الى أمي...

« If we knew what we were doing, it wouldn't be called research »

Albert Einstein

Avant-Propos

Je tiens tout d'abord à remercier Messieurs Mohamed OUKESSOU – Mohamed ERRITALI qui m'ont encadré tout au long de cette thèse, pour m'avoir guidé et fait confiance pendant ma formation à la Faculté des Sciences et Technique. Travailler avec eux à l'interface entre informatique et mathématiques appliquées a été un vrai plaisir. La liberté et la confiance qu'ils ont su m'accorder n'ont pu qu'accroître ma motivation. En dehors de leurs apports scientifiques, je n'oublierai pas aussi de les remercier pour leurs qualités humaines.

Mes remerciements vont également à Mr Said MELLIANI directeur du Laboratoire des Mathématiques Appliqués et Calcul Scientifique, qui m'a fait l'honneur de bien vouloir participer au jury de ma thèse.

Je tiens aussi à remercier Monsieur Mohamed FAKIR et Madame Najlae EL-IDRISSI professeurs à la Faculté des Sciences et Techniques de Beni Mellal, et Monsieur Bouabid EL OUAHIDI professeur à la Faculté des Sciences de Rabat pour avoir consacré le temps à la lecture de cette thèse ainsi pour avoir soumis leur précieux jugement sur la qualité et le contenu de ce travail.

Je tiens aussi à souligner le plaisir que j'ai eu à travailler au sein de Laboratoire des Mathématiques Appliqués et Calcul Scientifique, et j'en remercie ici tous les membres.

Enfin, j'adresse mes plus sincères remerciements à ma famille : à mes parents, mon frère, ma sœur et tous mes proches et amis dont l'affection, l'amour, le soutien et l'encouragement constants m'ont été d'un grand réconfort pour la réalisation de ce travail.

Liste des Publications et des Communications

1. Ezzikouri H., Oukessou M., Youness M., Erritali M. (2019) **Fuzzy Semantic-Based Similarity and Big Data for Detecting Multilingual Plagiarism in Arabic Documents**. In: Ezziyyani M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2018). AI2SD 2018. Advances in Intelligent Systems and Computing, vol 915. Springer, Cham.
2. Ezzikouri H., Erritali M., Oukessou M. (2019) **Plagiarism Detection in Across Less Related Languages (English-Arabic): A Comparative Study**. In: Khoukhi F., Bahaj M., Ezziyyani M. (eds) Smart Data and Computational Intelligence. AIT2S 2018. Lecture Notes in Networks and Systems, vol 66. Springer, Cham.
3. H. Ezzikouri, M. Oukessou, M. Youness, et M. Erritali, « **Fuzzy Cross Language Plagiarism Detection (Arabic-English) Using WordNet in a Big Data Environment** », in Proceedings of the 2018 2Nd International Conference on Cloud and Big Data Computing, New York, NY, USA, 2018, p. 22–27.
4. Ezzikouri H., Oukessou M., Erritali M., Madani Y. (2019) **Fuzzy Cross Language Plagiarism Detection Approach Based on Semantic Similarity and Hadoop MapReduce**. In: Melliani S., Castillo O. (eds) Recent Advances in Intuitionistic Fuzzy Logic Systems. Studies in Fuzziness and Soft Computing, vol 372. Springer, Cham
5. Hanane EZZIKOURI, Mohamed ERRITALI and Mohamed OUKESSOU, **"Fuzzy-Semantic Similarity for Automatic Multilingual Plagiarism Detection**" International Journal of Advanced Computer Science and Applications(IJACSA), 8(9), 2017. http://dx.doi.org/10.14569/IJACSA.2017.080912
6. H. Ezzikouri, M. Erritali, et M. Oukessou, « **Semantic Similarity/Relatedness for Cross Language Plagiarism Detection** », Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, no 2, p. 371-374, févr. 2016.
7. EZZIKOURI, & H. ERRITALI, M., (2015). **Pretreatment of web log files**. Journal of Information Sciences and Computing Technologies, 2(1), 108-121. Retrieved from http://www.scitecresearch.com/journals/index.php/jisct/article/view/29
8. International Conference on Cloud and Big Data Computing (ICCBDC 2018), du 3 au 5 Aôut 2018, à Barcelone, Espagne. **H. Ezzikouri, M. Erritali, et M. Oukessou**, participation avec une Communication oral intitulée " **Fuzzy Cross Language plagiarism Detection (Arabic-English) using WordNet in a big data environment documents**".
9. International Conference on Advanced Intelligent Systems for Sustainable Development, du 11 au 14 Juillet 2018, à la Faculté des Sciences et Techniques de Tanger, Maroc. **H. Ezzikouri, M. Erritali, et M. Oukessou**, participation avec une Communication orale intitulée "**Fuzzy Semantic-based Similarity and big data for detecting multilingual plagiarism in Arabic documents**".
10. International Conference on Intuistionistic Fuzzy Sets & Mathematics Sciences, du 11 au 13 Avril 2018, à université alAhawayn Ifrane, Maroc. **H. Ezzikouri, M. Erritali, et M. Oukessou,** participation avec une Communication orale intitulée "**Fuzzy Semantic Similarity for Cross Language Plagiarism Detection (Arabic-English) using big Data**".

11. **International Training Workshop on Application and Promotion of Traffic Information Management System**, du 12 Septembre au 1 Octobre 2016, à Anhui Keli Information Industry Co., Ltd. Hefei, Chine.
12. International Conference on Intuistionistic Fuzzy Sets Theory And Applications, du 20 au 22 Avril 2016, à université Sultan Moulay Slimane, Faculté des Sciences et Techniques de Beni Mellal.
13. 13th International Conference for Computer Graphics, Imaging and Visualisation, du 29 Avril au 1 Mars 2016, à université Sultan Moulay Slimane, Faculté des Sciences et Techniques de Beni Mellal. **H. Ezzikouri, M. Erritali, et M. Oukessou,** participation avec une Communication orale intitulée "**Semantic Similarity/Relatedness for Cross Language Plagiarism Detection**".
14. First Spring Conference on Applied Science and Computing, du 30 au 31 Mai 2015, à l'Ecole Supérieure de Technologie Essaouira Maroc. **H. Ezzikouri, M. Erritali, et M. Oukessou,** participation avec une Communication orale intitulée "**Developing a multilingual plagiarism detection corpus**".
15. The Second International Conference on Business Intelligence (CBI'15), du 23 au 25 April 2015, à université Sultan Moulay Slimane, Faculté des Sciences et Techniques de Beni Mellal, Maroc. **H. Ezzikouri, M. Erritali, et M. Oukessou,** participation avec une Communication écrite intitulée "**A Comparative study of methods for plagiarism detection in across less related languages (English-Arabic)**".

Résumé

Le plagiat multilingue fait référence à la réutilisation non reconnue d'un texte impliquant sa traduction d'une langue naturelle à une autre, sans référence appropriée à la source d'origine. L'un des problèmes courants du traitement des données réside dans l'efficacité de la comparaison de textes volumineux. Dans un cas d'une similarité sémantique floue, la complexité des langages naturels (en particulier l'Arabe), et le nombre croissant de publications, contribuent à l'augmentation du taux de documents suspects sources de plagiat. CLP (Cross-Language Plagiarism) est un processus plus compliqué que le plagiat monolingue. CLP est plus qu'une copie munie d'une traduction, c'est un changement sérieux du texte traduit sans indication de la source. Par conséquent, le processus de détection révèle le besoin des analyses et des techniques compliquées pour découvrir des pratiques malhonnêtes de plagiat caché dans des documents arabes traduits de sources anglaises ou françaises.

Dans cette thèse, nous proposons un système de détection de plagiat multilingue sémantique. D'abord on a construit un système de détection du plagiat multilingue CLPD (Cross-Language Plagiarism Detection) basé sur la similarité sémantique en utilisant WordNet. Puis, pour une analyse plus profonde des cas de plagiat multilingue, le système a été étendu en une similarité sémantique basée sur la théorie des ensembles flous. Ensuite, le travail est parallélisé en utilisant Apache Hadoop avec le système de fichiers HDFS et le modèle de programmation MapReduce, pour gérer les grandes masses d'informations et le nombre important d'opérations et des calculs faites dans un tel système.

Mots-clés : Plagiat multilingue ; théorie des ensembles floues ; traitement du langage naturel ; Arabe ; similarité sémantique.

Abstract

Cross-Language Plagiarism (CLP) refers to the unacknowledged reuse of a text involving its translation from one natural language to another without proper referencing to the original source. One of the common problems in data processing is efficient large-scale text comparison, especially in a fuzzy semantic based similarity due to the complexity of natural languages in particular Arabic, and the increasing number of publications which raise the rate of suspicious documents sources of plagiarism. CLP is more complicated than monolingual plagiarism, it goes beyond "copy+translate" and paste, consequently the detecting process exposes the need for complicated analysis and techniques to uncover dishonest practices of hidden plagiarism in Arabic documents translated from English or French sources.

In this thesis, we propose a Semantic Cross-Language Plagiarism Detection system (CLPD), The CLPD system is a semantic-based similarity using WordNet. This system is expanded to a fuzzy semantic-based similarity based on fuzzy set theory to analyse and compare hard multilingual plagiarism cases. The work is parallelized using Apache Hadoop with HDFS file system and MapReduce programming model.

Keywords—Cross Language Plagiarism; fuzzy set theory; natural language processing; Arabic; semantic similarity.

Table des matières

Liste des acronymes

BiPLS	:	Bilingual Partial Least Squares
BWESG	:	Bilingual Word Embeddings Skip-Gram
CDSDG	:	Copying Detection System of Digital Goods
CLIR	:	Cross lingual Information Retrieval
CLPD	:	Cross Language Plagiarism Detection
CNG	:	Character N-gramme
DAG	:	Diagramme Acyclique Dirigé
DPM	:	Détection du Plagiat Multilingue
HMM	:	Hidden Markov Model
IC	:	Information Content
IR	:	Information Retrieval
LCS	:	Longest Common Sequence
LSI	:	Indexation Sémantique Latente
MIS	:	Most Informative Subsume
MIS	:	Most Informative Subsume
MLMH	:	histogram- based multilevel matching
MLMS	:	Multilevel Matching
MT	:	Machine Translation
NLP	:	Natural Langage Processing
OWL	:	Ontology Web Language
POS Tagging	:	Part-Of-Speech Tagging
SCAM	:	Standard Copy Analysis Mechanism

SMT	:	Statistical Machine Translation
SOM	:	Self Organizing Map
SOM	:	Self-Organizing Map
SVD	:	Décomposition de Valeur Singulière
HDFS	:	Hadoop Distributed File System
VSM	:	Vector Space Model
WE	:	Word Embeddings
WNG	:	Word N-gramme

Liste des Figures

Liste des tables

Introduction Générale

Le développement d'internet et des technologies de l'information au niveau mondial a profondément transformé la gestion des documents. Avec la multitude de données et d'informations générées, nous avons trop souvent tendance à oublier les règles fondamentales du droit d'auteur. Ces révolutions technologiques ont engendré de nouvelles problématiques pour la recherche d'information et la détection de plagiat.

Le terme Plagiat (contrefaçon d'un point de vue juridique) a une connotation morale et éthique, par lequel on désigne le fait de prendre quelque chose (pensées, idées, texte, expressions, résultats, etc.) de quelqu'un d'autre de façon non avouée et infidèle.
Internet a annulé les frontières géographiques, en facilitant les échanges de documents et d'informations entre tous le globe. Par conséquent, les collections de documents écrites dans différentes langues peuvent être facilement et discrètement plagiés grâce à l'utilisation de la traduction sémantique (humaine ou machine), dans ce cas on parle de plagiat multilingue, de nature compliquée, hétérogène et de frontières floues. Ce dernier se reflète lors de la réutilisation non reconnue d'un texte impliquant sa traduction d'une langue à une autre, sans donner aucune référence à la source d'origine. Un autre problème qui rend la tâche de détection du plagiat plus laborieuse est la masse d'information et de documents existants sur le Web.

La langue arabe appartient au groupe linguistique afro-asiatique. Elle a beaucoup de spécificité, ce qui la rend très différente des autres langues indo-européennes. La détection du plagiat dans les documents arabes est particulièrement difficile, et ça devient encore plus dur, vu que la traduction est un processus flou et difficile à rechercher, en raison de la structure linguistique complexe de l'Arabe. Malgré les nombreuses recherches menées sur la détection du plagiat au cours des dernières décennies, celles concernant le texte en langue Arabe restent très limitées et concernent particulièrement le plagiat monolingue.

Dans le but de répondre à ces problématiques nous avons proposé une approche de détection du plagiat multilingue pour des langages distants qui ne partagent presque aucune caractéristique, allant de la morphologie jusqu'au lexique comme l'Arabe et l'Anglais, sans transformer le problème multilingue en monolingue via l'unification de la langue.

Afin de construire notre système de détection du plagiat sémantique multilingue, nous avons tout d'abord proposé une nouvelle approche sémantique basée sur un système de corpus

parallèles, pour trouver des co-occurrences multilingues, ou pour obtenir des modules de traduction. Les principes et les ressources de la traduction automatique (MT) sont appliqués sans que la traduction soit effectivement faite, pour garder le cadre de détection multilingue. Tout en introduisant la théorie des ensembles flous pour un confinement efficace d'ingéniosité du plagiaire, qui est sans doute quelqu'un de très dangereux, et qui fait un maximum d'effort pour cacher son crime. Cela en se basant sur l'hypothèse que la correspondance de deux phrases est approximative, ce qui peut être modélisé en considérant que chaque mot d'une phrase est associé à un ensemble flou qui contient des mots ayant la même signification.

Le traitement des informations dans un environnement Big Data retient beaucoup d'attention vu que la détection de documents similaires nécessite un nombre important d'opérations. Une étape préliminaire dans ce processus est la diminution du nombre de documents candidats (c'est-à-dire l'étape de la sélection des documents suspects comme source de plagiat). Cependant, le nombre de candidats reste énorme pour une analyse très approfondie d'un processus de similarité sémantique flou, car généralement, les collections de documents sont d'énormes corpus ou parfois le Web. Le grand nombre d'opérations et des calculs faites dans un tel système de DPM (Détection Plagiat Multilingue) et la gestion des grandes masses d'informations et de données nous a mené à utiliser Hadoop et Mapreduce dans le but de diminuer le temps d'exécution et de distribution du stockage.

Notre thèse traite la détection du plagiat multilingue. Elle est organisée comme suit :

Le **premier chapitre** introduit les termes et les concepts essentiels pour appréhender la notion du plagiat et ses types. Nous présentons les différences entre la détection du plagiat monolingue et multilingue, puis nous donnons un aperçu des modèles de sélection des documents candidats pour la phase de pré-détection. Ensuite, nous mettons l'accent sur les différentes méthodes existantes de détection du plagiat dans une partie d'état de l'art.

Le **deuxième chapitre** présente la partie du traitement du langage naturel. Les méthodes de traitement du texte (tokénisation, lemmatisation, suppression des mots vides…) appliquées sur les documents arabes et anglais/français de notre corpus. Une analyse détaillée des textes des langues a été effectué.

Le **troisième chapitre** donne les notions d'ontologie indispensable à notre système, et décrit l'approche de calcul de la similarité sémantique proposé pour la détection du plagiat multilingue, en commençant par le calcul de la similarité sémantique (Arabe-Anglais) mot à mot.

Le **quatrième chapitre** expose notre approche de calcul de la similarité sémantique floue pour la détection du Plagiat Sémantique Multilingue en se basant sur les principes de la théorie des ensembles floues.

La première partie du **cinquième chapitre** décrit notre choix de couplage de la détection sémantique floue multilingue du plagiat et l'utilisation d'un système Big Data. La deuxième partie est consacrée à l'application de cette approche et les comparaisons des résultats trouvés.

Enfin, une synthèse générale - qui reprend l'ensemble des contributions de ce travail de recherche - est présentée. De plus, un ensemble de perspectives sont identifiées et discutées.

Chapitre 1 Plagiat : Généralités et Etat de l'Art

I. Introduction

L'étymologie du mot plagiat remonte aux Grecs au mot « *plagios* », qui signifiait « *ce qui est oblique, de travers* ». Les Romains antiques changeaient l'appellation de ce mot est devenu « *plagiarius* » ou « *plagiator* », qui servait à qualifier celui qui vole ou kidnappe les enfants ou les esclaves de quelqu'un d'autre, ou quelqu'un qui vendait comme esclave une personne libre [1].

Le statut du plagiat a intensément changé avec les époques. Il était une pratique valorisée, qui définissait l'imitation à la fois comme inévitable et nécessaire. En effet, la copie était la seule façon d'apprendre. Le poète Marcus Valerius Martialis[1] a été le premier à avoir appliqué le terme « *plagiat* » dans un sens figuré pour dénoncer ceux qui s'appropriaient de ses textes [2]. Avec l'arrivée du concept de la propriété intellectuelle (Copyright) en XVIII^{e} siècle, accuser un auteur de plagiat est devenu l'insulte suprême, Il est souvent assimilé à un vol immatériel.

Le plagiat a toujours existé dans le monde arabe ancien durant la période antéislamique. Le célèbre vers du poète arabe Al Akhtal[2] : « *نحن معاشر الشعراء أسرق من الصاغة* » est un indice fort à ce propos. Toutefois, le plagiat était annoncé sous d'autres appellations telles que l'emprunt laudatif (الانتحال، الادعاء، الإغارة، الغصب، الاختلاس، الموازنة، التوارد) et l'emprunt péjoratif (الاجترار، الامتصاص، الشرح، التضمين، الحوار، التوليد). Il y avait une dépréciation ou condamnation du plagiat par plusieurs poètes de l'époque qui ont pris le plagiat comme un acte honteux, à savoir Al Hariri[3] qui déclare que " *Le vol des idées équivaut au vol de l'âme*", et aussi Ṭarafa[4] dans son célèbre vers :

*ولا أغير على الأشعار أسرقها *** عنها غنيت وشر الناس من سرقا*
*وإن أحسن بيت أنت قائله *** بيت يقال إذا أنشدته صدقا*

[1] Marcus Valerius Martialis, né vers l'an 40 et mort vers 104 à Bilbilis dans l'Empire romain. Un poète latin, connu pour ses Épigrammes, dans lesquelles il dépeint la société romaine de son temps.

[2] Al-Akhtal (الأخطل التغلبي) de son nom complet Ghiyath ibn Ghawth al-Taghlibi al-Akhtal (غياث بن غوث التغلبي), né en 640, mort en 710. Un poète chrétien issu de la tribu arabe des Taghlib.

[3] Abu Muhammad al-Qasim ibn Ali Al-Hariri (أبو محمد القاسم بن علي الحريري), dit aussi Al-Hariri de Basra, né en 1054 et décédé en 1122 à Bassora, en Irak, était un savant et écrivain arabe.

[4] Ṭarafa (arabe : طرفة ابن العبد) est le nom donné à un poète antéislamique, auteur d'une Muʿallaqa né en 543, mort en 569.

II. Plagiat :

II.1. Définition

Le terme plagiat a une connotation morale et esthétique, par lequel on désigne le fait de prendre quelque chose (pensées, idées, texte, expressions, résultats, etc.) de quelqu'un d'autre de façon non avouée et infidèle. En droit, on parle plutôt de « *contrefaçon »* et d'infraction à la loi du droit d'auteur (copyright) et non du « *plagiat »*.

Le plagiat peut être défini de plusieurs façons. Il est défini dans les dictionnaires comme « *appropriation injustifiée* » et « *imitation proche* ». L'une des définitions les plus intéressantes est celle de l'IEEE [3]:

« *Le plagiat est la réutilisation des processus antérieurs, idées, résultats ou mots de quelqu'un d'autre sans reconnaître explicitement l'auteur original et la source* ».

Comme s'est définit, le plagiat peut être de différentes natures, allant de la copie de texte jusqu'à l'adoption d'idées, alors en se basant sur le comportement du plagiaire on peut diviser le plagiat en deux types : littéral et intelligent.

II.2. Types du plagiat

II.2.1 Le plagiat littéral

Dans le plagiat littéral, syntaxique ou encore verbatim (copier-coller), est une pratique commune et majeure, où le plagiaire copie directement une partie ou la totalité d'un document source dans son propre travail, il peut aussi tenter de modifier légèrement le texte en supprimant, en ajoutant ou en remplaçant des mots / phrases et en réécrivant de courtes parties du passage qui affectent sa syntaxe.

II.2.2 Le plagiat intelligent

Le plagiat intelligent est une malhonnêteté académique sérieuse qui entraîne des effets négatifs importants, car elle mine toute la performance et la réputation d'une institution et d'un individu. Alors que le plagiat peut se produire incidemment et involontairement, en raison de phénomènes tels que la cryptomnésie[5] [4], il est souvent le résultat d'un processus conscient, où les plagiaires tentent de cacher, d'obscurcir et de modifier le travail original de diverses manières, y compris la manipulation de texte, la traduction et l'adoption d'idées.

[5] Le terme "cryptomnésie" signifie l'existence de souvenirs qui sont cachés de la conscience. Elle peut donner lieu à un plagiat involontaire, surtout lorsque les souvenirs logiques ne sont plus reconnus comme des souvenirs, mais sont vécus comme des idées nouvellement créées où la personne a l'impression erronée de les avoir produit.

a. Manipulation de texte

Le plagiat peut être caché en manipulant le texte et en changeant son apparence, par le biais de paraphrase[6] lexicale et syntaxique. Même si l'auteur ne copie pas directement l'œuvre originale et qu'une grande partie du texte soit modifiée, il est considéré comme plagiat sémantique [5].

b. Adoption d'idée

L'adoption d'une idée (une théorie, une conclusion, une hypothèse…etc.) est l'appropriation en totalité ou en partie d'une idée, avec des modifications superficielles sans donner de crédit à son auteur. Le plagiat d'idées peut être classé en trois types avec des limites floues : le sens sémantique, l'importance de la section, et le plagiat basé sur le contexte.

Le plagiat d'idées basé sur la sémantique peut être commis en paraphrasant, en résumant et en traduisant le texte, il peut être remarqué à travers le sens sémantique de deux. Le plagiat des idées peut être vu à travers l'importance des différentes sections dans les documents, par exemple plagier des segments substantiels d'un travail scientifique, tels que les résultats, les discussions et les conclusions. Même si toute une grande partie du texte sera réécrite, mais la structure du document maintient la séquence logique des idées de la source, cela entre dans le plagiat d'idées via l'adaptation du contexte.

c. Traduction – Cross Langage Plagiat

Le plagiat peut également être commis en traduisant partiellement ou totalement un texte d'une langue à une autre, en utilisant la traduction automatique (le traducteur Google, Bing…etc.), la traduction manuelle (par des personnes qui parlent les deux langues) ou bien combiner les deux afin d'améliorer la qualité de la langue avec l'introduction des vérificateurs d'orthographe et d'autres manipulations de texte. Parfois, pour dissimuler le plagiat, les plagiaires pourraient utiliser le retour plagiat BTP (Back Translated Plagiarism)[6] en traduisant automatiquement un texte d'une langue à une autre et le retraduire à la première.

III. Détection du plagiat

Les spécialistes – de chaque domaine – sont souvent capables de détecter des différents types de plagiat, soit en détectant des incohérences de texte, ou en ressemblant à du matériel préalablement consulté en utilisant leur propre expertise. Néanmoins, la grande quantité de textes sources potentiels disponibles de nos jours rend la détection manuelle du

[6] Paraphraser est une technique pour modifier la structure d'une phrase originale en changeant la structure de la phrase ou en remplaçant certains des mots originaux par ses synonymes. (Sans citation ou guillemets, il est considéré comme du plagiat.)

plagiat impossible. Afin d'aider les gens à découvrir le plagiat, différents modèles / solutions informatisés, regroupés sous le nom général de détection automatique du plagiat, ont été développés. Le but de ces modèles est de détecter d'éventuels cas de plagiat dans un document douteux et, si possible, de fournir la source alléguée pour faciliter le jugement de la qualité des différents travaux académiques.

Depuis les années 1970, la détection du plagiat a commencé par la découverte des clones de code et la détection des abus de logiciels en Pascal et C [7] [8]. La détection du plagiat dans les langues naturelles a débuté depuis les années 1990, par des méthodes statistiques ou informatisées [9] [10].

Les algorithmes de détection du plagiat dans les langages naturels et les langages de programmation ont des différences remarquables. Le premier aborde les différentes caractéristiques textuelles, tandis que le second se concentre principalement sur le suivi des métriques, telles que le nombre de lignes, de variables, de déclarations, de sous-programmes, d'appels aux sous-programmes… etc.

