Attribution d'auteur : Une approche basée sur l'allocation latente de Dirichlet (LDA)

Jacques Savoy

Institut d'informatique
Université de Neuchâtel – rue Emile Argand 11 - 2000 Neuchâtel - Suisse

Abstract

This paper describes and evaluates the use of *Latent Dirichlet Allocation* (LDA) as an new approach to authorship attribution. Based on this generative probabilistic model, each document is represented by a mixture of topic distributions with each topic specifying a given distribution over words. Based on author profiles (aggregation of all texts written by the same writer), we then propose computing a distance with a disputed text to determine its likely author. The smallest distance will define the most probable writer. To evaluate this approach together with three other attributions schemes, we develop an experiment based on 4,326 newspaper articles (*La Stampa*) written in Italian by twenty distinct columnists. This research demonstrates that the LDA-based classification scheme tends, under certain conditions, to perform better than the Delta rule, the χ^2 distance or the Kullbach-Leibler divergence (KLD) scheme. The computational cost however tends to penalize LDA method compared to other algorithms.

Keywords: Text categorization, authorship attribution, lexical statistics, latent Dirichlet allocation.

Résumé

Cette communication décrit et évalue l'emploi d'une nouvelle approche basée sur l'allocation latente de Dirichlet (*Latent Dirichlet Allocation*, LDA) en attribution d'auteur. A l'aide de ce modèle probabiliste, chaque document se représente comme un mélange de thèmes correspondant pour chacun d'eux à une distribution spécifique de mots. Sur cette base, nous proposons de calculer une distance entre un texte dont l'auteur est inconnu et les divers profils d'auteur (agrégation de tous les écrits d'un même écrivain). La distance minimale nous permettra de déterminer l'auteur probable. Afin d'évaluer cette solution et de la comparer avec trois autres stratégies d'attribution d'auteur, nous avons créé une collection-test composée de 4 326 articles écrits par vingt journalistes du journal *La Stampa*. Cette étude comparative démontre qu'une approche basée sur la LDA offre, sous certaines conditions, une qualité d'affectation supérieure à la règle Delta, à l'usage de la distance du χ^2 ou à une technique basée sur la mesure de divergence Kullbach-Leibler (KLD). Le temps de traitement pénalise toutefois la technique LDA en comparaison aux autres approches.

Mots-clés : Catégorisation de textes, attribution d'auteur, statistique lexicale, allocation latente de Dirichlet.

1. Introduction

La détermination du véritable auteur d'un écrit (œuvre littéraire, article de presse, lettre, courriel) a donné lieu à de nombreuses études au cours de ces deux dernières décennies (Juola, 2006).

Contrairement à d'autres domaines, l'attribution d'auteur dispose d'une longue tradition (Love, 2002), bien antérieure aux premières études recourant à la statistique (Mosteller & Wallace, 1964). Ainsi, dès la fin de l'Antiquité on s'est interrogé sur les épîtres de St Paul, en estimant que cet ensemble n'a pas forcément été écrit par le même auteur. Dans la littérature française, la dispute entre Molière et Corneille concerne plusieurs pièces de théâtre (Labbé, 2009), (Marusenko & Rodionova, 2010). La littérature anglaise connaît plusieurs débats concernant la paternité des parties de l'œuvre de Shakespeare (Craig & Kinney, 2009). La question de l'attribution d'auteur connaît de nouveaux prolongements comme la *vérification* d'auteur. Dans ce cas, on souhaite savoir si un texte a été écrit ou non par un auteur donné. De plus, au lieu d'obtenir le nom probable de l'auteur, on peut se limiter à déterminer des informations socio-économiques le concernant (*profilage*) comme le sexe, l'âge, la nationalité, le niveau d'éducation, etc. (Argamon *et al.*, 2009). Enfin, l'attribution d'auteur fait également partie des sciences forensiques, de débats légaux, mais surtout d'un intérêt grandissant sur le Web avec, comme variantes, l'analyse de la crédibilité des auteurs (blogs, Twitter), voire la détection de plagiat.

Dans notre approche, le système sélectionnera automatiquement l'auteur jugé le plus probable en fonction de la représentation du document, du profil des auteurs potentiels et du classifieur employé. Afin d'atteindre cet objectif, nous proposons de représenter les documents comme un mélange de thèmes, avec chaque thème correspondant à une distribution particulière de mots (*Latent Dirichlet Allocation* ou LDA) (Blei *et al.*, 2003), (Steyvers & Griffiths, 2007), (Blei & Lafferty, 2009). Basé sur cette représentation, nous proposons de définir une distance entre deux distributions de thèmes, l'une représentant le profil d'un auteur, l'autre le texte requête. La distance minimale indiquera l'auteur le plus probable.

Dans la suite de cet article, la deuxième section présente les principales méthodes proposées pour déterminer l'auteur d'un écrit. La troisième section expose les grandes lignes de notre corpus d'évaluation. La quatrième section décrit brièvement trois méthodes connues en attribution d'auteur ainsi que notre modèle basé sur le paradigme LDA. La cinquième section évalue et compare ces diverses approches.

