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

conférences ICDAR ou GREC depuis 1995. Ces approches ont tout d'abord cherché à exploiter des images binaires, pour ensuite évoluer vers des images de documents en couleurs (Ahmed, 2008; Nourbakhsh et al., 2011). Nous proposons un bref résumé des méthodes dédiées aux logos ci-dessous.

En 2008, Zeggari (Ahmed, 2008) développé un algorithme d'extraction de logo reposant sur deux propriétés principales des logos : leur compacité spatiale et leur uniformité colorimétrique. Tout d'abord, le contenu de l'image est simplifié et transformé à partir d'opérateur morphologiques pour uniformiser les logos appartenant à une même classe. Puis, la densité spatiale et chromatique des régions composant chaque logo sont calculés.

Une approche intéressante, proposée en 2012 par Sahbi et al (Sahbi et al., 2012), vise à définir un noyau sur la "similarité liée au contexte", qui intègre le contexte spatial de caractéristiques locales. Les points d'intérêts sont détectés à l'aide du détecteur *SIFT*, et le contexte est décrit à l'aide d'un *shape-context*. Une fonction mesurant la similarité est alors définie en s'appuyant sur trois critères principaux : la fidélité, le contexte et l'entropie du terme, pour rechercher des points d'intérêts similaires selon ces critères.

La plupart de ces méthodes présentent des restrictions particulières,soit elles reposent sur des heuristiques à priori, ou alors sont très coûteuses en temps de calcul. De plus, hormis la dernière approche citée, toutes les autres reposent sur des critères bas-niveau (descripteurs radiométriques) sans intégrer le contexte ni conceptualiser le contenu d'une image. L'objectif de cet article est de proposer une méthode applicable à la fois sur les images en couleurs et en noir et blanc, avec des documents de bonne qualité ou bruités, et capable d'intégrer un apprentissage incrémental (apprentissage de nouveau logo à la volée). Enfin, nous proposons une méthode qui combine des traitements d'images et des techniques de fouille de texte afin d'annoter un document par un ensemble de mots clés caractéristiques des différents concepts contenus dans le documents.

3 L'approche

Notre approche d'annotation de documents administratifs est composée de quatre étapes. La première étape consiste à extraire automatiquement les zones saillantesdu document afin d'extraire et d'identifier le(s) logo(s) contenu(s) dans le document. La seconde étape permet la construction automatique d'un corpus d'apprentissage en utilisant les vignette identifiées précédemment. La troisième étape sert d'apprentissage du vocabulaire relatif au logo. La dernière étape correspond à l'annotation du document par le lexique de descripteurs appris lors de l'étape précédente.

3.1 Extraction de logo

Un logo est une région particulière d'un document qui peut-être définie par les trois propriétés suivantes (Duthil et al., 2013) : Région visuellement saillante par rapport à son contexte, récurrent visuellement (même image dans différent contextes), élément graphique créé par l'Homme pour l'Homme. L'objectif de cette première étape est donc de rechercher et d'extraire des zones saillantes dans un document, zones qui seront analysées par la suite pour identifier les logos potentiellement présents. Pour réaliser cette étape, nous nous sommes appuyés sur l'approche proposée par (Perreira Da Silva et Courboulay, 2012) pour l'analyse visuel du

document. Ce modèle hybride permet d'étudier l'évolution temporelle du focus attentionel. Le système visuel est inspiré des modèles d'attention de Itti et al. (1998) et Frintrop (2005). La scène visuelle est décomposée en différentes caractéristiques selon une approche multirésolution et ces caractéristiques sont calculées à partir de filtres numériques. Le système génère, pour chacune des caractéristiques prises en compte (intensité, couleur et orientation), un certain nombre de cartes représentant les éléments les plus saillants. À partir de chacune des vignettes identifiées par le processus de saillance visuelle, un processus d'identification des logos est effectué. Chaque vignette fait l'objet d'une requête sous forme d'image sur un moteur de recherche web (*Google Images*) pour identifier le nom du logo. Si la vignette n'est pas reconnue, elle ne sera plus considérée dans la suite du processus (fouille de texte), sinon, le nom du logo est conservé et il sera ensuite utilisé lors de la phase d'apprentissage des descripteurs sous forme de "mot germe"(c.f. section 3.2.2).

3.2 Fouille de texte

Le processus de fouille de texte est composé des étapes 2 et 3. Cette section présente la méthode de construction automatique du corpus d'apprentissage et d'apprentissage automatique des descripteurs. L'approche *Synopsis* proposée par Duthil et al. ((Duthil et al., 2012)) entre dans ce cadre. D'une part, *Synopsis* nous permet de construire automatiquement un corpus d'apprentissage à partir de "mots germes", et d'autre part, l'approche nous permet un apprentissage automatique des descripteurs.

