A comparative Plagiarism Detection System methods between sentences

A. ZOUHIR¹, R. EL AYACHI¹, M. BINIZ²

¹Laboratory of Information Processing and Decision Support (TIAD), Faculty of Sciences and Technics, University of Sultan Moulay Slimane, Beni Mellal, Morocco

²Laboratory LIMATI, Pluridisciplinary Faculty Beni Mellal, Sultan Moulay Slimane University, Beni Mellal, Morocco

E-mail: assiazouhir94@gmail.com, rachid.elayachi@usms.com, mohamedbiniz@gmail.com

Abstract. After the era of the World Wide Web, information is easily accessible with a single click. But this progression has drawbacks despite the ease of access to information. Plagiarism has a growing challenge to society, which impact on the academic world, researchers, and students in particular. This work discusses the plagiarism process, types, and detection methodologies. It presents the different plagiarism detection techniques based on syntactic and semantic approaches. The result of this work is a comparative survey of plagiarism detection system methods using the identification of syntactic and semantic similarities based a sentence-to-sentence comparison, and no longer word-to-word like the classical systems because the similarity between the sentences is a complex phenomenon.

1. Introduction

For the past decade, the abundance of online resources and ease of internet access has increased the number of papers copied without asking permission or indicating the source, this practice is very popular, known as “plagiarism” where there is, therefore, no reliable protection of the originality of the contents.

For as long as there has been plagiarism in all century, but it wasn't considered a serious sort of violation. However, with the development that our world has known and especially by moving from analog to digital information and taking plagiarism with it, this concept becomes having other dimensions.

Although plagiarism affects many fields (political, literary, artistic, etc.), it is a phenomenon that spreads widely in the academic world (universities, schools, and institutions) as well as in all kinds of intellectual work.

Therefore, plagiarism detection has become one of the educational challenges because most students or researchers commit plagiarism without their knowledge, they have copied works of others without asking for authorization or indicating the source. So, the need to protect and keep the originality of the content gave birth to anti-plagiarism software.
In this paper, we are interested in overcoming the problems being faced in classical plagiarism detection systems based on syntactical similarity measures, by presenting a comparative study of the different techniques for the detection of Plagiarism, and secondly, by developing a plagiarism detection system based on syntactic and semantic similarity measures, that compares the two texts, sentence by sentence, and no longer word by word, by looking if a sentence in one of the texts has the same meaning as a sentence in the other text.

The rest of the paper is organized as follows:
The first section defines plagiarism, its detection, and its different types and methods. The second section is devoted entirely to the description of the notion of similarity, syntactic and semantic, between texts and particularly between sentences. The implementation of our plagiarism detection system and discussing the results obtained are presented in the third section.

2. Related works

2.1. Plagiarism
Plagiarism is an act of fraud, it's the act of appropriating the work of another and presenting it as one's own, such as copying from a book, or scientific articles, thus stealing ideas, images, videos, and music and integrate them into your own work without mentioning the source.

2.2. Plagiarism detection
Plagiarism detection is the process of localizing plagiarism cases in a paper or document using a software, based on searching, comparing large collections of documents with the suspect document, and detecting at the end if the suspect document is plagiarized or it is unique (Figure 1).

2.3. Plagiarism types
Textual plagiarism takes several forms which can be classified into five categories [1]:
- "Copy-Paste": copy word for word, all or part of a text into another based on one or more sources.
- Paraphrase: reformulate the ideas of another using our own words, (i.e.) by simply authorizing the change of grammar, the use of synonyms for words, the addition, deletion, and substitution of words.
- Plagiarism with translation: the contents are translated and used without reference to the original work.
- The use of false references: adding references that are false or that do not even exist.

![Figure 1. The plagiarism detection process](image-url)
Plagiarism of ideas: This is a type of plagiarism that is difficult to detect because it is not a matter of simple manipulations made on the text, but of reformulation which allows all textual modifications provided that the meaning of the sentence is preserved.