2. État des connaissances

Toute solution en attribution d'auteur repose sur une représentation des documents et un modèle ou règle de catégorisation (Juola, 2006). Afin de représenter un texte, les études quantitatives précédentes ont cherché à définir une mesure stylométrique unique devant être constante pour un auteur donné et différente d'un écrivain à l'autre (Holmes, 1998). Dans cette perspective, on a proposé de tenir compte de la longueur moyenne des mots ou des phrases, du nombre moyen de syllabes par mots, voire de la taille du vocabulaire V (notée |V|) par rapport à la longueur du document (Grieve, 2007). Comme alternative plus sophistiquée, on a également suggéré de calculer la valeur $R = |V| / sqrt(n)$ (avec n la taille du corpus), le rapport entre le nombre de *hapax legomena* ou le nombre de *dislegomena* (défini comme le nombre de mots apparaissant deux fois) et la taille du vocabulaire (Sichel, 1975). Toutefois, ces mesures ont l'inconvénient d'être instables (Baayen, 2008), (Hoover, 2003) en particulier face à des documents relativement courts (moins de 1 000 mots). De plus, on doit reconnaître que le genre (poésie, pièce de théâtre, roman, texte en vers ou en prose) influence de telles mesures, de même que, dans une moindre

ampleur, la chronologie (le style d'un auteur pouvant se transformer avec les années (Hoover, 2007)).

Comme deuxième stratégie de représentation, on a recouru à l'analyse du vocabulaire comme le démontre l'étude de Mosteller & Wallace (1964) marquant le début du recours à des méthodes statistiques dans l'attribution d'auteur. Dans ce cas précis, les méthodes privilégient l'analyse des mots fréquemment utilisés. On admet que ces vocables ne sont pas tous sous le contrôle conscient de l'auteur et que leurs fréquences varient d'un écrivain à l'autre. A l'aide de telles représentations, la différence entre auteurs peut se mesurer grâce à la distance du χ^2 (Grieve, 2007).

Comme troisième paradigme, on peut se limiter à une partie du vocabulaire et ne considérer que les mots très fréquents pouvant mieux refléter le style d'un auteur et demeurant plus indépendants des thèmes traités. Un tel ensemble peut se définir comme les *m* vocables les plus fréquents d'un corpus (Burrows, 2002).

Sur la base de ces représentations, on peut visualiser les similitudes et différences entre auteurs par des outils de statistiques multivariées comme l'analyse en composantes principales (Binonga & Smith, 1999) ou l'analyse des correspondances (Dixon & Mannion, 1993). D'autres approches ont également été proposées comme la classification automatique (*clustering*) (Hoover, 2007), (Labbé, 2007), parfois en complément à l'ACP (Holmes, 1992). Dans ces études, la distinction entre auteurs n'est pas parfaite et les attributions peuvent s'avérer difficiles, en particulier face à des auteurs ayant une culture commune (e.g., Goldsmith, Kelly & Murphy et leurs racines anglo-irlandaises (Dixon & Mannion, 1993)) ou lorsque le pouvoir discriminant s'avère faible (e.g., l'emploi de la fréquence des lettres pour distinguer entre les pièces de Shakespeare, Fletcher ou Dekker (Ledger & Merriam, 1992)).

Le recours à des méthodes tirées de l'apprentissage automatique par machine (*machine learning*) (Witten & Franck, 2005) ouvre une quatrième voie (Stamatatos, 2009). Par exemple, sur la base des 365 vocables parmi les plus fréquents de la langue anglaise, Zhao & Zobel (2005) cherchent à déterminer l'auteur d'articles de presse. Comme stratégie de catégorisation, ils démontrent qu'une approche Naïve Bayes tend à fournir une meilleure performance que les arbres de décision. Zheng *et al.* (2006) étudient l'application d'arbres de décision, de réseaux neuronaux et de machines à vecteurs de support (SVM) pour l'attribution de courts messages électroniques écrits en langue anglaise ou chinoise. Dans ces expériences, non seulement les éléments lexicaux sont pris en compte, mais les auteurs démontrent l'efficacité marginale de l'adjonction d'éléments syntaxiques, de contenu ou de présentation. Finalement, Jockers & Witten (2010) ont démontré que la règle Delta (Burrows, 2002) pouvait offrir des performances supérieures aux approches basées sur des SVM.

3. Corpus d'évaluation

Grâce à des collections-tests, nous pouvons évaluer et comparer divers modèles et représentations. Contrairement à la catégorisation automatique, les études en attribution d'auteur disposent d'un nombre restreint de corpus. De plus, les collections disponibles sont souvent écrites en langue anglaise et comprennent un nombre restreint d'auteurs potentiels et de documents (par exemple, les *Federalist Papers* (Mosteller, & Wallace, 1964) comprennent 85 articles et la paternité de 12 d'entre eux demeure incertaine (on hésite essentiellement entre deux auteurs possibles)).

