Une méthode flexible pour l'identification de la langue d'un texte dans un corpus hétérogène multilingue

Said Kadri[1] , Abdelouahab Moussaoui[2]

[1] Département des STIC, Faculté des Mathématiques et d'Informatique
Université de M'sila, 28000, Algérie
Kadri.said28@yahoo.fr

[2] Département informatique, Faculté des sciences
Université Farhat Abbes de Sétif, 19000 , Algérie
Moussaoui.abdel@gmail.com

Abstract. Identifying the text language means that we assign this text to a language in which it is written. This identification became important because of the increased diversity of textual data in different languages on the web. In addition, a real recognition of the text language is not possible if we only consider the word as a basic unit of information. It could be possible for some languages as French or English but very difficult for some other languages as German or Arabic. The approach of text segmentation into characteristic n-grams represents a very efficient alternative solution in this field. It also becomes a favorite tool to extract knowledge from texts.

In this paper, we present the most known identification methods and we propose a new method based on a new metric of similarity. We also evaluate the obtained results with other methods while adopting the two approaches respectively : the segmentation of texts into words and their segmentation into n-grams.

Keyswords: N-gram, language identification, text categorization, Text Mining, machine learning.

1 Introduction

Les besoins des utilisateurs en information ne cessent de croître à cause de l'extension massive du réseau internet et l'explosion informationnelle des documents sur ce réseau. La recherche documentaire multilingue est l'une des solutions proposées pour satisfaire ces besoins. Son objectif principal est de permettre à l'utilisateur de formuler une requête dans une langue et d'extraire les documents correspondants en d'autres langues, ce qui nécessite une phase préliminaire très importante qui est l'identification de la langue de chaque document avant d'en extraire l'information. L'identification de la langue d'un texte, appelée aussi « la catégorisation de la langue », peut être définie comme étant l'attribution d'un document appartenant à un corpus multilingue thématiquement hétérogène à une langue donnée. Plusieurs éléments caractéristiques de la langue peuvent être considérés dans ce cas, notamment : la présence de certains caractères [1], [2], la présence de certains mots [2], [3], la fréquence de n-grammes [4], [5], [6], [7], [8], [9].

2 Etat de l'art

2.1 Approches d'identification de la langue

En général, il existe quatre grandes d'approches :

- L'approche linguistique : S'appuie sur la présence de certaines chaînes de caractères spécifiques à une langue permettant la reconnaissance directe de cette langue [1], [2].

Par exemple les chaînes caractéristiques de l'anglais sont (they, ery, …), du français (eux, aux, …), de l'allemand (die, der, …).

Malheureusement, cette approche rencontre plusieurs problèmes [6] à noter : les chaînes spécifiques de la langue peuvent être rares ou absentes dans le texte. En plus, le texte dans une langue peut utiliser des chaînes d'autres langues.

- L'approche lexicale : est basée sur l'utilisation d'un lexique de mots pour chaque langue à identifier [2], [3]. C.à.d. comparer les mots du texte avec une liste fixée de mots pour chaque langue (un lexique), la langue dont le lexique contient tout ou la plupart des mots du texte est la langue effective du document. Plusieurs problèmes peuvent être soulignés ici, à savoir : l'incomplétude de tout lexique construit (absence des mots scientifiques), la présence de fautes d'orthographe, de frappe, et de reconnaissance pour un texte acquis par un OCR ce qui perturbe les résultats obtenus par cette approche.
- L'approche grammaticale : se base sur la présence de certains mots grammaticaux de la langue [7] tels que : les prépositions, les conjonctions, les déterminants, les pronoms, les adverbes, etc. qui représentent environ 50 % des phrases et des textes dans la plupart des langues [3]. Par exemple pour l'anglais on trouve les mots (the, they, are, he, she, …), pour le français (le, la, les, des, leur, leurs, …). Cette approche est plus rapide et plus efficace que les deux approches précédentes, mais souffre quand même de plusieurs inconvénients, à savoir : les textes doivent être segmentés en mots ce qui est difficile pour certaines langues, ces mots grammaticaux sont souvent éliminés lors d'un prétraitement effectué sur le texte (stopwords) [10], l'approche donne des mauvais résultats pour les textes courts à cause de l'absence de ces mots grammaticaux. En plus, elle est difficile à appliquer sur des séquences d'autres types (séquences ADN en biologie).
- L'approche statistique et probabiliste : cette approche ne demande pas de connaissances linguistiques préalables mais des calculs de probabilités, son principe est de capturer certaines régularités formelles (mots, n-grammes) des langues au moyen d'un modèle probabiliste à partir d'un corpus représentatif pour chaque langue et de lui associer une fréquence ou une probabilité d'apparition, puis calculer la probabilité qu'un texte appartient à l'une des langues en fonction de l'apparition de ces régularités observées. L'approche probabiliste se généralise de la reconnaissance de la langue à la classification thématique des textes [11]. Malheureusement le découpage des textes en mots dégrade les performances de ces approches du fait qu'un texte court ne contient pas forcément les mots les plus fréquents de la langue. En plus, pour certaines langues comme le chinois ou même pour les séquences ADN en biologie, il est difficile de découper le texte en mots, d'où l'utilisation de l'approche des n-grammes [8], [9], [12] paraît une solution très efficace.

