Vers une plate-forme d'indexation interactive de documents

N. Journet, J.-Y. Ramel, C. Cureau

Vers une Plate-Forme d'Indexation Interactive de Documents

Journet Nicholas[1] - Ramel Jean-Yves[1] - Clément Cureau[2]

1 LI-RFAI, 64 Avenue Jean Portalis 37200 TOURS – France

{njournet,Jean-yves}.ramel@univ-tours.fr, clement.cureau@gmail.com

Résumé *:Cet article décrit comment le logiciel AGORA a été amélioré afin de tenir compte des retours d'usages obtenus lors d'une utilisation intensive de la version 1 du logiciel par les etudiants et les membres du CESR. La nouvelle version propose un module d'assistance aux utiilisateurs afin de rendre plus simple, conviviale et intuitive la phase de mise en place des scenarios d'indexation. L'utilisateur (non expert en traitement d'images) construit simplement des scénarios permettant d'étiqueter, de fusionner et de supprimer les éléments de contenus à indexer. Il localise ainsi les entités qui l'intéressent en ignorant les autres régions de l'image. Les scénarios élaborés peuvent ensuite être sauvegardés, modifiés et appliqués sur différentes images lors de traitements par lots.*

Mots-clés *: :***Indexation, Images de documents, Rétroconversion, Interactions, Agora**

1 Introduction

Les travaux présentés dans cet article ont été effectués dans le cadre d'une collaboration étroite avec le Centre d'Etudes Supérieures de la Renaissance (CESR) de Tours qui souhaite disposer d'outils d'indexation d'images afin d'alimenter rapidement le contenu de sa Bibliothèque Virtuelle Humaniste. Cet article décrit comment le logiciel AGORA a été amélioré afin de tenir compte des retours d'usages obtenus lors d'une utilisation intensive de la version 1 du logiciel par les étudiants et les membres du CESR.

2 Agora Version 2

2.1 Objectifs

De nombreux détails sur les concepts de base d'Agora sont décrits dans [RAM 06]. Cependant, aussi bien l'étude de l'existant faisant ressortir les difficultés d'utilisation des systèmes actuels [BOU 00], que les retours d'usages venant du CESR témoignent de la complexité pour un utilisateur non spécialiste en informatique, de manipuler des logiciels d'indexation d'images. En effet, il reste extrêmement compliqué pour un non spécialiste en informatique de faire le lien entre son expertise en ouvrages anciens et le traitement d'images. Comment paramétrer au plus juste ces outils et comment les combiner pour obtenir les résultats espérés par l'utilisateur ?

La solution proposée par Agorav2 pour résoudre ce problème, se base sur deux apports originaux. Tout d'abord, l'interface a été entièrement pensée pour des utilisateurs non spécialistes en informatique. Au travers de 4 étapes successives, l'utilisateur est entièrement guidé à travers de phases de prétraitement, segmentation et extraction d'éléments de contenu (EdC) et application de scénarios à des corpus entiers d'images de documents anciens.

Le deuxième apport d'Agora est son moteur d'apprentissage automatique de paramètres utilisé dans les règles de labellisation des EdC. Ainsi, plutôt que de laisser l'utilisateur définir lui-même les seuils et critères utilisés dans les scénarios, il va manipuler un module lui permettant de montrer les éléments de contenu qu'il souhaite extraire. Sur la base des exemples donnés par l'utilisateur, le moteur d'apprentissage va proposer des règles d'extraction à l'utilisateur.

2.2 Assitant Pré-traitements

Avant de pouvoir manipuler les EdC, un premier assistant basique permet aux usagers de créer leur projet d'indexation. Il s'agit simplement de sélectionner des images types utilisées par la suite pour construire le scenario d'analyse. Durant cette étape, l'utilisateur a également la possibilité de tester différents algorithmes de binarisation et de nettoyage d'images prédéfinis afin de sélectionner la configuration la plus adaptée aux caractéristiques des images à traiter. Pour cela, l'utilisateur lance les prétraitements sur les images types puis adaptent ou valide ces derniers en fonction des résultats obtenus sur les images types. A la fin de cette étape, la liste des EdC a été créée et peut être manipulée par l'utilisateur durant la phase suivante.

2.3 Assitant à l'extraction d'EdC

La phase suivante a pour objectif de permettre à l'utilisateur de spécifier simplement quels sont les éléments de contenu (EdC) qui l'intéresse et leurs caractéristiques discriminantes associées. Pour cela, l'utilisateur sélectionne dans son corpus, un échantillon d'images sur lesquelles il pourra effectuer son apprentissage. Ensuite, pour chaque élément de contenu

qu'il souhaite labelliser, l'utilisateur saisit à la souris, en sélectionnant l'élément de contenu sur l'image, plusieurs exemples de cet élément de contenu et se constitue ainsi son ensemble d'apprentissage. Sur la base de ces exemples, un scénario de labellisation pour l'élément de contenu est créé et appliqué au reste des images de la base. Le fait d'appliquer le scénario au reste des images de la base permet à l'utilisateur de visualiser la qualité du scénario automatiquement généré. Si ce scénario lui convient, ce dernier est validé. Dans le cas contraire, l'utilisateur peut manuellement modifier le scénario ou ajouter de nouveaux exemples dans la base d'apprentissage afin de générer une nouvelle version du scénario.

