Représentation vectorielle pour l'indexation d'informations structurelles

Nicolas Sidère, Pierre Héroux, Jean-Yves Ramel

Représentation vectorielle pour l'indexation d'informations structurelles

Nicolas SIDERE[1,2] – Pierre HEROUX[1] – Jean-Yves RAMEL[2]

[1] Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS)
Université de Rouen
Avenue de l'université
76800 Saint-Etienne du Rouvray

[2] Laboratoire d'Informatique (LI)
Université François Rabelais
64 Avenue Jean Portalis
37200 Tours

{Nicolas.Sidere, Pierre.Heroux}@univ-rouen.fr
Jean-Yves.Ramel@univ-tours.fr

Résumé : *Cet article présente une représentation vectorielle des données structurées visant à réduire la complexité des calculs de dissimilarités dans un contexte de recherche d'information. Cette représentation permet via un calcul de distance adapté, d'approximer la distance entre représentations structurelles aussi bien dans un contexte de distance entre graphes que pour la recherche d'occurences de sous-graphes. De premières expérimentations montrent que la représentation proposée offre des performances comparables à celles de la littérature utilisées pour approximer des distances entre graphes.*

Mots-clés : Signature de graphes – Indexation et recherche d'information structurée

1 Introduction

L'évolution des techniques de numérisation, la facilité de diffusion par l'Internet et la volonté de conserver et d'accéder facilement aux documents numériques ont mis les thèmes de l'indexation et de la recherche de documents au centre de nombreux axes de recherches. Avec la diversité des thématiques (valorisation du patrimoine, archivage de données administratives, lecture automatique, aide à la recherche d'ouvrages,...) la masse de données documentaires ainsi générée augmente considérablement, ce qui contribue aussi à la multiplication des travaux dans ce domaine. Si des solutions commencent à voir le jour, ces dernières restent souvent ciblées et restreintes à un domaine d'application ou un corpus précis. Beaucoup utilisent une indexation basée sur des informations textuelles extraites grâce à des systèmes de reconnaissance optique de caractères inefficaces sur certains types de documents (archives du patrimoine, documents graphiques,...) ou sur une annotation manuelle limitée par le nombre d'ouvrages et par la subjectivité apportée par l'utilisateur. On voit donc apparaître un intérêt pour des interrogations utilisant de nouvelles modalités. En conséquence, nombre de travaux actuels s'orientent vers une caractérisation des documents par de nouveaux indices tels que la structure. Les travaux décrits dans cet article s'inscrivent dans cette démarche. Suivant le cadre applicatif, les informations structurelles décrivent différents aspects du document :

1. La description physique du document, l'agencement des différents paragraphes, des illustrations, des titres, etc... Par exemple, la mise en page d'une page d'un annuaire est significative et reconnaissable au premier coup d'oeil ;
2. L'organisation logique (titre, section, sous-section, paragraphe,...) permet également de différencier des ouvrages, un journal d'un roman par exemple ;
3. Certains types de formes sont souvent représentés par des informations structurelles. C'est le cas, en particulier, des symboles graphiques apparaissant sur les documents techniques.

La recherche d'information vise à établir la pertinence d'un document vis-à-vis d'une requête formulée par un utilisateur. Lorsque les documents sont décrits par des informations structurelles, cette mesure de la pertinence est souvent basée sur un calcul de distance entre les représentations structurelles des documents d'une part et de la requête d'autre part. L'objectif est alors de proposer les k documents dont les descriptions structurelles sont le plus en adéquation avec le graphe requête. En effet, la notion de structure d'un document étant sujette à différentes interprétations suivant l'utilisateur, il est important de pouvoir proposer à l'utilisateur de faire le choix final. Ces informations sont presque toujours représentées sous forme de graphes. On trouve d'ailleurs beaucoup de méthodes cherchant à valuer un graphe pour obtenir une représentation de ce type. Cependant, le calcul d'une distance graphe à graphe relève d'un problème NP-Complet. Cette complexité croît de façon exponentielle avec le nombre de nœuds et d'arcs. Cette complexité a souvent dissuadé de l'usage des graphes, mode de représentation pourtant apprécié en raison de son grand pouvoir d'expression.