2.2 Principe de la segmentation d'un texte en n-grammes de caractères

Un n-gramme est une séquence consécutive de *n* caractères [5] qui peuvent ne pas être ordonnées. Pour un document, l'ensemble des n-grammes qu'on peut générer est une collection de photos qu'on obtient en déplaçant par un caractère une fenêtre de n caractères sur le corps du texte. Par exemple, les 3-grammes de la phrase « bonjour monsieur » sont : *bon, onj, njo, jou, our, ur_, r_m, _mo, mon, ons, nsi, sie, ieu, eur* [13], [14], [15]. Un profil n-grammes d'un document consiste en la liste des contigüités de n-grammes les plus fréquentes dans le document accompagnées de leurs fréquences.

2.3 Avantages de la segmentation d'un texte en n-grammes

L'approche de segmentation de textes en n-grammes caractéristiques présente plusieurs avantages, notamment :

- L'approche évite de recourir à la phase de lemmatisation et de stemming sur le texte. Cette phase exige un effort linguistique et algorithmique considérable.
- Tolérante aux fautes d'orthographe, de frappe, et d'acquisition pour un OCR. [13] montre que des systèmes de recherche documentaires sur les n-grammes gardent leurs performances malgré des taux d'erreurs de 30 % ce qui n'est pas le cas pour les systèmes basés sur l'approche de segmentation en mots.
- Cette technique opère indépendamment des langues.
- La segmentation en mots est difficile pour certaines langues où il n'est pas facile de trouver des frontières claires entre les mots. Par exemple en langue arabe, les pronoms sujets et compléments sont dans certains cas attachés aux verbes et une seule chaîne de caractères représente ainsi une phrase (*katabtouhou*) (je l'ai écrit). La même chose peut être dite pour l'allemand, le chinois ou les séquences ADN en génétique[14], [15].

2.4 Méthodes d'identification de la langue basées sur les n-grammes

La plupart des systèmes d'identification de la langue utilisant les n-grammes se basent sur le même schéma consistant en [15] : la phase d'acquisition automatique des connaissances linguistiques dans laquelle on choisit un corpus représentatif pour chaque langue, puis on génère un profil caractéristique qui sera pris comme référence, ensuite on calcule les fréquences des différents n-grammes (avec $n = 3, 4, 5, \ldots$). La phase de diagnostique dans laquelle on construit pour chaque texte à identifier son profil n-grammes et on cherche le profil de référence le plus similaire. Plusieurs méthodes sont proposées pour mesurer cette similarité.

Méthodes des *k* plus proches voisins : Plusieurs algorithmes d'identification de la langue se basent sur la notion de distance ou de similarité. L'idée de base est de chercher le texte, parmi l'ensemble d'apprentissage, qui soit le plus proche en distance du texte à classer et de lui attribuer la même langue. On peut augmenter le nombre k de textes les plus proches au texte à identifier la langue si cela est nécessaire. Dans ce cas, la langue du nouveau texte est la même que la majorité des k plus proches voisins de ce texte (la langue majoritaire). Un défi pour ces approches est de définir une métrique de similarité. Pratiquement, il existe plusieurs distances, notamment :

