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 représentatifs de symboles sur des images de documents. Les symboles sont considérés comme des entités graphiques porteurs d'information et les images de document sont représentées par des graphes relationnels attribués. Dans un premier temps, la méthode réalise la découverte de sous-graphes disjoints fréquents et fait correspondre pour chacun d'eux un symbole différent. Une recherche des règles d'association entre ces symboles permet alors d'accéder à une partie des connaissances du domaine décrit par ces symboles. L'objectif à terme est d'utiliser les symboles découverts pour la classification ou la recherche d'images dans un flux hétérogène de document là où une approche supervisée n'est pas envisageable.

1 Introduction

Dans 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 une interprétation plus fine du contenu, et d'autre part pour la recherche des occurrences plus ou moins fréquentes d'un symbole particulier dans un ou plusieurs documents.

La nature des symboles utilisés permet de reconnaître le domaine dont relève le document. Nous considérons comme un 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, puis le principe de recherche de parties fréquentes adopté.

La section 2 présente le contexte et les travaux existants dans le domaine abordé. La section 3 détaille l'algorithme permettant la recherche de sous-graphes fréquents. La section 4 traite de la découverte des règles d'association entre les symboles. La section 5 illustre l'application de la méthode à travers un exemple. Enfin, la section 6 dresse

FIG. 1 – Images représentant un symbole apparaissant dans des contextes différents

des conclusions et énonce quelques perspectives permettant de prolonger le travail.

2 Contexte

« 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 »(Washio et Motoda 2003). Les premiers systemes issus de ce champ de recherche furent SUBDUE (Cook et Holder 1994) et GBI (Yoshida et al. 1994). Ils s'appuient sur la méthode GLOUTON et peuvent 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 (Yan et Han 2003). Les techniques de fouille de graphes ont par le passé été appliquées à l'analyse de scènes, aux bases de données de composés chimiques et aux flux de travail.

Si on représente 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 fréquemment dans différents contextes (cf. Fig. 1).

Le problème de la découverte de règles d'association a été introduit par Agrawal dans le domaine de l'analyse du contenu des chariots de supermarchés (Agrawal et Srikant 1994). Une règle d'association signifie 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 alors une implication. La qualité de cette approximation peut être donnée par différents indices. Parmi ceux-ci, 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 que l'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 du symbole S_3 dans un *document* implique la probable présence des symboles S_4, S_5 et S_6 dans ce document. »

Considérant une transaction T et une règle d'association $X \Rightarrow Y$, notée R, R est

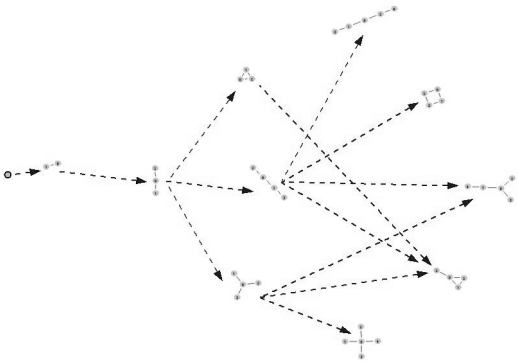


FIG. 2 – Réseau des graphes non-isomorphes

confirmée si X n'apparaît pas dans la transaction ou si Y y apparaît. La confirmation des règles dans les transactions peut être utilisée pour construire des méta-règles. Nous indiquons en section 4 ce 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 trouvés par des techniques de recherche de sous-graphes fréquents, nous définissons une transaction comme un partie du document (par exemple un paragraphe) ou le document dans sa totalité.

3 Recherche des sous-graphes fermés fréquents

Considérant un graphe composé de n nœuds et de e arcs, 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 peut alors être calculé à partir d'une approximation de ce nombre maximum de sous-graphes disjoints qu'il est possible de construire à nombres de nœuds et d'arcs donnés.

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

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

– le nombre de nœuds 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. Ses 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 2 représente un réseau de graphes 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 un graphe n'est pas fréquent, tous ses descendants (les graphes pouvant être engendrés par l'ajout de nœuds à ce graphe) ne peuvent être fréquents. Dans notre application, le réseau 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 présentée précédemment, une recherche des règles d'association entre ces symboles est effectuée en utilisant l'algorithme *A priori* (Agrawal et Srikant 1994).

En appliquant une partition du 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. Les transactions peuvent également être définies en se 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 d'un même niveau d'inclusion.

