Utilisation des méthodes de classification pour la construction automatique d'un corpus de Comptes Rendus de lecture

C. Benkoussas[1] P. Bellot[1] H. Hamdan[1] E.Faath[2]

[1]Domaine universitaire de Saint Jérôme,
Avenue Escadrille Normandie Niemen,
13397 Marseille
LSIS, Aix - Marseille Université

[2]École centrale de Marseille – Technopole de Château-Gambert
38, rue Frédéric Joliot-Curie
13451 Marseille
OpenEdition (Cléo)

chahinez.benkoussas@lsis.org

Domaine principale de recherche: IA
Papier soumis dans le cadre de la journée commune: NON

Résumé

Avec la croissance régulière des documents numériques, le besoin de les classer automatiquement devient indispensable. Le travail mené dans ce contexte consiste à construire automatiquement un corpus de critiques de livres (Comptes Rendus de lecture) dans la bibliothèque numérique Revues.org d'OpenEdition par un processus de classification de texte. Trois approches différentes pour la représentation des documents ont été élaborées, l'approche populaire de sac de mots, l'approche de sélection des caractéristiques (Features) et l'approche proposée qui se base sur la répartition des entités nommées dans le texte. Par la présente étude, nous montrons que les méthodes de classification habituelles, généralement efficaces pour la classification des thèmes, le sont aussi pour l'identification des Comptes Rendus de lecture considérée comme une tâche de classification des documents par genres.

Mots Clef

Classification, fouille de texte, SVM, sélection d'attributs, Apprentissage, Entropie Maximale.

Abstract

With a growing amount of numerical documents available, there is a need to classify them automatically. Our contribution in this context aims to build automatically a corpus of Reviews of Books in the digital library Revues.org from OpenEdition portal using text classification process. Three different approaches for representing documents have been used, Bag-of-Words approach, features selection and present approach based on the repartition of named entities in the text. In this study, we show that the usual methods of classification, generally effective for themes classification so too are to identify Reviews of Books considered as a task of classifying documents by genre.

Keywords*Classification, text mining, SVM, attributes selection, Learning, Maximal Entropy.*

1 Introduction

La recherche en classification automatique supervisée de textes vise à développer des modèles pour attribuer des étiquettes de catégories à des documents ou à des segments de documents en se basant sur un ensemble de documents d'apprentissage pré-classés manuellement par un expert. La majorité des études sur la classification automatique de textes sont orientées thématique. Dans cet article, nous nous intéressons à une classification non-thématique qui inclut analyse de sentiment et classification par genre.

L'analyse de sentiments ou fouille d'opinion a pour but d'identifier la polarité (positive, ou négative ou neutre) d'une phrase, d'un paragraphe ou d'un document entier .

Fouiller les opinions à partir des critiques en ligne de différents produits (caméras, téléphones, voitures, livres, etc.) est un processus complexe. En premier, les documents qui présentent des critiques doivent être récupérés des sites Web, les moteurs de recherche jouent un rôle très important dans cette étape. En second, les données nécessitent d'être mises en forme avant d'être analysées. Dans le processus de préparation des données, il est très probable que les pages web collectées contiennent d'autres types d'information que des critiques. Dans ce cas, il est important de séparer les documents porteurs d'opinion des documents non porteurs d'opinion pour améliorer le processus de classification de sentiment [4, 9]. L'affectation manuelle des étiquettes de catégories à un large ensemble de documents est inappropriée. Il existe plusieurs techniques pour distinguer les informations subjectives des informations factuelles ou objectives [1, 12]. En particulier, lors d'une combinaison des systèmes de classification automatique de sentiment avec des moteurs de recherche pour fouiller les documents qui présentent des critiques d'utilisateurs, la précision peut, largement, dépendre de la reconnaissance des documents porteurs d'opinion (critiques).