Désirant fonder nos conclusions sur une base plus large et au moyen d'une collection stable et facilement accessible, nous avons sélectionné un sous-ensemble de la collection CLEF- 2003 (Peters *et al.*, 2004) disponible auprès de ELRA (www.elra.info). Cette partie comprend les articles écrits en italien et publiés durant l'année 1995 dans le journal *La Stampa*. Si le corpus complet compte 58 051 documents, nous connaissons le ou les auteur(s) que pour 37 682 d'entre eux. De ce dernier sous-ensemble, nous avons sélectionné les articles rédigés par un seul auteur et écarté les journalistes ayant écrit peu d'articles durant l'année 1995. Finalement, nous avons obtenu un corpus de 4 326 articles écrits par vingt auteurs différents.

Dans le tableau 1 nous avons indiqué le nom des journalistes, le thème principal correspond à chaque auteur, puis le nombre d'articles qu'il / elle a rédigé. On constate que le nombre d'articles par journaliste varie fortement entre le minimum de 52 (F. Nirenstein) et le maximum de 434 (O. Del Buono). En dernière colonne, nous avons indiqué la longueur moyenne (en nombre de mots) des articles rédigés, subdivisés par auteur. Globalement, la taille moyenne d'un article de *La Stampa* s'élève à 777 mots (minimum : 60 ; maximum : 2 935, médiane : 721, écart-type : 333). Si on analyse cette moyenne par journaliste, on constate que cette valeur varie fortement entre auteurs, avec une valeur moyenne minimale de 612 (A. Conti) jusqu'à un maximum de 1 478 (B. Spinelli).

	Nom du journaliste	Thème	Nombre d'articles	Longueur moyenne
1	Ansaldo Marco	Sports	287	812
2	Battista Pierluigi	Politique	231	840
3	Beccantini Roberto	Sports	364	831
4	Beccaria Gabriele	Social	71	686
5	Benedetto Enrico	Politique	252	732
6	Del Buono Oreste	Sports	434	799
7	Comazzi Alessandra	Social	223	616
8	Conti Angelo	Social	198	612
9	Gavano Fabio	Politique	347	738
10	Gramellini Massimo	Politique	118	955
11	Meli Maria Teresa	Politique	215	857
12	Miretti Stefania	Social	63	793
13	Nirenstein Fiama	Politique	52	1 090
14	Novazio Emanuele	Politique	249	750
15	Ormezzano Gian Paolo	Sports	232	738
16	Pantarelli Franco	Politique	202	692
17	Passarini Paolo	Politique	303	720
18	Sacchi Valeria	Business	203	776
19	Spinelli Barbara	Politique	57	1 478
20	Torabuoni Lietta	Social	225	784

***Tableau 1** : Répartition des articles sélectionnés par journaliste (*La Stampa*, 4 326 articles)*

Dans tous les articles, nous avons évidemment supprimé le nom de l'auteur, de même que certaines expressions récurrentes comme *Dal nostra* (ou *nostra*) *corrispondente*, *nostro*

servizio, etc. Afin de former un profil d'auteur, nous avons concaténé tous les articles écrits par le même journaliste.

Afin d'obtenir une évaluation non biaisée, le document-requête (celui dont on cherche à déterminer l'auteur) ne devra jamais être pris en compte lors de l'apprentissage. Dans nos évaluations, nous avons choisi la stratégie *leaving-one out*. Dans ce cadre, chaque article de presse parmi les 4 326 à notre disposition sera utilisé, à tour de rôle, comme document-requête. Si, par exemple, nous avons un article de B. Spinelli comme requête, le profil auteur de cette journaliste sera formé uniquement des 56 autres articles qu'elle a écrit.

4. Modèles d'attribution d'auteur

Afin de concevoir un système automatique d'attribution d'auteur, nous devons définir une représentation des textes et un modèle de catégorisation. La section 4.1 décrit la règle Delta (Burrows, 2002) s'appuyant sur les *m* vocables les plus fréquents d'un corpus. La section 4.2 expose une stratégie qui sélectionne les termes à retenir selon une fréquence documentaire minimale par auteur et qui s'appuie sur la métrique du χ^2 (Grieve, 2007). Dans la section 4.3 nous aborderons le recours au calcul de la divergence Kullbach-Leibler (KLD) sur la base d'un ensemble restreint de vocables définis *a priori* (Zhao & Zobel, 2007). La section 4.4 présente notre nouveau modèle d'attribution basé sur une allocation latente de Dirichlet (LDA) (Blei *et al.*, 2004).

4.1. La règle Delta

Afin de déterminer l'auteur probable d'un écrit, Burrows (2002) propose de tenir compte des vocables les plus fréquents et, en particulier, des mots fonctionnels, tout en ignorant les signes de ponctuation. Afin de représenter chaque document ou profil d'auteur, Burrows (2002) tient compte de 40 à 150 termes les plus fréquents, cette dernière valeur donnant souvent les meilleures performances. Pour être précis, Burrows distingue entre les homographes comme, par exemple, *to* comme préposition (e.g., *to you*) ou partie de l'infinitif (e.g., *to be*). Une telle distinction complexifie le traitement des documents sans générer une amélioration significative des résultats (Hoover, 2004). De même, la limite de 150 termes peut être repoussée, et Hoover (2007) suggère de considérer jusqu'à 4 000 mots.

