Recherche de documents similaires sur le web par segmentations hiérarchiques et extraction de mots-clés

Alain Simac-Lejeune*

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

Résumé. La recherche de documents similaires est un processus qui consiste à trouver les documents présentant des similitudes, comme la copie ou la reformulation, sur des bases documentaires ou sur internet. Elle est utilisée notamment pour protéger la propriété intellectuelle de productions issues de l'enseignement, de la recherche ou de l'industrie. Dans cet article, nous définissons une approche automatique pour permettant d'extraire des mots-clés d'un document en effectuant un bouclage sur une succession de découpage de plus en plus petit. Cette approche permet d'obtenir des mots-clés impossibles à obtenir par une approche globale notamment quand la thématique, le style ou le contenu d'un document varient dans le document. L'objectif est de permettre la détection des documents présentant des similitudes en utilisant uniquement des mots-clés.

1 Introduction

Actuellement, de nombreuses recherches (Stein et al., 2007) traitent de la recherche de similitudes notamment à cause de l'augmentation importante du plagiat sous toutes ses formes et dans tous les domaines : l'enseignement avec les élèves et étudiants (quatre étudiants sur cinq déclarent avoir recours au copier-coller), la recherche scientifique avec les publications et thèses (Bao et Malcolm, 2006) (plagiat de thèses notamment) et l'industrie avec les problèmes de copie de brevets ou de codes sources. Les outils existants pour rechercher des documents similaires sont principalement basés sur la recherche de segments dits *n-gram* (n représentant la taille en mots du segments) identiques (Oberreuter et al., 2010) pour détecter les copies et commencent tout juste à proposer la détection de copie par traduction dite la copie inter-langue (Kent et Salim, 2009) à travers les travaux de ces dix dernières années.
L'approche proposée consiste à prendre en compte le document comme une agrégation de documents plus petits et récursivement que chacun des documents le composant soit lui-même l'agrégation de documents plus petits. Cette hiérarchie permet de déterminer des mots-clés à chacun des niveaux et ainsi détecter des similitudes normalement indétectables à l'échelle globale. Cette approche repose sur l'hypothèse que lorsqu'*on paraphrase ou reformule un texte, on garde le sens de celui-ci et ainsi on garde les mots-clés principaux, porteurs du plus haut niveau sémantique du texte*.