# An n-gram based Method for Nearly Copy Detection in Plagiarism Systems

Behrouz Minaei
Iran University of Science and Technology
Tehran, Iran
b_minaei@iust.ac.ir

Mahdi Niknam
University of Qom
Qom, Iran
niknam.znt@gmail.com

## ABSTRACT

There has been plagiarism as a concept of "intellectual-property-theft" form the time that human and artistic research activities have been created. But easy access to the web, the massive database of information and communications system in recent years has led to the issue of plagiarism as a serious issue for publishers, researchers and the research institutions. In this paper, we introduce a method based on n-gram to identify similar textual parts between two documents. Evaluation of our method shows that our method has obtained both high accuracy and proper efficiency simultaneously.

## CCS Concepts

• **Information systems → Near-duplicate and plagiarism detection**
• **Information systems→ Evaluation of retrieval results**

## Keywords

Plagiarism detection; text alignment; n-gram

## 1. INTRODUCTION

Several definitions for plagiarism have been mentioned in scientific resources. In [9] plagiarism has been described as the "wrongful appropriation and stealing and publication of another author's language, thoughts, ideas, or expressions and the representation of them as one's own original work." Plagiarism, intellectual robbery or idea thievery has been defined as "attribution of others' research or literary creativity or a part of it or its derivative text to oneself, as if they have created it by themselves." Intellectual robbery is called literary plagiarism if committed in the sphere of literature, artistic plagiarism if perpetrated in the sphere of arts, and academic plagiarism if done in scientific fields.

Plagiarism, as "stealing of intellectual property", has a history coincided with the emergence of man's research and artistic activities. But easy access to the web, the massive databases of information and, in general, communication means in recent years has caused the issue of plagiarism to be a serious problem for publishers, researchers and research institutions. On the other hand, the country's rapid scientific growth in recent years has caused an increase in the possibility of intentional and unintentional academic plagiarism.

## 2. DEFINITION OF PLAGIARISM DETECTION

In order to develop the plagiarism algorithm, the article [7] defines the issue of plagiarism as follows. A case of plagiarism can be shown as $s = \langle s_{plg}, d_{plg}, s_{src}, d_{src} \rangle$ that is composed of the following parts:

- $d_{src}$ the original document from which the plagiarized work has been derived.
- $s_{src}$ part of the original document that has been stolen.
- $d_{plg}$ the document where plagiarism has been detected.
- $s_{plg}$ part of the document $d_{plg}$ in which plagiarism has occurred.

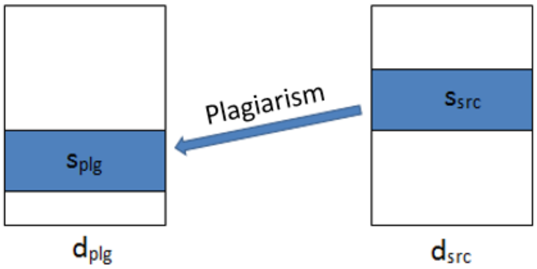

**Figure 1. Elements constituting a plagiarism**

A plagiarism detector's task is to report a case of plagiarism as $r = \langle r_{plg}, d_{plg}, r_{src}, d'_{src} \rangle$. The relation $r$ shows the part $r_{plg}$ from the document $d_{plg}$ which has been plagiarized from the part $r_{src}$ of the document $d_{src}$ and it is aimed at giving a maximum estimation of $s$. if the following circumstanced are realized between $s$ and $r$, it can be said that $r$ has managed to recognize $s$:

$$\boldsymbol{s_{plg} \cap r_{plg} \neq \emptyset\ , s_{src} \cap r_{src} \neq \emptyset\ , d_{src} = \acute{d}_{src}}$$

## 3. THE GENERAL APPROACH IN DETECTING PLAGIARISM

Article [8] has offered a general categorization of activities required to identify academic plagiarism using a huge amount of external resources; so that most of works in the field of academic plagiarism detection, although too different in implementation details, have used the general approach offered in the aforementioned article. This approach has generally the following three stages:

- Heuristic Retrieval: since the plagiarism suspect document $d_{plg}$ may have used any document available in dataset D and the total amount of data is usually very huge, comparing the suspicious document with each document in dataset isn't possible. Therefore, in the first step, for each $d_{src}$, a collection $D_x$ is so selected as a subset of D that the original document is most probable to occur in this dataset.

- Detail analysis: in this phase the document $d_{plg}$ is compared with any of the documents in $D_x$ to identify $(s_{plg}, s_x)$ so that they have a high similarity with each other and also $s_{plg}$ ϵ $d_{plg}$ and $s_x$ ϵ $d_x$.