La recherche sur la détection automatique du plagiat dans les langues naturelles a évolué activement, profitant des développements récents dans de domaines connexes comme la recherche d'information (IR), la Recherche d'Information en Langage Croisé (CLIR), traitement du langage naturel [11].

La détection du plagiat dans les langages naturels peut être classée en monolingue et multilingue en se basant sur l'homogénéité du langage ou l'hétérogénéité des documents comparés.

III.1 Détection du plagiat monolingue

La détection du plagiat monolingue concerne l'identification automatique et l'extraction du plagiat dans un cadre de langue homogène. La plupart des systèmes de détection de plagiat ont été développés pour la détection monolingue, qui est divisé en deux tâches : extrinsèques et intrinsèques.

III.1.1 Détection de Plagiat extrinsèque

La détection de plagiat extrinsèque vise à trouver les passages suspects plagiés et leurs contreparties correspondantes dans un ensemble de collections sources. Chaque document suspect est comparé aux sources disponibles pour détecter si elles sont copiées ou manipulées. Ces collections peuvent être un corpus, certaines bibliothèques ou des bases de données spécifiques à certains domaines ou peut-être tout le Web [12] [13].

La détection de plagiat extrinsèque devient inefficace lorsqu'on n'a pas accès aux documents sources du plagiat dans le corpus fouillé. Cependant, il existe un autre type de détection, la détection intrinsèque qui exploite des données extraites de l'intérieur du même document.

III.1.2 Détection de Plagiat intrinsèque

La détection de plagiat intrinsèque et la vérification d'auteur sont des tâches similaires où le style d'écriture est quantifié et la complexité des caractéristiques est analysée.

Dans un problème de vérification d'auteur on analyse un document dont la paternité est douteuse par rapport à un ensemble d'exemples d'écriture d'un auteur spécifique. Pour la détection intrinsèque du plagiat le but est d'identifier des passages de texte qui s'écartent dans leur style d'écriture du reste du document. L'analyse du plagiat intrinsèque peut être considérée comme la généralisation de la vérification de l'auteur, car on analyse un seul document de manière isolée, et on est censé de trouver les sections suspectes [14].

III.2 Détection du plagiat multilingue

Le problème du plagiat multilingue (inter-langues/CLP ou plagiat traduit) a acquis une importance considérable ces derniers temps, puisque le contenu sémantique d'un document peut être facilement et discrètement plagié par l'utilisation de la traduction (humaine ou automatique).

Le plagiat multilingue est la réutilisation non reconnue d'un texte impliquant sa traduction d'une langue à une autre.

La détection de plagiat multilingue (CLPD) consiste à discriminer des textes sémantiquement similaires indépendamment des langues dans lesquelles ils sont écrits, lorsqu'aucune référence à la source d'origine n'est donnée [15].

IV. Méthodes de détection du plagiat

IV.1 Modèle de sélection des documents candidats

La détection du plagiat extrinsèque se déroule en plusieurs étapes (figure 1), la phase de sélection des documents candidats est une partie essentielle du processus, son but est d'identifier les documents sources pour chaque document suspect. Elle réduit l'espace de recherche pour l'étape d'analyse détaillée en délimitant le nombre total de documents à traiter issus des grandes collections de références, telles que celles accessibles via Internet.

La sélection des documents candidats a démontré une amélioration de l'efficacité des systèmes de détection du plagiat [16].

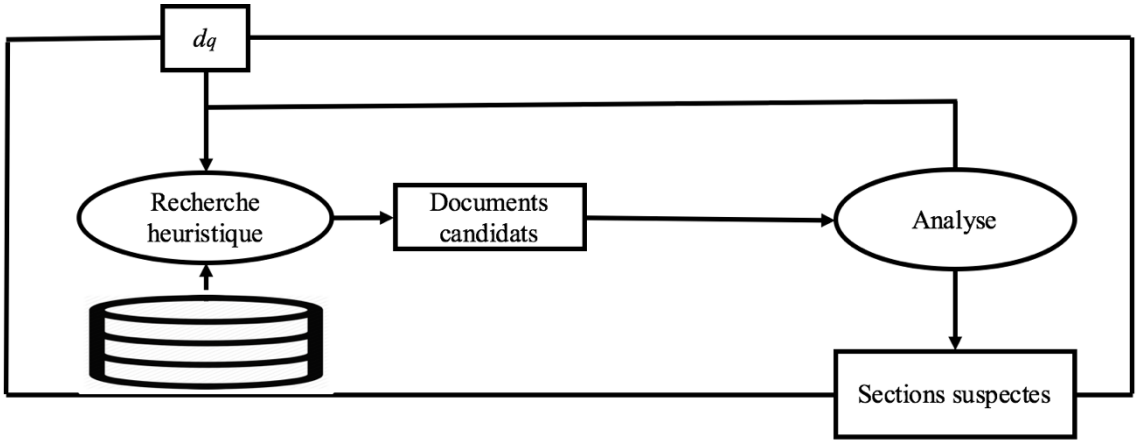

Figure 1: Processus de détection de plagiat

Appart la détection de plagiat intrinsèque, qui peut exploiter les résultats de la sélection des documents candidats d'une manière semi-automatique, en fournissant à un expert humain un ensemble de documents suffisamment petit pour être analysé manuellement. Alors que la détection de plagiat extrinsèque peut être considérée comme une tâche de Recherche d'Information (RI).

Les auteurs dans [11] ont considéré la recherche sur la détection de plagiat de texte analogue à la récupération de texte. Dans la récupération de texte, une liste de documents est récupérée sur la base d'une requête qui peut être petite (mots-clés), ou grande (document entier). De même, dans la détection de plagiat, une collection de documents source D est recherchée pour récupérer une liste de documents globalement similaires à un document suspect d_q.

Avec des grands ensembles de données [17], et l'utilisation du Web comme corpus, la recherche sur la détection du plagiat utilise des modèles RI traditionnels, afin de récupérer une liste relativement réduite de documents candidats pour chaque document suspect avant d'appliquer une analyse localement exhaustive.

IV.1.1 Modèles de recherches d'informations

Les modèles de recherches d'informations monolingues ont une large théorie et recherche [18] [19] [20], depuis la portée de notre travail est au-delà de la théorie RI, nous nous concentrons sur les modèles appliqués comme une étape préalable à la détection du plagiat.

Le modèle booléen [20] est parmi les premiers modèles utilisés en RI, la collection source est représentée comme un vecteur binaire de termes, et les termes de la requête sont associés

aux opérateurs booléens. Un terme est présent s'il se produit au moins une fois dans la représentation du document, ou absent s'il ne se produit pas du tout, c'est une recherche booléenne.

La recherche booléenne nécessite plusieurs intermédiaires et de gros moyens, elle est susceptible de rejeter des éléments utiles, ou de récupérer des éléments inutiles en réponse à des requêtes trop restrictives ou excessivement générales. Bien que sa simplicité, le modèle booléen n'a pas d'utilisation directe dans la détection de plagiat, car la requête est généralement de grande taille.

Les empreintes digitales (ou Shingling) [21] [22] et le hachage [23] [24] sont des modèles de récupération heuristique courants, où les documents sources D et le document de requête d_q sont divisés en petites unités (caractères, mots…) appelées empreintes digitales de longueur k. L' empreinte digitale est une autre nomination du modèle N-gramme :

- « Character N-gram » (CNG) est la forme la plus simple par laquelle un document *d* est représenté comme une séquence de caractères $d = \{(c_1, d), (c_2, d), ..., (c_n, d)\}$, où (c_i, d) fait référence au $i^{ème}$ caractère de d et $n = \|d\|$ est la longueur de la séquence *d* (en caractères).
- « Word N-gram » (WNG) le n-gramme basé sur les mots, représente *d* comme une collection de mots $d = \{(w_1, d), (w_2, d), ..., (w_n, d)\}$, où (w_i, d) se réfère au $i^{ème}$ mot en *d* et $n = \|d\|$ est la longueur de *d* (en mots) en ignorant les limites structurelles. WNG peut être construite en utilisant des bigrammes (mot 2 grammes), des trigrammes (mot-3 grammes) ou plus.

Le modèle basé sur le hachage utilise une fonction pour transférer des empreintes digitales dans des valeurs de hachage qui peuvent être triées et comparées à d'autres documents. La liste des empreintes digitales uniques de chaque document est considérée comme son vecteur.

Le Modèle de l'Espace Vectoriel (VSM) [25] est un modèle de récupération qui représente des documents textes dans un espace vectoriel. Le VSM tient compte du schéma de pondération de la fréquence des termes et de la fréquence des documents inverse (IDF-TF). La similarité entre les vecteurs pondérés de deux documents est calculée en utilisant des métriques de similarité vectorielle.

Plusieurs travaux de recherche ont utilisé VSM pour la recherche de candidats à partir de collections sources. Dans [26], les auteurs ont utilisé VSM à 1-gramme (W1G) et la similarité cosinus, dans [27] Grozea et al. ont opté pour le VSM de caractères de 16-grammes et la similarité cosinus, et dans [28] les auteurs ont utilisé VSM de 8-grammes et une mesure de distance personnalisée (équation 1).

Etant donné deux textes x et y leur distance n-gram est :

$$d_n(x,y) = \frac{1}{|D_n(x)| + |D_n(y)|} \sum_{w \in D_n(x) \cup D_n(y)} \left(\frac{f_x(w) - f_y(w)}{f_x(w) + f_y(w)}\right) \quad (1)$$

Avec

- $f_x(w)$ = fréquence du (caractère) n-gram de w dans x ;
- $D_n(x)$ = Ensemble de tous les n-grammes avec non-zéro fréquence dans x ;

Avec des grandes collections de données, les vecteurs à terme du VSM sont de taille importante, par conséquent l'indexation sémantique latente (LSI) [29] a été développée. Le schéma de pondération LSI est basé sur la réduction du VSM d'origine (c'est-à-dire des vecteurs de pondération TF-IDF) en utilisant une décomposition de valeur singulière (SVD). LSI a été utilisé pour coder la sémantique [30], pour élargir le vocabulaire en incorporant des dictionnaires de thésaurus [31] et dans [32] [33], pour la recherche de candidats et détection de plagiat.

Les plagiaires essayent toujours de cacher et d'ambigüer le plagiat commis, par conséquent de nombreux documents globalement similaires peuvent ne pas contenir de sections (paragraphes ou phrases) explicitement plagiés. Le VSM et LSI ne prennent en compte que la similitude globale des documents, Zhang et al. [34] ont donc proposé l'incorporation de caractéristiques structurales (document-paragraphe-phrase) dans la phase de la recherche des candidats deux méthodes de récupération ont été utilisées : l'appariement multiniveau basé sur l'histogramme (MLMH) et la correspondance à plusieurs niveaux (MLMS).

Dans la MLMH, la similarité globale au niveau du document et la similarité locale au niveau du paragraphe sont hybridées en une seule mesure, où chaque similarité est obtenue en faisant correspondre les histogrammes de mots de leurs représentants (document ou paragraphe) ; alors que la MLMS ajoute un paramètre de poids aux mots-histogrammes afin de considérer

ce que l'on appelle la capacité d'information, ou la proportion de mots dans le vecteur d'histogramme.

Fuzzy retrieval [35] [36] a été développé pour généraliser le modèle booléen en considérant une pertinence partielle entre la requête et la collecte de données. La théorie des ensembles flous (Zadeh 1965) traitent de la représentation de classes dont les limites sont mal définies [37]. Les ensembles flous sont reconnus comme un outil majeur en ingénierie de l'information dans le but de combler l'écart entre les connaissances formalisées d'origine humaine et les données numériques. Chaque élément de la classe est associé à une fonction d'appartenance qui définit le degré d'appartenance de l'élément dans la classe.

Dans [38], le modèle IR de l'ensemble flou a été appliqué pour extraire les documents qui partagent des instructions similaires, mais pas nécessairement les mêmes, au-dessus d'une valeur de seuil. Dans de nombreuses approches de représentation floue, la fonction TF-IDF du modèle vectoriel pondéré est utilisée comme fonction d'appartenance floue.

IV.1.2 Techniques de regroupement (Clustering)

Le Clustering désigne le regroupement des documents similaires entre eux et différents des documents appartenant à d'autres groupes. Les algorithmes de clustering cherche à segmenter l'ensemble des données en sous-groupes relativement homogènes à l'aide de mesures de distances, y compris le clustering plat et hiérarchique [20].

Les auteurs dans [39] ont introduit des techniques pour développer des clusters et des caractérisations de cluster en utilisant le point de vue de l'utilisateur, ce dernier est obtenu à travers une interview structurée basée sur une technique d'acquisition de connaissances. Dans [40], les auteurs proposent des modèles pour la récupération par clusters pour des collections de taille réaliste de manière entièrement automatique.

Kohonen [41] présente une forme de réseaux neuronaux non supervisés SOM (Self-Organizing Map), et montre des caractéristiques intéressantes d'une collecte de données telles que l'auto-organisation et l'apprentissage compétitif. SOM a été utilisé pour la projection de fonctionnalités, le regroupement de documents et la visualisation de cluster [42][43]. WEBSOM [44] a étudié l'utilisation de SOM dans le regroupement et la classification de grandes collections basées sur des histogrammes de mots statistiques, ce qui a permis de réduire les caractéristiques de grande dimension à des cartes bidimensionnelles. CHECK [45] est un système de détection du plagiat qui utilise le clustering pour trouver des documents

similaires, les documents du même groupe ont été comparés jusqu'à ce que deux paragraphes similaires soient trouvés.

Les techniques de regroupements peuvent être utilisées dans la phase de recherche de candidats. Pourtant, elles ne présentent pas un outil performant pour juger du plagiat.

IV.1.3 Modèles de récupération multilingue (CLIR)

La recherche d'information inter-linguistique (Cross Lingual Information Retrieval CLIR) désigne les activités de recherche d'informations dans lesquelles la requête est posée dans une langue, et dont le système doit renvoyer des documents écrits dans une deuxième langue. Un scénario qui aurait pu être considéré improbable il y a une dizaine d'années, mais la croissance explosive du Web a brouillé les frontières nationales au point où on peut avoir intérêt à récupérer des documents dans une langue étrangère.

Ces dernières années, le problème de la récupération multilingue a bénéficié d'un intérêt marqué de la part de la communauté de la recherche. Alors, plusieurs techniques ont été proposées pour résoudre le problème [46][47][48][49]. La plupart de ces techniques s'articulent autour d'une idée commune : elles tentent de traduire la requête de la langue de l'utilisateur vers la langue des documents. Les auteurs dans [50] ont proposé une approche de l'IR inter-linguistique basée sur des modèles de Markov cachés. Le système estime la probabilité qu'une requête dans une langue puisse être générée à partir d'un document dans une autre langue.

Le modèle de récupération basé sur un dictionnaire bilingue a pu réduire la dégradation des performances due à l'ambiguïté de la traduction. Les expériences ont été réalisées en utilisant les ensembles de tests chinois TREC5 et TREC6 et l'ensemble de test espagnol TREC4. Alors que le travail présenté dans [51] a combiné l'alignement basé sur le modèle de Markov caché (HMM) et la Traduction Automatique Statistique (SMT).

BiPLS [52] (Bilingual Partial Least Squares), est un modèle bilingue de corrélation des moindres carrés partiels basé sur une corrélation de sujet bilingue. BiPLS est un modèle de sujet bilingue non probabiliste. Il traite deux documents alignés bilingues comme deux vues représentant le même objet sémantique. Le modèle n'extrait pas la structure latente de la combinaison des espaces documentaires originaux bilingues, mais construit un espace sémantique latent unique pour chaque langue, et intègre une relation sémantique entre les langues. BiPLS vise à maximiser la covariance des sujets correspondants. Il découvre plus la

structure sémantique latente des corpus parallèles, mais ne synthétise pas les informations de traduction automatique ou de dictionnaire bilingue.

Dans [53], les auteurs ont présenté un modèle d'apprentissage de la représentation des mots BWESG (Bilingual Word Embeddings Skip-Gram), qui repose sur l'induction de vecteurs de mots réels à densité réelle connus sous le nom d'inclusion de mots (Word Embeddings (WE)) à partir de données comparables.

Certaines études antérieures ont suggéré que la traduction de la requête n'est pas une approche efficace de l'IR inter-linguistique [54]. Enfaite [55] est un modèle formel d'extraction de l'information multilingue qui ne repose ni sur la traduction de la requête ni sur la traduction de documents. Il s'appuie sur les avancées récentes en modélisation linguistique pour estimer directement un modèle de sujet précis dans la langue cible, en commençant par une requête dans la langue source. Le modèle intègre des techniques populaires de désambiguïsation et d'expansion de requêtes.

IV.2 Méthodes d'analyse de la détection du plagiat

Les méthodes pour comparer, manipuler et évaluer les caractéristiques textuelles afin de trouver le plagiat peuvent être classées en six types. Les sections suivantes décriront chaque type en détail.

IV.2.1 Méthodes basées sur les caractères (CNG)

La majorité des algorithmes de détection de plagiat reposent sur des caractéristiques lexicales et sur des fonctionnalités de syntaxe pour comparer le document de requête d_q avec chaque document candidat ($d_x \epsilon D_x$).

Un n-gramme est une suite de n éléments construits à partir d'une séquence donnée. L'idée peut être reliée au travail de Claude Shannon[7] en théorie de l'information [56]. Son idée était : "*À partir d'une séquence de lettres donnée, il est possible d'obtenir la fonction de vraisemblance de l'apparition de la lettre suivante*". D'où, à partir d'un corpus d'apprentissage, il est facile de construire une distribution de probabilité pour la lettre suivante avec un historique de taille n. Cette modélisation correspond en fait à un modèle de Markov d'ordre n, où seules les n dernières observations sont utilisés pour la prédiction de la lettre suivante.

[7] Claude Elwood Shannon, né en 1916 et décédé en 2001, est un ingénieur en génie électrique et mathématicien américain. Il est l'un des fondateurs, de la théorie de l'information.

Pour un document d_q, l'ensemble des n-grammes qu'on peut générer est le résultat qu'on obtient en déplaçant une fenêtre de n cases sur le corps de texte. Faire correspondre les chaînes peut être exactes ou approximatives. La correspondance de chaîne exacte entre deux chaînes *x* et *y* signifie qu'elles ont exactement les mêmes caractères dans le même ordre. Par exemple, le caractère 5-grammes *x = "aabcc"* diffère de *y = "aacbc"*.

Différentes techniques de plagiat utilisent la correspondance exacte des chaînes, soit avec le caractère n-gramme ou mot n-gramme. Barròn-Cedeño et al. [57] ont utilisé le caractère 3-grammes. Grozea et al. [27] ont utilisé le caractère 16-grammes. Basile et al. [28] ont utilisé le mot 8-gram, et Kasprzak et al. [58] ont utilisé le mot 5-grammes. La correspondance approximative de chaînes montre que deux chaînes *A* et *B* sont similaires / dissemblables. Par exemple, le caractère 8-grammes *A = "aaabbbcc"* et *B = "aaabbbcd"* sont très similaires car toutes les lettres correspondent sauf la dernière. Les opérations possibles qui pourraient transférer une chaîne en une autre sont [59], [60]:

- L'insertion *(s, a)* : insertion de la lettre a dans la chaîne s.
- La suppression *(s, a)* : suppression de la lettre a de la chaîne s.
- La substitution ou le remplacement *(a, b)* : remplacer la lettre a par b.
- La transposition *(ab, ba)* : permuter les lettres adjacentes a et b.

De nombreuses métriques mesurent la distance entre les chaînes de différentes manières. La distance $d = (A, B)$ entre deux chaînes *A* et *B* est définie comme suit :

"*La « distance » entre deux séquences est définie comme la séquence minimale (c'est-à-dire la plus probable) d'opérations pour transformer l'une en l'autre.* "[59].

L'approximation de la correspondance de chaînes et les mesures de similarité ont été largement utilisées dans la détection du plagiat. Scherbinin et al. [61] ont utilisé la distance de Levenshtein pour comparer le mot n-gramme et combiner des grammes similaires adjacents en sections. Su et al. [62] ont utilisé la distance de Levenshtein et l'algorithme de Smith-Waterman[63] simplifié pour l'identification et la quantification des similitudes locales dans la détection de plagiat. Elhadi et al. [64] ont utilisé la distance LCS (la Plus Longue Séquence Commune) combinée à d'autres fonctions syntaxiques pour identifier localement des chaînes similaires. Le tableau suivant (Tableau1) donne des descriptions des exemples de métriques de similarité de chaînes (Hamming, levenshtein [61], [62] et la Plus Longue Séquence Commune (LCS) [64], [65]).

Métrique de similarité	Description	Exemples
Hamming	Définit le nombre de caractères différents entre deux chaînes x et y de longueur égale.	*A= "aaabbbccc"* *B= "aaabbbcc**d**"* *d (A, B) = 1*
Levenshtein	Définit la distance d'édition minimale qui transforme A en B. Les opérations d'édition incluent : - supprimer un caractère - insérer un caractère - remplacer un char par un autre	*A= "aaabbbccc"* *B= "aaabbbcc**d**"* *C= "aaabbbcc**df**"* *D= "aaabbbcc"* $d(A, B) = 1$ $d(A, C) = 2$ $d(A, D) = 1$
Plus longue séquence commune (LCS)	Mesure la longueur de l'appariement le plus long des caractères qui peuvent être faits entre A et B par rapport à l'ordre des caractères. - permet des insertions, - permet des suppressions	*A = "**aaabbbcc**c"* *B = "**aaabbbcc**d"* $d(A, B) = 8$

Tableau 1: Exemples de métriques de similarité de chaînes

L'intérêt d'utiliser des n-grammes de caractères vient des possibilités qu'ils offrent, en particulier dans le cas des langues autres que l'anglais [15], [66]. Il fournit un moyen de substitution pour normaliser les formes de mots et il ne repose pas sur un traitement spécifique à une langue. Les n-grammes capturent automatiquement les racines des mots les plus fréquents [67] et opèrent indépendamment des langues[68], plus une tolérance aux fautes d'orthographes et aux manques des informations et des ressources linguistiques. Dans [69] on montre que des systèmes de recherches documentaires basés sur les n-grammes ont gardé leurs performances malgré des taux de déformations de 30%, situation dans laquelle aucun système basé sur les mots ne peut fonctionner correctement.

IV.2.2 Méthodes vectorielles (VEC)

Les caractéristiques lexicales et syntaxiques peuvent être comparées en tant que vecteurs de termes plutôt que de chaînes. La similarité peut être calculée en utilisant des coefficients de similarité vectorielle [60]. Autrement dit, le mot n-gramme est représenté comme un vecteur de *n* termes. Les phrases et les segments sont présentés comme des vecteurs de mots ou des vecteurs de caractères. La similarité peut être évaluée en utilisant l'appariement, Jaccard (ou Tanimoto), Dice, chevauchement (overlap), cosinus, euclidien ou des coefficients de Manhattan (Tableau 2).

Pouliquen et al. [70] proposent un système qui identifie les traductions et les documents très similaires parmi un grand nombre de candidats. Les contenus des documents sont représentés comme des vecteurs de termes d'un thesaurus multilingue. La mesure de similarité pour les documents est la même, indépendamment du langage du document.

Ekbal et al. [71] ont utilisé le modèle d'espace vectoriel traditionnel (VSM) et applique ensuite une technique basée sur un graphique pour trouver des passages similaires dans des documents suspects et des documents sources sélectionnés. De même, Bao et al. [72] ont utilisé le modèle vectoriel pour détecter la similarité dans le système de détection de copie de biens numériques (The Copying Detection System of Digital Goods (CDSDG)).

Murugesan et al. [73] ont utilisé la similarité cosinus sur l'ensemble du document ou sur des fragments de documents pour permettre la détection globale ou partielle du plagiat. En raison de sa simplicité, l'utilisation du cosinus avec d'autres métriques de similarité était efficace pour la détection du plagiat dans les environnements sécurisés.

Zhang et al. [34] ont utilisé la distance cosinus exponentielle comme mesure de la similarité des documents qui converge globalement vers 0 pour les petites distances et vers 1 pour les grandes distances. Pour englober des déclarations qui sont localement similaires dans la décision finale de détection de plagiat, le coefficient de Jaccard a été utilisé pour estimer le chevauchement entre les phrases.