* La distance de Beesly [4] : dans cette méthode, l'identification comporte deux phases : la phase d'apprentissage qui consiste à découper les textes de chaque langue en mots, puis segmenter chaque mot en bi-grammes pour construire à la fin un profil bi-grammes et l'utiliser comme profil de référence, ensuite calculer la fréquence ou la probabilité d'apparition de chaque bi-gramme dans ce profil. La phase de diagnostique qui consiste à établir le profil bi-gramme du nouveau texte T, puis rechercher le profil de référence le plus proche en utilisant comme mesure de distance avec la langue L le produit des probabilités des bi-grammes du nouveau texte T qui apparaissent dans le profil de la langue L. Cette méthode suppose la possibilité de découpage des textes en mots ce qui n'est pas le cas pour certaines langues. En plus, elle se base sur les bi-grammes, alors qu'il est nécessaire de ne pas négliger les trigrammes ou les quadri-grammes pour conserver la spécificité de chaque langue. Par exemple le quadri-gramme « tion » caractérise le français et l'anglais, si on le découpe en bi-grammes comme suit : « ti » et « on », le système trouve des difficultés pour les différencier avec «ti» et «on» de l'espagnol et du portugais [9], [15], [16].

* La distance de Cavenar et Trenkle CT [5] : cette méthode comporte aussi deux phases : la phase d'acquisition qui consiste à établir pour chaque langue L un profil trigramme pour l'utiliser comme profil de référence. La phase de diagnostique qui consiste à construire le profil trigramme du texte T à identifier la langue, puis calculer les distances entre ce profil et les profils de référence des différentes langues. La distance à calculer repose sur la somme des écarts de positions (les rangs) entre chaque trigramme dans le profil du nouveau texte T et ce même trigramme dans le profil de référence de chaque langue L si le trigramme est présent, sinon on lui attribue un écart maximal. La langue du nouveau texte est celle dont la distance est minimale. Formellement, la distance entre le profil du nouveau texte P_T et le profil de la langue P_L se calcule comme suit :

$$CT(P_T,\ P_L) = Min \begin{cases} \sum_{ng \in P_T} \left| Pos_{P_L(ng)} - Pos_{P_T(ng)} \right|. & si < ng > est\ présent \\ DMAX. & si\ < ng > est\ absent \end{cases} \quad (1)$$

Où : ng : un trigramme
P_T, P_L : profil du nouveau texte T, profil de la langue L
$Pos_{P_T(ng)}$, $Pos_{P_L(ng)}$: la position du trigramme «ng» dans les profils P_T , P_L si « ng » appartient à ce profil.

* La distance de Kullbach-Leibler (KL) [17] : basée sur l'entropie relative de Kullbach et Leibler comme mesure de distance. Formellement cette distance est calculée par la relation suivante :

$$KL(T_1\ , T_2) = \sum_{ng} f_2\ (ng).\ Log\left(\frac{f_2(ng)}{f_1(ng)}\right) \quad (2)$$

Où : T_1 , T_2 : des textes
$f_1(ng)$, $f_2(ng)$: les fréquences des n-grammes «ng» dans les textes T_1 , T_2
si le n-gramme « ng » est absent dans un texte T_i une demi-fréquence est alors ajoutée pour éviter que le score tombe vers -ω

* Distance de khi2 (χ^2) [15] : formellement cette distance est présentée comme suit :

$$\chi 2\ (T_1\ , T_2) = \sum_{ng} \left[\frac{(f_1(ng) - f_2(ng))^2}{f_2(ng)}\right] \quad (3)$$

avec : $f_1(ng)$, $f_2(ng)$: les fréquences du n-gramme « ng » dans les textes T_1 , T_2

$$f_i(ng) = \frac{Nb.Occurrences\ de\ «\ ng\ »\ dans\ Ti}{Total\ n-grammes\ dans\ Ti} \quad (4)$$

Méthodes classiques utilisées dans la catégorisation : Plusieurs autres méthodes existent dans le domaine de catégorisation de documents, leur difficulté commune est la très grande dimension. Parmi ces méthodes, on cite : les arbres de décision (ID3, C4.5, CART, …) [18] qui nécessitent la réduction de dimension, les réseaux de neurones avec rétro-propagation, les SVM et l'approche RBF [19], [20].

