Segmentation de flux de documents Application aux documents administratifs

Hani Daher, Abdel Belaïd, Vincent Poulain d'Andecy

Segmentation de flux de documents

Application aux documents administratifs

Hani Daher*, Abdel Belaïd et Vincent Poulain d'Andecy*****

**LORIA - **Université de Lorraine – LORIA, UMR 7503*
Vandoeuvre-Lès-Nancy, F-54506, France
****ITESOFT groupe, 30470, Aimargues, France*
{hani.daher, abdel.belaid}@loria.fr, Vincent.PoulaindAndecy@itesoft.com

RÉSUMÉ. *Cet article propose une approche de segmentation supervisée de flux de documents. L'approche traite le flux de documents comme une suite de paires de pages et étudie la relation qui existe entre elles pour déceler une continuité de documents ou une rupture. Dans un premier temps, des descripteurs sont extraits des pages et une approche est proposée pour fusionner ces descripteurs en un seul vecteur qui modélise la relation entre les paires de pages. Cette représentation est fournie à un classifieur binaire qui la classifie comme étant une rupture (synonyme de segmentation) ou une continuité. Dans le cas d'une rupture, nous considérons que nous avons atteint la limite d'un document complet et l'analyse du flux continue en commençant par un nouveau document. En cas d'une continuité, les deux pages sont considérées comme appartenant à un même document. S'il y a une incertitude sur la classe de la limite, un rejet est décidé et les pages analysées jusqu'à ce point sont considérées comme un « fragment » on réalise ici une sur-segmentation. Cette classification donne de bons résultats approchant 90% sur certains documents, ce qui est élevé à ce niveau du système.*

ABSTRACT. *This paper proposes a document flow supervised segmentation approach. Our algorithm treats the flow of documents as couples of consecutive pages and examines the relationship that exists between them in order to present a document continuity or rupture. In a first step, descriptors are extracted from the pages and an approach is proposed to merge these descriptors into a single vector that models the relationship between pairs of pages. This representation is provided to a binary classifier that classifies it as either a rupture (synonymous with segmentation) or continuity. In case of a rupture, we consider that the limit of a complete document has been reached and the stream analysis continues by starting a new document. In case of continuity, the two pages are considered to belong to the same document. If there is an uncertainty on the class of the limit, a rejection is decided and the pages analyzed until this point are considered as a "fragment" and an over-segmentation is applied. The classification provides good results approaching 90% on certain documents, which is high at this level of the system.*

MOTS-CLÉS : *Segmentation de flux de documents, descripteurs textuels, classification.*

KEYWORDS: *Document Flow segmentation, Textual descriptors, classification.*

1. Introduction

Chaque jour, différents types de documents sont traités par des organismes, comme des formulaires, des factures, des contrats, etc. Ces documents arrivent au niveau de la chaîne de traitement sous la forme d'un flux continu. La segmentation manuelle de ce flux pour isoler les documents est une tâche coûteuse en temps et est sujette à des erreurs. Une solution consiste à introduire des séparateurs de pages ou des marques lisibles par la machine, comme des codes-barres pour indiquer la fin du document. Dans le cas de séparateurs de pages, les marques doivent être insérées avant le parcours des pages et retirées par la suite, ce qui nécessite beaucoup de temps. Dans le cas des codes-barres, ceux-ci offrent une identification plus précise des documents, mais leur marquage a aussi un coût.

L'objectif de notre travail est de développer une approche automatique de segmentation de flux de documents sans aucune connaissance a priori sur le nombre de pages et les classes de documents et où chaque document représente un ensemble de pages successives bien ordonnées.

Le reste de l'article est organisé comme suit. Dans la section 1.1, nous présentons l'état de l'art. Dans la section 2, nous décrivons notre approche. Dans la section 3, nous détaillons les résultats des expérimentations. Enfin, dans la section 4, nous donnons une rapide conclusion et présentons quelques perspectives pour prolonger ce travail.

1.1. État de l'art

À notre connaissance, très peu de méthodes ont été proposées pour aborder ce problème de segmentation de flux et trouver des solutions. Dans notre recherche, nous avons identifié trois catégories de méthodes utilisées dans le traitement des documents :

- Segmentation de documents : Ces méthodes se basent sur la caractérisation des pages. L'objectif est d'identifier les limites entre les documents
- Recherche de documents : Ces méthodes se basent sur la caractérisation des documents. Les limites sont connues entre les documents. L'objectif est de chercher dans une base de données, les images les plus proches d'une image requête
- Classification de documents : Ces méthodes se basent sur la caractérisation des documents. Les limites sont connues entre les documents. L'objectif est d'affecter un document à une ou plusieurs classes