Dans cette méthode, la comparaison de deux textes ne s'appuie pas directement sur les fréquences relatives d'occurrence mais sur des fréquences standardisées (score Z). Comme l'indique l'équation 1, une telle valeur est obtenue depuis la fréquence relative (notée tfr_{ij} pour le terme t_i dans le document D_j) par soustraction de la moyenne (notée $mean_i$) et division par l'écart-type (sd_i), moyenne et écart-type estimés en considérant le corpus sous-jacent (Hoover, 2004).

$$Z\ score(t_{ij}) = \frac{tfr_{ij} - mean_i}{sd_i} \quad (1)$$

Cette valeur est associée à chaque vocable retenu pour chaque document ou profil d'auteur. A l'aide de ces valeurs, on peut calculer la distance entre un document requête noté Q (dont on cherche à déterminer l'auteur) et les divers profils d'auteurs, profil-type noté A_j. La distance Delta Δ se calcule selon la formule 2.

$$\Delta(Q, A_j) = 1/m \cdot \sum_{i=1}^{m} \left| Z\ score(t_{iq}) - Z\ score(t_{ij}) \right| \quad (2)$$

Dans cette formulation, nous attachons la même importance à chaque terme t_i. Une différence importante entre Q et A_j apparaît lorsque, pour un vocable donné, les deux scores Z sont élevés et de signe opposé. Dans ce cas, l'un des journalistes a tendance à employer ce terme de manière plus fréquente que la moyenne tandis que l'autre tend à le négliger. A l'inverse, si le terme est usité avec la même fréquence relative dans les deux textes, la différence des scores Z sera faible, indiquant un rapprochement possible des deux textes. Finalement, si pour les *m* termes retenus les différences entre les scores Z demeurent faibles, la distance Δ résultante sera minime, indiquant que les deux textes sont probablement écrits par la même personne.

4.2. La distance du chi-carré

Comme deuxième modèle d'attribution d'auteur, nous avons repris l'une des meilleures solutions empiriques testées par Grieve (2007). Dans ce cas, la meilleure représentation des documents se basait sur la fréquence relative des vocables comprenant également huit signes de ponctuation (. , : ; - ? ('). Afin de réduire l'univers lexical, Grieve (2007) impose comme critère de sélection une *k*-limite dans laquelle l'entier *k* indique le nombre minimum d'articles par auteur dans lesquels le terme doit apparaître. Ainsi, le critère 5-limite implique que les termes retenus doivent apparaître dans au moins cinq articles écrits par chacun des auteurs. Lorsque l'on augmente la valeur de ce paramètre *k*, on réduit le nombre de termes considérés. Selon les statistiques de notre corpus présenté dans le tableau 1, nous pourrions augmenter la valeur de *k* jusqu'à 52, soit le plus petit nombre d'articles rédigés par un journaliste (F. Nirenstein). Avec ce critère très restrictif, nous devons retenir seulement 20 termes à savoir {*a al che da del della di e è i il in l la non per un . , '* }. Selon l'étude de Grieve (2007), les meilleures performances s'obtiennent avec une valeur faible de *k*, soit 2, 5 ou 10. Avec notre corpus de *La Stampa*, le critère 2-limite génère 720 termes pour représenter chaque texte.

Afin de définir une distance entre un document requête Q et un profil d'auteur A_j, Grieve (2007) s'appuye sur la distance du χ^2 définie par l'équation 3. Dans cette formulation, $qr(t_i)$ indique la fréquence relative du $i^{ème}$ terme dans la requête, et $ar_j(t_i)$ la fréquence relative du même vocable dans le $j^{ème}$ profil d'auteur.

$$\chi^2(Q, A_j) = \sum_{i=1}^{m} \left(qr(t_i) - ar_j(t_i) \right)^2 / ar_j(t_i) \quad (3)$$

Afin de déterminer l'auteur probable d'un écrit, nous sélectionnons celui présentant la distance du χ^2 la plus faible. Finalement, nous limiterons notre critère de sélection à 2-limite car ce choix assure un plus grand nombre de vocables, redonne la meilleure performance et facilite le calcul de l'équation 3. En effet, cette limite impose que chaque terme apparaisse au moins dans deux textes écrits par chacun des auteurs. Comme le profil d'auteur n'inclut pas le texte requête, la fréquence relative $ar_j(t_i)$ ne sera jamais nulle, et la formule 3 pourra toujours être calculée.

4.3. La divergence Kullbach-Leibler (KLD)

Zhao & Zobel (2007) proposent de définir *a priori* une liste restreinte de vocables permettant de discriminer le style des divers auteurs. Pour la langue anglaise, leur liste comprend 363 termes

comprenant principalement des mots fonctionnels (e.g., *the*, *in*, *but*, *not*, *am*, *of*, *can*, …), de même que certaines formes assez fréquentes (e.g., *became*, *nothing*). Pour la langue italienne usitée dans notre corpus, nous avons repris une liste de 399 mots-outils employés par un moteur de recherche ayant obtenu d'excellents résultats dans cette langue (Savoy, 2001). La sélection des termes étant effectuée, la probabilité d'occurrence de chaque terme et pour chaque document (ou profil d'auteur) doit encore être estimée.