3.2.1 Construction du corpus d'apprentissage

La construction du corpus d'apprentissage a pour objectif d'obtenir des documents web qui ont un contenu similaire à la vignette requête. À chaque vignette saillante est associé un corpus de documents. Plus formellement, à chaque vignette q, q variant de 1 à k, k étant le nombre de vignettes identifiées, un corpus Doc^q de n documents est associé tel que $Doc^q = doc^q_n, n = 1 \dots n^q$.

3.2.2 Apprentissage des descripteurs

L'objectif de l'apprentissage est de construire un lexique de descripteurs textuels (mots) décrivant sémantiquement chacun des logos. À chaque logo est associé un lexique L^q. Pour faire le lien entre les éléments graphiques contenus dans le document et les différents concepts auxquels ils font référence, nous utilisons le corpus d'apprentissage constitué à l'étape 2. L'approche *Synospis* est principalement basée sur deux éléments clés : la notion de fenêtre et la notion de classe/anti-classe. La fenêtre permet d'effectuer un apprentissage des descripteurs tout en assurant leur cohérence sémantique avec le mot germe (Duthil et al., 2012). La notion de classe/anti-classe permet de filtrer le bruit web. Les noms communs et les noms propres sont les deux classes grammaticales apprises car elles sont reconnues comme porteuses de sens.

Plus formellement une fenêtre de taille sz centrée sur un mot germe g pour un document doc est définie par $F(g, sz, doc) = \{m \in doc / d^{doc}_{NC}(g, m) \leq sz\}$ où $d^{doc}_{NC}(g, m)$ est la distance entre le mot germe g et le mot m.

Le principe général est de calculer la représentativité (Duthil et al., 2012) d'un mot M dans chacune des deux classes (fréquence d'apparition normalisée $\rho(M)$ dans la classe (mots

présents dans les fenêtres) et dans l'anti-classe $\overline{\rho}(M)$ (mots en dehors des fenêtres). La représentativité d'un mot M dans chacune des classes est défini tel que :
$\rho(M) = \sum_{doc} \sum_{\gamma \in \mathcal{O}(g,doc)} |\mathcal{O}(M, F(\gamma, sz, doc))|$ et $\overline{\rho}(M) = \sum_{doc} |\mathcal{O}(M, \bigcap_{\gamma \in \mathcal{O}(g,doc)} \overline{F}(\gamma, sz, doc))|$

À partir de la représentativité d'un descripteur dans chacune des classes, il devient possible de déterminer la proximité sémantique du descripteur M considéré en appliquant une formule de discrimination f tel que cela est proposé dans (Duthil et al., 2012). Un score $Sc(M)$ est alors attribué à chaque descripteur. Chaque descripteur constitue une entrée du lexique L^q propre au logo q considéré. Le score d'un descripteur est calculé tel que : $Sc(M) = f^2(\rho(M), \overline{\rho}(M))$ où f est définie tel que : $f^2(x, y) = \frac{(x-y)^3}{(x+y)^2}$

3.2.3 Annotation des documents

L'étape d'annotation consiste à rattacher à un document l'ensemble des concepts qu'il contient (Lexique de mots précédemment construits). Annoter sémantiquement un document au format image revient à rechercher, et à identifier, les logos qu'il contient afin de lui rattacher les lexiques associés (sémantique). Annoter un document à partir de son contenu textuel (résultat d'OCR) consiste à identifier les concepts contenus dans le document. L'approche *Synopsis* permet d'identifier (segmentation), à partir d'un lexique, les zones du document qui traitent du concept (logo) considéré. L'approche utilise une fenêtre glissante (Duthil et al., 2012) centrée sur les noms communs pour identifier les segments de textes pertinents. Cette méthode nous permet également de connaître l'intensité ((Duthil et al., 2012)) du discours. À chaque document est associé un fichier xml qui contient un ensemble d'informations sémantiques : lexiques associés (identifiant du lexique correspondant au nom du logo), intensité du discours et l'importance du discours pour chacun des concepts (lexiques) qui ont été rattachés.

4 Expérimentations

Dans cette section nous évaluons la qualité de l'apprentissage sur un corpus de 1766 documents administratifs dans un contexte de classification. Ce corpus est composé de 4 classes de documents : acte de mariage (A-M), certificat d'assurance (C-A), relevé d'identité bancaire (RIB) et certificat de naissance (C-N). Le tableau 1 montre la répartition de ces quatre classes (Nombre de logos différents dans la classe et nombre de document de la classe). Cette base confidentielle provient d'un des leaders mondiaux de la dématérialisation documentaire. Chaque document contient un des 196 logos identifiés dans le corpus. Les documents sont scannés en 200 dpi noir et blanc. Nous utilisons les indicateurs classiques de mesure pour évaluer la classification : *Précision*, *Rappel*. La *Précision* est calculée en considérant les erreurs d'identification du logo : le système identifie un logo qui n'est pas le bon, le lexique associé ne correspond donc pas au logo à identifier. Le *Rappel* est calculé en utilisant le nombre de logos correctement identifié par le système. Nous utilisons durant tout le processus comme moteur de recherche web *Google Images*. Nos résultats sont résumés dans le tableau 2.

