Analyse de graphes de données textuelles et règles d'association

Bangaly Kaba[*]**, Eric SanJuan**[**]

[*]*LIMOS, Université Blaise Pascal Clermont 2, France*
kaba@isima.fr
[**]*LIA & IUT STID, Université d'Avignon, France*
eric.sanjuan@univ-avignon.fr

RÉSUMÉ. *Les matrices de données textuelles sont par nature très creuses et il est courant de préférer les représenter sous forme de graphes. Dans ce formalisme, la classification à lien simple ou ses variantes consistent à sélectionner un sous-ensemble d'arêtes (sous-graphe) et à calculer les composantes connexes de ce dernier. Le concept d'atome de graphe permet de désarticuler une composante connexe en une famille de sous graphes non-disjoints dont les intersections correspondent à des cliques maximales. Nous présentons ici des résultats expérimentaux sur une large variété de données textuelles qui montrent les propriétés de ces atomes vis-à-vis des ensembles d'items fréquents et des règles d'association. La désarticulation des graphes en atomes induit alors une nouvelle méthode de classification en classes non disjointes compatible avec les règles d'association.*

MOTS-CLÉS : *Algorithmes de graphes, fouille de données textuelles, classification non supervisée, désarticulation de graphes, règles d'association*

1. Introduction

Depuis quelques années, de nombreuses méthodes de décomposition de graphes sont proposées. Il existe les méthodes basées sur les calculs de coupe minimale pour partitionner de façon récursive un graphe valué en composantes tel que celle utilisée par Shamir et al. ([RSS00]). Voy et al. ([VSP06]) explorent par contre toutes les cliques maximales d'un graphe. Seno et al. ([STT04]) définissent la notion de sous graphes p-quasi complets. Tous ces travaux récents partagent la recherche des parties fortement connectées d'un graphe et les résultats montrent l'importance de ces parties. Cependant, l'un des problèmes rencontrés reste le grand nombre de cliques ou de p-cliques qu'il peut y avoir dans un graphe. Ainsi, ces méthodes sont coûteuses et demandent des heuristiques pour assurer un bon choix.

En suivant Tarjan [TA85], Berry [ABER98] et les travaux de thèse en [KABA08], nous proposons d'utiliser les séparateurs minimaux complets pour décomposer un graphe. Les groupes de sommets définis par la décomposition appelés atomes ne sont pas disjoints, les séparateurs minimaux complets étant recopiés dans le but de préserver la structure du graphe. Les séparateurs minimaux complets sont définis de façon unique et il en existe par définition moins que de sommets. Les algorithmes de triangulation permettent de les calculer de façon efficace et la décomposition qui en résulte est unique : pour un graphe donné, les atomes sont les mêmes et indépendants de l'algorithme qui les calcule. Par contre un graphe quelconque, même très grand, peut n'avoir qu'un seul atome, lui-même, c'est le cas par exemple de tout cycle sans corde. Après avoir défini exactement la notion de graphes d'atomes, nous montrons ici des exemples de graphes de données textuelles pour lesquels non seulement les atomes existent en grand nombre, mais de plus s'avèrent être stables vis-à-vis des règles d'association.

2. Graphe des atomes

Pour caractériser cet objet nous avons besoin au préalable de quelques définitions générales sur les graphes. Un **graphe non orienté** est défini par un ensemble fini de sommets $V = \{v_i\}$ et un ensemble fini d'arêtes $E = \{e_k\}$. Chaque arête est caractérisée par une paire $\{v_i, v_j\}$ de sommets, appelés extrémités. On note $G = (V, E)$. Deux sommets x et y sont dits adjacents lorsqu'ils sont reliés par une arête, x voit y. L'arête est dite incidente aux deux sommets. Un graphe est dit **complet** si tous ses sommets sont deux à deux adjacents. On dit que $G' = (V' E')$ est un sous graphe de $G = (V, E)$ si $V' \subseteq V$ et $E' \subseteq E$. Une **clique** est un sous-graphe complet. Dans un graphe, une chaîne est constituée d'une suite de sommets adjacents. Un cycle est une chaîne simple fermée d'un graphe non orienté. Une corde dans un cycle est une arête entre deux sommets non consécutifs. Un graphe $G = (V, E)$ est dit **connexe** lorsque pour tout $(x, y) \in V^2$, il existe une chaîne les reliant x et y. Une composante connexe d'un graphe est un sous graphe connexe non vide, maximal au sens du nombre de sommets.