Copy/paste is considered as a literal plagiarism. On the other hand, paraphrase, plagiarism by translation, and plagiarism of ideas are considered as cases of intelligent plagiarism [2].

2.4. Types of plagiarism detection approaches
Plagiarism detection can be classified into monolingual and multilingual based on the homogeneity or heterogeneity of the languages of the compared textual documents [3]:
- Monolingual plagiarism detection treats the identification and extraction of plagiarism in a homogeneous linguistic environment, for example, English-English plagiarism.
- Multilingual plagiarism detection deals with the identification and extraction of plagiarism in a multilingual environment, for example, English-Arabic plagiarism.

The detection of monolingual plagiarism is divided into two tasks, extrinsic and intrinsic detection [4]:
- Extrinsic detection is the fact of finding similarities between a document and a corpus of probable sources.
- Intrinsic detection is based on stylometric analysis which allows studying the style of the author.

However, the detection of extrinsic plagiarism breaks down into several major types of methods, the most interesting for us, those are methods based on syntactic comparison and the others which are based on semantic comparison.

3. Similarity measures

3.1. Definition of similarity
Similarity is the state of being similar, but similar refers to a resemblance in appearance, shape, character, or quantity, without being identical. There are two types of similarity: syntactic similarity and semantic similarity.

3.2. Syntactic similarity

3.2.1. Algorithms based on Edit distance. The edit distance [5] is used to compare two words, two strings of characters, and, more generally, two sequences by calculating the edit distance between them, which is mean by calculating the cost of the best sequence of editing operations to transform one string into another. Editing operations are insertion, deletion, and substitution. There are several variants to calculate the editing distance, Levenshtein Distance was the interesting one. Levenshtein's distance (1966) [5] is defined as the minimum number of edit operations to transform one string into another, the authorized edit operations are the insertion, deletion, and substitution of a single character.

3.2.2. Algorithms based on vector space model
The measures based on the vector space model are used to represent text documents (and all objects, in general) as vectors in multidimensional space.
Among the most common similarity measures that are based on the vector space model are Cosine similarity and Jaccard coefficient.
- TF-Idf & Cosine similarity:
The similarity of cosine [6] between two documents a and b is the similarity of their vector representations which measures the cosine of the angle Θ between the two vectors.
Cosine \( \vec{d}_A, \vec{d}_B \) = \( \frac{\vec{d}_A \cdot \vec{d}_B}{\|\vec{d}_A\| \|\vec{d}_B\|} = \frac{\sum_{t \in V} w(t, d_A) \times w(t, d_B)}{\sqrt{\sum_{t \in V} w(t, d_A)^2} \sqrt{\sum_{t \in V} w(t, d_B)^2}} \) \( (1) \)

where \( w(t, d_A) \) is the weight of word \( t \) in \( d_A \) that is, the \( tf \times idf \) weight, based on the number of occurrences of \( t \) in \( d_A \) and \( V \) indicate the size of the vocabulary of the document collection.

To calculate the cosine similarity between the documents, we must first convert the documents / sentences / words into vectors with real values. We choose the TF-IDF method because with the inverse frequency of the documents of a term (IDF), we evaluate the rarity of a word, we compare the use of the keywords of an individual page to a large corpus of documents which contain the keyword. It decreases the weight of commonly used words and increases the weight of the original keyword.

The TF-IDF method \([7]\) (in English: Term Frequency-Inverse Document Frequency) is a metric used to evaluate the importance of the terms in documents. Its formula is \( (2) \), with: \( tf(t) \): is the frequency of the term in a document, \( idf(t) \): is the reverse document frequency.

\[
\text{tf}(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}
\]

\[
\text{idf}(t) = \log_{\text{e}}(\text{Total number of documents} / \text{Number of documents with term } t \text{ in it})
\]

- **Jaccard’s coefficient:**

  For two sets, it is defined as the cardinality of their intersection divided by the cardinality of their union. The documents \( d_A \) and \( d_B \) are therefore represented, not as vectors, but as sets of terms \([8]\).

