Identification d'auteurs par apprentissage automatique

Jordan FRERY*, Christine LARGERON*, Mihaela JUGANARU-MATHIEU**

* Université Jean Monnet, Saint-Etienne, France.
Jordan.Frery, christine.largeron@univ-st-etienne.fr
**Institut H. Fayol, École Nationale Supérieure des Mines
mathieu@emse.fr

Résumé. Etant donné un ensemble de documents rédigés par un même auteur, le problème d'authentification d'auteurs consiste à décider si un nouveau texte a été rédigé ou non par cet auteur. Pour résoudre ce problème, nous avons proposé et implémenté différentes approches : comptage de similarité, techniques de vote et apprentissage supervisé qui exploitent différents modèles de représentation des documents. Les expérimentations réalisées à partir des collections de la compétition PAN-CLEF 2013 et 2014 ont confirmé l'intérêt de nos approches et leur performance en termes de temps de traitement.

1 Introduction

Qui n'a pas dit, un jour, en écoutant la radio : "Mais ça ressemble à Supertramp ou à une musique de Chopin ou à une musique baroque" ? Sur la base d'un court morceau écouté on peut en effet identifier directement l'auteur ou le placer dans une catégorie même si on ne connait pas forcément le morceau. Si c'est un chanteur, on le reconnait facilement au timbre de sa voix, pour un morceau de musique classique, l'interprétation peut varier et on détecte plutôt la ligne musicale. Pour les documents textuels, ce problème d'authentification d'auteur présumé est récurrent, et la fouille de texte peut s'avérer très utile. Ainsi, par exemple pour authentifier une élégie de Shakespeare en 1995 [1] des techniques telles que le comptage exclusif des mots et la prise en compte de mots rares ont été employés avec succès (Foster (1996)). Le champ littéraire n'est cependant pas le seul concerné. Le problème d'authentification d'un auteur apparait aussi dans bien d'autres applications, comme dans le domaine juridique par exemple pour l'authentification d'un testament ou dans le cadre des investigations anticriminelles ou antiterroristes pour identifier la provenance d'une demande de rançon ou de posts émis sur des forums de discussion du Dark Web (Abbasi et Chen (2005)). Le marketing peut également être intéressé par le profiling des auteurs des blogs ou des commentaires sur le Web.

Dans le cas de textes écrits, on peut plus généralement distinguer trois variétés de problèmes liés à la détermination d'un auteur inconnu :

— l'extraction de profil (Author Profiling) : il s'agit d'indiquer à partir d'un texte des éléments du profil de son auteur comme, par exemple, la tranche d'âge et le genre

1. http://www.lexpress.fr/informations/c-est-shakespeare-qu-on-ressuscite_614521.html

(Rangel et al. (2013)) ou l'appartenance à une catégorie particulière comme celle des criminels potentiels (Inches et Crestani (2012))
— la reconnaissance de l'auteur (Author Verification ou Author Recognition) : il s'agit de vérifier parmi une liste des auteurs possibles lequel est le bon
— l'identification d'un auteur : il s'agit de décider si un texte donné a été écrit par l'auteur d'un autre groupe de documents

Ainsi, dans ces trois problèmes, on s'interroge sur l'auteur d'un document écrit, et dans ces trois cas pour répondre, il est nécessaire de représenter de façon appropriée le document à explorer et de pouvoir le comparer avec d'autres. Toutefois, nous pensons qu'il est illusoire de rechercher une *empreinte* d'un auteur sur un texte qu'on pourrait comparer avec des empreintes extraites d'autres textes et qui serait unique au même titre qu'une empreinte digitale. Nous pensons qu'il faut utiliser divers espaces de représentation pour les textes à analyser selon la langue d'origine ou encore le genre ou la qualité du document. Dans cet article, où nous nous intéressons plus spécifiquement à des problèmes d'identification d'auteurs à partir de documents rédigés dans diverses langues et de différents types (textes littéraires courts ou longs, articles de presse ou publications, blogs) nous avons exploré différents modes de représentation des documents. Nous avons ensuite proposé de formaliser l'identification d'auteurs comme un problème de classement que nous avons résolu de trois façons : à l'aide d'un algorithme original de comptage de similarité (DCM), puis avec deux autres méthodes qui exploitent ce comptage, par une technique de vote (DCM-voting), par apprentissage automatique (DCM-classifier).

