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

Direction de thèse :

Bertrand COÜASNON, Irisa/Intuidoc, bertrand.couasnon@irisa.fr

Co-direction :

Aurélie LEMAITRE, Irisa/Intuidoc, aurelie.lemaitre@irisa.fr ,
Sébastien ADAM, LITIS, sebastien.adam@univ-rouen.fr

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

L'équipe de recherche Intuidoc de l'Irisa (http://www.irisa.fr/intuidoc) travaille notamment sur la reconnaissance du contenu et de la structure de documents anciens, manuscrits ou dégradés (partitions musicales, registres d'archives, journaux, courriers manuscrits, schémas électriques...). Dans ce contexte, les travaux de l'équipe ont abouti à des chaînes de traitement complètes (méthode DMOS-P [1]), conduisant à une reconnaissance de la structure des documents, décrite par l'introduction de grammaires bidimensionnelles permettant d'exprimer la connaissance visuelle sur les documents et des mécanismes perceptifs, qui reposent sur la reconnaissance d'éléments de base par l'intermédiaire de reconnaisseurs de symboles et d'écritures, l'ensemble étant validés sur plus de 700 000 documents et de nombreux types de documents différents. Ces travaux ont été évalués lors de différentes compétions internationales : RIMES 2007, RIMES 2008, ICDAR 2009, ICFHR 2010, ICDAR 2011, ICDAR 2013, HIP 2013, MAURDOR 2013.

Ce sujet de thèse s'inscrit dans le contexte du projet ANR HBDEX qui a été soumis pour l'appel à projet 2016-2017, accepté en juillet 2017. Ce projet a été soumis en partenariat avec l'équipe DFIH (Données Financières et Historiques) de l'Ecole d'Economie de Paris et le laboratoire LITIS qui travaille également sur l'analyse d'images. L'objectif de ce projet est de produire un système capable d'effectuer une reconnaissance automatique d'images de documents historiques : des tableaux de cotations boursières. L'enjeu est important au niveau des SHS, il s'agit d'obtenir un instrument numérique intelligent et collaboratif d'extraction de Big Data historiques par lecture automatique de données tabulées provenant de registres papier, afin de générer des bases de données financières.

La reconnaissance automatique de tels registres est complexe, et les systèmes actuels d'OCR commerciaux ne produisent pas des résultats de qualité suffisante. En revanche, les pages de documents contiennent des redondances d'informations entre pages successives (stabilité des noms des titres, règles de cohérences des valeurs d'une journée à l'autre). L'objectif du travail est donc d'exploiter ces règles de cohérence entre pages pour fiabiliser la reconnaissance des contenus.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

Plus précisément, l'objectif de cette thèse est de coordonner de manière intelligente la reconnaissance des contenus de pages d'une collection. Il s'agira de modéliser et de comparer des stratégies globales de lecture du corpus, afin d'exploiter de manière optimale les propriétés de la collection et les interactions avec les utilisateurs.

La stratégie sera basée sur l'organisation chronologique des documents, qui permet par exemple de mettre en avant : une redondance de la mise en forme de la structure tabulaire entre pages successives ; une redondance entre les champs textuels entre plusieurs jours de cotations ; une cohérence entre les champs numériques entre plusieurs jours de cotation.

Pour exploiter ces informations, ce système devra orchestrer :

- l'appel à un système de reconnaissance de la structure tabulaire de chaque page, en indiquant les propriétés des pages successives dans une collection. Le système devra ici transmettre au module de segmentation les informations provenant d'autres pages traitées, ou de l'interaction avec l'utilisateur;
- l'appel à des mécanismes d'interaction avec l'utilisateur lorsque ceci est requis par le système de segmentation/reconnaissance. Le rôle du système sera dans ce cadre de regrouper les demandes d'interventions similaires pour limiter le nombre de sollicitations de l'utilisateur.

Le système procédera à une construction par itérations successives d'une représentation cohérente et structurée des données que l'on cherche à produire.

Le système se basera sur des travaux précédents [2, 3, 4, 5, 6] qui ont permis d'exploiter une redondance textuelle entre pages d'une même collection dans un cadre applicatif de documents historiques.

L'enjeu scientifique est ici de chercher à produire une modélisation de la stratégie, de manière plus générique. Ceci permettra d'externaliser les connaissances liées à la stratégie. Cette modélisation devrait permettre de comparer facilement plusieurs approches. On pourra par exemple évaluer l'impact de la mise en place de règles de cohérences spécifiques. Cette modélisation de la stratégie devra également permettre de s'adapter à plusieurs corpus. Elle sera validée pour cela sur 3 corpus : "La Coulisse", les cotes du "Parquet de Paris" et les "Bourses de Province".

La modélisation de la stratégie devra être la plus simple et explicite possible, afin, dans l'idéal, de permettre à un non-informaticien de pouvoir décrire une nouvelle stratégie pour un nouveau corpus de document à traiter.

Modalités pratiques

Démarrage de la thèse : à partir de janvier 2018

Références

[1] Aurélie Lemaitre, Jean Camillerapp, Bertrand Coüasnon. **Multiresolution Cooperation Improves Document Structure Recognition**. International Journal on Document Analysis and Recognition (IJDAR), 11(2):97-109, Novembre 2008.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr

[2] Cérès Carton, Aurélie Lemaitre, Bertrand Coüasnon. **Eyes Wide Open: an interactive learning method for the design of rule-based systems**. In International Journal on Document Analysis and Recognition, Springer Verlag, 2017, 20 (2), pp.91-103.

[3] J. Chazalon, B. Coüasnon: **Iterative analysis of document collections enables efficient human-initiated interaction**, Document Recognition and Retrieval XIX, 2012.

[4] J. Chazalon, B. Coüasnon, A. Lemaitre: **A Simple And Uniform Way To Introduce Complimentary Asynchronous Interaction Models In An Existing Document Analysis System**, DAS, Pages 399-403, 2012.

[5] L. Guichard, J. Chazalon, B. Coüasnon: **Exploiting Collection Level for Improving Assisted Handwritten, Words Transcription of Historical Documents**, ICDAR, pp 875-879, 2011.

[6] Baptiste Poirriez, Aurélie Lemaitre, Bertrand Coüasnon. **Visual perception of unitary elements for layout analysis of unconstrained documents in heterogeneous databases**. In *14th International Conference on Frontiers in Handwriting Recognition (ICFHR-2014)*, 2014.

INSA RENNES
20, Avenue des Buttes de Cöesmes
CS 70839 - 35 708 Rennes Cedex 7
Tél.+ 33 (0)2 23 23 82 00 - Fax + 33 (0)2 23 23 83 96
www.insa-rennes.fr