  The similarity obtained \( \text{Jaccard} (d_A, d_B) \in [0, 1] \).

\[
\text{Jaccard} (A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (3)
\]

It is also possible to use the vector representation:

\[
\text{Jaccard} (\vec{A}, \vec{B}) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \|\vec{B}\|} \quad (4)
\]

So, Syntactic similarity has many advantages but also disadvantages, the methods based on the syntactic approach do not take into board the semantics. Therefore, taking into consideration the semantics seems more important.

### 3.3. Semantic similarity

In linguistics, semantic is the study of the meaning of words, their structure, and their relationships with other words. So, Semantic similarity consists in quantifying or measuring the degree of resemblance of two terms or concepts compared to their signification or semantic content, and which have no syntactic resemblance. It can be based on calculations using only the semantic network or the semantic network and the corpus \([9]\). Several measures to determine the semantic similarity between terms or sentences have been proposed in the literature and most of them have been tested on WordNet as a semantic network.

WordNet is a lexical database with large coverage, in which words are divided into categories and organized in a hierarchy of nodes. Each node has a unique identifier and represents a concept, or Synset (set of synonyms), it is used to record the different semantic relationships between these sets of synonyms \([10]\).

- **Mihalcea's measure:**

  Based on a metric for word-for-word similarity and a measure of word specificity, we define the semantic similarity of two sentences \( S_1 \) and \( S_2 \) using a metric that combines the semantic similarities of each sentence compared with the other sentence. Here, the talking is about Mihalcea's measure, a measure based on the principle of calculated similarity by maximizing the sum of the similarities between the terms of the two sentences according to this formula \( (5) \) \([11]\).

\[
\text{Sim}_{\text{Mihalcea}} (S_1, S_2) = \frac{1}{2} \left( \frac{\sum_{w \in [S_1]} \text{max} \{\text{sim}(w, S_2) \times \text{idf}(w)\}}{\sum_{w \in [S_1]} \text{idf}(w)} + \frac{\sum_{w \in [S_2]} \text{max} \{\text{sim}(w, S_1) \times \text{idf}(w)\}}{\sum_{w \in [S_2]} \text{idf}(w)} \right) \quad (5)
\]
Where maxSim \((w, T)\) is the maximum score between the word \(w\) and the words in the sentence \(T\) according to a measure of word-for-word similarity, and idf \((w)\) is the reverse document frequency of word \(w\).

But, the semantic similarity scores of this measure are calculated only between words of the same syntactic nature (part of speech) using lexical databases such as WordNet, because the lexical database used (WordNet) is unable to calculate the semantic similarity between terms of different syntactic nature.

4. Developed system of plagiarism detection
In this work, we tried to overcome the problems come up in classical plagiarism detection systems based on syntactic similarity measures, by comparing different plagiarism detection methods not only based on syntactic similarity measures but also semantic ones by developing a system that uses the previous measures which going to compare the two texts (i.e. comparing the plagiarized document with each document in the corpus), sentence by sentence and find if a sentence in one of the texts has the same meaning as a sentence in the other text.

4.1. Architecture of developed system
The developed PDS contains three modules, namely:

- **Preprocessing.** is transforming a text into a structured and formatted representation, which is more practical for the plagiarism detection process, and each type of similarity uses a precise set of these preprocessing methods which will help facilitate the processing of the developed system and will not affect the result. Preprocessing is a key step to obtain satisfactory results in the face of the difficulties of automatic natural language processing (NLP) [12].