Soient donc $G(V, E)$ un graphe connexe et x, y deux sommets non adjacents. Un ensemble de sommets S est un xy**-séparateur** si la suppression des sommets de S place x et y dans deux composantes connexes différentes du graphe $G[V - S]$. S est un xy-séparateur minimal s'il n'existe pas de xy-séparateur S' proprement inclus dans S. S est un **séparateur minimal** s'il existe x et y pour lesquels S est un xy-séparateur minimal. Nous avons alors les propriétés suivantes pour tout $S \subset V$ et $x, y \in V - S$. Si S est un séparateur d'un graphe $G = (V, E)$ et et C une composante connexe de $G = (V, E)$, alors C est dit **composante pleine** pour S quand $N(C) = S$. En fait, si $G = (V, E)$ est un graphe alors un séparateur $S \in V$ de G est un séparateur minimal si et seulement si $G(V - S)$ admet au moins deux composantes connexes pleines. De plus si C_1 et C_2 sont des composantes connexes pleines induites par S alors tout sommet de S a un voisin dans C_1 et un voisin dans C_2.

Un séparateur minimal S est appelé **séparateur minimal complet** si le sous graphe induit par S est une clique. On appelle **décomposition par séparateurs minimaux complets** d'un graphe G les sous-graphes obtenus en répétant l'étape de décomposition précédente sur chacun des sous-graphes engendrés, jusqu'à ce qu'aucun de ces sous-graphes n'admette de séparateur minimal complet.

Définition 2.1 *Nous appellerons* **atome** *de $G = (V, E)$ un sous-ensemble de sommets A de V qui induit un sous-graphe $G(A)$ connexe, sans séparateur minimal complet, et maximal pour ces deux propriétés. Pour un atome A de G on dira aussi que $G(A)$ est un atome.*

La décomposition par séparateurs minimaux complets a la propriété d'être cohérente, en ce sens que chaque séparateur minimal complet choisi dans l'un des sous-graphes obtenus en cours de décomposition est aussi un séparateur minimal complet du graphe de départ quelque soit l'ordre dans lequel on procède à la décomposition. Berry dans sa thèse [ABER98] a démontré qu'une décomposition par séparateurs minimaux complets était unique (voir aussi [LE93]). Apparu dans la littérature comme un résultat de la décomposition d'un graphe triangulé, le graphe des atomes n'a pas constitué jusqu'ici l'objet d'une étude systématique, encore moins lorsqu'il résulte d'une décomposition d'un graphe non triangulé. Pour nous par contre, le graphe des atomes, structure unique sous-jacente à tout graphe, constitue un outil d'analyse de graphes réels issus en particulier de matrices creuses de corrélations, lorsque l'on se donne un seuil en dessous duquel la corrélation est ignorée.

Définition 2.2 *Le graphe des atomes que nous notons $G_{At} = (V_{At}, E_{At})$ est défini comme suit : Les sommets de G_{At} sont les atomes obtenus après la décomposition du graphe G. Une arête $e = (w_1, w_2)$ est définie entre deux atomes A_1 et A_2 s'il existe un séparateur minimal complet S dans G qui sépare l'atome A_1 de l'atome A_2.*

G_{At} a donc autant de sommets que le graphe d'intersection des atomes mais moins d'arêtes. Il y a une relation forte entre séparateurs minimaux complets et triangulation de graphes. Un graphe $G = (V, E)$ est triangulé s'il ne contient pas de cycle sans corde de longueur supérieure à trois. Il s'ensuit que les seuls séparateurs minimaux d'un graphe triangulé sont complets. On appelle **triangulation minimale** de G tout graphe triangulé $H = (V, E + F)$, tel que pour toute partie propre F' de F, le graphe $H' = (V, E + F')$ n'est pas triangulé. Il se trouve que [PS95] un séparateur minimal d'un graphe connexe G est un séparateur minimal complet si et seulement si il est un séparateur

minimal de toute triangulation minimale de G. Il est alors possible de calculer les séparateurs minimaux d'un graphe en $O(|V||E|)$ en calculer une triangulation quelconque H de G, puis calculer les séparateurs minimaux de H et terminer en testant parmi les séparateurs minimaux de H ceux qui étaient déjà complets dans G.

