

# La relaxation probabiliste pour l'étiquetage logique des documents : Application aux tables des matières

Souad Souafi-Bensafi<sup>1,2</sup>, Franck Lebourgeois<sup>1</sup>, Hubert Emptoz<sup>1</sup>, Marc Parizeau<sup>2</sup>

<sup>1</sup>Laboratoire de Reconnaissance de Formes et Vision  
I.N.S.A. de LYON - Bât 403  
20 Av. A. Einstein 69621 Villeurbanne Cedex FRANCE  
Emails : [souafi,flebourg,emptoz]@rfv.insa-lyon.fr

<sup>2</sup>Laboratoire de Vision et de Systèmes Numériques  
Département de Génie-électrique, Université Laval  
Ste-Foy, Québec, CANADA G1K 7P4  
Emails : [bensafi,parizeau]@gel.ulaval.ca

## Résumé

*La relaxation probabiliste est une technique permettant d'associer des étiquettes à un ensemble d'objets en tenant compte de leur voisinage. Nous proposons une application de cette méthode, à une échelle différente, dans le domaine de l'Analyse et de la Reconnaissance des Documents. L'analyse d'un document permet de segmenter son image et d'en extraire une structure physique. Dans la phase de reconnaissance, la structure logique est établie à partir de la structure physique. Il s'agit de réaliser l'étiquetage logique des blocs de texte dans des documents, en leur affectant des fonctions logiques. C'est cette phase de reconnaissance, qui constitue le cadre de notre étude.*

*Notre expérimentation a porté sur les sommaires et les tables des matières des périodiques, et nous a conduit à des résultats intéressants et significatifs.*

## 1. Introduction

L'Analyse et la Reconnaissance des Documents (ARD) a connu une avancée importante ces dernières années, tant au niveau des différentes approches méthodologiques qu'au niveau des applications. L'objectif actuel de l'ARD [2] est de traduire le document sous une forme électronique compréhensible et réutilisable. Notre objectif est de développer des méthodes de reconnaissance des structures des documents, à partir de leurs images numérisées [10].

Il existe deux principaux niveaux de structuration des documents, le niveau physique et le niveau logique. Le niveau physique porte sur le résultat de la segmentation des images de documents, qui consiste selon les besoins de l'application, à séparer le texte des images, et fournit des blocs de différents types : caractères, mots, lignes, etc.; ces blocs constituent des entités physiques qui peuvent être décrites par des caractéristiques géométriques, typographiques et/ou topologiques. Au niveau logique, une sémantique est attribuée aux entités physiques.

Nous travaillons sur des documents dont le contenu est structuré de manière hiérarchique selon différentes fonctionnalités. Celles-ci sont traduites par des étiquettes logiques qui composent ainsi la structure logique à reconnaître. Nous nous sommes intéressés à une famille des documents particuliers, les documents à typographie riche et récurrente parmi lesquels nous pouvons citer : les sommaires de périodiques, les tables des matières, les dictionnaires, les pages jaunes, les guides, etc.

Différentes méthodes d'apprentissage et de reconnaissance de structures de documents ont été proposées; parmi celles-ci nous pouvons citer quelques références décrivant des approches variées. Dans [7], des règles de reconnaissance sont fournies par un apprentissage inductif supervisé basée sur la logique des prédicats. [3] utilise les grammaires d'arbres et [12] décrit une méthode basée sur un système neuro-flou. [4] modélise chaque classe de documents par un arbre

trigrams, et [14] utilise les relations d'Allen dans une représentation par graphes.

Les travaux sur la reconnaissance des structures des documents s'orientent généralement vers des approches structurelles. Dans notre étude nous adoptons une approche probabiliste justifiée par la variabilité dans l'organisation des structures; nous proposons un modèle probabiliste de structure logique qui repose sur des relations spatiales entre les blocs de texte; ces relations entre les propriétés physiques des blocs de texte et les étiquettes logiques seront construites dans le cadre d'un apprentissage supervisé.

Notre méthode de reconnaissance de la structure logique utilise la technique de relaxation probabiliste qui définit itérativement la fonctionnalité de chaque bloc de texte de telle façon qu'elle demeure compatible avec celles du voisinage, sur la base des caractéristiques physiques et logiques.

Nous décrirons d'abord le processus de relaxation d'une manière générale. Nous présenterons ensuite les modèles des structures physique et logique, ainsi que les processus d'apprentissage et de reconnaissance en utilisant la relaxation. Enfin, nous discuterons les résultats de l'application sur les tables des matières des périodiques.

## 2. Relaxation probabiliste

La relaxation probabiliste est une méthode utilisée pour affecter des étiquettes à un ensemble d'objets sur la base des propriétés des objets en tenant compte de leur voisinage, cette affectation étant ajustée de manière itérative.

