Fouille de graphes et découverte de règles d'association : application à l'analyse d'images de document

Eugen Barbu, Pierre Héroux, Sébastien Adam, Eric Trupin

Fouille de graphes et découverte de règles d'association : application à l'analyse d'images de document

Eugen Barbu*, Pierre Héroux*
Sébastien Adam*, Éric Trupin*

*Laboratoire PSI
CNRS FRE 2645 – Université de Rouen
UFR des Sciences et Techniques
F-76 821 Mont-Saint-Aignan cedex
{Prenom.Nom}@univ-rouen.fr,
http ://www.univ-rouen.fr/psi

Résumé. Cet article présente une méthode permettant la découverte non supervisée de motifs fréquents représentant des symboles graphiques sur des des images de documents. Sur les documents, les symboles sont des entités graphiques porteurs d'information. Les images de document sont représentées par des graphes relationnels attribués. La méthode réalise dans un premier temps la découverte de sous-graphes disjoints fréquents, faisant correspondre pour chacun d'eux un symbole différent. Une recherche des règles d'association entre ces symboles permet d'accéder à une partie des connaissances du domaine. L'objectif à terme est d'utiliser les symboles découverts pour la classification ou la recherche d'images dans un flux hétérogène dans lequel une approche supervisée n'est pas envisageable.

1 Introduction

Sur un document, un symbole est un signe (élément graphique) qui, selon certaines conventions relatives au domaine, encode une unité élémentaire de message. Dans ce contexte, la classification non supervisée de symboles et la recherche des règles d'association entre ces symboles sont utiles d'une part, pour la classification des images de documents, et donc, pour interprétation plus fine du contenu et d'autre part pour la recherche des occurrences d'un symbole particulier dans un ou plusieurs documents.

Nous considérons comme symbole, toute partie de l'image du document apparaissant avec une certaine fréquence. Nous présentons dans un premier temps les méthodes permettant le partitionnement de l'image du document et puis comment s'effectue la recherche de parties fréquentes.

La section **??** présente le contexte et les travaux précédents dans le domaine abordé. La section 3 détaille l'algorithme permettant la recherche de sous-graphes fréquents. La section **??** traite de la découverte des règles d'association entre les symboles. La section **??** illustre l'application de la méthode à travers un exemple. Enfin, la section **??** dresse les conclusions et énonce quelques perspectives permettant de prolonger le travail.

FIG. 1 – Extraits d'image représentant deux symboles appartaissant dans des contextes différents

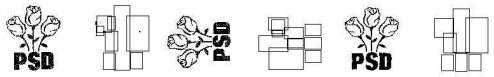

FIG. 2 – Représentations structurelles de plusieurs occurences d'un même symbole

2 Contexte

2.1 Fouille de graphes et analyse d'images de document

« L'objectif principal de la fouille de graphes est de fournir de nouveaux principes et des algorithmes efficaces pour la découverte de sous-structures topologiques incluses dans des données décrites sous forme de graphes »[**?**]. Les premiers systems issus de ce champ de recherche furent SUBDUE [**?**] et GBI [**?**]. Ils s'appuient sur la méthode GLOUTON et peuvent donc aboutir à la découverte de motifs. WARMR est une méthode basée sur l'induction logique programmable permettant la recherche complète des sous-graphes fréquents. Une évolution importante a été l'introduction du concept de sous-graphe fermé. Un sous-graphe est dit fermé s'il ne possède pas de graphe l'incluant avec le même nombre d'occurences dans les données traitées [**?**]. Les techniques de fouille de graphes ont par le passé été appliquées à l'analyse de scènes, au bases de données de composés chimiques et aux Traduire workflows.

Si on donne une représentation des images de documents à base de graphes, les symboles sont représentés par des sous-graphes fermés car même s'ils sont connectées à d'autres parties du document, seules les parties correspondant aux symboles apparaissent fréquement dans différents contextes (cf. Fig. 1).

