Découverte de motifs fréquents : Application à l'analyse de documents graphiques

Eugen Barbu, Pierre Héroux, Sébastien Adam, Eric Trupin

Découverte de motifs fréquents : Application à l'analyse de documents graphiques

Eugen Barbu – Pierre Héroux – Sébastien Adam – Éric Trupin

Laboratoire PSI
Université et INSA de Rouen
F-76821 Mont-Saint-Aignan, France

{Eugen.Barbu, Pierre.Heroux, Sebastien.Adam, Eric.Trupin}@univ-rouen.fr

Résumé : *Cet article présente une méthode pour l'extraction des motifs fréquents sur des images de documents graphiques. Cette méthode, appliquée à une représentation structurelle bas niveau des dessins techniques, utilise les concepts de la fouille de données et de l'extraction de connaissance. La découverte de sous-graphes fréquents et de règles d'association entre ces sous-graphes a pour objectif l'extraction automatique des symboles d'un document et de leurs relations. La principale contribution de ce travail réside dans la proposition, d'une part, d'un algorithme qui extrait des sous-graphes fréquents, et d'autre part, d'une méthode permettant la découverte de règles de plusieurs niveaux associant les motifs découverts. De premiers résultats montrent que cette approche est adaptée à la reconnaissance de formes, de symboles et des relations qu'ils entretiennent.*

Mots-clés : document graphique, extraction de motifs, reconnaissance de symboles, fouille de données dans les graphes, extraction de connaissances

1 Introduction

Sur un document technique, un symbole est un signe (élément graphique) qui, selon certaines conventions, encode une unité élémentaire de message. Dans cet article, nous présentons comment extraire automatiquement des symboles et les relations qu'ils entretiennent. Le signe et les règles associées peuvent être considérés comme une approximation du message véhiculé par le symbole.

L'extraction automatique de symboles sur les images de document sans connaissance du domaine d'usage ou du contexte est une tâche ardue. Cette approche a été abordée par Altamura et al. [ALT 00] et Messmer [MES 95]. Concernant les documents techniques (schémas architecturaux, mécaniques, électriques…), elle peut être abordée en détectant les occurrences fréquentes de motifs au sein des documents. Ces motifs peuvent être, selon le niveau de segmentation, des composantes connexes, des formes géométriques élémentaires (lignes, arcs de cercle) ou composites, ces dernières étant représentées alors par des graphes exprimant les relations de voisinage entre formes géométriques élémentaires ([BER 03], [COR 04], [ORD 99]). Une extension possible à cette approche est l'extraction des relations qui existent entre ces symboles. De telles relations peuvent être considérées comme de nouvelles entités, pouvant se révéler, elles-mêmes fréquentes, et participer à leur tour à des relations plus complexes. Dans ce contexte, l'algorithme *A priori* est largement utilisé et reconnu pour la découverte d'items fréquents [AGR 94]. Cependant, lorsque les objets sont des graphes, quelques modifications doivent être apportées à l'algorithme original. Par exemple, Kuramochi et Karypis [KUR 01] et Inokuchi et al. [INO 00] présentent des adaptations de l'algorithme *A Priori* pour la découverte de sous-graphes fréquents. La fouille de données à base de graphes [WAS 03] est un domaine de recherche qui décrit de nouveaux principes et algorithmes pour la découverte de sous-structures topologiques dans des données représentées sous forme de graphes.

Cet article propose un algorithme permettant l'extraction de sous-graphes fréquents au sein d'un graphe, ainsi qu'une méthode pour la découverte de plusieurs niveaux de règles d'association entre les symboles. Le principe de l'approche est donnée sur la figure 1. Il est possible de catégoriser un ensemble d'images de document selon les symboles qu'elles contiennent et selon les règles qu'entretiennent ces symboles. Deux documents peuvent alors être rattachés à la même classe s'ils respectent les mêmes règles. Une distance entre documents peut également être évaluée en examinant si un document se conforme aux règles d'un autre.