La relaxation probabiliste est une généralisation de la relaxation discrète [15], qui a été développée dans le cadre d'une interprétation simplifiée de dessins de traits[11]. Cette généralisation correspond à l'association d'un degré de certitude à l'affectation d'une étiquette à un objet.

Soit  $\mathbf{B}$  un ensemble d'objets et  $\Omega$  l'ensemble des étiquettes. Une fonction d'étiquetage  $\mathcal{E}$  est définie sur l'ensemble des objets; elle permet d'associer une étiquette  $\omega = \mathcal{E}(b)$  à un objet  $b$ .  $V(b)$  constitue l'ensemble des objets voisins de l'objet  $b$ . Notons  $P(i, j)$  la probabilité  $P[\mathcal{E}(b_i) = \omega_j]$  pour que  $\omega_j$  soit l'étiquette du bloc  $b_i$ .

La compatibilité de l'étiquetage d'un objet avec l'étiquetage d'un objet voisin est mesurée par une fonction  $C(i, j; h, k)$ , pour les étiquetages :  $\mathcal{E}(b_i) = \omega_j$  et  $\mathcal{E}(b_h) = \omega_k$  avec  $b_i, b_h \in \mathbf{B}$ ,  $b_h$  étant un voisin de  $b_i$ , et  $\omega_j, \omega_k \in \Omega$ .

Dans le cas de fonctions de compatibilité non négatives, les valeurs élevées correspondent à une haute compatibilité, et l'incompatibilité est exprimée par des valeurs proches de zéro. Comme exemple de fonction de compatibilité, nous pouvons prendre la probabilité conditionnelle de l'étiquetage de deux objets voisins  $b_i, b_h$  [13] :

$$C(i, j; h, k) = P[\mathcal{E}(b_i) = \omega_j / \mathcal{E}(b_h) = \omega_k] \quad (1)$$

A partir de la fonction  $C$ , la compatibilité moyenne, sur l'ensemble des voisins, pour l'étiquetage d'un objet  $b_i$  avec l'étiquette  $\omega_j$ , peut être calculée par :

$$Q(i, j) = \sum_{\omega_k \in \Omega} \sum_{b_h \in V(b_i)} w_{ih} C(i, j; h, k) \cdot P(h, k), \quad (2)$$

avec  $w_{ih}$  étant le poids du voisinage reliant les deux objets  $b_i$  et  $b_h$ . Ce poids peut être égal à  $1/|V(b_i)|$  dans le cas où les poids sont égaux pour tous les voisins considérés.

Pour un objet  $b_i$ , l'ensemble des  $Q(i, j)$  pour toutes les étiquettes  $\omega_j$  permet de définir un étiquetage probable en fonction du voisinage.

Soit  $P^0(i, j)$  une distribution initiale des probabilités d'affecter l'étiquette  $\omega_j$  à l'objet  $b_i$ . Il en résulte une distribution des étiquettes  $Q^0(i, j)$  en fonction du voisinage.

La relaxation consiste à calculer, pour chaque objet  $b_i$  et chaque étiquette  $\omega_j$ , de manière itérative, les deux suites :  $P^0(i, j), \dots, P^m(i, j)$  et  $Q^0(i, j), \dots, Q^m(i, j)$ . Le processus d'estimation des probabilités d'étiquetage doit tenir compte du fait que les mesures  $P^t(i, j)$  doivent rester des probabilités, c'est-à-dire, qu'elles doivent être non négatives, et que l'égalité  $\sum_{\omega_j \in \Omega} P^t(i, j) = 1$  soit vérifiée pour tout  $b_i \in \mathbf{B}$ .

La mise à jour itérative des probabilités d'étiquetage dans le cas des fonctions de compatibilités non négatives est donnée par :

$$P^{t+1}(i, j) = \frac{P^t(i, j) \cdot Q^t(i, j)}{\sum_{\omega_k \in \Omega} P^t(i, k) \cdot Q^t(i, k)} \quad (3)$$

Ce processus itératif correspond à la diminution d'une mesure d'incertitude de l'étiquetage, qui peut être définie à l'aide d'une fonction d'entropie  $H$ . Prenons par exemple l'entropie quadratique pour chaque bloc  $b_i$  :

$$H(i) = \sum_{\omega_j \in \Omega} P(i, j) \cdot (1 - P(i, j)) \quad (4)$$

Si  $H(i) = 0$ , l'étiquetage est certain, et ambigu si  $H(i)$  est maximum. De nombreux types d'entropies pondérées ont été introduits dans [6].