Les représentations à base de graphe utilisées en analyse d'image sont sensibles aux déformations. Une illsutration est donnée sur la figure 2. Pour construire un graphe à partir d'une image de document, il est nécessaire de définir une procédure pour segmenter l'image en parties (représentées par des nœuds du graphe) et de définir les relations entre ces parties comme les arcs du graphe. Dans le cas général, les nœuds et les arcs sont étiquetés par des réels ou des vecteurs numériques. On obtient alors un graphe relationnel attribué. Lorsque les graphes représentent des workflows ou des composés chimiques, les étiquettes possibles des nœuds sont en nombre fini. En revanche, si la représentation structurelle utilise les composantes connexes de l'images comme nœud du graphe, des images ne présentant que peu de différences peuvent être représentées par des graphes totalement différents.

Les techniques de fouille de graphe appliquées aux images de documents pour la recherche de symboles ne donneront de bons résultats qu'avec une représentation sous

forme de graphes relationnels attribués adaptée et la recherche de sous-graphes disjoints fermés et fréquents.

2.2 Régles d'association et analyse d'image de document

Le problème de la découverte de règles d'association a été introduit par Agrawal dans le domaine de l'analyse des market-basket [**?**]. Une règle d'association exprime une quasi implication. Le sens d'une règle d'association est que, si dans un certain contexte, on trouve un ensemble X d'éléments, alors on trouvera probablement également les éléments de l'ensemble Y également. Une règle d'association approxime une implication. La qualité de cette approximation peut être donnée par différents indices. Parmi ceux-là, le support et la confiance sont communément utilisés, même s'ils ne sont pas les plus riches du point de vue sémantique. Le support est la probabilité de trouver X, la confiance est la probabilité de trouver Y dans le même contexte quand X est présent. La notion de contexte est également désigné par le terme « transaction ». La sémantique des régles découvertes dépend de la façon dont les transactions sont définies. Les règles qu'on souhaite pouvoir extraire sur un document peuvent avoir une des formes suivantes : « La présence du symbole S_1 dans un *paragraphe* implique la probable présence du symbole S_2 dans le même *paragraphe*. », « La présence de symbole S_3 dans un *document* implique la probable présence des symboles S_4, S_5 et S_6 dans ce document. »

La première application de la recherche des règles d'association dans le domaine de l'analyse d'image a été menée par Ordonez [**?**]. Il s'agit de la recherche sans connaissance du domaine de régles d'association entre composantes connexes sur image générée de façon synthétique. À chaque image est associée une transaction contenant les ocmposantes connexes de l'image.

Les règles d'association ont été utilisées dans le contexte de l'analyse d'image de documents, et plus précisément pour l'analyse de la structure physique des documents, au sein du projet WISDOM++ [**?**]. Les règles d'association sont ici utilisées, après traitement des images, pour construire les structures physique et logique du document.

Considérant une transaction T et une règle d'association $R : X \rightarrow Y$, il est possible de savoir si R est confirmé ou pas dans la transaction T. R est confirmé si X n'apparaît pas dans la transaction ou si Y y apparaît. La confirmation de règles dans des transactions peut être utilisée pour construire des méta-règles. Nous indiquons en section 4 en quoi les méta-règles différent des règles simples, et justifions le fait qu'elles apportent de la connaissance à la description.

Pour découvrir les règles d'association entre les symboles trouvé par les techniques de recherche de sous-graphes fréquents, il est nécessaire de définir les transactions de telle sorte qu'elles véhiculent de l'information utile dans le contexte du document. Ces transactions peuvent représenter soient des parties du document, soit le document dans sa totalité. Par exemple, une transaction peut être associée à chaque paragraphe dans un document à dominante textuelle, ou à chaque scale-view échelle d'une carte.

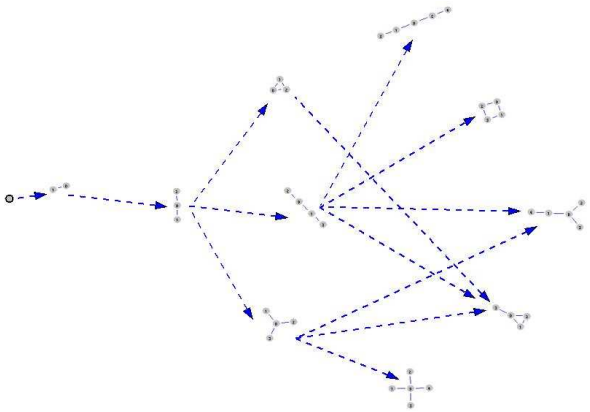

FIG. 3 – Réseau des graphes non-isomorphes

3 Recherche des sous-graphes fermés fréquents

Considérant un graphe composé de n nœuds et de e arcs de représentant l'image d'un document, un symbole dont la représentation est un graphe de n' nœuds et de e' arcs ne peut apparaître à plus de $\min(\frac{e}{e'}, \frac{n}{n'})$ emplacements différents dans le document.