![Image](image_url)

**Figure 2.** The architecture of developed PDS

the first is dedicated to the preprocessing of the suspect document and the corpus of source documents, the second is intended for the measurement of syntactic or semantic similarity based on the algorithms described above, the third is devoted to post-processing where there are the preparation and analysis of results to detect if the suspect document is plagiarized or not, if so, we identify the plagiarized sentences and which document. The figure 2 above describes the architecture of the developed system. Description of the components of the plagiarism detection system:
• Document segmentation: Splitting document text into sentences.
• Tokenization: chopping the sentence into significative units called tokens.
• Punctuation removal: Remove punctuation symbols or marks.
• Lowercase: Replace every uppercase letter with a lower case.
• Stop-word removal: Remove words that are so common that there is no need to index them or use them in a search.
• Part-of-Speech tagging (POS): Consists in assigning to each word of the text a grammatical label or tag providing certain information such as “noun”, “verb”, etc, in the context in which it appears.
• Stemming: Transform words into their stems by removing suffices, like “ing”, “ly”, “s”. For example, both « consigned » and « consignment » are normalized as « consign ».
• Lemmatization: Transform words into their canonical forms. For example, « walked », « walks », « walking », normalized as « walking ».

For syntactical similarity measures: Document segmentation into sentences, Tokenization, Punctuation removal, Transformation into lower case, Stop-word removal.

For Semantic Similarity Measures: In addition to Segmentation of the document into sentences and Tokenization, we are interested in the following methods: Part-of-speech tagging, Stemming, Lemmatization.

4.1.2. Similarity measures used

• Syntactic similarity: For syntactic similarity, the measures used in our system are Cosine similarity, Jaccard's coefficient, and Levenshtein distance.

Take the example of the following sentences:
Original sentence (OS): “A gem is a jewel or stone that is used in jewellery.”
Sentence 1 (S1): “A gem is a jewel that is used in trinkets.”
Sentence 2 (S2): “A jewel is a precious stone used to decorate valuable things that you wear, such as rings or necklaces.”

So, the result for each syntactic measure:

|               | Levenshtein distance | Jaccard's coefficient | cosine similarity |
|---------------|----------------------|-----------------------|-------------------|
| Sim(SO,S1)    | 0.7                  | 0.7                   | 0.66              |
| Sim(SO,S2)    | 0.3                  | 0.29                  | 0.32              |

Table 1. Result of syntactic similarity measures

• Semantic similarity: when the system has become unable to detect whether the document is plagiarized or not based on syntactic similarity measures, it should resort to semantic similarity measures. Among the different measures of semantic similarity, the interested one is the measure of Mihalcea [R. Mihalcea, 2006].

For the previous example, the measure of semantic similarity” Mihalcea”:

|               | Mihalcea’s measure |
|---------------|--------------------|
| Sim(SO,S1)    | 0.82               |
| Sim(SO,S2)    | 0.67               |

Table 2. result of semantic similarity measure
4.1.3. Post-processing. For preparing the results of the plagiarism detection task. We are then able to
decide whether the suspect document is plagiarized or not. The likelihood is usually determined by
predefining a threshold on the resulted similarity scores [13]. The sentences with higher similarity
scores are selected as the plagiarized sentences and the rest are discarded.

4.2. Plagiarism Detection System results:
We take 500 documents as dataset derived from the corpus of Webis Crowd Paraphrase 2011 (Webis-CPC-11).
From 30 documents in our test dataset, we manually build 18 plagiarized query documents.
And these query documents can be categorized into six cases of plagiarism, each case has three
documents:

| 1st Case          | Three Plagiarized documents with copy / paste |
|-------------------|---------------------------------------------|
| 2nd Case          | Three Plagiarized documents with copy / paste + external sentences |
| 3rd Case          | Three Plagiarized documents with paraphrase. |
| 4th Case          | Three Plagiarized documents with paraphrase + external sentences |
| 5th Case          | Three Plagiarized documents that combines copy / paste, paraphrase and external sentences |
| 6th Case          | Three Pure documents |