- Post-processing: to reduce the output inaccuracy, in this stage, the results of the previous phase are filtered so that the overall efficiency of algorithm is optimized. Filters can include a range of activities such as deleting the cases with low length, combining two or more cases, or removing the cases that have true references.

These three steps are shown in Figure 2.

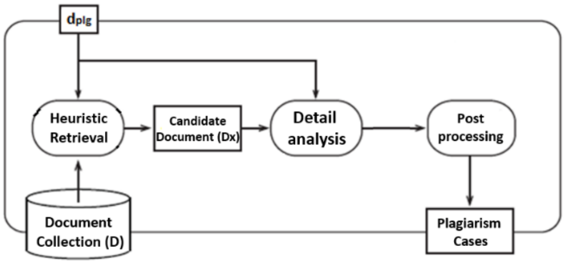

**Figure 2. Three Phases of Plagiarism Detection System**

In this paper, we present a method to compare two documents and to identify similar parts of them. This method can be used in "detail analysis" phase.

## 4. TEXT FRAGMENTATION UNIT IN PLAGIARISM DETECTION ALGORITHMS

In all text processing tasks, the text must be split into chunks so that processing algorithms can be applied on them. It is also necessary in plagiarism detection. Various processing units have been used in plagiarism detection researches, such as character, word, sentence or paragraph. The document is chunked based the chosen unit. The smaller the unites are (eg, word and character), the more chunks the document is divided into. This leads to an increase in computation time; however, it provides the possibility to identify the snippets of the copied text. If the length of the pieces is high (as in sentence and paragraph), the number of pieces decreases and the speed of performance rises; but the accuracy of the algorithm in identifying small pieces of the text disappears.

COPS (Copy Protection System) is a part of Digital Library Project at Stanford University. This system uses sentence units for comparison between documents and it is unable to identify overlaps within sentences [3]. To cover COPS inefficiencies, the SCAM system was developed using word units to compare documents.

The CHECK system uses the structural information of text to compare documents. Its comparison unit is paragraph. Its complete dependence to document structure was its constraint [8].

There has been introduced a system in [5] as PPChecker which uses sentence unit to compare documents. In their study, [6] uses n-grams with the value n=3 to convert each document to a set of tri-grams and then compares two documents by calculating the number of common tri-grams of them.

There has been introduced a system in [2] which uses a combination of the two methods [6] and [5]. This system's search strategy is to divide the suspicious document to separate sentences and converts original documents to tri-grams. In order to search, firstly sentences of the suspicious document are converted to tri-grams and then are searched for in the tri-grams of the original texts.

## 5. MODELING THE PROBLEM IN TWO-DIMENSIONAL SPACE

For better understanding of the plagiarism detection problem, we can show the comparison between the two documents $d_{plg}$ and $d_{src}$ using the Scatter Plot. In this visualization, index of the document $d_{src}$ is displayed on the horizontal axis and the index of the document $d_{plg}$ is displayed on the vertical axis. If a word has occurred in both documents, its index in the documents $d_{plg}$ and $d_{src}$ is shown as a spot on the graph. If a piece of text is copied from another document without any modification, the image plagiarized area creates a diagonal line. Figure 3 shows the similar parts of the two non-plagiarized documents. As shown in the graph, some several words are normally similar in the two documents. This similarity has shown as a dark page in the Scatter Plot. Number of similar spots and level of darkness of the graph differ according to similarity of subjects of the two documents. Figure 4 displays two documents in which a part of one document has occurred exactly in the other one. The diagonal line in the graph represents a list of words used in the other text in the same order.

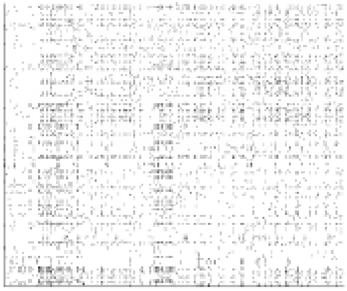

**Figure 3. Visualization of similarities in two documents without plagiarism**

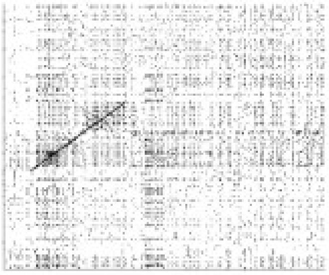

**Figure 4. Visualization of exact copy of a text to another document**

By modifying the plagiarized text, one causes changes in the diagonal line. In the following, we will examine different states of text modification and their effects in the corresponding picture.

