Intrinsic Plagiarism Detection Using Character *n*-gram Profiles

Efstathios Stamatatos
University of the Aegean
83200 - Karlovassi, Samos, Greece
stamatatos@aegean.gr

Abstract: The task of intrinsic plagiarism detection deals with cases where no reference corpus is available and it is exclusively based on stylistic changes or inconsistencies within a given document. In this paper a new method is presented that attempts to quantify the style variation within a document using character *n*-gram profiles and a style change function based on an appropriate dissimilarity measure originally proposed for author identification. In addition, we propose a set of heuristic rules that attempt to detect plagiarism–free documents and plagiarized passages, as well as to reduce the effect of irrelevant style changes within a document. The proposed approach is evaluated on the recently-available corpus of the 1st Int. Competition on Plagiarism Detection with promising results.

Keywords: Plagiarism detection, Character *n*-grams, Stylistic inconsistencies

1 Introduction

Textual plagiarism (the unacknowledged use of the work of another author either as an exact copy or a slightly modified version) is a major problem in modern world affecting education and research mainly. The rapid development of WWW made billions of web pages easily accessible to anyone providing plenty of potential sources for plagiarism. As a result, automated plagiarism analysis and detection receives increasing attention in both academia and software industry (Maurer et al, 2006).

There are two basic tasks in plagiarism analysis. In *external plagiarism detection* a reference corpus is given and the task is to identify pairs of identical or very similar passages from a suspicious document and some texts of the reference corpus. Most of the research studies in plagiarism analysis deal with this task (Hoad and Zobel, 2003; Stein, 2005). On the other hand, *intrinsic plagiarism detection* is more ambitious since no reference corpus is given (Meyer zu Eissen et al., 2007; Stein and Meyer zu Eissen, 2007). This task is applied in cases where it is not possible to have a representative reference corpus. In addition, the comparison of a suspicious document with all the texts of a very large corpus may be impractical in terms of computational time cost. It can also serve as a preprocessing step to an external plagiarism detection tool in order to reduce the time cost.

To handle the intrinsic plagiarism detection task one has to detect plagiarized passages of a suspicious document exclusively based on irregularities or inconsistencies within the document. Such inconsistencies or anomalies are mainly of stylistic nature.

The attempts to quantify writing style, a line of research known as ‘stylometry’, have a long history (Holmes, 1998). A great variety of measures that represent some kind of stylistic information have been proposed especially in the framework of authorship attribution research. In a recent survey, Stamatatos (2009) distinguishes the following types of stylometric features: lexical features (word frequencies, word *n*-grams, vocabulary richness, etc.), character features (character types, character *n*-grams), syntactic features (part-of-speech frequencies, types of phrases, etc.), semantic features (synonyms, semantic dependencies, etc.), and application-specific features (structural, content-specific, language-specific).

Although the lexical features are still the most popular, a number of independent recent studies have demonstrated the effectiveness of character *n*-grams for quantifying writing style (Keselj et al., 2003; Stamatatos, 2006;

Stamatatos, 2007; Kanaris and Stamatatos, 2007; Koppel et al., 2009). This type of features can be easily measured in any text and it is language and domain independent since it does not require any text pre-processing. These measures are also robust to noise. Note that in plagiarism analysis the efforts of an author to slightly modify a plagiarized passage may be considered as noise insertion. Graham et al. (2005) were the first to use character *n*-grams to detect stylistic inconsistencies in texts. However, their results were poor. One reason for this is that they only used character bigrams. Another reason is that the distance measure they used (cosine distance) was unreliable for very short texts. Note also that Graham et al. (2005) were based on predefined text segments (paragraphs) and their task was to identify whether two consecutive paragraphs differ in style or not.

In this paper, we propose a method for intrinsic plagiarism detection based on character *n*-gram profiles (the set of character *n*-gram normalized frequencies of a text) and an appropriate dissimilarity measure originally proposed for author identification. Our method automatically segments documents according to stylistic inconsistencies and decide whether or not a document is plagiarism-free. A set of heuristic rules is introduced that attempt to detect plagiarism on either the document level or the text passage level as well as to reduce the effect of irrelevant stylistic changes within a document.