Les travaux que nous présentons dans cet article visent à réduire le calcul de la complexité de la comparaison de représentations structurelles. Pour ce faire, plusieurs approches de la littérature extraient du graphe un vecteur de caractéristiques numériques encapsulant une partie de l'information topologique qu'il véhicule. La comparaison de graphes se ramène alors à un calcul de distance entre vecteurs dans un espace euclidien. De plus, dans le contexte d'une application de recherche d'information où les documents sont décrits par des informations structurelles, ce travail d'indexation peut être effectué hors-ligne. Quelques travaux abordent déjà cette question :

1. Une première méthode est présentée dans [LOP 03]. Lopresti et Wilfong proposent une représentation basée uniquement sur le degré des nœuds dans le cas de graphes non orientés et non étiquetés. Dans le cas où les graphes sont orientés, la représentation distingue les demi-degrés intérieur et extérieur. La représentation peut être adaptée pour prendre en considération les graphes étiquetés. Sa simplicité de mise en œuvre alliée à une comparaison de deux graphes réduite à un temps linéaire permet de retrouver des graphes topologiquement ressemblant dans la majorité des cas ;
2. Une seconde approche est présentée dans [BAR 05]. La représentation vectorielle extraite du graphe est construite en dénombrant dans le document les occurrences de symboles auxquels sont associés des sous-graphes fréquents dans le corpus. La représentation vectorielle est appelée "sac de symboles" par analogie à la représentation par sac de mots des documents textuels utilisée en recherche d'information. Ce mode de représentation nécessite que soit tout d'abord extrait un lexique de symboles par recherche de sous-graphes fréquents. Cette notion de fréquence rend le lexique dépendant du corpus utilisé. Ainsi, un même document n'aura pas la même description selon le corpus dont il est extrait.

La caractérisation d'un graphe par un vecteur apporte, dans le cadre décrit ci-dessus, des avantages, mais apporte également quelques contraintes. Par exemple, aussi performante (en complexité) soit elle, la description proposée par Lopresti et Wilfong ([LOP 03]) n'est pas bijective. Ainsi, deux graphes non isomorphes peuvent avoir des descriptions vectorielles identiques pour lesquelles la distance est donc nulle. Cette ambiguïté est dûe à la méthode de construction du vecteur. Celle-ci est basée uniquement sur le degré des nœuds. Or, il peut exister plusieurs configurations d'un ensemble de nœuds donnant la même signature mais correspondant à des graphes différents. Cette description n'intègre donc que de façon superficielle les informations concernant la topologie du graphe.

La description en sac de symboles proposée par Barbu [BAR 05] nécessite une forte connaissance du domaine, celle du lexique des symboles. Or, s'il offre la possibilité de constituer automatiquement un tel lexique via une recherche des sous-graphes fréquents, ce dispositif est inopérant lorsque l'ensemble des documents présente une forte hétérogénéité.

Notre approche se positionne dans la continuité de ces travaux. L'idée de base est de contruire une représentation vectorielle dénombrant les occurrences de motifs présents dans le graphe à décrire. Or, en présence d'un corpus hétérogène, il est impossible d'extraire la connaissance du lexique en utilisant un critère de fréquence. Par ailleurs, l'utilisation d'un lexique générique présente l'avantage de donner une représentation unique d'un document indépendament du corpus dont il est extrait.

Dans la section suivante nous présentons le lexique que nous avons choisi d'utiliser et la façon dont il peut être construit. La section 3 présente le processus de construction de la représentation vectorielle d'un graphe. La section 4 présente deux distances exploitant cette description vectorielle ; ces deux distances correspondent à deux cas d'usage différents. La section 5 présente les premières expérimentations montrant que cette représentation offre des performances équivalentes à celle de Lopresti et Wilfong pour l'approximation de calcul de distance entre graphes dans un contexte de classification. Enfin, la section 6 dresse un bilan et énonce un certain nombre de perspectives pour la poursuite de ces travaux.

2 Construction du lexique

Comme décrit précédemment, le lexique sert de base à la construction de la signature vectorielle d'un graphe. La composition de ce lexique est donc déterminante dans la pertinence de la représentation vectorielle. Nous avons vu ([BAR 05]) qu'il était envisageable de le constituer à partir des n sous-graphes les plus occurents dans la base par exemple. Cependant, ceci n'est seulement possible que dans le cas d'une base présentant une homogénéité forte. La fréquence est révélatrice d'une certaine sémantique. Ainsi, dans les travaux de Barbu, les sous-graphes fréquents sont associés à des symboles graphiques, entités porteuses de sens dans les documents techniques.