1.1.1. Segmentation de documents

Collins-Thompson et Nickolov (Collins-Thompson et Nickolov, *2002*) proposent une approche de calcul de similarité entre pages qui s'appuie sur des similitudes structurelles et textuelles. Les auteurs traitent la séparation de documents comme un problème de classification ascendante où chaque page est initialement considérée

comme un cluster, puis ils procèdent par étape, en fusionnant les paires de clusters par l'emploi d'un critère de lien unique. Cette méthode dépend énormément du résultat de l'OCR et de la bonne localisation des descripteurs structurels comme la boîte englobante du numéro de page. La méthode proposée par Meilender et Belaïd (Meilender et Belaïd, *2009*) est similaire aux méthodes utilisées dans la reconnaissance vocale, utilisant les n-grammes. La méthode consiste à séparer les pages consécutives qui ont une relation indépendantes entre elles et à regrouper les pages qui ont une relation de dépendance. Les relations de dépendance et d'indépendance sont modélisées à partir d'un algorithme progressif-rétrogressif (forward-backward) afin d'obtenir la meilleure segmentation. Etant donné que le calcul de la probabilité de toutes les pages du flux est *NP* complet, les auteurs utilisent des fenêtres glissantes de petites tailles pour segmenter le flux de documents. Les premiers résultats obtenus sur un flux homogène de documents ont produit plus de *75%* de précision et *90%* de rappel.

1.1.2. Recherche de documents

Rusiñol et al. (Rusiñol et al. *2012*) ont étudié différentes approches pour la recherche de documents multipages. Deux stratégies de fusion de pages sont proposées. La première consiste à représenter les pages par un histogramme qui correspond au nombre d'occurrences des mots dans l'ensemble des pages. La seconde consiste à calculer la similarité entre la page en cours et une page de la base; les pages consécutives affectées au même type de document sont finalement regroupées ensemble pour former un document dans le flux analysé. Kumar et Doermann (Kumar et Doermann, *2012*) présentent une approche basée sur les sacs de mots. Cette méthode s'appuie fortement sur la structure du document et est appliquée sur des documents mono-pages, ce qui n'est pas notre cas, où nous avons affaire à des documents multipages et où la structure donne peu d'informations sur les classes de documents. Shin et Doermann (Shin et Doermann, *2006*) segmentent les pages en blocs qui sont caractérisés par des descripteurs conceptuels et géométriques. La distance entre l'image requête et les autres images dans la base de données est obtenue en mettant en correspondance les blocs des images dont la distance entre leur vecteur descripteur est petite. L'inconvénient de cette méthode est que les descripteurs extraits sont très spécifiques à une classe de documents.

1.1.3. Classification de documents

Gordo et Perronnin (Gordo et Perronnin, *2010*) traitent le problème de la classification de documents multipages. Chaque document multipage est représenté par un histogramme qui correspond au nombre d'occurrences des types de pages (assurance, facture, etc.) qui forment ce dernier. Shin et al. (Shin et al. *2001*) se basent sur la structure du document pour extraire des descripteurs tels que les structures de colonnes, et le pourcentage du texte manuscrit et imprimé, afin de classifier des documents mono pages.

Notre méthode se situe dans la catégorie de la segmentation de flux de documents. Les autres catégories nous aident à chercher des descripteurs qui peuvent être utiles pour résoudre notre problème. En général, la majorité des approches vues précédemment, utilisent des méthodes basées sur des sacs de mots visuels et textuels pour classer les documents. Les méthodes de segmentation n'offrent pas une stratégie pour traiter les cas d'erreurs de segmentation. La majorité des méthodes montrent que les descripteurs textuels donnent de meilleurs résultats que les descripteurs visuels quand l'analyse se fait sur des documents qui contiennent des similarités au niveau du contenu que la structure, ce qui est le cas dans notre problématique.

2. Approche proposée

La Figure 1 illustre les trois principaux modules de l'approche proposée qui sont: l'extraction de descripteurs, la modélisation des relations entre pages successives et la classification. Toutes les images de documents sont OCR-isées et les mots vides sont écartés. Chaque page est présentée par un arbre XML. Un arbre est composé d'un ensemble de blocs. Chaque bloc comprend une séquence de lignes. Chaque ligne est composée d'une séquence de mots qui constituent les nœuds racines. Chaque bloc, ligne et mot, est identifié par les coordonnées de sa boîte englobante.

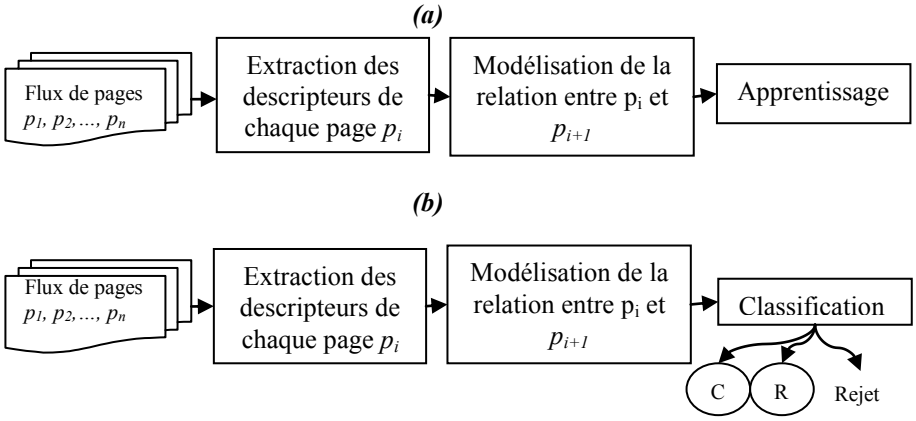


Figure 1. *Organigramme de la segmentation de flux de documents, basée sur la classification supervisée : (a) apprentissage, (b) test*