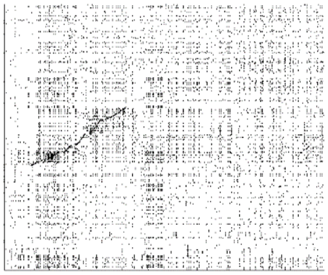

**Figure 5. Visualization of copying text with low replacement rate**

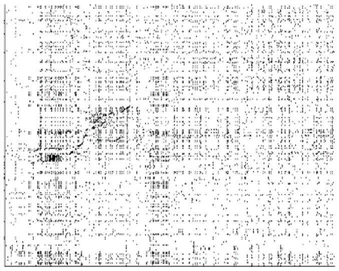

**Figure 6. Visualization of copying text with high replacement rate**

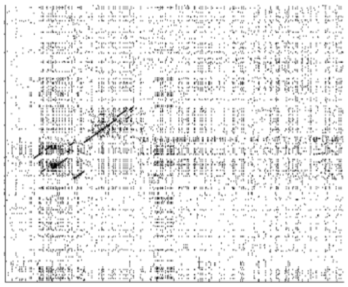

**Figure 7. Visualization of copying text with sentence displacement**

Taking into consideration the presented pictures, it can be said that "the process of detecting plagiarism is the identification of areas with higher density in the picture."

# 6. SUGGESTED METHOD

In this section, we present the proposed method for identifying similar parts in two documents.

To identify cases of plagiarism in the two documents $d_{src}$ and $d_{plg}$, we used the visualized model presented in the previous section. The algorithm steps can be outlined as follows:

- The list of n-grams is generated out of each document and each n-gram, in accordance with its display order in the document, is assigned an integer counter starting at 1.

- All n-grams occurred in both documents are shown as (i, j); that is i represents the n-gram position in the document $d_{plg}$ and j represents the n-gram position in the document $d_{src}$. For example, Figure 8 shows how these two sentences are modeled for detecting similar parts:

" تحقیق حاضر نشان می‌دهد میانگین نمرات کسب شده توسط کشتی‌گیران هر دو
تحقیق حاضر نشان داده است که " and "گروه مورد آزمایش بسیار بالا است.
میانگین نمره‌های کسب شده توسط کشتی‌گیران هر دو گروه مورد بررسی بسیار
بالاست".

- The process starts from the beginning of the list. If both (i, j) and (i+1, j+1) occur in the list then they are put in one plagiarism case and this task continues until the end of the list. Using this method, all n-grams occurring consecutively in both documents $d_{plg}$ and $d_{src}$ are identified as one plagiarism case. The purpose of this step is to identify all similar cases in both documents. Plagiarism cases that occur without any changes will be completely identified at this step. Two parts of documents that complete resemble are shown in Figure 9.

-

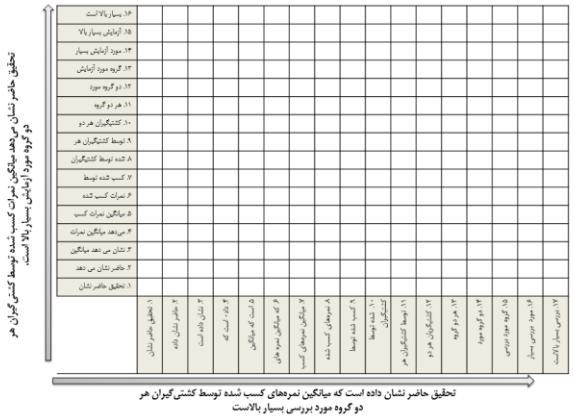

**Figure 8. Modeling two pieces of text for detecting similar parts**

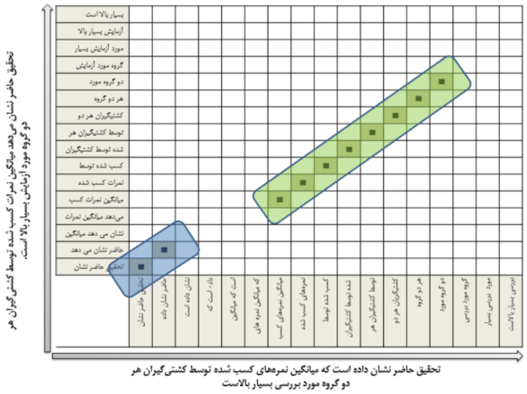

**Figure 9. Detecting parts that are identical in two documents**

- To overcome the changes made to the text and reduce the number of cases detected, they will be merged. For this purpose, if the distance between two cases is lower than a threshold in regard to their length, the cases are merged.

