Apport de la prise en compte du contexte structurel dans les modèles bayésiens de classification de documents semi-structurés

Pierre-François Marteau, Gildas Ménier, Leopold Ekamby
VALORIA, Université de Bretagne Sud, Campus de Tohannic, 56 000 Vannes
{Pierre-François.Marteau, Gildas.Menier}@univ-ubs.fr

Résumé. Nous nous intéressons dans cet article au problème de la classification supervisée de documents semi-structurés. Un modèle formel basé sur des hypothèses simples et originales à notre connaissance est proposé. Ce modèle puise ses fondements dans les modèles de classification bayésiens, en ciblant la prise en compte de la structure des documents dans les tâches de classification. Ce modèle permet d'envisager la fusion de données numériques ou symboliques structurées et de données non structurées qui peuvent faire l'objet d'une modélisation spécifique. Des versions simplifiées de ce modèle sont implémentées pour évaluer de manière comparée à d'autres approches l'impact de la prise en compte de la structure documentaire dans des tâches de classification de documents textuels. Les premiers résultats, qui confirment ceux déjà obtenu dans le cadre de travaux similaires, montrent que la prise en compte du contexte structurel d'occurrence des mots améliore de manière significative les performances d'un classifieur bayésien naïf multinomial. Cette implémentation conduit à des performances comparables à celles atteintes par les classifieurs SVM sur la tâche considérée. Une implémentation plus complète de ce modèle doit permettre d'envisager des expérimentations ou des applications plus complexes et plus riches. Ces résultats ouvrent des perspectives autour de l'exploitation d'heuristiques de pondération des estimateurs associés aux composantes structurelles des documents.

1 Introduction

Les volumes croissants d'information stockée tant dans les bases documentaires traditionnelles que sur le World Wide Web constituent une motivation de plus en plus pressante pour le développement de méthodes de classification et de filtrage automatique de documents performantes, ceci afin de mieux organiser et structurer ces données et en faciliter l'accès. Dans cet article, nous nous intéressons à la classification de documents textuels semi-structurés au format XML. La recherche sur les méthodes de classification de documents textuels a conduit à une production impressionnante d'articles depuis les années 1990 (Cf. [Yang, 1999] pour une synthèse récente de ces travaux). Une liste non exhaustive d'approches basées sur les techniques d'apprentissage automatique inclut le Classifieur Bayésien Naif (CBN) [Duda & Hart, 1973], [Lewis et Ringuette 1994], [McCallum et Nigam 1998], les k plus proches voisins (k-NN) [Yang, 1999], les Machine à support vectoriel (SVM) [Vapnik, 1995,98], [Joachims, 1999], [Dumais et al., 1998], le boosting appliqué à la

classification de textes (BCT) [Shapire et Singer, 2000], et les algorithmes basés sur l'apprentissage de règles de décision (ARD) [Apté et al., 1994].

Il est notable que ces approches conduisent à des performances de classification similaires dès lors que les données d'apprentissage sont en nombre suffisant [Yang et Xin, 1999]. Dans cet article nous nous intéressons plus particulièrement a la méthode du Classifieur Bayesien Naïf (CBN). L'algorithme du CBN ne prend pas en compte l'ordre d'apparition des mots dans le document, et s'appuie même sur l'hypothèse apparemment irréaliste d'indépendance des mots conditionnellement à la classe d'appartenance. En dépit de ces simplifications grossières et non satisfaites dans la pratique, cet algorithme se montre d'une précision comparable aux autres méthodes plus sophistiquées d'apprentissage (SVM, kNN …). Domingos et Pazzanni [Domingos et Pazzanni, 1997] ont notamment montré expérimentalement que l'erreur de classification de cet algorithme reste proche des valeurs optimales, y compris dans des cas où l'hypothèse d'indépendance n'est manifestement pas vérifiée.

Cependant, si l'on considère les documents semi-structurés (de type XML), le contexte d'occurrence des mots prend une importance encore plus grande, et à ce titre, l'impact de l'influence de ce contexte sur les risques d'erreur de classification mérite d'être analysé et estimé en détail. C'est l'objet du travail rapporté dans cet article. Deux formes de prise en compte du contexte sont développées dans la littérature :

- Premièrement, il s'agit de tirer profit du supplément d'information qu'apporte la localisation des mots dans la structure du document XML. Il a été montré par [Cline, 1999] sur des documents HTML que cette approche améliore sensiblement la performance du modèle CBN. Dans une logique similaire, le modèle que nous proposons intègre une composante structurelle associée à chaque mot du texte traité. Plus précisément, chaque mot d'un document XML est contextualisé par l'élément XML ou l'arborescence des éléments XML qui le contiennent.

- Deuxièmement, et éventuellement de manière conjointe, il s'agit de prendre en compte le fait que certains mots apparaissent en séquence (et non plus de manière indépendante) au sein du même élément structurel du document XML. Autrement dit, l'hypothèse d'indépendance conditionnelle des mots par rapport à la classe que nous évoquions ci-dessus est relâchée. Cette piste a été largement explorée pour améliorer les performances du CBN dans le cadre de l'exploitation de modèles n-gram de langage [Brown et al., 1992], [Cavnar et Trenkle, 1994], [Huffman et Damashek, 1994]. Plus récemment, un travail de Rish [Rish, 2001] a montré que l'exploitation des dépendances fonctionnelles entre les attributs pouvait accroître la précision des classifieurs Bayesien dans certains cas. Mais l'étude la plus générale et la plus convaincante dans cette voie est sans doute celle de [Peng et Schuurmans, 2003] qui en introduisant le *Chain Augmented Naive Bayes* (CAN) ont montré qu'en relâchant l'hypothèse d'indépendance et en lui substituant une dépendance de Markov d'ordre n (à fixer) entre les variables on améliore sensiblement la précision du classifieur Bayesien.

Notre objectif étant la prise en compte la notion de contexte au sens le plus large dans une tâche de classification, nous développons un modèle formel général de classification bayésienne, susceptible d'incorporer les deux aspects du contexte que nous venons de souligner, en permettant par exemple de fusionner des modèles n-gram associés à des éléments textuels avec des descripteurs caractérisant la structure des documents. Nous validons une implémentation simplifiée de ce modèle en l'évaluant de manière comparative avec des modèles classiques tels que les k plus proches voisins, le classifieur bayésien naif, les SVM, etc. Cette évaluation repose sur des tâches de classification de référence construites à partir d'extraits de la base de données Reuters [Reuters-21578].

