Aide à la gestion des processus de numérisation en vue de l'OCRisation des ouvrages

Ahmed Ben Salah

LITIS Laboratoire d'informatique, du traitement de l'information et des systèmes
Bibliothèque nationale de France et Université de Rouen
Site François Mitterrand T1 07 14
Quai Francois-Mauriac 75706 Paris Cedex 13
Tel : 0153794164
Fax : 0153795045

ahmed.ben-salah@bnf.fr

RÉSUMÉ. *Dans cet article, nous étudions deux pistes afin d'améliorer le processus de numérisation des documents de la Bibliothèque nationale de France en vue de leur OCRisation. Dans la première partie, nous étudions les corrélations qui existent entre les données bibliographiques du document et les décisions de sélection des documents. Dans la deuxième partie, nous présentons une méthode basée sur la précision et le rappel qui va nous servir à estimer le taux de reconnaissance caractères pour vérifier les résultats de l'OCRisation sans recours à une vérité terrain. Nous présentons également un état de l'art des méthodes de segmentation dans le but de vérifier la qualité de celle issue de l'OCRisation.*

ABSTRACT. *In this paper, we investigate how to improve the digitization process at the French national Library. We propose in the first part a study on the relationship between the bibliographic data of the document and the selection decisions of the documents to help in this task. In the second part, we present an existing approach to estimate precision and recall without ground truth that could be used to estimate the OCR recognition rates. Finally, we present a short state of the art of segmentation methods that can help us to validate the quality of the segmentation.*

MOTS-CLÉS : *Analyse Factorielle des Correspondances, Analyse des Correspondances Multiples, Estimation de la précision et du rappel, Extraction des zones de texte.*

KEYWORDS: *Factorial Correspondence Analysis, Multiple Correspondence Analysis, Estimation of Precision and Recall, Text extraction.*

1. Introduction

Les projets de numérisation de masse à la Bibliothèque nationale de France (BnF) se heurtent à des difficultés notamment lors de deux étapes clés. Ils commencent par une étape de sélection manuelle des documents à numériser. Cette opération est très lourde et coûteuse puisqu'elle se base sur plusieurs critères physiques, documentaires et intellectuels que doivent posséder les ouvrages. Par ailleurs, à la fin du processus de numérisation, la BnF reçoit les textes numériques correspondant aux images du document avec une estimation du taux de reconnaissance caractère global. L'absence de vérité terrain rend le processus de contrôle de ces résultats très difficile et coûteux. Pour améliorer le processus de numérisation des documents à la BnF, nous avons réalisé une étude sur les relations entre les données bibliographiques et la décision de sélection du document, ceci pour essayer de prédire le résultat du processus de sélection sans passer par les critères physiques du document. Ensuite, nous avons commencé à développer une méthode de contrôle automatique de résultats de l'OCR sans vérité terrain en nous basant sur une méthode d'estimation de la précision et du rappel.
Cet article[1] s'articule donc autour de deux parties. Dans un premier temps nous présentons le processus de numérisation des documents à la BnF. Nous exposons alors notre étude qui concerne le problème de la prédiction de la décision issue du processus de sélection du document, notre objectif étant de proposer un outil permettant d'automatiser ou d'aider le service responsable de la sélection. Dans un second temps nous présentons une méthode proposée par Bart Lamiroy qui permet initialement d'estimer des taux de précision et de rappel sans vérité terrain et que nous pouvons utiliser pour estimer le taux de reconnaissance des textes. Nous présentons également quelques méthodes de traitement d'image qui vont nous servir à contrôler les résultats renvoyés par l'OCR de façon automatique.

2. Le processus de numérisation à la BnF

Le processus de numérisation à la BnF commence par une phase de programmation et de sélection des documents physiques à OCRiser. Puis tous les titres choisis sont envoyés aux prestataires de numérisation. Après la numérisation des documents, trois types de contrôle sont effectués :

– Contrôle automatique de structure de fichier XML,
– Contrôle de la qualité de l'image,
– Contrôle des métadonnées.

A la fin de la tâche de contrôle et selon le type et le nombre d'erreurs, le document est accepté ou renvoyé au prestataire pour une nouvelle OCRisation.

