Hybrid System for Plagiarism Detection

Javier R. Bru
University of Alicante
javier.r.bru@gmail.com

Patricio Martínez-Barco
University of Alicante
patricio@dlsi.ua.es

Rafael Muñoz
University of Alicante
rafael@dlsi.ua.es

Abstract

The Internet boom in recent years has increased the interest in the field of plagiarism detection. A lot of documents are published on the Net everyday and anyone can access and plagiarize them. Of course, checking all cases of plagiarism manually is an unfeasible task. Therefore, it is necessary to create new systems that are able to automatically detect cases of plagiarism produced. In this paper, we introduce a new hybrid system for plagiarism detection which combines the advantages of the two main plagiarism detection techniques. This system consists of two analysis phases: the first phase uses an intrinsic detection technique which dismisses much of the text, and the second phase employs an external detection technique to identify the plagiarized text sections. With this combination we achieve a detection system which obtains accurate results and is also faster thanks to the prefiltering of the text.

1 Introduction

Plagiarism detection is a topic that has always received some interest. Authors have worried about other people stealing their intellectual property, in other words, having their work plagiarized. With the recent increase in the importance of the Internet plagiarism has become a real problem. Anyone anywhere in the world can access any document, plagiarize and publish it as their own. Each author cannot spend all his or her time watching that nobody copies his or her work, so it is very important to create systems that can automatically detect cases of plagiarism.

The research in this field is mainly divided into two branches: external plagiarism detection and intrinsic plagiarism detection. Each one has its own advantages and disadvantages. In this paper we introduce a new plagiarism detection system that combines these two detection techniques, joining their main advantages and avoiding their disadvantages. This system has a first phase that uses an intrinsic detection technique to identify text sections that are most likely to be plagiarism. This phase helps us to filter the text and discard much of it, thus the next phase must analyze less text. The second phase is based on an external detection technique, which employs text comparisons to identify plagiarized sections. This technique, although slow, is very precise for plagiarism detection. Moreover, the problem of slowness is mainly solved thanks to the filtering of text done in the previous phase.

The benefits of this combination of detection techniques are the merge of the speed of intrinsic detection and the precision of external detection. We also avoid their disadvantages. In intrinsic detection we improve precision with the second analysis phase. About external detection, which is a very slow technique, we increase speed thanks to the filtering of text in the first phase.

The remainder of this paper is organized as follows. In Section 2 we detail how the first phase of intrinsic plagiarism detection is implemented. The future implementation of the second phase of external detection is described in Section 3. In Section 4 we show the preliminary results obtained with the system developed so far. In Section 5 conclusions are presented. Finally, future work, especially external detection phase, is included in Section 6.

2 Intrinsic Detection

Intrinsic plagiarism detection technique does not require a reference collection with original documents. This technique only analyzes the suspicious document trying to find changes in the author's writing style. For that purpose, we use stylometry, which is the application of the study of linguistic style. Stylometry is based on the idea that each author has an individual writing style depending on unconscious habits. There are

Proceedings of Recent Advances in Natural Language Processing, pages 527–532,
Hissar, Bulgaria, 12-14 September 2011.

many stylometric features, for instance, count ing the number of punctuation marks, sentence length, or number of stopwords.

Our system employs the Averaged Word Frequency Class (Meyer zu Eissen et al, 2007) as writing style measure. A document's averaged word frequency class quantifies the style complexity and the size of the author's vocabulary. This measure has the advantage that is independent of the length and structure of text. This is suitable for our system because we take sentences as text units and these are of variable length and structure. Another salient property is it works with word frequencies, so this measure can be used with documents written in different languages.

In order to make the intrinsic analysis of the suspicious document we must first calculate the document's averaged word frequency class. To this end, we divide the document into sentences and calculate the averaged word frequency class each of them. The measure of a sentence is the average of the word frequency class of every word of the sentence. Then, there only remains to calculate the average of measures of all sentences.

The next step is to identify the plagiarized sections of the text. We calculate the averaged word frequency class of all the sentences of the document following the process described above. These measures are compared with the document's averaged word frequency class. Those sentences which have a significantly different value from the document's averaged value are considered as plagiarism.