3. Applications à des graphes de données textuelles

Nous montrons comment la décomposition de graphes permet de calculer et de mettre en évidence une famille de clusters non disjoints particulièrement stables vis à vis des règles d'association. Nous évaluons notre approche en utilisant des corpus de notices extraits du WebOfScience. Le premier porte sur le texte mining et le second sur le terrorisme. Pour le premier nous utilisons directement les listes de mots clefs fournies par le WebOfScience. Pour le second nous utilisons une extraction de termes produite par le système TermWatch [SI06].

Nous considérons d'abord le corpus "datamining" utilisé en [PS07] contenant $3,671$ enregistrements extraits de la base de données SCI (Science Citation Index) accessible via le WebOfScience, traitant du datamining et du textmining sur la période 2000-2006 indexée par un ensemble de 8040 mots clés. La moyenne des mots clés par enregistrement est de 5. Le nombre de mots clés d'une fréquence supérieure à 1 est de $1,524$ et indexent $2,615$ enregistrements. Chaque notice est représentée par l'ensemble de ses mots clefs. Cela constitue un hypergraphe H de $3,171$ hyperarêtes (documents indexés par au moins un mot clé) sur 3.671 hyper-sommets (les mots clés). Cet hypergraphe conduit à 7.082 règles d'association ayant un support supérieur à $0,1\%$ et ayant une confiance supérieure à 80%. Toutes les règles d'association sont de la forme $k_1, ..., k_n \rightarrow c$ avec $1 < n \leq 5$ et signifie que plus de 80% des arêtes de H contenant $k_1, ..., k_n$, contiennent aussi c, $k_1, ..., k_n, c$ étant tous de mots clefs différents.

De l'hypergraphe dual de H (les mots-clefs sont représentés par les ensembles des notices qu'ils indexent), nous dérivons le graphe d'association G_1 des mots clefs apparaissant ensemble dans au moins deux notices. Ce graphe a 645 sommets, 1057 arêtes et est divisé en une principale composante connexe avec 413 sommets et 15 petites composantes connexes avec moins de 16 sommets. Nous avons basé notre étude sur la principale composante que nous avons assimilé au graphe entier G_1. G_1 est un graphe petit monde (SWG) avec un coefficient de clustering moyen de 0.47. Cette valeur est loin de la valeur attendue pour un graphe aléatoire ayant le même degré moyen qui est la moyenne sur l'ensemble des arêtes (4.43/2335). Comme dans les graphes aléatoires, le chemin moyen est faible : 2.321.

Le graphe des atomes $G_1(At)$ a été calculé en moins d'une minute sur un PC standard. Il a 404 atomes dont un central contenant 298 sommets. Les autres 403 atomes ont moins de 13 sommets. Donc G_1 est clairement divisé en un noyau central et une périphérie. Le résultat remarquable que nous obtenons est que 96% de ces 403 petits atomes sont stables vis à vis des règles d'association.

Les règles d'association sont calculées à partir des ensembles fréquents d'items (EFI) $k_1, ..., k_n$ (sous-ensembles de mots clefs apparaissant ensemble dans plus de $0,1\%$ de notices). Lorsque le seuil utilisé pour construire le graphe d'association G_1 correspond ou est proche du support minimal d'un EFI, alors chaque arête correspond à un EFI de taille 2 et tout EFI de taille n génère une clique dans G_1. Il y a donc une relation directe entre EFI et cliques. En particulier, on peut s'attendre à ce que la majorité des séparateurs minimaux corresponde à des EFI et aient des propriété de stabilité vis à vis des règles d'association. Ce que nous constatons expérimentalement et qui est plus surprenant, est que les atomes de ces graphes quelconques qui ne se réduisent pas à des cliques soient aussi stables vis à vis de ces règles.