Barrón-Cedeño et al. [74] ont estimé la similarité entre les termes n-grammes de différentes longueurs $n = \{1,2, ..., 6\}$ en utilisant le coefficient de Jaccard, et dans [66] ils ont utilisé la mesure de confinement pour comparer des morceaux de documents basé sur le mot n-gramme $n = \{2,3\}$. Les vecteurs résultants des mots n-grammes et la similarité du confinement ont été utilisé pour montrer le degré de chevauchement entre deux fragments. De même, Lyon et al. [75] ont exploité l'utilisation de mots 3-grammes pour mesurer la similarité de courts passages dans de grandes collections de documents. White et al. [76] ont utilisé le coefficient correspondant avec un seuil pour noter des énoncés similaires.

IV.2.3 Méthodes Basées sur la Syntaxe (SYN)

Certains travaux de recherche ont utilisé des caractéristiques syntaxiques pour évaluer la similarité du texte et la détection du plagiat.

Les approches dans [64], [85], [86] réduisent le texte en un plus petit ensemble de balises syntaxiques en utilisant la majeure partie du contenu, contrairement à de nombreux systèmes de détection de plagiat existants qui réduisent le texte en un ensemble de jetons en supprimant les mots d'arrêt et les bourrages.

Métrique de Similarité Vectorielle	**Description**	**Équation**	**Variation**
coefficient d'appariement [77]	Similaire à la distance de Hamming mais entre des vecteurs de longueur égale	$M(x,y) = \|x\| - \|x \cap y\|$	0 à $\|x\|$ où $\|x\| = \|y\|$
Coefficient de Jaccard [78][74], [79], [80]	Définit le nombre de termes partagés par rapport au nombre total de termes.	$J(x,y) = \frac{\|x \cap y\|}{\|x \cup y\|}$	0 à 1
Coefficient de dés	Similaire à Jaccard mais réduit l'effet des termes partagés entre les vecteurs.	$D(x,y) = \frac{2\|x \cap y\|}{\|x \cup y\|}$	0 à 2
Coefficient de chevauchement (ou de confinement) [81]	Si v1 est un sous-ensemble de v2 ou l'inverse, alors la similitude est une correspondance complète	$O(x,y) = \frac{\|x \cap y\|}{min(\|x\|,\|y\|)}$	0 à 1
Coefficient de cosinus [80], [82]–[84]	Trouve l'angle cosinusoïdal entre deux vecteurs	$\cos(x,y) = \frac{\sum_i (x_i, y_i)}{\sqrt{\sum_i (x_i)^2}\sqrt{\sum_i (y_i)^2}}$	0 à 1
Distance euclidienne	Mesure la distance géométrique entre deux vecteurs	$Ec(x,y) = \sqrt{\sum_i \|x_i - y_i\|^2}$	0 à ∞
Distance euclidienne au carré	Passe progressivement plus de poids sur les vecteurs plus éloignés	$SEc(x,y) = \sum_i (x_i - y_i)^2$	0 à ∞
Distance de Manhattan	Mesure la différence moyenne entre les dimensions et les rendements, similaire à la distance euclidienne simple	$Manh(x,y) = \sum_i \|x_i - y_i\|^2$	0 à ∞

Tableau 2: Métriques de Similarité Vectorielle

Les différentes études [64], [85], [86] se sont basés sur l'intuition que des documents similaires (copies) auraient une structure syntaxique similaire (séquence d'étiquettes morpho-syntaxique POS-tagging). Ils ont utilisé des caractéristiques de balises POS-tagging suivies d'autres métriques de similarité de chaînes dans le calcul de la similarité entre les textes.

Elhadi et al. [85] ont proposé une approche qui consiste à utiliser des étiquettes POS-tagging pour représenter une structure de texte comme base de comparaison et d'analyse. Les étiquettes POS ont été utilisées pour le classement des documents. La méthodologie dans [85] a été amélioré en [64] en utilisant la plus longue sous-séquence commune (LCS) pour calculer la similarité entre les documents et les classer en fonction des documents extraits les plus pertinents.

Vani et al. [86] ont combiné différentes métriques de similarité pour la détection de plagiat extrinsèque en incorporant des informations d'étiquette POS-tagging, et en utilisant Stanford Tagger [87] pour évaluer l'impact de l'utilisation d'étiquetage morpho-syntaxique dans le modèle de détection de plagiat.

IV.2.4 Méthodes basées sur la sémantique (SEM)

Deux phrases peuvent être sémantiquement identiques mais différent dans leur structure ou mots. Une phrase peut être traitée comme un groupe de mots disposés dans un ordre particulier.

Plusieurs approches sémantiques ont été proposé, mais dû aux difficultés de représentation de la sémantique et à la complexité des algorithmes représentatifs, ils ont eu moins d'attention.

Trois facteurs associés à la hiérarchie taxonomique de l'ontologie peuvent affecter la mesure de la distance sémantique : La longueur de chemin, la profondeur et la densité local. La densité de deux concepts *C1* et *C2* est le nombre de fils des concepts qui appartiennent au chemin le plus court de la racine à la sous-séquence commune la plus spécifique des deux concepts *C1* et *C2*.

Les mesures de similarité et la taxonomie sont liées par des relations taxonomiques, c'est-à-dire la position des concepts dans la taxonomie, le nombre de liens hiérarchiques et le contenu informationnel des concepts sont considérés. Les mesures sémantiques proposées sont classées en trois catégories principales :

a. Mesures basées sur la structure

Les mesures basées sur la structure –ou sur le comptage des arêtes– calculent la similarité sémantique en se basant sur la structure de la hiérarchie de l'ontologie (IS-A, PartOf), c'est-à-dire calculent la longueur du chemin reliant les termes et la position des termes dans la taxonomie. Ainsi, plus les deux concepts sont similaires, plus il y a de liens entre les concepts et plus ils sont étroitement liés [88], [89].

Soient *C1* et *C2* deux concepts pour lesquels on calcule la mesure de similarité dans la structure hiérarchique.

- **Le plus court chemin**

C'est une mesure simple [88], puissante et principalement conçue pour fonctionner avec les hiérarchies.

$$Sim(C1, C2) = 2 * Max(C1, C2) - SP \qquad (2)$$

Où

Max est la longueur de chemin maximale entre *C1* et *C2* dans la taxonomie

SP (Shortest path) est le court chemin reliant *C1* au *C2*.

- **Liens pondérés :**

Cette mesure est une extension de la mesure ci-dessus [89]. Elle propose des liens pondérés pour calculer la similarité entre deux concepts. Deux facteurs qui affectent le poids d'un lien: La profondeur d'une hiérarchie spécifique et la force de la connotation[8] entre les nœuds parents et enfants. Par la suite, la distance entre deux concepts est obtenue en additionnant les poids des liens traversés au lieu de les compter.

- **Hirst et St-Onge (HSO)**

La mesure HSO [90] calcule la relation entre les concepts en utilisant la distance du trajet entre les nœuds des concepts, le nombre de changements dans la direction du chemin reliant les deux concepts et l'admissibilité[9] du chemin. S'il y a une relation étroite entre les significations de deux concepts, alors on dit que les concepts sont sémantiquement liés les uns aux autres [91]. Soit *d* le nombre de changements de direction dans le chemin qui relie deux concepts *C1* et *C2*. *C et k* sont des constantes dont les valeurs sont dérivées par des expériences. La similarité HSO est :

$$Sim_{HSO}(C1, C2) = C - SP - k * d \quad (3)$$

- **Wu & Palmer (WuP)**

La mesure WuP [92] rapporte la profondeur des synsets des mots dans la taxonomie DAG (Directed-Acyclic-Graph pour WordNet [93]) (Figure 2) et la profondeur de leur concept commun le plus spécifique (LCS).

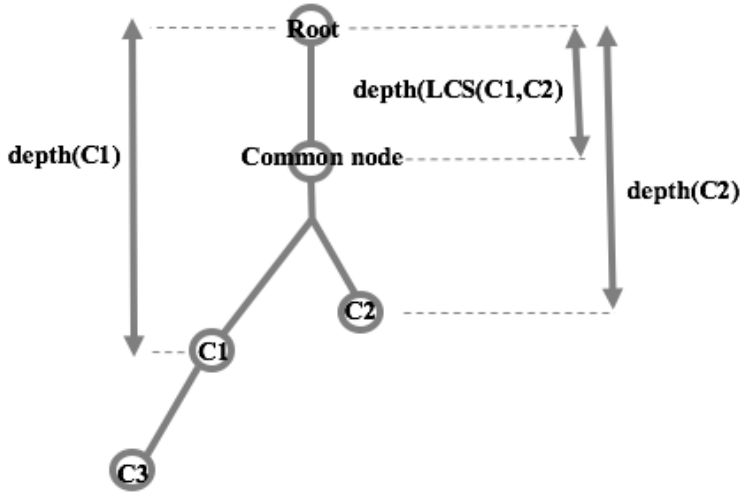

Figure 2: Diagramme acyclique dirigé (DAG) pour WordNet

[8] La connotation d'un terme est la liste des conditions d'adhésion pour la dénotation. La dénotation d'un terme est la classe de choses à laquelle le terme s'applique correctement.

[9] Un chemin admissible est un chemin qui ne s'écarte pas de la signification du concept source et doit donc être pris en compte dans le calcul de la parenté.

Plusieurs parents peuvent être partagés par *C1* et *C2* par plusieurs chemins. Le concept commun le plus spécifique est l'ancêtre commun le plus proche *C* (le parent commun lié au nombre minimum de liens IS-A avec les concepts *C1* et *C2*).

$$SIM_{WuP}(C1, C2) = 2\times\frac{depth(LCS(C1, C2))}{depth(C1) + depth(C2)} \quad (4)$$

Où $depth(C1)$ et $depth(C2)$ sont les profondeurs des deux synsets dans WordNet [94] (c-à-d la distance (nombre de liens IS-A) qui sépare, respectivement, le concept *C1* et *C2* du concept commun spécifique). $depth(LCS(C1, C2))$ est la profondeur de la *LCS* (distance) qui sépare l'ancêtre commun le plus proche de *C1* et *C2* du nœud racine.

- **Li et al.**

La similarité de Li et al. [95] combine la plus courte longueur de chemin *(SP)* entre deux concepts *C1* et *C2*, et la profondeur dans la taxonomie *(N)* du concept commun *C* le plus spécifique, dans une fonction non-linéaire.

$$Sim_{li}(C1, C2) = e^{-\infty * SP} * \frac{e^{\beta * N} - e^{-\beta * N}}{e^{\beta * N} + e^{-\beta * N}} \quad (5)$$

Où α>=0 et β>=0 sont les paramètres de mise à l'échelle de la contribution de la longueur du chemin le plus court et la profondeur respectivement. Cette mesure se situe entre 1 et 0. D'après [95] les paramètres optimaux sont *α=0,2 et β=0,6.*

- **Leacock et Chodorow**

La mesure de similarité (LCH) [96] est:

$$Sim_{LC}(C1, C2) = -Log(\frac{length}{2.D}) \quad (6)$$

Où *length* est la longueur du plus court chemin entre les deux concepts (en utilisant le comptage de nœuds) et *D* est la profondeur maximale de la taxonomie.

	Les sources de données	Sémantiques	Facteurs		
			SP	Densité Conceptuelle	N
Le plus court chemin	Ontologie	Distance	✓		
Liens pondérés	Ontologie	liens pondérés	✓	✓	
Hirst et St-Onge	Ontologie	parenté	✓		
Wu et Palmer	Ontologie	similarité	✓	✓	✓
Li et al.	Ontologie	similarité	✓		✓
Leachock et Chodorow	Ontologie	similarité	✓		

Tableau 3: Typologie des mesures sémantiques basées sur la structure

Les mesures précédemment décrites sont basées uniquement sur des liens hiérarchiques (IS-A) entre concepts, en tenant compte du fait que les liens dans la hiérarchie représentent des distances. Le tableau ci-dessus (Tableau 3) donne une topologie de comparaison des approches décrites.

b. Mesures du Contenu de l'Information

Les mesures basées sur le Contenu de l'Information *(IC)* sont celles qui utilisent le contenu informationnel des concepts pour mesurer la similarité sémantique entre deux concepts. La valeur du contenu d'information d'un concept est calculée en fonction de la fréquence du terme dans une collection de documents donnée.

Un nombre important de mesures de similarité sémantique utilisent le contenu informationnel du parent partagé entre deux termes *C1* et *C2* (équation 7), où *S(C1,C2)* est l'ensemble des concepts qui subsument *C1* et *C2*. Les deux concepts peuvent partager les parents par plusieurs chemins. Le minimum *p(C)* est utilisé lorsqu'il y a plus d'un parent partagé où *C* est le subsume le plus informatif (MIS) (Most informative subsume).

$$P_{MIS}(C1, C2) = min_{C \in S(C1,C2)}\{p(C)\} \quad (7)$$

- **Resnik**

Rensik et al. [97] utilisent le contenu d'information des parents partagés pour calculer la similarité. Deux concepts *C1* et *C2* sont plus similaires s'ils présentent une information plus partagée, cette dernière est indiquée par le contenu informationnel des concepts qui les subsument dans la taxonomie.

$$Sim_{Resnik}(C1, C2) = -\ln\left(p_{mis}(C1, C2)\right) \quad (8)$$

- **Lin et al.**

Lin et al. [98] [99] ont proposé une mesure basée sur une ontologie restreinte aux liens hiérarchiques et à un corpus. Cette similitude prend en compte les informations partagées par deux concepts comme Resnik, mais la différence entre eux est dans la définition. La définition contient les mêmes composants que la mesure Resnik mais la combinaison n'est pas une différence mais un rapport.

$$Sim_{Resnik}(C1, C2) = \frac{2 * \ln\left((p_{mis}(C1, C2)\right)}{\ln\big(p(C1)\big) + \ln\left(p(C2)\right)} \quad (9)$$

Par conséquent, l'utilisation de cette mesure pour comparer les termes d'une ontologie présente un meilleur classement de la similarité par rapport à la mesure Resnik.

- **Jiang et Conrath**

De la même manière que Resnik, les auteurs dans [100] ont utilisé un corpus en plus d'une ontologie hiérarchique (liens taxonomiques).La distance entre deux concepts *C1* et *C2* est la différence entre la somme du contenu informationnel des deux concepts et le contenu informationnel de leur subsumer le plus informatif.

$$Sim_{Jiang}(C1, C2) = 2 * \ln p_{mis}\,(C1, C2) - (\ln P(\,C1) + \ln P(C2)) \qquad (10)$$

	Source de données	Sémantiques	Facteurs		
			SP	Densité de concept	N
Resnik	Ontologie + Corpus	Similarité		✓	
Lord et al.	Ontologie + Corpus	Similarité		✓	
Lin et al.	Ontologie + Corpus	Similarité	✓	✓	✓
Jiang & contrath	Ontologie + Corpus	Distance	✓		✓

Tableau 4: Typologie des mesures sémantiques du contenu de l'information

c. Mesures basées sur les caractéristiques

Les mesures basées sur les caractéristiques supposent que chaque terme est décrit par un ensemble de termes indiquant ses propriétés ou caractéristiques. La mesure de similarité entre deux termes est définie en fonction de leurs propriétés (définitions ou «glosses» dans WordNet [94]) ou en fonction de leurs relations avec d'autres termes similaires dans la structure hiérarchique.

- **Tversky**

Tversky [101] prend en compte les caractéristiques des termes pour calculer la similarité entre différents concepts, en ignorant la position et le contenu informationnel des termes dans la taxonomie. Chaque terme devrait être décrit par un ensemble de mots indiquant ses caractéristiques.

$$Sim_{tvsk}(C1, C2) = \frac{|C1 \cap C2|}{|C1 \cap C2| + \alpha|C1 - C2| + (\alpha - 1)|C2 - C1|} \qquad (11)$$

Où *C1* et *C2* représentent les ensembles de description correspondants de deux termes. $\alpha \epsilon [0,1]$ est l'importance relative des caractéristiques non communes. La valeur de α augmente avec la communité et diminue avec la différence entre les deux concepts. La

détermination de α est basée sur l'observation que la similarité n'est pas nécessairement une relation symétrique.

- **X-Similarité**

Les fonctions X-similarité [102] proposent une correspondance entre les mots extraits de WordNet [94] en analysant les définitions des termes. Deux termes sont similaires si les mots et les concepts dans leurs voisinages (basés sur des relations sémantiques) sont lexicalement similaires.

Soit *A* et *B* deux synsets ou ensembles de descriptions de termes. Les termes au voisinage d'un terme ne présentent pas forcément un lien avec la même relation, la similarité est calculée par type de relation sémantique (par exemple, IS-A et Part-Of). La mesure est comme suit :

$$Sim_{xsim}(A,B) = \begin{cases} 1, if\ S_{synsets}(A,B) > 0 \\ max\{S_{neighb}(A,B), S_{descr}(A,B)\}, if\ S_{synsets}(A,B) = 0 \end{cases} \quad (12)$$

Soit *i* un type de relation, la similitude pour les voisins sémantiques est formulée comme suit :

$$S_{neight}(A,B) = \max_{i \subset SR} \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (13)$$

IV.2.5 Méthodes floues (FUZZY)

Dans les méthodes floues : « *Faire correspondre deux phrases peut être approximatif ou vague, ce qui peut être modélisé en considérant que chaque mot d'une phrase est associé à un ensemble flou qui contient les mots ayant la même signification. Et il y a un degré de similarité (généralement moins de 1) entre les mots (dans une phrase) et l'ensemble flou* » *Yerra et Ng (2005)(p:563)* [38].

Dans [38], une matrice de corrélation à terme est construite avec des mots et leurs facteurs de corrélation correspondants pour mesurer les degrés de similarité (degré d'appartenance entre 0 et 1) deux phrases différentes. Le facteur de corrélation terme-à-terme, $F_{i,j}$ définit une similitude floue entre deux mots w_i et w_j comme suit :

$$F_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \quad (14)$$

Où $n_{i,j}$ est le nombre de documents dans une collection avec les deux mots w_i et w_j, et n_i (n_j respectivement) est le nombre de documents avec w_i (w_j respectivement) dans la collection.

Dans [12], le facteur de corrélation à terme a été remplacé par une similitude floue définie comme suit:

$$F_{i,j} = \begin{cases} 1 \; si \; w_i = w_j \\ 0.5 \; si \; w_i = synset(w_j) \\ 0 \; sinon \end{cases} \quad (15)$$

Le synset du mot w_i a été extrait en utilisant WordNet [94]. Pour obtenir le degré de similarité entre deux phrases S_i et S_j, le facteur de corrélation mot-phrase $\mu_{i,j}$ de w_i dans S_i avec tous les mots de S_j est calculé [38]:

$$\mu_{i,j} = 1 - \prod_{w_k \in S_j} (1 - F_{i,k}) \quad (16)$$

Où w_k est chaque mot dans S_j et $F_{i,k}$ est le facteur de corrélation entre w_i et w_k.

Sur la base de la valeur μ de chaque mot d'une phrase S_i, qui est calculée par rapport à la phrase S_j, le degré de similarité de S_i par rapport à S_j peut être défini comme suit [38]:

$$Sim(S_i, S_j) = (\mu_{1,j} + \mu_{2,j} + \cdots + \mu_{n,j})/n \quad (17)$$

Où w_k $(1 \leq k \leq n)$ est un mot dans S_i, et n est le nombre total de mots dans S_i. $Sim(S_i, S_j)$ est une valeur normalisée. De même, $Sim(S_j, S_i)$ est le degré de similarité de S_j par rapport à S_i.

En utilisant l'équation telle que définie ci-dessus, deux phrases S_i et S_j doivent être traitées de la même manière, selon l'équation suivante [38]:

$$EQ(S_i, S_j) = \begin{cases} 1 \; si \; \min(Sim(S_i, S_j), Sim(S_j, S_i) \geq p \wedge |Sim(S_i, S_j) - Sim(S_j, S_i)| \leq v \\ 0 \; Sinon \end{cases} \quad (18)$$

Où p: $(= 0{,}825)$ valeur de seuil de permission et v : $(=0{,}15)$ valeur de seuil de variation [38]. Le seuil permission est la similitude minimale entre deux phrases S_i et S_j, qui est partiellement utilisé pour déterminer si les deux phrases doivent être traités comme égaux. D'autre part, la valeur de seuil de variation est utilisée pour diminuer les faux positifs (les déclarations qui sont traitées comme égales mais ne le sont pas) et les faux négatifs (les déclarations qui sont égales mais traitées comme différentes).

En continuant, [103] ont utilisé l'approche IR floue pour déterminer le degré de la similarité entre deux documents Web et regrouper des documents Web ayant un contenu similaire mais pas nécessairement le même. En outre, dans [104] les auteurs ont adapté le modèle IR flou pour l'utiliser avec des scripts arabes.

Dans [105] les auteurs ont proposé une approche basée sur un système d'inférence floue et l'étiquetage de rôle sémantique (FIS-SRL). Cette approche analyse et compare le texte basé sur une allocation sémantique pour chaque terme à l'intérieur de la phrase. La méthode proposée génère des arguments sémantiques pour chaque phrase, puis choisit pour chaque argument généré par FIS les arguments importants. FIS sélectionne les arguments les plus importants et utilise les résultats dans le processus de calcul de similarité. Les auteurs ont évalué la méthode en utilisant le corpus PAN-09.

Le travail dans [106] apporte des améliorations en efficacité et précision au système proposée par Alzahrani et al [12] en utilisant différentes méthodes de prétraitement basées sur les techniques PNL, et des mesures de similitude sémantique floue. Le système a été évalué à l'aide du corpus PAN-2012.

IV.2.6 Méthodes interlingual (CROSS)

Les méthodes inter-lingues sont basées sur la mesure de la similarité entre les sections du document suspect d_q et les sections du document candidat d_x à l'aide des fonctionnalités de textes inter-langues. Les méthodes comprennent: (i) des méthodes basées sur la syntaxe multilingue (lexique), (ii) des méthodes basées sur des dictionnaires multilingues [107], et (iii) des méthodes basées sur la sémantique interlinguale utilisant des corpus comparables ou alignés qui exploite les corrélations de vocabulaire[15], [108].

a. Méthodes basés sur le lexique

Ces méthodes s'appuient sur les similitudes lexicales entre les langues (par exemple anglais-français) et sur l'influence linguistique (par exemple anglais-espagnol). Les similitudes entre les mots dans différentes langues peuvent être reflétées lors de la composition de termes courts. Les premiers modèles de similarité de ce genre sont Cognateness [109] basé sur des préfixes et d'autres jetons, et Dot-Plot [110] basé sur le modèle CL-CNG (caractères 4-grammes). Bien qu'ils aient été initialement proposés pour aligner les bis-textes, ces modèles sont utiles pour détecter la réutilisation dans plusieurs langues [15], avec certaines limitations [57].

b. Méthodes à base de thésaurus

Les méthodes à base de thésaurus mappent des mots ou des concepts, tels que des entités nommées, dans un espace de représentation commun au moyen d'un thésaurus multilingue (par exemple Eurovoc [111] ou EuroWordnet [112]).

c. Méthodes à base de corpus parallèles

Ces méthodes sont formées sur des corpus parallèles, soit pour trouver des co-occurrences multilingues [113], soit pour obtenir des modules de traduction. Les principes et les ressources de la traduction automatique (MT) sont appliqués.

d. Méthodes de traduction automatique

Ces méthodes sont en vogue dans le CLPD. Pereira et al. [107], Kent et al. [114], Nawab et al. [115] et Oberreuter et al. [116] simplifient la tâche en la transformant en un problème monolingue. Le processus est le suivant :

(i) Un détecteur de langage est appliqué pour déterminer le langage le plus probable de d_q;
(ii) d_q est traduit s'il n'est pas écrit dans le langage de comparaison;
(iii) Une comparaison monolingue est effectuée entre d_q et d_x.

Les systèmes de Grman et Ravas [117] ont obtenu de bons résultats dans PAN-2011, parce que le même traducteur automatique (Google Translator) est utilisé à la fois pour la génération et la détection.

V. Conclusion

La détection de plagiat emploi diverses techniques de prétraitement afin d'améliorer la précision et/ou de diminuer le nombre de caractéristiques qui doivent être traitées. Nous avons effectué un prétraitement de tous nos documents d'entrainement et de test avant d'appliquer les approches proposées. Le prétraitement des documents consiste à nettoyer (Suppression des mots vides, ponctuation…etc.) et structurer (Tokenisation, lemmatisation, POS…etc.) les données contenues dans ces documents afin de les préparer à une future analyse de plagiat, cette étape est très cruciale pour filtrer les informations non pertinentes.