The difference between the value of the sentences and the value of the whole document is determined by a percentage set by the user. We have defined this difference as the Percentage Deviation (PD), which determines the results obtained by the intrinsic analysis. If PD is low, much plagiarism is detected because the difference between the values is low. Many false positives are also detected and little text is discarded. However, if PD is high we detect less plagiarism but the amount of discarded text is higher.

The benefits from this analysis phase are mainly two. First, we achieve to identify the text sections most likely to be plagiarized. Those sections are confirmed in the next analysis phase. Second and more important, we discard much of the text. Only plagiarized sentences are stored, so the next phase must process less text. This is important because external detection is a very slow technique.

3 External Detection

The second analysis phase of our system uses an external plagiarism detection technique. This technique is based in a reference corpus of source documents. The suspicious document is compared with all the source documents to find identical or similar text sections. If the comparison is successful we can confirm a plagiarism in the suspicious document and the source document which has been copied from. In our system only the probably plagiarized sentences identified in the previous phase are compared with the reference corpus. This speeds up the process considerably.

Currently, we are working on this phase and only an initial part is completed about the verbatim plagiarism. This type of plagiarism is known as a copy word for word without any change on the text. To identify verbatim plagiarism we compare the plagiarized sentences obtained in the previous intrinsic phase with every sentence of every document of the reference corpus. The comparison is made word for word. If the number of equal words is greater than 90 % the suspicious sentence is considered plagiarism and the reference sentence is its source. This method has a high accuracy as long as the plagiarized sentence has not been modified from its source.

But the verbatim plagiarism is the less common case. As expected, the plagiarist does not want his or her copy to be detected, for which he uses obfuscation methods in the text. These obfuscation methods try to hide the copies changing the plagiarized text. There are different obfuscation techniques such as: (i) removing, inserting, or replacing the words of the sentence, (ii) changing the words by their synonyms, antonyms, hyponyms, or hypernyms, and (iii) changing the structure of the sentence. In short, any technique that prevents a direct comparison between the plagiarized sentence and the source sentence is an obfuscation technique.

Our next step is to continue working to detect this type of more complex plagiarisms. Among the papers that can inspire us, we emphasize two which are appropriate for us. Firstly, the application PPChecker (Nam Oh Kang et al, 2006) is interesting because it also works on sentence level and is based on plagiarism pattern checking. This application is able to find subtle changes in the words and structure of the sentences. Secondly, an algorithm which works with sentences too (White and Joy, 2004). It measures the similarity between sentences based on the number of words

in common and the length of the sentences. If a certain threshold is exceeded, the sentences are considered equal, in other words, one sentence is the plagiarism of the other. This algorithm is able to detect sophisticated obfuscation like paraphrasing, reordering, or merging sentences.

Another possibility considered is not utilizing a reference corpus. The comparisons between the suspicious document and source documents can be made through the Internet. This is the method used by the application SNITCH (Niezgoda and Way, 2006). Thus, the text sections of the suspicious document are searched on the Internet. If one section is found, the section is plagiarism because someone had to copy it. This is an interesting technique because we do not have to build the reference corpus, which is a complex and long task in many cases.

Whatever the used method, the objective of this analysis phase is to confirm the plagiarism detected in the previous phase thanks to the precision of the external detection techniques. In addition, the false positives detected in the intrinsic phase are easily discarded in this phase.