Notre article est organisé de la manière suivante : après la section 2 consacrée aux travaux relatifs à l'identification d'auteurs, nous définissons plus formellement le problème dans la section 3, puis nous décrivons les trois méthodes proposées pour le résoudre dans la section 4. La section 5 présentera les résultats des expériences réalisées afin d'évaluer l'intérêt de ces approches et de les comparer à celles de l'état de l'art. Des conclusions seront présentées dans la dernière section.

2 Etat de l'art

L'identification d'auteur peut être définie comme un problème de classification de textes : *Etant donné un ensemble, grand ou réduit à un seul élément, de documents d'un même auteur, il faut déterminer si un nouveau document a été écrit par le même auteur que les autres ".*

Il s'agit donc d'un problème de classement supervisé binaire dont la réponse attendue est binaire ("oui" ou "non") ou une probabilité d'appartenance à l'ensemble de documents fournis. Toutefois, une des spécificités de ce problème de classement est que seuls des exemples d'une des deux classes sont donnés : les documents rédigés par l'auteur, mais la seconde classe n'est pas explicitée. De plus, parfois le nombre d'exemples positifs est réduit à un seul document, ce qui rend la tâche particulièrement difficile.

Pour pallier l'absence d'exemples négatifs, on peut essayer d'en produire. C'est la voie explorée par différents auteurs parmi lesquels figurent (Seidman (2013)) qui construisent une classe d'"imposteurs" choisis aléatoirement sur la base des dix mots les plus fréquents figurant dans les documents disponibles pour remplir la classe du "non". D'autres auteurs, comme Zhang et al. (2014) et Halvani et al. (2013), transforment ce problème de classification à deux classes en un problème avec plusieurs classes, soit en rajoutant des classes extérieures,

soit en transformant la classes initiale en plusieurs. Les même auteurs (Halvani et al. (2013)) augmentent la taille de la classe des documents connus quand celle-ci est réduite à un seul document. Ainsi, ces approches permettent de revenir à un problème classique de classement supervisé mais, lors de la construction des exemples négatifs, elles sont confrontées au risque de choisir des documents trop proches ou trop éloignés des documents déjà fournis.

Outre la question des données disponibles pour résoudre le problème, l'identification d'auteurs est ensuite confrontée à deux autres questions classiques en fouille de texte : comment représenter les documents et, une fois l'espace de représentation choisi, quelles méthodes appliquer pour résoudre le problème de classement ?

Comme nous l'avons déjà remarqué, l'identification d'auteurs peut être réalisée à partir de documents très différents : méls (de Vel et al. (2001); Chaurasia et Kumar (2010)), programmes, parties des oeuvres littéraires ou parties des documents de la vie de tous les jours, texte plat (Zhang et al. (2014)), extraits de chat (Inches et Crestani (2013)) ou séquences de commandes Unix (Szymanski et Zhang (2004)). Le choix des caractéristiques examinées, on parle des caractéristiques *stylométriques*, dépend du type de document, parfois de la langue et aussi de la qualité du texte initial. Les caractéristiques dites "spécifiques" aux applications portent plutôt sur le comptage des tabulations et autres séparateurs ou l'analyse des caractéristiques spécifiques, telles que la position des parenthèses et des crochets fermants pour les programmes, et des lignes vides pour les méls. Les caractéristiques sémantiques sont prises en compte plutôt pour des textes issus du web (forum et chat), comme, par exemple, l'usage des abréviations ou des mots démonstratifs fréquents ("well") ou des transcriptions concentrées des expressions orales ("sse u"). On peut également considérer des caractéristiques syntaxiques comme des fautes d'orthographe ou les abréviations. Si on se place dans un cadre générique (authentification d'auteur dans diverses langues et dans divers genres), on utilise plutôt des caractéristiques de type caractère ou mot, ou suites de caractères ou de mots (n-grams) (Chaurasia et Kumar (2010); Szymanski et Zhang (2004)). On peut aussi avoir recours à un étiqueteur lexical et syntaxique mais son usage augmente considérablement le temps de traitement (Juola et Stamatos (2013)) et les résultats vont dépendre de sa qualité (Vilariño et al. (2013)). Lorsque le choix de ces caractéristiques est fait, les documents peuvent être transformés en vecteurs en utilisant, le plus souvent, tf-idf comme pondération ou uniquement la fréquence. Ensuite, selon la représentation du texte adoptée, on peut comparer les documents à l'aide de fonctions "classiques" de similarité telles que le cosinus, la corrélation, moins souvent des mesures de compression de données comme la Fast Compression Distance (Cerra et al. (2014)) ou la Common N-Gram dissimilarity (Layton et al. (2013)).