Comme première estimation, nous pouvons recourir au principe du maximum de vraisemblance et estimer la probabilité d'occurrence du terme t_i par sa fréquence relative, soit tfa_i / n. Dans cette formulation, tfa_i indique la fréquence absolue du terme t_i dans un document ayant une taille n (en nombre de mots). Cette première solution a tendance à surestimer la probabilité des termes apparaissant au détriment des mots n'ayant pas encore été rencontrés. Comme la fréquence d'occurrence de ces derniers est nulle, leur probabilité sera de même. Or, la distribution des mots suit une loi de type LNRE (*Large Number of Rare Events* (Baayen, 2008)), avec l'apparition constante de nouveaux termes.

Afin de tenir compte de ce phénomène récurrent, nous devons lisser ces estimations. Des expériences antérieures (Savoy, 2010) ont démontré que l'application de la technique de Lidstone (Manning & Schütze, 2000) permettait d'obtenir de bons résultats, tout en étant simple à implémenter. Dans ce cas, l'estimation précédente est remplacée par $(tfa_i+\lambda) / (n+\lambda \cdot |V|)$, avec $|V|$ indiquant la taille du vocabulaire. Sans autre information a priori, nous proposons de fixer λ à 0,01, un choix certes arbitraire mais qui attribue une faible probabilité aux mots rares, ces derniers ne devant pas avoir une importance majeure en attribution d'auteur.

Basé sur ces estimations, nous devons calculer le degré de divergence entre deux distributions discrètes. Dans ce but, Zhao & Zobel (2007) suggèrent de recourir à la divergence de Kullbach-Leibler (KLD), aussi connue sous le nom d'entropie relative (Manning & Schütze, 2000). Cette mesure est indiquée dans l'équation 4 dans laquelle la distribution obtenue à partir du document requête est notée Q et celle du profil du j^e auteur par A_j.

$$KLD(Q \parallel A_j) = \sum_{i=1}^{m} q(t_i) \cdot \log_2 \left[\frac{q(t_i)}{a_j(t_i)} \right] \quad (4)$$

dans laquelle $q(t_i)$ et $a_j(t_i)$ représentent les probabilités d'occurrence du terme t_i dans le document requête Q, respectivement dans j^e profil d'auteur A_j. Si deux distributions sont identiques, la valeur retournée sera nulle. Dans tous les autres cas, cette valeur sera positive. Comme règle d'attribution, nous assignons l'auteur ayant le profil retournant la valeur la plus faible.

4.4. Allocation latente de Dirichlet (Latent Dirichlet Allocation, LDA)

La technique d'allocation latente de Dirichlet (Blei *et al.*, 2003), (Blei & Lafferty, 2009) propose un modèle probabiliste de génération de documents abordant plusieurs sujets. Dans ce cadre, chaque document d'un corpus donné se modélise comme une distribution de différents thèmes, avec chaque thème représentant une distribution spécifique de mots (l'ordre de ces derniers n'ayant pas d'importance, l'hypothèse du *sac de mots* est donc admise). Certes un texte peut couvrir un seul thème, mais ceci constitue l'exception et non la norme.

Par exemple, le document D_1 peut traiter essentiellement du premier thème, un peu du second et marginalement du troisième (et ignorer tous les autres). Un mélange équitable des deux

premiers thèmes se retrouve dans le texte D_2, et ainsi de suite. Chaque vocable peut apparaître sous plusieurs thématiques afin d'indiquer soit sa polysémie (comme, par exemple, *Jaguar* comme un animal, une voiture ou un logiciel), soit le fait qu'il corresponde à un mot fonctionnel présent dans presque tous les documents.

Thème 1	Thème 2	Thème 6	Thème 12
milan	pds	borsa	baggio
capello	alema	lira	sacchi
gullit	occhetto	dollaro	arrigo
savicevic	segretario	mercati	italia
gol	leader	mercato	signori
desailly	sinistra	tassi	roberto
lentini	veltroni	marco	mondiale
baresi	quercia	affari	codino

Tableau 2 : Vocables les plus probables selon quatre thèmes parmi les vingt extraits du corpus La Stampa, après élimination des mots fonctionnels

En appliquant cette approche non supervisée au journal *La Stampa*, quelques exemples de thèmes accompagnés de leurs mots les plus probables sont repris dans le tableau 2. Ainsi, le football tient une place non négligeable avec l'AC Milan, ses vedettes (R. Gullit, D. Savicevic, M. Desailly, G. Lentini, F. Baresi) et son entraîneur (F. Capello). L'équipe nationale se retrouve sous le douzième thème (Arrigo Sacchi, Roberto Baggio (*il Divin Codino*), G. Signori). La politique apparaît clairement sous le deuxième thème avec le parti PDS (*Partito Democratico della Sinistra*) et ses personnalités, comme M. D'Alema, le successeur de A. Occhetto, ou W. Veltroni. Le sixième sujet aborde le monde de la finance avec des vocables liés aux marchés, à la lire, au dollar, aux taux, et aux personnes de ce milieu comme Marco Biagi.

Habituellement, nous sommes plus intéressés par le problème dual (inférence statistique). Etant donné un nombre fixé de *k* thèmes, et la fréquence observée des *m* termes t_i dans les *n* document D_j, on doit déterminer la distribution la plus probable des thèmes par rapport aux documents, et des vocables par rapport aux thèmes.