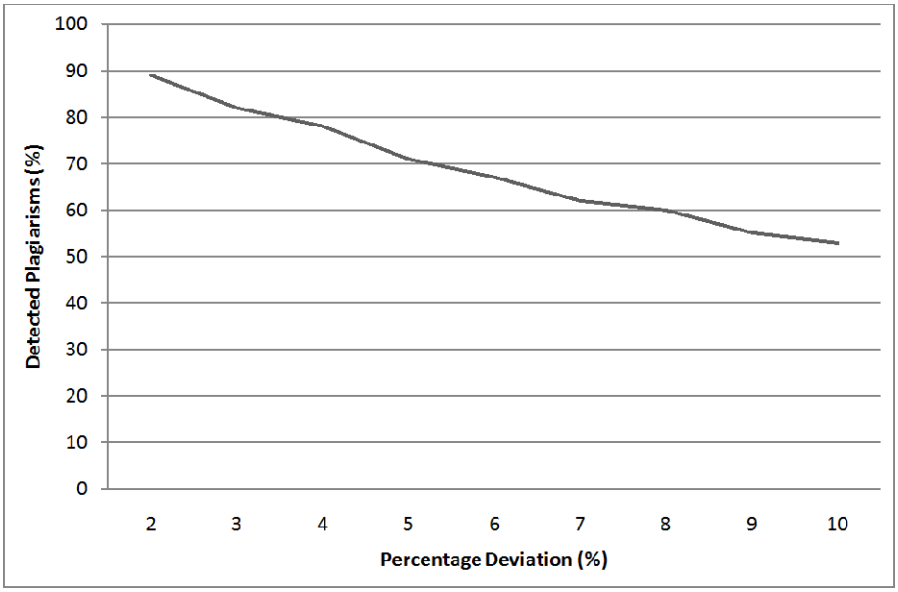

Figure 1: Detected plagiarisms depending on PD parameter value.

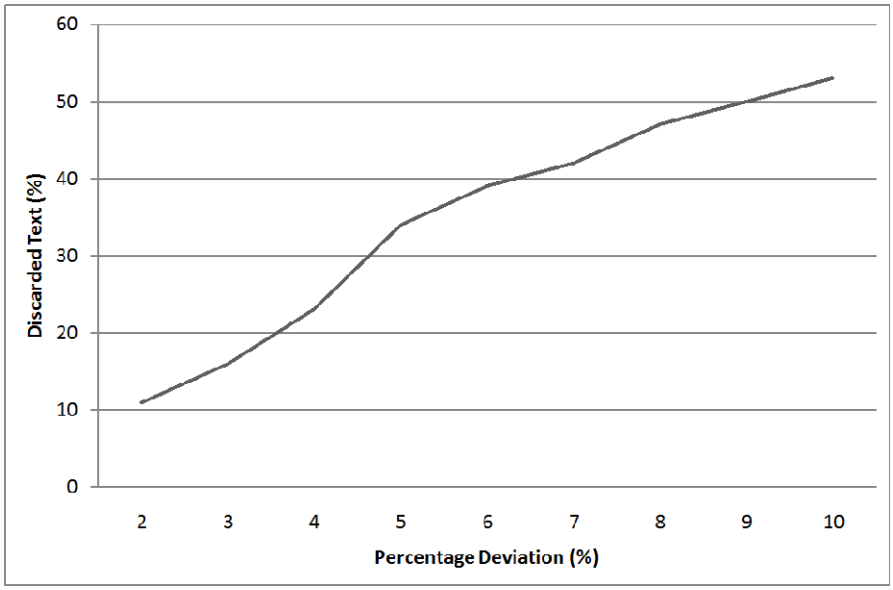

Figure 2: Amount of discarded text depending on PD parameter value.

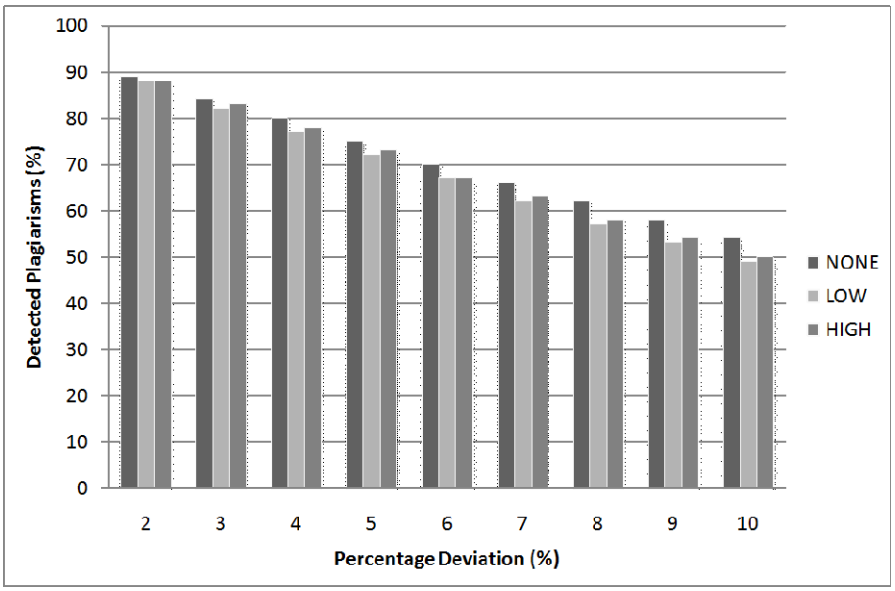

Figure 3: Detected plagiarisms according to PD parameter and corpus complexity.

