# 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*.

Après avoir présenté rapidement l'état de l'art et l'approche, nous décrivons d'abord comment extraire et utiliser des mots-clés, puis nous présenterons les niveaux hiérarchiques proposés (taille, organisation, obtention). Enfin, nous présentons l'évaluation de notre approche en la comparant à la méthode classique n-gram.

# 2 État de l'art et approche

## 2.1 La notion de similitude

Une similitude est un rapport, une relation qui existe entre deux choses semblables. Cela peut aller de la simple ressemblance jusqu'à l'identité. Lorsqu'on parle de similitude textuelle ou de document similaire, on distingue plusieurs types de similitudes allant de la ressemblance à l'identité : la **reformulation** qui consiste à reprendre la sémantique d'un texte et à l'exprimer différemment ; la **paraphrase** qui consiste à reprendre les éléments d'un texte, dans l'ordre d'origine mais en les formulant différemment normalement dans le but d'éclairer, d'expliciter, ou pour développer certains points ; la **citation** qui consiste à faire la copie mots à mots d'une portion de texte en informant le lecteur de l'origine extérieure de celle-ci ; la **copie** qui consiste à faire la copie mots à mots d'un texte ou d'une partie d'un texte sans citer la source ; et la **traduction** qui consiste à faire la copie mots à mots d'un texte ou d'une partie d'un texte sans citer la source et en traduisant dans une autre langue. Selon la langue, la forme finale est donc potentiellement différente de l'originale.

## 2.2 Recherche et comparaison de documents

La recherche de documents s'effectue selon deux approches. La première est basée sur **le style de l'auteur**, elle est dite stylistique (Iyer et Singh, 2005), part du postulat (Jardino et al., 2005) que *deux textes du même auteur contiennent un grand nombre de correspondance et qu'inversement, deux textes de deux auteurs différents contiennent un petit nombre de correspondances*. On recherche de documents présentant un style identique ou approchant en utilisant la structure de phrase, la grammaire ou par observation stylistique (Stamatatos, 2009). Cette approche est plutôt récente et bien que très rapide, elle pose encore un certain nombre de problème notamment sur l'aspect de la précision. La seconde approche est basée sur **le contenu du document** (White et Joy, 2004; Iyer et Singh, 2005; Eissen et Stein, 2006) et consiste à rechercher des similitudes exactes ou approchées entre documents par analyse de contenu en comparant les caractéristiques extraordinaires extraites de chaque document. Les caractéristiques souvent utilisées sont des mots, des fautes d'orthographe, des syntagmes ou des suites exactes. La plupart des approches et outils disponibles utilisent donc l'approche contenu.
La comparaison entre deux documents (le document à vérifier et le document cible) s'effectue par analyse de similarités entre deux ensembles (différence ensembliste), entre deux vecteurs (calcul de distance d'Euclide ou Cosin), entre deux listes (distance de Levenshtein ou calcul de la sous-chaîne partagée la plus longue dit LCS) ou entre deux représentations composées. Les éléments constituants sont soient les mots, soient des blocs de mots, soient des syntagmes. Un syntagme est un ensemble de mots formant une seule unité catégorielle et fonctionnelle, constituant une unité sémantique, mais dont chaque constituant, parce que dissociable, conserve sa signification et sa syntaxe propre.

| Nombre mots-clés | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Nombre groupes | 10 | 45 | 120 | 210 | 252 | 210 | 120 | 45 | 1 | 1 |
| En premier (en %) | 46 | 43 | 72 | 73 | 59 | 38 | 26 | 14 | 11 | 6 |
| 3 premiers (en %) | 49 | 48 | 79 | 79 | 61 | 39 | 28 | 14 | 12 | 8 |
| 5 premiers (en %) | 51 | 53 | 81 | 83 | 63 | 39 | 29 | 15 | 12 | 9 |
| 10 premiers (en %) | 56 | 57 | 87 | 89 | 67 | 40 | 32 | 16 | 12 | 9 |

TAB. 1 – *Détermination du nombre de mots-clés à utiliser dans les recherches de document. Nombre de combinaisons de mots-clés par quantité de mots-clés et pertinence de recherche. Il s'agit d'une moyenne sur l'ensemble des 100 documents utilisés.*

### 2.3 Notre approche

Notre approche consiste à extraire sur chaque document les mots-clés principaux d'un document dans sa globalité mais également dans les sous-documents construits par découpages successifs. Pour chaque niveau hiérarchique, on combine les mots-clés 3 par 3 pour former des triplets de recherche.

