Détection automatique de reformulations - Correspondance de concepts appliquée à la détection du plagiat

Jérémy Ferrero*, Alain Simac-Lejeune**

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

Résumé. Dans le cadre de la détection du plagiat, la phase de comparaison de deux documents est souvent réduite à une comparaison mot à mot, une recherche de « copier/coller ». Dans cet article, nous proposons une approche naïve de comparaison de deux documents dans le but de détecter automatiquement aussi bien les phrases copiées de l'un des textes dans l'autre que les paraphrases et reformulations, ceci en se focalisant sur l'existence des mots porteurs de sens, ainsi que sur leurs mots de substitution possibles. Nous comparons trois algorithmes utilisant cette approche afin de déterminer la plus efficace pour ensuite l'évaluer face à des méthodes existantes. L'objectif est de permettre la détection des similitudes entre deux textes en utilisant uniquement des mots clefs. L'approche proposée permet de détecter des reformulations non paraphrastiques impossibles à détecter avec des approches conventionnelles faisant appel à une phase d'alignement.

1 Introduction

Actuellement, la recherche et la détection de similitudes s'effectuent en deux phases : une première phase de recherche de sources candidates, suivie d'une seconde de comparaison de ces sources possibles avec le document que l'on suspecte d'être un plagiat. La phase de collecte est de plus en plus optimale grâce à l'amélioration de l'efficacité des moteurs de recherche en local et sur le Web. C'est à la seconde phase que cet article s'intéresse. Une fois qu'une source candidate est trouvée, elle doit être comparée avec le document sur lequel pèse les soupçons. À l'heure actuelle, la plupart des logiciels anti-plagiat, une fois une liste de sources candidates constituée, se contentent de comparer mot à mot le document analysé avec chaque source possible. Cette technique permet seulement de détecter les similitudes de types « copier/coller ». Bien que cette approche ait prouvé son efficacité et suffise la plupart du temps, en France près d'un étudiant sur deux a déjà eu recours au « copier/coller » (Gibney, 2006), une énorme faille persiste. En effet, le fait de reformuler ou tout simplement de paraphraser un texte, en utilisant des synonymes par exemple, rend la plupart des techniques actuelles caduques. Certains articles (Callison-Burch et al., 2008; Bannard et Callison-Burch, 2005) se sont tout de même