
Un modèle statistique pour la classification de documents structurés

Trang Huyen Vu, Ludovic Denoyer, Patrick Gallinari
LIP6 - Université Paris 6 - 8 rue du capitaine Scott 75015 PARIS
{vu,denoyer,gallinari}@poleia.lip6.fr

RÉSUMÉ. : Nous présentons un modèle d'apprentissage général pour la classification de documents structurés permettant de prendre en compte simultanément la structure et le contenu. Pour cela, nous définissons tout d'abord un modèle génératif de documents structurés à l'aide de réseaux Bayésiens. Nous transformons ensuite ce modèle génératif en un modèle discriminant en utilisant la méthode du noyau de Fisher. Nous détaillons enfin une instance de ce modèle dédié à la classification de pages HTML. Les expériences sur un corpus de référence montrent que la prise en compte de la structure permet un gain de performance par rapport aux modèles classiques de classification génératifs et discriminants.

ABSTRACT. : We present a learning model for categorization of structured documents that takes into account both structural information and textual information. We first define a generative model of structured documents using belief networks. Then we transform the generative model into a discriminant one using the Fisher kernel. Finally, we describe an instance of this model applied to the categorization of HTML documents. The experimental application to a classical corpus shows that the use of structural information outperforms other classical models.

Le développement du document électronique et du Web ont vu émerger puis s'imposer des formats de données structurés, tels que le SGML et le HTML, permettant de représenter l'information sous une forme plus riche que le simple contenu et adaptée à des besoins spécifiques. Aujourd'hui, des propositions de format comme RdF et des langages de descriptions comme XML sont en train de s'imposer. Ces nouveaux formats permettent de représenter conjointement l'information textuelle et l'information de structure d'un document. A coté de cela, les modèles classiques de recherche d'information et de classification de documents ont été principalement conçus pour traiter des documents plats sans prendre en compte d'aucune manière les informations de structure. En classification de documents qui est le sujet de cet article, la structure du document joue un rôle important. D'une part les mots n'auront pas le même rôle ni la même importance suivant leur place dans le document (titre, mots clé, profondeur, méta-donnée, etc). D'autre part, des documents complexes peuvent appartenir à une classe même si une seule de leurs composantes est pertinente pour cette classe, or, cette information est souvent noyée dans les codages classiques. Quelques travaux commencent cependant à aborder ce problème.

Nous nous intéressons ici à la classification de documents structurés avec prise en compte simultanée du contenu et de la structure et proposons deux modèles originaux pour cela. Le premier est un modèle génératif qui utilise le formalisme des réseaux Bayésiens. En

s'appuyant sur ce premier modèle, nous construisons ensuite un modèle discriminant en utilisant la technique du noyau de Fisher.

L'article est organisé comme suit. Nous faisons tout d'abord un état de l'art sur la classification de documents structurés. Ensuite, nous introduisons la notion de structure de document et proposons un modèle génératif associé à cette structure. Nous montrons ensuite comment créer à partir de ce modèle génératif un modèle discriminant à l'aide du noyau de Fisher. Enfin, nous présentons une série d'expériences sur une base de données de référence. Nous montrons que les deux méthodes utilisant la structure permettent une diminution des erreurs de classification par rapport aux modèles classiques qui travaillent sur des représentations plates, puis nous discutons les perspectives des modèles présentés.

1. Etat de l'art

La classification de textes est une des tâches classiques de la recherche d'information et, en tant que telle, elle a suscité de nombreuses études. Les modèles de classification de texte, considèrent, pour la plupart, des documents plats avec des représentations « sac de mots » qui ne prennent pas en compte l'ordre des mots dans les documents ni leur structure. Parmi les méthodes génératives (i.e. qui estiment les densités), le modèle naïve Bayes est l'un des plus utilisés. Parmi les méthodes discriminantes, différentes techniques ont été utilisées comme les Réseaux de Neurones [Schu95] ou les Machines à Vecteurs Support (MVS) [Joa98]. Les modèles discriminants offrent en général de meilleures performances. Plus récemment, quelques modèles prenant en compte l'information de séquence dans les documents notamment à l'aide de Modèles de Markov Cachés (MMC) ont été proposés [Den00]. On pourra consulter [Seb02] et dans [Sah98] pour une revue des modèles de classification des documents plats.