2.1. Flux de pages

Le flux sur lequel se fait l'analyse est composé de documents multipages hétérogènes (facture, feuille de soin, procès-verbal, etc.) formant au total *57* classes. La Figure 2 illustre un exemple de quatre classes de documents. L'ordre d'arrivée des

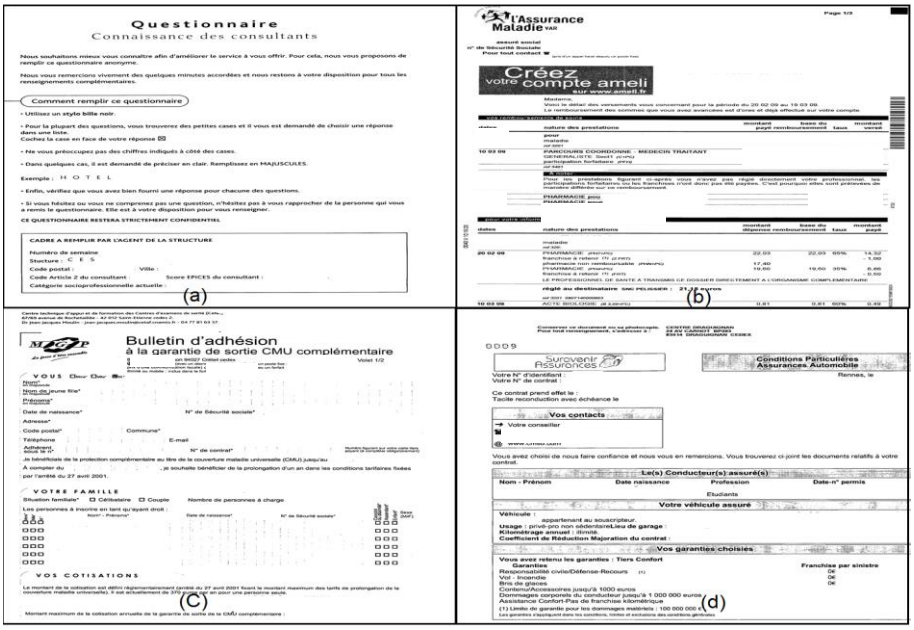


Figure 2. *Quatre exemples de documents : (a) Questionnaire, (b) Décompte, (c)Bulletin, (d) Contrat Assurance*

documents multipages dans le flux est aléatoire. Dans un même document, les pages sont toujours ordonnées, ce qui nous permet d'extraire des éléments de continuité comme les numéros de page, les items et les justifications afin de décider si deux pages successives dans le flux appartiennent à un même document. La Figure 3 montre deux documents successifs composés de deux pages. L'ordre des documents est aléatoire mais les pages à l'intérieur du document sont toujours ordonnées.

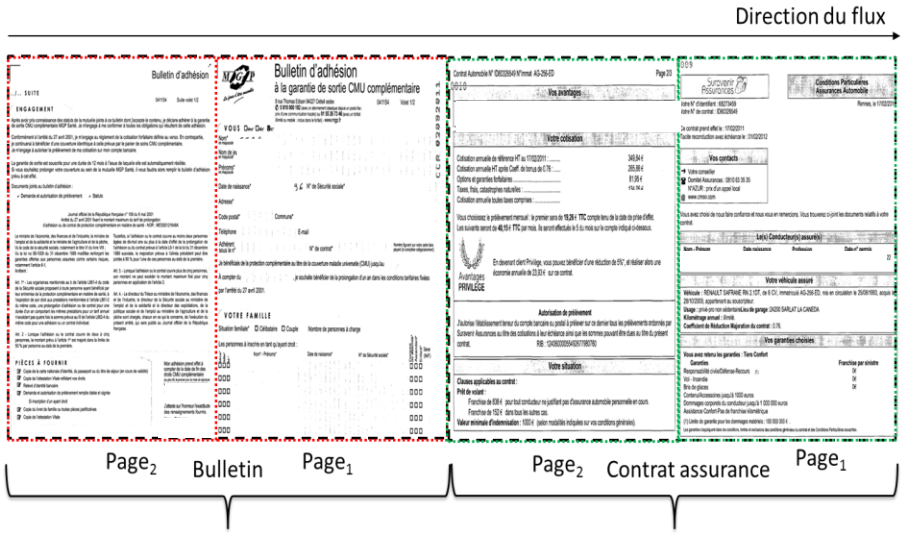


Figure 3. *Deux documents multipages successifs*

2.1.1. Variabilité de la structure des pages dans un document

La Figure 4 montre deux pages d'un même document. On voit que la structure des pages qui forment un document est variable. Il est difficile de décider si deux pages appartiennent au même document en se basant sur les descripteurs visuels. C'est à partir du numéro de référence, qui est un descripteur textuel, qu'on peut savoir que ces deux pages appartiennent au même document.

Figure 4. *Variation de la structure des pages d'un document*

Le flux de documents multipages à segmenter est formé de documents hétérogènes avec des pages ordonnées à l'intérieur de chaque document (voir Figure 4). Leurs structures sont très variées. Ce qui empêche d'utiliser un modèle générique pour les représenter. Pour toutes ces raisons, il n'est pas facile de modéliser un document dans sa globalité, et la solution trouvée sera de modéliser les paires de pages successives à l'aide de descripteurs de continuité et de rupture que nous définirons par la suite.