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

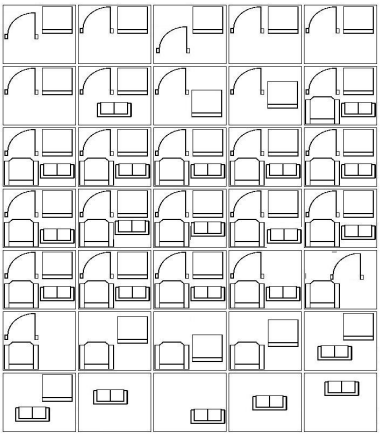
$$\begin{array}{c} (O_{i1}, O_{i2}, \ldots, O_{in}) \Rightarrow (O_{j1}, O_{j2}, \ldots, O_{jm}) \\ \text{avec} \quad (O_{i1}, O_{i2}, \ldots, O_{in}) \cap (O_{j1}, O_{j2}, \ldots, O_{jm}) = \emptyset \end{array} \tag{2}$$

Dans chaque transaction, il est possible de vérifier si les règles obtenues par l'algorithme *A priori* sont confirmées ou non. 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.

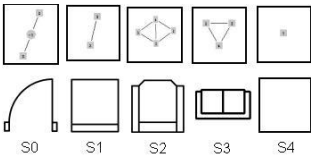
Les méta-règles ainsi découvertes ajoutent des connaissances autres que celle qu'apportent les règles simples. Pour le montrer, nous présentons une méta-règle ne pouvant être réduite à une règle simple. La méta-règle $(O_1 \Rightarrow O_2) \Rightarrow (O_3 \Rightarrow O_4)$ s'écrit sous la forme normale disjonctive $O_1\overline{O_2} + \overline{O_3} + O_4$ mais il n'existe pas de règle simple telle que $(O_1, O_2) \Rightarrow (O_3, O_4)$ ou $O_1 \Rightarrow (O_2, O_3, O_4)$ qui, écrite sous forme normale disjonctive, 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. 3(a)). Dans un premier temps, les composantes connexes, les occlusions et les relations de voisinage



(a) Image synthétique utilisée pour l'illustrer l'approche



(b) Symboles fréquents découverts

FIG. 3 – Expérimentation

sont extraites. Un graphe d'adjacence est construit et chaque nœud est étiqueté. Pour ce faire, on extrait de chaque forme (composante connexe et occlusion) les 9 premiers moments de Zernike (Khotanzad et Hong 1990). Pour étiqueter les nœuds, une classification ascendante hiérarchique est construite sur les caractéristiques extraites. Le critère de Calinsky-Harabasz (Milligan et Cooper 1985) est utilisé pour déterminer automatiquement le nombre de regroupements proposés par la CAH (et donc d'étiquettes).

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

La figure 3(b) montre les symboles fréquents extraits. Par exemple, une transaction $T_1(S_0, S_1)$ signifie que les symboles S_0 et S_1 sont présents au même niveau d'inclusion (dans notre cas, le rectangle en haut à gauche de la figure 3(a)).

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.

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'association 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 illustrer les connaissances liées au domaine d'usage du document.

La méthode exposée peut être appliquée à d'autres représentations structurelles de document, la seule restriction étant que les objets présents sur le documents doivent être représentés par des sous-graphes dont les nœuds doivent être distincts.

Même si cette approche novatrice est intéressante car elle permet sans *a priori* d'extraire des motifs fréquents spécifiques des connaissances liées au domaine 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.

Références

Agrawal, R. et Srikant, R. (1994), Fast algorithms for mining association rules. In Bocca,J.B., Jarke, M., Zaniolo, C., eds., Proc. 20th Int. Conf. Very Large Data Bases, VLDB, Morgan Kaufmann, pp 487–499.

Washio, T., Motoda, H. (2003), State of the art of graph-based data mining. SIGKDD Explor. Newsl. 5, pp 59–68.

Yan, X. et Han, J. (2003), Closegraph : mining closed frequent graph patterns, Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp 286–295.

Cook, J. et Holder, L. (1994), Substructure discovery using minimum description length and background knowledge. J. Artificial Intel. Research, Vol. 1, pp 231–255.

Yoshida, K., Motoda, H., et Indurkhya, N. (1994), Graph based induction as a unified learning framework. J. of Applied Intel., Vol. 4, pp 297–328.

Milligan, G. W. et Cooper, M.C. (1985), An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika, 58(2),pp 159–179.

Khotanzad, A. et Hong, Y.H. (1990), Invariant Image Recognition by Zernike Moments. IEEE Trans. on PAMI, 12(5). pp 289–497.