La génération automatique d'un scénario est basée sur l'étude des caractéristiques des éléments de contenu montrés par l'utilisateur. Pour chaque élément de contenu, on extrait des caractéristiques de forme, position et de niveaux de gris. C'est l'étude de la variabilité de ces caractéristiques qui permet de décider quelles caractéristiques utiliser pour constituer le scénario de labellisation. Ainsi, pour plusieurs exemples montrés par l'utilisateur on peut, pour chaque caractéristique extraire une valeur minimum, maximum, moyenne et enfin un écart type. On connaît ainsi la position, la taille, la densité de niveaux de gris minimum, maximum et moyen de l'élément de contenu que souhaite labelliser l'utilisateur.

La sélection des caractéristiques est réalisée selon l'heuristique que les caractéristiques stables d'un exemple sur l'autre sont celles qui sont les plus pertinentes. Ainsi, dans le cas où une caractéristique a un écart-type supérieur à un seuil alpha, alors cette caractéristique n'est pas retenue.

Si N exemples saisis et M caractéristiques extraites

$$Carac_{i \in M} \text{ sélectionnée si } Std_{j \in N}(Carac_i) < \alpha$$

Une fois la caractéristique retenue, il faut définir la plage de valeurs de cette caractéristique. Si le rapport largeur sur hauteur est par exemple retenu, il faut définir les valeurs minimums et maximums entre laquelle devra se situer un EdC à reconnaitre. Pour cela, un seuil beta permet de définir les limites des valeurs autour de la moyenne. Par défaut ce seuil beta est fixé à la valeur de l'écart type.

Valeur_Min_Carac_sélectionnée = moyenne - beta
et *Valeur_Max_Carac_sélectionnée = moyenne + beta*

Cette sélection est effectuée sur chacune des caractéristiques extraite. Une modification manuelle est toujours possible pour l'utilisateur. L'interface propose d'accéder à toutes les caractéristiques et fait varier, les seuils alpha et beta afin que l'utilisateur voit les changements que cela opère sur la création du scénario. D'un point de vu pratique, plus le seuil alpha est faible plus la caractéristique devra être stable pour être retenue dans le scénario final. Plus le seuil beta sera faible, plus la plage de valeurs de la caractéristique sélectionnée sera petite. Une fois le scénario d'indexation considéré comme pertinent et complet, l'utilisateur doit enfin définir quelles types d'information il désire obtenir en sortie : des imagettes des EdC sous forme binaire ou avec leurs couleurs d'origine, des coordonnées de chaque EdC extraits dans chaque page au format texte, XML ou HTML, des dictionnaires des formes similaires pour chaque type d'EdC, des extractions des lignes et mots dans les blocs de texte…

Cette étape termine le scénario qui peut ensuite être archivé et exécuté sur différents lots d'images lors de traitement en batch. Cette version d'Agora a été évaluée sur 300 pages de documents anciens, ce qui représente plus 5000 éléments de contenu à identifier.

Le tableau suivant résume les résultats obtenus sur la reconnaissance de 59 lettrines et 564 éléments de type « notes en marge ». Les chiffres indiques le nombre d'éléments reconnus en appliquant le scénario généré automatiquement (SA) ainsi que les résultats après modification manuelle du scénario généré automatiquement (SM).

	Bonnes détection		Fausses détection		Non détectés	
	SA	SM	SA	SM	SA	SM
Lettrines	59	59	16	1	0	0
Marges	473	551	0	0	94	13

3 Conclusion et perspectives

Nous avons présenté dans cet article la nouvelle version d'Agora qui est un logiciel d'indexation d'images de documents. Agora a la particularité de s'appuyer sur une approche interactive, permettant à un utilisateur non spécialiste en informatique de mettre en place des processus d'indexation d'images de documents. A court terme, et dans l'attente des futurs retours d'usages sur la nouvelle version d'AGORA, nous souhaitons orienter nos travaux de recherche vers l'évaluation de performances. Nous travaillons actuellement sur la mise en place d'un logiciel permettant de comparer les résultats obtenus par le module de labellisation d'Agora aux bases de vérité terrain UWII et PRIMA [ANT 07]. Enfin, une version d'Agora est disponible sur internet[1].

4 Bibliographie

[BOU 00] Bouché, Emptoz, Lebourgeois, and Metzger. Debora projet européen. Technical report, LIRIS,université de Lyon, 2000.

[RAM 06] J.Y. Ramel, S. Busson, and M.L. Demonet. Agora : the interactive document image analysis tool of the bvh project. DIAL, 0 :145–155, 2006.

[ANT 07] A. Antonacopoulos, B. Gatos, D. Bridson Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR2007), Curitiba, Brazil, September 2007, pp. 1279-1283.

[1] http://njournet.free.fr ou http://www.rfai.li.univ-tours.fr/pagesperso/ramel/fr/work1.html