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