Le développement du Web a motivé plusieurs études sur la classification de pages HTML. Des informations de nature différente sont présentes dans ces pages (titre, liens qui pointent vers la page, texte de la page, etc....) et il semble logique de les traiter de façon différenciée, c'est ce qu'ont tenté plusieurs auteurs. Les modèles proposés sont généralement des combinaisons de classifieurs de base spécialisés pour les différentes sources d'information disponible en HTML (titre, liens, etc....). Par exemple, [Quek97] compare trois classifieurs qui opèrent sur des corpus de documents au format HTML : un classifieur prend en compte uniquement les textes plats, un deuxième exploite les titres des documents et des sections et le dernier utilise les textes d'hyperliens. Les expériences sur le corpus WebKB [WebKB], un des rares corpus structuré et étiqueté disponible, démontrent que les titres et les textes d'hyperliens peuvent fournir des informations pertinentes pour la classification de documents. En général, ils résument assez bien le contenu du document. Cependant, ce type de classifieur semble peu robuste face à des liens trop généraux (Return, Homepage, etc.). Dans [Cli99], Cline représente l'information de structure dans un vecteur en codant chaque composante du document, (e.g. le titre ou le corps ou les hyperliens), dans une zone spécifique du vecteur. Le codage utilisé est un classique TF-IDF calculé par composante. Ce vecteur sera ensuite exploité par un classifieur classique. Les expériences sur la base WebKB ne montrent pas d'amélioration de performances par rapport à un classifieur qui opère sur le document plat. Dumais et Chen dans [Dum00], utilisent la structure HTML pour sélectionner des parties de documents sur lesquelles ils utilisent des

codages et des méthodes classiques de classification. La structure est ici utilisée pour détecter les parties les plus pertinentes à une prise de décision sur le document. Yang et al. [YSG02] abordent la classification des hypertextes par l'étude du comportement de trois modèles classiques, (Naïve Bayes, kppv et FOIL) qui utilisent différentes informations contenues dans l'hypertexte comme le contenu des pages filles, des tags HTML, et des méta données.

Globalement, pour tous les travaux concernant la classification de pages HTML, les méthodes proposées utilisent l'information de structure des pages de façon rudimentaire, soit afin de construire une nouvelle représentation du document, soit en combinant des classifieurs simples qui opèrent sur les différentes parties du document. Ces méthodes sont dédiées au traitement de pages HTML et ne permettent pas d'exploiter des structures plus complexes.

Au-delà de ces travaux sur le HTML dont la portée est limitée, quelques auteurs ont commencé à proposer des modèles dédiés à la classification de documents structurés et prenant naturellement en compte les informations de structure et de contenu. Ces modèles ne sont pas spécifiques au HTML, même si, faute de bases de données disponibles, les tests sont actuellement pratiqués sur des bases HTML. Ils peuvent être utilisés sur divers types de documents structurés (e.g. XML). Diligenti et al [DGF01] utilisent un modèle génératif nommé Arbre de Markov Caché (HTMM). Ils considèrent que l'arbre correspondant à un document structuré est généré par un arbre d'états cachés. Piwowarski et al. [PDG02] ont proposé un modèle de Réseau Bayésien pour la classification des documents structurés. Il s'agit d'un modèle discriminant qui calcule directement la pertinence d'un document pour une classe.

Les modèles que nous présentons appartiennent à cette deuxième catégorie dans le sens où ils peuvent traiter n'importe quel type de structure arborescente (typiquement des pages XML) et qu'ils peuvent opérer sur des données de différentes natures (e.g. multi-média) à la seule condition que l'on soit capable de calculer un score pour chaque type de données. Le modèle génératif que nous proposons a des points communs avec [DGF01] même si le formalisme adopté est différent. Il est cependant plus général et permet de prendre en compte différents types de relations entre éléments structurels. Le modèle discriminant que nous construisons dessus constitue une des premières utilisations opérationnelles des noyaux de Fisher dans le domaine du texte et n'a pas d'équivalent à notre connaissance.

2. Définition d'un document structuré

Nous allons représenter un document structuré sous la forme d'un graphe orienté sans cycle (DAG : Direct Acyclic Graph) : un nœud correspond à un élément structurel du document, un arc représente la relation d'inclusion d'un élément dans un autre (e.g. un paragraphe est inclus dans une section). Pour des raisons de simplicité, nous ne prenons pas en compte les éventuelles relations « circulaires » qui peuvent exister au sein de certains types de document et dans ce cas nous nous ramenons à une représentation sous forme d'arbre. Chaque nœud du graphe est constitué :

- d'une étiquette ; par exemple, les étiquettes peuvent être *section, paragraphe, titre* et ainsi représenter la sémantique structurelle du document.