Pour ce qui concerne la résolution du problème de classement lui-même, on peut appliquer des méthodes "classiques" telles que les k plus proches voisins (k-NN) (Zhang et al. (2014); Ghaeini (2013); Halvani et al. (2013)) ou les SVM (Vilariño et al. (2013)). Certains auteurs (Dam (2013); Layton et al. (2013); Jankowska et Milios (2013)) proposent des méthodes basées sur le choix d'un seuil ou d'un vote et des formules de calcul de l'éloignement entre le document d'auteur inconnu et les autres. Les différences entre ces approches résident dans la phase de prétraitement, dans l'extraction des caractéristiques et dans le choix du seuil et de la fonction de dissimilarité.

Par rapport aux travaux antérieurs, notre contribution se situe dans un cadre plus large avec l'objectif de proposer une méthodologie générique, applicable à des collections très différentes tant par le genre des documents que par le langage. Ceci nécessite la mise en place d'une

approche permettant de choisir automatiquement la représentation textuelle la mieux adaptée pour un corpus donné.

3 Définition du problème et représentation des documents

Le problème d'identification d'auteur peut être défini de la façon suivante. Etant donné un corpus composé de documents d'un même type (mel, blog, roman, code, *etc.*) écrits dans un même langage (anglais, français, Java, *etc.*) on dispose pour chaque problème p d'un ou plusieurs documents A_p du corpus qui ont été rédigés par un même auteur et d'un document u_p dont l'auteur est inconnu. L'objectif est de déterminer si u_p a été écrit ou non par le même auteur que les documents de A_p. Si on dispose d'un échantillon d'apprentissage, autrement dit d'un ensemble de problèmes P tel que pour chaque problème $p \in P$ on sait si le document inconnu u_p a été rédigé ou non par le même auteur que les documents associés A_p, alors on peut formaliser le problème comme un problème de classement supervisé binaire et le résoudre à l'aide de méthodes d'apprentissage automatique. La difficulté consiste alors à déterminer d'une part un ou des espaces de représentation des documents appropriés et d'autre part à construire à partir de ces représentations des facteurs descriptifs des documents inconnus permettant de prédire efficacement si chaque document u_p a été ou non produit par le même auteur que les documents de A_p qui lui sont associés.
Parmi les modèles les plus connus et les plus utilisés pour représenter des documents figure le modèle tf-idf introduit par Salton et al. (1975). Un document d est représenté par un vecteur $(w_1, \ldots, w_j, \ldots, w_{|T|})$ tel que le poids w_j du terme t_j dans d correspond au produit de la fréquence tf_j du terme t_j dans d par le pouvoir discriminant $idf(j)$ de t_j. Ce modèle est très efficace notamment pour identifier des termes (caractères, mots ou séquences de mots ou caractères correspondant à des n-grams) qui sont fréquents dans un document et rares dans les autres. Mais, comme nous l'avons souligné en introduction, d'autres caractéristiques peuvent être prises en compte pour représenter les documents. De plus, nous pensons qu'il n'existe pas un modèle de représentation universel adapté à tous les documents mais que le choix de cet espace de représentation doit dépendre du type de documents et du langage.

Ceci nous a conduit à considérer d'autres espaces de représentation indiqués dans le tableau 1. Outre le modèle tf-idf défini à partir des mots, avec élimination des mots outils à l'aide d'un dictionnaire (R5) ou en considérant leur fréquence (R4), des suites de mots ou de caractères (R1, R2, R3), nous avons introduit trois autres modèles de représentation (R6, R7 et R8) visant à caractériser le style d'écriture du document. Dans le modèle R6, la moyenne et l'écart type du nombre de mots par phrase sont associés au document. Le modèle R7 attribue à chaque document une mesure de diversité du vocabulaire définie comme le nombre de mots différents employés divisé par le nombre total d'occurrences de mots (*i.e.* la longueur du document). Le modèle R8 correspond au modèle de Salton dans lequel on considère les caractères de ponctuation au lieu des termes (mot ou caractère n-grams). Enfin, le modèle R678 correspond à la concaténation des trois modèles précédents : chaque document est représenté par un vecteur indiquant la moyenne par phrase des caractères de ponctuation " :" , " ;" , ",", la moyenne et l'écart type du nombre de mots par phrase et la diversité du vocabulaire.