Cet article est organisé comme suit. La section 2 présente l'algorithme d'extraction des sous-graphes fréquents. La section 3 décrit notre approche pour l'extraction des règles d'association. Enfin, un exemple et une étude préliminaire de la robustesse de notre approche sont présentés en section 4.

2 Un algorithme pour l'extraction de sous-graphes fréquents

L'approche que nous proposons se base sur le fait que les symboles sur les documents techniques traduisent graphiquement des éléments de message selon certaines conventions. Ainsi, au sein d'une même classe de documents, le même motif représente toujours la même entité. Les symboles d'une classe de documents, ou tout au moins un certain nombre, apparaissent donc avec une certaine fréquence sur les documents relevant de ce domaine. L'algorithme que nous proposons a pour objectif de trouver des sous-graphes connexes fréquents au sein de graphes extraits depuis les

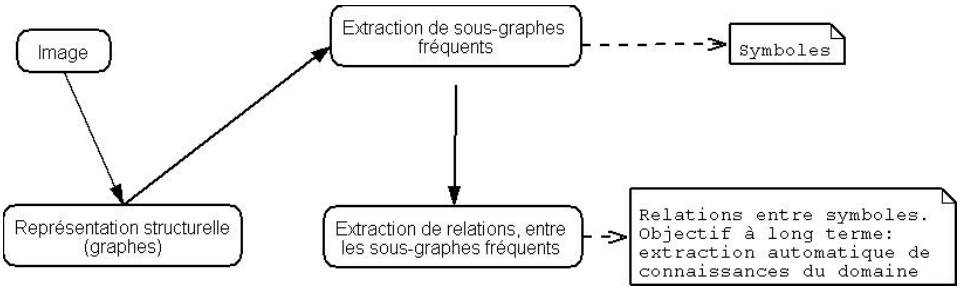


FIG. 1 – Principe de l'approche

images de document. De plus, les sous-graphes représentant les symboles sont des graphes fermés (un graphe est dit fermé s'il n'est pas inclus dans un super-graphe ayant le même nombre d'occurences) [YAN 03].
Ces graphes sont construits en représentant par des nœuds des occlusions (formes 2D) ainsi que des formes linéaires (formes 1D) et par des arcs les relations d'adjacence. Ce type de graphe est un super-graphe du RAG (Region Adjency Graph) décrit par Pavlidis dans [PAV 82] car il contient, en plus, les nœuds décrivant les formes non fermées (formes 1D) et leurs relations de voisinage. Chaque nœud du graphe est étiqueté par les 8 premiers moments de Zernike extraits de la forme. De cette façon, deux graphes représentent le même symbole s'ils sont isomorphes et si les nœuds appariés deux à deux ont des valeurs d'invariants équivalentes.

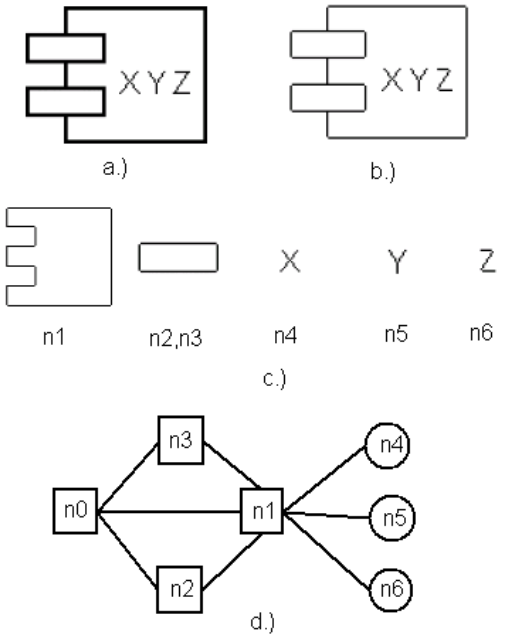


FIG. 2 – Un symbole a.) et son graphe associé d.) intégrant la représentation du fond (n0). Les formes 1D sont représentées par des cercles et les formes 2D sont représentées par des carrés. Les relations d'adjascence sont calculées à partir de le la version squelettisée de l'image b.)