Chapitre2 Prétraitement du Langage Naturel

I. Introduction

Le plagiat écrit est un problème très répandu qui nécessite un nombre important d'opérations du prétraitement de texte pour être révéler. Le prétraitement du texte a une influence significative sur la performance de nombreuses tâches de traitement du langage naturel (NLP), y compris la détection du plagiat. La détection du plagiat est un domaine distinct qui mérite une attention particulière, car il nécessite d'appliquer un large éventail de techniques de prétraitement (par exemple : tokénisation, lemmatisation, Étiquetage morphosyntaxique, suppression de mots vides… etc.) (Figure 3). Divers prétraitements ont des effets différents, certains améliorent la précision, d'autres diminuent simplement les exigences de temps, et d'autres font les deux.

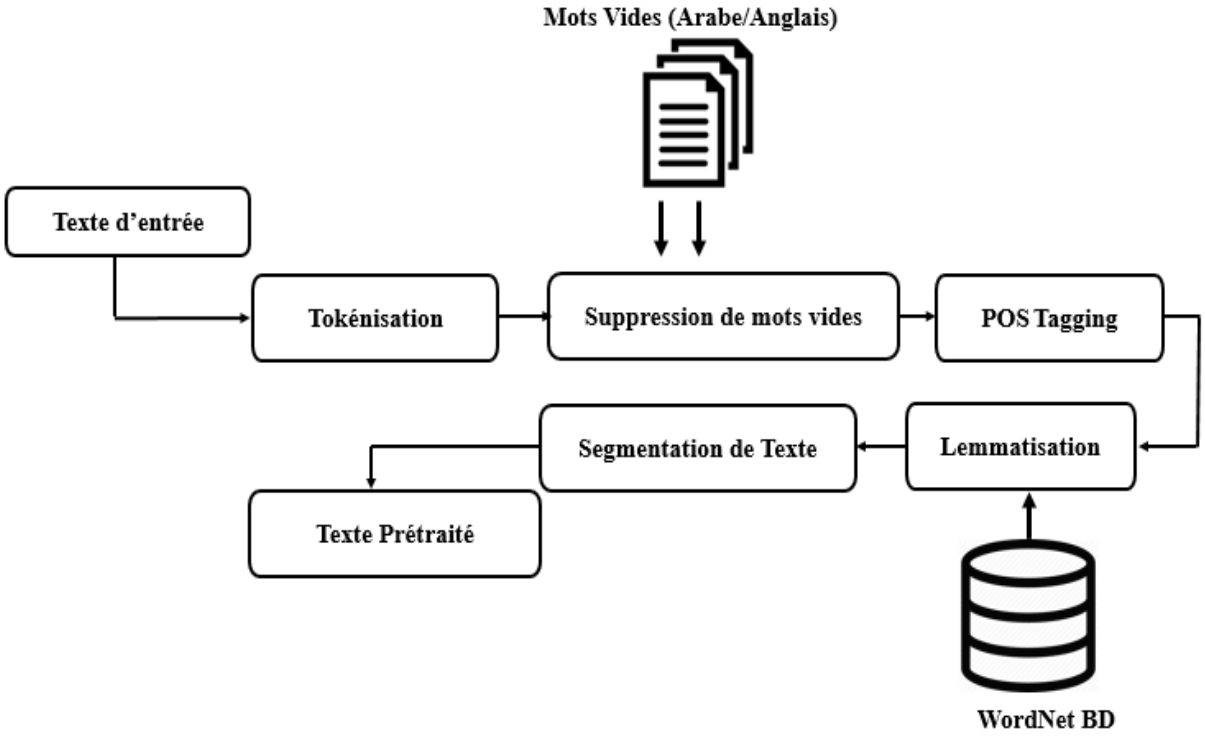

Figure 3: Schéma de prétraitement de texte pour

II. Techniques de prétraitement

II.1. Tokénisation

La notion de jeton (token) doit être définie avant n'importe quel traitement informatique. Le problème ne se limite pas à l'identification des chaînes délimitées des deux côtés par des espaces ou des signes de ponctuation. Différentes notions dépendent d'objectifs

différents, et souvent de contextes linguistiques différents, un jeton doit être linguistiquement significatif et méthodologiquement utile.

Le texte électronique est une séquence linéaire de symboles (caractères, mots…). Naturellement, la première étape dans la majorité des applications de traitement de texte consiste à segmenter le texte en unités linguistiques, ce processus est appelé tokénisation.

La tokénisation est généralement considérée comme facile par rapport à d'autres tâches en langage naturel, et l'une des tâches les plus inintéressantes (par exemple l'anglais). Cependant, les erreurs commises dans cette phase peuvent entrainer une influence grave sur des phases ultérieures et peuvent causer des problèmes sérieux.

Un autre défi pour la tokénisation est les "textes sales" [118]. Les textes extraits automatiquement des fichiers PDF, des champs de bases de données ou d'autres sources, peuvent contenir des jetons mal composés, des fautes d'orthographe ou des caractères inattendus.

La segmentation des mots peut sembler simple dans un langage qui sépare les mots par un caractère spécial "espace" (tokénisation d'un bas niveau). Cependant, ce n'est pas le cas pour toutes les langues. Un examen plus profond permettra de comprendre que la tokénisation avec l'espace tout seul n'est pas suffisant même pour l'anglais.

II.1.1 Tokénisation et mots composé

Par définition, « *un mot composé est une juxtaposition de deux lexèmes libres permettant d'en former un troisième qui soit un lemme (mot) à part entière et dont le sens ne se laisse pas forcément deviner par celui des deux constituants* » [119]. Il y en a plusieurs types :

- Le composé unifié
- Le composé à apostrophe
- Le composé à trait d'union
- Le composé détaché.

Déterminer si deux mots ou plus doivent être regroupés pour former un seul jeton serait une tâche de tokénisation de haut niveau. L'étiquetage morphosyntaxique traite généralement les mots coupés comme une unité syntaxique unique, il les préfèrent donc être tokeniser comme des jetons simples. Ce genre de segmentation est beaucoup plus motivé linguistiquement que le premier et nécessite relativement un traitement linguistique plus profond.

II.1.2 Manipulation des abréviations

Généralement, la ponctuation et la casse permet de séparer les phrases d'un texte, mais en Anglais, Arabe et d'autres langues, des complications peuvent être causées par les abréviations utilisant un point, ou les citations incluant des ponctuations à l'intérieur.

Par exemple la phrase « Dr. Mohamed a reçus un prix Nobel. » sans relever le défi posé par l'abréviation, va être délimitée en deux phrases : « Dr » et « Mohamed a reçus un prix Nobel ».

Malheureusement, les normes universellement acceptées pour les abréviations et les acronymes n'existent pas. Alors l'approche la plus adoptée pour la reconnaissance des abréviations est de maintenir une liste des abréviations les plus utilisées.

II.1.3 Expressions Numériques et Extraction d'Entité Nommée

Les adresses mail, les URL, les numéros de téléphone et les dates… peuvent produire beaucoup de confusion pour un tokenizer, car ils impliquent généralement une syntaxe alphanumérique et de ponctuation complexe.

Il est presque impossible de séparer la tokénisation de la reconnaissance d'entités nommées (noms de personnes, d'organisations ou lieux…). Il n'est vraiment pas possible de proposer un ensemble de règles génériques qui traitent tous les cas ambigus en Anglais, Arabe ou Français. L'approche la plus simple consiste généralement à utiliser des dictionnaires d'expression multi-mots et des bases de données (listes de noms d'organisations, de villes ou de pays…), en plus des systèmes d'étiquetages utilisent des grammaires formelles avec des modèles statistiques. Un tokenizer doit être alors personnalisé pour les données en question.

II.2. Lemmatisation

Pour des raisons grammaticales, les documents vont utiliser différentes formes d'un mot. Il existe des familles de mots alliés de manière dérivée avec des significations similaires, par exemple la *démocratie*, *démocratique*, et la *démocratisation*.

Dans de nombreuses situations, il semble utile que la recherche de l'un de ces mots renvoie des documents contenant un autre mot de l'ensemble. Le but de la racisation et de la lemmatisation est de réduire les formes flexionnelles et parfois les formes dérivées d'un mot à une forme de base commune.

La lemmatisation est l'analyse lexicale d'un texte dans le but de regrouper les mots d'une même famille en une unique entité appelée " lemme[10] " ou "forme canonique". Par exemple, tous ces mots ont le même lemme "Définir" : *Définir, définition, définitions, définissons*.

Lemmatisation ou Stemming ?

Parfois, la lemmatisation est erronée pour le stemming. Cependant, il y a une différence essentielle. La racinisation (stemming) ou désuffixation est la transformation des flexions en leur radical ou racine[11].

La racinisation est un processus heuristique brut qui coupe les extrémités des mots sans connaissance du contexte. Un processus de racinisation ne peut pas distinguer entre des mots ayant plusieurs significations. La racine qui en résulte n'est pas nécessairement un mot significatif.

La lemmatisation, de l'autre côté, utilise l'analyse morphologique des mots et le contexte pour désambiguïser la signification. Elle vise à supprimer uniquement les terminaisons flexionnelles et à retourner la forme de base d'un mot (lemme). Ceci est particulièrement important pour les langues qui ont des systèmes d'inflexion riches, tels que le l'Arabe.

II.3. Élimination des mots vides

L'une des techniques les plus importantes de prétraitement est la suppression de mots fonctionnels qui affectent les performances des tâches d'exploration de texte. Les mots vides sont des mots généraux et communs du langage, n'ayant généralement aucune relation sémantique significative avec le contexte dans lequel ils existent [120]. Dans les processus d'organisation des textes non structurés, la suppression des termes non significatifs est essentielle. La tendance générale dans les systèmes IR a été l'utilisation de listes de mots vides assez grandes (200-300 termes) - en raison de la richesse morphologique des langues. La liste contient toutes les variantes morphologiques possibles de chaque mot vide [121].

II.3.1 Liste des mots vides Arabe

L'arabe est une langue lexicalement riche, cela signifie que les mots d'arrêt sont disponibles en grandes quantités. Les mots d'ordre en arabe ont certaines propriétés [121] :

- Ils n'ont aucune signification s'ils sont utilisés séparément.
- Nécessaires pour la construction de la langue.

[10] Un lemme correspond à un terme issu de l'usage ordinaire des locuteurs de la langue.

[11] La racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son (ses) préfixe(s) et suffixe(s), à savoir son radical.

- Ils sont principalement des adjectifs.
- Ne forment jamais une phrase complète lorsqu'ils sont utilisés d'une façon isolée.
- Les mots vides en arabe comprennent certains liens grammaticaux tels que :
  - L'article défini (ال)
  - Les prépositions jointes (اللام، الكاف، الباء…) et séparées (من، إلى، حتى، في، عن، على…)
  - Les conjonctions (ثمّ ، الواو ، أو، الفاء،)
  - Les mots interrogatifs (ماذا، متى، مَنْ، أينَ، كيف…)
  - Les adverbes de temps et de lieu (… خلال، بعد، وراء، طَوال، تحت، بين، عند، حيثُ، أمس، الآن)
  - Les pronoms (… نحن، أنتَ، هو، أنا)
  - Les cinq noms distinctifs (أب، أخ، حم، فو، ذو).

Les mots vides peuvent être séparés ou attachés sous forme de préfixes ou de suffixes. Il existe une liste générale de mots vides en Arabe. Cependant, en raison de la nature fortement inflexionnelle de la langue Arabe, ces mots peuvent prendre différentes formes selon les préfixes et les suffixes qui leur sont attachés.

II.3.2 Approche de suppression des mots vides

En général, il n'y a pas de liste de mots vide fixe dépendante du domaine. Ce type de liste dépend principalement du corpus d'où les mots d'arrêt ont été extraits. Les listes de mots vides génériques utilisés contiennent les 173 mots les plus fréquents la langue anglaise, les 163 pour la langue française et 104 mots de la langue arabe. Ces listes sont créés en utilisant l'approche hybride utilisée dans [122].

L'élimination des mots vides entraîne une amélioration en réduisant le bruit [123], et la taille du corpus de 20% à 30%, ce qui conduit à une plus grande efficacité [120]. L'approche utilisée pour supprimer les mots d'arrêt est illustré dans la figure ci-dessous (Figure 4).

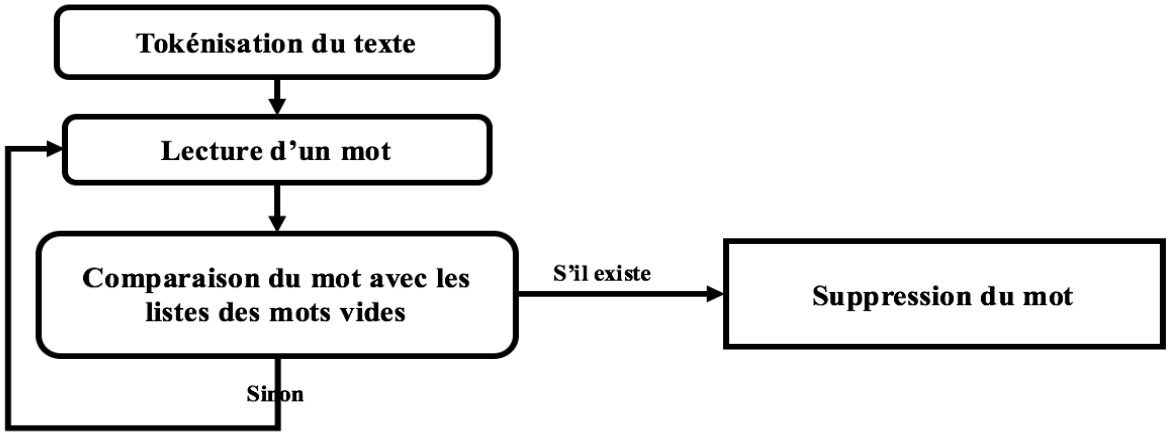

Figure 4: Approche utilisée pour supprimer les mots d'arrêt.

II.4. Étiquetage morphosyntaxique (POS Tagging)

Habituellement, dans le plagiat intelligent, le plagiaire remplace les mots contenus par ses synonymes ou manipule les informations grammaticales, le remplacement changera les mots mais ne changera pas leurs classes. L'étiquetage morphosyntaxique (POS-Tag) consiste à associer à chaque mot d'une phrase les informations grammaticales correspondantes, à savoir sa partie correspondante du discours, genre, nombre…etc.

Dans le système proposé, l'étiquetage morphosyntaxique est fait à l'aide de tagueur Stanford CoreNLP [87]. Les balises POS-Tag de base comprennent les verbes, les noms, les pronoms, les adjectifs, les adverbes, les prépositions, les conjonctions et les interjections.

II.5. Segmentation de texte

Le texte est segmenté en petites parties (sections). La segmentation du texte dans notre cas est effectuée au niveau des mots. Le texte est segmenté en phrases et en mots de 3-grammes (W3G) basées sur le travail réalisé dans [5], où les auteurs comparent plusieurs segmentations (mot 3-grammes (W3G), mot 5-grammes (W5G), mot 8-grammes et phrases (S2S)) pour voir quelle approche peut mieux gérer les cas de plagiat intelligent et ont conclu que W3G donne de meilleurs résultats.

III. Conclusion

Le prétraitement est une tâche importante pour les systèmes de traitement de texte. La nature des documents traités force le besoin de techniques différentes pour nettoyer et structurer les données textuelles, afin d'améliorer la précision ou de diminuer le nombre de caractéristiques qui doivent être traitées, surtout pour un problème complexe comme la détection de plagiat intelligent. La phase de prétraitement réduit la taille du texte et peut même entraîner une amélioration en raison de la réduction du bruit (l'élimination de mots vides, la désambiguïsation du sens du mot à l'aide de l'étiquetage morphosyntaxique… etc.).

Chapitre3 Similarité Sémantique pour la Détection du Plagiat Multilingue

I. Introduction

L'arabe est une langue morphologiquement riche, dans laquelle un mot porte non seulement des inflexions mais aussi des clitiques, tels que les pronoms, les conjonctions et les prépositions. Cette complexité morphologique donne un espace vague au plagiaire de choisir parmi des centaines de possibilités une reformulation de la façon la plus sombre et presque impossible d'être détecter par les systèmes actuels.

Le corpus de test est construit de 600 documents anglais et arabes provenant de différentes sources (nouvelles, articles, tweets et travaux universitaires). Pour des raisons de détection du plagiat multilingue, 200 documents sont simplement traduits (traducteur automatique) sans changement, et 400 documents sont traduits de l'anglais/français vers l'arabe avec un pourcentage de plagiat caché (paraphrase, back-traduction, etc.).

Tout système conçu traiter les langues naturelles doit avoir des informations sur les mots et leurs significations. Ces informations sont traditionnellement fournies via des dictionnaires. Les dictionnaires lisibles par machine sont désormais largement disponibles. Cependant les entrées du dictionnaire ont évolué pour la commodité des lecteurs humains, non pas pour des machines. WordNet [124] offre une combinaison efficace d'informations lexicographiques traditionnelles et moderne.

II. Ontologie : WordNet

II.1. Notion d'ontologie

L'ontologie dans son sens le plus général s'interroge sur la signification du mot *" être",* le terme est utilisé par analogie avec le concept philosophique, qui est " *l'étude de l'être en tant qu'être* ", c'est-à-dire l'étude des propriétés générales de ce qui existe. Avec l'émergence de l'ingénierie des connaissances et du web sémantique, ce terme a pris une autre tournure pour designer la problématique de représentation et de manipulation des connaissances dans un système informatique [125]. L'objectif premier d'une ontologie est de modéliser un ensemble de connaissances dans un domaine donné. En effet, les ontologies permettent de décrire la structure et la sémantique des données. Quand des données sont présentées ou annotées par des ontologies, les logiciels peuvent mieux comprendre leurs

sémantiques, ce qui facilite la localisation et l'intégration des données pour des objectifs divers.

II.1.1 Définition

L'ontologie est une conceptualisation d'un domaine dans un format compréhensible par l'homme, lisible par une machine, composé d'entités, d'attributs, de relations et d'axiomes [126]. Il est utilisé comme représentation standard de connaissances pour le Web sémantique [127]. L'utilisation des ontologies pour surmonter les limites de la recherche par mots clés a été avancée comme l'une des motivations du Web sémantique [128][129].

Une des définitions de l'ontologie qui fait autorité est celle de Gruber [130]:
« *Une ontologie est la spécification d'une conceptualisation. […] Une conceptualisation est une vue abstraite et simplifiée du monde que l'on veut représenter* ». Une ontologie n'est en fin de compte qu'une modélisation du monde réel en concepts et relations entre ces concepts (Figure 5). Elle est donc la manifestation d'une compréhension partagée d'un domaine de connaissance pour plus d'interopérabilité, de réutilisation et du partage [131].

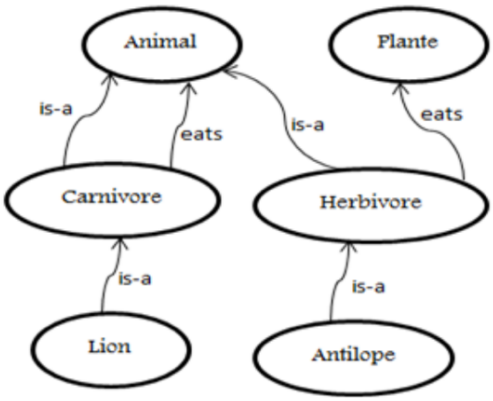

Figure 5: Exemple d' une partie d'une ontologie

Le but principal d'intégration d'ontologie dans les systèmes informatiques n'est pas seulement de servir d'argument déductif pour définir une réalité, mais de permettre une explication des termes et des significations pour définir une base consensuelle pour l'interopérabilité dans un domaine. Les principales composantes d'une ontologie que nous pouvons distinguer sont donc l'objet de la section suivante.

II.1.2 Composants d'une ontologie

D'après[125] [131], une ontologie fournit le vocabulaire d'un domaine et définit le sens des termes et les relations qui les relient. La connaissance dans une ontologie est

formalisée en utilisant cinq composantes : les concepts, les relations, les fonctions, les axiomes et les instances [131].

- **Concepts** : Un concept peut représenter un objet matériel, une notion, une idée. Les concepts constituent les objets de base manipulés par les ontologies. Ils sont représentés dans le langage OWL[12] [132] par owl : Class.

- **Relations** : Traduisent les interactions existant entre les concepts présents dans le domaine traité. Ces relations incluent la relation de spécialisation (subsomption), la relation de composition (méronymie)…etc. Elles sont représentées dans OWL par owl : ObjectProperty.

- **Fonctions** : sont des cas particuliers de relations dans lesquelles le $n^{ième}$ élément de la relation est défini de manière unique à partir des n-1 éléments précédents.

- **Axiomes** : Permettent de modéliser des assertions toujours vraies, à propos des abstractions du domaine, traduites par l'ontologie. Ils permettent de combiner des concepts, des relations et des fonctions pour définir des règles d'inférences.

- **Instances** : (ou individus) représentent des éléments singuliers véhiculant les connaissances à propos du domaine.

II.1.3 Classification des ontologies

Les ontologies peuvent être classé selon le degré de généralité ou le niveau décroissant d'abstraction.

- Ontologies globales (Top-Level Ontology) : présentent un plus haut niveau d'abstraction, décrivent des concepts très généraux et fournissent des notions générales sous lesquelles tous les termes racines dans les ontologies existantes devraient être liés. Elles sont destinées à des utilisations générales (ex : WordNet). Guarino [133] l'a défini: « *Une ontologie formelle est donc une théorie des distinctions formelles entre les éléments d'un domaine, indépendamment de leur réalité* » .

- **Ontologies de domaine** sont réutilisables dans un domaine spécifique donné. Ces ontologies fournissent les vocabulaires sur les concepts et leurs relations dans un domaine. Exemple : commerce, médecine, chimie…etc.

[12] OWL (Ontology Web Language) est un langage de représentation des connaissances construit sur le modèle de données de RDF [132]. Il fournit les moyens pour définir des ontologies web structurées.

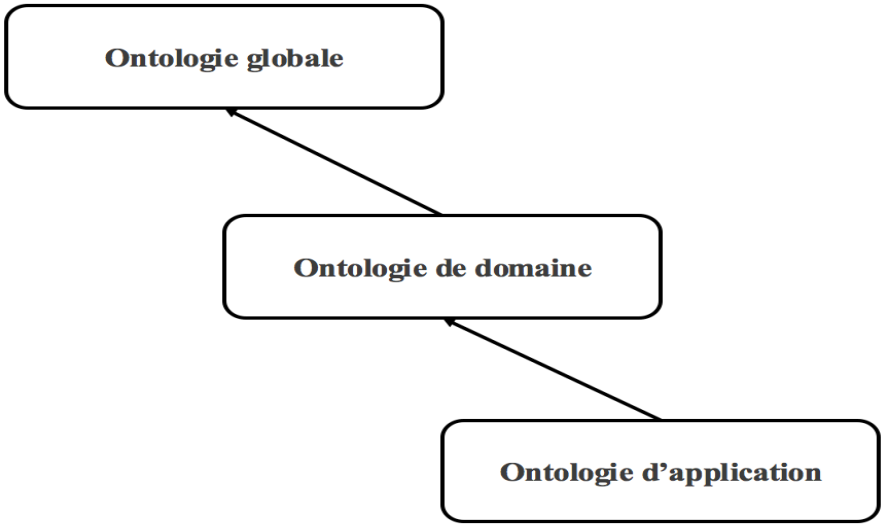

Figure 6:Classification des ontologies

WordNet [124] est l'une des ontologies globales les plus utilisés pour répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue. Elle est parmi les ressources ontologiques les plus exploités dans la désambiguïsation des mots. WordNet [124] est à la base de nombreux travaux et projets récents de détection du plagiat sémantique.

II.2. WordNet : Définition et Structure

WordNet est une base de données lexicale conçue pour être utilisée sous le contrôle d'un programme [93] (Miller 1995). WordNet est une ontologie élaborée initialement pour la langue anglaise, et pour d'autres langues dans les versions ultérieures.