	Espace de représentation	
	Terme	Modèle
$R1$	Caractère 8-grams	tf-idf
$R2$	Caractère 3-grams	tf-idf
$R3$	Mot 2-grams	tf-idf
$R4$	Mot 1-gram	tf-idf sans les 30% des plus fréquents mots
$R5$	Mot 1-gram	tf-idf sans les mots vides
$R6$	Phrases	mots par phrase en moyenne et en écart type
$R7$	Diversité de vocabulaire	nombre total de termes divisé par le nombre total des occurrences des mots
$R8$	Ponctuation	nombre moyen de signe de ponctuation par phrase caractères pris en compte : "," " ;" " :" "(" ")" " !" " ?"
$R678$	Concaténation	R6 + R7 + R8

TAB. 1: La liste de espaces de représentation considérés

4 DCM, DCM-voting et DCM-classifier

Ayant choisi un des espaces de représentation, on peut comparer les documents deux à deux à l'aide de mesures de similarité comme le cosinus et le coefficient de corrélation ou avec la distance euclidienne et appliquer une des trois méthodes (DCM, DCM-voting, DCM-Classifier) que nous avons proposées pour résoudre le problème d'identification d'auteur. La première méthode DCM permet de traiter directement le problème d'identification $p \in P$, en considérant uniquement les similarités entre les vecteurs décrivant les documents suivant un des espaces. Les deux autres méthodes, basées sur DCM, permettent de combiner différentes représentations des documents, par une méthode de vote dans le cas de DCM-voting, à l'aide d'une méthode d'apprentissage supervisée nécessitant la construction d'attributs prédictifs pour DCM-classifier. Ces différentes méthodes sont décrites dans les sections suivantes.

4.1 Méthodes de comptage de similarités : DCM et DCM-voting

Etant donné un problème $p \in P$ défini par un ensemble A_p de documents rédigés par un même auteur et un document u_p dont l'auteur est inconnu, représentés dans un même espace, et un seuil de décision δ , la méthode *DCM*, décrite par l'algorithme suivant, fournit en sortie la valeur *True* si l'auteur de u_p est le même que celui des documents de A_p ou la valeur *False* dans le cas contraire. Cette méthode exploite les similarités (ou distances) entre tous les documents disponibles. Elle consiste à assigner le document u_p au même auteur que les documents de A_p si la plupart d'entre eux sont plus proches de u_p qu'ils ne le sont des autres documents de A_p. Plus précisément la plus grande similarité de chaque document $d_x \in A_p$ aux autres documents de A_p est calculée puis comparée à la similarité de d_x à u_p. Si la première est inférieure à la seconde, un compteur est incrémenté. Après examen de tous les documents de A_p (fin de la boucle **for** extérieure), il comptabilise la proportion de documents de A_p qui sont les plus proches de u_p que des autres de A_p et si cette proportion est supérieure au seuil fixé δ, alors l'auteur du document inconnu u_p est le même que celui des autres documents.

procedure DISSIMILARITY COUNTER METHOD
Input data : A_p, u_p, δ
Result : *True* si A_p et u_p sont du même auteur, *False* sinon
A_p : ensemble de documents connus (même auteur)
u_p : document d'auteur inconnu
δ : seuil fixé
smin : similarité minimale dans sous-ensembles deA_p

count $\leftarrow 0$
for $d_x \in A_p$ **do**
  smax $\leftarrow 0$
  for $d_y \in A_p - \{d_x\}$ **do**
    compute $s(d_x, d_y)$ // similarité entre d_x et d_y
    if *smax* $< s(d_x, d_y)$ **then**
      smax $\leftarrow s(d_x, d_y)$
    end if
  end for
  if $s(u, d_x) >$ *smax* **then**
    $count \leftarrow count + 1.$
  end if
end for
if $count > \delta$ **then return** *True*
else return *False*
end if
end procedure

Cette méthode présente l'avantage d'être simple et rapide à mettre en oeuvre et elle permet de traiter un problème d'identification $p \in P$ indépendamment des autres. En revanche, elle n'exploite qu'un seul mode de représentation des documents.

Pour pallier ce défaut, on peut avoir recours à une méthode de vote *DCM-voting* consistant simplement à appliquer la méthode *DCM* en considérant plusieurs espaces de représentation des documents, de préférence en nombre impair, puis à affecter le document inconnu à la classe majoritairement retournée par les différentes exécutions. Cependant, comme nous l'avons souligné précédemment, tous les espaces de représentation ne sont pas équivalents et il serait souhaitable de pouvoir ajuster leur poids dans la décision finale ; ce qui est difficile à faire en pratique même pour un expert ; de même que le choix du seuil δ. Pour toutes ces raisons, nous proposons une autre méthode plus générale permettant d'exploiter simultanément plusieurs modes de représentation des documents et d'ajuster automatiquement, par apprentissage automatique, leur importance dans l'identification de la classe des documents inconnus.