Définition 1 *Transaction*
Soit $I = \{i_1, i_2, i_3, \ldots, i_n\}$ *un ensemble d'objets appelés items.*

Une transaction T *est un ensemble d'items tel que* T *est inclus dans* I.

Dans ce contexte, un sous-graphe est considéré fréquent si le nombre de ses occurrences (non inclus dans un autre sous-graphe) est supérieur à un seuil s. Ce seuil ne peut pas être défini par rapport au nombre de transactions comme c'est le cas dans d'autres algorithmes ([KUR 01], [INO 00]). En effet, un sous-graphe peut avoir plusieurs occurrences au sein d'un même graphe. Une possiblilité pour fixer ce seuil est de le mettre en rapport avec le nombre maximum de sous-graphes (à nombres de nœuds et d'arcs donnés) composés de nœuds distincts qu'il est possible de construire sur le graphe représentant le document.
L'algorithme proposé est une adaptation de l'algorithme *A priori* exploitant deux hypothèses :
– les symboles sont rarement représentés par des graphes dont le nombre de nœuds est supérieur à 10 ;
– les symboles d'un document sont représentés par des graphes dont les nœuds sont distincts.
L'idée principale de l'algorithme *A priori* est qu'un objet est fréquent si tous ses sous-éléments sont également fréquents. Appliquée au cas des graphes, cette proposition n'est vérifiée que si, comme dans notre cas, les sous-graphes fréquents ont des nœuds distincts. Sur le figure 3, le graphe c) n'a qu'une occurrence dans le graphe a). Si on considère que les sous-graphes peuvent avoir des nœuds communs, on trouve trois occurences de b) dans a). Dans notre cas, les noeuds ne participent à la représentation que d'un seul symbole, on cherche donc des sous-graphes dont les nœuds sont distincts. On ne peut alors avoir qu'un seul occurrence de b) dans a).

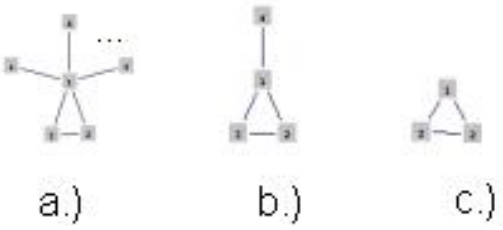


FIG. 3 – Exemple

Dans le but de réduire la complexité algorithmique, un réseau de graphes non isomorphes est construit. Ce réseau est un graphe orienté acyclique dont les nœuds sont tous les graphes non isomorphes dont le nombre d'arcs est inférieur à un paramètre MAX. La figure 4 présente le réseaux des sous-

graphes non isomorphes lorsque MAX vaut 5. Si à un certain niveau de ce réseau, un graphe n'est pas fréquent, ses successeurs, c'est-à-dire les graphes qu'il est possible de contruire en lui ajoutant des arcs (et éventuellement des nœuds), ne peuvent pas être fréquents.

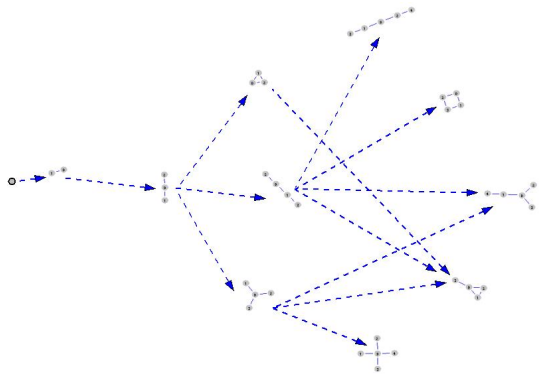


FIG. 4 – Réseau des graphes non isomorphes dont le nombre d'arcs est inférieur à $MAX = 5$

Dans la mise en œuvre implantée, MAX est fixé à 9, car la taille du réseau croît plus de que façon exponentielle. Par ailleurs, les symboles sur les documents traités sont représentés par des graphes dont le nombre d'arcs est inférieur à 9. La recherche des graphes fréquents en utilisant ce réseau s'effectue en temps polynomial.