### 3. Analyse de la structure physique

Nous travaillons en général sur des images en niveaux de gris, réalisées avec une résolution de 400ppp [9]; nous effectuons d'abord la segmentation puis nous générons l'image binaire en appliquant un seuillage adaptatif. A ce niveau, nous disposons d'un ensemble de boîtes rectangulaires correspondant aux blocs textuels et non textuels. Nous construisons une modélisation physique du document qui est composée :

- d'une structure hiérarchique des blocs textuels sous forme de paragraphes, lignes, mots et caractères ;
- des relations spatiales entre les différents blocs ;
- des descriptions géométrique et typographique de chaque bloc de texte en fonction de son niveau hiérarchique.

#### 3.1. Représentation hiérarchique

Les blocs textuels sont organisés en 4 niveaux hiérarchiques, liés par une relation d'inclusion, donnant pour chaque caractère le mot qui le contient, pour chaque mot la ligne où il se trouve, et pour chaque ligne le paragraphe auquel elle appartient. La relation d'inclusion est représentée par un arbre (figure 1). L'ensemble des blocs textuels est noté  $B$ .

#### 3.2. Typographie des mots

La typographie est porteuse d'information sur la structure d'un document. A partir des données géométriques des blocs de texte nous pouvons extraire un certain nombre de caractéristiques typographiques

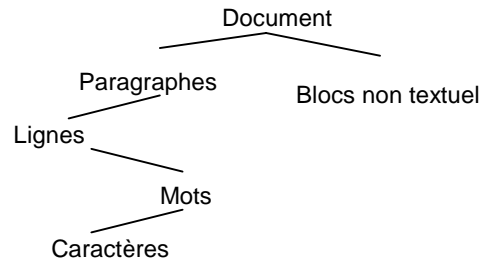


Figure 1 : Arbre d'inclusion

telles que l'alignement au niveau des lignes, l'espacement interlignes au niveau des paragraphes etc.

En ce qui concerne l'extraction de la typographie au niveau des mots et des caractères, il convient de noter que peu de travaux de recherches se sont intéressés à la reconnaissance de fontes [1]; dans les documents que nous traitons, nous trouvons des styles et fontes très variés; cela a donné naissance à une nouvelle approche développée au sein de notre laboratoire [9]; elle consiste à former des groupes de mots ayant la même police. Dans un premier temps, un prototypage des caractères est réalisé en utilisant la méthode d'appariement des formes; on obtient ainsi un ensemble de prototypes avec pour chaque caractère le prototype associé.

Ensuite, en se basant sur l'hypothèse que deux mots ayant un prototype en commun ont la même police, on effectue un regroupement des mots en familles qu'on appelle familles typographiques. Une famille de rejet est prévue pour les mots qui ne peuvent pas être classés. A l'issue de cette étape, la structure physique est complétée par : l'ensemble  $P$  des prototypes des caractères avec les formes binaires correspondantes, l'ensemble  $T$  des familles typographiques, les fonctions  $\pi : B \rightarrow P$ , associant à chaque caractère  $c \in B$  un prototype  $\pi(c) \in P$ ,  $\tau : B \rightarrow T$ , associant à chaque mot  $m \in B$ , une famille typographique  $\tau(m) \in T$ .

#### 3.3. Relations spatiales

A ce niveau nous décrivons les relations entre des blocs de même niveau hiérarchique : il s'agit de relations spatiales entre les mots ou entre les lignes. Nous classons ces relations en deux catégories : relations quantitatives, telles que les distances (espacements horizontaux et verticaux), et des relations qualitatives telles que le voisinage.

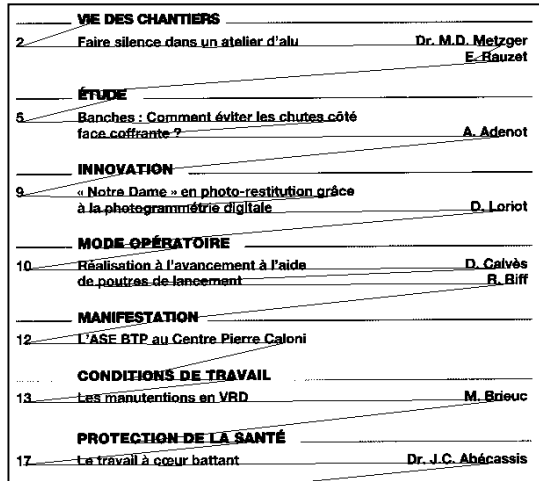
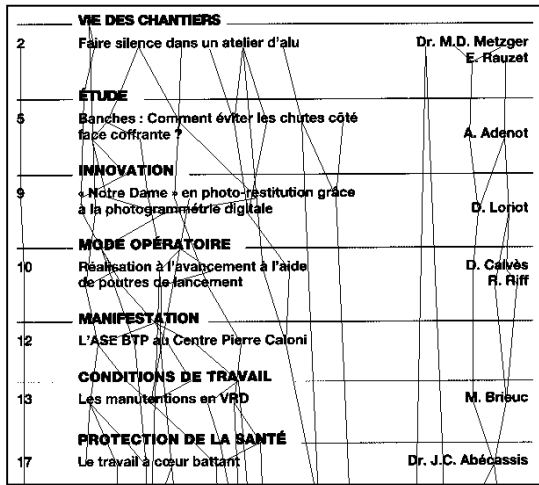