4.2 La méthode DCM-classifier

Dans le cadre de l'apprentissage supervisé, on suppose que pour un sous-ensemble P_A de problèmes de P, on sait en fait si les documents inconnus u_p ont ou non été produits par le même auteur que les documents qui lui sont associés *i.e.* on dispose en plus de la classe $class(u_p)$ des documents inconnus à savoir *même auteur* ou *auteur différent*. Ce sous-ensemble P_A est décomposé en un échantillon d'apprentissage P_a utilisé pour construire un

modèle de décision et un échantillon test employé pour l'évaluer. La phase d'apprentissage permet de mettre en relation des facteurs descriptifs (ou attributs) des documents avec leur classe de façon à pouvoir ensuite identifier l'auteur d'un nouveau document dont la classe est inconnue uniquement à partir de ces facteurs descriptifs. Il est clair que la qualité du modèle dépend largement du pouvoir prédictif de ces facteurs que nous proposons de définir de la façon suivante.

Pour chaque espace de représentation R_v, $v \in \{1, .., V\}$, chaque document u_p est décrit par deux attributs $count_v(u_p)$ et $mean_v(u_p)$ respectivement définis à l'aide d'une mesure de similarité s par :

$$count_v(u_p) = |\{di \in A_p / min\{s(d_i, d_j), d_j \in A_p - d_i\} > s(d_i, u_p)\}|$$

$$mean_v(u_p) = \frac{1}{|A_p|} \times \sum\nolimits_{d_i \in A_p} s(d_i, u_p)$$

Un dernier attribut $TOT_{count}(u_p)$, basé sur tous les espaces de représentation est également calculé afin d'avoir une description plus synthétique. Il est défini par :

$$TOT_{count}(u_p) = \sum_{v=1}^{V} count_v(u_p)$$

Ainsi lors de l'apprentissage, on considère les documents u_p de chaque problème p de P_a décrit par ces attributs descriptifs prédictifs et par leur classe réelle ($count_v(u_p)$ et $mean_v(u_p)$, $\forall v \in \{1, .., V\}$, $TOT_{count}(u_p)$ et $class(u_p)$). Compte tenu du caractère numérique de ces attributs descriptifs, plusieurs méthodes d'apprentissage supervisé peuvent alors être employées (SVM, *etc*). Dans le cadre des expérimentations, nous avons privilégié les arbres de décision qui ont l'avantage d'intégrer une phase de sélection des attributs en fonction de leur pouvoir prédictif ; ce qui permet de favoriser selon la famille de problèmes considérés tel ou tel espace de représentation. De plus, ils permettent aussi d'ajuster automatiquement les paramètres du modèle.

5 Expérimentations et résultats

Nous avons évalué les trois méthodes DCM, DCM-voting et DCM-classifier présentées dans la section précédente sur les corpus créés pour les challenges PAN CLEF 2013 et PAN CLEF 2014 [2]. Pour chaque année, on dispose d'une collection d'apprentissage (app) et d'une collection de compétition qui sert pour l'évaluation (eval). Chaque collection de PAN est composée de plusieurs corpus et un corpus contient des problèmes d'une même langue (européenne) et d'un même genre rédactionnel. En 2013 les deux collections, app2013 et ev2013, contenaient trois corpus de langues différentes : anglais, espagnol et grecque alors qu'en 2014, elles contenaient six corpus (EN : Romans anglais, EE : Essais anglais, SP : Articles espagnols, GR : Articles grecques, DR : Revues néerlandaises, DE : Essais néerlandais). Chaque corpus comporte un certain nombre de problèmes. Chaque problème est composé de documents "connus" (écrits par un même auteur), au moins un, et d'un document "inconnu", dont

2. `http://clef2014.clef-initiative.eu/`

	EN	EE	SP	GR	DR	DE	Total
app13	10	-	4	20	-	-	34
eval13	20	-	10	20	-	-	50
app14	100	200	100	100	100	96	696
eval14	200	200	100	100	100	96	796

TAB. 2: La taille des corpus utilisés

on doit indiquer s'il a été rédigé par le même auteur que les autres. Le tableau 2 résume le nombre de problèmes contenus par corpus.

