcul de la distribution : les petites valeurs de β privilégieront plus les voisins directs.

3 Exemples

Le but de cette section est de présenter l'approche sur un jeu de données très simple. Pour cela, nous avons choisi de l'appliquer à un petit graphe dont les sommets sont des livres de politique américaine, tous publiés autour de l'élection présidentielle américaine de 2004. Les sommets de ce graphe sont les livres, étiquetés selon leur orientation politique (« conservateur », 49 livres, « liberal », 43 livres ou « neutral », 13 livres). Les arêtes du graphe correspondent à des co-occurrences fréquentes d'achats (achats par le même acheteur) entre les deux livres sur le vendeur en ligne Amazon.com. Le graphe a été extrait par Valdis Krebs et peut être récupéré sur son site web [4].

Les résultats des deux approches sont donnés dans, respectivement, les figures 1 et 2. Dans les deux cas, l'interprétation est relativement similaire : le premier axe représente plus de 80% de l'inertie et toute l'inertie est incluse sur les deux premiers axes [5].

Le 1$^{\text{er}}$ axe du graphique des variables oppose l'étiquette « conservateur » et l'étiquette « liberal » alors que le second axe est lié à la variable « neutral ». On constate que les livres sont regroupés en fonction de leur orientation politique ce qui montre que les livres d'une même orientation politique ont des distributions des étiquettes voisines semblables (en fait, les co-occurrences d'achat se font très majoritairement à l'intérieur d'une même famille politique). Parmi les livres ayant une composante forte sur le premier axe, on constate un défaut dans la représentation des livres qui ont pratiquement la même distribution d'étiquettes voisines et qui sont donc superposés. À l'inverse, un certain nombre de livres ont une position plus nuancée : si ils appartiennent à une certaine catégorie politique leurs coordonnées factorielles sont proches de celles d'une autre catégorie. Par exemple, dans la figure 1 "Rise of the Vulcans" a été étiqueté « conservateur » mais ses coordonnées factorielles ne sont pas trop éloignées de celles d'autres livres étiquetés « liberal ». Enfin, le livre "Sleeping With the Devil" est quant à lui représenté au centre du graphique ce qui montre que les personnes ayant acheté ce livre étaient ouverts à tous les ouvrages quel que soit le type de l'orientation politique.

Si les deux approches donnent des résultats assez similaires, elles présentent aussi des différences : en particulier, la représentation du graphe dans le cas de l'utilisation du noyau (figure 2) regroupe de manière plus forte les individus selon leur appartenance politique : c'est l'effet de la prise en compte de la valeur même de l'étiquette du sommet dans le

4. `http://www-personal.umich.edu/~mejn/netdata/polbooks.zip`

5. C'est une conséquence directe du fait qu'il y a ait 3 étiquettes, une étiquette unique par sommet que les lignes ait été normalisées par le nombre de sommets : toutes les lignes somment donc à 1 et il y a un facteur trivial dans l'ACP.