Figure 2 : Voisinages vertical, horizontal et retour chariot entre les mots

De nombreux travaux de recherche s'intéressent à ce type de relations, principalement dans le domaine des systèmes d'information géographiques. [5] propose une modélisation permettant une représentation quantitative des relations spatiales entre deux objets, tandis que [8] utilise une approche algébrique pour représenter les distances et directions dans un espace géographique de manière qualitative.

Nous avons défini 3 types de voisinage pondérés : horizontal (à gauche et à droite), vertical (en haut et en bas), et retour chariot (ligne suivante et ligne précédente). La figure 2 montre des résultats de ce calcul de voisinage. Ces relations sont représentées par :

- un graphe non orienté et valué, dont les sommets sont des blocs, et dont les arêtes expriment les distances entre les blocs d'un même niveau. Ce graphe est valué par les fonctions  $D_H : B \times B \rightarrow IR^+$ , et  $D_V : B \times B \rightarrow IR^+$ , telles que  $D_H(b_i, b_j)$  et  $D_V(b_i, b_j)$  sont respectivement les distances horizontale et verticale entre les deux blocs  $b_i, b_j \in B$ .
- un graphe orienté et étiqueté décrivant le voisinage entre les blocs de même niveau. Les sommets sont les blocs, et le voisinage entre 2 blocs est exprimé par la présence d'un arc direct entre les 2 sommets correspondants avec une étiquette indiquant le type de voisinage. Les étiquettes correspondant aux types de voisinage choisis sont désignées par:  $G, D, H, B, LP$  et  $LS$  (pour gauche, droite, haut, bas, ligne précédente et ligne suivante); soit  $T_V$  l'ensemble de ces étiquettes. Ce graphe est représenté par la fonction de voisinage  $V_T : B \times T_V \rightarrow B$ , telle que  $V_T(b, v)$  donne le voisin du bloc  $b \in B$  en considérant le type de voisinage  $v \in T_V$ . Nous définissons aussi une fonction  $V : B \rightarrow \wp(B)$  qui donne, pour chaque bloc  $b \in B$ , l'ensemble de ses voisins.

Nous avons présenté les différents modules nous permettant de caractériser chaque bloc textuel par des propriétés géométriques et typographiques, qui peuvent être quantitatives ou qualitatives. Un bloc textuel est ainsi décrit par :

- son niveau hiérarchique en fonction duquel, s'ajoutent d'autres caractéristiques plus spécifiques,
- l'ensemble de ses voisins avec les distances qui les séparent.

Nous pouvons obtenir, pour chaque caractère, les couleurs des fond et forme, le prototype associé ; pour chaque mot, l'espacement moyen inter-caractères, et la famille typographique à laquelle il appartient ; pour chaque ligne, l'espacement moyen entre les mots qui la constitue, l'alignement et l'indentation; pour chaque paragraphe, l'inter-lignes moyen.

#### 4. Modélisation de la structure logique

La reconnaissance de la structure logique consiste à élaborer une correspondance entre les entités physiques,

	<b>DISPOSITIFS ET MATÉRIELS</b>	
2	Plate-forme de protection pour pose de dalles alvéolées	D. Lejeune
7	A propos de coffrage en bois	J.C. Turpin
	<b>SALON</b>	
8	Intermat	M. Deroide
	<b>ACCIDENT</b>	
14	Les accidents liés à l'utilisation de matériels et engins de terrassement	A. Martinez
18	Stabilité des tours d'étalement	D. Calvès
	<b>PROTECTION DE LA SANTÉ</b>	
21	Viellir dans le bâtiment	Dr. J.-Y. Dubré
	<b>DISPOSITIFS ET MATÉRIELS</b>	
25	A découvrir...	
	<b>ENSEIGNEMENT - FORMATION</b>	
28	Calendrier des stages du Centre Pierre-Caloni Octobre - novembre - décembre 1994	
29	Référentiels de formation et sécurité intégrés au CFA du Morbihan (Comité régional Bretagne)	

Figure 3 : Etiquettes logiques

et un ensemble d'étiquettes logiques exprimant la sémantique du document. Cette opération est dite étiquetage logique. Le choix des étiquettes logiques dépend des différentes fonctionnalités de l'application. Dans notre étude, nous considérons des étiquettes logiques associées aux différentes catégories de texte constituant l'information à extraire. Dans le cas des tables des matières des périodiques, ces étiquettes correspondent par exemple, aux sections, titres, sous-titres, noms d'auteurs, etc (figure 3).