2.2. Extraction des descripteurs

Soit $P = p_1,...,p_n$ le flux de pages. Nous illustrons dans cette section les étapes que nous avons suivies afin de caractériser chaque page p_i dans le flux.

2.2.1. Classes des descripteurs

L'analyse du flux de documents a fait apparaître plusieurs types de descripteurs tels que les données de Fax, les dates, les codes, les numéros, les ID et les pages (voir Tableau 1). Chaque colonne du tableau indique les différentes instanciations du type dans les documents. Ces descripteurs nous renseignent par leur répétition d'une page à l'autre dans un même document, sur la continuité du document ou le changement d'un document au suivant.

Tableau 1. *Types de descripteurs extraits à partir des documents administratifs.*

Fax	Dates	Codes	Numérique	ID	Page
Date	Ré-édition	APE	Numéro de facture	Sinistre	Numéro
Numéro de fax	Assignation	Compte	Dossier	Global	Police
Numéro de page	Mission	Destinataire	Sécurité S.	Mission	Marge
Télécopieur	Dépôt	Ré-éditeur	Client	utilisateur	Item
Heure fax	Facture	OTC	Commande		Logo

2.2.2. Descripteurs retenus

Comme les instances de descripteurs sont nombreuses, difficiles à extraire, et peuvent encore s'élargir avec l'arrivée dans le flux de nouvelles classes, nous nous contentons d'extraire uniquement les valeurs des descripteurs quel que soit leur instance. Ceci nous permet de réduire les descripteurs à 9 entités (voir Tableau 2). Par exemple, pour le descripteur f_1 (date) représentera toutes les instances de dates écrites suivant un format général.

Tableau 2. *Description et format des descripteurs retenus*

Descripteur	Description	type
f_1	Date	Chaîne de caractères
f_2	Heure	Chaîne de caractères

f_3	Téléphone	chaîne numérique
f_4	Code postal	chaîne numérique = ***5***
f_5	Alphanumérique	chaîne alphanumérique
f_6	Numérique	chaîne numérique > ***6***
f_7	Salutation	Chaîne de caractères
f_8	Numéro de page	Numérique
f_9	Marge	Numérique

2.2.3. Extraction des descripteurs

Chaque descripteur parmi $f_1,...,f_7$ est extrait par une ou plusieurs expressions régulières qui traduisent ses formats. Ainsi, on se limite uniquement à la valeur à l'exception du numéro de page où un contexte est recherché pour le localiser efficacement. En effet, le numéro de page n'est pas facile à extraire tout seul car il peut apparaître comme un simple chiffre pouvant être confondu avec d'autres. Des contextes comme "page", "feuillet", "volet", etc. peuvent aider à son extraction. La Figure 5 illustre un exemple d'extraction de descripteurs par expressions régulières : f_1 est composé d'un ensemble de *4* dates, $f_1 = \{d_1, ..., d_4\}$. L'expression régulière n'a pas pu extraire de descripteur de type f_2, alors $f_2 = \emptyset$. Pour le numéro de page f_8. L'expression régulière trouve le contexte "Feuillet". La valeur *1* qui est située sur le côté droit du contexte est alors attribuée à f_8. Le descripteur f_9 représente la largeur du bloc de texte le plus large.

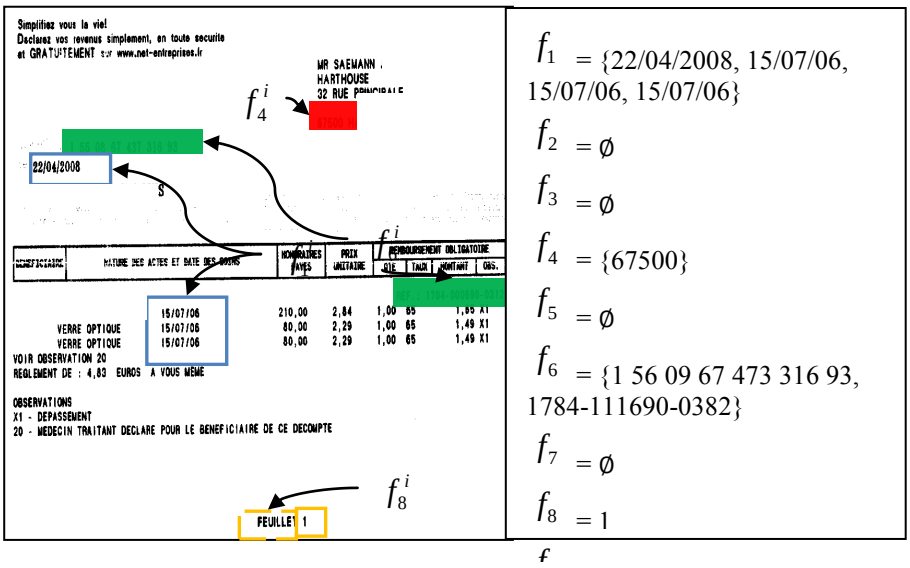


Figure 5. *Exemple d'extraction de descripteurs par expressions régulières*

2.2.3.1. Expressions régulières

Le Tableau 3 illustre les expressions régulières choisies pour extraire les descripteurs.

