Méthode alternative à la détection de « copier/coller » : intersection de textes et construction de séquences maximales communes

Jérémy Ferrero*, Alain Simac-Lejeune**

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

Résumé. La détection du plagiat passe le plus souvent par la phase de recherche de similitudes la plus naïve, la détection de « copier/coller ». Dans cet article, nous proposons une méthode alternative à l'approche standard de comparaison mot à mot. Le principe étant d'effectuer une intersection des deux textes à comparer, récupérant ainsi un tableau des mots qu'ils ont en commun et de ne conserver que les séquences maximales des mots se suivant dans l'un des textes et existant également dans l'autre. Nous montrons que cette méthode est plus rapide et moins coûteuse en ressources que les méthodes de parcours de textes habituellement utilisées. L'objectif étant de détecter les passages identiques entre deux textes plus rapidement que les méthodes de comparaison mot à mot, tout en étant plus efficace que les méthodes n-grammes.

1 Introduction

Lorsque l'on cherche à comparer deux documents, on recherche tout élément présent dans l'un qui est également présent dans l'autre, ces éléments sont dénommés "similitudes". Les plus évidentes à voir à l'œil humain sont les similitudes exactes, les parties copiées d'un document directement dans l'autre. Cependant, reproduire informatiquement cette capacité humaine est une opération délicate. Ce procédé est souvent gourmand en temps, car passant par une comparaison mot à mot afin d'identifier les séquences de mots identiques dans les deux textes. De ce fait, des méthodes beaucoup moins gourmandes ont vu le jour. Basées sur un système de n-grammes, elles extraient des séquences de n mots se suivant d'un texte et en cherche la présence dans l'autre. C'est dans l'optique de proposer une alternative à ces méthodes que nous allons décrire dans cet article une nouvelle approche de construction de séquences communes.

Après avoir présenté l'état de l'art, nous décrirons dans un premier temps le processus d'intersection des deux textes, ensuite, la phase de construction des plus longues séquences communes et pour finir, nous présenterons l'évaluation de notre approche en la comparant aux méthodes naïves de comparaison mot à mot et à la méthode classique n-grammes.