Table 3. the eighteen plagiarized query documents

4.3. Experiment and results:
The developed plagiarism detection system PDS is based on four similarity measures. Consequently, a
percentage threshold is generally predefined to determine the cases of plagiarism to be detected. This
threshold can be configured in the developed SDP. So, we set a value for this threshold $\Theta$ to find
plagiarism. If the percentage of plagiarism detected is greater than $\Theta_{syn} = 30\%$ for syntactic similarity
measures and $\Theta_{sem} = 50\%$ for semantic similarity measures, we can say that there is plagiarism, if not,
the plagiarism will not be found.
Tables from 2 to 4 show average of the recall, precision, F-measure, and accuracy rates for each case of
plagiarism, and for each text similarity measure when the 18 plagiarized query documents were
compared to the dataset.

|                  | 1st Case | 2nd Case | 3rd Case | 4th Case | 5th Case | 6th Case |
|------------------|----------|----------|----------|----------|----------|----------|
| Precision        | 0.99     | 0.99     | 0.92     | 0.99     | 0.87     | 0.99     |
| Recall           | 0.99     | 0.96     | 0.99     | 0.67     | 0.46     | 0.69     |
| F-measure        | 0.99     | 0.97     | 0.95     | 0.81     | 0.6      | 0.81     |
| Accuracy         | 0.99     | 0.96     | 0.94     | 0.69     | 0.52     | 0.75     |

Table 4. Performance measures of the syntactic similarity measure « Cosine »

|                  | 1st Case | 2nd Case | 3rd Case | 4th Case | 5th Case | 6th Case |
|------------------|----------|----------|----------|----------|----------|----------|
| Precision        | 0.99     | 0.99     | 0.99     | 0.99     | 0.99     | 0.99     |
| Recall           | 0.99     | 0.96     | 0.99     | 0.66     | 0.32     | 0.7      |
| F-measure | 0.99 | 0.97 | 0.99 | 0.79 | 0.49 | 0.82 |
|----------|------|------|------|------|------|------|
| Accuracy | 0.99 | 0.96 | 0.99 | 0.66 | 0.46 | 0.76 |

Table 5. Performance measures of the syntactic similarity measure « Jaccard »

|          | 1st Case | 2nd Case | 3rd Case | 4th Case | 5th Case | 6th Case |
|----------|----------|----------|----------|----------|----------|----------|
| Precision| 0.99     | 0.99     | 0.99     | 0.99     | 0.87     | 0.99     |
| Recall   | 0.99     | 0.93     | 0.99     | 0.63     | 0.46     | 0.68     |
| F-measure| 0.99     | 0.96     | 0.99     | 0.76     | 0.6      | 0.8      |
| Accuracy | 0.99     | 0.93     | 0.99     | 0.63     | 0.52     | 0.74     |

Table 6. Performance measures of the syntactic similarity measure « Levenshtein »

|          | 1st Case | 2nd Case | 3rd Case | 4th Case | 5th Case | 6th Case |
|----------|----------|----------|----------|----------|----------|----------|
| Precision| 0.99     | 0.99     | 0.87     | 0.99     | 0.92     | 0.93     |
| Recall   | 0.95     | 0.99     | 0.99     | 0.94     | 0.94     | 0.96     |
| F-measure| 0.97     | 0.99     | 0.93     | 0.97     | 0.93     | 0.94     |
| Accuracy | 0.95     | 0.99     | 0.9      | 0.94     | 0.9      | 0.91     |

Table 7. Performance measures of the semantic similarity measure « Mihalcea »

For very similar texts which are based on the plagiarism of copy/paste, the performance of all the measures was high and almost equal, and the Jaccard coefficient gave the best results.

On the other hand, the performance of syntactic similarity measures with paraphrased texts was relatively poor compared to the semantic similarity measure Mihalcea which was presented better results for this case of plagiarism.

But the fact of adding the external sentences to the plagiarized text is a bit confusing for the measure of Mihalcea. This suggests that adding external sentences to the plagiarized text does not affect the performance of measures of syntactic similarity as much as for the other semantic similarity measure.