3 La méthode proposée

Notre méthode est inspirée de la méthode de [5] avec les améliorations suivantes :
[5] dans leur méthode exige le tri des profils des différentes langues, ainsi que le profil du nouveau texte selon l'ordre décroissant des fréquences avant tout calcul ce qui n'est pas nécessaire pour notre méthode. Cela permet de gagner un temps considérable exigé par l'opération de tri. [5] travaillent seulement avec des trigrammes bien que notre méthode est plus générale (n=3, 4, 5, …). La distance utilisée par [5] repose sur la somme des écarts des

positions (des rangs) entre chaque trigramme du profil du nouveau texte et ce même trigramme dans le profil de référence de chaque langue si le trigramme est présent, sinon on lui attribue un écart maximal. Ici, on peut souligner deux inconvénients pour cette distance : le premier est que le calcul de la somme des écarts de rangs demande un effort algorithmique énorme surtout pour des corpus de grandes tailles, le deuxième est au niveau du choix de l'écart maximal lorsque le trigramme est absent, aucune méthode n'est spécifiée pour trouver cet écart maximal. Pour pallier aux deux inconvénients notre méthode propose le suivant : on prend chaque n-gramme (*n*=3, 4, 5) du profil du nouveau texte et on cherche dans le profil de chaque langue, s'il existe on lui affecte la valeur 1, sinon on lui affecte la somme des fréquences de tous les n-grammes du corpus, puis on calcule la somme qui représente la distance. Le texte sera affecté à la langue dont la distance est minimale.
Formellement, la distance proposée par notre méthode est calculée comme suit :

$$CT(P_T, P_L) = Min \begin{cases} \sum_1^{ng \in P_T} a_i \quad a_i = 1 \quad si < ng > est\ présent \\ som\_freq \qquad\qquad si < ng > est\ absent \end{cases} \tag{5}$$

4 Expérimentations

4.1 Corpus d'apprentissage

Nous avons utilisé un corpus hétérogène multilingue constitué de 425 documents rédigés en 06 langues (Ar, Fr, En, All, Esp, Ita).

4.2 Corpus de test

Le corpus de test est à son tour constitué d'une collection de 90 documents rédigés dans les six langues précitées.

Les deux corpus d'apprentissage et de test ont été perfectionnés avec un effort personnel, leurs documents sont issus des dépêches de journaux internationaux (le Monde, Newsweek, Der Spiegle, Aljazeera.net,…). On a préféré utiliser des documents de taille raisonnable (entre 1 et 3ko) pour faciliter leur traitement et leur segmentation en mots et en n-grammes. Ces deux corpus peuvent être répartis comme suit :

Table 1. Corpus d'apprentissage et de test utilisés dans les expérimentations

	Corpus d'apprentissage	Corpus de test
Langue	Nombre de textes	Nombre de textes
Arabe	68	17
Français	80	19
Anglais	79	20
Allemand	54	10
Espagnol	73	19
Italien	46	5
Totaux	425	90

4.3. Prétraitement effectué sur les corpus d'apprentissage et de test

Avant de procéder à la phase d'identification de la langue proprement dite, une phase de prétraitement automatique sur les corpus d'apprentissage et de test est indispensable. Cette phase comporte plusieurs taches, à noter :

- L'élimination des caractères inutiles (signes de ponctuation, chiffres, caractères spéciaux, abréviations et caractères isolés, …).

- La conversion des majuscules en minuscules.
- Un traitement morphosyntaxique sur le texte (pour l'arabe : utilisation d'un seul hamza, ya et alif maksoura avec et sans hamza, signes de vocalisation tels que : shadda et attanwin, … etc).
- Le découpage de textes en mots et en n-grammes (n=3, 4, 5).
- La fixation d'un seuil minimal pour la fréquence et l'élimination des n-grammes hapax dont les fréquences sont inférieures au seuil choisi.

4.4. Traitements effectués

Après avoir terminé la phase de prétraitement automatique sur les deux corpus de travail, on effectue les traitements suivants :

- Pour le découpage des textes on a utilisé les deux approches : sac de mots, n-grammes de caractères.
- On a testé les résultats pour des seuils de fréquences variables (s=2, 3, 4).
- Pour le découpage de textes en n-grammes, on a testé les résultats pour des valeurs variables de n (n=3, 4, 5).
- Pour l'algorithme d'apprentissage appliqué, on a choisi l'algorithme des k- plus proches voisins avec k=1, ainsi que l'algorithme de naïve Bayes. Puis on a appliqué l'algorithme 1-PPV avec plusieurs distances telles que : la distance CT [5], la distance KL [15], la distance χ^2 [15].
- En fin, on a appliqué notre méthode proposée avec la nouvelle distance associée.