Comme un dictionnaire standard, WordNet contient les définitions des mots, mais elle diffère de ce dernier dans la manière d'organisation. La principale relation entre les mots dans WordNet est la synonymie, comme entre les mots *" voiture "* et " *automobile* ". Les noms, verbes, adjectifs et adverbes sont regroupés en ensembles non ordonnés de synonymes appelés synsets [124]. Chacun des 117 000 synsets de WordNet est lié à d'autres synsets au moyen des " *relations conceptuelles* ". Les relations les plus courantes sont l'hyponyme / Hypernym (c'est-à-dire les relations IS-A) et les relations Meronym / Holonym (c'est-`a-dire les relations Part-Of).

Les formes de mots ayant plusieurs significations distinctes sont représentées dans autant de synsets distincts. La figure 7 ci-dessus illustre un fragment de la hiérarchie de WordNet.

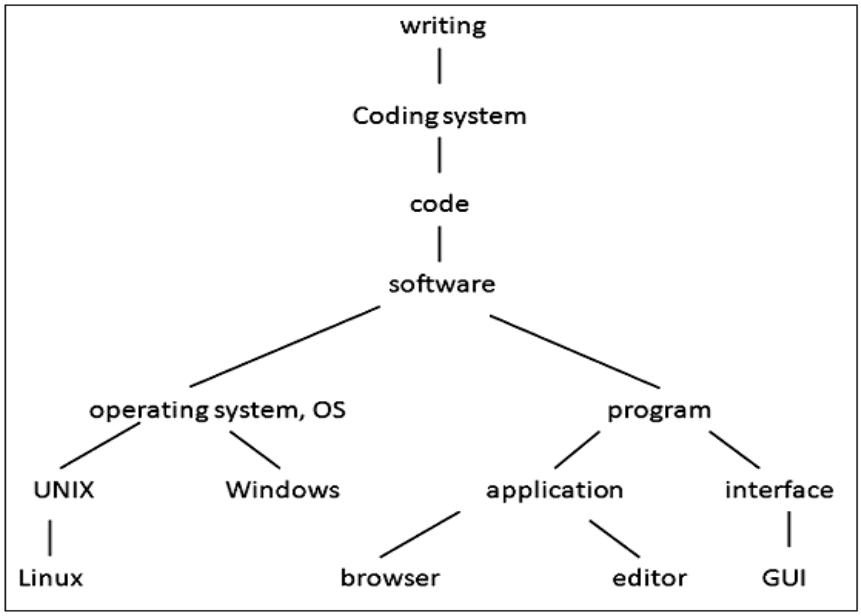

Figure 7: Fragment de la hiérarchie de WordNet

La base de données WordNet est stockée dans un format ASCII composé de huit fichiers, deux pour chaque catégorie syntaxique.

III. Similarité sémantique pour la DPM

Mesurer la distance entre les concepts est un domaine d'étude important du traitement du langage naturel. Comparer (sémantiquement) deux concepts exprimés lexicalement est un problème qui envahit une grande partie des applications de traitement du langage naturel (NLP), spécialement dans un cadre multilingue (par exemple la désambiguïsation des mots, la réponse aux questions, l'extraction d'informations…). WordNet [124], qui comprend une grande variété de concepts associés aux mots (synsets), est souvent utilisé comme source pour calculer ces distances.

Dans cette partie de la thèse, nous présentons une distance de similarité pour deux concepts multilingues. Cette distance est appliquée en particulier pour la détection du plagiat en langues Arabes et Anglais/Français. Plusieurs mesures [92], [98], [134], [135] avaient été proposées à cette fin, principalement basées sur les distances définies dans la hiérarchie WordNet [124] [93].

Conformément à ces travaux, nous proposons une nouvelle approche définissant une méthode de similarité sémantique multilingue basée sur la similarité de deux synsets WordNet [93], en utilisant les séquences d'étiquettes morpho-syntaxique (POS-tagging). Pour définir cette distance, nous utilisons les différentes distances citées précédemment, qui reposent sur

plusieurs techniques. La méthode proposée offre une nouvelle perspective aux tâches mentionnées précédemment et pourrait être utilisée pour étendre les informations fournies par les distances existantes.

III.1 Similarité sémantique mot à mot

L'arabe "*abjad*" nécessite un traitement morphosyntaxique spécial, c'est ce qui nous a incité à modifier et réorganiser d'un côté le texte arabe afin d'appliquer les mesures de similarité, et de l'autre côté faire adapter ces dernières avec la nature linguistique des textes arabes. Le processus de calcul de similarité sémantique est composé de plusieurs tâches. Chacune des tâches séquentielles accepte les données d'une étape précédente, effectue une transformation sur ces données, puis les transmet à l'étape suivante (Figure 9).

- Les entrées sont deux mots (Arabe et Anglais).
- Un prétraitement est effectué sur les entrées.
  - Les diacritiques arabes "*Harakat/Tashkeel*" sont très importants pour définir le sens correct d'un mot écrit en arabe "*abjad*". Pour cela, on va translittérer les mots en utilisant un API Java développé par Tim Buckwalter[13] [136].
  - L'application des divers traitements des textes naturels à savoir : la lematisation, étiquetage morphosyntaxique (POS Tagging) …etc.

L'objectif est de calculer la similarité sémantique entre deux mots sans avoir recours à la traduction de l'un des deux. L'idée se repose principalement sur l'utilisation des méthodes à base de corpus parallèles. De ce fait la localisation des deux concepts dans les bases de données WordNet-Arabe et WordNet-Anglais est spéciale dans notre approche. D'abord, on repère les deux concepts dans la base correspondante à leur langue, puis on cherche l'ancêtre commun le plus proche CCA (Closest Common Ancestor).

Généralement dans le calcul de similarité, il faut préciser les concepts et la base fouillée à l'algorithme de recherche. Dans notre cas, on a deux bases différentes en nature de langue mais analogues dans leurs structures. D'où la construction des synsets est presque la même (avec des différences négligeables). Durant la recherche de l'ancêtre commun des deux concepts on utilise le même algorithme de fouille pour les deux bases. La similarité est obtenue à la base de l'emplacement des ancêtres communs dans WordNet (Figure 8). Si on n'arrive pas à trouver l'ancêtre de l'un des concepts, on prend le synonyme le plus proche dans la synset du l'ancêtre manquant (car il s'agit d'une similarité sémantique).

[13] La translittération arabe de Buckwalter a été développée à Xerox par Tim Buckwalter dans les années 1990, il comprend des classes Java pour l'analyse morphologique des fichiers texte arabes, quel que soit leur encodage.

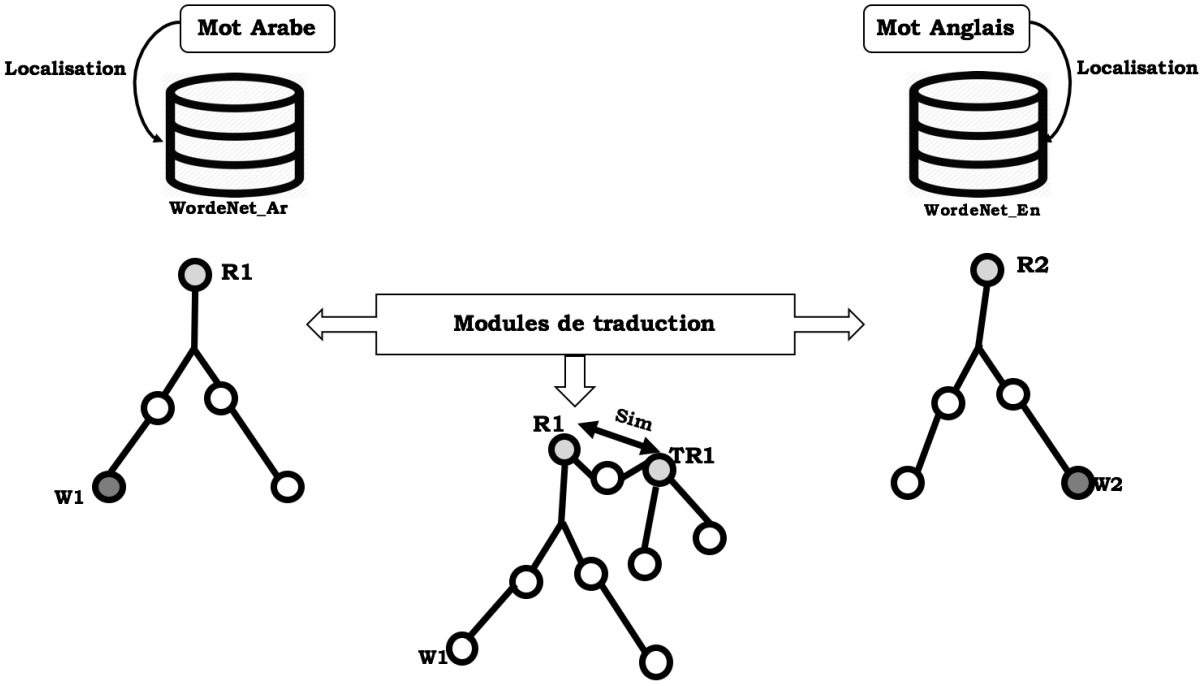

Figure 8: l'approche de calcul de similarité Mot à Mot

Quoique WordNet est une large base, l'un des problèmes qui peuvent survenir est l'absence de l'un des concepts (ou parfois les deux). Dans ce cas, on procède (obligatoirement) à la traduction des concepts en prenant le premier synonyme.

```
Algorithm: W2WSS
Inputs: Word Ar, Word En
Output: SS(Ar,En)
BEGIN
Preprocessing for Word Ar
Preprocessing for Word En
Localization of Ar in  Arabic_WordNet
Localization of En in  English_WordNet
Look for CCA_Ar
Look for CCA_En
Compute Sim(CCA_Ar,CCA_En)
END
```

Algorithme 1: Similarité sémantique Mot à Mot

L'implémentation d'un nombre important des mesures de similarité est dû au manque d'une bibliographie relative à notre problématique. Pour choisir les mesures adéquates, nous étions obligés d'en tester le maximum possible.

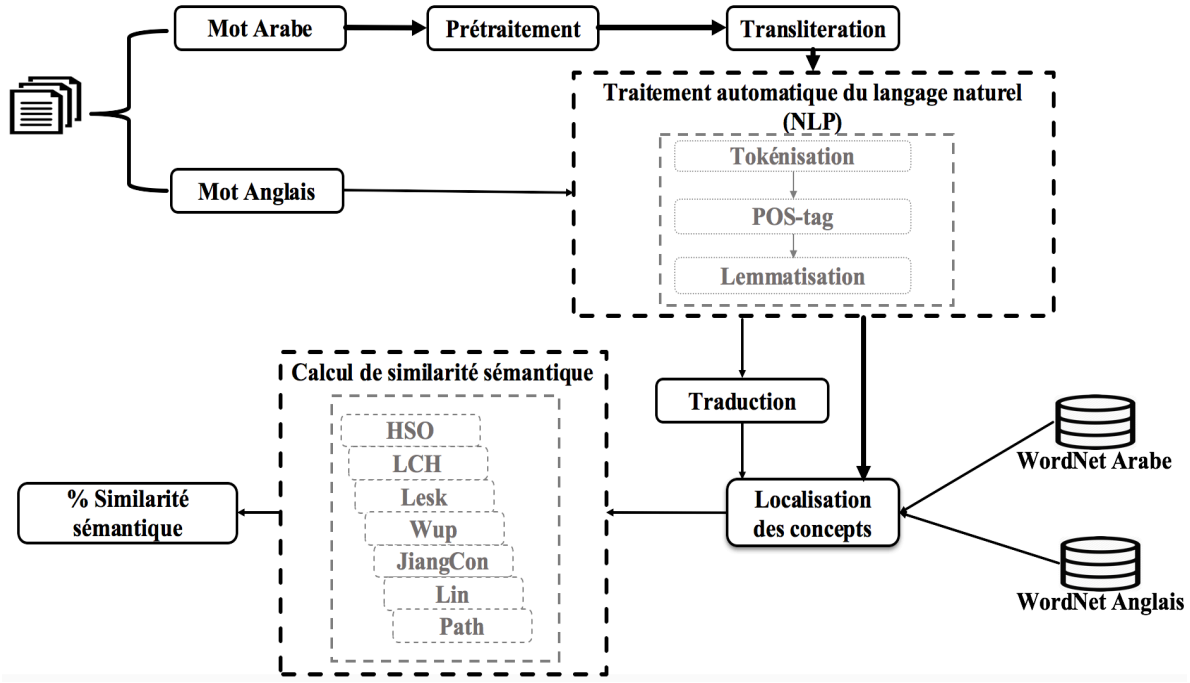

Figure 9: Calcul de similarité Mot à Mot

Les résultats de cette expérience sont présentés en une simple interface GUI.

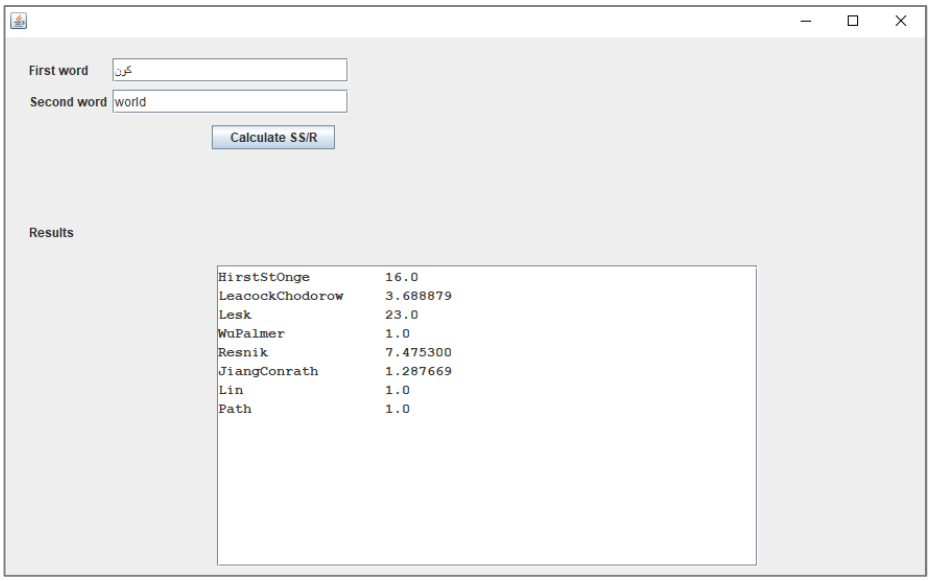

Figure 10: Exemple de similarité sémantique pour deux mots

III.2 Comparaison des différentes mesures de similarité sémantique utilisées

La similarité sémantique entre deux documents est un concept selon lequel un ensemble de métriques sont calculées en se basant sur la similitude de leurs significations (contenu sémantique). Le choix des approches basées sur WordNet de celles présentées précédemment est dû à l'efficacité du calcul sémantique pour la détection du plagiat. Compte tenu de deux

mots, leur degré de similarité peut être estimée à partir de leur position relative dans la hiérarchie de WordNet.

Le tableau ci-dessous (Tableau 5) présente les résultats de calcul de la similarité sémantique entre des mots correspondants.

		Mesure de similarité sémantique			
Terme1	**Terme 2**	**WuP**	**LCH**	**Lin**	**JCN**
الكون	universe	1.0	3.68	1.0	--
عملية	process	0.4615	1.3862	0.0967	0.0687
شبكة	web	0.6153	1.8971	0.2117	0.0592
حديث	conversation	0.4	1.3962	0.0976	0.0694
دائم	perpetual	0.1818	0.7444	0.0	0.0
مخزن	stored	0.6	1.4916	0.2901	0.08197
تلاشى	vanish	1.0	3.3322	1.0	--
لانهائي	infinite	0.7273	2.3982	0.0	0.0
الفضاء	space	1.0	3.6889	1.0	--
أخطاء	mistakes	1.0	3.6889	1.0	--
ظاهري	face	0.2105	0.9162	0.0	0.0
معاناة	pain	0.3333	1.1239	0.0989	0.0704
نمّام	gossip	0.1818	0.7444	0.0	0.0

Tableau 5: Mesures de similarité sémantique selon WordNet entre deux mots extraits de notre corpus

D'après les résultats des calculs, l'approche proposée combinée avec WuP et Lin donne des rapports de similarité raisonnables. Les résultats de Leacock&Chodorow sont également acceptables. Jusqu'à maintenant on ne s'intéresse pas au temps d'exécution, l'essentiel est de trouver le taux de similarité correcte entre deux mots de deux langues différentes dans un système multilingue, en se basant sur un système à base de corpus parallèles.

Le tableau suivant (Tableau 6) illustre les résultats de notre approche, en se basant sur WordNet. Le calcul de similarité est effectué entre deux documents parallèles

(Arabe/Anglais) (le calcul a été répété sur plusieurs documents extraits de notre corpus) :

	Similarité	**Temps(ms)**
Wup	0.2883	6059
LCH	0.1885	6680
Lin	0.1262	4500

Tableau 6: Comparaison des différentes mesures de calculs de similarité

Les résultats de cette comparaison confirment la validité de notre choix de mesures (WuP et Lin) pour le calcul de la similarité sémantique et la détection du plagiat multilingue entre deux documents parallèles, avec un point de plus : un temps d'exécution minimal.

Les documents de tests sont des documents parallèles avec un taux élevé de similarité sémantique (plagiat) alors que les résultats obtenus (Figure 11) ne satisfassent pas exactement nos besoins de calcul de similarité sémantique bien évidemment une détection fiable de plagiat multilingue.

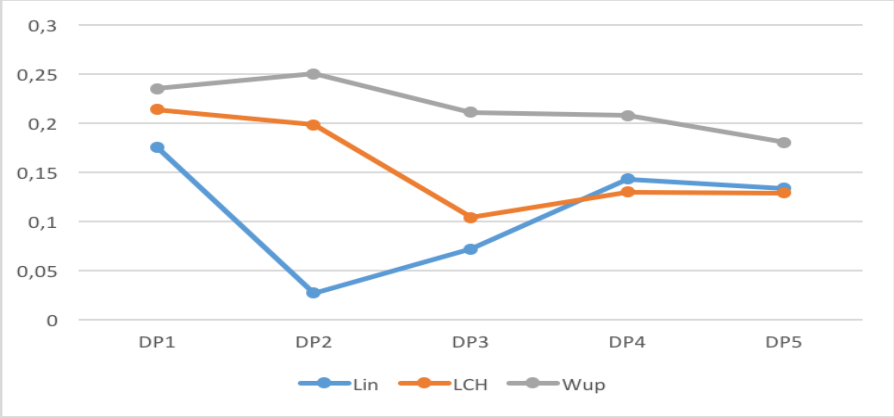

Figure 11: Comparaison des différentes mesures de calculs de similarité pour des documents parallèles

Une chose à admettre, les plagiaires sont des gens intelligents, ils essaient leur maximum pour voiler le plagiat soit par généralisation ou spécialisation des termes. Par conséquent, ils n'utilisent jamais le premier synonyme du mot. D'où, dans un processus de détection de plagiat, il faut élargir l'espace de recherche et de calcul afin d'englober tous les synonymes associés du mot à l'aide de la " *Synonymie WordNet* " qui se présente dans WordNet sous forme de synsets. Pour cette raison, on calcule la similarité sémantique pour le mot donné et tout le synset, c.à.d. tous les synonymes figurants dans le synset un par un.

Exemple :

"*Software is a description of what a computer does, the term software refers to everything that is made up of a computer except for computer hardware.*"

'' البرمجيات عبارة عن وصف لما يقوم به الحاسوب من عمليات ، مصطلح البرمجيات يشير إلى كل ما يتكوّن منه جهاز الحاسوب باستثناء مكوّنات الحاسوب المادية.''

'' البرنامج المعلوماتي عبارة عن وصف لما يقوم به الكمبيوتر ، يشير المصطلح إلى كل شيء مكون للكمبيوتر ماعدا الأجزاء المادية.''

Le mot " *software* " peut être plagié en plusieurs mots arabes qui partagent un sens commun à savoir "*برمجيات*"، "*برنامج الحاسوب*"، "*نظم التشغيل*"، "*برامج النظم*" selon le choix et la volition du plagiaire. De ce fait, si on calcule la similarité sémantique juste pour le premier sens de " *software* " qui est "*برمجيات*" alors que dans le texte existe un des autres sens, on va surement avoir un résultat erroné.

IV. Conclusion

Le plagiat multilingue peut être plus compliqué que nous ne pourrions attendre. Ainsi, l'utilisation de la théorie des ensembles flous dans la détection du plagiat multilingue parait la solution idéale pour remédier au problème de l'utilisation de synonymie (c.à.d. le remplacement d'un mot par des mots de même signification) pour s'échapper du plagiat. En effet, une approche sémantique floue peut être modélisée en considérant que les mots d'une phrase (à partir de deux textes comparés) ont un ensemble flou qui contient des mots ayant une signification similaire.

Chapitre 4 Détection Floue pour le Plagiat Sémantique Multilingue

I. Introduction

Un mot peut avoir plusieurs significations ou sens. La correspondance de deux phrases de langues différentes est souvent approximative, ce qui peut être modélisé en considérant que chaque mot d'une phrase est associé à un ensemble flou qui contient les mots ayant une signification similaire (Application du principe de Zadeh) (Figure 12).

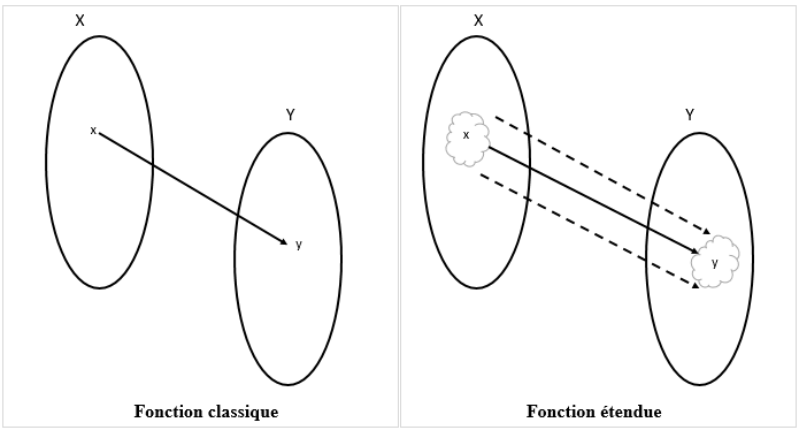

Figure 12: Principe d'extension de Zadeh

La langue Arabe est connue par sa richesse, ainsi sa diversité de construction et de signification des mots. Le changement de textes de/vers l'Arabe est une tâche complexe, par conséquent, adopter une approche basée sur la sémantique floue semble être la meilleure solution.

II. Similarité sémantique floue pour la DPM

II.1. Relation Mot-à-Mot

Les relations mot-à-mot peuvent être basées sur différentes hypothèses (Figure 13) :

- ⇒ Les mots sont identiques ;
- ⇒ Les mots sont dans le même synset (synonymes) ;
- ⇒ Les mots ne sont pas dans le même synset mais leur synset contient au moins un mot commun ;
- ⇒ Les mots ont au moins un hypernym partagé ;
- ⇒ Les mots sont différents.

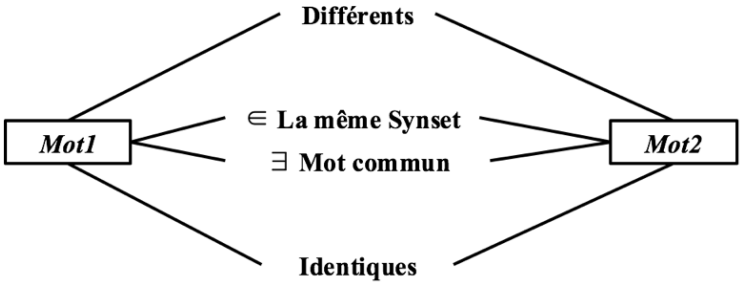

Figure 13: Relation entre deux mots dans WordNet

Les synsets dans WordNet sont liés à d'autres synsets plus ou moins élevés dans la hiérarchie par différents types de relations. Dans notre approche de calcul de similarité sémantique on s'est basé sur les principales relations entre les synsets dans WordNet. La relation la plus fréquemment encodée entre les synsets est la relation super-subordonnée (également appelée hyperonymie ou relation IS-A). Elle combine des synsets plus généraux avec des synsets de plus en plus spécifiques. Toutes les hiérarchies nominales remontent finalement au nœud racine. La relation IS-A est transitive.