Basé sur ce modèle, (Blei *et al.*, 2003) et (Steyvers & Griffiths, 2007) proposent un survol général des différentes techniques d'estimations et des applications qui ont été proposées. En particulier, Rosen-Zvi *et al.* (2004) et Griffiths *et al.* (2004) indiquent comment on peut inclure le nom des auteurs au moyen d'une distribution sur les thèmes retenus (sous l'hypothèse que les auteurs écrivent régulièrement sur les mêmes sujets). L'objectif visé est de connaître l'évolution thématique d'un chercheur, d'un groupe d'auteurs, ou les relations qui s'établissent entre scientifiques en fonction des thèmes communs abordés ou publications rédigées conjointement. L'attribution d'auteur n'est mentionnée que comme développement futur possible.

Pour résoudre cette question à l'aide d'un modèle LDA, nous avons retenu une implémentation écrite en C et disponible gratuitement (écrit par D.M. Blei). A l'aide de ce logiciel, nous donnons en entrée les vingt documents (profil d'auteur) représentés par la fréquence absolue des divers vocables retenus. En sortie, le système retourne la distribution des *m* termes en fonction des *k* thèmes, ainsi que la distribution des *k* thèmes sur les *n* (= 20) documents. Cette première phase

correspond à l'apprentissage ou à l'estimation du modèle. Cette estimation requiert que l'on spécifie le nombre *k* de thèmes. Dans notre exemple, nous pouvons faire l'hypothèse qu'un thème correspond au style particulier d'un auteur. La valeur du paramètre *k* devrait donc être proche de 20.

A l'aide de ces deux ensembles de distributions, le système peut ensuite estimer la distribution des thèmes pour un nouveau document (le texte requête dans notre cas). Enfin, pour déterminer l'auteur probable de ce nouvel article, nous calculons la distance entre la distribution thématique Q du document requête et ceux des 20 profils d'auteur A_j selon la formule suivante :

$$sKLD(Q \parallel A_j) = \frac{1}{2}\left(\sum_{i=1}^{m} q(t_i) \cdot \log_2 \left[\frac{q(t_i)}{a_j(t_i)} \right] + \sum_{i=1}^{m} a_j(t_i) \cdot \log_2 \left[\frac{a_j(t_i)}{q(t_i)} \right] \right) \quad (5)$$

Cette équation correspond à la distance symétrique de Kullbach-Leubler (Abdi, 2007) entre deux distributions discrètes.

Contrairement aux trois autres stratégies d'attribution, notre application basée sur le LDA ne spécifie pas explicitement une phase de sélection les vocables à retenir. Nous avons donc décidé de reprendre les 720 termes sélectionnés par la contrainte 2-limite (section 4.2) et, dans une seconde évaluation, la liste des 399 mots-outils utilisés avec le calcul KLD (section 4.3). Ces deux solutions reposent sur des critères assez similaires et les deux listes comprennent essentiellement des mots fonctionnels.

Comme alternative, nous pouvons également inclure plus de termes et considérer l'ensemble du vocabulaire de *La Stampa* (soit 106 680 vocables). Il s'avère plus économique et plus efficient de réduire cet espace lexical sachant qu'environ 50 % des vocables n'apparaissent qu'une ou deux fois. Habituellement, au lieu de considérer la fréquence d'occurrence, on préfère recourir à la fréquence documentaire (notée *df*) (nombre de documents dans lesquels apparaît un vocable). Ainsi, en considérant uniquement les termes présents dans trois articles ou plus, notre vocabulaire se réduit à 36 928 vocables (une réduction d'environ 65,4 %). Signalons finalement que cette technique d'élagage est considérée comme l'une des plus efficientes dans le domaine de la catégorisation automatique (Yang & Pedersen, 1997).

"This suggests that DF (document frequency) thresholding, the simplest method with the lowest cost in computation, can be reliably used instead of IG (information gain) or CHI (c2-test)".

5. Evaluation

Afin de mesurer la performance d'un système de catégorisation automatique, nous pouvons calculer son taux d'exactitude (*accuracy rate*) selon deux modes. Premièrement, on peut juger que chaque décision d'attribution possède la même valeur. Selon cette optique, la performance moyenne se calcule sur la moyenne de toutes les décisions (*micro-moyenne*). Comme alternative, nous pouvons considérer qu'il faut faire une moyenne sur la performance obtenue pour chaque catégorie (*macro-moyenne*). Les deux mesures, fortement corrélées, seront reprises dans nos évaluations.

Avec la règle Delta (Borrows, 2002), nous devons représenter les textes à l'aide des *m* vocables les plus fréquents. Selon nos expériences, le meilleur taux de succès s'obtient avec la valeur *m* = 400, avec une légère baisse si l'on considère les valeurs 200 ou 600. Pour l'approche basée sur

la distance du χ^2, la sélection des termes s'effectue selon une fréquence documentaire minimale par auteur. Dans ce cas, la meilleure performance s'obtient avec le critère de 2-limite imposant que chaque terme retenu apparaisse au moins dans deux articles écrits par chaque journaliste. Comme troisième modèle de référence, nous avons retenu le système d'attribution KLD (Zhao & Zobel, 2007) basé sur une liste de termes défini *a priori* (soit 399 mots dans notre cas).