Il existe malheureusement beaucoup de cas où ces hypothèses ne sont pas vérifiées. On peut prendre par exemple, les documents anciens où les ouvrages sont différents suivant l'auteur, l'éditeur, la date... Par conséquent, afin de conserver un caractère générique, il nous semble judicieux d'opter pour un lexique totalement indépendant des graphes présents dans la base. Il faut néanmoins que ce lexique soit suffisamment complet afin que ces termes puissent permettre de discriminer un graphe d'un autre.

Nous avons donc décidé de prendre comme référence le réseau des sous-graphes isomorphes, présenté dans [JAR 92]. Ce réseau présente l'ensemble des graphes composés de n arcs jusqu'au rang N (N étant le nombre d'arcs maximum) de manière totalement exhaustive. Ce réseau est construit de façon itérative à partir du graphe constitué d'un unique nœud. À chaque itération il est possible de construire un graphe de rang n en ajoutant un arc à un graphe de rang $n-1$ avec au besoin l'ajout d'un nœud supplémentaire. Toutes les solutions sont envisagées ce qui rend le réseau complet. Le graphe de rang n issu d'un graphe de rang $n-1$ est appelé successeur. Réciproquement, le graphe de rang $n-1$ est appelé prédécesseur. Un graphe de ce réseau peut avoir plusieurs successeurs. De même, plusieurs graphes de rang

Rang	Taille
0	1
1	2
2	3
3	6
4	11
5	23
6	51
7	117
8	276
9	669
10	1629

TAB. 1 – Effectif du lexique en fonction du rang des graphes non isomorphes

$n-1$ peuvent donner lieu à un même successeur. Les chemins de construction de ce réseau des graphes non isomorphes peuvent être conservés pour facilement reconsituer l'ensemble des prédécesseurs et des successeurs d'un graphe donné.

Par la suite, le terme *motif* désignera un sous-graphe élément du réseau des graphes non-isomorphes. Le lexique à la base de la représentation vectorielle des graphes est constitué de l'ensemble des motifs jusqu'au rang défini.

Par exemple, la figure 1 illustre le réseau des graphes non-isomorphes jusqu'au rang 4 donnant lieu à un lexique de 11 motifs. Les flèches en pointillé indiquent les chemins de construction du réseau, les flèches étant orientées du prédécesseur vers le successeur.

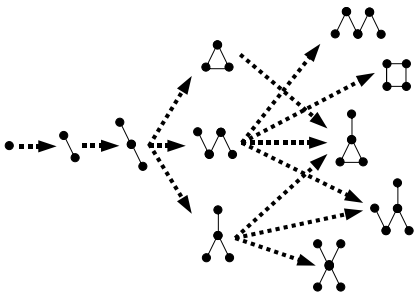


FIG. 1 – Réseau des sous-graphes isomorphes

Le tableau 1 donne le nombre d'éléments du lexique en fonction du rang maximal du réseau des graphes non-isomorphes.

Nous pouvons remarquer que le nombre de motifs augmente de façon exponentielle avec le rang. La taille du lexique est donc un paramètre à fixer suivant plusieurs critères. En effet, la complexité du passage en représentation vectorielle est directement dépendante du nombre de motifs. Cependant, plus la taille du lexique augmente, plus il intègre des motifs d'ordre important, la représentation vectorielle en découlant intègre alors davantage d'informations sur la topologie. Il s'agit donc de trouver un compromis entre l'expressivité et la complexité de la représentation.

3 La construction de la représentation vectorielle

La construction du vecteur consiste à déterminer la fréquence de chacun des motifs du lexique présents dans le graphe à décrire. La dimension du vecteur correspond à la taille du lexique. Sa construction peut alors devenir très coûteuse en temps. Cependant la phase d'indexation, qui consiste à contruire la représentation vectorielle de tous les graphes de la base, peut être effectuée hors-ligne. La complexité de construction du vecteur n'est uniquement critique que lors du traitement du graphe requête. En effet, ce graphe peut être obtenu à partir d'un document exemple présenté par un utilisateur. La méthode d'extraction de la représentation sera identique à celle utilisée pour la base et pourra donc être coûteuse.

