Approche par motifs pour l'analyse de données multi-résolution

Pierre-Nicolas Mougel, Fréderic Flouvat, Nazha Selmaoui-Folcher

Université de Nouvelle Calédonie
{pierre-nicolas.mougel,frederic.flouvat,nazha.selmaoui}@univ-nc.nc

Résumé. Dans cet article nous nous intéressons aux approches pour l'analyse de graphes pouvant évoluer dans le temps et tel qu'un sommet à un temps donné peut correspondre à plusieurs sommets au temps suivant et où les sommets sont associés à un ensemble d'attributs catégoriels. Dans ce type de données, nous proposons une nouvelle classe de motifs basée sur des contraintes permettant de décrire l'évolution de structures homogènes. Ce type d'approche est particulièrement adaptée pour l'analyse d'images multi-résolution sans perte d'information. Nous présentons un résultat qualitatif dans ce domaine.

1 Introduction

Ces dernières années, plusieurs travaux se sont intéressés à la fouille de graphes pour modéliser des phénomènes réels. Récemment, on trouve des travaux sur les graphes dynamiques (Pei et al. (2005); Borgwardt et al. (2006); Bilgin et Yener (2006); Robardet (2009); Rossi et al. (2013)) qui ont permis la modélisation de l'évolution d'objets au cours du temps ainsi que leurs relations. Dans ce type de méthodes, on analyse essentiellement les évolutions structurelles. Par exemple dans (Robardet (2009)), les auteurs étudient l'évolution de sous-graphes en considérant des opérations tels que le découpage, le regroupement, la suppression ou la création de quasi-cliques. De plus en plus de travaux de fouille de graphes s'orientent vers les graphes attribués (e.g., Moser et al. (2009); Silva et al. (2012); Mougel et al. (2012a)) qui sont des graphes dynamiques dont les sommets sont décrits par des attributs. Ces derniers ont permis l'étude de plusieurs domaines d'applications notamment les réseaux biologiques (Fukuzaki et al. (2010); Mougel et al. (2012b)). Mais à notre connaissance peu de travaux traitent la fouille de graphe dynamique dont le nombre de sommets évolue dans le temps.

Dans cet article, nous proposons une approche permettant d'étudier les graphes attribués dynamiques et dont le nombre de sommets varient dans le temps. Dans ce type d'approche on considère une séquence temporelle de graphes attribués dont le nombre de sommets peut varier dans le temps. Un domaine d'application pour lequel une telle approche est particulièrement intéressante est l'analyse d'objets dans une séquence d'images satellites à différentes résolutions. En effet, il est très difficile de constituer une longue séquence temporelle d'images à la même résolution notamment des images à très haute résolution qui sont coûteuses en acquisition.

Pour traiter ce problème, nous proposons un nouveau modèle de données basé sur une séquence de graphes attribuées [1] nommé *Graph Attribué Multi-résolution* (GAM). Un GAM est une séquence de graphes tels que les sommets entre deux temps consécutifs peuvent être reliés par des *arêtes temporelles*, les arêtes reliant des sommets à un temps donnée sont appelées *arêtes structurelles*. A partir de ce modèle, nous proposons de rechercher des motifs nommés *Graphes Multi-résolutions Homogènes* (GMH). Un GMH est formé par une collection de sous-graphes connexes et homogènes nommés Graphe Connexe Homogène (GCH). La propriété d'homogénéité assure que le graphe est formé par des sommets partageant des propriétés similaires. La contrainte structurelle de connexité permet de trouver des graphes sans forme définie a priori. Dans le cadre de l'analyse d'images, ces deux conditions sont basées sur les hypothèses suivantes : (1) un GCH représente un segment de l'image correspondant à un objet réel, (2) un objet segmenté dans une image est généralement décrit par des pixels ayant des propriétés similaires (e.g., rouge, vert, bleu) et (3) la forme des objets segmentés n'est pas connue à l'avance mais reste connexe.

2 Contexte et définition des motifs

Définition 2.1 (GAM) *Un GAM est un tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E}_s, \mathcal{E}_t, \mathcal{A})$ avec :*
- $\mathcal{V} = \langle \mathcal{V}_1, \ldots, \mathcal{V}_n \rangle$ *une séquence d'ensembles de sommets tel que* $\forall_{i,j|i\neq j} \mathcal{V}_i \cap \mathcal{V}_j = \emptyset$ *;*
- $\mathcal{E}_s \subseteq \cup_{t=1}^{n} \{\{v_1, v_2\} \mid v_1 \in V_t \wedge v_2 \in V_t\}$ *un ensemble d'arêtes structurelles ;*
- $\mathcal{E}_t \subseteq \cup_{t=1}^{n-1} (V_t \times V_{t+1})$ *un ensemble d'arêtes dirigées temporelles ;*
- $\mathcal{A} = \langle \mathcal{A}_1, \ldots, \mathcal{A}_n \rangle$ *une séquence d'ensembles d'attributs catégorielles tel que* $\forall \mathcal{A}_i \in \mathcal{A}, \bigcap_{A \in \mathcal{A}_i} Dom(A) = \emptyset$.