Méthode	Paramètre	Micro	Macro
Delta	400 termes	76.07%	75.08%
χ^2	2-limite, 720 termes	68.28%	65.79%
KLD	$\lambda = 0{,}01$, 399 termes	84.84%	82.84%
LDA	720 termes, $k = 20$	72.81%	73.79%
LDA	720 termes, $k = 40$	73.13%	74.80%
LDA	720 termes, $k = 50$	72.34%	74.25%
LDA	399 termes, $k = 20$	55.60%	56.52%
LDA	399 termes, $k = 40$	53.15%	55.79%
LDA	399 termes, $k = 50$	48.97%	52.57%
LDA	36 928 termes, $k = 20$	85.86%	82.11%
LDA	36 928 termes, $k = 40$	**89.39%**	**83.79%**
LDA	36 928 termes, $k = 50$	89.34%	83.85%

Tableau 3 *: Evaluation de quatre stratégies d'attribution sur notre corpus* La Stampa *(4 326 articles, 20 auteurs)*

Comme l'illustre le tableau 3, la méthode KLD offre la meilleure performance par rapport à la règle Delta ou à celle du χ^2. Si l'on analyse les résultats obtenus par l'approche LDA, on se rend compte de l'importance du choix d'une valeur appropriée pour le paramètre k (nombre de thèmes). En se basant sur les mêmes vocables que l'approche KLD (399 mots), les variations de performance sont importantes dans le modèle LDA en fonction de k. Fixer cette valeur à 20 correspond à un choix qui se justifie au niveau théorique (chaque auteur étant alors représenté par un thème, soit une distribution particulière de vocables).

Pour des représentations basées sur un nombre restreint de termes (399 ou 720), l'approche KLD s'avère la meilleure. De surcroît, si l'on considère un ensemble plus important de termes (36 928 mots ayant une fréquence documentaire strictement supérieure à deux), le modèle d'attribution fondé sur la LDA permet d'obtenir une plus grande performance.

Le temps de traitement doit aussi être pris en compte. Avec une version écrite en C, l'approche LDA requiert environ 2,3 minutes pour estimer les distributions lexicales et thématiques ($k = 20$). Ce délai est supérieur au temps moyen requis par les approches Delta, χ^2 ou KLD (environ une minute). De surcroît, si l'on accroît la taille du vocabulaire choisi (soit 36 928 mots), le temps de traitement nécessaire pour l'approche LDA s'élève, en moyenne, à environ 30 minutes (paramètre $k = 20$).

6. Conclusion

Dans cette communication, nous avons traité de l'attribution d'auteur dans laquelle le nom du véritable auteur était l'un des candidats connus. Comme stratégie, nous avons décrit et évalué la règle Delta (Burrows, 2002) s'appuyant sur les 400 vocables les plus fréquents d'un corpus. Comme deuxième modèle, nous avons repris la meilleure représentation analysée par Grieve (2007) dans une étude empirique regroupant un nombre conséquent d'approches. Cette solution, recourant à la métrique du χ^2, sélectionne les termes à retenir selon une fréquence documentaire minimale de deux par auteur (avec notre corpus, 720 vocables sont ainsi retenus). Comme troisième approche, nous avons repris l'idée de Zhao & Zobel (2007) définissant *a priori* une liste de termes pour la langue anglaise (liste que nous avons adaptée pour la langue italienne et qui contient 399 mots). La comparaison entre textes et profils d'auteur se base sur la divergence de Kullbach-Leibler (KLD).

Enfin nous avons présenté une nouvelle stratégie d'attribution d'auteur incorporant l'allocation latente de Dirichlet (LDA). Dans ce paradigme, nous fixons un nombre k de thèmes pouvant décrire le contenu d'un corpus. Un exemple illustre ces propos à l'aide des articles du journal *La Stampa*. Nous avons ensuite décrit comment adapter ce paradigme pour l'attribution d'auteur en comparant des distributions sur les thèmes, une pour le document requête, l'autre pour chaque profil d'auteur.

Afin de comparer ces quatre stratégies, nous avons recouru à un sous-ensemble de 4 326 articles de *La Stampa* écrits durant l'année 1995 par vingt journalistes différents. Cet ensemble, correspondant à une partie du corpus de la campagne CLEF-2003, est stable et facilement accessible. De plus, comme les articles sont écrits durant la même année, pour la même audience et par des auteurs qui partagent la même culture, cette collection présente un intérêt indéniable pour l'évaluation en attribution d'auteur.

L'évaluation comparative indique qu'une approche KLD permet d'obtenir des performances supérieures (84,8 %) à la règle Delta (76,1 %) ou à la métrique du χ^2 (68,3 %). En utilisant les mêmes termes et en fixant adéquatement le nombre k de thèmes, le paradigme LDA apporte un taux de réussite supérieur (73,1 % avec 720 termes, 55,6 % avec 399 mots) à la règle Delta et à la métrique du χ^2. Lorsque la représentation des textes se fonde sur un nombre restreint de termes (399 ou 720), le modèle KLD apporte la meilleure performance. Par contre notre modèle LDA permet une performance moyenne supérieure si l'on dispose d'un vocabulaire plus étendu. Toutefois, le temps de traitement requis par le LDA s'avère être un inconvénient majeur à cette solution.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n^0 #200021-129535/1). L'auteur remercie les relecteurs anonymes pour leurs remarques pertinentes ayant permis d'améliorer cette communication.