2 Le Classifieur Bayesien Naïf (CBN)

Si l'on considère un ensemble $\Omega = \{\omega_1, \omega_2, ..\omega_{|\Omega|}\}$ de catégories, le problème de classification posé est celui d'attribuer à un document *d* l'une des catégories précédentes. En général, on utilise un algorithme d'apprentissage supervisé pour entraîner le classifieur ; à partir d'un ensemble de *N* documents pré-étiquetés $\{(d_i, \omega_i), 1 \le i \le N\}$, l'algorithme d'apprentissage va élaborer une fonction de classification $F : D \rightarrow \Omega$, ou *D* représente l'ensemble des documents. L'algorithme d'apprentissage dans le cas du CBN est basé sur une simple application de la formule de Bayes qui pour un document *d* donné exprime la probabilité d'appartenance de celui-ci à une classe ω:

$$P(\omega \mid d) = \frac{P(d \mid \omega) * P(\omega)}{P(d)} \qquad (1)$$

Si l'on représente un document, comme il est souvent d'usage en classification textuelle, par un vecteur de *K* attributs $d=(v_1, ..., v_K)$, et si l'on exploite au second membre de (1) l'hypothèse d'indépendance de ces attributs conditionnellement à la classe ω, l'égalité (1) devient :

$$P(\omega \mid d) = \frac{P(\omega) * P((v_1, ..., v_K) \mid \omega)}{P(d)} = \frac{P(\omega) * \prod_{j=1}^{K} P(v_j \mid \omega)}{P(d)} \qquad (2)$$

A partir de l'équation (2) on finalise l'élaboration de la fonction discriminante du CBN en prenant le maximum des probabilités conditionnelles $P(\omega|d)$ sur l'ensemble Ω des classes par la formule :

$$\omega^* = \arg\max_{\omega \in \Omega} \{P(\omega \mid d)\}$$

$$\omega^* = \arg\max_{\omega \in \Omega} P(\omega) * \left\{ \frac{\prod_{j=1}^{K} P(v_j \mid \omega)}{P(d)} \right\} = \arg\max_{\omega \in \Omega} \left\{ P(\omega) * \prod_{j=1}^{K} P(v_j \mid \omega) \right\} \qquad (3)$$

(car $P(d)$ est une constante pour toute classe ω) ;

Il existe plusieurs approches pour estimer les paramètres ($\{P(v_j \mid \omega)\}_{1\le j\le K}$) du CBN dont le modèle binaire indépendant, le modèle multinomial, le modèle de Poisson, le modèle binomial négatif etc. Les représentations du document diffèrent dans chacun de ces modèles. Dans le modèle binaire indépendant par exemple, un document *d* se représente par un ensemble de K attributs $d=(v_1, \ldots\ldots, v_K)$, ou l'attribut v_j a la valeur 1 si le mot est présent dans le document, et 0 si le mot est absent. Tandis que dans le modèle multinomial un document se représente par un ensemble de *K* attributs $d=(v_1,\ldots\ldots,v_K)$, tel qu'un attribut v_j vaut le nombre d'occurrences du mot *j* dans le document *d*. Synthétiquement, on dit dans ce cas qu'un document est un ensemble de mots avec leurs occurrences (« bags of words »). Plusieurs études ont montré que le modèle multinomial est le meilleur pour les applications de classification de documents textuels [Eyheramendy et al., 2003], [McCallum et Nigam, 1998]. C'est donc ce dernier modèle que nous allons considérer.

L'ensemble des paramètres qui caractérisent le CBN multinomial pour *K* attributs et $|\Omega|$ classes est le suivant :

$$\left\{\theta_j^{\omega} = P(v_j \mid \omega) : 1 \le j \le K; \quad 1 \le \omega \le |\Omega|\right\}$$

Pour un document *d* et une classe ω donnée, la probabilité *a priori* (probabilité du document conditionnellement à la classe) :

$$P(d \mid \omega) = \frac{N^d!}{\prod_{j=1}^{K} N_j^d!} \quad \prod_{j=1}^{K} \left(\theta_j^{\omega}\right)^{N_j^{\omega}} \tag{4}$$

où N_j^d est la fréquence de l'attribut *j* dans *d* ; et $N^d = \sum_{j=1}^{K} N_j^d$.

Remarquons qu'en pratique on évalue θ_j^{ω} par l'estimateur de Laplace :

$$\theta_j^{\omega} = \frac{N_j^{\omega} + 1}{N^{\omega} + K} \tag{5}$$

où N_j^{ω} est la fréquence de l'attribut *j* dans ω, et $N^{\omega} = \sum_{j=1}^{K} N_j^{\omega}$

On évite en général un estimateur du type maximum de vraisemblance $\theta_j^{\omega} = \frac{N_j^{\omega}}{N^{\omega}}$, car pour un document d_p la probabilité $P(d_p|\omega)$ s'annule dès qu'un attribut *j* n'apparaît pas dans d_p, ce qui rend impossible la classification d'un tel document.