- et si c'est le cas, d'un texte associé qui correspond à la séquence de mots associée à ce nœud.

La figure 1 donne un exemple de document structuré. Nous allons utiliser le graphe ainsi constitué afin de construire le modèle génératif.

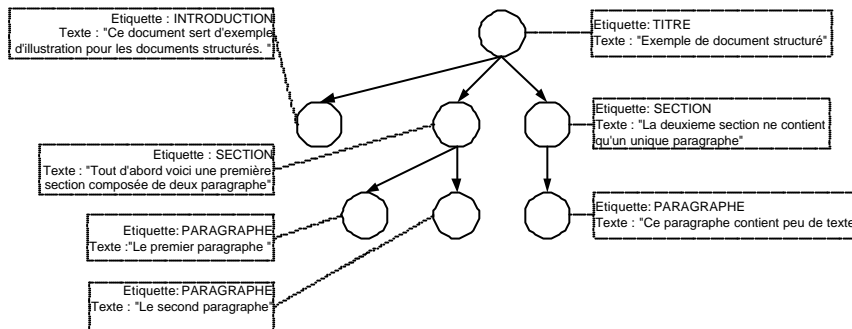


Figure 1. Représentation graphique d'un document structuré. Ce document se compose d'une introduction et de deux sections de respectivement deux et un paragraphe chacune. Chaque partie du document est représentée par un nœud composé d'une étiquette et d'un contenu textuel.

3. Modèle génératif

Nous montrons comment construire pour chaque document, un modèle génératif décrit par un réseau bayésien. Nous décrivons ensuite une instance particulière de ce modèle utilisée pour la tâche de classification de documents HTML sur laquelle s'appuient nos expériences.

3.1 Idées et hypothèses du modèle génératif

Nous allons modéliser un document structuré par un réseau Bayésien. Dans ce modèle, un document d sera la réalisation d'un ensemble de variables aléatoires associées aux nœuds du réseau. On notera $d = \{ni\}$, $i \in \hat{I} [1..|d|]$ la réalisation associée à d . Ce réseau a pour paramètres les probabilités conditionnelles $P(ni/pa(ni))$ notées \mathbf{q} par la suite d'observer ni connaissant son père $pa(ni)$. Les ni prendront leur valeur dans l'espace des étiquettes $L = \{1, \dots, p\}$ - on a p étiquettes différentes - ou des termes T qui constituent le corpus.

Pour chaque document rencontré, il nous faut donc apprendre ces paramètres \mathbf{q} afin de permettre le calcul de $P(d/\mathbf{q})$.

Les réseaux Bayésiens permettant de représenter des relations de dépendance entre les variables aléatoires utilisées pour décrire un problème ou un phénomène. L'idée à la base de ces méthodes est de spécifier, à l'aide de connaissances a priori du phénomène que l'on veut modéliser, un certain nombre de dépendances conditionnelles entre variables aléatoires. Cela permet de réduire la complexité de l'inférence et de l'apprentissage par rapport à un modèle où toutes les dépendances statistiques sont prises en compte. Un

réseau bayésien peut être représenté graphiquement par un « DAG ». Si l'on considère un ensemble de variables aléatoires $\{x_1, \dots, x_n\}$, la vraisemblance jointe

$$p(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i / x_{i-1} \dots x_1) \text{ s'exprimera sous la forme simplifiée } \prod_{i=1}^n P(x_i / pa(x_i))$$

où $pa(x_i)$ représente l'ensemble des parents de la variable x_i dans le réseau bayésien. Nous renvoyons à [Jen96] ou à [Mur01] pour une introduction à ces modèles.

Classiquement, un réseau bayésien unique est utilisé pour modéliser un problème. Par exemple, en diagnostic médical, le réseau construit décrit l'ensemble des connaissances et des données du problème. Dans notre cas, il n'est pas possible de construire un unique réseau bayésien pour tous les documents car chacun possède une structure particulière. Pour cette raison, nous nous plaçons dans un cadre proche des réseaux bayésiens dynamiques [Russell]. Nous allons construire un réseau bayésien par document. Pour permettre une estimation robuste des paramètres de ces réseaux, nous utilisons un ensemble fini de paramètres partagés par l'ensemble des réseaux. On étend ainsi au cas des structures, une approche couramment utilisée pour la modélisation de séquences avec, par exemple, des Modèles de Markov Cachés dans les domaines de la parole ou du texte [Den00].

