Une méthode hybride pour la prédiction du profil des auteurs

Seifeddine Mechti, Maher Jaoua, Lamia Hadrich Belguith

mechtiseif@gmail.com
maher.jaoua@fsegs.rnu.tn
l.belguith@fsegs.rnu.tn
Faculté des Sciences Economiques et de Gestion (FSEGS),
Laboratoire MIRACL, B.P 1088, 3018, Sfax, Tunisie

Résumé. Dans cet article, nous nous intéressons à la détection du profil des auteurs (âge, genre) à travers leurs discussions. La méthode proposée s'appuie sur la classification automatique qui utilise certaines données extraites d'une manière statistique à partir de corpus source. Nous présentons une méthode hybride qui combine l'analyse de surface dans les textes avec une méthode d'apprentissage automatique. A fin d'obtenir une meilleure gestion de ces données, nous nous sommes basés sur l'utilisation des arbres de décision. Notre méthode a donné des résultats intéressants pour la détection du genre.

1 Introduction

De nos jours, les réseaux sociaux ont connu une croissance importante. Sur Twitter ou sur Facebook la plus part des utilisateurs renseignent seulement 20% de leurs profils. La détection du profil peut être utilisée dans plusieurs domaines, par exemple du point de vue marketing, les entreprises peuvent être intéressées à déterminer quels types de personnes préfèrent leurs produits. Dans la littérature, beaucoup de travaux ont focalisé sur la classification d'une conversation ou d'un texte donné et plus précisément la détection de l'âge de l'auteur, de son genre, sa personnalité, sa langue native, etc. Argamon et al. (2009); Schler et al. (2006); Koppel et al. (2003); Pennebaker (2011).

Les travaux réalisés par Koppel et al. (2003) ont montré qu'au niveau du genre il y a des différences linguistiques entre les hommes et les femmes. En effet, les hommes qui préfèrent catégoriser les choses, utilisent plus de déterminants (le/la, cette/ce, un/une, etc.) et de quantificateurs (deux, plus, peu, etc.). Les femmes, s'intéressent aux relations et plus que les hommes recourent aux pronoms personnels (je, tu, moi, etc.).
La suite de ce papier est organisée comme suit, dans la section 2 nous présentons notre méthode d'apprentissage en focalisant sur le choix des classes et l'algorithme employé. La dernière section présente notre étude expérimentale.

2 Méthode proposée

La méthode proposée s'appuie sur la classifcation de discussions en fonction du genre et de l'âge en se basant sur les probabilités d'apparitions des mots. La dimension genre est repré-

sentée par la classe homme et la classe femme. Les classes d'âges ont été défnies selon Schler et al. Schler et al. (2006). La classe 10s présente les individus entre 13 et 17 ans, 20s ceux entre 20 et 33 ans et enfin 30s ceux entre 33 et 47ans. Nous avons commencé par calculer le nombre d'occurrences de tous les termes trouvés dans le corpus en les classant par ordre décroissant de leurs apparitions, néanmoins nous nous sommes contentés des 200 premiers attributs. Nous avons calculé CF(class frequency) pour chaque classe d'attributs dans la perspective de mesurer la fréquence d'apparition de chaque classe d'attributs dans chaque document du corpus. Les approches les plus répandues dans la littérature, distinguent deux principaux types d'attributs qu'on peut utiliser pour la détection du profil de l'auteur : les attributs stylistiques et les attributs basés sur le contenu Pennebaker (2011). Nous avons groupé manuellement les termes appartenant à la même classe d'attributs. nous avons déterminé 25 classes à savoir : Prepositions, Pronouns, Determiners, Adverbs, Verbs, He, She, No, Of, I, Me, Medecine, Chemistry, Music, Sport, Tv, Phone, Beer, Sleeping, Eating, Sex, Love, Money, Internet, Marketing. Nous avons utilisé les classes d'attributs purement stylistiques (basés sur le style), En outre, nous avons choisi d'utiliser trois attributs : les prépositions, les pronoms et les déterminants. Une fois les classes sont fixées, il s'agit d'effectuer l'apprentissage. Nous avons eu recours au logiciel libre d'apprentissage « Weka [1] ». Nous avons commencé par la construction des fchiers ARFF (Attribute Relation File Format), un fchier pour la dimension genre et un autre pour la dimension âge.

3 Expérimentation et évaluation

Nous avons utilisé les corpus discernés de la conférence CLEF [2] 2013. Nous avons effectué l'expérimentation avec un extrait du corpus d'entrainement. En faite, pour la dimension genre, qui a comme moyenne (baseline) de précision 0.5, nous avons obtenu de bons résultats, 58,16% des documents ont été bien classés. Pour la dimension âge qui a comme moyenne 33% les résultats sont prometteurs et témoignent de l'efficacité de la méthode. En effet, 57% des documents ont été bien classés. Comme le montre la figure 1, nous avons trouvé que la méthode d'apprentissage fondé sur les arbres de décision donne de meilleurs résultats.