L'objectif est de réaliser cet étiquetage de manière automatique. Cependant, d'un document à l'autre, l'ensemble des étiquettes ainsi que leurs relations avec la structure physique risquent de varier considérablement. Il est donc nécessaire de construire un modèle pour chaque classe de documents, celle-ci regroupant les documents des structures voisines. Nous avons donc choisi de procéder par apprentissage à partir d'une base d'exemples.

#### 4.1. Apprentissage

Initialement, une phase d'étiquetage logique manuel est nécessaire sur les documents de la base d'apprentissage; cela permettra d'extraire les données qui serviront dans la suite de l'apprentissage. Un ensemble de vecteurs de données doit être construit, chaque vecteur correspondant à un bloc de texte, et contenant l'ensemble des attributs physiques le décrivant et l'étiquette logique qui lui a été affectée. Les blocs de texte considérés sont les mots, car même dans une même ligne, des mots peuvent avoir des fonctions logiques différentes.

Nous proposons un modèle probabiliste, exprimant par des estimations de probabilités, le lien entre les

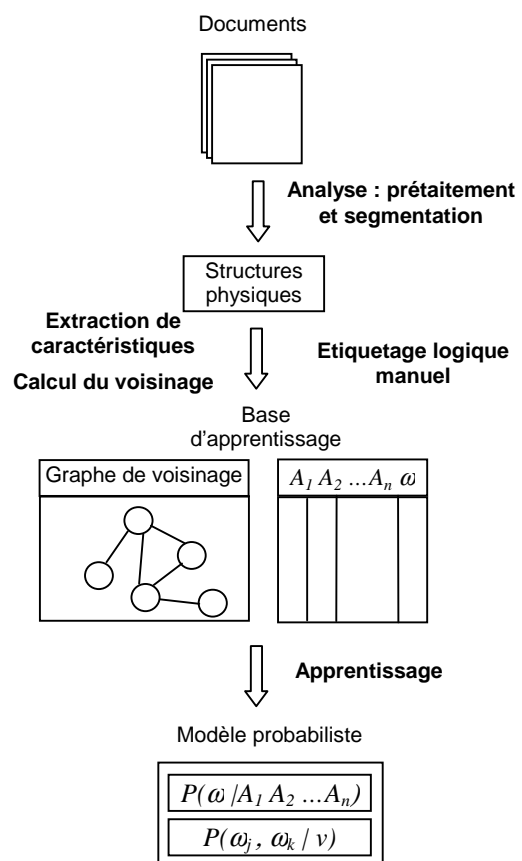


Figure 4 : Processus d'apprentissage

caractéristiques physiques d'un mot et son étiquette logique, et l'étiquetage en fonction du voisinage (figure 4). Ces estimations sont réalisées par comptage de nombres d'occurrence dans la base d'apprentissage.

Une partie du modèle correspond à une table des probabilités conditionnelles :  $P(\omega / A_1, A_2, \dots, A_n)$ , et ce pour toutes les étiquettes et pour toutes les valeurs possibles des attributs. La seconde partie porte sur le lien entre l'étiquetage et le voisinage. Pour chaque paire d'étiquettes logiques  $\omega_j, \omega_k$  et chaque type de voisinage  $v$ , l'apprentissage consiste à estimer la probabilité conditionnelle  $P(\omega_j, \omega_k / v)$ , d'avoir les étiquettes voisines  $\omega_j, \omega_k$  connaissant le type de voisinage  $v$ .

Le processus d'apprentissage consiste à mettre à jour ces probabilités décrivant la structure d'un document (figure 4), mais le nombre de caractéristiques ou attributs physiques que nous pouvons extraire est plus

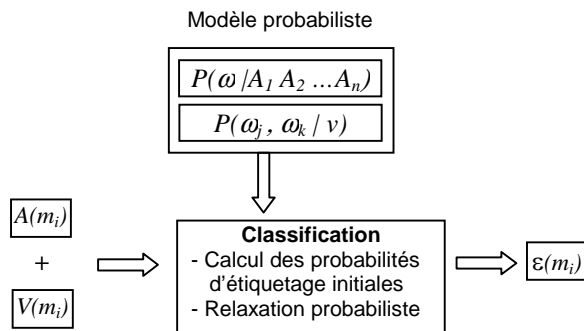


Figure 5 : Processus de reconnaissance

élevé que celui nécessaire pour modéliser la structure. Nous projetons d'améliorer le processus d'apprentissage par la sélection automatique de la combinaison des caractéristiques les plus pertinentes décrivant la structure d'une classe de documents.