Dans cet article, nous nous intéressons à l'identification d'un type particulier de documents ou pages Web: les critiques de livres ou Comptes Rendus de lecture (CR). S'il existe de très nombreux travaux autour de l'analyse d'opinion, il n'existe pas à notre connaissance de service permettant de rechercher automatiquement des documents de type «critiques» sur le Web. Les méthodes d'analyse d'opinion sont la plupart du temps éprouvées sur des corpus de documents déjà constitués et issus de sites Web spécialisés dans l'échange d'opinion (réseaux sociaux tels LibraryThing, plateformes de micro-blogging., forums...). Nous proposons dans ce qui suit une approche de détection de critiques (classification en deux classes : Compte-rendu de lecture (CR) et Non-Compte-Rendu de lecture (NCR)) en exploitant des méthodes habituellement utilisées pour la classification thématique de documents . La question est de savoir à quel point des approches de classification thématique sont adaptées à ce problème et quels sont les descripteurs linguistiques les plus utiles à cette fin. Nous appliquons ces approches dans le contexte d'une bibliothèque numérique en sciences humaines au travers des plateformes OpenEdition.org.

L'article est organisé de la façon suivante, la section 2 présente quelques travaux sur la classification en genre, dans la section 3, nous définissons le corpus utilisé et sa construction. La section 4 présente les approches utilisées pour la représentation des textes. La section 5 décrit les différents algorithmes de classification. Dans la section 6, nous présentons les expérimentations réalisées ainsi que les résultats obtenus avec différentes familles de descripteurs (traits linguistiques et modèles de langue) et une méthode de sélection automatique de ces descripteurs.

2 État de l'art

La classification en genre des documents Web a été largement étudiée au cours des dernières années [13-15, 17, 18]. La difficulté de cette tâche s'inscrit dans la recherche des caractéristiques les plus appropriées pour représenter les documents.

Maeda et Hayashi [3] ont proposé une méthode pour classer les documents Web par genre, ils se sont basé sur les mots et les balises HTML comme caractéristiques et ils ont calculé l'efficient discriminant pour chaque pair de mots et balise HTML pour trouver les balises les plus efficaces pour la classification. Leurs expériences sur un corpus de documents Web japonais ont montré que l'utilisation des mots avec les balises HTML améliore de 8 % la valeur du Rappel, en utilisant la méthode SVM pour la classification.

Santini [16] a analysé les effets de composition du corpus pour la phase d'apprentissage, les différents genres et la représentativité des caractéristiques dans la classification en genre et a montré les limitations de l'exportabilité des modèles de classification à une autre collection.

Ferizis et al. [6] ont effectué une analyse de performance de la classification en genre de documents sur le Web et ont proposé une méthode pour améliorer l'efficacité de l'analyse linguistique.
L'originalité de notre approche est d'utiliser des méthodes de classification thématiques pour classer les documents en genre, précisément identifier les Comptes Rendus de lecture, ce qui n'a, à notre connaissance, jamais été réalisé.

3 Description et construction du corpus

Un compte rendu de lecture inclut la présentation d'un livre, une analyse des arguments principaux et des opinions (négatives, positives ou neutres) sur différents éléments du livre. Voici un exemple extrait de la revue « transatlantica -845» qui illustre les différentes parties d'un ouvrage.

```
<p> L'ouvrage du professeur Uviller et de William G. Merkel apporte une nouvelle contribution au débat sur le second amendement. Un débat qui s'est développé parmi les historiens et les juristes depuis la parution, en 1991, […] </p>

<p> L'ouvrage est divisé en trois parties : la première permet au lecteur de découvrir ou de mieux connaître les fondements historiques […] </p>

<p> Avec minutie, rigueur et brio H. Richard Uviller et W. G. Merkel démontrent que dans la période qui précède la Déclaration d'Indépendance, […] </> […]

<p> Que l'on partage ou non le point de vue défendu par H. Richard Uviller et William G. Merkel, la lecture de cet ouvrage s'impose pour plusieurs raisons : tout d'abord à cause de la qualité de l'analyse de la pensée politique de la période […] </p>
```