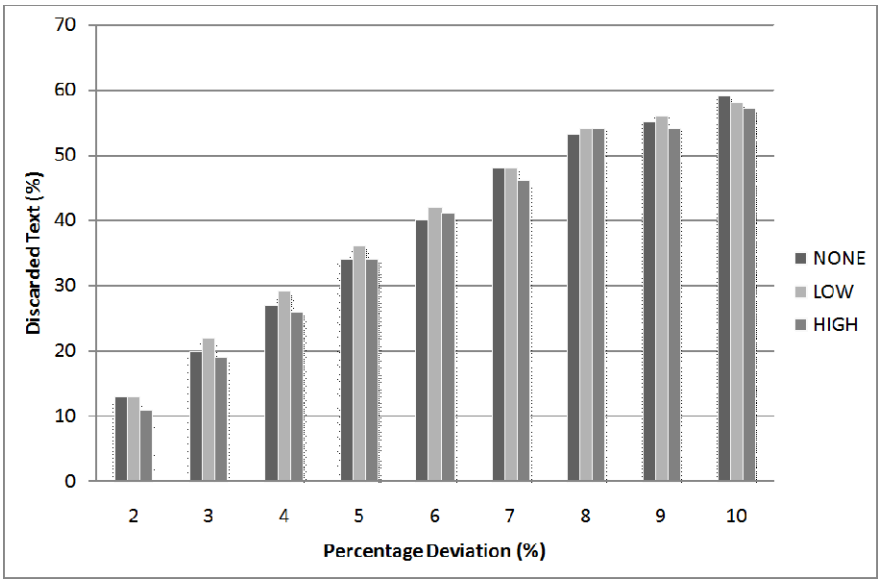

Figure 4: Discarded text according to PD parameter and corpus complexity.

4 Experiments

This section presents the preliminary tests performed with the developed system so far. The tests have been carried out with the PAN-PC-10 corpus (Potthast et al, 2010), which was created for the 4th International Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse. This is a detailed corpus which contains 64,558 artificial and 4,000 simulated plagiarism cases spread over nearly 6,000 suspicious documents. It also contains over 11,000 source documents to make comparisons. All the documents have an extension between 10 and 1,000 pages. The included plagiarisms are very varied and can be verbatim or obfuscated copies. Several obfuscation strategies have been used: (i) manual obfuscation realized by a human, (ii) random text operations, (iii) semantic word variations, and (iv) word shuffling. Therefore, this is a good

corpus to do different tests to achieve exhaustive results.

For the intrinsic detection phase the system tries to find the plagiarism cases in the suspicious documents collection of the corpus without utilizing a reference collection. The Percentage Deviation (PD) parameter seen in Section 2 has been established to a 5% value.

Intrinsic Phase Results		
Sentences plagiarized	4,182,604	
Sentences detected	2,999,834	(71%)
Total text (characters)	3,495,686,760	
Discarded text (characters)	1,208,801,781	(34%)

Table 1: Results of the Intrinsic Detection Phase

As shown in Table 1, the intrinsic phase of the system is able to detect the 71% of the plagiarisms included in the suspicious collection and discards the 34% of all text (more than a third of the text). Therefore, the next phase of external detection only has to compare the 66% of text of all suspicious collection. The time taken to process more than 3GB of the suspicious collection has been 47 minutes, which shows the speed of this intrinsic technique.

As said in Section 2, varying the value of PD parameter we can change the results of the intrinsic detection phase. If PD is decreased, the difference between plagiarized sentence's value and document's averaged value is lower. This makes the plagiarism detection task more restrictive. Detection percentages regarding the PD values are shown in Figure 1. On the other hand, the PD parameter affects the amount of discarded text. Unlike before, more text is discarded when PD value is high. Values of discarded text are represented in Figure 2.