Tableau 3.*Expressions régulières utilisées pour extraire les motifs*

Descripteurs	Expressions régulières
Date$_1$ (f_1)	^\d{1,2}[-/.]\d{1,2}[-/.]\d{4}$
	^\d{4}[-/.]\d{1,2}[-/.]\d{1,2}$
	^d{2,4}[-/.](JANVIER|…|[dD][eé]cembre)[-/.]\d{2,4}$
Heure (f_2)	^([0-1][0-9]|[2][0-3]):([0-5][0-9])$
Téléphone (f_3)	^0[1-9](([\.\-\s])?[0-9]){8,20}$
Code postal (f_4)	^(F-)?((2[A|B])|[0-9]{2})[0-9]{3}$
Alphanumérique (f_5)	^(?=.*\d)(?!.*(?:JANVIER|…|Dec))[][\w@_(){}ç\\ù%-]{3,90}$
Numérique (f_6)	^[^a-z]{6,}$
Salutation (f_7)	^((c|C)ordialement|… |consid(e|é|\u00E9)ration)$
Numéro de page (f_8)	((p|P)age|(f|F)olio|(P|p)g|((v|V)olet)|(f|F)euillet)

Quand une date est extraite, le mois est converti en numérique et les caractères spéciaux sont écartés. Cette étape nous garantit d'avoir le même format de date durant le calcul de la continuité. L'expression régulière qui correspond aux descripteurs alphanumériques f_5 ignore les chaînes de caractère contenant les mois de l'année pour faire la séparation avec le descripteur f_1. Pour le descripteur numériques f_6, toutes les formes numériques dont la taille est inférieure à *6* sont ignorées car il est difficile de les distinguer des montants dans des tableaux et qui ne sont pas significatifs.

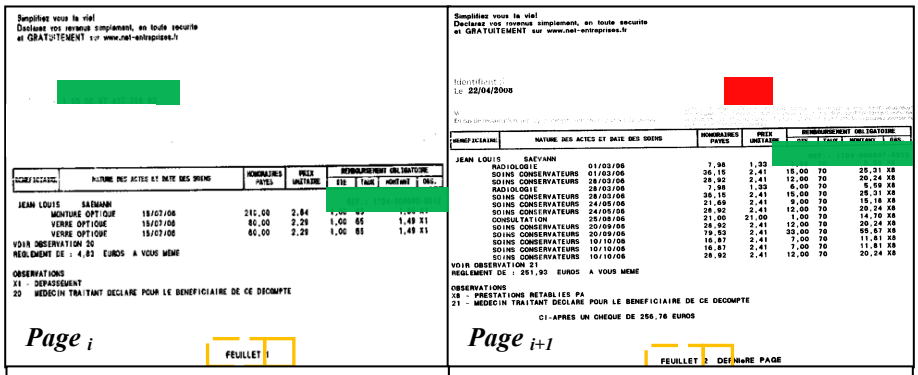
2.3. Modélisation de la relation entre deux pages successives

Une fois les descripteurs de pages individuelles extraits, nous cherchons à modéliser la relation entre les couples de pages consécutives p_i et p_{i+1} telle une continuité ou rupture. Nous avons remarqué qu'une comparaison deux à deux des descripteurs des deux pages avec une majorité de continuités ou de ruptures ne conduit pas forcément à une continuité ou rupture au niveau des deux pages. C'est

pour cette raison que nous avons créé un vecteur dont les composantes sont des résultats de comparaison deux à deux des descripteurs et utilisé un classifieur global. Les valeurs des composantes appartiennent à l'ensemble : *{-1, 0, 1}* où *-1* traduit l'inégalité des composantes et indique donc la rupture, 1 traduit l'égalité des composantes et indique la continuité et 0 traduit l'absence de l'un des descripteurs. La relation R entre le couple de pages p_i et p_{i+1} est alors définie comme suit :

$$R = \{v_j : v_{j=1,\ldots,8} \in \{-1, 0, 1\},\ v_{j=9} \in \{-1, 1\}\} \quad [1]$$

La Figure 6 illustre comment la relation R est construite. Dans le cas de f_1, comme il existe au moins une intersection entre f_1^i et f_1^{i+1} alors $v_1 = 1$. $f_2^i, f_3^i, f_4^i, f_5^i$ n'existent pas dans les deux pages, alors $v_{j=2,3,4,5} = 0$. Comme $f_8^i < f_8^{i+1}$ alors $v_8 = 1$.



f_1^i = {*22/04/2008, 15/07/06, 15/07/06, 15/07/06*}
$f_2^i = \emptyset$, $f_3^i = \emptyset$
f_4^i = {*67500*}, $f_5^i = \emptyset$
f_6^i = {*1 56 09 67 473 316 93, 1784-111690-0382*}
$f_7^i = \emptyset$, $f_8^i = 1$
f_9^i = *Taille du bloc le plus large*

f_1^{i+1} = {*22/04/2008, 01/03/06,…,10/10/06*}
$f_2^{i+1} = \emptyset$, $f_3^{i+1} = \emptyset$, f_4^{i+1} = {*67500*}
$f_5^{i+1} = \emptyset$
f_6^{i+1} = {*1 56 08 67 437 316 93, 1708-000690-0316*}
$f_7^{i+1} = \emptyset$, $f_8^{i+1} = 2$
f_9^{i+1} = *Taille du bloc le plus large*