Le lexique étant trié en suivant l'ordre du réseau des sous-graphes, la première valeur du vecteur décrivant un graphe correspondra alors au nombre de nœuds, la deuxième au nombre d'arcs, la troisième au nombre de sous-graphes présentant deux arcs,...

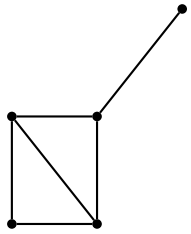
La recherche des motifs peut se faire suivant différentes contraintes. La comptabilisation (ou non) de motifs partageant un ou plusieurs éléments (nœuds ou arcs) influe directement sur la représentation vectorielle issue d'un même graphe. L'utilisation de ce type de contraintes peut se justifier par la volonté de trouver une bijection plus juste entre le graphe et sa représentation vectorielle. En effet, pour que la description vectorielle soit au plus proche du graphe, un motif découvert dans un graphe doit être retiré de celui-ci. Les éléments (arcs et/ou nœuds) qui le composent ne peuvent donc pas appartenir à une autre occurence du même motif. Cependant, la complexité de l'extraction de la représentation vectorielle avec ces contraintes augmente. Suivant les cas d'usage, il n'est toutefois pas nécessaire d'appliquer de telles contraintes. Nous détaillerons ce point dans la section suivante.

En guise d'exemple didactique, la figure 2 représente la description vectorielle associée à un graphe simple non-orienté et non-étiqueté (Fig.2(a)) pour un lexique de taille 6 (ordre maximum des motifs : 3) sans contrainte (Fig. 2(b)), avec la contrainte des arcs disjoints (Fig. 2(c)) et avec celle des noeuds disjoints (Fig. 2(d)).

Dans la section suivante, nous présentons deux possibilités d'exploiter la description vectorielle obtenue. Nous pourrons donc montrer l'avantage de pouvoir décliner la représentation suivant l'utilisation souhaitée.

4 Exemples de mesures en fonction du cas d'usage

Un calcul de distance dans l'espace des graphes permet de quantifier la différence entre 2 graphes. Cependant, la complexité lié au calcul de distance entre graphes interdit son utilisation dans un cas de recherche d'information où il s'agit d'ordonner tous les documents de la base (ou les k plus proches) en fonction de leur proximité à une requête. La représentation vectorielle que nous proposons va permettre d'approximer cette distance par une mesure de dissimilarité



(a) Graphe G

Motif						
Fréq.	5	6	10	2	10	3

(b) Nombre d'occurrences de chacun des motifs dans G

Motif						
Fréq.	5	6	3	1	1	1

(c) Nombre d'occurrences de chacun des motifs dans G ne présentant aucun arcs communs

Motif						
Fréq.	5	2	1	1	1	1

(d) Nombre d'occurrences de chacun des motifs dans G ne présentant aucun nœuds communs

FIG. 2 – Un graphe simple et ses descriptions vectorielles

entre les représentations structurelles des documents d'une part, et une requête également exprimée de façon structurelle, d'autre part. Si l'extraction de cette représentation vectorielle requiert un coût important, ce travail d'indexation peut être effectué hors-ligne et peut donc être toléré.

Les différentes dimensions de la représentation vectorielle que nous proposons présentent une certaine redondance. En effet, si un motif d'ordre n est dénombré, tous ses prédécesseurs sont également intégrés dans le décompte. Cette redondance correspond à l'intégration d'une certaine robustesse dont nous avons souhaité doter notre représentation. Il nous a semblé pertinent de pouvoir prendre en considération les perturbations susceptibles d'intervenir sur les représentations structurelles souvent extraites automatiquement : apparition ou disparition de nœuds ou d'arcs. Ainsi, si deux graphes identiques ont des descriptions vectorielles strictement identiques, deux graphes dont un serait une version bruitée de l'autre ont au moins des motifs d'ordre inférieur en commun.