La fonction $AtbV$ associe un ensemble de valeurs d'attributs à chaque sommet.

Le nombre de pas de temps dans $\mathcal{G}$ (i.e., $|\mathcal{V}|$) est noté $\mathcal{T}_\mathcal{G}$. L'ensemble de tous les sommets de $\mathcal{G}$ est noté $\mathcal{V}_\mathcal{G}$, i.e., $\mathcal{V}_\mathcal{G} = \cup_i \mathcal{V}_i$. Le domaine de tous les attributs de la collection $\mathcal{A}_t \in \mathcal{A}$ est noté $\mathcal{D}_t$, i.e., $\mathcal{D}_t = \cup_{A \in \mathcal{A}_t} Dom(A)$. Le sous-graphe de $\mathcal{G}$ induit par l'ensemble de sommets V est noté $\mathcal{G}[V]$ et correspond au couple de sommets et d'arêtes structurelles $(V, \{\{v_1, v_2\} \in \mathcal{E}_s \mid v_1 \in V \wedge v_2 \in V\})$. Nous définissons une fonction Hmg associant l'ensemble des valeurs d'attributs partagés par un ensemble de sommets et la fonction inverse $Vert$. Etant donné un ensemble de sommets V et un ensemble de valeurs d'attributs A, $Hmg(V) = \bigcap_{v \in V} AtbV(v)$ et $Vert(A) = \{v \in \mathcal{V}_\mathcal{G} \mid AtbV(v) \subseteq A\}$.

Nous proposons une nouvelle classe de motifs nommée *Graphes Homogènes Multi résolutions* (GMH). Un GMH est formé par une collection de sous-graphes connexes et homogènes nommés Graphe Connexe Homogène (GCH).

Définition 2.2 (Graphe Connexe Homogène) *Soit $\mathcal{G}$ un GAM et $\mathfrak{h} \in [0,1]$, $\mathfrak{s} \in [0,1]$ deux seuils définis par l'utilisateur. Un ensemble de sommets H tel que $H \subseteq \mathcal{V}_t \in \mathcal{V}$ est un Graphe Connexe Homogène (GCH) si et seulement si (1) tous les sommets de H partagent au moins $\mathfrak{h} \cdot |A_t|$ valeurs d'attributs en commun avec les autres sommets de H, i.e., $|Hmg(H)| >= \mathfrak{h} \cdot |A_t|$; (3) il existe un chemin passant par des arêtes structurelles entre chaque paires de sommets de H ; (2) la collection contient au moins $|H| / |\mathcal{V}_t| \geq \mathfrak{s}$ sommets ; et (4) il n'existe pas de sommets $v \in \mathcal{V}_\mathcal{G}$ n'appartenant pas à H tel que $H \cup \{v\}$ vérifie les conditions précédentes.*

1. Dans cet article, nous considérons des attributs catégoriels.

La première condition assure l'homogénéité de l'ensemble des sommets par rapport aux attributs. La condition (2) s'intéresse à la structure du graphe. Dans le contexte de l'analyse d'images, la contrainte de connexité permet de trouver des régions de forme arbitraire. La condition (3) permet de filtrer les petits graphes qui peuvent ne pas être intéressant pour les experts. Enfin, la condition (4) assure la maximalité des GCH par rapport à l'inclusion.

Un GCH décrit une partie d'un GAM pour un temps donné. Afin d'étudier l'évolution des GCH nous proposons de grouper les GCH connectés entre des pas de temps consécutifs.