5 Evaluation des résultats obtenus

Table 2. Segmentation du corpus d'apprentissage en mots (le seuil s = 2)

Lng	# txt	# mots bruts	# mots épurés	# mots fréq
Ar	68	19010	4960	1933
Fr	80	14698	5757	2646
En	79	16672	5472	2761
All	54	2792	1845	437
Esp	73	15249	5748	2498
Ita	46	2259	1507	332

Table 3. Segmentation du corpus d'apprentissage en n-grammes (s = 2)

Lng	# txt	# 3g bruts	# 3g épurés	#3g fréq	# 4g bruts	# 4g épurés	#4g fréq	# 5g bruts	# 5g épurés	# 5g fréq
Ar	68	53609	5276	3318	52707	10397	5005	52490	16162	6755
Fr	80	63956	7271	4862	63172	15998	6777	63048	24637	10946
En	79	61128	6796	4608	59347	15828	8813	59104	25016	10706
All	54	29906	5052	2895	28370	10405	4364	28134	15714	4659
Esp	73	58399	6261	4138	58222	14157	7654	58108	21791	9454
Ita	46	21066	3638	2085	20007	7647	3307	19849	11566	3360

Table 4. Calcul de : *taux-suc, taux-err* pour tous les algorithmes d'apprentissage (App. sac de mots)

Algo	Taux_suc	Taux_err
N.Bayes	95,55	4,45
1ppv avec CT	98,88	1,12
1ppv avec KL	97,03	2,97
1ppv avec χ^2	98,74	1,26
la nouvelle méthode	98,87	1,13

Table 5. Calcul de : *taux-suc, taux-err* pour tous les algorithmes d'apprentissage et (App. n-grammes)

Algo	N=3		N=4		N=5	
	Taux_ suc	Taux_ err	Taux_ suc	Taux_err	Taux_ suc	Taux_err
N.Bayes	96,66	3,34	98,72	1,28	100	0
1ppv avec CT	98,91	1,09	99,20	0,80	98,75	1,25
1ppv avec KL	97,89	2,11	98,14	1,86	96,37	3,63
1ppv avec χ^2	98,86	1,14	100	0	97,33	2,67
la nouvelle méthode	98,91	1,09	100	0	98,85	1,15

Après les prétraitements automatiques effectués sur les deux corpus d'apprentissage et de test comme c'est indiqué dans la section 4.3, on a procédé à la phase de segmentation des textes en unités de base (*tokens*), on a utilisé pour cela la segmentation en mots, en 3-grammes, en 4-grammes, et en 5-grammes. Les tables 2, 3 résument les résultats obtenus par la phase de segmentation. On note ici que l'apprentissage est appliqué sur les mots et les n-grammes les plus fréquents comme c'est montré dans la table 2 (colonne 5), et la table 3 (colonnes 5, 8, 11) à cause de leur nombre restreint et parce qu'ils donnent les meilleurs taux de succès en reconnaissance de la langue. Pour l'apprentissage on a appliqué l'algorithme de naïve Bayes qui est le plus connu dans ce domaine et qui donne de bons résultats par rapport à d'autres méthodes, plus un algorithme 1-ppv en faisant référence à plusieurs pseudo-distances. Et enfin, on appliqué notre nouvelle méthode dotée de sa propre pseudo-distance. Les résultats obtenus montrent que notre méthode proposée donne toujours des résultats égales ou meilleurs par rapport aux résultats obtenus en appliquant d'autres pseudo-distances (avec un taux de reconnaissance > 98%). Pour un découpage en 4-grammes, le taux de reconnaissance est de 100% (les tables 4 et 5).

Vu les données utilisées, notamment les corpus de textes utilisés dans les expérimentations qui ont été perfectionnés d'une manière arbitraire, ainsi que les implémentations faites pour tous les algorithmes utilisés, nous estimons que les résultats obtenus sont très significatifs et encourageants et peuvent être améliorés dans d'autres travaux.

6 Conclusion et perspectives

Dans ce papier, nous avons traité le problème d'identification de la langue d'un texte dans un corpus de textes hétérogène multilingue. Nous avons exposé les deux approches les plus connues dans ce domaine pour segmenter un texte en unités de base, notamment : l'approche sac de mots et l'approche n-grammes. Les implémentations réalisées montrent la limitation de l'approche sac de mots en faveur de l'approche n-grammes qui reste plus générale, indépendante de la langue et donne toujours les meilleurs résultats. Pour l'identification de la langue, on a implémenté plusieurs algorithmes basées sur de différentes distances à savoir : la distance CT, la distance KL, la distance χ^2. A la lumière de toutes ces distances et les résultats obtenus, nous avons proposé une nouvelle méthode dotée de sa propre distance. Cette méthode est inspirée de la méthode CT avec l'avantage que la nouvelle méthode ne nécessite pas de tri, elle est basée sur une distance très simple à calculer et moins couteuse du coté algorithmique surtout pour des corpus de grandes tailles. Les résultats obtenus en appliquant la nouvelle méthode sont très significatifs en terme temps de calcul, et exactitude de résultats obtenus pour l'identification de la langue. Comme perspectives à ce travail, nous proposons d'appliquer la nouvelle méthode sur des

