Big-data historique : modélisation de stratégies d'analyse de collections de document

Camille Guerry, Bertrand Coüasnon, Aurélie Lemaitre, Sebastien Adam

Big-data historique : modélisation de stratégies d'analyse de collections de document

Camille Guerry[1], Bertrand Coüasnon[1], Aurélie Lemaitre[1], and Sebastien Adam[2]

[1]Univ Rennes, CNRS, IRISA, [2]LITIS, université de Rouen

Type de soumission : jeune-chercheur

Mots clés : analyse d'images de documents, reconnaissance de structure de documents, collection, stratégie, analyse itérative, modélisation de connaissances, redondance d'informations, big data historique

Les travaux que nous présentons ici sont menés dans le cadre du projet ANR HBDEX en collaboration avec l'équipe DFIH de l'école d'économie de Paris et le LITIS (Rouen). Ce projet a pour but de faire progresser la compréhension des dynamiques des marchés financiers sur le long terme. C'est dans ce contexte que nous élaborons un système de reconnaissance automatique de collections de tableaux de cotations boursières provenant de différents marchés financiers de la fin du 19^{e} et du début du 20^{e} siècle. Les tableaux de cotations sont des documents qui étaient imprimés de manière journalière et qui recensent l'ensemble des titres boursiers échangés la veille sur le marché financier, les cours auxquels chacun de ces titres ont été échangés ainsi que diverses informations concernant ces titres. Les OCR et les outils de reconnaissance actuels, tel que l'on peut en trouver dans le commerce, ne permettent pas une reconnaissance suffisamment fiable pour que leurs résultats soient utilisés tels quels sans passer par une phase de correction manuelle fastidieuse. L'enjeu est donc de proposer un système de reconnaissance fiable qui minimise les interventions de l'utilisateur humain. Pour cela nous explorons différentes stratégies qui exploitent le contexte apporté par la collection comme par exemple la redondance d'informations et les règles financières entre différentes données. Nous nous sommes focalisés sur les documents issus du marché de la Coulisse de Paris. Une fois développée, la stratégie que nous élaborons pourra facilement s'adapter à d'autres collections de documents.

Les travaux effectués jusqu'à présent ont consisté à d'une part développer un premier prototype qui reconnait la structure des documents de type tableau de cotations et d'autre part à l'élaboration d'une stratégie de reconnaissance globale de la collection. Notre prototype permet de reconnaitre l'organisation structurelle des lignes de texte selon les colonnes des tableaux de cotations grâce à une description grammaticale écrite dans le langage EPF de la méthode générique DMOS-PI [1]. Les terminaux de notre grammaire sont : les lignes de texte (obtenues grâce à de l'apprentissage profond [3]) et les segments verticaux et horizontaux. Ce prototype est ensuite intégré dans un processus de reconnaissance global itératif. Chaque itération validera un type de données dans l'ordre hiérarchique donné par la structure des documents : colonnes, sections, titres boursiers, autres champs. Les itérations seront composées d'une première étape de reconnaissance structurelle pour extraire des éléments dans l'image, suivie d'une phase de validation transversale de l'information extraite. Nous effectuerons cette phase de validation grâce à la modélisation du contexte lié à la collection. Ce contexte nous est donné par les spécifications des documents produites par nos partenaires économistes. À partir de ces spécifications, nous avons pu énoncer et catégoriser les différentes règles applicables à la collection des documents de la Coulisse. Nous proposons une première manière d'exploiter ces différentes règles pour : apporter du contexte aux OCR utilisés, fiabiliser l'information extraite, permettre de préciser la description structurelle des documents et repérer les informations qui devront être vérifiées par un expert. Pour mettre en place notre stratégie, nous nous appuierons sur les travaux de thèse de Chazalon [2].

References

[1] Cöuasnon B. Dmos, a generic document recognition method: Application to table structure analysis in a general and in a specific way. IJDAR, 2006.

[2] Chazalon J. *Interprétation contextuelle et assistée de fonds d'archives numérisées : application à des registres de ventes du XVIIIe siècle.* PhD thesis, INSA de Rennes, 2013.

[3] Kaplan F. Oliveira S. A., Seguin B. dhsegment: A generic deep-learning approach for document segmentation. ICFHR, 2018.