Algorithme 1

```
Initialisation du réseau de graphes non isomorphes
dont le nombre d'arcs est inférieur à MAX.
Entrée : Graphe(s) étiqueté(s) non orienté(s)
Sortie : Liste de sous-graphes fréquents et de ses
occurrences
Début
  k<-0
  tant que k<MAX faire
  Début
    pour tous les graphes pouvant être fréquents
    possédant k arcs
    Début
      Soit G le graphe courant
      La liste des occurrences de G est contruite à
        partir de la liste des occurences de ses
        prédécesseurs
      Si le cardinal de la liste est supérieur au
        seuil s
      Alors G est considéré fréquent
      Si G est fréquent
      Alors mise à jour des listes de prédécesseurs
        en précisant qu'ils sont inclus dans un
        graphe fréquent
      Sinon mise à jour des successeurs de G en
        précisant qu'ils ne peuvent pas être
        fréquents
    Fin
    pour tous les graphes fréquents du niveau
      anterieur possédant k-1 arcs
      mise à jour de la listes des occurences en
        prenant  en compte l'inclusion dans un
        graphe fréquent
  k<-k+1
  Fin
Fin
```

Le seuil s, au dessus duquel un sous-graphe est considéré fréquent, est calculé en utilisant la formule suivante, où p représente un ratio, e et n sont les nombres d'arcs et de nœuds du graphe représentant le document, e' est le nombre maximum d'arcs des sous-graphes fréquents et n' est le nombre de nœuds correspondant.

$$s = p \min\left(\frac{e}{e'}, \frac{n}{n'}\right) \tag{1}$$

Cette formule est une approximation du nombre maximum de sous-graphes qu'on peut trouver dans un graphe. On considère qu'un sous-graphe est fréquent si son nombre d'occurrences est supérieur à un ratio p du nombre maximal de sous-graphes potentiels. L'algorithme est alors appliqué à un graphe ou à un ensemble de graphes représentant respectivement un document ou un ensemble de documents. Les symboles n'apparaissant que rarement ne sont pas localisés.

3 Règles et méta-règles

Après l'extraction des sous-graphes fréquents représentant potentiellement des symboles par l'algorithme présenté précédemment, nous considérons les relations entre ces symboles. La recherche des règles d'association est effectuée par l'algorithme *A priori*. Les symboles partageant des propriétés communes participent à une transaction. L'extraction des relations entre symboles est effectuée en examinant l'ensemble des transactions. Par exemple, en examinant l'ensemble des transactions suivant, on peut extraire la règle "*Si l'objet O_1 est présent, l'objet O_2 est probablement présent également*" vérifiée dans les transactions T_1, T_2 et T_4 et non infirmée dans T_3.

$$T_1(O_1, O_2, O_3); T_2(O_1, O_2); T_3(O_2, O_3); T_4(O_1, O_2, O_4);$$

Les transactions peuvent également être définies en utilisant d'autres critères. Par exemple, une transaction peut-être associée à un document. Les relations extraites signifient alors que si un symbole apparaît sur un document, les objets associés y apparaissent probablement. L'algorithme *A priori* trouve alors des règles d'association telles que celles données en [2].

$$\begin{gathered}(O_{i,1}, O_{i,2}, \ldots, O_{i,n}) \Rightarrow (O_{j,1}, O_{j,2}, \ldots, O_{j,m}) \\ \text{avec } (O_{i,1}, O_{i,2}, \ldots, O_{i,n}) \cap (O_{j,1}, O_{j,2}, \ldots, O_{j,m}) = \emptyset\end{gathered} \tag{2}$$