Les résultats montre la pertinence du système. Les différences de résultats entre chacune des classes s'expliquent par la qualité des documents. Cependant, les résultats sont remarquables, nous obtenons un *Rappel* de ***80,6*** et une précision de ***100*** toutes classes confondues. La *Précision* (100) met en évidence la robustesse de l'approche.

	nombre de logos	**nombre de documents**
acte de mariage	36	40
certificat de naissance	113	169
certificat de naissance	30	822
RIB	15	735

TAB. 1 – *Répartition des classes*

	C-A	A-M	C-N	RIB	Toutes classes
Rappel	**95,5**	22,5	60,4	**71,7**	***80,6***
Précision	**100**	**100**	**100**	**100**	***100***

TAB. 2 – *Résultats de classification pour chaque classe : acte de marriage (A-M), certificat de naissance (C-N), certificat d'assurance (C-A), relevé d'identité banquaire (RIB)*

5 Conclusion et perspectives

Dans cet article nous avons présenté une nouvelle approche d'annotation de documents administratifs. L'annotation sémantique qui est proposée permet d'annoter un document sous deux angles : à partir du texte contenu dans le document et/ou du logo qu'il contient. Cette approche permet de réduire le fossé sémantique qu'il existe entre les données bas-niveau contenues dans une image (pixels) et leurs sens. Les perspectives sont nombreuses. Tout d'abord, les annotations fournies peuvent être utilisées par un système de recherche d'information. Dans un contexte de classification, les annotations permettent d'attaquer le problème en considérant l'image (logo) du document ou bien d'un point de vue textuel en utilisant les lexiques associés au document. Considérer l'annotation sémantique permettrait de raffiner la classification et ainsi, plutôt que de considérer uniquement comme classe de document le logo d'une entreprise, de pouvoir identifier de nouvelles classes de documents comme par exemple tous les document d'un secteur d'activité.

Références

Ahmed, Z. (2008). Logos extraction on picture documents using shape and color density. In *IEEE International Symposium on Industrial Electronics, ISIE*, pp. 2492–2496.

Duthil, B., M. Coustaty, V. Courboulay, et J.-M. Ogier (2013). Visual saliency and terminology extraction for document annotation. In *Proceedings of the 13th ACM Symposium on Document Engineering*.

Duthil, B., F. Trousset, G. Dray, P. Poncelet, et J. Montmain (2012). Extraction d'opinions appliquée à des critères. In Y. Lechevallier, G. Melançon, et B. Pinaud (Eds.), *EGC*, Volume RNTI-E-23 of *Revue des Nouvelles Technologies de l'Information*, pp. 483–488. Hermann-Éditions.

Frintrop, S. (2005). *VOCUS :AVisual Attention System for Object Detection and Goal-Directed Search*. Phd, University of Bonn.

Itti, L., C. Koch, E. Niebur, et Others (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*(11), 1254–1259.

Nourbakhsh, F., D. Karatzas, E. Valveny, et J. Llados (2011). Color Logo Detection and Retrieval in Document Collections. In *Ninth IAPR International Workshop on Graphics RECognition - GREC*.
Anglais

Perreira Da Silva, M. et V. Courboulay (2012). Implementation and evaluation of a computational model of attention for computer vision. In *Developing and Applying Biologically-Inspired Vision Systems : Interdisciplinary Concepts*, pp. 273–306. Hershey, Pennsylvania : IGI Global.

Sahbi, H., L. Ballan, G. Serra, et A. Del Bimbo (2012). Context-dependent logo matching and recognition. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society 22*(3), 1018–31.

Summary

Document dematerialization is a priority to economic and ecological issue in the current social context . The mass of digitalized documents is not without consequence and raises problems of organization, storage and access to information. The challenge is no longer scanning the document, but the extraction of information they contain. Documents are produced by human, for human. This property allows you to locate information in prominent areas of the document (logos). Salience and recognition are essential elements for the rapid classification of documents. In contrast, the search for a document or set of documents is almost always plain text, it is necessary to make a correspondence between a text query and the document. This paper presents a new approach to automatic annotation of documents that uses a visual approach and text-mining approach.