La difficulté de la tâche d'identification des compte-rendus (CR) est dans la diversité des types de documents que nous traitons. Les documents n'étant pas des compte-rendus (NCR) sont, dans certain cas, proches aux CR en terme de style d'écriture, structuration, et mots utilisés. De fait un compte-rendu de lecture d'un livre traitant d'une thématique donnée aborde de fait cette thématique et nous devons veiller à ce que la classification se fasse selon le genre du texte et non son thème principal. A titre d'exemple de document qui ressemble à un compte-rendu de lecture mais qui n'en n'est pas un, voici un extrait d'un éditorial (extrait de la revue « mimmoc-748 »):

```
[…]<p> En ouverture, Kenneth O. Morgan, Baron of Aberdyfy, Fellow of the British Academy, nous a livré un exposé intitulé 'The new Liberal party from dawn to downfall, 1906-1924'. L'étude de Lord Morgan sur le parti Libéral s'est structurée comme suit : […] </p>

[…]<p> Avant d'entrer dans le vif du sujet, Lord Morgan, […], a dressé une brève comparaison du parti Libéral sous Asquith et du parti Travailliste d'Attlee de 1945. Sur cet aspect comparatif, il a montré qu'il y a une grande différence d'autant […] </p>
```

```
[…]<p>Poursuivant son analyse du parti Libéral avant la guerre, Kenneth Morgan a présenté le contexte historique […] </p> […]
```

Notre corpus est constitué de données en langue française provenant de deux sources, la plateforme *Revues.org* d'OpenEdition.org d'une part et du Web général d'autre part (ce dernier corpus est collecté via des requêtes sur le moteur de recherche Google). Le corpus est subdivisé en deux catégories, Comptes Rendus (CR) et Non-Comptes Rendus (NCR) qui représentent une mixture de documents scientifiques (articles, billets, appels à contribution, éditoriaux, etc.).

Le premier ensemble de documents est issu de Revue.org[1] une des quatre plateformes d'OpenEdition[2] (Hypothèses.org, Calenda et Books) le portail de ressources électroniques en sciences humaines et sociales. Revues.org est la plus ancienne plateforme française de revues et collections de livres. Elle offre actuellement plus que 300 revues de différentes disciplines de la science humaine et sociale (dans plusieurs langues). Le mode de collecte pour cet ensemble de document se base sur le système Lucene Solr[3] d'Apache. L'index Solr d'OpenEdition (contenant plus de 140 000 documents de Revues.org, plus de 40 000 documents de Hypothèses.org et plus de 20 000 documents de Calenda) a été interrogé avec deux requêtes précisant le type et la langue du document. La première requête extrait les Comptes Rendus français de livres (CR) et la seconde requête extrait des documents divers Non-Comptes Rendus (NCR). À la fin, nous avons sélectionné une collection de 498 CR et 244 NCR. Tous les documents de la collection sont en format XML TEI[4].

Le second ensemble de documents est issu du web. Le but est de construire une collection de CR à partir du web et non seulement des plateformes d'OpenEdition, en retrouvant les CR en ligne pour chaque livre d'OpenEdition. Nous avons choisi 127 livres dans 3 catégories : 49 livres d'environnement, 63 de sociologie et 15 livres d'informatique. Via des requêtes posées à Google Web Search, comportant la concaténation du titre du livre (garder la même phrase du titre) et le(s) nom(s) de son(ses) auteur(s). Ensuite, nous avons téléchargé et annoté manuellement les 20 premières pages Web pour chaque requête. 2000 pages ont été évaluées et classées en plusieurs classes : CR, interview, annonce, bibliographie ou encore « accès refusé ». Après filtrage des pages, nous avons une collection de CR qui comporte 95 CR et 240 NCR. Il s'avère difficile de trouver des CR de lecture dans le web pour plusieurs raisons :absence de méta-données indiquant le type de l'article (document) présence notable de sites de vente en ligne (notamment dans le cas où les requêtes utilisées pour la collecte ne contiennent que le titre d'un livre).

1 : http://www.revues.org/
2 : http://www.openedition.org/
3 : http://lucene.apache.org/solr/
4 : **T**ext **E**ncoding **I**nitiative, http://www.tei-c.org/index.xml

Nous illustrons dans le tableau suivant le peuplement des deux catégories du corpus et la longueur moyenne (en nombre de mots) des documents.

Source	CR	NCR
Revues.org	498	244
Web	95	240
Total	**593**	**484**
Longueur moyenne (mots)	**1561**	**5043**

Tableau 1. Nombre de CR et NCR collectés de Revues.org et du Web

Le tableau 2 introduit le nombre de documents CR et NCR pour le corpus d'apprentissage et de test.