3 Modélisation du contexte structurel XML

Un document XML peut être représenté sous la forme d'une structure arborescente appelée DOM (Document Object Model) dont les feuilles contiennent des éléments d'information textuels ou binaires (TEXT ou CDATA) et dont les nœuds correspondent à des éléments XML auxquels sont éventuellement associés un ensemble de couples (attributs, valeurs). Les feuilles de cet arbre peuvent faire référence à des éléments externes tels que des images, des bandes sonores, des vidéos ou du texte. Pour tout document XML bien formé, cet arbre peut-être extrait en utilisant un analyseur syntaxique (parseur) XML [IBM, 2000], [XERCES, 2002]. Un exemple de document XML contenant une dépêche de la base de données «Reuters » est présenté en figure 1, l'arbre DOM qui lui correspond est présenté en figure 2.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<FILE>
 <REUTERS TOPICS="NO" LEWISSPLIT="TEST" CGISPLIT="TRAINING-SET"
OLDID="204" NEWID="210">
    <DATE>19-OCT-1987 15:27:23.12</DATE>
   <TOPICS> <D>earn</D> </TOPICS>
   <PLACES> <D>usa</D> </PLACES>
   <PEOPLE/> <ORGS/> <EXCHANGES/> <COMPANIES/>
   <UNKNOWN> 5; 5; 5;F 22; 22; 1;f2832 31;reute h f BC-
      LANE-TELECOMMUNICATIO 10-19 0080 </UNKNOWN>
   <TEXT> 2;
      <TITLE>LANE TELECOMMUNICATIONS PRESIDENT
      RESIGNS</TITLE>
      <DATELINE> HOUSTON, Oct 19 - </DATELINE>
      <BODY>
Lane Telecommunications Inc lt;LNTL.O> said Richard Lane, its
president and chief operating officer, resigned effective Oct 23.
Lane founded the company in 1976 and has been its president since
its inception, …;
      </BODY>
   </TEXT>
 </REUTERS>
</FILE>
```

FIG. 1 - *Exemple de fichier XML extrait du corpus Reuters ([Reuters 2000]).*

La structure des documents XML varie en général d'un document à l'autre. Elle est implicitement véhiculée par le document lui-même et pas systématiquement déterminée par une organisation structurelle globale stable (DTD). Pour cette raison, les documents XML sont considérés plutôt comme des documents semi-structurés, par opposition aux documents structurés pour lesquels aucune irrégularité structurelle n'existe [Abiteboul, 1997].

Un document *d* semi-structuré est représentable par un arbre (graphe connexe acyclique) noté T_d. constitué d'un ensemble de sommets *Sd* et d'un ensemble d'arcs A_d.

En suivant les développements initiés dans le domaine de l'indexation et de la recherche d'information dans les banques semi-structurées [Ménier & Marteau 2002], [Marteau & Ménier 2003], nous considérons que dans la structure de l'arbre DOM associé à un document XML, chaque nœud *n* (en particulier chaque feuille) peut être rattachée à la racine de l'arbre par l'intermédiaire d'un chemin que l'on note *c(n)*.

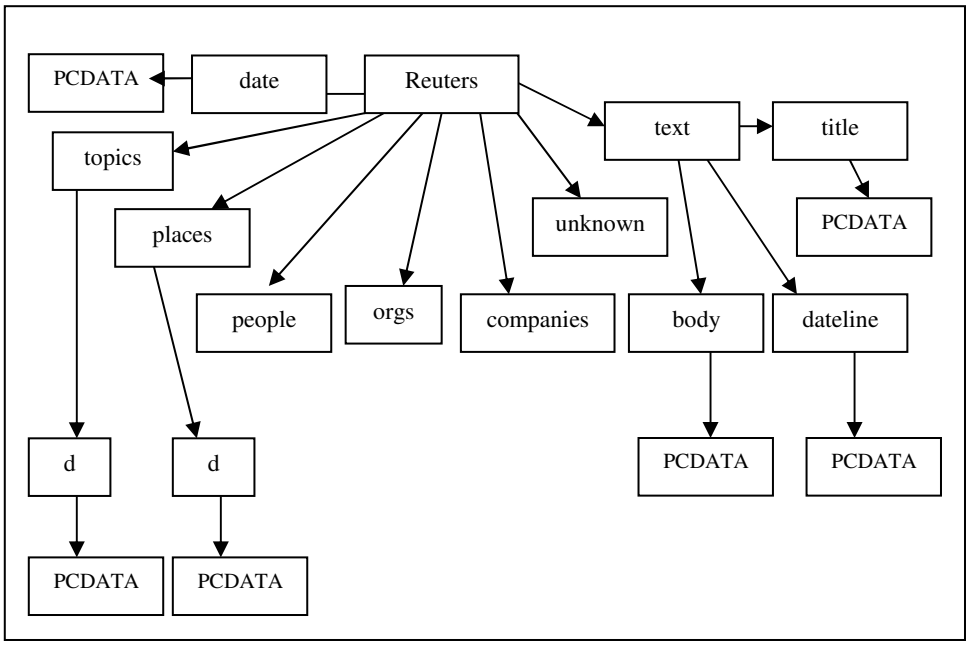

FIG. 2 - *Arbre DOM correspondant à l'exemple présenté en figure 1. L'élément PCDATA correspond à des feuilles décomposables (éléments textuels par exemple)*

Ce chemin est une suite ordonnée d'éléments XML. Il constitue le contexte d'occurrence du nœud *n* dans le document considéré. Dans le cas où une feuille *l* de l'arbre DOM peut être décomposée en sous éléments $\{v_i\}$, on considérera que chaque sous élément v_i est rattaché au contexte XML *c(l)*. En particulier, si une feuille de l'arbre DOM est identifiée à une entité textuelle, chaque mot v_i (lemme ou chaîne de caractères) présent dans l'entité textuelle pourra être également rattaché au contexte XML *c(l)*. Ce contexte XML associé à la position de v_i dans la feuille *l* caractérise l'occurrence de v_i dans le document.

Plus précisément, c*(n)* est identifiable à la séquence ordonnée des éléments XML attachés aux nœuds sur le chemin qui conduit du nœud n à la racine de l'arbre DOM :

$c(n) =< e(n_0), a(n_0) >< e(n_1), a(n_1) > ... < e(n_p), a(n_p) >$ avec :

n_0 est le nœud racine de l'arbre,
n_p est le nœud père de n_{p-1},
$p+1$ est la longueur de la séquence de nœuds contenus dans le chemin $c(n)$.
$e(n_i)$ est l'élément XML attaché au nœud n_i (« TITLE », « BODY », « PCDATA », etc.)
$a(n_i)$ est l'ensemble des couples <attribut, valeur> éventuellement attaché au nœud n_i.

Dans le cas où le nœud considéré est une feuille l décomposable, chaque sous-élément v_i qu'il contient est considéré comme un nœud terminal ne possédant pas d'attribut. Dans ce cas, on considèrera que le contexte XML d'occurrence du sous élément v_i est identifiable au contexte d'occurrence du nœud l :

$c(v_i) =< e(n_0), a(n_0) >< e(n_1), a(n_1) > ... < e(n_p), a(n_p) >= c(l)$

avec n_p nœud père de l.

On pourra distinguer deux cas pour la décomposition d'un nœud feuille l:

- l est assimilable à un ensemble de sous éléments $\{ v_i \}$ et dans ce cas on écrira : $e(l) = \{v_i\}$.
- l est assimilable à une séquence de sous éléments $v_1 v_{2...} v_k$ et dans ce cas on écrira : $e(l) = v_1 v_{2...} v_k$.

C'est sur la base de cette définition du contexte structurel que nous développons notre modèle de classification bayésienne à contexte augmenté.

4 Le Classifieur Bayésien à Contexte Structurel (CBCS)

Nous proposons et décrivons dans cette section un modèle formel qui intègre dans le cadre général de la classification bayésienne le contexte structurel d'occurrence des éléments de données non structurés.

Dans le cadre d'une classification bayésienne d'un ensemble de documents, les probabilités a *posteriori* de choisir une classe ω étant donnée un document d sont reliées par la loi de Bayes aux probabilités conditionnelles $P(d|\omega)$ conformément à l'équation (1).

Les probabilités conditionnelles $P(d|\omega)$ doivent donc être estimées. Si l'on accepte pour les documents semi-structurés la représentation sous la forme d'arbre, on est amené à assimiler $P(d|\omega)$ à la probabilité conditionnelle $P(T_d|\omega)$ où T_d est l'arbre associé au document d.

Deux difficultés émergent à ce stade :

i) des hypothèses doivent être formulées pour décomposer et simplifier l'estimation de cette probabilité,

ii) compte tenu de la variabilité et de l'hétérogénéité qui caractérisent les documents semi-structurés, le nombre de paramètres à prendre en compte croît très rapidement en fonction, de la taille du domaine traité. La tâche d'apprentissage nécessite alors une masse importante de données pas toujours disponibles.

Pour répondre à ces deux difficultés liées à la complexité des données, plusieurs approches simplificatrices ont été proposées pour tenir compte simultanément du contenu et de la structure documentaire. Parmi elles, on peut citer l'approche dite de « splitting » qui consiste a exploiter autant de classifieurs qu'il y a de composantes structurelles à prendre en compte. La classification finale est obtenue grâce à une fonction discriminante qui combine les prédictions des classifieurs élémentaires. L'approche du « modèle vectoriel structuré » (MVS) [Yi et Sundaresan, 2000] et l'approche développée par [Denoyer et al. 2003] pour classer des documents multimédia structurés (CDMS) : Strucutured Multimedia Document Classifier) sont des exemples de méthodes qui généralisent le principe du splitting.

Le MVS est une structure vectorielle pour la représentation d'arbres. Les auteurs ont développé un modèle probabiliste caractérisé par des fréquences locales de termes qui dépendent de la localisation précise du contenu textuel au sein de la structure documentaire. Les auteurs montrent que sur deux tâches de classification distinctes pour lesquelles un grand nombre de données d'apprentissage, leur modèle diminue de manière très probante les taux d'erreur comparativement aux résultats obtenus par un modèle vectoriel classique ne prenant pas en compte la structure du document.

Le CDMS est principalement basé sur deux hypothèses simplificatrices qui facilitent le problème d'estimation des paramètres en relâchant certaines dépendances entre variables du modèle. La première hypothèse considère que les contenus d'information associés aux différents noeuds de l'arbre sont indépendants les uns des autres, étant donnée la structure du document. La seconde hypothèse stipule que le contenu d'information dépend seulement du nœud auquel il est rattaché. Les auteurs montrent que sur des données HTML collectées sur la toile (WEB) [Netproject, 2001], les performances de classification sur une tâche binaire (2 classes) de leur modèle surpasse de manière très significative celles d'un CBN ne prenant pas en compte la structure des documents.

Le modèle de prise en compte de la structure documentaire que nous proposons se situe entre les deux approches précédentes.

Nous proposons de décomposer la probabilité $P(T_d|\omega)$ en posant deux hypothèses de la manière suivante :

$$\text{(H1) :} \quad \begin{aligned} P(T_d/\omega) &= P(r,T_1,T_2,...,T_k/\omega) \\ &= P(T_1,T_2,...,T_k/r\,\omega).P(r/\omega) \end{aligned} \tag{6}$$

où r est le nœud racine de l'arbre T_d et $\{T_i\}$ l'ensemble des sous arbres fils issus de r. Par simplification d'écriture et abus de langage, nous identifions en fait par r le couple $<e(r),a(r)>$ constitué de l'élément XML et de l'ensemble des attributs associés au nœud r. Cette première hypothèse stipule que l'ordre des sous arbres fils du nœud r n'a pas d'importance.

$$\text{(H2) :} \quad \begin{aligned} P(T_d/\omega) &= P(T_1/r\,\omega).P(T_2/r\,\omega)...P(T_k/r\,\omega).P(r/\omega) \\ &= P(r/\omega).\prod_{i=1}^{k} P(T_i/r\,\omega) \end{aligned} \tag{7}$$

Cette hypothèse admet l'indépendance conditionnelle des sous arbres fils étant données la racine r et la classe ω.

Moyennant l'hypothèse relativement faible (H1) qui permet de décomposer la probabilité d'avoir l'arbre T_d étant donné la classe ω en la probabilité d'avoir la racine r de l'arbre et conjointement l'ensemble des sous arbres issus de r conditionnellement à la classe ω et moyennant l'hypothèse assez forte (H2) qui sous-tend l'indépendance conditionnelle des sous arbres, on aboutit ainsi la formulation récursive traduite par l'équation (7).

Ainsi, $P(T_i|r\ \omega)$ est récursivement décomposable en $P(r_i/r\,\omega).\prod_j P(T_{i,j}/r\,r_i\omega)$, où r_i est la racine du sous arbre T_i et $T_{i,j}$ les sous arbres fils issus de r_i.

On peut noter que, r_i étant un fils de la racine r de l'arbre T_d, r est identifiable au contexte XML du nœud r_i tel que défini au paragraphe précédent, i.e. $c(r_i)$. De même, $r\,r_i$ est le contexte XML du nœud racine du sous arbre $T_{i,j}$.

D'où : $P(r_i/r\,\omega).\prod_j P(T_{i,j}/r\,r_i\omega) = P(r_i/c(r_i)\,\omega).\prod_j P(T_{i,j}/c(r_j)\,\omega)$

Nous obtenons donc en fin de récursion la formulation simple suivante :

$$P(T_d / \omega) = \prod_{n_i \in S_d} P(n_i / c(n_i)\,\omega) \qquad (8)$$

La formule (8) exprime donc simplement, que sous les hypothèses (H1) et surtout (H2), l'estimation de la probabilité de l'arbre T_d associé au document semi-structuré d conditionnellement à la classe ω se réduit au produit des probabilités pris sur l'ensemble des nœuds n_i de l'arbre T_d de la probabilité d'occurrence du nœud n_i étant donnée la classe ω et le contexte d'occurrence XML du nœud n_i dans l'arbre T_d. L'exemple suivant, présenté en FIG. 3, précise à la fois le mécanisme récursif proposé et sa portée :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<FILE>
  <REUTERS NEWID="210">
            <TITLE>
                  texte1…
            </TITLE>
            <BODY>
                  texte2…
            </BODY>
      </TEXT>
  </REUTERS>
</FILE>
```