## 3 Extraction et utilisation de mots-clés

### 3.1 La notion de mots-clés

Le terme *mot-clé* désigne de manière générale un mot qui a une importance particulière et il est notamment utilisé lors des recherches d'informations. Il existe plusieurs manières d'extraire des mots-clés afin de garantir leur pertinence. La plus 'traditionnelle' est la méthode fréquentielle. Plus un mot-clé apparaît souvent, plus il est important. En ajoutant le regroupement par racine d'un même mot (regroupement de 'cheval' et de 'chevaux' sous le lemme 'cheva*' par exemple) et en n'utilisant que les mots les plus informatifs sémantiquement, on obtient les mots-clés théoriquement les plus représentatifs du contenu sémantique d'un texte.

### 3.2 Les groupes de mots-clés

A partir de la liste des mots-clés trouvés, il est important de savoir si il faut chercher les documents similaires avec un, deux, trois ou plus, mots-clés combinés. Afin de déterminer ce nombre, nous avons utilisé une petite base de documents sur laquelle nous avons extrait les 10 principaux mots-clés extraits sur l'analyse de l'intégralité du document. Puis ils ont été utilisés pour retrouver le document.

Le tableau 1 présente la pertinence de recherche d'un document par nombre de mots-clés utilisés dans la recherche. Le nombre de requêtes de recherche variant d'un minimum de 100 requêtes à un maximum de 25200. Au final, on observe un bon compromis entre résultats et nombre de requêtes effectuées en utilisant des groupes de 3 mots-clés et avec des résultats allant de 72 à 87% de récupération du document tout en étant à 120 requêtes par document. Le passage à 4 mots-clés ne fait gagner qu'0,5% de récupération supplémentaire alors qu'il double le nombre de requêtes (passage de 120 à 210 par document).

# 4 Niveaux hiérarchiques

## 4.1 Méthodes de segmentation

Il existe de nombreuses manières de procéder à la segmentation d'un texte. Bien qu'il existe des méthodes indépendante du formalisme, ou des méthodes par modèles thématiques, les principales sont encore le segmentation sémantique qui consiste à découper en fonction de bloc logique lié à la rédaction ou à la lecture (découpage par parties, par paragraphes, par sous-paragraphes, par phrases, découpage par ponctuation) et la segmentation dimensionnelle qui consiste à découper en bloc de dimension donnée sans prendre en compte le contenu du document. Cette segmentation peut elle-même être effectuée de manière absolue (découpage par bloc de $1000$ mots, $500$ mots...) ou de manière relative (découpage par bloc couvrant $25\%$, $10\%$...).

## 4.2 Taille et pertinence des différents niveaux

Les mots-clés sont extraits principalement par méthode fréquentielle ce qui implique qu'il soit nécessaire d'avoir un nombre minimum de mots pour générer des mots-clés. Statistiquement, des segments de moins de 40 mots ne comportent jamais de mots-clés intéressants, ceux de moins de 70 mots très rarement et au mieux, des mots représentés 2 fois au maximum. En effet, les auteurs ont tendance à éviter de faire des répétitions dans des zones trop proches. Ainsi, la méthode fréquentielle ne peut s'appliquer que sur des segments d'au moins 70 mots. Il est impossible de travailler avec des phrases, rarement avec des paragraphes. La taille minimale des segments doit donc être entre 70 et 100 mots afin d'obtenir un nombre minimal d'occurrences des mots.

## 4.3 Distribution des segments

Dans le cas de la segmentation dimensionnelle absolue, on peut légitimement se demander la pertinence de la position d'un segment de taille donnée dans l'ensemble du document qui pourrait avoir une multitude de positions différentes. Une approche naïve consiste à positionner un découpage donné comme une sous-partie directe d'une partie plus grande et de procéder par découpage successif c'est à dire un découpage par partie disjointe. Dans le but de couvrir des segments différents et potentiellement intéressants, une alternative est d'utiliser la taille du segment pour créer un masque de récupération que l'on utilise en balayant le texte pour générer plusieurs segments non disjoints i.e. avec une intersection non vide. Dans l'approche par segments non disjoints, il convient de définir le déplacement du masque (pas de masque) ou le nombre maximal de parties acceptables par niveau hiérarchique (le pas étant alors égal au rapport taille du document / nombre de parties). L'expérimentation montre qu'un pas de masque du tiers de la taille du document permet d'obtenir des résultats satisfaisants sans augmenter de manière trop importante le nombre de segments générés.

# 5 Évaluation et tests

## 5.1 La base de tests et protocole

La base de tests est composée de 200 textes de 2500 mots environ dont on dispose des sources effectivement utilisées pour les écrire. Certains reprises sont de la copie au mot près (copie ou citation), d'autres sont des paraphrases ou même des reformulations. La taille des sections copiées allant de la simple phrase (6-8 mots) à celle d'une partie complète (1000 mots). On peut avoir plusieurs reprises dans un même document. Les sources utilisées sont disponibles en ligne et d'origine diverses : Wikipédia, publications de recherche, mémoires de stage, article de presse. Le protocole est composé d'une extraction de mots-clés de manière hiérarchique sans et avec glissement puis la recherche de documents en utilisant les triplets de mots-clés parmi ceux extraits. L'ensemble des documents utilisés est accessible librement sur internet.
La recherche est également effectuée en utilisant la méthode traditionnelle des n-gram avec n égal à 8.