Nous avons aussi appliqué cette méthode de décomposition au corpus issu de Chaomei et al.[CZZV07] extrait de la même base de données bibliographiques de SCI (Science Citation Index). Ce corpus a été choisi afin d'étudier l'évolution structurelle des réseaux de recherche sur le terrorisme. Nous avons utilisé le systeème TermWatch [1] pour extraire et sélectionner 57.855 syntagmes nominaux à partir de 3.366 résumés bibliographiques. Ces syntagmes nominaux ont été regroupés en 3.293 composantes de termes variants ayant au moins 2 syntagmes nominaux presques synonymes. La taille maximale de ces composantes de termes est de 30. 8.357 termes isolés mais

1. http ://index.termwatch.es

apparaissant dans au moins deux notices distinctes ont été ajoutés à l'ensemble des composantes de termes. A cette liste de termes nous avons encore ajouté le nom de tous les auteurs des articles. Nous avons alors considéré l'hypergraphe induit par la réunion des relations document-terme et document-auteur.

Le graphe d'associations auteurs-termes qui en découle a $16,258$ arêtes. Sa principale composante a $9,324$ arêtes et $1,070$ sommets. La décomposition de cette composante nous donne 489 atomes. L'atome central a $2,070$ arêtes et 307 sommets. Comme dans le cas du graphe des mots clés extraits du corpus sur le datamining, nous obtenons après la décomposition un atome central et une périphérie constituée de multiples petits atomes de moins de 20 éléments chacun. Bien que ce graphe d'association ait été obtenu par une toute autre méthode d'indexation, on retrouve une proportion similaire d'atomes 95% qui sont stables vis à vis des règles d'association.

4. Conclusion

Cette étude est une première tentative d'application de la décomposition de graphes à la cartographie des domaines de connaissance. L'avantage de la décomposition en atomes est qu'elle est unique. Elle est basée sur la seule structure du graphe. Son principal inconvénient est que les petits atomes n'existent pas toujours dans un graphe. Les deux expérimentations présentées ici tendent à montrer la décomposition en atomes s'adapte bien au corpus de thématiques bibliographiques puisque on trouve un large ensemble de petits atomes qui s'avèrent être stables à plus de 95% vis à vis des règles d'association de confiance élevée (80%). D'autres expérimentations sur des corpus plus artificiels de thématiques bibliographiques traitant de : Information Retrieval, genomics ou Organic Chemistry ont confirmé ces résultats.

5. Bibliographie

[ABER98] A.Berry. Désarticulation d'un graphe. Thèse de doctorat, LIRMM, Montpellier, décembre 1998.

[CZZV07] C. Chen, W. Zhu, B. Tomaszewski, A. MacEachren (2007). Tracing conceptual and geospatial diffusion of knowledge. HCI International 2007. Beijing, China. July 22-27, 2007. LNCS, 4564. pp. 265-274.

[KABA08] B.Kaba. Décomposition de graphes comme outil de clustering et de visualisation en fouille de données. Thèse de doctorat, LIMOS, Clermont Ferrand, novembre 2008.

[LE93] H. G.Leimer. Optimal Decomposition by Clique separators, Discrete Mathematics archive, 113(1-3), pp 99-123, 1993.

[PS07] X.Polanco,E.SanJuan. Hypergraph modelling and graph clustering process applied to co-word analysis. In : 11th biennial International Conference on Scientometrics and Informetrics, 2007.

[PS95] A.Parra and P. Scheffler. How to use the minimal separators of a graph for its chordal triangulation. Proc.22nd International colloquium on automata, Languages and Programming (ICALP'95) ; Lecture Notes in Computer Science, 944 : 123-134,1995.

[STT04] S.Seno, R.Teramoto, Y.Takenaka and H.Matsuda. A Method for Clustering Gene Expression Data Based on Graph Structure, Genome Informatics 2004, 15(2), pp 151-160.

[TA85] R.E.Tarjan.Decomposition by clique separators. Discrete Math :55 : 221-232, 1985.

[RSS00] R.Sharan and R.Shamir. CLICK : A Clustering Algorithm with Applications to Gene Expression Analysis, Proc. ISMB'00, AAAI Press, Menlo Park (CA, USA), pp 307-316, 2000.

[SI06] E.SanJuan, F.Ibekwe-SanJuan. Text mining without document context. Information Processing and Management, 42 1532-1552, 2006.

[VSP06] B.H.Voy, J. A.Scharff, A. D.Perkins, A.M.Saxton, B. Borate, E.J. Chesler, L.K.Branstetter and M.A.Langston. Extracting Gene Networks for Low-Dose Radiation Using Graph Theoretical Algorithms, PLOS Computational Biology, 2006.