Pour modéliser nos documents et rendre notre modèle calculable, nous allons faire un ensemble d'hypothèses sur les dépendances conditionnelles concernant les différentes parties d'un document :

Hypothèse 1 : les dépendances statistiques entre les noeuds du modèle respectent la structure du document, i.e. le réseau Bayésien associé à un document est calqué sur la structure arborescente de ce document. Cette hypothèse rend compte de la relation entre entités structurelles du document. Ainsi, la probabilité d'une sous-partie dépend uniquement de la probabilité de la partie qui la contient dans l'arbre.

Hypothèse 2 : le processus de génération du texte dépend uniquement de l'étiquette du nœud dans lequel le texte apparaît. Un même processus génératif est ainsi associé à des nœuds de même étiquette dans des documents différents. C'est ce qui permet de mettre en œuvre le partage de paramètres.

Hypothèse 3 : l'étiquette d'un nœud dépend uniquement de l'étiquette du nœud père et pas du texte de ce nœud père. Cela induit une simplification supplémentaire par rapport à l'hypothèse 1.

On peut résumer ces trois hypothèses précédentes par : « *La structure d'un document ne dépend pas du texte contenu dans ce document tandis que le texte du document dépend uniquement de l'unité structurelle qui le contient* ».

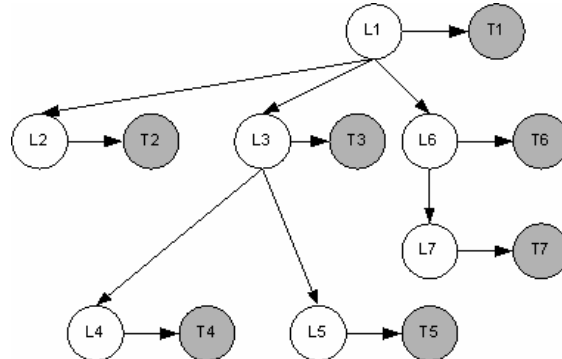


Figure 2. Le modèle génératif final du document structuré présenté dans la figure 1. Un nœud sera soit une étiquette l_i soit un contenu textuel t_i .

Ces hypothèses induisent bien sûr des simplifications sur la nature du processus génératif d'un document, mais elles vont nous permettre de construire des réseaux adaptés à notre problème, qui soient en même temps suffisamment simples pour permettre l'inférence et l'apprentissage sur de grosses bases de données textuelles.

Sous ces hypothèses, le réseau bayésien construit pour le document de la figure 1 est représenté en figure 2. Les nœuds sont soit des étiquettes l_i soit du texte t_i . La vraisemblance du réseau correspondant au document d est après ces hypothèses :

$$P(d/q) = P(l_1/q) \prod_{i=2}^{|d|} P(l_i / pa(l_i), q) \prod_{i=1}^{|d|} P(t_i / l_i, q)$$

où q désigne l'ensemble des paramètres du réseau. D'après l'hypothèse 2, les nœuds t_i dont l'étiquette possède la même valeur partageront le même paramètre, i.e. la même densité $P(t/l, q)$.

3.2. Modèles pour les pages HTML

Pour la classification de pages HTML qui est la tâche testée ici, nous allons utiliser une instance particulière du modèle général présenté précédemment. La structure d'un document HTML est donnée par les tags HTML. Par exemple, le tag <H1> signifie que le texte suivant est le titre d'un paragraphe de premier niveau ; le tag <TITLE> permet de spécifier le titre de la page HTML etc...