Cependant, la mise à jour des probabilités n'améliore pas immédiatement la reconnaissance si les observations antérieures, plus nombreuses, sont contradictoires. Afin de réduire le temps d'apprentissage dans des limites raisonnables, nous avons préféré mettre à jour les probabilités jusqu'à ce que les blocs corrigés soient correctement reconnus. Cette technique permet de négliger le poids du passé et d'adapter plus rapidement le modèle aux nouvelles situations.

## 4.2. Reconnaissance par la relaxation

La phase de reconnaissance correspond à l'étiquetage logique automatique. Elle doit attribuer à l'ensemble des mots pour un document donné, les étiquettes correspondantes. Il s'agit d'une opération de classement, les étiquettes logiques étant les classes selon lesquelles les mots doivent être classés.

Soit  $A(m)$  le vecteur des valeurs prises par les attributs physiques  $A_1, A_2, \dots, A_n$  décrivant le mot  $m$ , et soit  $\mathcal{E}(m)$  sont étiquette logique. Le modèle fourni par la phase d'apprentissage, permettre le calcul de  $\mathcal{E}(m)$  à partir de  $A(m)$  et de l'ensemble de ses voisins  $V(m)$ .

La reconnaissance se fait en deux étapes (figure 5) : une phase se basant sur l'ensemble des attributs  $A(m)$  pour chaque mot  $m$ , et utilisant les probabilités

conditionnelles  $P(\omega / A_1, A_2, \dots, A_n)$ , pour calculer les probabilités d'étiquetage pour ce mot. Ces probabilités servent dans la seconde étape comme probabilités d'étiquetage initiales dans le processus de relaxation.

Nous nous sommes limités dans un premier temps, pour l'ensemble des attributs physiques, à la famille typographique. Pour chaque mot  $m_i$  et chaque étiquette  $\omega_j$ , la probabilité d'étiquetage est initialisée par l'estimation de probabilité conditionnelle d'étiquetage connaissant la famille typographique du mot  $\tau(m_i)$ :  $P^0(i, j) = P(\omega_j / \tau(m_i))$ . Cette probabilité exprime le risque de classer un mot  $m_i$  dans  $\omega_j$  en fonction de sa typographie. La fonction de compatibilité utilisée est celle introduite à la section 2 (équation 1). Elle est estimée par la probabilité liant les deux étiquettes :  $C(i, j; h, k) = P[\omega_j, \omega_k / v_{ih}]$  telle que  $V_T(m_i, v_{ih}) = m_h$ . La compatibilité moyenne se déduit directement par :

$$Q^t(i, j) = \frac{1}{|V(m_i)|} \sum_{\omega_k \in \Omega} \sum_{m_h \in V(m_i)} P[\omega_j, \omega_k / v_{ih}] \times P^t(h, k) \quad (5)$$

Les probabilités d'étiquetage sont mises à jour par la formule itérative (section 2, équation 3). A chaque itération, l'entropie quadratique (équation 4) est calculée, et une valeur proche de 0 correspondra à la satisfaction du critère d'arrêt. Les étiquettes affectées aux mots à classer, à la fin du processus de reconnaissance, seront celles ayant les plus fortes probabilités d'étiquetage.

## 5. Expérimentation et Résultats

Nous avons appliqué notre système d'apprentissage et de reconnaissance aux sommaires de revues. Nous avons défini 5 classes logiques focalisées sur la notion d'article :  $\Omega = \{\text{Titre, résumé, auteur, référence, section}\}$  auxquelles on doit ajouter une sixième classe qui est celle des mots rejetés. Pour les sommaires de périodiques, la définition du bloc de texte est réduite à celle du mot, car les sommaires peuvent comporter dans la même ligne, des mots de fonctions logiques différentes. Dans les sommaires, le mot est donc la plus petite entité indivisible qui ne peut pas avoir plus d'une fonction logique.

Les caractéristiques utilisées se limitent à la famille typographique du mot. Nous pourrions étendre l'ensemble à d'autres caractéristiques descriptives dans un prochain travail.

Nous avons appliqué la reconnaissance dans la phase d'apprentissage, pour semi-automatiser l'étiquetage logique, c'est-à-dire qu'à partir du deuxième exemple, un étiquetage automatique est tenté, en appliquant la reconnaissance, et sera éventuellement corrigé manuellement, afin de servir à l'apprentissage du modèle. De cette façon, la base d'apprentissage ainsi que le modèle, sont mis à jour de manière incrémentale.