	CR		NCR	
	Revues	*Web*	*Revues*	*Web*
Apprentissage	298	60	144	150
Test	200	35	100	90

Tableau 2. Nombre de CR et NCR dans le corpus d' apprentissage et de test

4 Quels descripteurs pour une classification efficace ?

Comme nous l'avons précisé précédemment, le but de notre étude est d'identifier des CR de lecture par des méthodes de classification automatique généralement utilisées pour classer les documents par thème. Dans ce contexte, le choix des descripteurs peut se révéler décisif. Nous allons utiliser différents descripteurs pour la tâche de classification.

4.1 Représentation en sac de mots

La représentation des textes la plus simple a été introduite dans le cadre du modèle vectoriel, et porte le nom de « sac de mots ». Les textes sont transformés simplement en vecteurs de poids qui correspondent à une fonction de l'occurrence des mots dans le texte. Cette représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots.

L'approche de sac de mots est généralement utilisée pour la classification des documents en plusieurs thématiques, puisqu'on se base sur les mots présents dans le texte pour identifier la classe d'appartenance. Nous avons utilisé cette représentation dans le cadre de l'identification des CR. L'idée est de supposer que les CR contiennent des mots communs qui sont représentatifs à la classe CR par rapport à la classe NCR ce qui veut dire que nous exploitons l'approche pour classer les documents en classes qui représentent plutôt des genres (CR et NCR) et pas des thématiques.

4.2 Approches de sélection de caractéristiques

Les algorithmes de sélection permettent d'extraire une information non redondante et pertinente, en vue d'une exploitation efficace des bases de données. Ils font l'objet d'une littérature abondante depuis une dizaine d'années [5, 7, 8] ; ils sont répartis en trois groupes principaux : les « *filters* », les « *wrappers* » [5] et les « *embedded* » [8]. Les méthodes filtres (*filter methods*) opèrent directement sur le jeu de données et fournissent une pondération, un classement ou un ensemble de variables en sortie. Ces méthodes ont l'avantage d'être rapides et indépendantes du modèle de classification, mais au prix de résultats inférieurs. Les méthodes enveloppes (*wrapper methods*) effectuent une recherche dans l'espace des sous-ensembles de variables, guidées par le résultat du modèle, par exemple les performances en validation croisée sur les données d'apprentissage. Elles ont souvent de meilleurs résultats que les méthodes de filtrage, mais au prix d'un temps de calcul plus important [5] . Enfin, les méthodes embarquées *(Embedded methods*) utilisent l'information interne du modèle de classification (par exemple, le vecteur de poids dans le cas des SVM – Support Vector Machine) [10] , ces méthodes sont donc proches des méthodes d'enveloppes, du fait qu'elles combinent le processus d'exploration avec un algorithme d'apprentissage sans étape de validation, pour maximiser la qualité de l'ajustement et minimiser le nombre d'attributs [2, 5, 8].

La sélection des mots réduit la dimension de l'espace vectoriel (dans notre cas plus de 164000 mots) dont chaque document est un vecteur et en même temps augmente la capacité de discrimination entre les documents. La dimension de l'espace vectoriel correspond à la complexité du modèle d'apprentissage, c'est-à-dire que plus il y a de mots à prendre en compte, plus il y a de variables à évaluer. Pour ces raisons, nous avons procédé à une réduction de l'espace vectoriel en supprimant en premier les balises XML(TEI) et HTML, les mots vides et les mots outils de la langue française.

Une deuxième étape de filtrage a été appliquée sur le sac de mots en utilisant l'algorithme RFE-SVM (Recursive Feature Elimination – Support Vector Machine) [5] qui est une méthode de type *embedded* basée sur l'élimination *backward* en utilisant les SVM pour sélectionner un sous-ensemble d'attributs optimal non redondants. La méthode repose sur l'estimation de poids relatifs à l'optimisation d'un problème de discrimination linéaire, ce problème étant résolu à l'aide d'une machine à vecteurs de support (SVM). Il est montré dans [5] que le coût de suppression d'une caractéristique est de l'ordre de $\{w_j^2\}_{j=1}^{D}$ ou w_j est le poids attribué à l'attribut *j* lors de la phase de construction du SVM [10]. La procédure de sélection est décrémentale et élimine progressivement les attributs de faible poids [2].