---

1. Ce travail a été encadré par : Thierry Paquet (Université de Rouen, LITIS), Geneviève Cron (Bibliothèque nationale de France) et Nicolas Ragot (Université François Rabelais Tour, LI). L'auteur est affilié à l'Université de Rouen.

3. Aide à la sélection des ouvrages pour OCRisation

La sélection des documents est une étape préliminaire au processus de numérisation et d'OCRisation des documents à la BnF. Notre étude a montré que le nombre des documents acceptés à cette étape (i.e. qui vont être effectivement OCRisés) ne dépasse pas 10% du nombre total des documents traités. Pour améliorer cette opération, nous avons étudié les relations qui existent entre les décisions de sélection de documents[2] et leurs données bibliographiques avec une *Analyse Factorielle des Correspondances (AFC)* et une *Analyse des Correspondances Multiples (ACM)*. Les modalités de la variable de décision de sélection sont regroupées en deux : (*Accepté* et *Refusé*). En effet, nous avons fusionné deux modalités car les documents *Validés en HQ* sont minoritaires. Nos analyses ont été appliquées sur un corpus de données homogène au niveau du *Format*[3]. Elle regroupe 20411 documents choisis de façon aléatoire. Les informations bibliographiques recueillies dans le corpus d'analyse sont : *Cotes, Date d'édition, Langue, Type de notice, Format, Code support physique, Type traitement.* Le choix de ces variables à étudier vient d'une part du constat des personnels de sélection qui affirment que les documents petits et anciens possèdent plus de défauts physiques que les autres documents, et d'autre part de la politique de numérisation de la BnF qui vise principalement les documents en Français.
L'AFC est une méthode statistique qui utilise en entrée un tableau croisé dynamique afin de déterminer la relation entre les deux variables croisées. Dans notre analyse nous avons construit deux tableaux de contingence qui croisent les *dates d'édition* et les *formats* de documents avec les décisions de sélection. D'après les *tests d'indépendance* et les mesures de $v - cramers$ [4] réalisés sur les deux tableaux de contingence (tableau des dates d'édition : $\chi^2 = 359,703$ et $V = 0.1370$; tableau des formats : $\chi^2 = 1313.304$ et $V = 0.25366$), les variables étudiées sont bien dépendantes de la décision de sélection puisque les valeurs χ^2 sont différentes de 0. Par contre, ces relations sont faibles puisque les valeurs de V sont faibles. Les résultats de l'AFC en eux-mêmes montrent que la modalité *Accepté* a un rôle important dans la construction des axes factoriels (96%). Nous avons remarqué aussi que le format 4 et les dates d'édition *1871-1930* et *1961-1990* ont plus de poids que les autres modalités dans la construction des axes factoriels. La supériorité du nombre des documents refusés par rapport au nombre des documents acceptés influence le calcul du modèle d'indépendance ce qui rend la représentation de la modalité *Refusé* voisine de l'origine du premier axe (*cf. figure 1 et 2*).

Etant donné que l'AFC ne montre pas de relations fortes entre les variables étudiées, nous avons également effectués une ACM qui analyse plusieurs variables qua-

---

2. les décisions de la sélection de documents à la BnF varient entre *Refusé*, *Validé en Brut* c'est-à-dire que le résultat de l'OCRisation est non corrigé et *Validé en Haute Qualité (HQ) : le texte résultant de l'OCR doit être corrigé par le prestataire avant livraison.*
3. La base d'étude est constituée par le même nombre de documents pour tous les formats possibles (*in-plano (GR FOL), in-folio (FOL), in-4°(4), in-8° (8) et in-16 (16)*) sauf pour le format " GR FOL " qui ne contient que 338 documents.
4. La valeur de $v - cramers$ calcule l'intensité de la relation qui lie les deux variables du tableau de contingence.