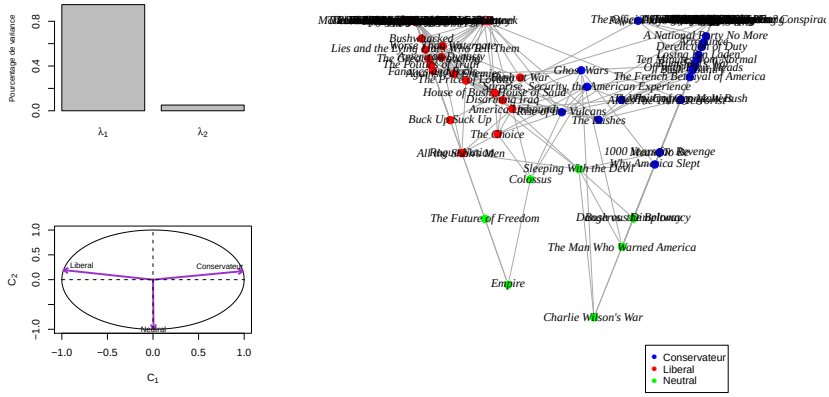

FIGURE 1 – Graphiques issus de l'ACP de la distribution des étiquettes voisines

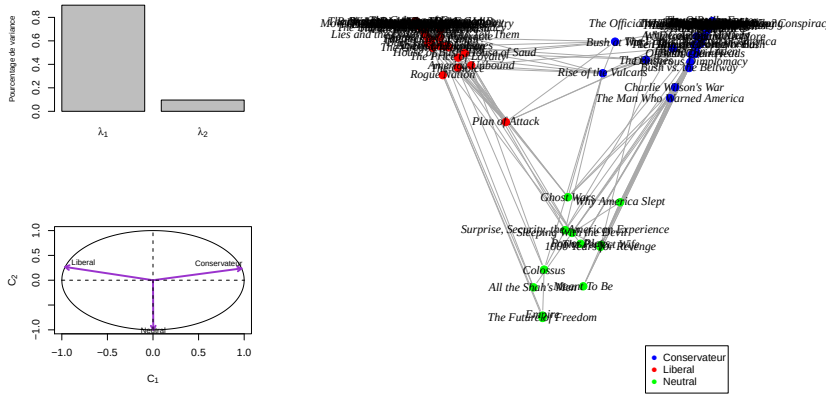

FIGURE 2 – Graphiques issus de l'ACP de la diffusion par noyau ($\beta = 0,1$) des étiquettes

calcul de la distribution. Un livre ressort du graphique : "Plan of Attack". Ce dernier appartient à la classe « liberal » et compte trois voisins de type « conservateur », dix de type « liberal » et aucun de type « neutral ». Pourtant, ses coordonnées factorielles semblent indiquer que la distribution de ce livre est au carrefour des trois types d'orientation politique. Ceci pourrait s'expliquer par le fait que des livres qui lui sont connectés soient eux-mêmes connectés à des livres de type « neutral » et « conservateur » et que la méthode de diffusion d'étiquettes par noyau ait ainsi tenu compte pas seulement des voisins directs, mais aussi des voisins des voisins.

4 Conclusion

Cette proposition de communication présente deux méthodes permettant d'obtenir une représentation du graphe qui privilégie l'interprétabilité selon un critère d'étiquetage des sommets sur la pure esthétique de représentation. Ces propositions sont en cours de test sur des graphes de plus grande taille, contenant plus d'étiquettes et avec éventuellement plus d'une étiquette par sommet. Par ailleurs, les perspectives d'amélioration de ces approches sont multiples : tout d'abord, la question du choix du paramètre β pourrait être abordée, soit par un critère externe (lisibilité, ...) ou par un choix guidé par l'utilisateur, selon l'importance qu'il souhaite donner aux voisins directs du sommet. Quelques itérations d'un algorithme de forces pourrait également être appliquées à la projection du graphe sur le premier plan factoriel : sans nuire à l'interprétabilité de la représentation, cela pourrait permettre d'en améliorer sa lisibilité.

Références

[1] D. Auber. Tulip : A huge graph visualisation framework. In P. Mutzel and M. Jünger, editors, *Graph Drawing Softwares*, Mathematics and Visualization, pages 105–126. Springer-Verlag, 2003.

[2] M. Bastian, S. Heymann, and M. Jacomy. Gephi : an open source software for exploring and manipulating networks. In E. et al. Adar, editor, *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, pages 361–362. Menlo Park : AAAI Press, 2009, 2011.

[3] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related laplacian methods for social network analysis. *Neurocomputing*, 71(7-9) :1257–1273, 2008.

[4] R. Bourqui, D. Auber, and P. Mary. How to draw clustered weighted graphs using a multilevel force-directed graph drawing algorithm. In *Proceedings of the 11th International Conference Information Visualization, 2007. IV'07.*, pages 757–764, July 2007.

[5] C. Bouveyron, H. Chipman, and E. Côme. Supervised classification and visualization of social networks based on a probabilistic latent space model. In *Proceedings of 7th International Workshop on Mining and Learning with Graphs*, Leuven, Belgium, 2009.

[6] P. Eades and Q.W. Feng. Multilevel visualization of clustered graphs. In Stephen C. North, editor, *Proceedings of International Conference on Graph Drawing, Symposium on Graph Drawing*, volume 1190 of *Lecture Notes in Computer Science*, pages 101–112, Berkeley, California, USA, September 18-20 1996. Springer.

[7] P. Eades and M.L. Huang. Navigating clustered graphs using force-directed methods. *Journal of Graph Algorithms and Applications*, 4(3) :157–181, 2000.

[8] T. Fruchterman and B. Reingold. Graph drawing by force-directed placement. *Software-Practice and Experience*, 21 :1129–1164, 1991.

[9] R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[10] B. Schölkopf, K. Tsuda, and J.P. Vert. *Kernel methods in computational biology.* MIT Press, London, 2004.

[11] Q.D. Truong, T. Dkaki, and P.J. Charrel. An energy model for the drawing of clustered graphs. In *Proceedings of Vème colloque international VSST*, Marrakech, Maroc, 21-25 octobre 2007.