Après avoir sélectionné les attributs avec l'algorithme RFE-SVM (Recursive Feature Elimination with Support Vector Machine) en utilisant la méthode SVMAttributeEval[5] de l'outil Weka[6] (valeurs des paramètres par défaut), nous avons transformé les documents du corpus en descripteurs contenants des caractéristiques de valeurs binaires. Ces caractéristiques représentent les mots sélectionnés par l'algorithme RFE-SVM, et leurs valeurs correspondent à la présence « 1 » ou l'absence « 0 » des caractéristiques dans les documents.

4.3 Représentation basée sur les entités nommées

La plupart des travaux consistent à supprimer les descripteurs non pertinents. Nous pouvons également nous focaliser sur la détermination des types de descripteurs les plus représentatifs pour chaque document en sélectionnant les éléments caractérisant une ou plusieurs classes (catégories).

Après une analyse linguistique et statistique du corpus, nous avons identifié des caractéristiques (illustrées dans la figure 1, figure 2 et figure 3) globalement communes entre les documents CR et d'autres entre les documents NCR ce qui permet de les distinguer. Nous citons quelques-unes :

- le titre des CR est souvent, soit la référence bibliographique du livre sur lequel porte le compte rendu, soit contient des éléments de la référence, par exemple : le nom de l'auteur (personne), l'année de publication du livre (date), le lieu (location), etc. (autrement dit, une référence incomplète) ;

- nous trouvons, dans la plupart des CR, vers le début du texte une description générale du livre où sont indiqués le nom de l'auteur (personne) et la date de publication de l'ouvrage ;

- dans les publications scientifiques qui ne sont pas des compte-rendus (NCR) et qui représentent des articles scientifiques, une partie « bibliographie » est présente à la fin du document. Elle contient des noms d'auteurs (personne), des lieux (location), des dates, etc. Contrairement aux comptes-rendus qui ne comportent, généralement pas, cette partie.

En se basant sur cette étude, nous avons annoté le corpus avec l'outil de reconnaissance d'entités nommées TagEN [11] pour identifier les « *noms de personnes* », « *dates* » et « *lieux* » dans le texte après avoir supprimé toutes les balises XML(TEI) pour le corpus issu d'OpenEdition.org et les balises HTML pour le corpus Web. Ensuite nous avons divisé chaque texte en 10 parties (taille de chaque partie = nombre de mots dans le texte / 10) et calculé pour chaque partie le taux de répartition de chacune des 3 entités nommées (personnes, date et lieu).

5 http://weka.sourceforge.net/doc.packages/SVMAttributeEval/weka/attributeSelection/SVMAttributeEval.html

6 http://www.cs.waikato.ac.nz/ml/weka/

Les graphes ci-dessous montrent la répartition des entités nommées « person », « date » et « location » dans les documents CR et NCR.