Si on considère une règle R issue de l'algorithme *A priori*, on peut observer pour chaque transaction si cette règle est vérifiée ou pas. De cette façon, une règle devient un motif représentatif de classe de documents si elle est confirmée avec une certaine fréquence. De façon récursive, on peut également obtenir des règles telles que celles présentées en [3]. Ce type de règles est plus difficile à exprimer en langage naturel mais elles sont plus proches de l'expression des connaissances du domaine. Une implication entre deux règles ($R_1 \Rightarrow R_2$) signifie que toutes les transactions qui vérifient R_1 vérifient également R_2.

$$\begin{gathered}R_i \Rightarrow R_j \\ R_i \Rightarrow (R_j \Rightarrow R_k) \\ (R_i \Rightarrow R_j) \Rightarrow R_k \\ (R_i \Rightarrow R_j) \Rightarrow (R_k \Rightarrow R_t)\end{gathered} \tag{3}$$

4 Exemples

4.1 Exemple didactique

Cette section présente une mise en œuvre didactique de notre approche sur un document synthétique (Fig. 5) composé de symboles architecturaux. Une extraction des composantes connexes, des occlusions et des voisinages donne le graphe représenté sur la figure 6(a). Il est possible de traiter ce graphe afin d'extraire le voisinage particulier qu'est l'inclusion. On obtient alors un arbre d'inclusion (Fig. 6(b)).

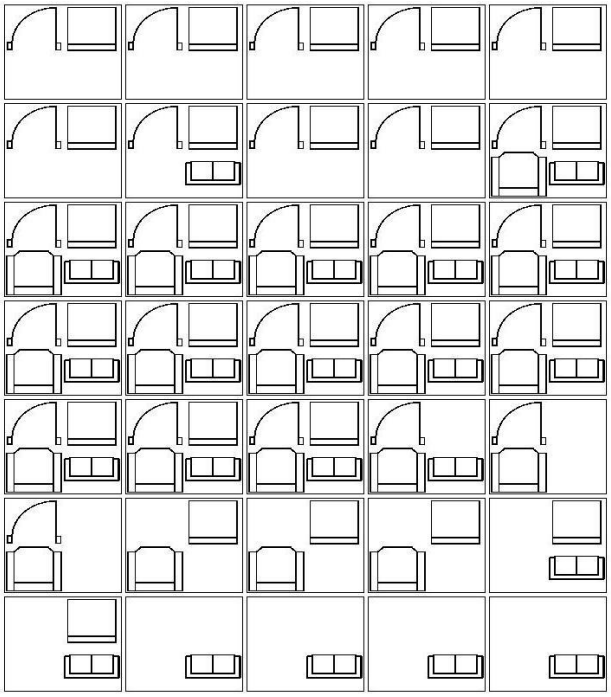


FIG. 5 – Exemple d'un document technique

Une valeur du paramètre p égale à 0,2 donne, en appliquant l'équation 1 une valeur du seuil s de 6. Un sous-graphe est considéré fréquent s'il dispose donc de 6 occurrences au minimum. Les résultats de la recherche des sous-graphes fréquents ainsi que les symboles correspondants sont donnés sur la figure 7. Les transactions contenant les symboles découverts sont alors construites en donnant à chaque transaction la liste des symboles se trouvant sur les feuilles de l'arbre d'inclusion (Tableau 1).