Enfin, nous justifions notre proposition de description vectorielle par le fait qu'elle peut être utilisée dans deux cas d'usage distincts. Dans le premier usage, il s'agit de trouver dans la base de graphes indexée les plus proches du graphe requête. Une seconde application consiste à trouver les graphes de la base contenant le plus grand nombre d'occurences du graphe requête. Les deux sous-sections qui suivent présentent les mesures exploitant notre description vectorielle et s'appliquant à ces deux cas d'usage.

4.1 Mesure de dissimilarité graphe à graphe

Dans ce cas d'usage, il s'agit de trouver, parmi les graphes d'une base, ceux se rapprochant le plus d'un graphe requête. Les graphes de la base et le graphe requête sont représentés par un vecteur de caractéristiques numériques. Il s'agit alors d'approximer les distances entre les graphes de la base d'une part et le graphe requête d'autre part par une mesure de dissimilarité. Cette mesure de dissimilarité correspond à la distance entre les représentations vectorielles des graphes.

Cette problématique a déjà été soulevée dans la littérature. Nous avons choisi d'utiliser une distance euclidienne. Dans un espace à n dimensions, la distance s'exprime sous cette forme :

$$D(G_1, G_2) = \sum_{i=1}^{N} \sqrt{(k_{1i} - k_{2i})^2}$$

avec G_1 et G_2 les 2 graphes à comparer, k_{i1} et k_{i2} les occurences des motifs i dans G_1 et G_2.

Suivant le cadre applicatif, des évolutions sur cette distance sont possibles. Dans la recherche de documents techniques, par exemple, les experts fournissent des connaissances *a priori* sur la sémantique des symboles présents dans le plan. Cela peut permettre de "sanctionner" ou de "primer" la présence ou l'absence d'un symbole. On peut bien sûr élargir notre démarche à tous types de documents où une connaissance suffisante permet de juger de la pertinence des motifs. Celà se traduit par la pondération α_i de chaque motif i.

$$D(G_1, G_2) = \sum_{i=1}^{N} \sqrt{\alpha_i (k_{1i} - k_{2i})^2}$$

Les pondérations α_i peuvent alors être déterminées par des algorithmes d'optimisation ou d'apprentissage artificiel sur des cas d'usage précis. De même, si cette représentation vectorielle est utlisée sur une base présentant une certaine homogénéité, il est possible d'appliquer des techniques de sélection de caractéristiques visant la réduction de la dimensionnalité. Le but de cette réduction du vecteur permettra d'améliorer les performances en supprimant des motifs aberrants ou inutiles. Il est aussi possible de combiner des motifs complémentaires. Il existe dans la littérature plusieurs méthodes, telles que l'analyse en composantes principales ou toute autre sélection de caractéristiques.

Nous remarquerons que dans ce cas d'usage, l'utilisation des représentations sans contrainte apporte une redondance de l'information permettant d'augmenter la précision de la représentation vectorielle. Nous pouvons la considérer comme un ensemble de caractéristiques nous permettant aussi de faire de la classification, du clustering,...

4.2 Recherche d'occurences du graphe requête

D'autres cas d'application ne nécessitent pas de mesurer une distance entre deux vecteurs pour quantifier les similarités entre eux, mais de rechercher la présence et le nombre d'occurences du graphe requête dans un graphe, voire même dans une base préindexée.

On peut par exemple vouloir retrouver un schéma électrique contenant un composant spécifique. Il s'agit donc de

retrouver le nombre u d'occurrences de S, le graphe requête, à l'intérieur du graphe G.

On note V_G la représentation vectorielle de G et V_S celle de S. On note v_{Gi} (respectivement v_{Si}) le nombre d'occurrences du motif i dans G (respectivement dans S).

$$V_G = \begin{bmatrix} v_{G1} \\ \vdots \\ v_{GN} \end{bmatrix}$$

$$V_S = \begin{bmatrix} v_{S1} \\ \vdots \\ v_{SN} \end{bmatrix}$$

On a alors

$$\forall i \in \mathbb{N}, 1 \leq i \leq N, v_{Gi} \geq u.v_{Si}$$

En d'autres mots, si un graphe G inclut u occurences d'un graphe requête S alors chaque sous-graphe de S est au moins présent u fois dans G.