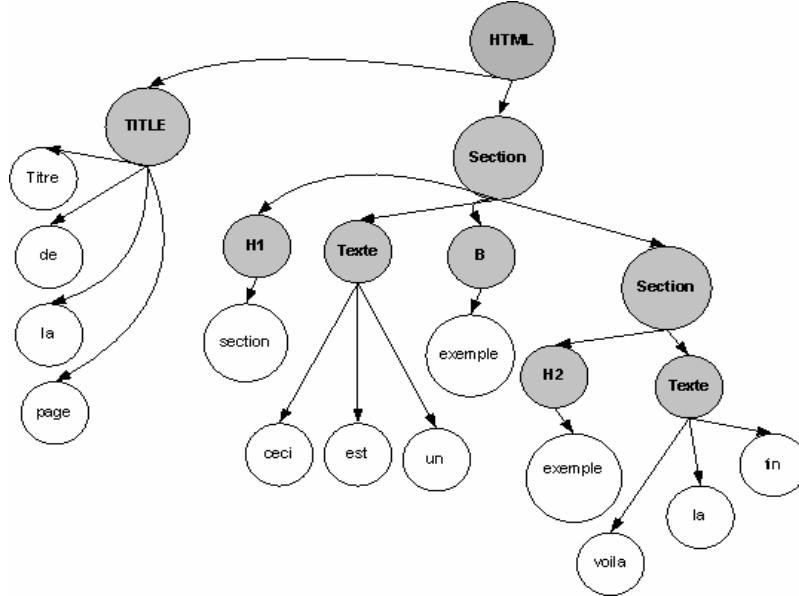


Figure 3 : le réseau Bayésien construit pour un document HTML simple.

Nous allons considérer que les étiquettes des nœuds de notre réseau sont les tags du document HTML. Ainsi, la probabilité de générer un certain texte dépend du tag qui contient ce texte et les distributions de probabilité des mots seront estimées indépendamment pour chacun des tags HTML. Un même texte sera généré avec une probabilité différente selon qu'il est un titre ou un texte en gras par exemple, les textes qui possèdent un même tag (e.g. titre) seront modélisés par la même densité. De plus, nous considérons que les termes du texte sont indépendants conditionnellement au tag (hypothèse Naive Bayes). Dans ce modèle pour le HTML, des tags « Section » et « Texte » sont rajoutés afin d'obtenir une représentation hiérarchique cohérente des documents HTML (cf. figure 3). La figure 3 représente le réseau construit pour un document HTML simple.

Notons $q_{\mathbf{y}(i), n_i, pa(n_i)} = P(n_i / pa(n_i))$ où $\mathbf{y}(n_i)$ est une fonction identifiée de manière unique la valeur de l'étiquette du nœud n_i (pour le réseau de la figure 3, cette valeur peut être « titre », « H1 », etc.) et n_i désigne un nœud du réseau (étiquette ou texte). $\mathbf{q} = \{q_{k,n,pa(n)}\}$ constitue l'ensemble des paramètres du réseau.

Avec ces notations, la vraisemblance de notre modèle de document s'écrit :

$$P(d / \mathbf{q}) = P(n_1, \dots, n_{|d|} / \mathbf{q}) = \prod_{i=1}^{|d|} q_{\mathbf{y}(i), n_i, pa(n_i)}$$

Remarque :

Dans le cas où la structure est absente du document, on retrouve un modèle naïve Bayes classique. Prenons le cas d'un document composé de la séquence de mots (w_1, \dots, w_n) . L'équation précédente devient :

$$P(d / \mathbf{q}) = \prod_{i=1}^n P(w_i / \mathbf{q})$$

Ainsi, notre modèle génératif de document structuré constitue une extension au structuré du modèle Naive Bayes qui opère sur des documents plats.

4. Un modèle discriminant basé sur les réseaux Bayésiens et le Noyau de Fisher

Les modèles génératifs permettent de modéliser des données complexes comme des séquences ou encore des arbres comme nous venons de le voir. Par contre, ils abordent le problème de discrimination de façon indirecte via l'estimation de densité. Les modèles discriminants résolvent directement le problème de discrimination et se révèlent en général bien plus efficaces pour cela, en revanche, la plupart de ces modèles ne permettent que de traiter des données vectorielles. Récemment, pour la classification de séquences biologiques, Jaakkola a proposé d'utiliser ([JH98], [JDH98]) l'information capturée dans les paramètres d'un modèle génératif pour entraîner un modèle discriminant. Cette idée a été reprise par différents auteurs [Hof00],[VG01]. Elle est attractive car elle permet d'utiliser sur des données complexes toute la palette des classifieurs vectoriels classiques. Nous allons voir que cette méthode proposée pour les séquences, s'étend aux arbres. Tout d'abord, nous introduisons le noyau Fisher qui est au cœur de cette méthode.