$R =$

v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
1	0	0	1	0	1	0	1	1

Figure 6. *Exemple de calcul des relations entre deux pages consécutives*

2.4. Classification

Commençons d'abord avec l'hypothèse qu'il est préférable de sur-segmenter un document que de le fusionner. La fusion par erreur implique qu'un document est perdu (sous-segmentation) car il est sûrement combiné avec le précédent. En cas de sur-segmentation, aucun document ne se perd. Même si un document est fragmenté, le destinataire recevra au moins un fragment. Les autres fragments seront probablement classés correctement et renvoyés plus tard. S'appuyant sur cette hypothèse, l'étape de classification consiste à prédire la classe de la relation ***R***. Elle peut être classée comme ***continuité***, ***rupture*** et dans le cas d'une ***incertitude,*** elle est rejetée, et dans ce cas, on obtient un fragment.

2.4.1. Différents cas de reconnaissance

- **Continuité :** On considère que ces deux pages appartiennent à un même document et on passe à la page suivante pour étudier la relation
- **Rupture franche :** On arrête la lecture du flux et on pause l'hypothèse qu'on a un document complet
- **Incertitude :** On considère qu'il existe un rejet qui est présenté par une sur-segmentation, alors on crée un fragment de document avec les pages précédentes. On poursuit la classification avec le reste des pages jusqu'à une rupture franche. Le résultat est une suite de fragments et un document à la fin

2.4.2. Fonction de rejet

Soit *P(c|R)* la probabilité qu'une relation *R* appartienne à la classe continuité *(c)*. Soit *P(r|R)* la probabilité qu'une relation *R* appartienne à la classe rupture *(r)*. La fonction de rejet est alors définie comme suit : $Q_x = \left| P(c|R) - P(r|R) \right|$. Si $Q_x \prec \delta$ alors il existe une incertitude et donc un rejet. Le seuil δ est défini manuellement et adapté à toutes les bases

3. Expérimentations

Toutes nos expériences ont été réalisées sur les bases de données fournies par la société ITESOFT. Pour tester la stabilité de l'approche, nous avons utilisé *quatre* bases de données contenant des nombres différents de documents et de pages.

- Base *1* : *761* documents (*1857* pages et *15* classes)
- Base *2* : *2021* documents (*3630* pages et *23* classes)
- Base *3* : *3159* documents (*8584* pages et *137* classes)
- Base *4*: *5184* documents (*13448* pages et *164* classes)

La base *1* représente un flux de documents hétérogènes avec des classes qui sont facilement séparables. Les bases *2*, *3* et *4* contiennent des documents avec des

structures complexes. Les classes de documents se chevauchent et ne sont pas faciles à séparer.

3.1. Extraction des descripteurs

Pour tester la stabilité de l'algorithme d'extraction des descripteurs, *393* pages ont été choisies au hasard parmi les quatre bases de données. La vérité du terrain a été construite en marquant manuellement tous les mots dans les fichiers XML selon leurs types (date, heure, téléphone etc.). Le système compte automatiquement le nombre de descripteurs f_i correctement identifiés sur chaque page en comparant les descripteurs attendus avec les étiquettes de la vérité de terrain. La Figure 7 montre la stabilité de l'algorithme d'extraction de descripteurs proposé, basé sur les expressions régulières.

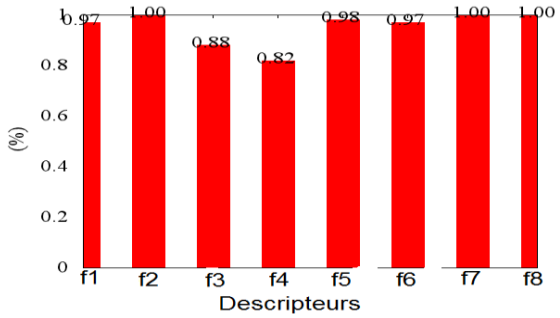


Figure 7. *Précisions de l'algorithme d'extraction des descripteurs*

3.2. Évaluation globale du système

L'objectif est d'évaluer la pertinence de la relation *R* à prédire correctement la relation (continuité ou rupture). L'évaluation est représentée par la moyenne pondérée du rappel pour la relation *R*. La moyenne pondérée est calculée en pondérant la mesure de la classe (Rappel) par la proportion des instances qui sont dans cette classe. Comme l'ensemble des validations explicites n'est pas disponible, la fiabilité du modèle est estimée par la validation-croisée avec $k = 5, 10, 15$ et *20*. Quatre classifieurs sont choisis pour cette validation : SVM, Voted perceptron, Multi-layer perceptron et multi-boost (voir Figure 8). Pour la base *1,* la valeur moyenne du rappel sur les *4* classifieurs est *92.5%*. Les classes des documents de la base *1* sont très hétérogènes donc faciles à séparer, c'est pour cette raison là qu'on a un bon résultat de classification sur les *4* classifieurs.

