Détection et regroupement automatique de style d'écriture dans un texte

Jérémy Ferrero*, Alain Simac-Lejeune**

Compilatio
276, rue du Mont-Blanc
74520 Saint-Félix, France
*jeremyf@compilatio.net
**alain@compilatio.net

Résumé. La détection de plagiat extrinsèque devient vite inefficace lorsque l'on n'a pas accès aux documents potentiellement sources du plagiat ou lorsque l'on se confronte à un espace aussi vaste que le Web, ce qui est souvent le cas dans les logiciels anti-plagiat actuels. Dès lors la détection intrinsèque devient nettement plus efficace. Dans cet article, nous traitons justement de la détection automatique d'auteurs qui permet de savoir si un passage d'un texte n'appartient pas au même auteur que le reste du texte et donc en théorie de repérer les passages plagiés d'un document. Nous expliquons notre contribution aux procédures déjà existantes et évaluons les limites de notre approche. L'objectif est de permettre la détection et le regroupement de passages d'un document par auteur.

1 Introduction

La plupart des logiciels anti-plagiat se concentrent sur une détection extrinsèque de plagiat, c'est-à-dire sur le fait de trouver des similitudes entre un document et un corpus de sources probables. Or ce système est inutile si le document ayant été plagié ne se trouve pas dans le corpus fouillé. Néanmoins, il existe un autre type de détection, la détection intrinsèque qui exploite des données extraites de l'intérieur même du document. La détection d'auteurs par étude du style d'écriture du document est la forme de détection de plagiat intrinsèque la plus répandue. Cette approche diverge selon les travaux car elle soulève plusieurs problèmes, allant du découpage du texte de façon pertinente, au choix et à la collecte des données stylistiques à surveiller, en passant par la manière de découper et de classer les différents passages du document par auteur. C'est sur ce dernier point que l'article va essentiellement se concentrer.

2 La détection d'auteur

2.1 La notion de stylométrie

La stylométrie ou l'étude stylométrique d'un texte est une analyse à mi chemin entre une analyse linguistique et statistique. Elle exploite des variables stylométriques, qui sont des