We can therefore conclude that when it comes to comparing documents for a copy/paste plagiarism, it suffices to choose one of the syntactic similarity measures. On the other hand, for the plagiarism of paraphrase, the measure of semantic similarity Mihalcea seems relatively better for the measuring.

5. Conclusion

The privilege which brought to this developed system compared to the other traditional systems resides in: the semantic similarity approaches added to the syntactic similarity approaches which allow, in the case of the absence of syntactic similarity, we resort to analyze and filter the semantic similarities, and the comparison which consists in comparing the two texts, sentence by sentence, and no longer word by word like the classic systems because The meaning of a sentence does not only depend on the words that it contains, but also on the way they are combined.

Each finished work is the beginning of another, so this work accepts several perspectives, among them we have:

- Adapt this system to detect translation plagiarism.
- Evolve the test corpus.
- Good detection of plagiarism requires a massive collection of source texts as a corpus for a global overview. In this context, we need Big Data which makes it possible to process and analyze big data efficiently.
- Work with documents on the Web or in an internal document base.
References

[1] H. Mostafa, F. Benabbou, H. Ben Lahmar, Comparaison Des Techniques De Détecton Du Plagiat Académique, La quatrième journée sur les Technologies d'Information et de Modélisation TIM’16, Laboratoire Traitement de l’Information et Modélisation, Faculté des sciences Ben M’SIK, Juin 2016.

[2] S.M. Alzahrani, N. Salim, and A. Abraham, Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods, Senior Member, IEEE, April 2012.

[3] D.R. Bhalerao, S. S. Sonawane, A Survey of Plagiarism Detection Strategies and Methodologies in Text Document, International Journal of Science, Engineering and Technology Research (IJSETR) Volume 4, Issue 12, December 2015.

[4] V. Shukla, F. Khan, K. Mody, Prof S. Rathod, Plagiarism detection for document, International Journal on Recent and Innovation Trends in Computing and Communication, 2016.

[5] H. Chen, String Metrics and Word Similarity applied to Information Retrieval, Master Thesis, School of Computing, University of Eastern Finland, 2012.

[6] L. Huang, D. Milne, E. Frank, I. H. Witten, Learning a Concept-based Document Similarity Measure, Journal of the American Society for Information Science and Technology, August 2012.

[7] T. Mouflakir, Comparaison de TF-IDF et BM25 pour le repérage de l'information, Mémoire. Montréal (Québec, Canada), Université du Québec à Montréal, Juillet 2014.

[8] E. Negre, Comparaison de textes: quelques approches..., Laboratoire d’Analyses et Modélisation de Systèmes pour l’Aide à la Décision UMR 7243, Université Paris, avril 2013.

[9] I. Hamada, M. Fahci, Utilisation de "WordNet" pour indexation sémantique et recherche d'information, Mémoire de MASTER, Département de Mathématiques & informatique, Université Dr. MOULAY TAHAR DE SAÏDA, 2013.

[10] D. Fiser, B. Sagot, Construction d'un wordnet libre du français à partir de ressources multilingues, Actes de 15 e conférences sur le Traitement Automatique des Langues Naturelles, Avignon, France, 9 au 13 juin 2008.

[11] R. Mihalcea, C. Corley, and C. Strapparava, Corpus-based and Knowledge-based Measures of Text Semantic Similarity, In Proceedings of the 21st national conference on Artificial intelligence - Volume 1, AAAI Press, 2006.

[12] M. Chong, L. Specia, R. Mitkov, Using Natural Language Processing for Automatic Detection of Plagiarism, Research Group in Computational Linguistics, University of Wolverhampton, UK, 2010.

[13] R. Küppers, S. Conrad, A Set-Based Approach to Plagiarism Detection, Notebook for PAN at CLEF 2012, Published 2012.