Le seuil de fréquence permettant de considérer un symbole comme fréquent est calculé à partir d'une approximation du nombre maximum de sous-graphes disjoints qu'il est possible de construire à nombre de nœuds et nombre d'arcs donnés.

$$seuil = p.\min\left(\frac{e}{e'}, \frac{n}{n'}\right), \text{si } e' > 0, \text{sinon } seuil = p.\frac{n}{n'} \quad (1)$$

On considère qu'un symbole est fréquent s'il apparaît plus de p fois le nombre maximum de sous-graphes (avec n' nœuds et e' arcs) pouvant être contenus dans le graphe (p est un pourcentage).

L'algorithme que nous proposons est basé sur le principe de l'algorithme *A priori* et exploite également deux hypothèses de simplification :

- le nombre nœud contenu dans la représentation d'un symbole est rarement important ;
- les occurences d'un même symbole sont représentés par des sous-graphes disjoints.

Afin de réduire la complexité temporelle de notre algorithme, un réseau de graphes non-isomorphes est prédéterminé. Ce réseau est un graphe orienté acyclique dont les nœuds sont les graphes non-isomorphes dont le nombre d'arcs est inférieur à un paramètre MAX et dont les arcs représentent des relations d'inclusion.

La figure 3 représente un réseau de graphe non-isomorphes pour lequel le paramètre MAX a été fixé à 4. Ce réseau est parcouru en partant de la recherche des graphes fréquents n'ayant qu'un nœud, puis à chaque itération si un graphe est fréquent on cherche à lui ajouter un nœud de telle sorte que le graphe obtenu soit lui même fréquent.

Si à une certaine étape, un graphe n'est pas fréquent, tous ses descendants (les graphes engendrés par l'ajout de noeud à ce graphe) ne peuvent pas être fréquents.

Dans notre application, le réseaux a été calculé avec $MAX = 9$.

L'algorithme utilise les informations données par le réseau de graphes non-isomorphes (relations d'inclusion et automorphismes pour chaque graphe) pour une recherche efficace des sous-graphes fréquents.

4 Règles d'association et méta-règles

Après la découverte de symboles par l'algorithme précédent, une recherche des règles d'association entre ces symboles est effectuée en utilisant l'algorithme *A priori* [**?**].

En appliquant une partition d'un graphe initial, il est possible d'associer une transaction à chacune des parties issue de cette partition. Une partition du graphe peut être obtenue en appliquant un algorithme de classification non supervisée aux nœuds du graphe. La partition obtenue permet de déterminer k zones d'intérêt sur l'image du document. Les transactions peuvent également être définies en basant sur les relations d'inclusion entre les composantes. En effet, à partir du graphe relationnel attribué décrivant le document, il est possible de retrouver les relations d'inclusion et de considérer qu'une transaction contient tous les objets de même niveau d'inclusion.

L'algorithme *A priori* appliqué dans ce contexte permet d'extraire des règles entre objets de la transaction telles que :

$$\begin{array}{c} (O_{i1}, O_{i2}, \ldots, O_{in}) \Rightarrow (O_{j1}, O_{j2}, \ldots, O_{jm}) \\ \text{avec} \quad (O_{i1}, O_{i2}, \ldots, O_{in}) \cap (O_{j1}, O_{j2}, \ldots, O_{jm}) = \emptyset \end{array} \tag{2}$$