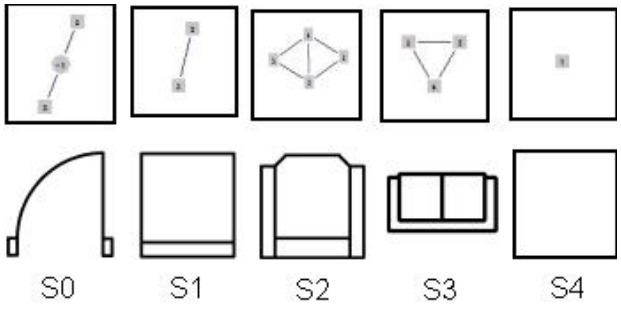


FIG. 7 – Sous-graphes fréquents découverts et les symboles correspondant

	S_0	S_1	S_2	S_3
T_1	1	1	0	0
T_2	1	1	0	0
T_3	1	1	0	0
T_4	1	1	0	0
T_5	1	1	0	0
T_6	1	1	0	0
T_7	1	1	0	1
T_8	1	1	0	0
T_9	1	1	0	0
T_{10}	1	1	1	1
T_{11}	1	1	1	1
T_{12}	1	1	1	1
T_{13}	1	1	1	1
T_{14}	1	1	1	1
T_{15}	1	1	1	1
T_{16}	1	1	1	1
T_{17}	1	1	1	1
T_{18}	1	1	1	1

	S_0	S_1	S_2	S_3
T_{19}	1	1	1	1
T_{20}	1	1	1	1
T_{21}	1	1	1	1
T_{22}	1	1	1	1
T_{23}	1	1	1	1
T_{24}	1	0	1	1
T_{25}	1	0	1	0
T_{26}	1	0	1	0
T_{27}	0	1	1	0
T_{28}	0	1	1	0
T_{29}	0	1	1	0
T_{30}	0	1	0	1
T_{31}	0	1	0	1
T_{32}	0	0	0	1
T_{33}	0	0	0	1
T_{34}	0	0	0	1
T_{35}	0	0	0	1

TAB. 1 – Liste des transactions (1 indique la présence de l'item dans la transaction, 0 indique son abscence)

Les indices de support et de confiance sont traditionnellement utilisés pour qualifier les règles d'implication. Pour une règle $a \Rightarrow b$, ces indices sont définis par :

$$support = \frac{n_a}{n} \text{ et } confiance = \frac{n_{ab}}{n_a}$$

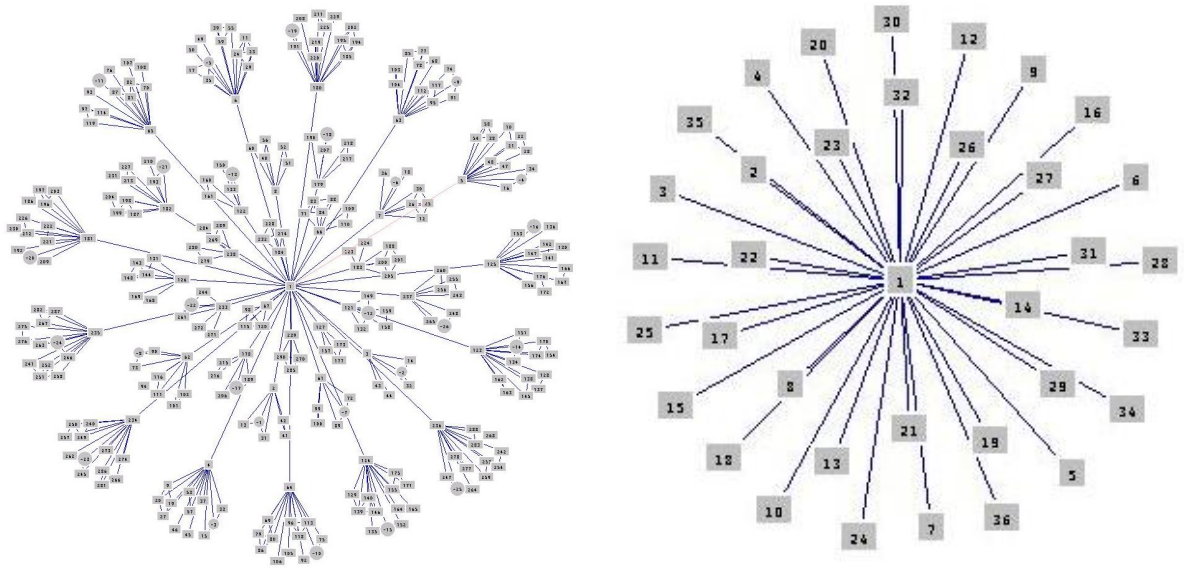
où n est le nombre de transactions, n_a est le nombre de transactions où a est présent n_{ab} est le nombre de transactions où a et b sont présents. Les transactions du tableau 1 permettent d'extraire les règles et méta-règles suivantes (les seuils sont fixés à 0,5 pour le support et 0,8 pour la confiance) :