The rest of the paper is organized as follows. Section 2 describes the method of quantifying stylistic changes within a document. Then, Section 3 includes the plagiarism detection heuristics while Section 4 describes the evaluation procedure. Finally, Section 5 discusses the main points of this study and proposes future work directions.

2 *The style change function*

The main idea of the proposed approach is to define a sliding window over the text length and compare the text in the window with the whole document. Thus, we get a function that quantifies the style changes within the document. Then, we can use the anomalies of that function to detect the plagiarized sections. In particular, the peaks of that function (corresponding to text sections of great dissimilarity with the whole document) indicate likely plagiarized sections. Therefore, what we need is a means to compare two texts knowing that one of the two (the text in the window) is shorter or much shorter than the other (the whole document).

Following the practice of recent successful methods in author identification (Keselj et al., 2003; Stamatatos, 2006; Stamatatos, 2007; Koppel et al., 2009; Stamatatos, 2009), each text is considered as a bag-of-character *n*-grams. That is, given a predefined *n* that denotes the length of strings, we build a vector of normalized frequencies (over text length) of all the character *n*-grams appearing at least once in the text. This vector is called the *profile* of the text. Note that the size of the profile depends on the text length (longer texts have bigger profiles). An important question is the value of *n*. A high *n* corresponds to long strings and better capture intra-word and inter-word information. On the other hand, a high *n* considerably increases the dimensionality of the profile. To keep dimensionality relatively low and based on preliminary experiments as well as on previous work on author identification (Stamatatos, 2007; Koppel et al., 2009) we used character 3-grams in this study. The complete set of parameter settings for the proposed method is given in Table 1. These settings were estimated using a small part (~200 documents) of the evaluation corpus (see section 4).

Description	Symbol	Value
Character *n*-gram length	n	3
Sliding window length	l	1,000
Sliding window step	s	200
Threshold of plagiarism-free criterion	t_1	0.02
Real window length threshold	t_2	1,500
Sensitivity of plagiarism detection	a	2

Table 1: Parameter settings used in this study.

Let $P(A)$ and $P(B)$ be the profiles of two texts A and B, respectively. Stamatatos (2007) studied the performance of various distance measures that quantify the similarity between two character *n*-gram profiles in the framework of author identification experiments. The following distance (or dissimilarity) measure has been found to be both accurate and robust when the two texts significantly differ in length.

$$d_1(A,B)=\sum_{g\in P(A)}\left(\frac{2(f_A(g)-f_B(g))}{f_A(g)+f_B(g)}\right)^2$$

where $f_A(g)$ and $f_B(g)$ are the frequency of occurrence (normalized over text length) of the *n*-gram *g* in text *A* and text *B*, respectively, Note that d_1 is not a symmetric function (typically, this means it cannot be called distance function). That is, only the *n*-grams of the first text are taken into account in the sum. This function is designed to handle cases where text *A* is shorter than text *B*. Stamatatos (2007) showed that d_1 is quite stable even when text *A* is much shorter than text *B*. This is exactly the case in the proposed method for intrinsic plagiarism detection where we want to compare a short text passage with the whole document that may be quite long. In this paper, we modified this measure as follows:

$$nd_1(A,B)=\frac{\sum_{g\in P(A)}\left(\frac{2(f_A(g)-f_B(g))}{f_A(g)+f_B(g)}\right)^2}{4|P(A)|}$$

where $|P(A)|$ is the size of the profile of text *A*. The denominator ensures that the values of dissimilarity function lie between 0 (highest similarity) and 1. We call this measure *normalized* d_1 (or nd_1).

Let *w* be a sliding window of length *l* (in characters) and step *s* (in characters). That is, each time the window is moved to the right by *s* characters and the profile of the next *l* characters is extracted. If $l>s$ the windows are overlapping. Then, we can define the *style change* function (*sc*) of a document *D* as follows:

$$sc(i,D)=nd_1(w_i, D),\ i=1\ldots|w|$$

where $|w|$ is the total amount of windows (it depends on text-length). Given a text of *x* characters $|w|$ is computed as follows:

$$|w|=\left\lfloor 1+\frac{x-l}{s}\right\rfloor$$

Examples of style change functions can be seen in figures 1, 2, 3, and 4.