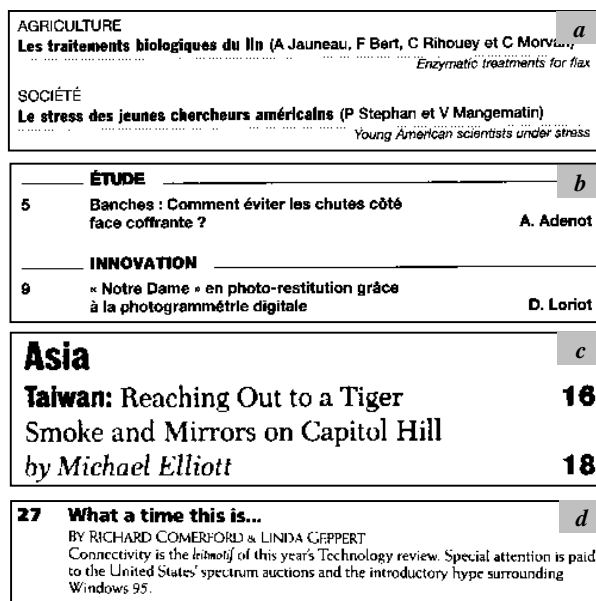
Le tableau ci-dessous, présente les taux de reconnaissance obtenus en phase d'apprentissage (2<sup>ème</sup>, 3<sup>ème</sup> et 4<sup>ème</sup> exemples), et le taux moyen de reconnaissance après l'apprentissage (sur 5 exemples au maximum). Les colonnes correspondant aux nombres de familles typographiques et étiquettes logiques montrent les différents cas possibles. Les revues ont été choisies pour leur difficulté dans l'association d'étiquettes logiques multiples à une même famille typographique.

La différence dans les taux de reconnaissance pour les deux classes 1 et 2 est due à l'organisation spatiale des étiquettes correspondant à une même famille. La classe 2 nécessiterait de considérer en plus de la typographie des mots et du voisinage, les distances entre les mots, afin d'améliorer les résultats. Dans les classes 2, 3 et 4, les différents types de voisinage jouent des rôles d'importance inégale, il serait donc plus intéressant de leur associer des poids différents, pouvant être calculés en phase d'apprentissage.

Classes (Revues)	(a)	(b)	(c)			(d)
			2 <sup>ème</sup>	3 <sup>ème</sup>	4 <sup>ème</sup>	
1 (Biofutur)	4	5	99,3%	100%	88,7%	97,3%
2 (Cahiers des ...)	3	4	80,5%	76,0%		80,6%
3 (NewsWeek)	5	4	76,7%	84,4%	89,1%	86,0%
4 (Spectrum)	3	4	81,9%	90,3%	86,2%	88,1%
5 (L'Express)	2	2	95,1%	98,1%		97,8%

- (a) Nombre De familles typographiques
- (b) Nombre d'étiquettes logiques
- (c) Taux de reconnaissance durant l'apprentissage (sur les exemples)
- (d) Taux de reconnaissance moyen après apprentissage

**Tableau 5** : Taux de reconnaissance



**Figure 6** : Exemples de revues : -a- Biofutur, -b- Cahiers des préventions, -c- NewsWeek et -d- Spectrum

Les périodiques qui présentent un taux de reconnaissance bas, indiquent une instabilité de la structure logique qui ne peut pas être apprise avec les seules informations typographiques et spatiales dont le système dispose actuellement. Pour ces périodiques, nous n'avons pas d'autres solutions que d'augmenter le nombre d'attributs physiques.

Nous avons constaté avec satisfaction que l'apprentissage est très rapide puisque seulement 3 ou 4 revues sont nécessaires pour obtenir des densités de probabilités fiables et un dictionnaire complet des différentes formes de caractères afin d'avoir la certitude de posséder toutes les typographies présentes dans le modèle de document. Contrairement à ce qui se passe pour la segmentation d'image, l'application de la relaxation en reconnaissance de la structure logique des documents est rapide; le nombre de mots est très limité et donc les informations nécessaires aux calculs, occupent un espace mémoire très réduit. Le nombre d'itérations dépend de la complexité de la structure logique et de l'adéquation entre la structure physique et la structure logique décrite par le modèle. Dans la plupart des cas, moins de 10 itérations sont nécessaires pour la convergence et l'obtention d'une classification finale, cela donne un temps de reconnaissance extrêmement court de l'ordre d'une seconde pour une page de document.

## 6. Conclusion et perspectives

Nous avons présenté un système d'apprentissage et de reconnaissance de structures de documents basé sur un modèle probabiliste et utilisant la technique de relaxation pour la reconnaissance. Ce système a été élaboré pour la classe de documents à typographie riche et récurrente. Il utilise des caractéristiques spatiales et typographiques liées aux structures physique et logique pour chaque classe de documents.