Il est possible dans chacune des transactions de vérifier si les règles obtenues par l'algorithme *A priori* sont confirmer ou pas. Une règle d'association est ensuite considérée comme un motif qui apparaît dans la transaction si elle y est confirmée. Cette approche peut être appliquée de façon récursive pour obtenir des méta-règles telles que :

$$((O_{i1}, \ldots, O_{k1}) \Rightarrow (O_{i2}, \ldots, O_{k2})) \quad \Rightarrow \quad ((O_{i3}, \ldots, O_{k3}) \Rightarrow (O_{i4}, \ldots, O_{k4})) \tag{3}$$

$$(O_{i1}, \ldots, O_{k1}) \quad \Rightarrow \quad ((O_{i2}, \ldots, O_{k2}) \Rightarrow (O_{i3}, \ldots, O_{k3})) \tag{4}$$

$$((O_{i1}, \ldots, O_{k1}) \Rightarrow (O_{i2}, \ldots, O_{k2})) \quad \Rightarrow \quad (O_{i3}, \ldots, O_{k3}) \tag{5}$$

Les méta-règles ainsi découverte ajoute des connaissances autres que celle qu'apportent les règles simples. Pour le montrer, nous présentons une méta-règle ne pouvant être réduite à une règle simple. La méta-règle $(O_1 \Rightarrow O_2) \Rightarrow (O_3 \Rightarrow O_4)$ s'écrit sous la forme normale disjontive $\overline{O_1}O_2 + \overline{O_3} + O_4$ mais il n'existe pas de règles simples telles que $(O_1, O_2) \Rightarrow (O_3, O_4)$ ou $O_1 \Rightarrow (O_2, O_3, O_4)$ s'écrivant sous forme normale disjonctive qui contienne la conjonction du positionnement d'un objet et de la négation d'un autre comme c'est le cas pour les méta-règles.

5 Exemple didactique

Cette section présente en guise d'exemple les résultats de notre approche appliquée à des images synthétiques contenant des symboles architecturaux (Fig. **??**). Dans un premier temps, les composantes connexes, les occlusions et les relations de voisinage

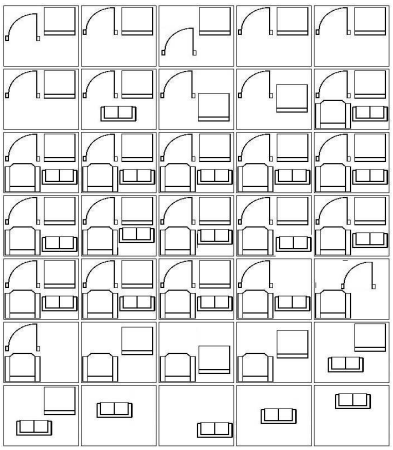

FIG. 4 – Image synthétique utilisée pour l'illustrer l'approche

sont extraites. Un graphe d'adjascence est construit et chaque nœud est étiqueté. Pour chaque forme extraite (composante connexe et occlusion) nous calculons les 9 premiers moments de Zernike [**?**]. Pour étiqueter les nœuds, une classification ascendante hiérarchique est construite sur les caractéristiques extraites. Le critère de Calinsky-Harabasz [**?**] est utilisé pour déterminier automatiquement le nombre de regroupements (et donc d'étiquettes).

Le seuil concernant le nombre d'occurrence défini par l'équation (1) est calculé avec $p = 0,2$. Un sous-graphe est dans notre cas considéré fréquent s'il apparaît au moins 6 fois.

La figure **??** montre les symboles fréquents extraits. En se basant sur la relation d'inclusion, on obtient les transactions suivantes :

$$T_1(S_0,S_1), T_2(S_0,S_1), T_3(S_0,S_1), T_4(S_0,S_1), T_5(S_0,S_1), T_6(S_0,S_1), T_7(S_0,S_1,S_3),$$
$$T_8(S_0,S_1), T_9(S_0,S_1), T_{10}(S_0,S_1,S_2,S_3), T_{11}(S_0,S_1,S_2,S_3), T_{12}(S_0,S_1,S_2,S_3),$$
$$T_{13}(S_0,S_1,S_2,S_3), T_{14}(S_0,S_1,S_2,S_3), T_{15}(S_0,S_1,S_2,S_3), T_{16}(S_0,S_1,S_2,S_3),$$
$$T_{17}(S_0,S_1,S_2,S_3), T_{18}(S_0,S_1,S_2,S_3), T_{19}(S_0,S_1,S_2,S_3), T_{20}(S_0,S_1,S_2,S_3),$$
$$T_{21}(S_0,S_1,S_2,S_3), T_{22}(S_0,S_1,S_2,S_3), T_{23}(S_0,S_1,S_2,S_3), T_{24}(S_0,S_2,S_3), T_{25}(S_0,S_2),$$
$$T_{26}(S_0,S_2), T_{27}(S_1,S_2), T_{28}(S_1,S_2), T_{29}(S_1,S_2), T_{30}(S_1,S_3), T_{31}(S_1,S_3), T_{32}(S_3),$$
$$T_{33}(S_3), T_{34}(S_3)$$