FIG. 3 - *exemple simplifié de document XML extrait de la base REUTERS ([Reuters 2000]).*

Ainsi, l'estimation de la probabilité d'occurrence pour l'exemple simplifié de document XML présenté en figure 3, étant donnée une classe ω sera estimée de la manière suivante :

$$P(T_d / \omega) = P(< FILE, \varnothing > \mid \omega).$$
$$P(< \text{REUTERS}, \{\text{NEWID} = « 210 »\} > \mid < FILE, \varnothing >, \omega).$$
$$P(< TITLE, \varnothing > \mid < FILE, \varnothing >< \text{REUTERS}, \{\text{NEWID} = « 210 »\} >, \omega).$$
$$P(<"texte1...", \varnothing > \mid < FILE, \varnothing >< \text{REUTERS}, \{\text{NEWID} = « 210 »\} >< TITLE, \varnothing >, \omega).$$
$$P(< BODY, \varnothing > \mid < FILE, \varnothing >< \text{REUTERS}, \{\text{NEWID} = « 210 »\} >, \omega).$$
$$P(<"texte2...", \varnothing > \mid < FILE, \varnothing >< \text{REUTERS}, \{\text{NEWID} = « 210 »\} >< BODY, \varnothing >, \omega)$$

La formulation récursive proposée permet d'appréhender de manière approchée l'arbre représentatif d'un document semi-structuré T_d sous la forme d'un ensemble de nœuds associés aux chemins les reliant à la racine de l'arbre $T_d : \{< n_i, c(n_i) >\}_i$. Cette formulation est établie en posant une hypothèse simple d'indépendance des sous arbres étant donné leur nœud père et la classe ω.

Dans l'équation (8), les grandeurs $P(n_i / c(n_i)\omega)$ traduisent des relations de dépendance complexes qui doivent être explicitées et qui peuvent faire l'objet d'hypothèses supplémentaires :

- par exemple, on peut considérer l'exploitation d'une loi markovienne pour limiter la dépendance du nœud n_i aux premiers éléments de la séquence définissant le contexte structurel $c(n_i)$,

$$P(n_i / c(n_i)\omega) = P(< e(n_i), a(n_i) > / c_p(n_i)\omega)$$

où $C_p(n_i)$ est le contexte du nœud i limité à l'ordre p.

- on peut considérer dans certains cas l'indépendance conditionnellement à la classe des attributs et du contexte $c(n_i)$:

$$\begin{aligned} P(n_i / c(n_i)\omega) &= P(< e(n_i), a(n_i) > / c(n_i)\omega) \\ &= P(< e(n_i) > / c(n_i)\omega).P(< a(n_i) > / < e(n_i) > \omega) \end{aligned}$$

- on peut également considérer des lois paramétriques pour modéliser les distributions associées aux attributs numériques, etc.