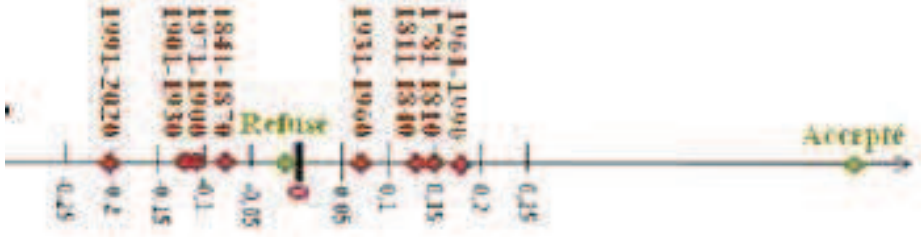

Figure 1. *Représentation du lien entre la décision de sélection et les dates d'éditions sur le premier axe factoriel*

Figure 2. *Représentation du lien entre la décision de sélection et le format sur le premier axe factoriel*

litatives en même temps. Nous avons utilisé un tableau de 5 variables qui sont : *Date, Langue, Pays, Format et Etat de sélection.* Les deux premier axes factoriels sont les plus importants en terme d'inertie cumulée, c'est pourquoi nous n'avons gardé ici que les deux premiers axes factoriels pour faire l'interprétation même s'il est intéressant d'interpréter les axes suivants. L'analyse montre que le premier axe sépare les documents en format *16* et *4* acceptés, des documents en format *8* refusés. De la même manière, le deuxième axe sépare les documents en formats *FOL* et *GR FOL* refusés, des documents en format *4* et *16* acceptés. Pour déterminer le lien entre la variable *état de sélection* et les autres variables, nous avons appliqué un test d'indépendance χ^2 sur nos données, comme pour l'AFC. Les résultats montrent que la variable *Format* a la plus petite probabilité critique et que donc elle est la plus liée à la variable *état de sélection.* Nous détaillerons plus notre travail dans (Salah *et al.*,, 2012).

A travers les AFC et l'ACM, nous constatons que la variable format joue un rôle important dans l'acceptation ou le refus des documents. D'autres variables comme la date d'édition jouent également un rôle mais dans tous les cas, ces variables bibliographiques ne suffisent pas à prédire la décision de sélection. En effet, d'une part les critères bibliographiques sont assez pauvres et d'autre part la décision de sélection dépend également de l'état physique du document[5] et de critères intellectuels.

5. Une base contenant ces informations est disponible mais elle n'est pas exploitable par les services de la BnF.

4. Contrôle qualité des résultats de l'OCR

4.1. *Estimation des taux de reconnaissance de caractères sans vérité terrain*

Suite à l'OCRisation, la BnF reçoit les textes numériques accompagnés d'un indicateur représentant le taux de reconnaissance des caractères sur tout l'ouvrage. Un des premiers problèmes auquel nous devons faire face pour contrôler la validité de ce résultat est l'absence de vérité terrain. Une piste pour estimer ce taux serait de considérer un OCR comme un outils de recherche d'information qui doit retrouver tous les caractères d'un ouvrage. Sa performance s'évalue alors en termes de précision et de rappel des caractères. Dans ce cadre, *Bart Lamiroy* et *Tao Sun* ont proposé dans (Lamiroy *et al.,*, n.d.) une méthode d'estimation de la précision et du rappel sans utilisation de vérité terrain. Les résultats obtenus sont très proches des valeurs réelles obtenues avec la vérité terrain. Le principe de cette méthode consiste à utiliser plusieurs systèmes de recherche d'information pour déterminer les fréquences d'apparition de chaque information recherchée. Ces fréquences servent par la suite pour l'estimation de la précision et du rappel. Nous pensons appliquer cette méthode à notre problème en utilisant plusieurs OCRs qui feront office d'outils de recherche d'information.