Par exemple, $T_1 S_0, S_1$ signifie que les symboles S_0 et S_1 sont présent au même niveau d'inclusion, dans le même rectangle dans notre cas.

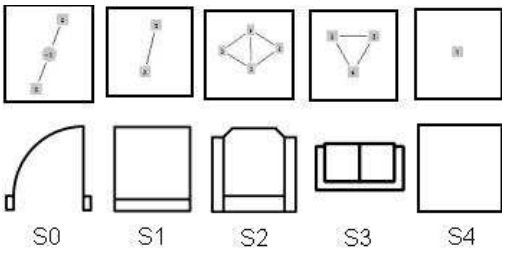

FIG. 5 – Symbole fréquents découverts

À partir de ces transactions, les règles et méta-règles suivantes sont extraites :

$$\begin{array}{lll} R_1 : (S_0 \Rightarrow S_1) & support = 0,74 & confiance = 0,88 \\ R_2 : (S_2 \Rightarrow S_0) & support = 0,57 & confiance = 0,85 \\ R_3 : (S_3 \Rightarrow (S_2 \Rightarrow S_0)) & support = 0,62 & confiance = 1,0 \end{array}$$

Les règles ont été sélectionnées en choissant un seuil de 0,8 pour la confiance et de 0,5 pour le support.

Notre approche a également été testée sur 108 images de fax. Les images de ce corpus contiennent un certain nombre d'objets d'intérêt pour la communauté de l'analyse d'image de document : logos, tableau, écriture manuscrite et texte imprimé. Ces images sont affectées de bruit caractéristique sur les fax et ont subi le plus souvent une rotation. Dans ce cas, les seuls symbols découvert comme fréquents sur l'ensemble de la base sont des logos. La faiblesse des résultats s'explique par le niveau important de bruit présent sur les images. La représentation sous forme de graphe des images semble ne pas être dans ce cas bien adaptée.

6 Conclusion

Cet article présente une approche novatrice dans le domaine de l'analyse des documents. Elle utilise les concepts de fouille de graphes et de recherche de règles d'assocuiation pour l'extraction de connaissances. Elle vise, sans connaissance du modèle de document, à l'extraction de symboles et de plusieurs niveaux de règles d'association. Les motifs fréquents découverts automatiquement peuvent être rapprochés des connaissances liées au domaine d'usage du document.

La méthode exposée peut être appliquée à d'autres représentations structurelles de documents, la seule restriction étant que les objets présents sur le documents doivent être représentés par des sous-graphes dont les nœuds doivent être distincts. Nous envisageons en particulier de tester cette approche sur des résultats de segmentation de documents structurés (à dominante textuelle) afin d'extraire automatiquement les règles relatives à la mise en page.

Même si cette approche novatrice semble intéressante dans le sens où elle permet sans *a priori* d'extraire des motifs fréquents pouvant être rapprochés des connaissances liées au domaine spécifique du document, les travaux doivent être poursuivis pour une

application à des données réelles souvent bruitées et dégradées. Pour tendre vers cet objectif, plusieurs perspectives peuvent être formulées. En particulier, un post-traitement devra pouvoir être appliqué au graphe de voisinage afin d'atténuer les effets liés au bruitage des images et des effets de bord des outils d'extractions de traitement d'image. Une utilisation d'un algorithme d'appariement de graphes tolérant aux erreurs permettrait également de s'abstraire des erreurs provenant des traitements antérieurs. Par ailleurs, des indices plus performants [**?**] devront pouvoir être trouvés pour atteindre des règles de niveau sémantique, ces règles pouvant alors être hiérarchisées par une approche similaire de celle développée par Gras et al. [**?**].