Ce raisonnement peut être tenu pour tous les sous-graphes de S et en particulier pour les motifs du lexique. Ainsi pour toutes les dimensions i de la représentation vectorielle on a

$$\forall i \in \mathbb{N}, 1 \leq i \leq N, u \leq u_i = \frac{v_{Gi}}{v_{Si}} \text{ avec } v_{Si} \neq 0$$

Par conséquent,

$$u \leq \min_{i/v_{Si} \neq 0} \left(\frac{v_{Gi}}{v_{Si}} \right)$$

Il s'ensuit que la quantité $\min_{i/v_{Si} \neq 0} \left(\frac{v_{Gi}}{v_{Si}} \right)$ peut être utilisée pour approximer u, le nombre d'occurrences de S dans G.

Il faut noter que dans ce cas d'usage, il faut utiliser une représentation vectorielle extraite avec de fortes contraintes, comme nous l'avons vu dans le troisième paragraphe. En effet, la recherche des occurences d'un graphe sous-entend souvent que toutes les occurences doivent être présentes intégralement dans le graphe. Elles ne peuvent donc pas partager d'arcs ou de nœuds. Le choix de la contrainte dépend du contexte et nécessite alors les connaissances *a priori* d'un expert.

5 Expérimentations

Nous décrivons dans cette section les premières expérimentations relatives à la description vectorielle de graphes que nous avons menées. Lors de ces tests, notre description vectorielle de graphe est comparée à la description vectorielle proposée par Lopresti et Wilfong.

Nous nous sommes limités pour les deux descriptions vectorielles à 6 dimensions. Ainsi, la description vectorielle de Lopresti et Wilfong dénombre dans les graphes les nœuds de degré 1 à 6 et notre description dénombre les motifs disposants de 3 arcs ou moins.

Les deux descriptions sont comparées sur une tâche de classification par plus proches voisins en utilisant la technique dite "leave one out", c'est-à-dire que chaque élément est classé en considérant que tous les autres appartiennent à la base d'apprentissage.

Une première comparaison a été effectuée sur une base de graphes synthétiques. Dans cette base, à chaque classe est associée un modèle de génération aléatoire disposant de deux paramètres que sont n le nombre de nœuds du graphe et d le degré moyen des nœuds. 20 classes ont été générées pour n prenant les valeurs 5, 10, 20, 50 et 100 et pour d prenant les valeurs 0.2, 0.5, 1 et 2. Ainsi, 20 graphes ont été générés pour chacune des classes, soit une base de 400 graphes.

Le taux de bonne classification utilisant la description de Lopresti et Wilfong atteint 99.45%. Seuls quelques confusions sont intevenues entre les classes de graphes C(n=5, d=0.2) et C(n=5, d=0.5). La classification utilisant la description proposée donne des résultats strictement identiques, les mêmes confusions s'opérant sur les mêmes éléments.

Dans une seconde expérimentation, les graphes correspondent à des représentations structurelles de symboles issus du concours de reconnaissance de symboles GREC. Nous avons utilisé 10 classes de symboles ayant subi des rotations et des déformations vectorielles. Les représentations structurelles représentent des graphes d'adjacence de régions correspondant aux composantes connexes noires et blanches. Les classes contiennent entre 5 et 19 éléments par classe, l'effectif total de la base s'élevant à 88 éléments. La figure 3 illustre des exemples de symboles issus de la base.

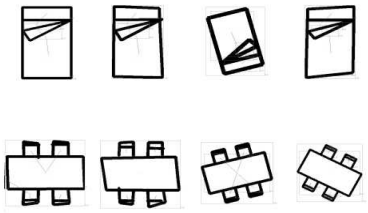


FIG. 3 – Exemple de symboles extraits de la base GREC

Le taux de bonne classification obtenu grâce à la representation de Lopresti et Wilfong s'élève à 51.14% alors que celui obtenu grâce à notre représentation atteint 53.41%.

Les bons taux de reconnaissance obtenus dans la première expérimentation doivent être relativisés en considérant que les modèles génératifs de graphes aléatoires tels qu'ils ont été définis distinguent les classes clairement. En effet, les classes pourraient être discriminées sans aucune confusion en considérant uniquement le nombre de nœuds et le nombre d'arcs.