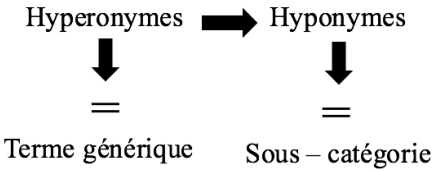

Figure 14: Relation hyperonymie/ hyponymie

La Meronymie (Holonym ou relation Part-Of) ou la relation partielle entre les synsets. WordNet distingue les types (noms communs) et les instances (personnes, pays et entités géographiques spécifiques). Les instances sont toujours des nœuds terminaux (terminaux) dans leurs hiérarchies.

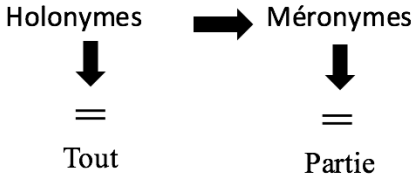

Figure 15: Relation Meronymie / Holonymie

Dans un processus de détection de similarité multilingue basé sur la sémantique, où les mots frontières ne sont pas clairs, et l'intersection des significations des mots est floue. La théorie des ensembles flous semble être la bonne façon de traiter un tel cas.

La théorie des ensembles flous (Lofti Zadeh en 1965) est une généralisation de la théorie des ensembles classiques. Elle permet d'évaluer progressivement l'appartenance des

éléments à un ensemble à l'aide d'une fonction d'appartenance valorisée dans l'intervalle des unités réelles [0,1]. La théorie des ensembles flous pourrait être utilisée dans un large éventail de domaines, en particulier pour le traitement de données incertaines et imprécises liées à la DPM.

Chaque mot d'un document est associé à un ensemble flou qui contient des mots de même signification avec un degré de similarité (généralement inférieur à 1) (Figure 16). L'utilisation de la théorie des ensembles flous dans la DPM apparaît donc comme un moyen évident pour résoudre le problème de détection de plagiat multilingue.

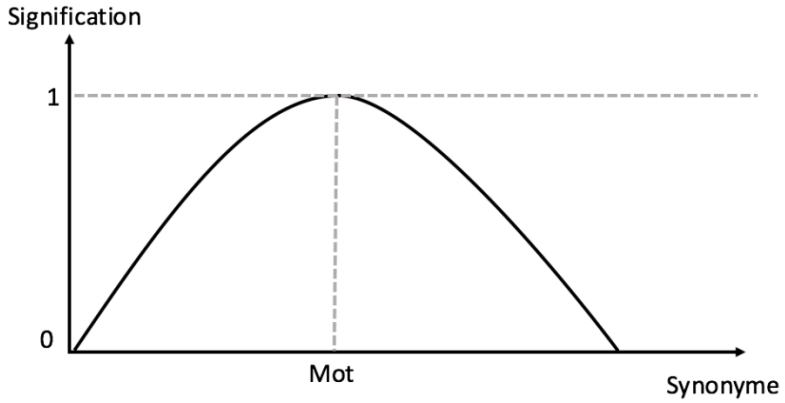

Figure 16: Schématisation de l'ensemble flou des synonymes d'un mot

II.2. Démarche proposée pour la détection du Plagiat multilingue

Diverses mesures de similarité sémantique de mots ont été proposées en tenant compte de leur relation dans la base de données lexicale WordNet [93], sur la base de nos résultats précédents [137], [138] et de travaux connexes [137] la mesure WuP donne des résultats intéressants. Par conséquent, pour fuzzifier la relation des paires de mots (à partir des textes d'entrée), on prend la mesure de similarité de WuP comme fonction d'appartenance, (4) sera exprimée comme suit :

$$\mu_{a_i,b_j=Wup(a_i,b_j)} \quad (19)$$

Les couples de mots de deux textes d'entrés sont considérés comme des variables floues. Cette relation évalue le degré de similarité (sémantique) entre deux mots. La relation floue entre deux mots est comprise entre 0 pour des mots totalement différents (il n'y a pas d'hypernemie partagée entre les mots), et 1 pour des mots identiques ou appartient au même synset (synonymes). Un système d'inférence floue a été construit pour évaluer la similarité de deux textes et en déduire le plagiat. Pour évaluer la relation d'un mot dans un texte par rapport aux

mots de l'autre texte, on a choisi d'utiliser l'opérateur PROD flou comme dans les formules suivantes :

$$\mu_{a_1,B} = 1 - \prod_{b_j \in B, j \in [1,m]} (1 - Wup(a_1, b_j))$$

$$\vdots \tag{20}$$

$$\mu_{a_n,B} = 1 - \prod_{b_j \in B, j \in [1,m]} (1 - Wup(a_n, b_j))$$

Ensuite, nous calculons la somme moyenne :

$$\mu_{A,B} = (\sum_{i=1}^{n} \mu_{a_i,B}\ )/n \tag{21}$$

Les entrées sont des documents parallèles (de langues différentes) passés par plusieurs étapes de prétraitement et une segmentation mot-3-grammes (W3G). Les textes d'entrée passent au prétraitement avec quelques différences, considérant que l'arabe est une langue assez difficile à traiter. Les textes résultants sont utilisés comme entrées dans le système d'inférence floue, puis la mesure de similarité sémantique WuP est modélisée comme une fonction d'appartenance. La sortie est un score de similarité entre les textes d'entrée (Figure 17). Cela peut être modélisé dans l'algorithme ci-dessous (Algorithme 1) :

```
Algorithm: FCLPD
Inputs: Text A, Text B
Output: CPD(A,B)
BEGIN
Preprocessing for Text A
Preprocessing for Text B
For each segment a_i ∈ A Do
 For each Segment b_j ∈ B Do
  Input a_i and b_j to fuzzy inference
system
      Compute Wup(a_i,b_j)
      CPD(a_i,b_j) ← Wup(a_i,b_j)
      If CPD(a_i,b_j) ≥ threshold
            Add (a_i,b_j) to Output
      End If
 End of loop For
End of loop For
END
```

Algorithme 2: Algorithme de détection floue du plagiat multilingue

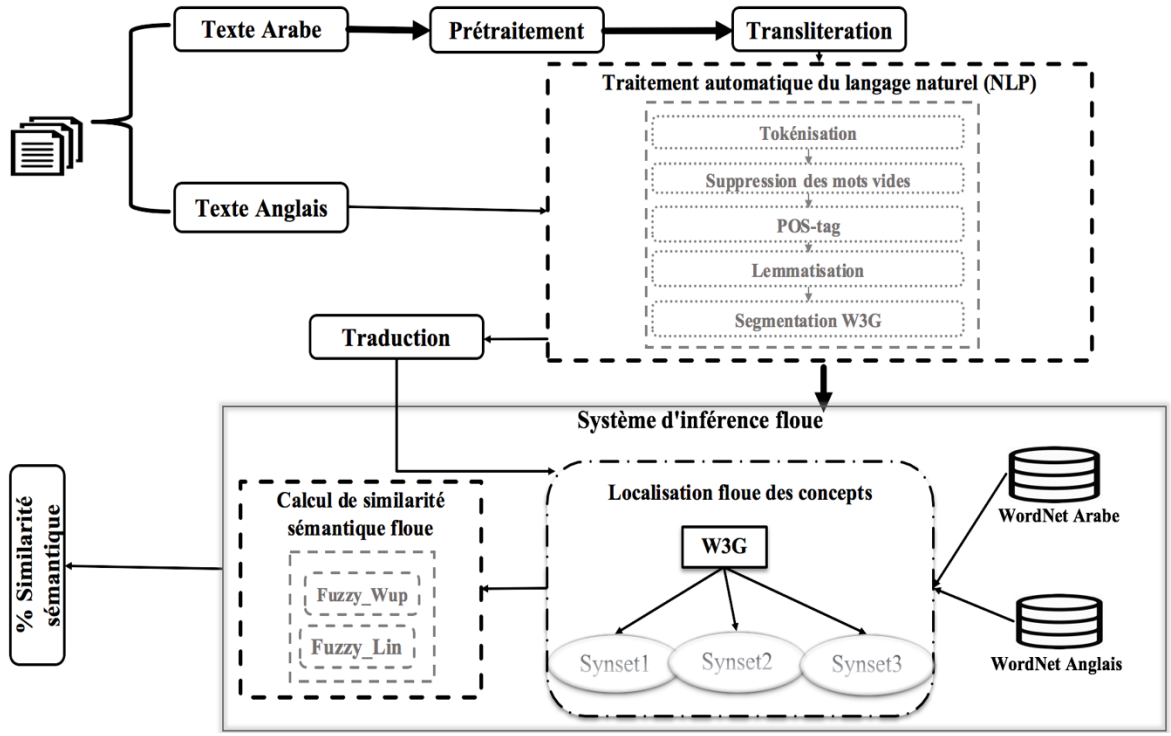

Figure 17:Détection floue du plagiat multilingue

Exemple :

Dans cet exemple, le deuxième texte est reformulé à partir du premier, le sens reste le même. Les textes *Ar* et *En* passent d'abord par le prétraitement puis par notre système. L'analyse des deux textes signifie que chaque segment du texte *Ar* sera comparé à chaque segment du texte *En*.

"*Wikipedia is an online, universal, multilingual and wiki-based online encyclopedia project. Wikipedia aims to provide freely reusable, objective and verifiable content that everyone can modify and improve. All the editors of Wikipedia articles are volunteers. They coordinate their efforts in a collaborative community, without a leader*".

"ويكيبيديا هي مشروع موسوعة رقمية، متعددة اللغات، حرة المحتوى. يستطيع أي شخص التحرير فيها بدون تسجيل، ويستطيع أي شخص الاستفادة من المحتوى، واستغلاله بهدف تجاري أو غيره وفقًا لترخيص الموسوعة. ويجري آلاف الزوار، من مختلف أنحاء العالم، الكثير من التعديلات، وينشئون الكثير من المقالات الجديدة يوميًا "

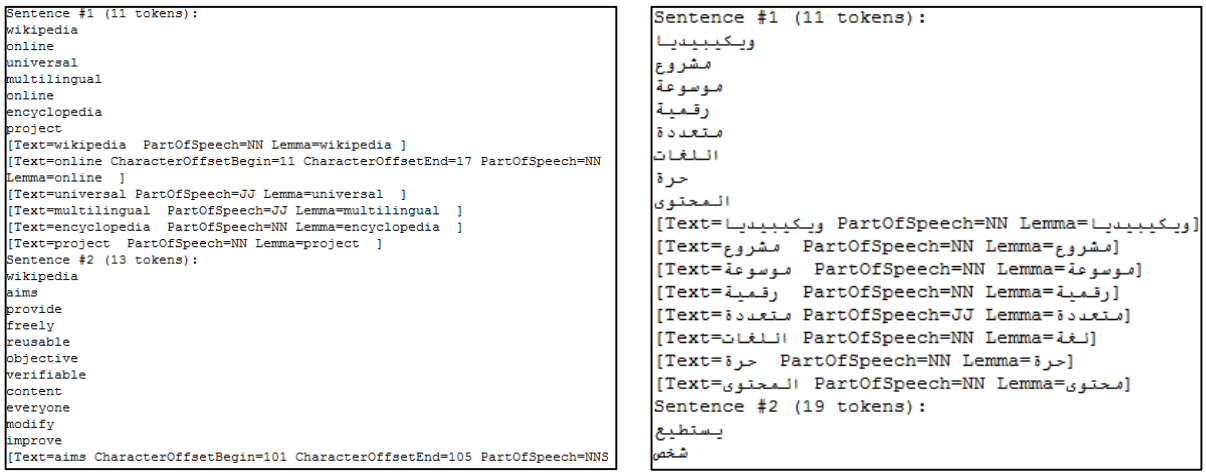

Figure 18: Texte (Exemple) prétraité

Il est clair que les deux textes sont identiques. Les segments des premières phrases sont semblables d'un pourcentage de 88,13%. Les segments *Ar2* et *En2* sont similaires d'un degré de 64,86%. *Ar4* est identique à *En3* d'un degré de 99,6%. La même chose peut être remarquée pour les derniers segments. Si on compare les deux textes entiers, cela donnera un pourcentage de similarité sémantique de 76,75%, ce qui représente un taux élevé de plagiat.

II.3. Résultats expérimentaux et comparaisons

La théorie des ensembles flous et la DPM semblent d'être le couple parfait. Les résultats de cette expérience sont présentés dans ce qui suit (Tableau 7 et Figure 19).

	Similarité	Temps(ms)
Fuzzy_Wup	0.5661	10139
Fuzzy_Lin	0.2523	9234

Tableau 7: Comparaison de notre approche combinée avec différentes méthodes de calculs de similarité

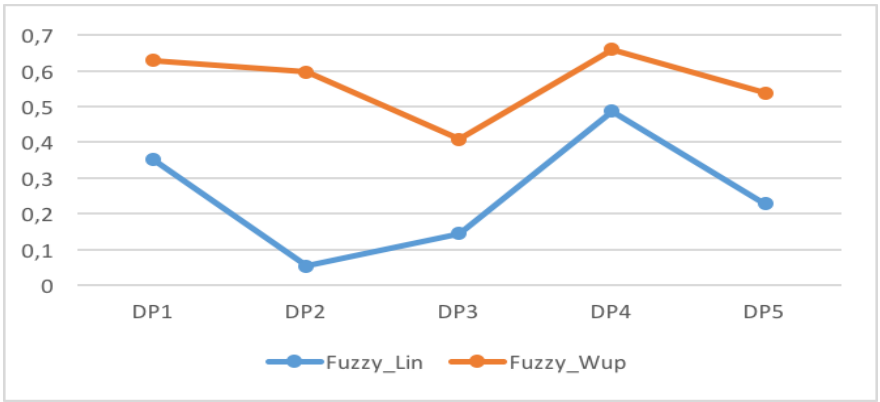

Figure 19: Comparaison de notre approche combinée avec des mesures floues de détection du plagiat pour des documents parallèles

L'évaluation de l'efficacité des deux approches se base sur le taux de similarité calculé et le temps pour avoir ce résultat, nous avons utilisé différentes collections de documents de test de tailles différentes, de contenu hétérogène et d'un pourcentage varié de plagiat dans chaque collection. Les résultats des mesures de temps d'exécution et de similarité sont illustrés dans les figures 19 et 20. L'application de notre approche avec la mesure de WuP donne des résultats intéressants en terme de similarité (pour les gros documents), alors qu'avec Lin on remarque une diminution du temps d'exécution pourtant une similarité médiocre.

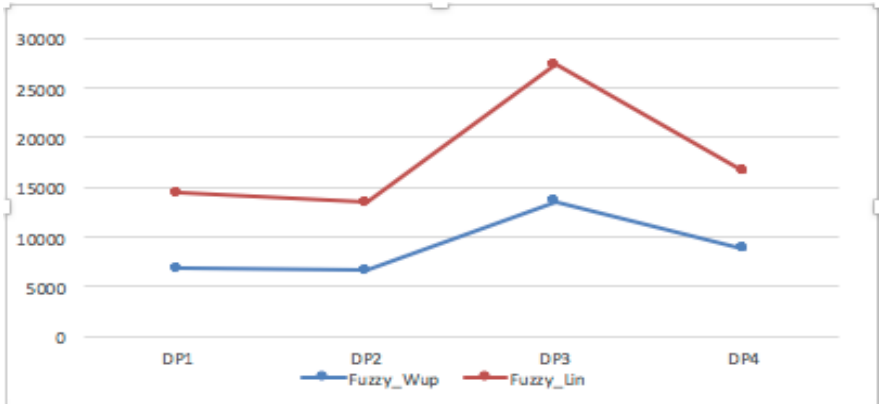

Figure 20: Comparaison de temps d'exécution des mesures de similarité Fuzzy_WuP et Fuzzy_Lin

En comparant les résultats obtenus et ceux présentés précédemment (Tableau 6 et Figure 11), le taux de détection de plagiat a augmenté de plus que 50% avec un temps d'exécution plus grand. Puisque le calcul est fait pour des documents parallèles le taux de similarité sémantique multilingue est satisfaisant.

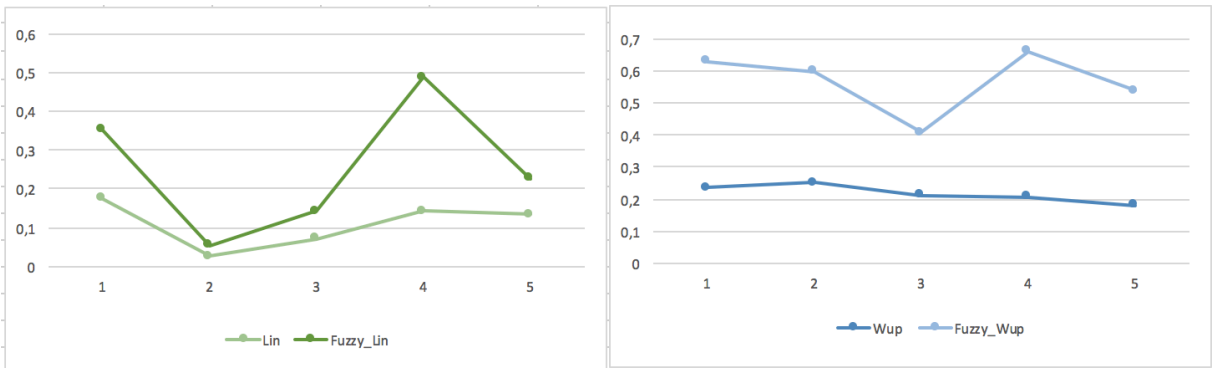

Figure 21: Comparaison des résultats avant et après la fuzification de notre approche

D'après ces résultats, notre approche est efficace pour un système de détection du plagiat multilingue (Figure 21). Toutefois, le nombre important d'opérations et la durée d'exécution (Figure 20) impliquent la recherche d'une solution pour améliorer la performance en ce terme.

III. Conclusion

L'utilisation de la théorie des ensembles floues est basée sur l'hypothèse que la correspondance de deux phrases sémantiquement peut être modélisé en considérant que chaque mot d'une phrase est associé à un ensemble flou qui contient les mots ayant la même signification. La théorie des ensembles flous permet d'évaluer progressivement l'appartenance des éléments à un ensemble à l'aide d'une fonction d'appartenance valorisée dans l'intervalle des unités réelles [0,1]. Dans l'approche proposée, la mesure de similarité de WuP a été prise comme fonction d'appartenance, et n système d'inférence floue a été construit pour évaluer la similarité de deux textes et en déduire le plagiat.

L'un des problèmes les plus importants en matière de détection du plagiat à part l'énorme masse d'informations est le temps de recherches et de calcul surtout dans un système multilingue. Une solution qui continue de croître est l'utilisation des technologies Big Data pour paralléliser le travail et distribuer le stockage.

Chapitre5 Similarité Sémantique Floue et Big Data pour la Détection du Plagiat Multilingue

I. Introduction

Les problèmes qui se posent dans les systèmes de détection du plagiat multilingue sont le temps de recherche et la gestion du stockage. Le calcul dans un corpus volumineux prend un temps déraisonnable.

L'idée est d'utiliser les technologies Big Data en travaillant avec le framework Hadoop basé sur le modèle de programmation MapReduce et le système de fichier HDFS (Hadoop Distributed File System). Notre approche consiste à distribuer le stockage entre les différentes machines de notre cluster Hadoop, et à paralléliser l'algorithme de calcul avec MapReduce.

II. Pourquoi le Big Data pour la détection de plagiat ?

De nos jours, la quantité de données générées sur le Web est inimaginable. La quantité de données et la fréquence par laquelle sont générées ont produit le terme «Big Data» littéralement « *grosses données* » ou méga-données, défini par Gartner (Beyer et Douglas 2012)[139], comme étant « *un volume élevé, une vitesse et une variété de données qui exigent des formes de traitement novatrices et rentables pour une meilleure compréhension et prise de décision* » ; et la véracité, la valeur récemment ajoutées, fondées sur le fait qu'une analyse précise pourrait être affectée par la qualité des données capturées.

Le Big Data désignent des ensembles de données tellement volumineux qu'ils en deviennent difficiles à travailler avec des outils classiques de gestion de base de données ou de gestion d'information, comme dans notre cas de détection de plagiat multilingue. Il s'agit donc d'un ensemble de technologies, d'architectures, d'outils et de procédures permettant une organisation très rapide de captage, traitement de larges quantités de données, et extraction des informations pertinentes à un coût optimal.

III. Travaux existants

Dans tout système de détection de plagiat, un problème majeur commun au traitement des données est la comparaison efficace des textes. En particulier le calcul de la similarité sémantique, l'augmentation du nombre de publications implique directement une augmentation du nombre de documents suspects sources de plagiat.

La plupart des travaux de recherche dans ce domaine ont utilisé le Big Data dans la phase de recherche d'informations ou de sélection des documents candidats [140], [141]. Zhang et al. [142] ont présenté une méthode basée sur la séquence pour détecter la similarité partielle de pages Web à l'aide de MapReduce, composée de deux sous-tâches, la détection des doublons au niveau des phrases et la correspondance des séquences. Erritali et al. [143] ont proposé une approche de mesures de similarité sémantique et un nouvel algorithme MapReduce. Dwivedi et al. [144] introduisent un algorithme de détection de plagiat SCAM (Standard Copy Analysis Mechanism), le processus de détection proposé est basé sur le traitement de langage naturel, et un algorithme modifié de SCAM basé sur MapReduce.

IV. Application de notre approche dans un système Big Data

Comme indiqué précédemment, l'un des problèmes les plus importants en matière de détection de plagiat est l'énorme masse d'informations, en particulier lorsque la collection source est sous forme d'un corpus volumineux. La solution est de paralléliser notre système en partageant le travail entre plusieurs machines. Les entrées et la sortie du système proposé sont stockées en HDFS, et le développement de l'algorithmes est parallélisé avec MapReduce.

IV.1. Méthodologie de recherche et Algorithme

Dans cette partie, on a fuzzifié la méthode de détection du plagiat en utilisant les mesures de similarité sémantique floues (fuzzy_WuP, fuzzy_Lin) de manière parallèle en utilisant Apache Hadoop (HDFS et MapReduce). L'idée est de stocker les entrées (les documents candidats et le document suspect), et les résultats des méthodes proposées dans le système HDFS pour la répartition du stockage entre plusieurs machines (Hadoop Cluster). Dans le développement de notre proposition, nous utilisons le modèle de programmation MapReduce pour paralléliser l'approche.

Les entrées proviennent de deux langues différentes, un texte Arabe et un corpus de candidats sources de plagiat en Anglais/Français. Le texte Arabe et chaque texte du corpus sont utilisées comme entrées pour le système d'inférence floue, les mesures de similarité sémantique (WuP et Lin) sont alors modélisées comme des fonctions d'appartenances. Le résultat est un score de similarité entre le texte Arabe et chaque texte d'entrée du corpus. La figure suivante (Figure 22) montre les différentes étapes de notre travail.

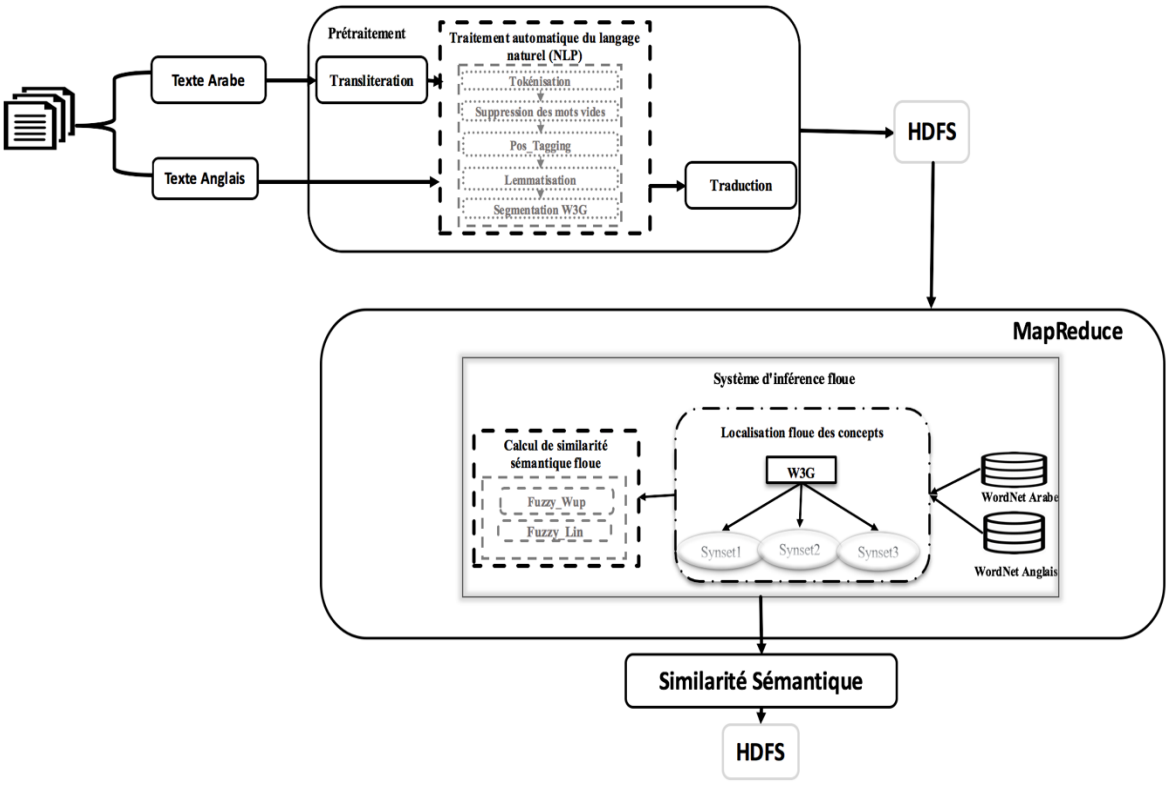

Figure 22: Système parallèle de détection de plagiat multilingue

Comme présenté précédemment et illustré dans l'algorithme ci-dessous (Algorithme 2), chaque entrée de l'algorithme MapReduce contient deux textes : Arabe et Anglais, permettant de calculer la similarité sémantique (détection du plagiat multilingue).