4.1. Noyau de Fisher

[JH98] définissent une transformation T qui associe à un exemple x et à un modèle génératif \mathbf{q} , un vecteur qui caractérise l'exemple et le modèle. La transformation utilisée est le score de Fisher $U_d = \nabla_{\mathbf{q}} \ln P(x/\mathbf{q})$, il s'agit du gradient de la log-vraisemblance de x pour le modèle génératif \mathbf{q} . Le score de Fisher mesure **combien un paramètre contribue au processus de générer un certain exemple**.

A partir de ce score ils utilisent une fonction noyau, qu'ils nomment pour l'occasion, noyau de Fisher, qui définit une distance entre exemple :

$$K(x, x') = U_x^T M^{-1} U_{x'} \text{ avec } M \text{ la matrice d'information de Fisher } M = E_x [U_x^T U_x].$$

Ils emploient ensuite un classifieur à noyaux pour classer les exemples. En pratique, il faut procéder à différentes approximations pour assurer le bon fonctionnement de cette méthode, différentes variantes (noyaux, scores) ont été proposées sans qu'il y ait pour l'instant de consensus.

4.2 Application à notre modèle

La démarche est transposable à notre modèle génératif d'arbre comme suit.

En reprenant les notations de la partie 3.2, nous avons :

$$\ln P(d / \mathbf{q}) = \sum_{i=1}^{|d|} \ln P(n_i / pa(n_i)) = \sum_{i=1}^{|d|} \ln \mathbf{q}_{\mathbf{y}^{(i), n_i, pa(n_i)}}$$

d'où :

$$\frac{\partial \ln P(d / \mathbf{q})}{\partial \mathbf{q}_{j,v,pv}} = \frac{n_{j,v,pv}}{\mathbf{q}_{j,v,pv}}$$

où $n_{j,v,pv}$ est le nombre de nœuds ni du réseau modélisant le document tels que $\mathbf{y}^{(i)} = j, n_i = v$ et $pa(n_i) = pv$, où v et pv peuvent être des étiquettes ou du texte, i.e. les nœuds dont l'étiquette vaut j , dont l'évidence et celle de son père sont respectivement v et pv . Pour nos réseaux bayésiens dans lesquels toutes les variables possèdent de l'évidence, le noyau de Fisher possède donc une expression simple et rapide à calculer.

La mise en œuvre de la méthode du noyau de Fisher est non triviale. Elle nécessite plusieurs ajustement, notamment quand le nombre de paramètres du modèle génératif est élevé. Nous avons donc effectué différentes approximations. Tout d'abord, la matrice \mathbf{M} est remplacée comme chez [JD98] par l'identité. Ensuite, nous avons calculé le gradient par rapport à $2\sqrt{\mathbf{q}_{j,v,pv}}$ comme dans [Hof00]. Si nous posons $\mathbf{r}_{j,v,pv} = 2\sqrt{\mathbf{q}_{j,v,pv}}$, nous avons :

$$\frac{\partial \ln P(d / \mathbf{q})}{\partial \mathbf{r}_{j,v,pv}} = \frac{\partial \sum_{i=1}^{|d|} \ln \frac{\mathbf{r}_{\mathbf{y}^{(i), n_i, pa(n_i)}}^2}{4}}{\partial \mathbf{r}_{j,v,pv}} = \frac{2n_{j,v,pv}}{\mathbf{r}_{j,v,pv}} = \frac{n_{j,v,pv}}{\sqrt{\mathbf{q}_{j,v,pv}}}$$

C'est cette dernière formule que nous avons utilisée afin de coder nos documents sous forme de vecteurs.

5 Expériences

5.1. Corpus

Pour nos expériences, nous avons utilisé le corpus *WebKB* fourni par CMU [WebKB]. C'est un ensemble de 8282 pages web de départements d'informatique d'universités aux Etats-Unis. Une page appartient à une parmi sept catégories: *student*, *faculty*, *course*, *project*, *department*, *staff* et *other*. Ce corpus est utilisé dans plusieurs travaux de classification de documents textuels [Joa98], [DGF01], [YSG02]. Etant donnée que la classe *other* est une classe «poubelle», nous avons fait le choix de l'exclure de nos expériences. Le corpus considéré contient alors 4520 pages.

Pour le prétraitement, nous avons effectué un stemming de Porter, utilisé une stop-list et enlevé les mots du vocabulaire présents dans moins de 5 documents. Nous avons obtenu un vocabulaire constitué de 8038 termes différents. Nous avons ensuite, à l'aide d'un analyseur que nous avons développé, conservé uniquement les tags les plus nombreux (H1...H3, TITLE, B, I, A). Ce sont ces tags qui servent d'étiquettes pour le modèle génératif.