## 5.2 Résultats

| **Attributs/Sources** | **Répart,** | **Mots-clés** | **Glissants** | **n-gram** |
|---|---|---|---|---|
| Copie | 122 | 101 | 107 | 108 |
| Cit./Para./Ref. | 43/42/57 | 12/29/21 | 13/32/23 | 10/0/0 |
| Phrase (6-8 mots) | 147 | 18 | 19 | 112 |
| 20/50/100 mots | 117/45/29 | 22/28/21 | 30/33/25 | 76/31/23 |
| 250/500/1000 mots | 27/18/10 | 25/17/10 | 26/18/10 | 24/16/10 |
| Wikipédia | 20 | 19 | 19 | 19 |
| Pub./Mém./Art. | 40/10/50 | 31/6/42 | 32/6/44 | 33/1/37 |
| Blog et divers | 100 | 94 | 96 | 95 |
| Nombre/temps requêtes | - | 26437 (21 ms) | 35811 (21 ms) | 62500 (32 ms) |

TAB. 2 – *Répartition des attributs et des sources des documents formant la base d'évaluation ainsi que le nombre de documents trouvés.*

Les résultats sont présentés dans le tableau 1 et permettent d'observer les différents résultats avec l'approche mots-clés et avec l'approche n-grammes. **L'approche mots-clés** présente plusieurs points plutôt positifs. Elle permet la détection des documents similaires présentant des similitudes supérieurs à 100 mots à 90% bien qu'il y ait des disparités entre types de document (83% copie, 69% paraphrase, 37% reformulation) mais aussi entre sources (94% internet, 75% articles et 60% mémoire). Elle permet également de trouver des similitudes de type paraphrase ou reformulations impossible à trouver avec l'approche n-gram (sauf si inférieur à 100 mots). Enfin, elle est trois fois plus rapide que l'approche n-gram en calcul et 40% plus rapide dans l'exécution des requêtes moteur (simplicité de requête). L'ajout du glissement permet l'amélioration des résultats de l'ordre de 5% mais en ajoutant environ 5% de requêtes.

# 6 Conclusions

La méthode proposée montre des résultats probants pour la recherche de documents similaires. Elle présente de très bonnes performances pour les similitudes supérieures à 100 mots et permet la détection des similitudes de type paraphrase ou reformulations. De plus, cette approche est plutôt rapide et permet de diminuer de manière importante le nombre de requêtes de recherche à effectuer. Cependant, elle présente de très mauvaises performances sur des similitudes inférieures à 50 mots, les phrases et les citations. Au final, les mots-clés extraits de manière hiérarchique proposent une alternative intéressante aux n-gram qui pourrait être une méthode à privilégier pour la détection de paraphrase/reformulations.

# Références

Bao, J.-P. et J. A. Malcolm (2006). Text similarity in academic conference papers.

Eissen, S. M. Z. et B. Stein (2006). Intrinsic plagiarism detection. In *Proceedings of the European Conference on Information Retrieval (ECIR-06*.

Iyer, P. et A. Singh (2005). Document similarity analysis for a plagiarism detection system. In *2nd Indian International Conference and Artificial Intelligence*.

Jardino, M., M. Hurault-Plantet, et G. Illouz (2005). Identification de thème et reconnaissance du style d'un auteur pour une tâche de filtrage de textes. In *DEFT05*.

Kent, C. K. et N. Salim (2009). Web based cross language plagiarism detection. *CoRR volume abs/0912.3959*.

Oberreuter, G., G. L. Huillier, S. A. Ríos, et J. D. Velásquez (2010). Fastdocode : Finding approximated segments of n-grams for document copy detection - lab report for pan at clef 2010. In *CLEF (Notebook Papers/LABs/Workshops)*.

Stamatatos, E. (2009). Intrinsic plagiarism detection using character n-gram profiles. In *Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN09)*.

Stein, B., M. Koppel, et E. Stamatatos (2007). Plagiarism analysis, authorship identification, and near-duplicate detection pan'07. *SIGIR Forum 41*(2), 68–71.

White, D. et M. Joy (2004). Sentence-based natural language plagiarism detection. In *ACM Journal on Educational Resources in Computing*, Volume 4.

# Summary

Research of similar documents is a process which consists in finding documents with similarities, such as copying or paraphrasing, on document databases or on Internet. It is used in particular to protect the intellectual property of productions from education, research and industry. In this article, we define an approach to automatic extraction of keywords from a document by performing a loopback on a series of cutting smaller and smaller. This approach provides keywords unobtainable by a classical approach especially when the theme, style or content of a document in the document vary. The objective is to enable the detection of documents with similar using only keywords.