Où :

- Text-Preprocessing() : Consiste à appliquer différentes méthodes de prétraitement (tokénisation, suppression des mots vides, lemmatisation, etc.).
- W3G() : Une fonction qui applique une segmentation de mots de 3-grammes.
- WordNet() : appliqué au segment arabe (translittération) ou à la traduction si le mot n'existe pas dans WordNet.
- Wup(mot, terme), Lin(mot, terme) : Deux fonctions qui calculent la similarité sémantique floue.
- Write () : C'est une fonction qui stocke le résultat de notre méthode dans HDFS sous forme de deux colonnes, le premier pour les entrées et le second pour la valeur calculée.

```
Inputs : Arabic Text, English text from the corpus
Require : Semantic Similarity between inputs
  S_Wup ← 0;  S_Lin ← 0;  C ← 0;
  AR ←  Text Preprocessing(Arabic Text);
  EN ←  Text Preprocessing(English Text);
  Segmentation1[] ← W3G(AR);
  Segmentation2[] ← W3G(EN);
  For all word In Segmentation1
      For all term In Segmentation2
         If  word In WordNet
           Then  word ← WordNet(word);
         Else    word ← Translate(word);
          End If
          Fuzzy_Wup ← 1 - Wup(word, term);
          Fuzzy_Lin ← 1 - Lin(word, term);
          S_Wup ← S_Wup + Fuzzy_Wup;
          S_Lin ← S_Lin + Fuzzy_Lin;
          C ← C + 1;
     End For
  End For
  Sim_Wup=S_Wup/C;
  Sim_Lin=S_Lin/C;
  Write(Arabic Text || English text, Sim_Wup or
Sim_Lin)
```

Algorithme 3 : Modèle de programmation MapReduce pour l'approche proposée

La distribution du stockage des entrées et la parallélisation de la détection du plagiat se font en construisant un cluster Hadoop contenant trois machines nœuds, ce cluster comprend une machine maître et deux machines esclaves. Chaque nœud est une machine Ubuntu 16.04. La figure ci-dessous montre la configuration de notre cluster (Figure 23).

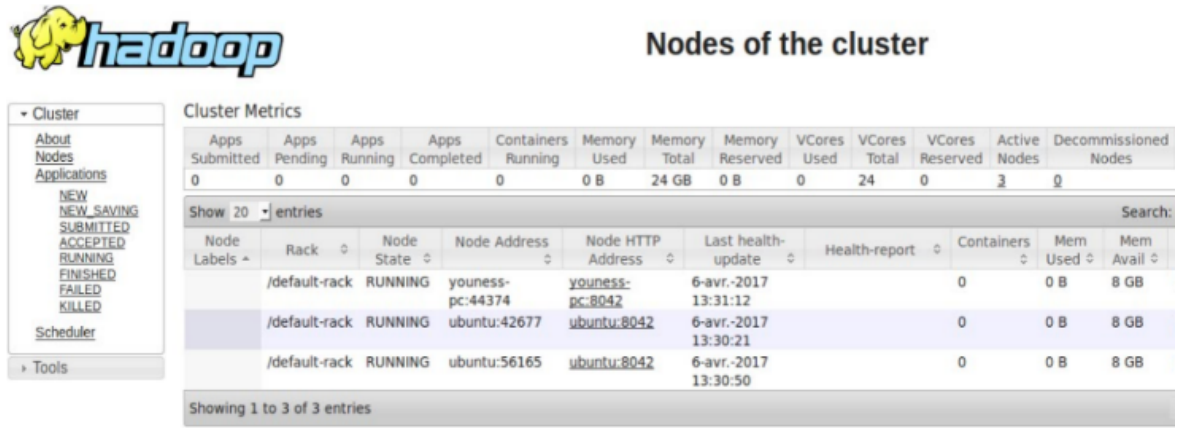

Figure 23: Configuration du Cluster

V. Résultats Expérimentaux et Discussion

V.1 Fuzzy Wup et Lin

Deux mesures sémantiques floues implémentées dans le système sont comparées : Fuzzy WuP et Lin. Les résultats de la DPM floue sont évalués en fonction de trois paramètres de test : précision, rappel et F-mesure. Les résultats présentés aux figures 24 et 25 et dans le

tableau 8 font partie des tests expérimentaux démontrant que Fuzzy_WuP est plus performant que Fuzzy_Lin.

	Fuzzy-WuP	Fuzzy-Lin
Précision	0.54	0.27
Rappel	0.66	0.37
F-Mesure	0.594	0.312

Tableau 8: Précision, rappel et F-mesure pour la détection parallèle floue du plagiat multilingue

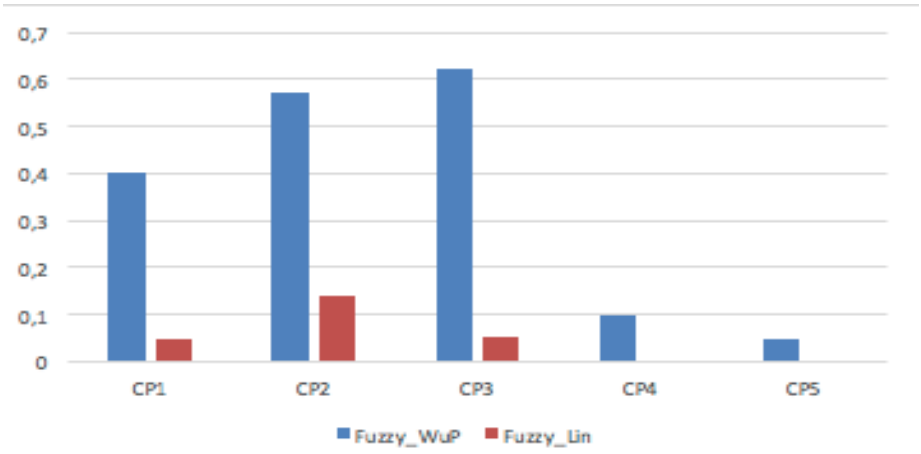

Figure 24: Comparaison de similarité sémantique floue pour des collections parallèles

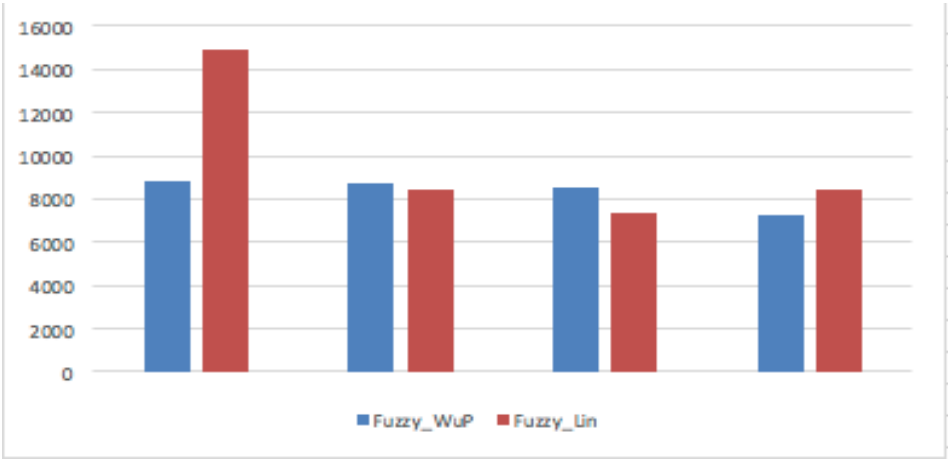

Figure 25: Comparaison de temps d'exécution des mesures de similarité dans un système parallèle

Sur la base des expériences précédentes, les résultats de Fuzzy_WuP sont meilleures que celles de Fuzzy_Lin (Tableau 8). Comme indiqué précédemment, notre corpus de test contient plusieurs formes de plagiat allant de la simple traduction à une autre qui fait changer sérieusement le texte. Lors des tests, nous avons constaté que notre approche combinée avec Lin n'est pas efficace pour détecter les documents plagiés dans une grande masse de données. Cependant, il a donné une performance moyenne de détection plagiat dans les petits textes. Par conséquent, Fuzzy_Lin n'est pas adapté à la détection de plagiat dans de grands volumes

d'informations. Fuzzy_WuP reste la meilleure mesure de similarité sémantique pour la détection de plagiat obscurci dans un environnement Big Data. Par conséquent, la raison principale d'appliquer la théorie floue dans la détection du plagiat (où on cherche tous les synonymes pour chaque mot du gramme) est dû à la façon dont le texte traduit est modifié. On calcule la similarité sémantique pour toutes les combinaisons possibles, d'où le plagiat est toujours détecté.

L'approche de détection de plagiat multilingue parallèle basée sur la logique floue en utilisant la taxonomie lexicale WordNet dans un environnement Big Data donne de bons résultats par rapport à certains modèles et approches existants, à savoir la méthode IR floue dans [145], car les facteurs de corrélation de mots obtenus à partir de grands corpus nécessitant une allocation d'espace disque pour sauvegarder les tables de facteurs de corrélation mot à mot.

V.2 Fuzzy Wup, Lin et LCH

Cette fois, le corpus de test est constitué de 600 documents anglais, français et arabes de différentes sources (actualités, articles, tweets et travaux universitaires). Pour tester le système de détection du plagiat multilingue, 200 documents sont traduits de l'anglais / français en Arabe (Automatique) sans changement, et 400 documents sont traduits et modifiés avec un pourcentage élevé de plagiat (paraphrasant, back-translation, etc.). Les mesures sémantiques floues WuP, Lin et LCH sont mises en œuvre et les résultats sont comparés.

Les résultats des tests expérimentaux sont présentés aux figures 26 et 27.

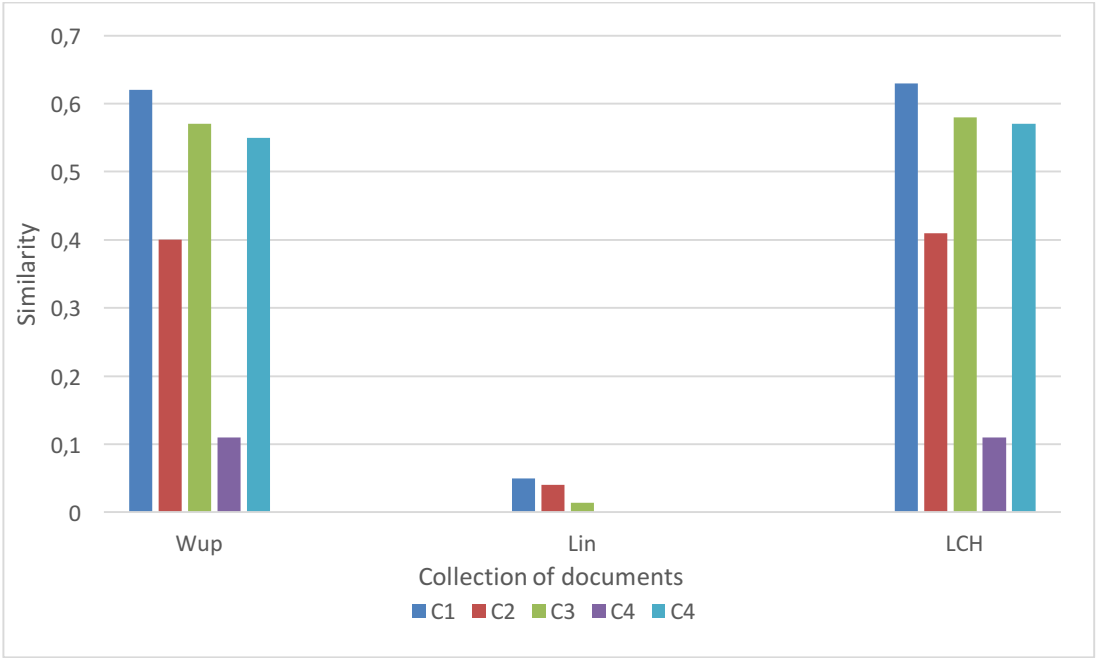

Figure 26: Comparaison de la similarité des mesures de similarité pour le système proposé

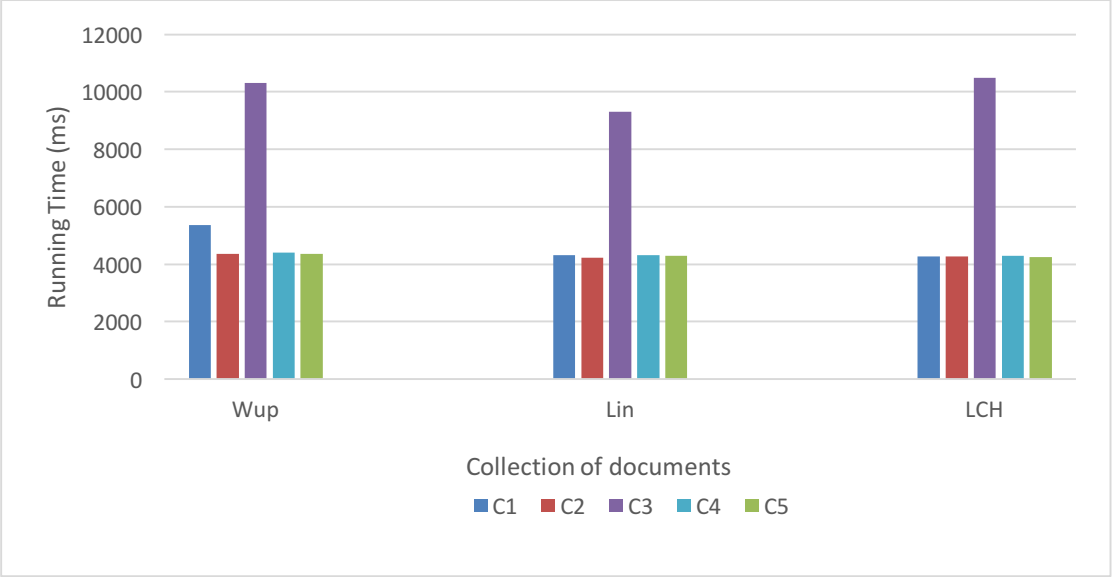

Figure 27: Comparaison de temps d'exécution des mesures de similarité pour le système proposé

Sur la base des résultats de l'évaluation et des résultats précédents dans [8] [17] [18], Fuzzy-Lin donne de bons résultats en termes de temps d'exécution mais de mauvais résultats pour la similarité multilingue, et aussi, il n'est pas efficace pour détecter le plagiat dans une grande masse de données. Par conséquent, les tests précédents sont confirmés, Fuzzy_Lin n'est pas adapté à la détection de plagiat multilingue dans de grands volumes d'informations. Fuzzy_WuP et LCH donnent des résultats similaires en termes de similarité avec des différences notables en temps d'exécution.

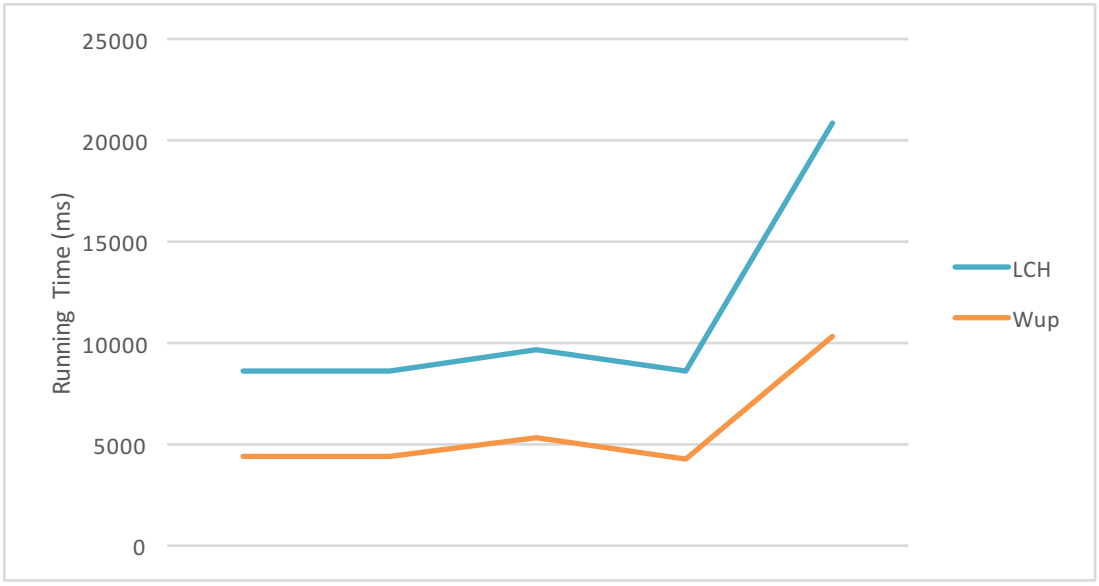

Figure 28: Comparaison du temps d'exécution par taille de document pour Fuzzy_WuP et LCH

Pour le temps d'exécution, nous remarquons un changement dans les résultats de LCH pour les collections des documents un peu volumineux. Nous avons examiné les deux

algorithmes en termes de temps avec des documents de tailles différentes (figure 28). Fuzzy_WuP apparaît être la meilleure mesure de similarité sémantique pour la détection du plagiat obscurci dans un grand corpus.

VI. Conclusion

L'utilisation des technologies Big a pour but la gestion des grandes masse d'informations, et aussi le grand nombre d'opérations et des calculs faites dans un tel système de détection du plagiat multilingue. Notre approche est de travailler dans un système Big Data avec le framework Hadoop basé sur le modèle de programmation MapReduce et le système de fichier HDFS.

Conclusion générale et perspectives

Le plagiat multilingue est la réutilisation non reconnue d'un texte impliquant sa traduction d'une langue naturelle à une autre sans référence appropriée à la source d'origine. Les modèles de détection automatique du plagiat visent à détecter des cas potentiels de réutilisation non autorisée de texte.

Ce travail de recherche a proposé une méthode de détection plagiat sémantique multilingue – pour des langages qui ne partagent aucune caractéristique – sans transformer le problème multilingue en monolingue via l'unification de la langue. Cette méthode est en particulier adaptée à la langue Arabe, qui présente des défis réels en raison de sa structure linguistique complexe.

Pour atteindre cet objectif, nous avons commencé par une étude exhaustive des méthodes de détection du plagiat monolingue et multilingue existants, surtout celles concernant la détection de plagiat multilingue, en se basant sur l'analyse de la relation sémantique des textes en question.

Pour que les textes écrits en des langages naturels soient adaptés à une analyse sémantique, ils doivent être soumis à plusieurs processus de prétraitement, à savoir : la tokénisation, la lemmatisation, l'étiquetage morphosyntaxique (étiquetage de parties de discours), la suppression de mots vides… etc. L'application des différentes techniques de traitement des langages naturels était une phase indispensable pour notre système. L'absence de ces traitements crée des problèmes sérieux, sans parler du nombre énorme des opérations et des calculs que consomment les parties non utiles du texte. Chose qui peut provoquer un résultat erroné des calculs de similarité, et par conséquent, une décision non objective du plagiat pour le document suspect. Ces techniques ont également une influence significative sur la taille et la structuration du texte, certaines améliorent la précision, d'autres diminuent les exigences de temps et d'autres font les deux.

L'intérêt principale de cette thèse est la « détection du plagiat sémantique multilingue dans des documents écrits en langue Arabe ». A cette fin, nous avons proposé une nouvelle approche sémantique implémentant un système à base de corpus parallèles. Pour garder le cadre multilingue de la détection, on s'est basé sur des co-occurrences multilingues et des

modules de traduction. Les principes et les ressources de la traduction automatique (MT) sont appliqués, mais aucune traduction n'est effectivement faite.

Ensuite, l'introduction de la théorie des ensembles floues s'était basée sur l'hypothèse que la correspondance de deux phrases sémantiquement équivalentes est approximative et vague, ce qui peut être modélisé en considérant que chaque mot d'une phrase (dans la 1ère langue) est associé à un ensemble flou qui contient les mots ayant la même signification (dans la 2ème langue). L'introduction de la théorie des ensembles floues dans notre cas de détection de plagiat sémantique multilingue est un besoin pour un confinement efficace d'ingéniosité du plagiaire, qui essaye son maximum de caché "son crime" en manipulant, de différentes méthodes, l'idée original (le sens) du texte.

Dans la partie du calcul, on ne calcule pas la similarité sémantique juste pour les mots du segment (W3G) mais pour toute la synset de chaque mot ; c'est à dire le calcul est fait pour le mot figurant dans le segment et pour tous ses synonymes trouvés en respectant bien sûr l'ordre de l'apparence des mots et des segments dans le texte. Le nombre d'opérations et des calculs faits dans un tel système pour chaque document du corpus avec le document suspect est incroyable, ce qui implique une solution immédiate de ce problème et aussi pour la gestion des grandes masse d'informations et de données, on a utilisé les technologies Big Data.

L'utilisation des technologies Big Data est une étape préliminaire dans de nombreux sujets de recherche et de récupération des informations, bien évidemment la détection de plagiat dans la partie de recherche des documents candidats (c'est-à-dire l'étape de la sélection des documents suspects source de plagiat). Cependant, le nombre de candidats reste énorme pour une analyse très approfondie d'un processus de similarité sémantique flou, car généralement, les collections de documents sont d'énormes corpus ou parfois le Web. Vue le grand nombre d'opérations et des calculs faites dans un tel système de détection du plagiat multilingue et aussi pour la gestion des grandes masse d'informations et de données, on a utilisé des technologies Big Data pour diminuer le temps d'exécution et la distribution du stockage. Notre approche est de travailler dans un système Big Data avec le framework Hadoop basé sur le modèle de programmation MapReduce et le système de fichier HDFS. De ce fait pour la détection du plagiat, nous proposons un algorithme MapReduce basé sur notre approche pour calculer la similarité sémantique floue entre les documents d'une manière parallèle en partageant le travail entre différentes machines (nœuds) du Cluster.

Le travail réalisé dans le cadre de cette thèse est un domaine de recherche important et ne cesse de croitre, ainsi en perspective, nous pensons qu'il est intéressant :

- D’ajouter d'autres informations figurant dans WordNet “Gloss“ afin d’améliorer le calcul de similarité de documents pour une détection de plagiat plus précise.
- D’améliorer notre approche en utilisant la désambiguïsation du sens des mots.
- D’intégrer notre approche dans un système complet de détection de plagiat multilingue (en gérant les citations et les références).
- D’améliorer notre approche dans le système Big Data en intégrant un traitement à temps réel basé sur un système de streaming capable de transférer les données textuelles vers notre système de détection de plagiat. Les données récupérées par cette technique peuvent être utilisées par la suite comme un corpus dans notre système, afin d’utiliser d’autres techniques de machine Learning comme les réseaux de neurones dans la détection de plagiat.

Bibliographie

[1] C. Vandendorpe et M. Bénabou, *Le Plagiat: actes du colloque tenu à l'Université d'Ottawa, du 26 au 28 septembre 1991*. Presses de l'Université d'Ottawa, 1992.

[2] Martial, *Les épigrammes de Martial*. Bibliothèque nationale de France. J. Chapelle (Paris).