Les taux de classification plus modestes observés sur l'application de reconnaissance de symboles doivent à l'inverse être relativisés en raison du fait que les représentations structurelles utilisées ne sont pas étiquetées et qu'elles ne reflètent que la topologie du graphe d'adjacence de régions sans que soit considérée la nature de la région.

Dans un cas comme dans l'autre, il est intéressant de

	Méthode de Lopresti & Wilfong	Notre Notre méthode
Graphes synthétiques	99.45%	99.45%
GREC	51.14%	53.41%

TAB. 2 – Tableau réapitulatif des premiers résultats obtenus

constater que les taux de classification obtenus avec les deux descriptions vectorielles sont comparables. Cela indique que la description que nous proposons peut être utilisée dans une recherche de plus proche voisin.

6 Conclusion et perspectives

Dans cet article, nous avons présenté une nouvelle description vectorielle des graphes visant à réduire la complexité du calcul de distance nécessité par des applications de recherche d'information en présence de données structurées. Cette représentation est proposée afin de pouvoir s'appliquer à deux cas d'usage différents :

1. La recherche de graphes similaires ;
2. La recherche de graphes contenant des occurences multiples d'un graphe requête.

Les résultats des premières expérimentations mettent en évidence que notre représentation vectorielle est équivalente en terme de performance à d'autres approches proposées dans la littérature pour approximer la distance entre graphes dans une tâche de classification.

Cependant, nous avons vu tout au long de cet article que certains points restent en suspend. Il semble maintenant important de mesurer l'influence de la taille du lexique sur la construction de la description et sa complexité. En effet, la figure 1 montre la nécessité de limiter la description à un rang faible pour réduire la complexité de sa construction. Les premiers tests ont toutefois prouvé que la pertinence de la description était meilleure avec un nombre de motifs élévés. Nos futurs travaux vont donc naturellement porter sur le compromis entre l'expressivité et la dimensionnalité de notre représentation.

Nous voulons aussi faire évoluer notre méthode. Les premiers tests ont été éffectués sur des graphes non-valués et non-attibués. Nous pensons cependant que cette approche peut s'adapter à tous types de graphes. Une de nos perspectives est de modifier la structure du vecteur pour y intégrer en plus des occurences les valeurs des attributs. Une autre solution serait d'utiliser la topologie comme un filtrage préliminaire réduisant le nombre de résultats possibles pour ensuite effectuer une analyse plus fine des étiquettes et attributs dans une base de taille réduite.

Afin d'évaluer la pertinence de notre représentation vectorielle dans un contexte de recherche d'information où on cherche à ordonner les graphes en fonction du nombre d'occurences du graphe requête nous avons utilisés des représentations structurelles d'images de documents anciens extraites à partir de la plateforme AGORA (cf. [RAM 06]). Il nous reste à constituer un jeu de requêtes avec pour chacune les documents pertinents pour pouvoir évaluer précisions et rappels pour les k premiers documents en fonction de la valeur de k.

Les premières expérimentations ont cependant montré que face à des représentations structurelles avec un grand nombre d'arcs, l'extraction de la représentation nécessitait un temps de calcul assez conséquent. Des optimisations algorithmiques devront être réalisées.

7 Remerciements

Les travaux décrit dans cet article ont été réalisés avec le soutien de l'ANR dans le cadre du projet NAVIDOMASS ANR-06-MDCA-12.

Références

[BAR 05] BARBU E., HÉROUX P., ADAM S., TRUPIN E., Clustering document images using a bag of symbols representation, *ICDAR*, pp. 1216-1220, 2005.

[JAR 92] JAROMCZYK J., TOUSSAINT G., Relative neighborhood graphs and their relatives, *Proceedings of the IEEE*, 1992.

[LOP 03] LOPRESTI D., WILFONG G., A fast technique for comparing graph representations with applications to performance evaluation, *Int. J. Doc. Anal. Recognit.*, vol. 6, n° 4, pp. 219–229, 2003, Springer-Verlag.

[RAM 06] RAMEL J. Y., BUSSON S., DEMONET M. L., AGORA : the Interactive Document Image Analysis Tool of the BVH Project, *DIAL '06 : Proceedings of the Second International Conference on Document Image Analysis for Libraries*, pp. 145–155, Washington, DC, USA, 2006, IEEE Computer Society.