En ce qui concerne les feuilles décomposables l de l'arbre T_d, nous avions considéré deux cas qui conduisent aux situations suivantes :

- l est assimilable à un ensemble de sous éléments $\{v_i\}$ et dans ce cas, $e(l) = \{v_i\}$ et :

$$\begin{aligned} P(l / c(l)\omega) &= P(< e(l), a(l) > / c(l)\omega) \\ &= P(e(l) / a(l)c(l)\omega).P(a(l) / c(l)\omega) \\ &= P(\{v_i\} / a(l)c(l)\omega).P(a(l) / c(nl)\omega) \end{aligned}$$

Si l'on considère par ailleurs l'indépendance des caractéristiques on obtient :

$$P(l / c(l)\omega) = P(a(l) / c(l)\omega) \prod_i P(v_i / a(l)c(l)\omega) \qquad (9)$$

- l est assimilable à une séquence de sous éléments $v_1 v_{2...} v_k$ et dans ce cas on aura : e(l) = $v_1 v_{2...} v_k$

$$P(l/c(l)\omega) = P(v_1 v_2 ... v_i .../a(l)c(l)\omega).$$
$$P(a(l)/c(l)\omega)$$

Dans ce cas, l'utilisation de modèles markoviens de type n-gram est envisageable pour tenir compte de la position d'occurrence des caractéristiques au sein de la séquence qui constitue la décomposition du nœud l. Par exemple, pour une dépendance markovienne à l'ordre n, on pourra construire le modèle caractérisé par la formule (12).

$$P(l/c(l)\omega) = P(a(l)/c(l)\omega).\prod_i P\left(v_i \mid v_{i-n+1} v_{i-n+2} .....v_{i-1}/a(l)c(l)\omega\right) \qquad (10)$$

On notera que, pour ce modèle, la localisation du contenu documentaire au sein de la structure est moins précis que dans l'approche MVS [Yi et Sundaresan, 2000] dans la mesure où l'information non structurelle est référencée de manière non unique par le chemin reliant la racine de l'arbre T_d au nœud auquel cette information est attachée. Cette localisation est par contre plus précise que dans l'approche CDMS [Denoyer et al. 2003] qui ne considère qu'une dépendance locale du contenu uniquement vis-à-vis du nœud de rattachement. De plus, le CDMS considère une dépendance structurelle des nœuds vis-à-vis de leurs parents (ordre 1), alors que le modèle proposé considère ici également une dépendance des nœuds vis-à-vis des chemins qui les relient à la racine de l'arbre T_d.

5 Modèle CBCS simplifié

Compte tenu de la complexité et de la portée assez générale du modèle CBCS développé en section 4, nous limitons les expérimentations à un sous modèle référencé par CBNCS, avec pour objectif l'analyse de l'impact de la prise en compte du contexte XML dans les tâches de classification de données textuelles.

Le modèle CBNCS (N pour naïf) repose sur les hypothèses (H1) et (H2) et sur l'hypothèse d'indépendance des caractéristiques de représentation des éléments non structurés, par exemple les mots des éléments textuels. Ce modèle n'implémente donc pas les modèles de langage de type n-gram pour la caractérisation des éléments textuels.

Les équations (3), (8) et (9) définissent le modèle CBNCS :

$$\omega^* = \arg\max_{\omega \in \Omega} \{P(\omega \mid T_d)\}$$

$$P(T_d / \omega) = \prod_{n_i \in S_d} P(n_i / c(n_i)\,\omega)$$

$$P(n_i / c(n_i)\,\omega) = P(< a(n_i), e(n_i) > / c(n_i)\,\omega) \qquad (11)$$

$$P(l / c(l)\,\omega) = P(a(l) / c(l)\,\omega) \prod_i P(v_i / a(l)\,c(l)\,\omega)$$

avec :

- T_d est l'arbre représentant le document semi-structuré d,
- n_i est un nœud de l'arbre T_d,
- l est une feuille de l'arbre décomposable en un ensemble de caractéristiques indépendantes,
- $c(n_i)$ est le chemin conduisant du nœud n_i à la racine de l'arbre T_d,
- $a(n)$ est l'ensemble des relations attributs/valeur attachées au nœud n.
- $e(n)$ est l'identificateur du nœud n.

L'estimation des probabilités $P(v_i / a(l)\,c(l)\,\omega)$ est effectuée conformément au modèle multinomial caractérisé par les équations (4) et (5).

$$\prod_i P(v_i / a(l)\,c(l)\,\omega) = \frac{N^{l,\omega}!}{\prod_i N_i^{l,\omega}!} \quad \prod_i \left(\theta_i^{l,\omega}\right)^{N_i^{l,\omega}} \qquad (12)$$

où $N_i^{l,\omega}$ est la fréquence de l'attribut i dans l'élément identifié par le nœud l de l'arbre T_d pour la classe ω ; et $N^{l,\omega} = \sum_i N_j^{l,\omega}$.

On évalue $\theta_i^{l,\omega}$ par l'estimateur de Laplace : $\theta_i^{l,\omega} = \frac{N_i^{l,\omega}}{N^{l,\omega} + K}$ (13)

Dans le cadre des expérimentations, compte tenu de la base de données exploitée, nous n'avons pas tenu compte des relations « attribut/valeur » ce qui conduit aux équations simplifiées suivantes :

$$\begin{aligned} &\omega^* = \arg\max_{\omega\in\Omega}\{P(\bar{\omega}/T_d)\} \\ &P(T_d/\omega) = \prod_{n_i\in S_d} P(n_i/c(n_i)\omega) \\ &P(n_i/c(n_i)\omega) = P(e(n_i)/c(n_i)\omega) \\ &P(l/c(l)\omega) = \prod_i P(v_i/c(l)\omega) \end{aligned} \quad (14)$$

En pratique, on exploite la fonction discriminante suivante :

$$\omega^* = \arg\max_{\omega\in\Omega}\{Log(P(T_d/\omega).P(\omega))\}, \text{ soit :}$$

$$\omega^* = \arg\max_{\omega\in\Omega}\left\{Log(P(\omega)) + \sum_{n_i\in S_d} Log(P(n_i/c(n_i)\omega))\right\} \quad (15)$$

6 Modèle CBCS Pondéré (CBCSP)

La formule (15) montre que chaque composante XML, associée à un nœud n_i, participe à l'élaboration de la fonction discriminante en apportant une contribution additive $Log(P(n_i/c(n_i)\omega))$ indépendante du pouvoir discriminant de la composante ou de sa qualité. Afin de prendre en compte l'inhomogénéité de la répartition des données d'apprentissage dans les composantes XML, il peut être souhaitable de pondérer les contributions précédentes par des poids w_i. Ces poids peuvent être déterminés soit par le biais de techniques d'apprentissage (méthode du stacking [Wolpert, 1992], méthode du Méta Classifieur [Ting and Witten, 1997]) ou grâce à des d'heuristiques qui cherchent à caractériser la qualité des estimations des grandeurs $Log(P(n_i/c(n_i)\omega))$. Le modèle CBCSP que nous proposons est défini par :

$$\omega^* = \arg\max_{\omega\in\Omega}\left\{Log(P(\omega)) + \sum_{n_i\in S_d} w_i.Log(P(n_i/c(n_i)\omega))\right\} \quad (16)$$

où w_i est le poids associé au nœud n_i.

