Annotation sémantique de documents administratifs

Benjamin Duthil, Mickaël Coustaty, Vincent Courboulay, Jean-Marc Ogier

Laboratoire L3I, Avenue M. Crépeau, 17042 La Rochelle cedex 01
prénom.nom@univ-lr.fr,

Résumé. La numérisation de documents administratifs est un enjeu économique et écologique prioritaire dans le contexte sociétal actuel. La dématérialisation massive de document n'est pas sans conséquence et soulève les problèmes d'organisation, de stockage et d'accès à l'information. Le défi n'est donc plus la numérisation du document, mais l'extraction des informations qu'ils contiennent. Les documents sont produits par l'Homme et pour l'Homme. Cette propriété permet de localiser des informations dans les zones saillantes du document (logos). La saillance et la reconnaissance sont deux éléments essentiels pour la classification rapide de documents. A l'opposé, la recherche d'un document ou d'un ensemble de documents repose presque toujours sur le texte brut, il est donc nécessaire de faire une correspondance entre une requête textuelle et le document. Cet article présente une nouvelle approche d'annotation automatique de documents administratifs qui utilise une approche visuel et une approche de fouille de texte.

1 Introduction

La numérisation de documents administratifs est un enjeu économique et écologique prioritaire dans le contexte sociétal actuel. Le défi n'est plus la numérisation du document, mais l'extraction des informations qu'ils contiennent. Cet article présente une nouvelle approche d'annotation automatique de documents administratifs (certificat d'assurance, acte de naissance, etc.) qui utilise le logo contenu dans les documents comme élément d'apprentissage. Le logo est un élément graphique riche de sens (Duthil et al., 2013) auquel il est possible de rattacher de multiples informations (secteur d'activité, etc.). L'objectif ne se limite donc pas à la reconnaissance d'un logo dans un document (élément graphique) mais également aux aspects sémantiques connexes. Cet article est organisé de la manière suivante : la section 2 présente un état de l'art des méthodes existantes de classification et d'extraction de logo pour l'annotation. La section 3 présente notre approche d'annotation automatique de documents administratifs. La section 4 est consacrée aux expériences et la section 5 conclue cet article et donne quelques perspectives.

2 État de l'art

La détection et l'extraction de symboles ou de logos est un sujet de recherche actif de ces deux dernières décennies, comme l'atteste les très nombreuses publications réalisées dans les