Définition 2.3 (Connectivité temporelle) *Soit $\mathcal{G} = (\mathcal{V}, \mathcal{E}_s, \mathcal{E}_t, \mathcal{A})$ un GAM, $\tau \in [0,1]$ un seuil d'extension temporelle et $H_1 \subseteq \mathcal{V}_i$, $H_2 \subseteq \mathcal{V}_{i+1}$ deux GCH. L'ensemble H_1 est temporellement connecté à H_2 si et seulement si $\frac{|\phi(H_1) \cap H_2|}{|\phi(H_1)|} \geq \tau$, avec $\phi(H)$ l'ensemble des sommets connectés par des arêtes temporelles par des sommets de H. I.e., $\phi(H) = \{v \in \mathcal{V}_{i+1} \mid v' \in H \wedge (v', v) \in \mathcal{E}_t\}$. Une séquence $S = \langle H_1, \ldots, H_n \rangle$ est une* chaîne connectée temporellement *si et seulement si $\forall i \in \{1, \ldots, n-1\}$, H_i est connecté temporellement à H_{i+1}.*

Définition 2.4 (Graphe Multi-résolution Homogènes) *Une collection P de GCH est un Graphe Multi-résolution Homogène (GMH) ssi. (1) pour chaque paire $H_1, H_2 \in P$, il existe un GCH $H' \in P$ et deux chaînes connectées temporellement S_1 et S_2 formées uniquement par des GCH de P tel que S_1 contient H_1 et H' et S_2 contient H_2 et H' et (2) pour tout GCH H, si H n'appartient pas à P alors il n'existe pas de GCH de P connecté temporellement à H.*

La condition (1) assure que chaque paire de GCH d'un GMH est connectée de manière transitive par la relation de connectivité temporelle. La condition (2) assure la maximalité du motif dans le sens qu'il n'existe pas de GCH n'appartenant pas au motif et pouvant lui être ajouté sans violer la condition (1).

3 Méthode d'extraction

Stratégie d'énumération et d'élagage des GCH. Nous présentons dans un premier temps une stratégie naïve d'énumération des GCH. Ensuite nous proposons plusieurs techniques permettant de réduire l'espace de recherche.

Lemme 3.1 *Soit $\mathcal{G}$ un GAM et $\mathfrak{h}, \mathfrak{s} \in [0,1]$ deux seuils. Un ensemble de sommets $H \subseteq \mathcal{V}_t \in \mathcal{V}$ est un GCH si et seulement si il existe un ensemble de valeurs d'attributs $X \subseteq \cup_{A \in \mathcal{A}_t} Dom(A)$ tel que (1) $|X| \,/\, |\mathcal{D}_t| \geq \mathfrak{h}$; (2) H est une composante connexe dans le sous graphe $\mathcal{G}[Vert(X)]$; et (3) $|H| \,/\, |\mathcal{V}_t| \geq \mathfrak{s}$.*

D'après cette propriété pour calculer la collection des GCH à un temps t un algorithme naïf peut énumérer tous les ensembles non vides de valeurs d'attributs $X \subseteq \mathcal{D}_t$. Si $|X| \,/\, |\mathcal{D}_t| \geq \mathfrak{h}$ alors la collection des composantes connexes dans le graphe $\mathcal{G}[Vert(X)]$ satisfaisant la condition de taille minimale est formée uniquement par des GCH. Cependant cette approche nécessite l'énumération de $2^{|\mathcal{D}_t|} - 1$ ensembles de valeurs d'attributs pour chaque temps. Les propriétés suivantes permettent de réduire l'espace de recherche.

Propriété 3.1 *Soit $\mathcal{G}$ un GAM et $V \subseteq \mathcal{V}_t \in \mathcal{V}$ un ensemble de sommets. Uniquement les sommets de V appartenant à une composante connexe de $\mathcal{G}[V]$ ayant au moins $\mathfrak{s} \times |\mathcal{V}_t|$ sommets peuvent former un GCH.*

Propriété 3.2 *Soit $V \subseteq \mathcal{V}_t \in \mathcal{V}$ un ensemble et x une valeur d'attribut partagée par moins de $\mathfrak{s} \times |\mathcal{V}_t|$ sommets de V. Aucun sous-ensemble de $V \cap Vert(\{x\})$ ne peut former un GCH.*

Description des algorithmes. L'algorithme principal se déroule en deux parties. La première partie calcule la collection des GCH notée $\mathcal{H}_t$ pour chaque temps $t \in \{1, \ldots, \mathcal{T}_\mathcal{G}\}$ en appelant l'algorithme 1. La deuxième partie de cet algorithme réalise l'extension temporelle. Pour chaque temps $t \in \{1, \ldots, \mathcal{T}_\mathcal{G} - 1\}$, les GCH de $\mathcal{H}_t$ qui n'ont pas été précédemment traités sont utilisés pour construire un nouveau GMH noté P. Si un GCH H n'a pas été précédemment énuméré, la fonction $\varphi(H)$ renvoie $\emptyset$, sinon le motif auquel il appartient. Enfin, P est étendu en utilisant les GCH au temps consécutif en appelant l'algorithme 2 .