7 Expérimentations

7.1 Base de données utilisée pour l'évaluation

La base de données Reuters est largement exploitée dans le cadre de la validation de système de classification de documents textuels. Cette base de données a été constituée initialement par une équipe de l'Université de Carnegie Mellon (the Carnegie group) à partir de dépêches de l'agence de presse Reuters en 1987. A ce jour, au moins cinq versions de cette base de données sont utilisées. Nous avons choisi la version ModApte version de la base Reuters-21578, téléchargée à partir de l'URL :

http://www.daviddlewis.com/resources/testcollections/reuters21578.

Pour nos expérimentations, nous avons sélectionné les 10 catégories les plus fréquentes à la fois pour les tâches d'apprentissage et de classification. 9035 documents sont ainsi répartis de manière inhomogène sur les 10 classes : la classe la plus grande contient 3964 documents, tandis que la plus petite contient 286. Ces catégories sont listées ci-dessous :

$$\Omega = \{Acq, Corn, Crude, Earn, Interest, Ship, Trade, Grain, Money\text{-}fx, Wheat\}.$$

7.2 Prétraitements

Tous les documents utilisés Durant les phases d'apprentissage et de test ont subi un prétraitement minimal qui consiste :

- à appauvrir la structure XML des documents en ne conservant que l'arborescence reliant les éléments textuels à la racine de l'arbre DOM, et en supprimant les attributs. Dans le cadre des données Reuters, seul les éléments TITLE, BODY, DATELINE sont donc considérés en tant qu'éléments textuels.

- à segmenter les éléments textuels en mots en utilisant les caractères séparateurs suivants : « `.;,:?<>=+}{()'\"^$#[]\\/\n` ».

Le choix de ne retenir que trois éléments XML a été effectué dans le but de confronter nos modèles avec les résultats d'autres approches évaluées sur les mêmes données et dans les mêmes conditions. Nous avons utilisé un parseur XML allégé pour extraire les données utiles nécessaire à la construction des modèles et à leur évaluation. Chaque nœud de l'arbre DOM associé à un élément XML TITLE, BODY ou DATELINE donne lieu à la construction d'une structure de données à base de tables de hachage. Celles-ci permettent de stocker les paramètres des modèles multinomiaux, principalement le chemin à la racine, le vocabulaire et la fréquence d'occurrence des mots. On peut considérer que la complexité globale des classifieurs CBCS et CBCSP est linéaire avec le nombre d'éléments XML pris en compte.

7.3 Heuristique de pondération pour le modèle CBCSP

Pour le modèle pondéré CBCSP, nous considérons l'heuristique qui consiste à pénaliser, pour une classe donnée, la contribution des nœuds n_i dont la taille du vocabulaire associé (nombre de paramètres multinomiaux) n'est pas en rapport avec la taille de l'information textuelle rattachée pour le document *d* considéré.

Nous proposons de choisir $w_i = \frac{|V_{n_i}|}{|d_{n_i}|}$, où V_{n_i} représente le vocabulaire associé au nœud n_i, et d_{n_i} représente l'élément textuel associé au nœud n_i pour le document *d*, la règle de décision devenant :

$$\omega^* = \arg\max_{\omega \in \Omega} \left\{ Log(P(\omega)) + \sum_{n_i \in S_d} \frac{|V_{n_i}|}{|d_{n_i}|} . Log(P(n_i / c(n_i)\,\omega)) \right\} \quad (17)$$

On notera que l'heuristique choisie ici est indépendante de la classe ω.

7.4 Mesures exploitées pour l'évaluation

A des fins d'évaluation comparative avec d'autres travaux dans le domaine de la classification de textes, les résultats expérimentaux présentés ci-dessous sont basés sur la mesure *F1* [van Rijsbergeng, 1979], définie comme la moyenne harmonique de deux mesures complémentaires : la *précision* et le *rappel* :

$$F_1(précision, rappel) = \frac{2.précision.rappel}{(précision + rappel)}$$

Dans la formule précédente, la *précision* et le *rappel* sont les deux mesures classiquement utilisés par la communauté de la Recherche d'Information (IR : Information Retrieval) pour évaluer les algorithmes de classification ou de filtrage d'information [van Rijsbergeng, 1979], [Apté et al., 1994] :

$$précision = \frac{nbVraisPositifs}{(nbVraisPositifs + nbFauxPositifs)}$$

$$rappel = \frac{nbVraisPositifs}{(nbVraisPositifs + nbFauxNégatifs)}$$

Dans les formules précédentes, pour une catégorie ω donnée, *nbVraisPositifs* correspond au nombre de documents appartenant à la classe ω correctement classés dans ω; *nbFauxPositifs* correspond au nombre de documents incorrectement classés dans ω;

nbFauxNégatifs correspond au nombre de documents appartenants à la classe ω mais non classés dans ω.

La procédure d'évaluation repose sur une validation croisée en dix étapes. Pour cette procédure, la base de données est préalablement découpée en 10 sous bases de tailles égales. Pour l'étape *i*, la i^{ème} sous base est sélectionnée pour l'évaluation, les 9 autres sous-bases servant pour l'apprentissage. Les données de la base Reuters exploitée pouvant posséder plusieurs étiquettes, chaque classe est évaluée de manière binaire par rapport au corpus constitué des documents appartenant aux autres classes.

7.5 Résultats comparatifs aux travaux de Guo, Wang et Bell

La procédure précédente a été appliquée sur les 7 classes (Ω = *{Acq, Corn, Crude, Earn, Interest, Ship, Trade}*) et sur les composants XML TITLE et BODY, conformément à la procédure décrite dans les travaux de Guo et al. [Guo et al., 2003] de manière à confronter les résultats obtenus par les modèles CBN, CBCS et CBCSP aux résultats obtenus en utilisant les modèles SVM (machine à vecteur support) , k-NN (k plus proches voisins), Rocchio (classification linéaire [Sebastiani, 2002]) et k-NN Guo-Model (k-NN exploitant un nombre limité de représentants de classe, homogène à des agrégats [Guo et al., 2003])

Nous calculons la moyenne de la mesure *F1* sur l'ensemble des éléments de Ω pour évaluer la performance globale des algorithmes sur les données de test :

$\hat{F}_1 = \frac{\sum_{i=1}^{|\Omega|} F_1(i)}{|\Omega|}$, où $F_1(i)$ est la valeur de la mesure F_1 pour la classe i ; $|\Omega|$ est le nombre de classes.

Les résultats expérimentaux donnant pour chaque classe la mesure F_1 sont synthétisés en table 1.

