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