Dans l'application que nous avons présentée, les calculs se sont limités à l'utilisation de la typographie et du voisinage au niveau des mots. Les taux de reconnaissance donnés par la relaxation sont assez satisfaisants.

Il est clair que pour d'autres familles de documents l'utilisation d'autres caractéristiques et d'autres sources d'information sera nécessaire. Cela engendrera des problèmes au niveau de l'apprentissage. Nous projetons dans l'immédiat, pour améliorer les résultats, d'utiliser des informations que nous sommes en mesure d'extraire, et ce au niveau des caractères, des mots ainsi que les lignes, qui peuvent contribuer à mieux différencier des fonctions logiques.

Un travail approfondi doit être aussi mené sur la reconnaissance de structures supérieures hiérarchiques qui regroupent plusieurs blocs logiques différents dans des blocs logiques de plus haut niveau, de façon à faire apparaître la notion d'article, de rubrique et enfin de retrouver réellement le sens de parcours et l'ordonnement des informations.

Nous avons adopté une approche statistique dont l'utilisation se fait rare dans le domaine, contrairement aux approches structurelles, qui sont limitées à des documents particuliers souvent très structurés. Notre objectif est de développer une méthode qui s'adapte à la variabilité dans les structures de documents. Il est donc intéressant de représenter de manière générique la structure d'un document. Nous étudions actuellement l'utilisation d'un modèle probabiliste appelé réseau bayésien, que nous comptons combiner avec la relaxation probabiliste.

## 7. Références

- [1] Bapst F., Ingold R., "Using Typography in Document Image Analysis", *RIDT'98*, Avril 1998.
- [2] Belaïd A., "Analyse et Reconnaissance de Documents", Le traitement électronique du document, Cours INRIA, Octobre 1994, Aix-en-Provence, Editions ADBS, pp.49-92.
- [3] Belaïd A., "Conception automatisée de modèles de page en vue de leur utilisation en reconnaissance de documents", *Lausanne-Atelier sur les modèles de pages Electroniques (LAMPE'97)*, Lausanne, 22 Septembre 1997.
- [4] Brugger R., Zramdini A., Ingold R., "Modeling documents for structure recognition using generalized N-grams", *4<sup>th</sup> ICDAR*, Ulm, Germany, August 1997, pp.56-60.
- [5] Del Bimbo A., Vicario E., "Using Weighted Spatial Relationships in Retrieval by Visual Contents", *Proceedings. IEEE Computer Society*, Los Alamitos, CA, USA, 1998, VIII+115, pp.35-39.
- [6] Emptoz H., *Informations utiles et pseudoquestionnaires*, Thèse de Doctorat Université Claude Bernard de Lyon, 12 Mars 1976, 130p.
- [7] Esposito F., Malerba D., Semeraro G., "Automated Acquisition of Rules for Document Understanding", *2<sup>nd</sup> ICDAR*, Tsukuba Science City, Japan, October 1993, pp.650-654.
- [8] Franck A.U., "Qualitative Spatial Reasoning about Distances and Directions in Geographic Space", *Journal of Visual Languages and Computing*, 1992, n°3, pp.343-371.
- [9] LeBourgeois F., Emptoz H., "Document Analysis in Gray Level and Typography Extraction Using Character Pattern Redundancies", *5<sup>th</sup> ICDAR*, 1999, pp.177-180.
- [10] Lebourgeois F., Emptoz H., Vigne H., "RASADE : Reconnaissance Automatique des Structures Associées aux Documents Ecrits", *CIFED 2000*, pp. 281-294.
- [11] Rosenfeld A., Hummel R., Zucker S.W., "Scene Labeling by Relaxation Operation", *IEEE Transactions on Systems, Man & Cybernetics*, 1976, Vol 6, pp. 420-433.
- [12] Sainz Palmero G.I., Dimitriadis Y.A., "Structure Document Labeling and Rule Extraction Using a New Recurrent Fuzzy-Neural System", *5<sup>th</sup> ICDAR*, Bangalore, India, September 1999, pp.181-184.
- [13] Simon J.C., *La reconnaissance des formes par algorithmes*, Editions Masson, Paris 1984, 251p.
- [14] Walischewski H., "Automatic Acquisition for Spatial Document Interpretation", *4<sup>th</sup> ICDAR*, Ulm, Germany, August 1997, Vol 1, pp.243-247.
- [15] Zucker S.W., "Relaxation Labelling, Local Ambiguity, and Low-Level Vision", Editor Chen C. H., *Pattern Recognition and Artificial Intelligence*, Academic Press, Inc, 1976, pp.593-616.