Therefore, the PD value must be low when the primary objective is to detect plagiarism as much as possible. If our priority is to discard much of the text we must assign a high PD value. It would be interesting for large corpuses when the second analysis phase should analyze the least amount of text. It is necessary to find an intermediate value for PD parameter that provides a balance between the number of detected plagiarisms and the amount of discarded text. Through various tests we have determined that the optimum value for PD is 5%.

Moreover, we have also tested how the PD parameter affects the results depending on the corpus complexity. To achieve this test we have divided the PAN corpus according to the level of plagiarism complexity included in each document. For this, we have used the own division made by its authors. The documents of the corpus are classified in three types depending on the obfuscation level: high, low or none. The tests done with these three groups are represented in Figures 3 and 4. It can be seen that percentage of detected plagiarisms is similar for each subcorpus. Only the group without obfuscation obtains slightly higher results. The discarded text is also constant for each group. With this it is shown that PD parameter influences the results but the parameter itself is not influenced by the corpus complexity. Thus, this is positive because we do not have to worry about the configuration of PD parameter in function of complexity of the corpus we work on.

Regarding the external detection phase, we have only tested the completed part so far. Tests have been carried out with verbatim plagiarism and results show that virtually 100% of plagiarism is detected. This is logical because this type of plagiarism is easily identified by direct comparisons of text. Now we are working with more complex types of plagiarism and all different obfuscation strategies.

5 Conclusions

The system which is being implemented shows promising results in the plagiarism detections field. The intrinsic detection phase has given good results in the detection of plagiarisms as well getting to discard a considerable part of the text. This benefits the next external phase and ultimately decreases the system runtime. The intrinsic phase has also been flexible and adjustable depending on our needs: more plagiarism detection or more discarded text. The number of detected plagiarisms and the amount of discarded text can change through Percentage Deviation parameter setting. The tests have proved that the system is able to detect nearly 90% of the plagiarism cases and discard more than half of the text. Because one thing is against the other finding a balance between both terms is recommended.

Moreover, we have tested that the results for a certain PD value are constant regardless of the corpus complexity. We only have to set PD parameter to obtain good results but we do not have

to previously check the obfuscation level of the corpus. This simplifies the intrinsic detection task and makes the system independent of the used corpus.

The external detection phase will make the system more precise in the task of plagiarism detection thanks to the high precision of the external detection techniques. The work being done at this phase will allow the system to detect all types of obfuscation strategies and therefore more plagiarism cases will be identified.

In conclusion, our system is able to offer good results in the plagiarism detection. Moreover, the detection is done at high speed, which is very interesting due to the large number of documents to analyze nowadays.

6 Future Work

In the short term our future work is concentrated in completing the second phase of the system. The external detection phase must be able to confirm nearly 100% of the detected plagiarisms in the intrinsic phase and remove as many false positives as possible. In order to do this, the system must identify a large number of obfuscation techniques like changing the word order or the sentence structure. The more techniques are identified, the more plagiarism is detected and more types of documents can be analyzed.

Once the system has been completed, we can improve the different phases of the system. The intrinsic phase can be perfected to detect more plagiarism without harming the amount of discarded text. It would also be interesting to reduce the number of false positives obtained in this phase.

The external detection phase can also be improved to detect more types of plagiarism. For instance, we can add another algorithm to detect translated plagiarisms, in other words, plagiarisms where the source text has been written in one language and the plagiarized text has been translated into another language.

References

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig. 2007. *Plagiarism detection without reference collections.* Advances in Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 359-366.

Nam Oh Kang, Alexander Gelbukh and Sang Yong Han. 2006. *PPChecker: Plagiarism Pattern Checker in Document Copy Detection.* Text, Speech and Dialogue Proceedings, Lecture Notes in Artificial Intelligence, volume 4188, pp. 661-667.

Daniel R. White and Mike S. Joy. 2004. *Sentence-based natural language plagiarism detection.* Journal on Educational Resources in Computing, volume 4, issue 4.

Sebastian Niezgoda and Thomas P. Way. 2006. *SNITCH: a software tool for detecting cut and paste plagiarism.* ACM SIGCSE Bulletin, volume 38, issue 1, pp. 51-55.

Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, Paolo Rosso. 2010. *An Evaluation Framework for Plagiarism Detection.* Proceedings of the 23rd International Conference on Computational Linguistics, COLING 2010, Beijing, China.