$$\begin{array}{ll} R_1 : (S_0 \Rightarrow S_1) & support = 0,74 \\ & confiance = 0,88 \\ R_2 : (S_2 \Rightarrow S_0) & support = 0,57 \\ & confiance = 0,85 \\ R_3 : (S_3 \Rightarrow (S_2 \Rightarrow S_0)) & support = 0,62 \\ & confiance = 1 \end{array} \quad (4)$$

R_1 indique la présence de S_1 à 88% quand S_0 est présent. R_2 signifie qu'en présence de S_2, S_0 apparaît à 85%. Enfin, R_3 précise que R_2 n'est jamais démentie en présence de S_3.

4.2 Évaluation de la robustesse

Nous présentons une première évaluation de la robustesse de notre approche au bruit. La figure 8(a) représente plusieurs occurrences du même document synthétique sur lequel ont été appliqué plusieurs niveaux de bruit. Deux catégories de bruit ont été introduites modélisant d'une part, les superpositions et connexion d'informations graphiques ($Vb1$), et d'autre part, la qualité de la numérisation des documents par l'ajout d'un bruit gaussien ($Vb2$). La figure 8(b), donne pour chaque niveau de bruit $Vb1$, l'évolution, en fonction de $Vb2$, de la proportion d'objets retrouvés. Même si on constate une chute des performances, en partie due à la qualité des prétraitements, notre objectif n'est pas l'extraction de tous



(a) (b)

FIG. 6 – Graphe de voisinage et arbre d'inclusion correspondant

les symboles mais plutôt d'identifier des redondances qui caractérisent le document. On peut cependant conclure que les seuils permettant de considérer un sous-graphe comme fréquent doivent être adaptés à la qualité de image et à la richesse de l'information graphique qu'elle supporte.

5 Conclusion

Cet article présente une approche novatrice dans le domaine de l'analyse des documents techniques. Elle utilise les concepts de fouille de données pour l'extraction de connaissance. 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 [FLE 95] 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. [GRA 03].

Références

[AGR 94] AGRAWAL R., SRIKANT R., Fast Algorithms for Mining Association Rules, BOCCA J. B., JARKE M., ZANIOLO C., Eds., *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, Morgan Kaufmann, 12–15 1994, pp. 487–499.