L'algorithme 1 calcule la collection des GCH pour un temps donné. Le test effectué à la ligne 1 vérifie si l'ensemble de sommets actuellement énuméré est homogène. Si c'est le cas, la collection de GCH est mise à jour aux lignes 2 ou 3. Le test de la ligne 2 permet de traiter le cas particulier où tous les sommets à un temps donné sont homogènes. Lors des appels récusrifs, l'ensemble de sommets V est nécessairement connexe donc la collection peut directement être mise à jour (ligne 3). D'après la condition de maximalité, si l'ensemble des sommets est homogène, l'énumération de la branche peut s'arréter, sinon l'énumération continue aux lignes 5 à 10. Le filtrage effectué à la ligne 5 retire de l'ensemble des valeurs attributs énumérées A_{cand} les valeurs d'attributs partagées par tous les sommets de V_{cand}. L'énumération des valeurs d'attributs restantes est effectuée ensuite, ainsi que le calcul des composantes connexes correspondantes (ligne 8, fonction **CC**). Pour chaque composante connexe satisfaisant la condition de taille minimale, l'algorithme est appelé de manière recursive.

Algorithme 1 : `EnumérerGCH`

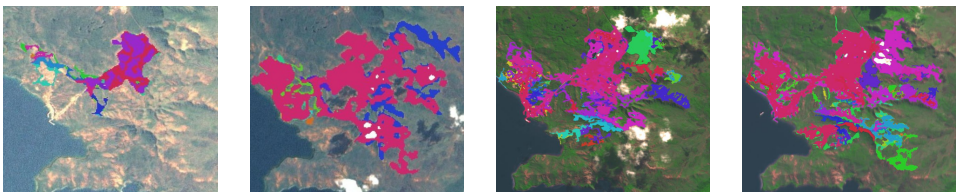
Input : V_{cand}, A_{cand}, $\mathcal{H}$, $premierAppel$
Output : $\mathcal{H}$

```
1  if Hmg(V) ≥ 𝔥 then
2      if premierAppel then H ← H ∪ CC(V)
3      else H ← H ∪ {V}
4  else
5      A_cand ← {a ∈ A_cand | V_cand ⊆ Vert({a})}
6      while A_cand ≠ ∅ do
7          Sélectionner un élément a de A_cand et le retirer
8          for V'_cand ∈ CC(V_cand ∩ Vert({a})) do
9              if |V'_cand| ≥ 𝔰 × |V_t| then
10                 H ← EnumérerGCH(V'_cand, A_cand, H, false)
```

L'algorithme 2 effectue le regroupement temporelle des GCH précédement extraits. étant donné un GCH H, la première ligne de l'algorithme calcule les sommets V_{suiv} connectés à H par des arêtes temporelles. La ligne 2 énumère pour chaque GCH H' du temps consécutif ceux qui sont connectés temporellement à H. Le motif P en cours de construction est ensuite mis à jour avec H' (ligne 3). Si H' appartenait déjà à un motif, les deux motifs sont regroupés (lignes 5 et 6), sinon, un appel récursif à l'algorithme est effectué à partir de H' (ligne 8).

Algorithme 2 : `ExtensionTemporelle`

Input : H, t, P
Output : P
$V_{suiv} \leftarrow \{v_2 \in V_{t+1} \mid (v_1, v_2) \in \mathcal{E}_t \wedge v_1 \in H\}$
for $H' \in \{H_{cand} \in \mathcal{H}_{t+1} \mid \frac{|H_{cand} \cap V_{suiv}|}{|V_{suiv}|} \geq \tau\}$ **do**
  $P \leftarrow P \cup \{H'\}$; // $\varphi(H') = P$
  if $\varphi(H') \neq \emptyset$ **then**
    $\mathcal{P} \leftarrow \mathcal{P} \setminus \varphi(H')$
    $P_{prec} \leftarrow \varphi(H')$; $P \leftarrow P \cup \varphi(H')$; // $\forall x \in P_{prec}, \varphi(x) = P$
  else if $t < \mathcal{T}_{\mathcal{G}}$ **then**
    $P \leftarrow$ `ExtensionTemporelle`$(H', t+1, P)$



(a) t2 (Landsat 7) (b) t3 (Landsat 7) (c) t4 (SPOT 4) (d) t5 (SPOT 5)