5.2 Méthodologie

Nous avons effectué nos expériences par validation-croisée en divisant aléatoirement le corpus en 5 parties égales, l'apprentissage est réalisé sur 4 parties, le test sur la 5^e. Les performances sont mesurées par le taux de bonne classification. Etant donné que notre problème concerne des classes disjointes, cette mesure est représentative de la performance du système.

Les résultats présentés dans la table 1 sont la moyenne des résultats obtenus sur les 5 ensembles de validation-croisée.

Les modèles utilisés à titre de comparaison sont le modèle Naive Bayes, le modèle MVS présenté dans [Joa98] qui utilise une représentation plate TF-IDF du texte et un modèle issu de l'application du noyau de Fisher sur le modèle naïve Bayes¹. Ces trois modèles sont des modèles de type sac de mots et ne prennent pas en compte la structure. Ils nous permettent de savoir si oui ou non la structure des pages HTML amène de l'information supplémentaire pour la tâche de classification

Pour les modèles génératifs (Modèle génératif structuré et Naive Bayes), nous avons appris un modèle par classe. On a donc un ensemble de paramètres \mathbf{q} par classe. Un document d sera ensuite classé dans la classe qui réalise le maximum de $P(c)P(d/c, \mathbf{q})$. Les modèles naïve Bayes sont appris par simple comptage sur les fréquences, les paramètres du modèle génératif structuré sont appris par l'algorithme EM. Pour les deux modèles avec noyau de Fisher, nous avons en plus utilisé comme classifieur à noyau les Machines à Vecteurs Support (MVS) et donc appris une MVS par classe qui apprend à discriminer entre cette classe et l'ensemble des autres.

	<i>course</i>	<i>department</i>	<i>Staff</i>	<i>faculty</i>	<i>student</i>	<i>project</i>	<i>Moyenne</i>
Naive Bayes	95.3	92.7	8.6	65.4	91.5	65.5	80.8
Modèle structuré génératif	96.3	81	3.4	70.6	92.6	74.8	83.2
MVS	88.2	77.8	18.7	85.2	90.6	75.3	84.9
Naive Bayes Fisher	95.2	74.9	18	82.8	90.2	70.2	84.6
Modèle structuré Fisher	94.3	82.3	15.2	83.4	93.5	70.7	86.2

Table 1 : taux de bonne classification des différentes méthodes testées dans les expériences.

¹ Il s'agit du cas particulier du modèle présenté en § 3.2 quand la structure du document est inexistante.

La moyenne des taux de bonne classification est une moyenne pondérée par la taille des classes (macro-moyenne).

5.3 Commentaires sur les résultats

Naïve Bayes est un cas particulier de notre modèle où la structure n'est pas prise en compte. On voit en comparant respectivement les lignes 1-2 et 4-5 de la table 1 que la prise en compte de la structure par notre modèle, aussi bien dans sa version générative que dans sa version discriminante apporte une amélioration des résultats. Le modèle utilise une partie de l'information structurelle contenue dans les documents telle que le titre, les liens de la page ou bien le texte mis en évidence afin d'augmenter sa capacité à classer des documents.

Comparé à MVS qui est un des classificateurs offrant les meilleures performances sur cette tâche, notre modèle offre des performances de même niveau et même légèrement supérieures dans sa version discriminante. L'instance du modèle génératif développée pour les pages HTML possède la même complexité calculatoire² qu'un classique modèle Naive Bayes et le gain de performance ne s'effectue pas au détriment du temps de classification. La base WebKB est de petite taille, ce qui a priori devrait nuire à nos modèles qui possèdent plus de paramètres à apprendre que les modèles classiques naïve Bayes et MVS. On peut espérer que le potentiel de ces modèles sera encore plus manifeste sur des bases de plus grande taille et sur des types de documents possédant une structure plus riche. D'autres types de structure pourraient être pris en compte même sur des simples documents HTML, l'utilisation des noyaux de Fisher peut également être mise en œuvre de différentes façons.