Références

Abdi H., (2007). Distance. In N. Salkind (Ed), *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks.

Argamon S., Koppel M., Pennebaker J.W. and Schler J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2), pp. 119-123.

Baayen H.R. (2001). *Word Frequency Distributions*. Kluwer Academic Press, Dordrecht.

Baayen H.R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Binonga J.N.G. and Smith M.W. (1999). The application of principal component analysis to stylometry. *Literary and Linguistic Computing*, 14(4), pp. 445-465.

Blei D.M., Ng A.Y. and Jordan M.I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, pp. 993-1022.

Blei D.M. and Lafferty J.D. (2009). Topic models. In A.N. Srivastava and M. Sahami (Eds). *Text Mining. Classification, Clustering, and Applications*. Chapman & Hall, Boca Raton, pp. 71-93.

Burrows J.F. (2002). Delta: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), pp. 267-287.

Craig H. and Kinney A.F. (Eds), (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, Cambridge.

Dixon P. and Mannion D. (1993). Goldsmith's periodical essays: A statistical analysis », *Literary and Linguistic Computing*, 8(1), pp. 1-19.

Grieve J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), pp. 251-270.

Griffiths T., Smyth P., Rosen-Zvi M. and Steyvers T. (2004). Probabilistic Author-Topic Models for Information Discovery. In *Proceedings ACM-KDD*, Seattle, pp. 306-315.

Holmes D.I. (1992). A stylometric analysis of Mormon scripture and related texts. *Journal of the Royal Statistical Society A*, 155(1), pp. 91-120

Holmes D.I., (1998). The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3), pp. 111-117.

Hoover D.L. (2003). Another perspective on vocabulary richness. *Computers and the Humanities*, 37, 151-178.

Hoover D.L. (2004). Delta Prime? *Literary and Linguistic Computing,* 19(4), pp. 477-495.

Hoover D.L. (2007). Corpus Stylistics, Stylometry, and the styles of Henry James. *Style*, 41(2), pp. 160-189.

Jockers M.L. and Witten D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2), pp. 215-223.

Juola P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3).

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1), pp. 33-80.

Labbé D. (2009). *Si deux et deux font quatre, Molière n'a pas écrit Dom Juan*. Max Milo, Paris.

Laroche A. (2010). Attribution d'auteur au moyen de modèles de langue et de modèles stylométriques. *Actes RECITAL*, 2010.

Ledger G. and Merriam R. (1994). Shakespeare, Fletcher, and the *Two Noble Kinsmen. Literary and Linguistic Computing*, 9(3), pp. 235-248.

Love H. (2002). *Attributing Authorship: An Introduction*. Cambridge University Press, Cambridge.

Marusenko M. and Rodionova E. (2010). Mathematical methods for attributing literary works when solving the "Corneille-Molière" problem. *Journal of Quantitative Linguistics*, 17(1), pp. 30-54.

Manning C.D. and Schütze H. (2000). *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge.

Mosteller F. and Wallace D.L. (1964). *Inference and Disputed Authorship, The Federalist*. Addison-Wesley, Reading.

Peters C., Braschler M., Gonzalo J. and Kluck M. (Eds). (2004). *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237, Springer, Berlin.

Rosen-Zvi M., Griffiths T., Steyvers M. and Smyth P. (2004). The author-topic model for authors and documents. In *Proceedings of the Uncertainty in Artificial Intelligence*. Banff, pp. 487-494.

Savoy J. (2001). Report on CLEF-2001 experiments. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds), *Cross-Language Information Retrieval and Evaluation*. Springer, Lectures Notes in Computer Science #2069, pp. 27-43

Savoy J. (2010). Discours électoral et discours présidentiel: Une étude lexicale comparative de B. Obama. Journées internationales d'Analyse statistique des Données Textuelles JADT 2010, Rome, pp. 827-838

Sichel H.S., (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70(351), pp. 542-547.

Stamatatos E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science & Technology*, 60(3), pp. 433-214.

Steyvers M. and Griffiths T. (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis and W. Kintsch (Eds). *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

Witten, I.H. and Franck, E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques*. Elsevier, Amsterdam.

Yang Y. and Pedersen J.O. (1997). A comparative study of feature selection in text categorization. In *Proceedings Conference on Machine Learning ICML*, pp. 412-420.

Zhao Y. and Zobel J. (2005). Effective and scalable authorship attribution using function words. In *Proceedings AIRS Asian Information Retrieval Symposium*, pp. 174-189.

Zhao Y. and Zobel J. (2007). Entropy-based authorship search in large document collection. In *Proceedings ECIR2007*, Springer LNCS #4425, pp. 381-392.

Zheng R., Li J., Chen H. and Huang Z. (2006). A framework for authorship identification of online messages: writing-style features and classification techniques. *Journal of the American Society for Information Science & Technology*, 57(3), pp. 378-393.