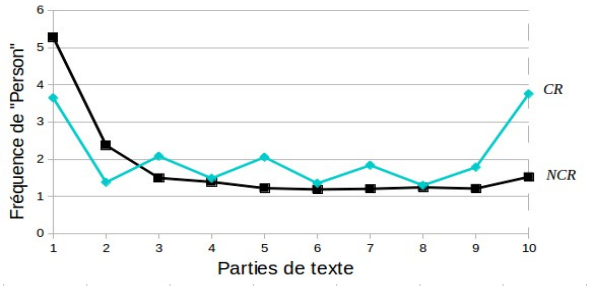

Figure 1. Répartition de l'entité « person » dans les documents CR et NCR

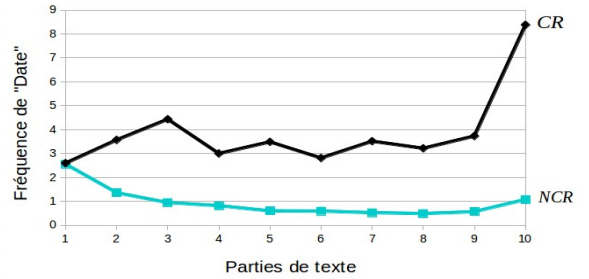

Figure 2. Répartition de l'entité « date » dans les documents CR et NCR

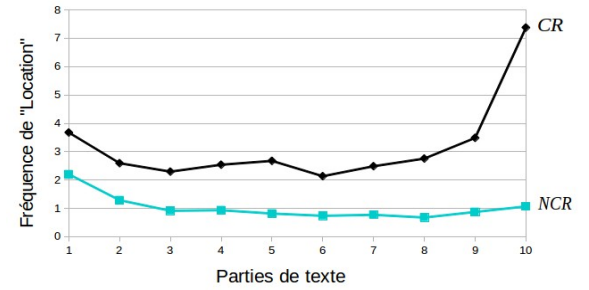

Figure 3. répartition de l'entité « location »dans les documents CR et NCR

Les documents de la collection ont été transformés en un ensemble de descripteurs. Chaque descripteur contient la classe (CR → 1 / NCR → 0) dans laquelle est classée le document et les taux de répartition des EN (entités nommées) définies précédemment dans le document. Ces taux représentent les caractéristiques (traits ou descripteurs) statistiques.

5 Quelle méthode de classification ?

Après avoir représenté les documents de la collection en plusieurs types de descripteurs :

1. représentation en sac de mots du corpus (en gardant le format original, XML-TEI pour le sous-corpus Revues.org et le HTML pour celui du web), ce qui induit à la présence des balises dans les vecteurs de mots.
2. représentation en descripteurs contenant des caractéristiques binaires issues du filtrage du sac de mots et de l'application de l'algorithme de RFE-SVM,
3. représentation en se basant sur la répartition des entités nommées dans le texte (toutes les balises XML-TEI et HTML ont été supprimées), nous procédons au calcul des modèles de classification pour chaque type de représentation. Ces modèles seront ensuite utilisés pour l'évaluation des textes du jeu de test.

Nous avons utilisé deux méthodes de classification :
- La méthode des Machines à Vecteurs de Support (SVM) en choisissant le noyau non-linéaire RBF (*Radial Basis Function*) avec (**gamma** = 0.01, **C**=1.0), généralement utilisé pour la classification de texte. Nous avons utilisé Weka (pour la représentation en sac de mots et la représentation après filtrage des attributs) et SvmLight[7] (pour la représentation basée sur la répartition des entités nommées).
- Entropie Maximale, utilisée pour la classification du corpus représenté avec les approches définies précédemment. La réalisation de la classification avec cette méthode était avec l'outil Megam[8].

6 Résultats

Le tableau 3 présente les résultats obtenus avec les différents descripteurs en appliquant le test avec le jeu de test défini dans la section 3. Pour les mesure d'évaluation, nous avons utilisé le rappel et la précision et la F-mesure (moyenne harmonique du rappel et de la précision).

	SVM			EntMax		
	Rappel	Précision	F-Mesure	Rappel	Précision	F-Mesure
Représentation en sac de mots	92,7 %*	93,0 %*	92,7 %*	**94,46 %**	**96,52 %**	**95,47 %**
Réduction de l'espace vectoriel	**97,17 %**	87,66 %	**92,17 %**	62,64 %	**94,78%**	75,42 %
Représentation basée sur les entités nommées	**95,02 %**	90,72%	**92,71 %**	83,75 %	**99,25 %**	90,84 %

Tableau 3. Évaluation de deux modèles de classification selon trois différentes représentations des documents

Dans cette table, les mesures d'évaluation de la baseline sont notées par **'*'**.
- Nous observons que les meilleurs résultats sont obtenus avec l'algorithme EntMax en utilisant la représentation en sac de mots (un gain de 3 % en taux de bonne classification comparé à la baseline).
- La réduction de l'espace vectoriel avec SVM a amélioré le rappel par rapport à la baseline et diminué la précision,contrairement à l'EntMax qui a sensiblement diminué le rappel et amélioré la précision.
- L'approche proposée basée sur la répartition des entités nommées dans le texte avec la méthode SVM donne des résultats encourageants et comparables à la baseline.

7 : http://svmlight.joachims.org/
8 : http://www.umiacs.umd.edu/~hal/megam/version0_3/

- Le temps d'apprentissage est nettement à l'avantage de l'approche proposée, puisque elle contient le plus petit nombre de caractéristiques(features) dans les descripteurs. Elle apporte un gain remarquable au niveau du temps d'exécution (20 fois plus rapide que la méthode classique de sac de mots et 7 fois plus rapide après la réduction de l'espace vectoriel).