FIG. 1 – *Exemple de motif. Les zones coloriées correspondent aux GCH.*

4 Résultats expérimentaux

Le jeu de données a été construit à partir d'images satellites à différentes résolutions (30 mètres et 10 mètres le pixel). Huit images ont été utilisées de la même région. Nous avons utilisés six attributs (discrétisés) dont trois correspondent à la radiomètre rouge, vert et proche infrarouge. Les trois autres propriétés sont des indices calculés à partir des attributs précédents : l'indice de rougeur, l'indice de brillance, et l'indice normalisé de végétation (NDVI). Dans le graphe, un sommet correspond à un pixel et les arêtes structurelles au 4 voisinages d'un pixel (i.e., sans considérer les diagonales). Une arête temporelle connecte deux sommets correspondant aux pixels situés dans la même zone entre des temps consécutifs. Le GAM construit contient 4 graphes avec 10 890 000 de sommets et 4 graphes avec 1 210 000 de sommets.

Dans ce jeu de données, nous recherchons des motifs homogènes sur la moitié des attributs (i.e., $\mathfrak{h} = 0.5$) et ayant une taille relativement petite, correspondant à une superficie d'environ 0.45 km^2 (i.e., $\mathfrak{s} = 2 \times 10^{-6}$). Plus précisément les GCH regroupés partagent au moins 80 % de leurs sommets avec un autre GCH du motif (i.e., $\tau = 0.8$). En utilisant ces seuils, 24 motifs ont été extraits en 5 minutes. Parmi ces motifs, nous présentons sur la figure 1 un motif représentant une zone qui évolue de la même manière correspondant à l'activité minière. De manière intéressante, on peut noter que regrouper les motifs avec la contrainte de connectivité temporelle permet de bien couvrir la zone.

5 Conclusion

Le problème de l'analyse de graphes multi-résolution est considéré et une proposition de modélisation de ce type de données est formalisé. Une nouvelle famille de motifs permettant d'étudier l'évolution de structures homogènes est également proposée. Un algorithme complet permet l'extraction de cette classe de motifs est détaillé, ainsi que plusieurs techniques permettant de réduire l'espace de recherche. Un résultat qualitatif sur une séquence d'image satellites est finalement présenté afin de justifier l'intérêt de l'approche. En terme de travaux futurs, une analyse des résultats en terme de performance reste à réaliser.

Références

Bilgin, C. C. et B. Yener (2006). Dynamic network evolution : Models, clustering, anomaly detection. *IEEE Networks*.

Borgwardt, K., H.-P. Kriegel, et P. Wackersreuther (2006). Pattern mining in frequent dynamic subgraphs. In *IEEE Int. Conf. on Data Mining (ICDM)*, pp. 818–822.

Fukuzaki, M., M. Seki, H. Kashima, et J. Sese (2010). Finding Itemset-Sharing Patterns in a Large Itemset-Associated Graph. In *PAKDD*, pp. 147–159.

Moser, F., R. Colak, A. Rafiey, et M. Ester (2009). Mining Cohesive Patterns from Graphs with Feature Vectors. In *SIAM International Conference on Data Mining (SDM)*, pp. 593–604.

Mougel, P.-N., C. Rigotti, et O. Gandrillon (2012a). Finding Collections of k-Clique Percolated Components in Attributed Graphs. In *Pacific-Asia Conf. on Knowl. Discov. and Data Mining (PAKDD)*, pp. 181–192.

Mougel, P.-N., C. Rigotti, M. Plantevit, et O. Gandrillon (2012b). Finding Maximal Homogeneous Clique Sets. *Knowledge and Information Systems (KAIS)*, 1–30.

Pei, J., D. Jiang, et A. Zhang (2005). On mining cross-graph quasi-cliques. In *Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pp. 228–238.

Robardet, C. (2009). Constraint-based pattern mining in dynamic graphs. In *IEEE Int. Conf. on Data Mining (ICDM)*, pp. 950–955.

Rossi, R. A., B. Gallagher, J. Neville, et K. Henderson (2013). Modeling dynamic behavior in large evolving graphs. In *ACM Int. Conf. on Web Search and Data Mining*, pp. 667–676.

Silva, A., W. J. Meira, et M. J. Zaki (2012). Mining attribute-structure correlated patterns in large attributed graphs. *Proceedings of the VLDB Endowment 5*(5), 466–477.

Summary

In this paper, we study the analysis of evolving graphs such that a vertex at a given timestamp can match several vertices at the consecutive timestamp. Moreover, vertices may be associated to a set of categorical attributes. In such dataset, we propose a new familly of pattern to study the evolution of homogeneous structures. This type of approach is valuable for the analysis of multi-resolution images. We present qualitative results in this domain showing the interest of our approach.