Classe ω_i	SVM	*k*-NN	Rocchio	*k*-NN Guo-Model	*CBN*	*CBCS*	*CBCSP*
Acq	90.74	78.81	84.03	86.08	94.31	94.28	97.56
Corn	94.74	87.08	87.74	91.85	88.92	89.59	92.99
Crude	87.30	84.12	84.51	84.72	87.24	87.11	91.45
Earn	95.14	88.02	90.39	89.45	94.84	94.98	98.01
Interest	92.78	79.67	80.84	83.26	93.39	92.89	96.79
Ship	88.02	84.02	86.22	86.73	83.89	84.60	88.18
Trade	91.81	80.69	81.64	80.00	81.07	81.28	90.78
$\hat{F}_1$	**91.50**	**83.20**	**85.05**	**86.01**	**89.09**	**89.25**	**93.68**

TAB 1 – *Analyse comparative basée sur la mesure F1 des performances des algorithmes sur la base de données Reuters-21578 pour les composants TITLE et BODY*

7.6 Résultats comparatifs aux travaux de Bratko et Filipic

La procédure précédente a été appliquée sur les 10 classes (*Ω = {Acq, Corn, Crude, Earn, Grain, Interest, money-fx, Ship, Trade, Wheat*), et sur les composants XML TITLE, DATELINE et BODY conformément à la procédure décrite dans les travaux de Bratko et Filipic [Bratko et Filipic, 2004a,b] de manière à confronter les résultats obtenus par les modèles CBN, CBCS et CBCSP aux résultats obtenus en utilisant les modèles SVM sur le texte mis à plat ou dans le cadre d'une méthode dite de « splitting » qui prend en compte les éléments structurels TITLE, DATELINE et BODY des documents. Les mesures $m \in \{recall, precision, F1\}$ utilisées pour évaluer les approches sont basés sur des macro-moyennes évaluées comme suit :

$$precision = \frac{\sum_{\omega_i \in \Omega} nbVraisPositifs(\omega_i)}{\sum_{\omega_i \in \Omega} nbVraisPositifs(\omega_i) + \sum_{\omega_i \in \Omega} nbFauxPositifs(\omega_i)}$$

$$recall = \frac{\sum_{\omega_i \in \Omega} nbVraisPositifs(\omega_i)}{\sum_{\omega_i \in \Omega} nbVraisPositifs(\omega_i) + \sum_{\omega_i \in \Omega} nbFauxNégatifs(\omega_i)}$$

Les résultats expérimentaux sont synthétisés en table 4. A titre de comparaisons, les résultats de Bratko et Filipic obtenus sur les mêmes expériences sont donnés en tables 2 et 3.

Mesure	Texte à plat	Splitting
Rappel	0.9623	0.9548
Précision	0.8280	0.8485
F1	**0.8901**	**0.8985**

TAB 2 – : Macro-moyenne rappel, précision, et mesure F1 pour le modèle « bayésien naïf » sur la base de données Reuters-21578, d'après les travaux de [Bratko et Filipic, 2004b]

Mesure	Texte à plat	Splitting
Rappel	0.9214	0.9660
Precision	0.9658	0.9178
F1	**0.9431**	**0.9413**

TAB 3 – : Macro-moyenne rappel, précision, et mesure F1 pour les modèles SVM sur la base de données Reuters-21578, d'après les travaux de [Bratko et Filipic, 2004b]

Mesure	CBN	CBCS	CBCSP
Rappel	0.9185	0.9241	0.9540
Précision	0.8540	0.8586	0.9294
F1	0.8851	0.8902	0.9415

TAB 4 – : Macro-moyenne rappel, précision, et mesure F1 pour les modèles CBN, CBCS et CBCSP sur la base de données Reuters-21578

Note : le modèle CBN est comparable dans sa structure au modèle bayésien naïf utilisé par Bratko et Filipic.

7.7 Analyse des résultats

Sur la première expérience, le modèle CBCS se rapproche des performances des modèles SVM appliqués sur le texte « à plat » en ne cédant sur la tâche considérée que 2.5% sur la mesure $\hat{F}_1$: il se situe devant les modèles k-nn et Rocchio.

Sur les deux expériences, nous constatons que le modèle CBCS se comporte de manière comparable (légèrement mieux) au modèle CBN, ce qui tendrait à montrer que l'apport de la structure du document dans les modèles bayésiens a une relative importance. Cela est confirmé par Bratko et Filipic qui retrouvent des résultats comparables entre le modèle bayésien naïf appliqué sur le texte « à plat » et découpé par composants XML selon la méthode « splitting ». Par contre, le modèle CBCSP, avec l'heuristique introduite au paragraphe 4.3 conduit à des améliorations très significatives par rapport au modèle CBN, de l'ordre de 4%. Sur la première expérience, on peut constater que l'amélioration sur la classe « Trade » est de l'ordre de 10%. Sur cette même expérience, le modèle CBCSP se comporte mieux que le modèle SVM appliqué sur le texte « à plat », tandis que sur la 2ème expérience, celui-ci se comporte de manière comparable au modèle SVM appliqué sur le texte « à plat » ou décomposé en composants XML selon la méthode dite du « splitting ». A notre sens, ces deux expériences montrent de manière significative l'intérêt d'une prise en compte de l'information structurelle, dans le contexte de la base de données Reuters-21578 et de la classification bayésienne de documents semi-structurés textuels. Cet intérêt nécessite l'utilisation d'une heuristique permettant de pondérer les estimations associées aux éléments structurés en fonction d'un critère quantitatif simple (rapport entre la taille de l'élément textuel analysé et la taille du dictionnaire associé au composant structurel). Ce type de critère traduit en partie la qualité des estimations produites par les modèles associés aux composants structurels.

8 Conclusion et perspectives

Nous nous sommes intéressés au problème de la classification supervisée de documents semi-structurés. Un modèle formel situé entre les approches de [Yi et Sundaresan, 2000], [Denoyer et al. 2003] ou [Bratko et Filipic, 2004a] a été proposé. Ce modèle, qui puise ses fondements dans les modèles bayésiens de classification, vise également à intégrer la prise en compte de la structure des documents dans les tâches de classification. Cette intégration permet la fusion de données numériques ou symboliques structurées d'une part, et de données non structurées qui peuvent faire l'objet d'une modélisation spécifique d'autre part. L'originalité du modèle proposé provient de sa formulation récursive qui permet d'appréhender de manière approchée l'arbre représentatif d'un document semi-structuré sous la forme d'un ensemble de nœuds associés aux chemins les reliant à la racine de l'arbre. Cette formulation est établie en posant une hypothèse simple d'indépendance des sous arbres étant donné leur nœud père. Une version simplifiée de ce modèle a été implémentée pour évaluer l'impact de la prise en compte de la structure documentaire dans des tâches de classification de documents textuels. L'introduction d'une heuristique permettant de pondérer les modèles de classification locaux associés aux nœuds de l'arbre représentant le document semi-structuré. Les premiers résultats semblent montrer que, pour le modèle pondré CBSCP, la prise en compte du contexte structurel d'occurrence des mots associée à l'heuristique de pondération proposée améliore de manière significative les performances d'un classifieur bayésien naïf multinomial évalué sur la base de données Reuters-21578. Ce modèle conduit à des résultats comparables aux modèles SVM associés à des techniques de « splitting ». Ces résultats, qui demandent à être confirmés sur d'autres sources de données plus adéquates à une validation approfondie, pourraient être éventuellement améliorés en :

- intégrant des modèles de langage de type n-gram, sur les traces des travaux initiés par [Peng et Schuurmans, 2003].

- En proposant d'autres heuristiques de pondération ciblant la qualité des estimations ou la minimisation des erreurs de classification par des techniques de stacking par exemple [Wolpert, 1992].

L'implémentation du modèle complet devra par ailleurs être entrepris pour évaluer plus finement l'apport de la structure dans le cadre d'expérimentations ou d'applications plus complexes, mais sans doute plus réalistes et intéressantes, pour lesquelles, la structure documentaire est très hétérogène (exploitation du WEB par exemple). Cette implémentation fait l'objet de développement en cours pour apporter des solutions au problème d'estimation des paramètres en présence d'un faible nombre de données d'apprentissage.