Les résultats obtenus sur chaque corpus des collections ev2013, app2014 et ev2014 ont été évalués à l'aide des indicateurs habituels de précision, de rappel et avec la mesure $F1$.

Le taux d'erreur indiqué par la mesure $F1$ étant très synthétique, pour comparer les méthodes, on a également utilisé l'indicateur de performance AUC qui mesure l'aire de la courbe ROC (Davis et Goadrich (2006)).

Pour la collection ev2014, seuls les indicateurs de performances calculés par la plateforme du challenge pour chaque corpus sont disponibles : AUC, l'indicateur $c@1$, le produit des deux indicateurs et le temps d'exécution. L'indicateur $c@1$ permet de donner plus d'importance à une réponse correcte par rapport à l'absence de décision (i.e. une probabilité d'appartenance à la classe de 0.5). Cet indicateur est défini par :

$$c@1 = \frac{1}{n}(n_c + \frac{n_c}{n} n_u)$$

où n est la taille du corpus, n_c est le nombre de réponses correctes, n_u le nombre de problèmes laissés sans décision.

5.1 Résultats sur la collection 2013

Pour la méthode DCM, nous avons utilisé la représentation $R1$ (caractère 8-grams) qui avait donné les meilleurs résultats sur la collection d'apprentissage de 2013 et fixé δ à $\frac{|A_p|}{2}$ alors que pour DCM-voting, nous avons privilégié les espaces de représentation R1, R2, R3, R4 et R678 (cf. Tableau 1). Pour les quatre espaces ayant trait aux mots ou aux n-grams caractères, les documents sont représentés sous format vectoriel avec la pondération $tf - idf$.

La table 3 présente les résultats produits par les trois méthodes sur la collection eval13 ainsi que les résultats des gagnants de la compétition par corpus puis sur l'ensemble de la collection. La faible précision de DCM-classifier pour le corpus espagnol peut s'expliquer par le manque de problèmes pour cette langue (seulement quatre) dans le corpus d'apprentissage rendant difficile la construction d'un modèle performant. Les trois méthodes produisent des résultats satisfaisants cependant, DCM et DCM-voting sont limités aux problèmes contenant au moins deux textes connus. Si on compare les résultats obtenus par nos méthodes avec ceux des gagnants de la compétition par corpus, alors il n'y a que sur le corpus grec que la méthode DCM-classifier l'emporte avec un score de 85%. Par contre, sur l'ensemble de la collection, DCM-voting comme DCM-classifier obtiennent des résultats meilleurs ou équivalents à ceux du gagnant pour tous les critères d'évaluation (F1, précision et rappel) .

	DCM	DCM-voting	DCM-classifier	Meilleur résultat 2013
Anglais	73.3%	76.7%	75%	80% Seidman (2013)
Espagnol	77.3%	81.8%	60%	84% Halvani et al. (2013)
Grecque	73.3%	82%	85%	83% Seidman (2013)
Mesure F1	73.1%	76.7%	76%	75.3% Seidman (2013)
Precision	74.4%	78.1%	76%	75.3% Seidman (2013)
Rappel	71.8%	75.3%	76%	75.3% Seidman (2013)

TAB. 3: Résultats en terme de mesure $F1$ des trois méthodes sur chaque corpus et de score $F1$, $precision$ et $rappel$ sur toute la collection d'évaluation de 2013.

Corpus	EN	EE	DR	DE	SP	GR
Problèmes#	100	200	100	96	100	100
Problèmes# avec $\|A\| = 1$	100	57	99	62	0	20
AUC	89%	70%	68%	91%	77%	76%

TAB. 4: Résultats de la 10-cross validation de DCM-classifier sur app2014

5.2 Résultats sur les collections 2014

La collection 2014 contient un nombre plus élevé de problèmes et de types de document que la collection de 2013 et elle s'avère plus difficile à traiter puisque plus de la moitié des problèmes ont un seul document connu ($|A| < 2$) ; ce qui rend les méthodes DCM et DCM-voting inadaptées et inefficaces. Pour cette raison, seule DCM-classifier a été évaluée, d'abord sur la collection d'apprentissage en utilisant la technique de 10-validation croisée qui consiste à séparer le corpus à traiter en deux groupes, un pour entrainer le modèle et l'autre pour le tester, puis sur la collection d'évaluation dans le cadre du challenge.

Les résultats obtenus par validation croisée sur l'ensemble d'apprentissage sont présentés dans la table 4. Ils confirment les performances de la méthode DCM-classifier.