corpus de textes semi-structurés, et de la généraliser avec la catégorisation thématique des textes ou toute autre tache liée.

7 Références

1. MUSTONEN S.: Multiple Discriminant Analysis in Linguistic Problems., 147, 195--197 (1981), *Statistical Methods in Linguistics*, vol. Page visitée le 15 juin 2000 à l'adresse http://www.nodali.sics.se/bibliotek/kval/smil, 1965.
2. SOUTER C., CHURCHER G., HAYES J., HUGHES J., JOHNSON S.: Natural Language Identification Using Corpus-Based Models, *Hermes Journal of Linguistics*, vol. 13, 1994.
3. GIGUET E. : Méthode pour l'analyse automatique de structures formelles sur documents multilingues , PhD thesis, Université de Caen, France, 1998.
4. BEESLEY K.: Language Identifier: A Computer Program for Automatic Natural Language Identification on On-Line Text , *Proceedings of the 29th Annual Conference of the American Translators Association*, 1988, p. 47–54.
5. CAVNAR W., TRENKL J.: Gram Based Text Categorization , *Symposium on Document Analysis and Information Retrieval*, Las Vegas, 1994.
6. DUNNING T.: Statistical Identification of Languages , rapport nMCCS 94-273, 1994, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.
7. GREFENSTETTE G.: Comparing Two Language Identification Schemes , *Proceedings of the 3rd International Conference on the Statistical Analysis of Textual Data*, Rome, Italy, 1995.
8. MILNE R.M., O'KEEFE R.A., TROTMAN A.: A study in language identification, ADCS *December 05 - 06 2012, Dunedin, New Zealand, 2012*
9. ZAMPIERI M., GEBRE B.G.: Automatic identification of language varieties: the case of portuguese, *Proceedings of KONVENS 2012 (Main track: poster presentations), Vienna, September 20, 2012*
10. SAHAMI M. : Using Machine Learning to Improve Information Access , PhD thesis, Computer Science Department, Stanford University, 1999.
11. GEIGER W.M., RAUCH J., HORNIK K .: Text categorization in R: A Reduced N-grams Approach, *Springer-Verlag Berlin Heidelberg, 2012*
12. SHANNON C.: The Mathematical Theory of Communication , *Bell System Technical Journal*, vol. 27, 1948, p. 379–423 and 623–656.
13. MILLER E., SHEN D., LIU J., C.NICHOLAS. : Performance and Scalability of a Large-Scale N-gram Based Information Retrieval System , *Journal of Digital Information*, vol.1, n5, 1999.
14. BISKRI I., DELISLE S.: les n-grammes de caractères pour l'aide à l'extraction de connaissance dans des bases de données textuelles multilingues, *TALN 2001, Tours, 2-5 juillet 2001*
15. JALAM R., TEYTAUD O.: Identification de la Langue et Catégorisation de Textes basées sur les N-grams, Journées Francophones d'extraction et de gestion de connaissances, 2002
16. BROWN R.D.: Finding and identifying text in plus de 900 languages, *Published by Elsevier Ltd. All rights reserved, 2012*
17. SIBUN P., REYNAR J.: Language Identification: Examining the Issues , *Symposium on Document Analysis and Information Retrieval*, Las Vegas, 1996, p. 125–135.
18. RAKOTOMALA R.: Arbres de décision, *Revue MODULAD, 2005, N°33*
19. JOACHIMS T.: Text Categorization with SVM: Learning with Many Relevant Features, *Machine Learning: ECML-98, 10th European Conference on Machine Learning*, 1998.
20. DIMITRIOS A., PRITSOS C., STAMATATOS E.: Open-Set Classification for Automated Genre Identification , *ECIR 2013, LNCS 7814, pp. 207–217, 2013. Springer-Verlag, Berlin Heidelberg, 2013*