Références

[Abiteboul, 1997] Abiteboul S. « Querying semi-structured data ». In F. Afrati and Ph. Kolaitis, editors, Proc. of the 6th Int. Conf. on Database Theory (ICDT), Lecture Notes in Computer Science 1186, pages 1-18. Springer, January 1997.

[Apté et al., 1994] Apté, C. d., Damerau, F. J., and Weiss, S. M. 1994. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems 12, 3, 233–251.

[Bratko et Filipic, 2004a] Bratko, A. and Filipic, B. Exploiting Structural Information in Semi-structured Document Classification. Proc. 4th International 13th Electrotechnical and Computer Science Conference, ERK'2004, September 2004.

[Bratko et Filipic, 2004b] Bratko, A. and Filipic, B. A Study of Approaches to Semi-structured Document Classification. Technical Report IJS-DP-9015, Department of Intelligent Systems, Jozef Stefan Institute, November 2004. http://ai.ijs.si/andrej/papers/ijs_dp9015.html

[Brown et al., 1992] Brown, P. F. et al., Class-based n-gram models of natural language. Computational Linguistics, 18(4):467-479.

[Cavnar et Trenkle, 1994] Cavnar, W. B. and Trenkle, J. M. 1994. N-gram based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval (Las Vegas, NV, 1994), 161–175.

[Cline, 1999] Cline, M. (1999) Utilizing HTML Structure and Linked Pages to Improve Learning for Text Categorization. Undergraduate Honors Thesis. University of Texas, Austin.

[Denoyer et al., 2003] Ludovic Denoyer, Jean-Noel Vittaut, Patrick Gallinari, Sylvie Brunessaux, Stephan Brunessaux, "Structured Multimedia Document Classification", in DocEng'03, November 20–22, 2003, Grenoble, France.

[Domingos et Pazzanni, 1997] Domingos, P., Pazzanni, M. (1997) Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning,29.

[Duda et Hart, 1973], Richard O. Duda and Peter E. Hart Pattern Classification and Scene Analysis John Wiley & Sons, 1973

[Dumais et al., 1998] S Dumais, J Platt, D.H., Sahami, M.: Inductive learning algorithms and representations for text categorization. In: CIKM, ACM (1998) 148–155

[Eyheramendy et al., 2003] Eyheramendy, S., Lewis, D., Madigan, D. (2003) On the Naive Bayes Model for Text Categorization. Artificial Intelligence and Statistics 2003.

[Guo et al., 2003] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "Using kNN Model-based Approach for Automatic Text Categorization". Submitted to Soft Computing Journal, October, 2003,
http://www.icons.rodan.pl/publications/%5BGuo2003a%5D.pdf

[Huffman et Damashek, 1994] Huffman, S. and Damashek, M. 1994. Acquaintance: A novel vector-space n-gram technique for document categorization. In D. K. Harman Ed., Proceedings of TREC-3, 3rd Text Retrieval Conference (Gaithersburg, US, 1994), pp. 305–310. National Institute of Standards and Technology, Gaithersburg, US.

[IBM, 200] IBM (International Business Machine Corporation), *XML Extender (Administration and Programming)*, 2000

[Joachims, 1999] Joachims, T.: Making large-scale svm learning practical. Advances in Kernel Methods - Support Vector Learning (1999)

[Lewis et Ringuette, 1994] Lewis, D., Ringuette, M.: A comparison of two learning algorithms for text categorization. In: SDAIR. (1994) 81–93

[Marteau & Ménier, 2003] Marteau P.F. & Ménier Gildas (2003), Alignement approximatifs d'arbres pour la recherche d'information en contexte dans les données XML hétérogènes - Fusion d'information structurées et textuelles, numéro spécial « Données Numériques et informations symboliques », revue Techniques et Sciences Informatiques (TSI), vol. 22 , no 7-8 , pp. 1011 – 1034, 24 pages, Hermès Ed. 2003.

[McCallum et Nigam, 1998] McCallum, A., Nigam, K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. In Proceedings of AAAI-98 Workshop on "learning for Text Categorization", AAAI Press.

[Ménier & Marteau, 2002] Gildas Ménier, Pierre-François Marteau (2002), Information retrieval in heterogeneous XML knowledge bases, The 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems , IEEE, 1-5 July, 2002, Annecy, France .

[Netproject, 2001] Netproject page, 2001, http://www.netproject.org

[Peng et Schuurmans, 2003] Peng, F., Schuurmans, D. (2003) Combining Naive Bayes and n-Gram Language Models for Text Classification. ECIC Pise conference April 2003.

[Reuters-21578] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[Rish, 2001] Rish, I. (2001) An Empirical Study of the Naive Bayes Classifier. In Proceedings of IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence

[Schapire et Singer, 2000] Schapire, R., Singer, Y.: BoosTexter: A boosting-based system for text categorization. Machine Learning 39 (2000) 135–168

[Ting and Witten, 1997] Ting K. M. and Witten I. H. (1997). Stacked generalization: When does it work? In Proceedings of IJCAI-97, 15th International Joint Conference on Artificial Intelligence, pages 866–873, 1997.

[Vapnik, 1995,1998] Vapnik, V. (1995) Support Vector Networks. Statistical Learning Theory. Wiley-Interscience.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol.34, No.1, March 2002, pages 1-47.

[van Rijsbergen, 19979] van Rijsbergen C. J. (1979). Information Retrieval. Butterworths, London, second edition, 1979.

[Wolpert, 1992] Wolpert D. H. (1992). Stacked generalization. Neural Networks, 5:241–259, 1992.

[XERCES, 2002] Xerces- XML parsers in Java, C++ (with Perl and COM bindings). http://xml.apache.org/, 2002.

[Yang et Chen, 2002] Yang Jianwu, Chen Xiaoou, "A semi-structured document model for text mining", Journal of Computer Science and Technology archive, Volume 17(5),603-610, May 2002.

[Yang et Xin, 1999] Yang, Y., Xin, L. (1999) A re-examination of text categorization methods. Information Retrieval, Vol 1.

[Yang, 1999] Yang, Y.: An evaluation of statistical approaches to text categorization. Journal of Information Retrieval 1 (1999) 67–88

[Yi et Dundaresan, 2000] Jeonghee Yi and Neel Dundaresan, A classifier for semi-structured documents. in Proc. of the 6th International Conference on Knowledge Discovery and Data mining (KDD) 2000.

Summary

The aim of this paper is the supervised classification of semi-structured data. A formal model based on bayesian classification is developed while addressing the integration of the document structure in classification tasks. This integration allows to propose an information fusion approach for structured (numeric or symbolic data) and unstructured (e.g. text) XML elements. Simplified versions of this formal model are implemented to carry out textual documents experiments. First results show that the structure context of word occurrences have a significant impact on classification results comparing to the performance of a simple multinomial naïve Bayes classifier, as stated already by similar studies. These implementations perform better than k-nn or Rocchio models. On the Reuters-21578 data base, the weighted version of the simplified model competes with the SVM classifier associated or not with the splitting of structural components. A complete implementation of the model should allow tackling more complex and useful tasks.

RNTI - E -