Le tableau 5 contient les résultats officiels obtenus lors de la compétition PAN14 in Author Identification, Stamatatos et al. (2014). Ils permettent de comparer notre méthode à celles des autres participants. DCM-classifier nous a permis d'être classé en deuxième position lors de la compétition. Elle fournit de bons résultats en un temps relativement court. Il convient de noter que les temps de traitement affichés par le gagnant de la compétition sont en moyenne supérieurs à trois heures alors que ceux de DCM-classifier sont de l'ordre de quelques secondes.

Un des avantages de la méthode DCM-classifier, basée sur les arbres de décision, est de mettre en évidence les caractéristiques qui permettent le mieux d'identifier l'auteur d'un do-

Corpus	EN	EE	DR	DE	SP	GR
AUC	61 %	72%	60%	90%	77%	68%
c@1	59 %	71%	58%	90%	75%	64%
Temps (minutes)	3 :10	0 :54	0 :08	0 :29	1 :00	0 :57
Position finale	7/13	**1/13**	6/13	**2/13**	4/13	7/12
Position en temps d'exécution	3/13	3/13	3/13	4/13	3/13	3/12

TAB. 5: Résultats de DCM-classifier sur la collection d'évaluation 2014

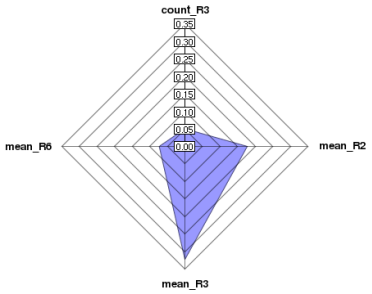

FIG. 1: Importance des attributs pour le corpus GR

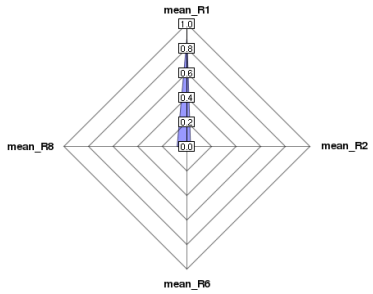

FIG. 2: Importance des attributs pour le corpus DE

Attribut	Utilisation	Taux d'importance	Corpus
$mean_{R1}$	4/6	29%	EE, EN, DE, DR
$mean_{R2}$	5/6	5%	EE, SP, GR, DE, DR
$mean_{R3}$	4/6	10%	EE, EN, SP, GR
$mean_{R4}$	2/6	2%	EE, EN
$mean_{R5}$	2/6	12%	DR, EE
$mean_{R6}$	5/6	7%	EE, SP, GR, DE, DR
$mean_{R7}$	2/6	7%	SP, DR
$mean_{R8}$	5/6	11%	EE, EN, SP, DE, DR
$count_{R2}$	1/6	1%	EE
$count_{R3}$	1/6	1%	GR
TOT_{count}	1/6	8%	SP

TAB. 6: Classement des attributs

cument selon le type de corpus considéré. En effet, les documents sont décrits par des attributs calculés sur différents espaces de représentation mais l'apprentissage intègre une phase de sélection de ceux qui sont les plus discriminants. Ainsi, on peut en déduire pour chaque corpus l'importance de chaque attribut.

Les figures 1 et 2 présentent les différents espaces de représentation utilisés pour deux corpus de langues différentes, on voit que ces espaces sont très différents et que les poids rattachés le sont aussi. La table 6 indique de manière synthétique la liste des attributs les plus utilisés sur l'ensemble des corpus de la collection d'évaluation 2014. Ce résultat confirme l'intérêt de combiner plusieurs espaces de représentation pour résoudre le problème d'identification d'auteurs.

6 Conclusion

Afin de résoudre le problème d'identification d'auteur, nous avons proposé trois méthodes. La première DCM, est une méthode de comptage qui exploite un seul mode de représentation des documents. Elle fournit de bons résultats mais son efficacité est d'autant plus grande que le nombre de documents connus est important. De plus, elle permet uniquement de traiter des problèmes ayant au moins deux documents connus d'un même auteur. Une première extension de cette méthode, DCM-voting permet en partie de pallier les limites de DCM puisqu'elle utilise plusieurs espaces de représentation mais ne traite pas les problèmes avec un seul texte connu. Une seconde extension, DCM-classifier basée sur les arbres de décision, remédie aux deux limites de DCM tout en conservant le principe.