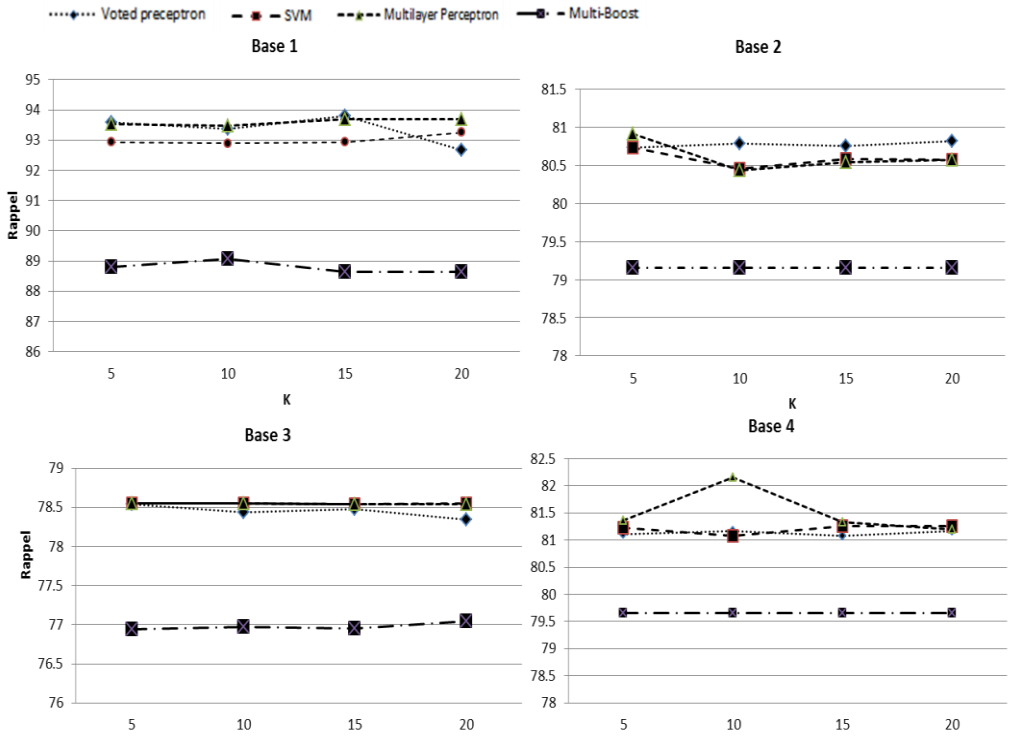


Figure 8. *Evaluation de la relation R sur les 4 bases en utilisant 4 classifieurs*

Pour la base 2, la valeur moyenne du rappel est de *80.25%*. La différence de *12.25%* avec la base *1* est dû au fait que les documents de la base 2 sont difficiles à séparés. Il existe parfois des similarités entre différentes classes, ce qui complique la classification. Pour les bases *3* et *4,* les moyennes sont respectivement *78.12%* et *82%*. Ces résultats montrent la stabilité de notre choix. Même avec l'augmentation de la taille des bases *2, 3* et *4*, l'algorithme reste stable.

3.3. Évaluation de la fonction de rejet

La fonction de rejet est calculée ***uniquement*** sur les cas où la relation *R* est classée comme continuité. Une fausse continuité produit une sous-segmentation « fusion » de documents. L'erreur de rupture signifie qu'il y a une sur-segmentation entre deux pages successives. Cette erreur est moins grave que l'erreur de sous-segmentation car tout le document n'est perdu et la partie qui reste sera probablement classée et envoyée après au destinataire.

Le Tableau 4 illustre la matrice de prédiction de la fonction de rejet. *FP* (Faux positif) représente une relation *R* faussement rejetée. Une sur-segmentation aura lieu. Dans le cas de *FN* (Faux négatif), le classifieur n'a pas rejeté la relation mal classée. Une fusion de deux pages appartenant à deux documents différents aura lieu. Plus *FN* est faible mieux est la segmentation. Ce cas est risqué car les documents peuvent être fusionnés et donc perdus.

Tableau 4. *Matrice de prédiction*

Rejet	**Prédiction**	**Vérité de terrain**	
Vrai	Continuité	Rupture	TP
	Continuité	**Continuité**	**FP sur-segmentation**
Faux	**Continuité**	**Rupture**	**FN fusion**
	Continuité	Continuité	TN

L'évaluation est faite en utilisant le classifieur Multi-layer perceptron qui a fourni sur les quatre bases le meilleur rappel et en utilisant *k-fold = 10*. La Figure 9 montre que plus le seuil de rejet augmente, plus le nombre d'erreurs de fusion ***FN*** diminue mais au profit des erreurs de sur-segmentation ***FP***. Pour les bases *1*, *2*, *3* et *4,* les seuils optimaux sont respectivement *0.6, 0.3, 0.4, 0.4.* Les documents de la base *1* sont facilement séparables alors la quantité Q_x est élevée signifiant que l'incertitude entre les classes est faible. Par contre, plus les documents deviennent complexes et plus la taille de la base augmente; il y aura plus d'incertitude. La quantité Q_x est alors plus difficile à choisir. C'est le cas pour les bases *2, 3* et *4*. Décider d'avoir un taux de rejet élevé ou non est une décision très importante. Si le nombre d'erreurs de sous-segmentation doit être minimisé alors le seuil de rejet doit être augmenté.