6. Conclusion

Nous avons présenté une méthode de classification de documents structurés décrits par des arbres. Pour cela, nous avons introduit deux modèles originaux, un premier modèle génératif qui modélise la distribution des différentes composantes du document dans la structure, et un modèle discriminant basé sur des noyaux de Fisher qui est construit sur ce premier modèle. Nous avons constaté dans chaque cas un gain de performance par rapport aux modèles classiques travaillant sur les documents plats.

L'information structurelle des pages HTML est une information relativement bruitée du fait du manque de rigidité de ce standard de données et de la diversité des concepteurs de pages. Il reste à tester ces modèles sur des bases de corpus de plus grande taille et des documents avec une structure plus riche. Cette étape est la prochaine de notre travail.

² Du fait que les documents possèdent beaucoup moins de tags que de mots, la complexité de notre modèle est la même que la complexité du modèle Naive Bayes.

7. Bibliographie

- [Bur98] BURGES, C. J. C., A Tutorial on Support Vector Machines for Pattern Recognition, *Knowledge Discovery and Data Mining*, 2(2), 1998,
- [Cli99] CLINE, M., Utilizing HTML Structure and Linked Pages to Improve Learning for Text Categorization, *Department of Computer Sciences, University of Texas, Undergraduate Honors Thesis, May 1999*,
- [Den00] Denoyer L, Zaragoza L, Gallinari P, HMM-Based passage models for document classification and ranking, in *Proceedings of the 23rd BCS-IRGS (ECIR 2001)*
- [DGF01] DILIGENTI, M., GORI, M., MAGGINI, M., SCARSELLI, F., Classification of HTML documents by Hidden Tree-Markov Models, in *Proceedings of the Int. Conference on Document Analysis and Recognition (ICDAR), Seattle, WA (USA), pp. 849-853, 2001*,
- [Dum00] DUMAIS, S., CHEN, H., Hierarchical Classification of Web Content, in *Proceedings of SIGIR'00, August 2000, pp.256-263*,
- [Hof00] HOFMANN, TH., Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization, in *Advances in Neural Information Processing Systems 12, S.A. Solla, T.K. Leen, K.-R. Müller (eds.), pp. 914-920, MIT Press (2000)*
- [JDH98] JAAKKOLA, T., DIEKHANS, M., HAUSSLER, D., A discriminative framework for detecting remote protein homologies, *Journal of Computational Biology* 95-114, 1998
- [JH98] JAAKKOLA, T.S., HAUSSLER, D., Exploiting generative models in discriminative classifiers, 1998. In *Advances in Neural Information Processing Systems 11, 1998*
- [Jen96] JENSEN, F., *An Introduction to Bayesian Networks*, UCL Press, 1996.
- [Joa98] JOACHIMS, TH., Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In *Proceedings of the European Conference on Machine Learning 1998*
- [Mur01] MURPHY K.P., A Brief Introduction to Graphical Models and Bayesian Networks
- [PDG02] PIWOWARSKI, B., DENOYER, L., GALLINARI, P., Un modèle pour la Recherche d'Information sur des Documents Structurés, *6emes Journées internationales d'Analyse statistique des Données Textuelles, JADT 2002, 2002*
- [Quek97] QUEK, CH. Y., Classification of World Wide Web Documents, *School of Computer Science, Carnegie Mellon University, Senior Honors Thesis*,
- [Russell] RUSSELL S. AND NORVIG P. Artificial Intelligence : A Modern Approach (*Book*)
- [Sah98] SAHAMI, M., Using Machine Learning to Improve Information Access, *dissertation for the degree of Doctor Philosophy, Stanford University, 1998*.
- [Schu95] SCHUTZE H. HULL D. A. AND PEDERSEN J.O. A comparison of Classifiers and Document Representation for the routing problem. In *proceedings of TREC 4 (1995)*
- [Seb02] SEBASTIANI, F., Machine Learning in Automated Categorization, in *ACM Computing Surveys*, 34(1), pp.1-47, 2002.
- [VG01] VINOKOUROV A., GIROLAMI M. Document Classification Employing the Fisher Kernel Derived from Probabilistic Hierarchic Corpus Representations, in *Proceedings of the 23^{ème} BCS-IRSG European Colloquim on IR Research(ECIR 2001) pp 24-40*
- [WebKB] <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb/>
- [YSG02] YANG, Y., SLATTERY, S., GHANI, R., A Study of Approaches to Hypertext Categorization, to appear in the *Journal of Intelligent Information systems - Special Issue on Automatic Text Categorization (2002)*