[3] « A Plagiarism FAQ ». [En ligne]. Disponible sur: https://www.ieee.org/publications/rights/plagiarism/plagiarism-faq.html. [Consulté le: 02-avr-2019].

[4] F. K. Taylor, « Cryptomnesia and plagiarism », *The British Journal of Psychiatry*, vol. 111, n° 480, p. 1111–1118, 1965.

[5] M. Roig, *Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing*. 2006.

[6] M. Jones, « Back-translation: The latest form of plagiarism », 2009.

[7] M. K. Summers, W. B. Evans, J. J. Fletcher, C. Hanchey, et L. J. Waguespack, « Program plagiarism revisited: Current issues and approaches », *ACM SIGCSE Bulletin*, vol. 20, n° 1, p. 224–224, 1988.

[8] K. J. Ottenstein, « An algorithmic approach to the detection and prevention of plagiarism », *ACM Sigcse Bulletin*, vol. 8, n° 4, p. 30–41, 1976.

[9] S. Brin, J. Davis, et H. Garcia-Molina, « Copy detection mechanisms for digital documents », in *ACM SIGMOD Record*, 1995, vol. 24, p. 398–409.

[10] N. Shivakumar et H. Garcia-Molina, « SCAM: A copy detection mechanism for digital documents », 1995.

[11] S. M. Alzahrani, N. Salim, et A. Abraham, « Understanding plagiarism linguistic patterns, textual features, and detection methods », *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, n° 2, p. 133–149, 2012.

[12] S. Alzahrani et N. Salim, « Fuzzy semantic-based string similarity for extrinsic plagiarism detection », *Braschler and Harman*, vol. 1176, p. 1–8, 2010.

[13] D. Gupta, « Study on Extrinsic Text Plagiarism Detection Techniques and Tools. »,

Journal of Engineering Science & Technology Review, vol. 9, n° 5, 2016.

[14] B. Stein, N. Lipka, et P. Prettenhofer, « Intrinsic plagiarism analysis », *Language Resources and Evaluation*, vol. 45, n° 1, p. 63–82, 2011.

[15] M. Potthast, A. Barrón-Cedeño, B. Stein, et P. Rosso, « Cross-language plagiarism detection », *Language Resources and Evaluation*, vol. 45, n° 1, p. 45–62, 2011.

[16] A. Barrón-Cedeño, P. Rosso, et J.-M. Benedí, « Reducing the plagiarism detection search space on the basis of the kullback-leibler distance », in *International conference on intelligent text processing and computational linguistics*, 2009, p. 523–534.

[17] Potthast, Stein, B., BarrónCedeno, A., & Rosso, P, *Pan plagiarism corpus pan-pc-11*. 2011.

[18] F. W. Lancaster, « Vocabulary control for information retrieval. », 1972.

[19] L. Cerulo et G. Canfora, « A taxonomy of information retrieval models and tools », *Journal of Computing and Information Technology*, vol. 12, n° 3, p. 175–194, 2004.

[20] C. D. Manning, P. Raghavan, et H. Schütze, *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge, 2008.

[21] N. Heintze, « Scalable document fingerprinting », in *1996 USENIX workshop on electronic commerce*, 1996, vol. 3.

[22] C. K. Kent et N. Salim, « Features based text similarity detection », *arXiv preprint arXiv:1001.3487*, 2010.

[23] B. Stein, « Principles of hash-based text retrieval », in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, p. 527–534.

[24] S. Schleimer, D. S. Wilkerson, et A. Aiken, « Winnowing: local algorithms for document fingerprinting », in *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, 2003, p. 76–85.

[25] M. Dillon, *Introduction to modern information retrieval: G. Salton and M. McGill. McGraw-Hill, New York (1983). xv+ 448 pp., $32.95 ISBN 0-07-054484-0*. Pergamon, 1983.

[26] M. Zechner, M. Muhr, R. Kern, et M. Granitzer, « External and intrinsic plagiarism detection using vector space models », in *Proc. SEPLN*, 2009, vol. 32, p. 47–55.

[27] C. Grozea, C. Gehl, et M. Popescu, « ENCOPLOT: Pairwise sequence matching in linear time applied to plagiarism detection », in *3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, 2009, p. 10.

[28] C. Basile, D. Benedetto, E. Caglioti, G. Cristadoro, et M. D. Esposti, « A plagiarism detection procedure in three steps: Selection, matches and "squares" », in *Proc. SEPLN*, 2009, p. 19–23.

[29] C. D. Manning, P. Raghavan, et H. Schütze, « Matrix decompositions and latent semantic indexing », *Introduction to Information Retrieval*, p. 403–417, 2008.

[30] C. H. Ding, « A similarity-based probability model for latent semantic indexing », in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, p. 58–65.

[31] H. Chen, K. J. Lynch, K. Basu, et T. D. Ng, « Generating, integrating, and activating thesauri for concept-based document retrieval », *IEEE Expert*, vol. 8, n° 2, p. 25–34, 1993.

[32] Z. Ceska, « Plagiarism detection based on singular value decomposition », in *Advances in natural language processing*, Springer, 2008, p. 108–119.

[33] Z. Ceska, « Automatic plagiarism detection based on latent semantic analysis », PhD Thesis, Ph. D. dissertation, Faculty Appl Sci., Univ. West Bohemia, Pilsen, Czech Republic, 2009.

[34] H. Zhang et T. W. Chow, « A coarse-to-fine framework to efficiently thwart plagiarism », *Pattern Recognition*, vol. 44, n° 2, p. 471–487, 2011.

[35] V. Cross, « Fuzzy information retrieval », *Journal of Intelligent Information Systems,* p. 29-56, 1994.

[36] Y. Ogawa, T. Morita, and K. Kobayashi, « A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method », *Fuzzy sets and systems, Elsevier*, p. 163-179, 1991.

[37] D. Dubois and H. Prade, « An introduction to fuzzy systems1 », *Clinica Chimica Acta*, vol. 270, n° 1, p. 3-29, févr. 1998.

[38] R. Yerra and Y.-K. Ng, « A sentence-based copy detection approach for web documents », présenté à International Conference on Fuzzy Systems and Knowledge

Discovery, Berlin, Heidelberg, 2005, p. 557-570.

[39] S. K. Bhatia and J. S. Deogun, « Conceptual clustering in information retrieval », *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics*, p. 427-436, juin-1998.

[40] Liu, X., & Croft, W. B., « Cluster-based retrieval using language models », présenté à 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom, 2004, p. 186-193.

[41] T. Kohonen, *Self-Organization and Associative Memory*, vol. 8. Springer Science & Business Media., 2012.

[42] M. K. M. Rahman, W. Pi Yang, T. W. S. Chow, and S. Wu, « A flexible multi-layer self-organizing map for generic processing of tree-structured data », *Pattern Recognition*, vol. 40, p. 1406-1424, mai 2007.

[43] S. K. Pal, S. Mitra, and P. Mitra, « Soft Computing Pattern Recognition, Data Mining and Web Intelligence », *Intelligent Technologies for Information Analysis,Springer, Berlin, Heidelberg.*, p. 475-512.

[44] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, « WEBSOM – Self-organizing maps of document collections », vol. 21, n° 1-3, p. 101-117, nov. 1998.

[45] S. Antonio, L. Hong Va, and W. H. L. Rynson, ", « CHECK: a document plagiarism detection system », présenté à 1997 ACM Symposium Applied Computing, San Jose, United States, p. 70-77.

[46] L. Ballesteros et W. B. Croft, « Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval », in *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 1997, p. 84–91.

[47] P. Virga et S. Khudanpur, « Transliteration of Proper Names in Cross-lingual Information Retrieval », in *Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition - Volume 15*, Stroudsburg, PA, USA, 2003, p. 57–64.

[48] A. Bellaachia et G. Amor-Tijani, « Proper nouns in English–Arabic cross language

information retrieval », *Journal of the Association for Information Science and Technology*, vol. 59, n° 12, p. 1925–1932, 2008.

[49] S. Pourmahmoud et M. Shamsfard, « Semantic cross-lingual information retrieval », in *Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on*, 2008, p. 1–4.

[50] Xu, J., & Weischedel, R, « Cross-lingual information retrieval using hidden Markov models », présenté à EMNLP '00 Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 2000, vol. 13, p. 95-103.

[51] E.-E. Jan, S.-H. Lin, et B. Chen, « Transliteration retrieval model for cross lingual information retrieval », in *Asia Information Retrieval Symposium*, 2010, p. 183–192.

[52] Luo, Y., Le, Z., & Wang, M., « Cross-lingual information retrieval model based on bilingual topic correlation », *Journal of Computational Information Systems*, p. 2433-2440, 2013.

[53] Vulić, I., & Moens, M. F., « Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. », présenté à 38th international ACM SIGIR conference on research and development in information retrieval. ACM, Santiago, Chile, 2015, p. 363-372.

[54] J. Allan, J. G. Carbonell, G. Doddington, J. Yamron, et Y. Yang, « Topic detection and tracking pilot study final report », 1998.

[55] Lavrenko, V., Choquette, M., & Croft, W. B., « Cross-Lingual Relevance Models », présenté à 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002.

[56] C. E. Shannon, A. D. Wyner, et N. J. Sloane, *Claude E. Shannon: collected papers*. John Wiley & Sons, 1993.

[57] A. Barrón-Cedeno, P. Rosso, E. Agirre, et G. Labaka, « Plagiarism detection across distant language pairs », in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, p. 37–45.

[58] J. Kasprzak, M. Brandejs, et M. Kripac, « Finding plagiarism by evaluating document

similarities », in *Proc. SEPLN*, 2009, vol. 9, p. 24–28.

[59] G. Navarro, « A guided tour to approximate string matching », *ACM computing surveys (CSUR)*, vol. 33, n° 1, p. 31–88, 2001.

[60] W. Cohen, P. Ravikumar, et S. Fienberg, « A comparison of string metrics for matching names and records », in *Kdd workshop on data cleaning and object consolidation*, 2003, vol. 3, p. 73–78.

[61] V. Scherbinin et S. Butakov, « Using Microsoft SQL server platform for plagiarism detection », in *Proc. SEPLN*, 2009, p. 36–37.

[62] Z. Su, B.-R. Ahn, K.-Y. Eom, M.-K. Kang, J.-P. Kim, et M.-K. Kim, « Plagiarism detection using the Levenshtein distance and Smith-Waterman algorithm », in *Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on*, 2008, p. 569–569.

[63] M. S. Waterman, « Identification of common molecular subsequence », *Mol. Biol*, vol. 147, p. 195–197, 1981.

[64] M. Elhadi et A. Al-Tobi, « Duplicate detection in documents and webpages using improved longest common subsequence and documents syntactical structures », in *2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, 2009.

[65] A. Parker et J. O. Hamblen, « Computer algorithms for plagiarism detection », *IEEE Transactions on Education*, vol. 32, n° 2, p. 94–99, 1989.

[66] A. Barrón-Cedeno, P. Rosso, D. Pinto, et A. Juan, « On Cross-lingual Plagiarism Analysis using a Statistical Model. », in *PAN*, 2008, p. 1–10.

[67] G. Grefenstette, « COMPARING TWO LANGUAGE IDENTIFICATION SCHEM ES », 1995.

[68] T. Dunning, *Statistical identification of language*. Computing Research Laboratory, New Mexico State University, 1994.

[69] E. Millar, D. Shen, J. Liu, et C. Nicholas, « Performance and scalability of a large-scale n-gram based information retrieval system », *Journal of digital information*, vol. 1, n° 5, 2006.

[70] B. Pouliquen, R. Steinberger, et C. Ignat, « Automatic identification of document translations in large multilingual document collections », *arXiv preprint cs/0609060*, 2006.

[71] A. Ekbal, S. Saha, et G. Choudhary, « Plagiarism detection in text using vector space model », in *Hybrid Intelligent Systems (HIS), 2012 12th International Conference on*, 2012, p. 366–371.

[72] S. Qinbao et S. Junyi, « On illegal coping and distributing detection mechanism for digital goods », *Journal of computer research and development*, vol. 38, n° 1, p. 121–125, 2001.

[73] M. Murugesan, W. Jiang, C. Clifton, L. Si, et J. Vaidya, « Efficient privacy-preserving similar document detection », *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 19, n° 4, p. 457–475, 2010.

[74] A. Barrón-Cedeño, C. Basile, M. Degli Esposti, et P. Rosso, « Word length n-Grams for text re-use detection », in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2010, p. 687–699.

[75] C. Lyon, J. Malcolm, et B. Dickerson, « Detecting short passages of similar text in large document collections », in *Proceedings of the*, 2001.

[76] D. R. White et M. S. Joy, « Sentence-based natural language plagiarism detection », *Journal on Educational Resources in Computing (JERIC)*, vol. 4, n° 4, p. 2, 2004.

[77] D. R. White et M. S. Joy, « Sentence-based natural language plagiarism detection », *Journal on Educational Resources in Computing (JERIC)*, vol. 4, n° 4, p. 2, 2004.

[78] J. Kasprzak, M. Brandejs, et M. Kripac, « Finding plagiarism by evaluating document similarities », in *Proc. SEPLN*, 2009, vol. 9, p. 24–28.

[79] C. Lyon, J. Malcolm, et B. Dickerson, « Detecting short passages of similar text in large document collections », in *Proceedings of the*, 2001.

[80] H. Zhang et T. W. Chow, « A coarse-to-fine framework to efficiently thwart plagiarism », *Pattern Recognition*, vol. 44, n° 2, p. 471–487, 2011.

[81] A. Barrón-Cedeño et P. Rosso, « On automatic plagiarism detection based on n-grams comparison », in *European Conference on Information Retrieval*, 2009, p. 696–700.

[82] M. Murugesan, W. Jiang, C. Clifton, L. Si, et J. Vaidya, « Efficient privacy-preserving

similar document detection », *The VLDB Journal—The International Journal on Very Large Data Bases*, vol. 19, n° 4, p. 457–475, 2010.

[83] M. Zechner, M. Muhr, R. Kern, et M. Granitzer, « External and intrinsic plagiarism detection using vector space models », in *Proc. SEPLN*, 2009, vol. 32, p. 47–55.

[84] A. Parker et J. O. Hamblen, « Computer algorithms for plagiarism detection », *IEEE Transactions on Education*, vol. 32, n° 2, p. 94–99, 1989.

[85] M. Elhadi et A. Al-Tobi, « Use of text syntactical structures in detection of document duplicates », in *Digital Information Management, 2008. ICDIM 2008. Third International Conference on*, 2008, p. 520–525.

[86] K. Vani et D. Gupta, « Investigating the impact of combined similarity metrics and POS tagging in extrinsic text plagiarism detection system », in *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on*, 2015, p. 1578–1584.

[87] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, et D. McClosky, « The Stanford CoreNLP natural language processing toolkit », in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, p. 55–60.

[88] R. Rada, H. Mili, E. Bicknell, et M. Blettner, « Development and application of a metric on semantic nets », *IEEE transactions on systems, man, and cybernetics*, vol. 19, n° 1, p. 17–30, 1989.

[89] R. Richardson, A. Smeaton, et J. Murphy, *Using WordNet as a knowledge base for measuring semantic similarity between words*. Technical Report Working Paper CA-1294, School of Computer Applications, Dublin City University, 1994.

[90] G. Hirst et D. St-Onge, « Lexical chains as representations of context for the detection and correction of malapropisms », *WordNet: An electronic lexical database*, vol. 305, p. 305–332, 1998.

[91] M. Choudhari, « Extending the hirst and St-Onge measure of semantic relatedness for the unified medical language system. », 2012.

[92] Z. Wu et M. Palmer, « Verbs semantics and lexical selection », in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, p. 133–138.

[93] G. A. Miller, « WordNet: a lexical database for English », *Communications of the ACM*, vol. 38, n° 11, p. 39–41, 1995.

[94] G. A. Miller, « WordNet: a lexical database for English », *Communications of the ACM*, vol. 38, n° 11, p. 39–41, 1995.

[95] Y. Li, Z. A. Bandar, et D. McLean, « An approach for measuring semantic similarity between words using multiple information sources », *IEEE Transactions on knowledge and data engineering*, vol. 15, n° 4, p. 871–882, 2003.

[96] C. Leacock et M. Chodorow, « Combining local context and WordNet similarity for word sense identification », *WordNet: An electronic lexical database*, vol. 49, n° 2, p. 265–283, 1998.

[97] P. Resnik, « Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language », *J. Artif. Intell. Res.(JAIR)*, vol. 11, p. 95–130, 1999.

[98] D. Lin, « An information-theoretic definition of similarity. », in *Icml*, 1998, vol. 98, p. 296–304.

[99] D. Lin, « Principle-based parsing without overgeneration », in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 1993, p. 112–120.

[100] J. J. Jiang et D. W. Conrath, « Semantic similarity based on corpus statistics and lexical taxonomy », *arXiv preprint cmp-lg/9709008*, 1997.

[101] A. Tversky, « Features of similarity. », *Psychological review*, vol. 84, n° 4, p. 327, 1977.

[102] E. G. Petrakis, G. Varelas, A. Hliaoutakis, et P. Raftopoulou, « X-similarity: Computing semantic similarity between concepts from different ontologies. », *Journal of Digital Information Management*, vol. 4, n° 4, 2006.

[103] J. Koberstein et Y.-K. Ng, « Using word clusters to detect similar web documents », in *International Conference on Knowledge Science, Engineering and Management*, 2006, p. 215–228.

[104] S. Alzahrani et N. Salim, « Statement-based fuzzy-set IR versus fingerprints matching for plagiarism detection in arabic documents », in *Proc. 5th Postgraduate Annu. Res. Seminar*, 2009, p. 267–268.

[105] A. H. Osman, N. Salim, Y. J. Kumar, et A. Abuobieda, « Fuzzy semantic plagiarism detection », in *International Conference on Advanced Machine Learning Technologies and Applications*, 2012, p. 543–553.

[106] D. Gupta, K. Vani, et C. K. Singh, « Using Natural Language Processing techniques and fuzzy-semantic similarity for automatic external plagiarism detection », in *Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on*, 2014, p. 2694–2699.

[107] R. C. Pereira, V. P. Moreira, et R. Galante, « A new approach for cross-language plagiarism analysis », in *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2010, p. 15–26.

[108] M. Potthast, B. Stein, et M. Anderka, « A Wikipedia-based multilingual retrieval model », in *European conference on information retrieval*, 2008, p. 522–530.

[109] M. Simard, G. F. Foster, et P. Isabelle, « Using cognates to align sentences in bilingual corpora », in *Proceedings of the 1993 conference of the Centre for Advanced Studies on Collaborative research: distributed computing-Volume 2*, 1993, p. 1071–1082.

[110] K. W. Church, « Char_align: a program for aligning parallel texts at the character level », in *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 1993, p. 1–8.

[111] R. Steinberger, B. Pouliquen, et J. Hagman, « Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc », in *International Conference on Intelligent Text Processing and Computational Linguistics*, 2002, p. 415–424.

[112] P. Vossen, *A multilingual database with lexical semantic networks*. Springer, 1998.

[113] M. L. Littman, S. T. Dumais, et T. K. Landauer, « Automatic cross-language information retrieval using latent semantic indexing », in *Cross-language information retrieval*, Springer, 1998, p. 51–62.

[114] C. K. Kent et N. Salim, « Web based cross language plagiarism detection », in *Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on*, 2010, p. 199–204.

[115] R. Nawab, M. Stevenson, et P. Clough, « University of sheffield: Lab report for PAN at CLEF 2010 », in *CLEF 2010 LABs and Workshops, Notebook Papers*, 2010.

[116] G. Oberreuter, G. L'Huillier, S. A. Rıos, et J. D. Velásquez, « Approaches for intrinsic and external plagiarism detection », *Proceedings of the PAN*, 2011.

[117] J. Grman et R. Ravas, « Improved implementation for finding text similarities in large collections of data », *Proceedings of PAN*, 2011.

[118] D. Jurafsky et J. H. Martin, *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, Pearson Education International, 2009.

[119] « Mot composé », *Wikipédia*. 01-janv-2018.

[120] M. Khosrow-Pour, *Encyclopedia of information science and technology*, vol. 1. IGI Global, 2008.

[121] B. Alhadidi et M. Alwedyan, « Hybrid Stop-Word Removal Technique for Arabic Language. », *Egyptian Computer Science Journal*, vol. 30, n° 1, p. 35–38, 2008.

[122] J. K. Raulji et J. R. Saini, « Generating Stopword List for Sanskrit Language », in *Advance Computing Conference (IACC), 2017 IEEE 7th International*, 2017, p. 799–802.

[123] R. Feldman et J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[124] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, et K. J. Miller, « Introduction to WordNet: An on-line lexical database », *International journal of lexicography*, vol. 3, n° 4, p. 235–244, 1990.

[125] J. Charlet, P. Laublet, et C. Reynaud, *Le web sémantique*. Cépaduès-Ed., 2003.

[126] D. Fensel, F. Van Harmelen, I. Horrocks, D. L. McGuinness, et P. F. Patel-Schneider, « OIL: An ontology infrastructure for the semantic web », *IEEE intelligent systems*, vol. 16, n° 2, p. 38–45, 2001.

[127] T. Berners-Lee, *Semantic web road map*. 1998.

[128] P. Castells, M. Fernández Sánchez, et D. J. Vallet Weadon, « An adaptation of the vector-space model for ontology based information retrieval », *IEEE transactions on knowledge and data engineering*, 2007.

[129] J. Li, J.-Y. Song, et H. Zhong, « Ontology-based query division and reformulation for heterogeneous information integration », *Ruan Jian Xue Bao(Journal of Software)*, vol. 18, n°

10, p. 2495–2506, 2007.

[130] T. R. Gruber, « A translation approach to portable ontology specifications », *Knowledge acquisition*, vol. 5, n° 2, p. 199–220, 1993.

[131] F. Fürst, « Contribution à l'ingénierie des ontologies: une méthode et un outil d'opérationalisation », PhD Thesis, Nantes, 2004.

[132] S. Bechhofer, « OWL: Web ontology language », in *Encyclopedia of database systems*, Springer, 2009, p. 2008–2009.

[133] N. Guarino, *Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy*, vol. 46. IOS press, 1998.

[134] A. Budanitsky et G. Hirst, « Evaluating wordnet-based measures of lexical semantic relatedness », *Computational Linguistics*, vol. 32, n° 1, p. 13–47, 2006.

[135] P. Resnik, « Using information content to evaluate semantic similarity in a taxonomy », *arXiv preprint cmp-lg/9511007*, 1995.

[136] « Buckwalter Arabic Transliteration ». [En ligne]. Disponible sur: http://www.qamus.org/transliteration.htm. [Consulté le: 22-sept-2018].

[137] E. Hanane, M. Erritali, et M. Oukessou, « Semantic Similarity/Relatedness for Cross language plagiarism detection », in *Computer Graphics, Imaging and Visualization (CGiV), 2016 13th International Conference on*, 2016, p. 372–374.

[138] H. Ezzikouri, M. Erritali, et M. Oukessou, « Fuzzy-Semantic Similarity for Automatic Multilingual Plagiarism Detection », *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 8, n° 9, p. 86–90, 2017.

[139] M. A. Beyer et D. Laney, « The importance of 'big data': a definition », *Stamford, CT: Gartner*, p. 2014–2018, 2012.

[140] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, et E. Muharemagic, « Deep learning applications and challenges in big data analytics », *Journal of Big Data*, vol. 2, p. 1, févr. 2015.

[141] B. Parhami, « A Highly Parallel Computing System for Information Retrieval », in *Proceedings of the December 5-7, 1972, Fall Joint Computer Conference, Part II*, New York, NY, USA, 1972, p. 681–690.

[142] Q. Zhang, Y. Zhang, H. Yu, et X. Huang, « Efficient partial-duplicate detection based on sequence matching », in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010, p. 675–682.

[143] M. Erritali, A. Beni-Hssane, M. Birjali, et Y. Madani, « An approach of semantic similarity measure between documents based on big data », *International Journal of Electrical and Computer Engineering*, vol. 6, n° 5, p. 2454, 2016.

[144] J. Dwivedi et A. Tiwary, « Plagiarism detection on bigdata using modified map-reduced based SCAM algorithm », in *Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on*, 2017, p. 608–610.

[145] R. Yerra and Y.-K. Ng, « A sentence-based copy detection approach for web documents », présenté à International Conference on Fuzzy Systems and Knowledge Discovery, Berlin, Heidelberg, 2005, p. 557-570.