4.2. *Recherche des zones textes non détectées par les OCRs*

Les erreurs de segmentation sont fréquentes dans les résultats de l'OCR en qualité Brut. Ces erreurs biaisent le calcul du taux d'OCR annoncé dans le fichier ALTO[6]. Nous allons travailler sur les zones non décrites dans le fichier ALTO (i.e. les zones non segmentées) en leurs appliquant des méthodes de segmentation. Celles-ci sont divisées par (Journet *et al.,*, 2008) en deux familles. La première famille utilise un a priori sur la structure supposée de l'écriture pour segmenter les images. Parmi celles-ci, les méthodes de segmentation par découpage utilisent l'image entière de la page pour la découper récursivement en analysant plutôt les espaces que les traits. Ces approches nécessitent toujours un modèle, soit pour raffiner la segmentation de la page dans le cas des méthodes ascendantes, soit pour découper l'image de la page en plusieurs zones homogènes dans le cas des méthodes descendantes. Cette propriété rend leur utilisation dans notre contexte[7] inefficace.
Les méthodes qui utilisent la texture ne nécessitent pas de modèle de document pour extraire la structure physique du document. Il existe 4 familles principales pour les approches texture :

– **les méthodes statistiques** se basent sur l'analyse du niveau de gris des images de document pour décrire la texture de l'écriture. Parmi les méthodes qui appartiennent à cette famille nous trouvons le Grey Level Co-occurence Matrix (GLCM) proposé par Haralick dans (R.M. Haralick *et al.,*, November 1973) ;

---

6. ALTO est un fichier XML qui contient les résultats de l'OCR. Ce format est maintenu par le Library of Congress (http ://www.loc.gov).
7. Notre corpus d'étude est très variable en termes de dégradation physique des documents et règles éditoriales.

– **les méthodes à base de modèles probabilistes** se basent sur la construction d'un modèle qui permet non seulement de décrire une texture mais aussi d'en générer. D'après (Journet *et al.,*, 2008), les Champs de Markov et les fractales sont souvent utilisés dans cette catégorie ;

– **les méthodes fréquentielles** utilisent des primitives qui viennent du domaine du traitement du signal pour décrire la texture de l'écriture.[(Raju *et al.,*, 2005) et (Chan *et al.,*, 2001)].

– **les méthodes géométriques** caractérisent les formes qui constituent la texture et elles décrivent les relations spatiales qui les relient, (Tuceryan,, 1994).

Pour notre problème, nous allons appliquer une méthode géométrique d'analyse de texture pour vérifier la qualité de la segmentation de notre image. Cette approche utilise 3 fenêtres avec des tailles différentes 256×256, 128×128 et 64×64 et 6 angles (0°, 30°, 60°, 90°, 120° et 150°) pour extraire des indices de texture à plusieurs résolutions et orientations. Ce travail est actuellement en cours de réalisation et de test.

5. Conclusion et perspectives

Dans cet article nous avons présenté une étude sur les relations qui existent entre les données bibliographiques et les décisions de sélection des documents. Nous avons également présenté une méthode pour contrôler les résultats de l'OCR sans recours à une vérité terrain, ainsi que des méthodes de segmentation pour vérifier la qualité de la segmentation. Les perspectives de notre travail consisteront à développer une application de contrôle automatique de la qualité de l'OCRisation à partir de ces outils.

6. Bibliographie

Chan W., Coghill G. G., « Text analysis using local energy », *Pattern Recognition*, vol. 34, n° 12, p. 2523-2532, 2001.

Journet N., Ramel J.-Y., Eglin V., Mullot R., « Analyse d images de Documents Anciens :une Approche Texture », *Traitement du Signal (TS)*, vol. 24, n° 6, p. 461-479, 2008.

Lamiroy B., Sun T., « Precision and Recall Without Ground Truth », *Ninth IAPR International Workshop on Graphics RECognition - GREC 2011*, n.d.

Raju S. S., Pati P. B., Ramakrishnan A. G., « Text Localization and Extraction from Complex Color Images », *ISVC*, p. 486-493, 2005.

R.M. Haralick K. S., Dinstein I., « Textural features for image classification », *Pattern Recognition*, vol. 3, n° 6, p. 610-621, November 1973.

Salah A. B., Paquet T., Cron G., Ragot N., « Prediction of selection decision of document using bibliographic data at the national library of France », *ISetT's Archiving 2012 Conference-Copenhagen 2012*, Copenhagen, Denmark, p. à paraître, Juin, 2012.

Tuceryan M., « Moment-based texture segmentation », *Pattern Recognition Letters*, vol. 15, n° 7, p. 659-668, 1994.