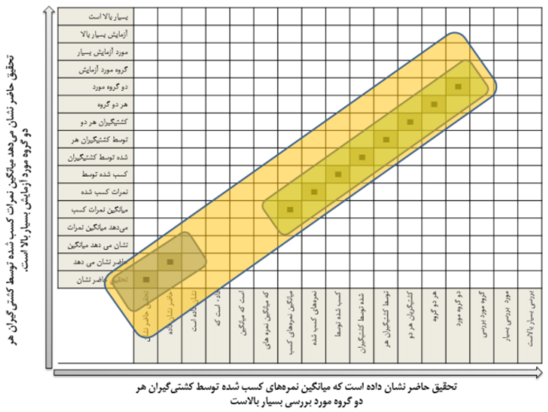

**Figure 10. Merging the adjacent identical cases**

- Finally, to reduce misidentified instances, the cases that are lower than a threshold in terms of length will be removed from the list.

## 7. EVALUATION

To evaluate the proposed method, the Persian dataset introduced in PAN 2016 conference and the evaluation method in [7] and [4] were used. Taking into consideration the explanations provided in the previous sections, the proposed method may have two adjustable parameters. The first parameter is the intimacy threshold limit for integrating proximate similar parts; and the second parameter is the length threshold limit for removing the cases that are shorter than a specified length. In order to find the best threshold values, we used the range of 20 to 50 characters with 5 intervals for proximity threshold limit and the range of 50 to 100 characters with 10 intervals for length threshold limit.

Then the algorithm was run per possible values for the parameters and in each case the output was calculated based on the evaluation metrics. After the algorithm was executed, the best results, considering the evaluations metrics, were gained with the values 45 characters for proximity threshold limit and 90 for length threshold limit. The outcome of the algorithm based on the evaluation metrics is shown in Table 1.

**Table 1. Algorithm result based on evaluation metrics**

| Plagdet | Granularity | Recall | Precision |
|---|---|---|---|
| 0.84 | 1.04 | 0.79 | 0.93 |

## 8. CONCLUSION

The process of identifying plagiarism has two steps: Heuristic Retrieval and detail analysis. In this paper, we presented a method for the second step of the plagiarism detection process in order to identify similar parts of two documents. In this method, by generating n-grams from each document and by comparing n-grams of two documents, the similar parts are identified. In the next step, the similar parts that are closer to each other than a threshold limit are merged with each other. The proposed method, without being engaged in complex text analysis technique, can detect copy and near-copy parts in two documents though with reasonable accuracy.

## 9. REFERENCES

[1] Asghari, H., Mohtaj, S., Fatemi, O., Faili, H., Rosso, P., and Potthast, M., 2016. Algorithms and Corpora for Persian Plagiarism Detection: Overview of PAN at FIRE 2016. In *Working notes of FIRE 2016 - Forum for Information Retrieval Evaluation,* Kolkata, India, December 7-10, 2016, CEUR Workshop Proceedings, CEUR-WS.org.

[2] Barrón-Cedeño, A. and Rosso, P. 2009. On automatic plagiarism detection based on n-grams comparison. *Advances in Information Retrieval*. Springer. 696–700.

[3] Brin, S., Davis, J. and Garcia-Molina, H. 1995. Copy detection mechanisms for digital documents. *ACM SIGMOD Record* (1995), 398–409.

[4] Improving the Reproducibility of PAN's Shared Tasks: - Springer: *http://link.springer.com/chapter/10.1007%2F978-3-319-11382-1_22*. Accessed: 2016-11-04.

[5] Kang, N., Gelbukh, A. and Han, S. 2006. PPChecker: Plagiarism pattern checker in document copy detection. *Text, Speech and Dialogue* (2006), 661–667.

[6] Lyon, C., Barrett, R. and Malcolm, J. 2004. A theoretical basis to the automated detection of copying between texts, and its practical implementation in the Ferret plagiarism and collusion detector. *Plagiarism: Prevention, Practice and Policies*. (2004).

[7] Potthast, M., Stein, B., Barrón-Cedeño, A. and Rosso, P. 2010. An evaluation framework for plagiarism detection. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (2010), 997–1005.

[8] Stein, B., zu Eissen, S.M. and Potthast, M. 2007. Strategies for retrieving plagiarized documents. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (2007), 825–826.

[9] 2016. Plagiarism. *Wikipedia, the free encyclopedia*.