Evaluée dans le cadre du challenge PAN-CLEF 2014, la méthode DCM-classifier avec un score général de 70.7 % a obtenu la seconde meilleure performance tout en ayant un temps d'exécution relativement bas. Les résultats obtenus lors de l'évaluation sont cohérents avec ceux obtenus lors de l'apprentissage.

Enfin, un des autres avantages de cette approche est de confirmer l'intérêt de combiner plusieurs espaces de représentation et d'adopter une méthode permettant à travers une phase d'apprentissage, de sélectionner ceux qui s'avèrent les plus discriminants pour le corpus considéré.

Références

Abbasi, A. et H. Chen (2005). Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE 20*(5), 67–75.

Cerra, D., M. Datcu, et P. Reinartz (2014). Authorship analysis based on data compression. *Pattern Recognition Letters 42*(0), 79 – 84.

Chaurasia, M. et D. S. Kumar (2010). Natural language processing based information retrieval for the purpose of author identification. *International Journal of Information Technology and Management Information Systems (IJITMIS) 1*(1), 45–54.

Dam, M. V. (2013). A basic character n-gram approach to authorship verification. In *Notebook for PAN at CLEF 2013*.

Davis, J. et M. Goadrich (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, New York, NY, USA, pp. 233–240. ACM.

de Vel, O., A. Anderson, M. Corney, et G. Mohay (2001). Mining e-mail content for author identification forensics. *SIGMOD Rec. 30*(4), 55–64.

Foster, D. W. (1996). A funeral elegy : William Shakespeare's "Best-Speaking Witnesses". *Publications of the Modern Language Association of America*, 1080–1105.

Ghaeini, M. R. (2013). Intrinsic Author Identification Using Modified Weighted KNN. In *Notebook for PAN at CLEF 2013*.

Halvani, O., M. Steinebach, et R. Zimmermann (2013). Authorship Verification via k-Nearest Neighbor Estimation Notebook for PAN at CLEF 2013. In *Notebook for PAN at CLEF 2013*.

Inches, G. et F. Crestani (2012). Overview of the international sexual predator identification competition at pan-2012. In *CLEF (Online Working Notes/Labs/Workshop)*, Volume 30.

Inches, G. et F. Crestani (2013). Overview of the international sexual predator identification competition at pan-2012. In *Proceedings of PAN at CLEF 2012*.

Jankowska, M. et E. Milios (2013). Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task Notebook for PAN at CLEF 2013. In *Notebook for PAN at CLEF 2013*.

Juola, P. et E. Stamatos (2013). Overview of the author identification task at pan 2013. In P. Forner, R. Navigli, et D. Tufis (Eds.), *Information Access Evaluation. Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, CLEF*, pp. 23–26.

Layton, R., P. Watters, et R. Dazeley (2013). Local n-grams for author identification notebook for pan at clef 2013. In *Notebook for PAN at CLEF 2013*.

Rangel, F., P. Rosso, M. Koppel, E. Stamatatos, et G. Inches (2013). Overview of the author profiling task at pan 2013. *Notebook Papers of CLEF*, 23–26.

Salton, G., A. Wong, et C.-S. Yang (1975). A vector space model for automatic indexing. *Communications of the ACM 18*(11), 613–620.

Seidman, S. (2013). Authorship Verification Using the Impostors Method. In *Notebook for PAN at CLEF 2013*, pp. 13–16.

Stamatatos, E., W. Daelemans, B. Verhoeven, M. Potthast, B. Stein, P. Juola, M. A. Sanchez-Perez, et A. Barrón-Cedeño (2014). Overview of the author identification task at pan 2014. In *Proceedings of CLEF PAN 2014*.

Szymanski, B. et Y. Zhang (2004). Recursive data mining for masquerade detection and author identification. In *Information Assurance Workshop, 2004. Proceedings from the Fifth Annual IEEE SMC*, pp. 424–431.

Vilariño, D., D. Pinto, H. Gómez, S. León, et E. Castillo (2013). Lexical-syntactic and graph-based features for authorship verification. In *Notebook for PAN at CLEF 2013*.

Zhang, C., X. Wu, Z. Niu, et W. Ding (2014). Authorship identification from unstructured texts. *Knowledge-Based Systems 66*(0), 99 – 111.

Summary

The problem of author identification is the following: for a given set of documents having the same author and a new document, we have to decide if this one was written or not by the author of the whole set. For solving this problem we have suggested and implemented various approches : similarity counting, vote technique and supervised learning which explore various model of document representations. The experiences are done using the collections of PAN-CLEF 2014 challenge and confirmed the interest of our approches and their performance in terms of time execution.