3.1 Comparaison entre notre méthode et celle de Koppel et al.

Pour mieux cerner les points forts et les points faibles de notre méthode, une comparaison avec la méthode de koppel et al. a été effectuée. Toujours en utilisant le logiciel d'apprentissage Weka nous avons essayé plusieurs jeux de test avec l'algorithme Winnow (Multi-Class Real Winnow). La Figure 2 montre les résultats obtenus :

Nous constatons que koppel et al. ont eu les meilleurs résultats en utilisant un grand nombre d'attributs. Notre méthode qui se base seulement sur l'utilisation de données stylistiques, se voit moins couteuse en ressources puisque elle n'utilise pas le contenu, elle utilise seulement les 200 termes les mieux classés alors que l'autre méthode emploi 467 termes.

Dans un autre jeu de test, nous avons aussi comparé la précision obtenue en fonction du nombre d'attributs utilisés pour chaque méthode (voir Figure 3). Il est certain que les résultats

1. http ://www.cs. waikato.ac.nz / ml/weka
2. www.pan.webis.de

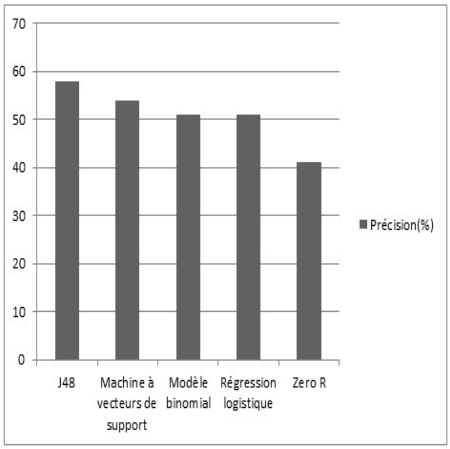

FIG. 1 – *Résultats des classifieurs utilisés*

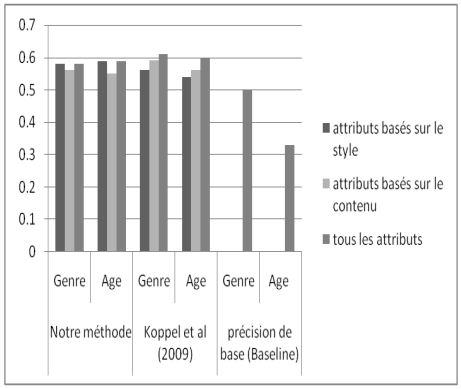

FIG. 2 – *Comparaison avec la méthode de Koppel et al.*

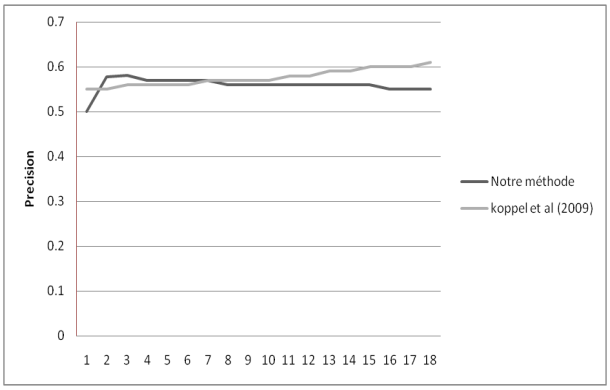

FIG. 3 – *Variation de la précision pour la dimension genre en fonction du nombre de classes*

trouvés par la méthode de Koppel et al. sont les meilleurs. Comme le montre la courbe, 58,16 % des documents ont été bien classés en utilisant trois attributs. Alors que, 61% des documents ont été bien classé pour la méthode de Koppel et al. En utilisant 18 attributs, ce qui témoigne de l'efficacité de notre algorithme et du choix minutieux des classes.

4 Conclusion

Nous avons effectué la catégorisation de documents en vue de fournir une classification de l'auteur d'un texte donné selon ses caractéristiques. Les résultats obtenus sont encourageants et surtout pour la dimension genre. La sélection manuelle du contenu des classes a montré ses limites face à des corpus de langues peu connues par le chercheur. L'automatisation de cette tâche s'avère d'une grande utilité, et l'utilisation de dictionnaires bilingues ou multilingues pourra faire face aux insuffisances linguistiques.

Il s'est avéré que l'utilisation des classes lexicales à elle seule n'est pas suffisante, cependant nous comptons intégrer d'autres aspects comme l'aspect syntaxique, morphologique, sémantique, etc. D'un autre coté, pour pouvoir mieux effectuer la détection du profil de l'auteur nous pensons s'ouvrir sur d'autres dimensions, à part l'âge et le genre nous allons aborder aussi la détection de la langue native, la détection des données géographiques de l'auteur et la détection du niveau linguistique, etc.

Références

Argamon, S., M. Koppel, J. Pennebaker, et J. Schler (2009). Automatically detection the author of an anonymous text. *Communications of the ACM*, 119–123.

Koppel, M., S. Argamon, et A. Shimoni (2003). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*.

Pennebaker, J. (2011). The secret life of pronouns: What our words say about us. pp. 401–412.

Schler, J., M. Koppel, S. Argamon, et J. Pennebaker (2006). Effects of age and gender on blogging. *AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Summary

In this paper, we present a method for profiling the author of an anonymous text. Our approach is based on learning the author profile with a focus on dimensions age and gender. First, we computed a ranked list of words that occur in the corpus and we grouped them into classes according to their similarities. Then, we calculated the CF (class frequency) score of each class for each document in order to find the stylistic differences between men and women, on the one hand, and those between different age intervals on the other hand. Our system has shown a high level of accuracy and effectiveness in treating the gender dimension.