3 *Detecting plagiarism*

3.1 Plagiarism on the document level

The first important question that must be answered is whether or not a given document contains any plagiarized passages. This is crucial to keep the precision of our method high. If we are unable to find documents that are plagiarism-free, it is quite likely for the plagiarism detection method to identify a number of text passages as the result of potential plagiarism for any given document. Thus, the credibility of the method would be very low.

There are two options to decide whether or not a document contains plagiarized sections:

By pre-processing: A criterion must be defined to indicate a plagiarism-free document. If this is the case, there is no further detection of plagiarized sections.

By post-processing: The algorithm detects any likely plagiarized sections and then a decision is taken based on these results.

Typically, the detected sections are compared to other sections of the document to decide whether there are significant differences between them (Stein and Meyer zu Eissen, 2007).

In this study we followed the former approach. The criterion we used is based on the variance of the style change function. If the document is written by one author, we expect the style change function to remain relatively stable. On the other hand, if there are plagiarized sections, the style change function will be characterized by peaks that significantly deviate from the average value. The existence of such peaks is indicated by the standard deviation. Let *S* denote the standard deviation of the style change function. If *S* is lower than a predefined threshold, then the document is considered plagiarism-free.

Plagiarism-free criterion: $S<t_1$

The value of the threshold t_1 was determined empirically at 0.02. Recall that the dissimilarity function we use is normalized. So, the definition of such a common threshold for all the documents is possible. However, the nd_1 measure is not independent of text length. Very short documents tend to have low style change function values. Moreover, very long texts are likely to contain stylistic changes made intentionally by the author. In both these cases this criterion will not be very accurate.

Figures 2 and 3 show the style change function of documents 00017 and 00034 of IPAT-DC (see section 4) that fall under the plagiarism-free criterion. The former is a successful case where no plagiarism exists. On the other hand, in the case of document 00034,

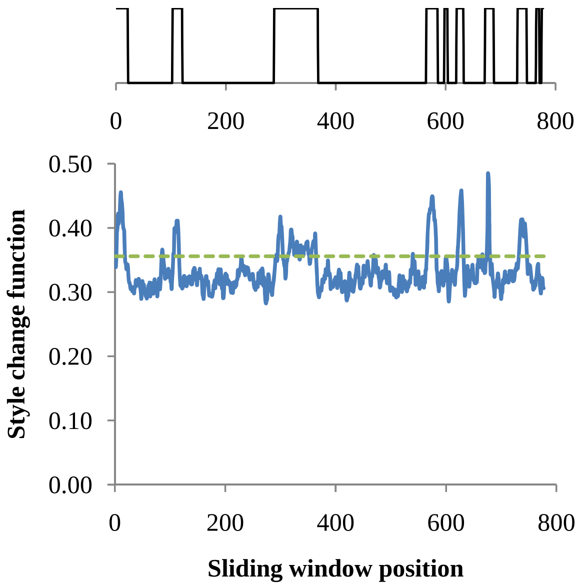

Figure 1: The style change function of document 00005 of IPAT-DC (solid line). The dashed line indicates the threshold of the plagiarized passage criterion. The binary function above indicates real plagiarized passages (high values).

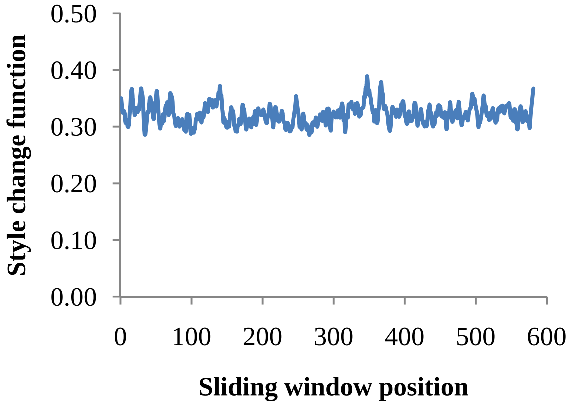

Figure 2: The style change function of document 00017 of IPAT-DC (a plagiarism-free document).