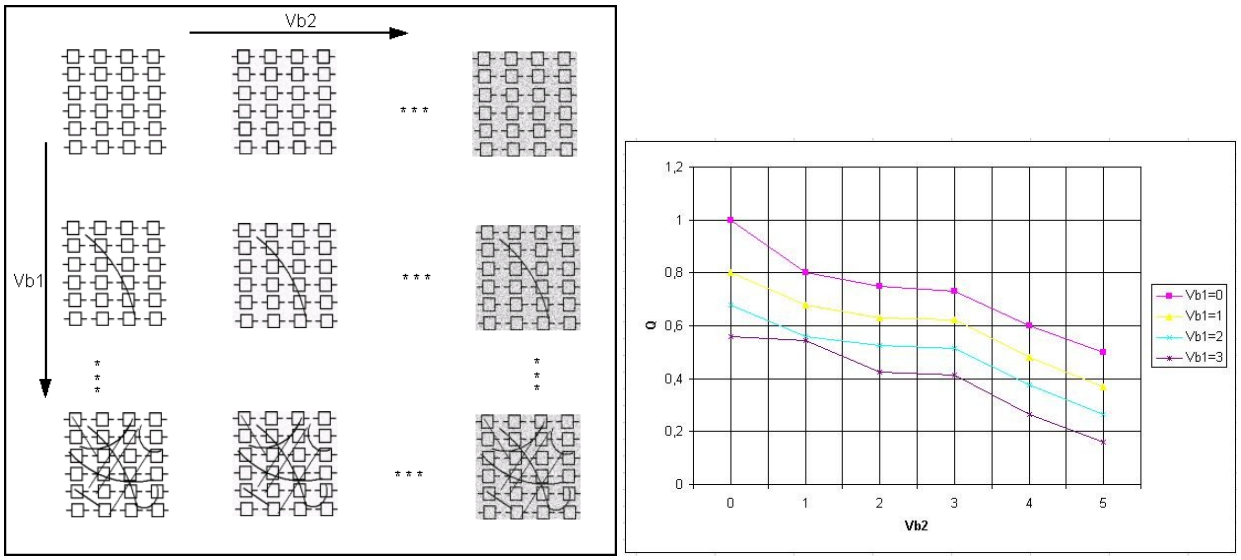
[ALT 00] ALTAMURA O., ESPOSITO F., MALERBA D., Transforming Paper Documents into XML Format with WISDOM++, *International Journal on Document Analysis and Recognition*, vol. 3, n° 2, 2000, pp. 175-198.

[BER 03] BERARDI M., CECI M., MALERBA D., Mining Spatial Association Rules from Document Layout Structures, *Proceedings of the Third International Workshop on Document Layout Interpretation and its Applications*, 2003.

[COR 04] CORNUÉJOLS A., MARY J., SEBAG M., Classification d'images à l'aide d'un codage par motifs fréquents, *Actes de la Journée analyse de données, statistique et apprentissage pour la fouille d'image du Congrès RFIA*, 2004, pp. 11-16.

[FLE 95] FLEURY L., MASSON Y., The intensity of implication, a measurement learning machine, *Proceedings of the eighth international conference on Industrial and engineering applications of artificial intelligence and expert systems*, 1995, pp. 621–629.

[GRA 03] GRAS R., KUNTZ P., BRIAND H., Hiérarchie orientée de règles généralisées en analyse implicative, *Actes des journées francophones d'extraction et de gestion des connaissances*, 2003.



(a) Différents niveaux de bruit

(b) Évolution de la performance

FIG. 8 – Évaluation de performance

[INO 00] INOKUCHI A., WASHIO T., MOTODA H., An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data, *Proceedings of the Conference on Principle and Practice of Knowledge Discovery in Databases*, 2000.

[KUR 01] KURAMOCHI M., KARYPIS G., Frequent Subgraph Discovery, *Proceedings of the International Conference on Data Mining*, 2001.

[MES 95] MESSMER B., Efficient Graph Matching Algorithms for Preprocessed Model Graphs, PhD thesis, University of Bern, CH, Institute of Applied Mathematics, 1995.

[ORD 99] ORDONEZ C., OMIECINSKI E., Discovering Association Rules based on Image Content, *Proceeding of the IEEE Advances in Digital Libraries Conference*, 1999.

[PAV 82] PAVLIDIS T., *Algorithms for Graphics and Image Processing*, Computer Science Press, 1982.

[WAS 03] WASHIO T., MOTODA H., State of the art of graph-based data mining, *SIGKDD Explor. Newsl.*, vol. 5, n° 1, 2003, pp. 59–68, ACM Press.

[YAN 03] YAN X., HAN J., CloseGraph : mining closed frequent graph patterns, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press, 2003, pp. 286–295.