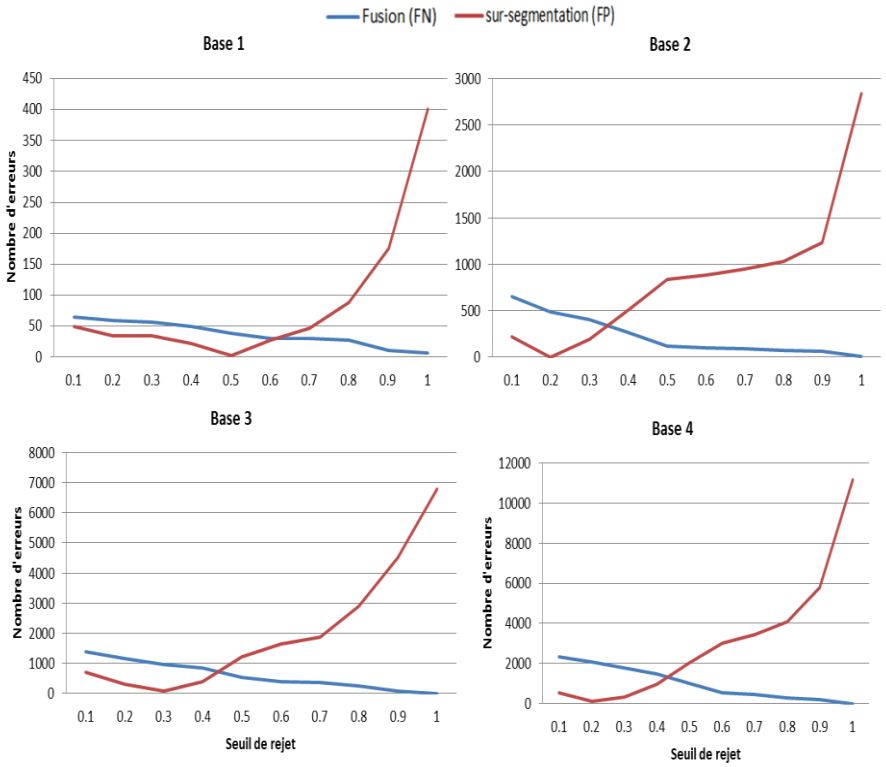


Figure 9. *Relation entre le seuil et la fonction de rejet*

4. Conclusion

Nous avons présenté dans cet article une approche de segmentation de flux hétérogènes de documents. Une méthode générique a été proposée pour modéliser la relation entre les pages successives du flux qui, ensuite, est utilisée par un classifieur pour décider si la relation est une continuité ou rupture. L'étape d'extraction des descripteurs à partir des expressions régulières a été validée et nous avons montré l'efficacité des expressions régulières à extraire les descripteurs. Les résultats de classification montrent la stabilité de notre approche. En effet, l'augmentation du nombre de documents n'a pas affecté les résultats. L'utilisation des probabilités d'appartenance aux classes pour étudier l'ambigüité d'une relation R est une couche très importante; elle permet de différencier les documents complets des documents qui ont une forte probabilité d'être des documents incomplets ou des fragments de documents. La méthode proposée utilise un classifieur statique, une base d'apprentissage doit être toujours disponible. Ce n'est pas le cas dans les applications du monde réel où les entreprises doivent faire face à de nouvelles classes de documents dans des flux continus. Pour résoudre ce problème, nous prévoyons utiliser un classifieur incrémental afin d'intégrer de nouvelles classes de documents sans la nécessité d'une base d'apprentissage fixe. Enfin, nous prévoyons d'intégrer la

méthode proposée dans un système plus complet comprenant un système expert pour corriger les cas d'erreurs de segmentation et un niveau de classification de dossiers avec des connaissances de haut niveau sur les classes de documents.

5. Bibliographie

(Collins et Nickolov, 2002) Collins-Thompson, K. et Nickolov, R., "A clustering-based algorithm for automatic document separation", SIGIR, p. 38-43, 2002.

(Gordo et Perronnin, 2010) Gordo, A. et Perronnin, F., "A bag-of-pages approach to unordered multi-page document classification", p. 1920-1923, ICPR, 2010.

(Kumar et Doermann, 2012) Kumar, J., Ye, P. et Doermann, D., "Learning Document Structure for Retrieval and Classification", p. 1558-1561, ICPR, 2012.

(Meilender et Belaïd, 2009) Meilender, T. et Belaïd, A., "Segmentation of continuous document flow by a modified backward-forward algorithm", DRR, pp. 724 705–724 705–10, 2009.

(Rusiñol et al. 2012) Rusiñol, M., Karatzas D. Bagdanov et Llados J., "Multipage document retrieval by textual and visual representations", p. 521-524, ICPR, 2012.

(Shin et Doermann, 2006) Shin, C. et Doermann, D.,"Document image retrieval based on layout structural similarity", p. 606-612 , ICIP, 2006.

(Shin et al. 2001) Shin, C., Doermann, D. et Rosenfeld, A., "Classification of document pages using structure-based feature", p.232-247, IJDAR, 2001.