- La représentation proposée est bien adaptée au contexte de l'identification des CR à partir d'une collection de documents provenant de deux sources de nature différentes (en terme de structure et style du contenu).
- Cette approche présente quelques lacunes dans les cas de ressemblances de répartition des entités nommées dans les documents de deux classes différentes. Pour faire face à cette difficulté, il est envisageable d'intégrer d'autres caractéristiques linguistiques (par exemple les 50 premiers mots les plus fréquents dans la classe CR) dans les descripteurs.

7 Conclusion

Nous avons montré dans cette article que les méthodes utilisées pour la classification thématique des documents s'adaptent aussi pour la classification des documents par genre.
L'approche de représentation des documents proposée dans la section 3.3, qui repose sur une schématisation vectorielle des documents, non plus axée sur les mots contenus mais sur la structure et la répartition des entités nommées dans le texte, donne des résultats concordants avec ceux des approches classiques notamment en temps de calcul. Les expérimentations sont faites sur le corpus issue de « Revues.org » et du web général dans l'intérêt d'identifier les Comptes Rendus de lecture (CR).

Bibliographie

[1] Ghose, A. and P. Ipeirotis. Towards an understanding of the impact of customer sentiment on product sales and review quality[M]. Proceedings of the Workshop on Information Technology and Systems (WITS), Milwaukee, December ,2006

[2] Samb, M., L., Camara F., Ndiaye S., "Approche de selection d'attributs pour la classification basée sur l'lgorithme RFE-SVM", 11ème Colloque African sur la recherche en Informatique et Mathèmatique (CART' 2012), 2012.

[3] Maeda, A.; Hayashi, Y., "Automatic genre classification of Web documents using discriminant analysis for feature selection," Applications of Digital Information and Web Technologies, 2009. ICADIWT '09. Second International Conference on the , vol., no., pp.405,410, 4-6 Aug. 2009

[4] Pimwadee Chaovalit and Lina Zhou. Movie review mining: A comparison between supervised and In: unsupervised classification approaches[M]. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.

[5] Guyon I., Weston J., Barnhill S., and Vapnik V., «Gene selection for cancer classification using support vector machines», *MACHLEARN: Machine Learning*, vol. 46, 2006.

[6] Ferizis, G., Bailey, P., "Towards practical genre classification of web documents", in Proceedings of the 15th international conference on World Wide Web, pp.1013-1014, 2006.

[7] Liu H. and Yu L., «Toward integrating feature selection algorithms for classification and clustering», *IEEE Trans. Knowl. Data Eng*, vol. 17, no. 4, 2005, p. 491–502.

[8] Lal T. N., Chapelle O., Weston J., and Elisseeff A., « Embedded Methods ». Springer, Nov. 20, 2004.

[9] Ye Qiang, Shi Wen and Li Yijun. Sentiment classification for movie reviews in chinese by improved semantic oriented approach[M]. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006.

[10] Vapnik V. N , «The nature of statistical learning theory». New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[11] Poibeau T.. «The multilingual named entity recognition framework». In EACL '03: Proceedings of the *tenth conference on European chapter of the Association for Computational Linguistics*, Morristown, NJ, USA, 2003, p. 155–158.

[12]Andreas Rauber and Alexander Muller-Kogler. Integrating automatic genre analysis into digital libraries[M]. In: Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, 2001.

[13] Lim, C.S., Lee, K.J., Kim, G.C., "Multiple sets of features for automatic genre classification of web documents", Information Processing & Management, Vol. 41, No. 5, pp.1263-1276, 2005.

[14] Dong, L., Watters, C., Duffy, J., Shepherd, M., "An Examination of Genre Attributes for Web Page Classification", in Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), p.133, 2008.

[15] Vidulin, V., Lustrek, M., Gams, M., "Training the Genre Classifier for Automatic Classification of Web Pages", in Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007), pp.93-98, 2007.

[16] Santini, M., "Some issues in Automatic Genre Classification of Web Pages", JADT 06 - Actes des 8 Journées internationales d'analyse statistiques des donnés textuelles, Vol. 2, pp.865-876, 2006.

[17] Chaker, J., Habib, O., "Genre Categorization of Web Pages", in Proceedings of Seventh IEEE International Conference on Data Mining (ICDM 2007) Workshops, pp. 455-464, 2007.

[18] Levering, R., Cutler, M., Yu., L., "Using Visual Features for Fine-Grained Genre Classification of Web Pages", in Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008), p.131, 2008.