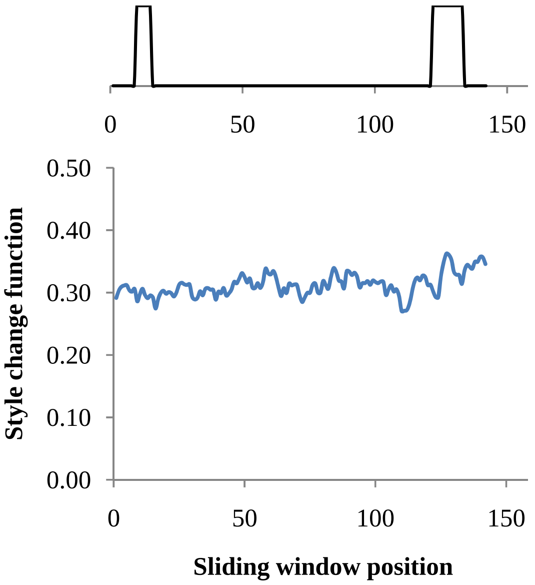

Figure 3: The style change function of document 00034 of IPAT-DC (false negative). The binary function above indicates real plagiarized passages (high values).

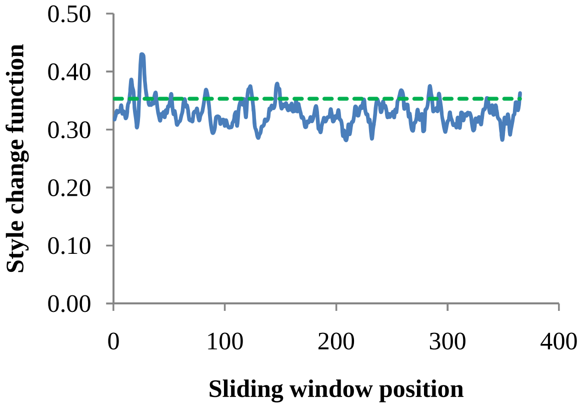

Figure 4: The style change function of the plagiarism-free document 00022 of IPAT-DC (a false positive). The dashed line indicates the threshold of the plagiarized passage criterion.

despite the presence of two plagiarized passages, the style change function fails to produce significant peaks that would increase its standard deviation. Note also that 00017 is longer than 00034 (more sliding windows in the *x*-axis) and the average style change function of 00017 is higher than that of 00034. Additionally, Figure 4 shows the style change of document 00022 of IPAT-DC. Although this document is plagiarism-free, the standard deviation of its style function is greater than the used threshold (false positive).

3.2 Identifying plagiarized passages

Given the style change function of a document, the task of plagiarism detection can be viewed as detecting peaks of that function corresponding to text sections that significantly differ from the rest of the document. One big problem in plagiarism detection is that it is not possible to estimate the percentage of plagiarized text beforehand. In intrinsic plagiarism detection the problem is much harder since if the plagiarized sections are too long the stylistic anomalies would correspond to the style of the alleged author rather than the plagiarized sections. In this study we suppose that at least half of the text is not plagiarized so that the average of style change function would indicate the style of that author. However, the calculation of the average *sc* value would inevitably involve the plagiarized passages as well.

Let *M* and *S* denote the mean and standard deviation of *sc*, respectively. To reduce this problem we first remove from *sc* all the text windows with value greater than *M*+*S*. These text sections are highly likely to correspond to plagiarized sections. Let *sc*(*i*′,*D*) denote the style change function after the removal of these sections. Let *M*′ and *S*′ be the mean and standard deviation of *sc*(*i*′,*D*). Then, we define the following criterion to detect plagiarism:

Plagiarized passage criterion:
sc(*i*′,*D*) >*M*′+*a***S*′

The parameter *a* determines the sensitivity of the plagiarism detection method. The higher the value of *a*, the less (and more likely plagiarized) sections are detected. The value of *a* was determined empirically at 2.0 to attain a good combination of precision and recall. Figures 1 and 4 show the result of applying the proposed criterion in two documents.

3.3 Detecting irrelevant style changes

An important factor that affects style using the character *n*-gram representation is the formatting of documents. A document written in uppercase with many space characters, punctuation symbols will have a quite different character *n*-gram profile than the same document in lowercase after the removal of any extra space and punctuation characters. The proposed method for the quantification of style changes is very general and is sensitive to such stylistic changes that are irrelevant to plagiarism. In fact, a very common technique to disguise plagiarism is to change the formatting of text. So, any plagiarism detection tool should attempt to reduce the formatting factor.

To deal with this problem, we performed a number of processes. First, each document is transformed to lowercase. Although the uppercase information is important for representing adequately the style of an author, it can be easily used to fool a plagiarism detection tool. Then, we removed from the profile of a text every character *n*-gram that contains no letter characters (a-z, or any lowercase character of foreign languages) at all. This way, any character *n*-gram that contains only digit, space, or punctuation characters, that is irrelevant to the content of text, is excluded and the formatting factor is reduced. Finally, the sliding window parameters operate on letter characters. That is, a window length of *l* characters means that the window should contain *l* letter characters. Note that all the other characters (digits, spaces, punctuation, etc.) are not removed. Therefore, if *l*=1,000, a window may contain 1,200 characters (this is the real window length) in total from which 1,000 are letter characters. Moreover, a step of *s* characters means that the window is moved to the right by *s* letter characters. This procedure ensures that all the text windows will have the same number of letter (or content) characters

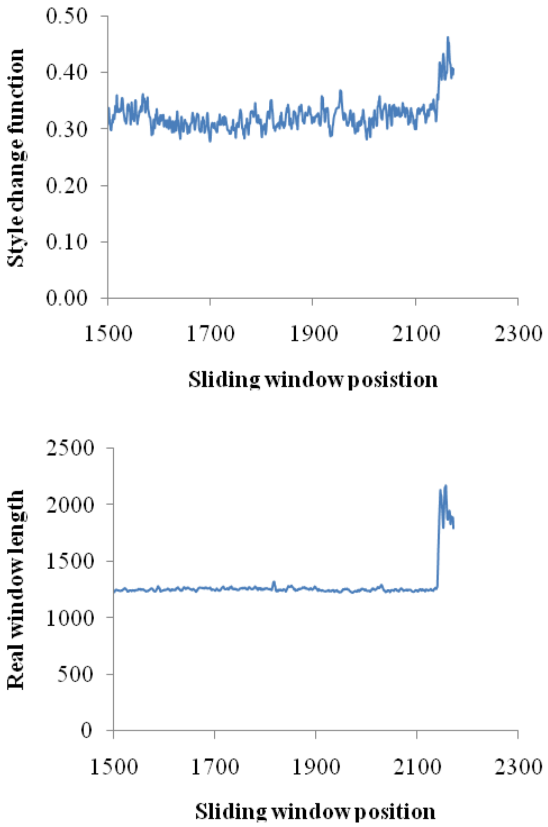

Figure 5: The style change function and the real window length of the last part of document 00046 of IPAT-DC.

and the formatting of the text will not significantly affect the style change function.

Since there is no prior knowledge on the genre of documents, a given document may be composed of several sections each one belonging to a different genre (or sub-genre) and therefore having different stylistic characteristics. For example, a table of contents has different style than the main document. The character *n*-gram representation is able to capture both the style of author and the style of genre but it is hard to distinguish these factors. To handle this problem, we make use of the real window length as defined above. In more detail, let l' be the real window length (the total number of characters included in a window that contains l letter characters) of a text section. The real window length is affected by some genres. For example, the l' of a table of contents is higher than the l' of the main document. This is demonstrated in figure 5 that shows the style change function and the real window length of the last part of document 00046 of IPAT-DC (for l=1,000). This document ends with an index. Note that the real window length of this special section is much higher than the rest of the document. The stylistic difference between the index and the rest of the document is captured by the style change function. However, this difference has nothing to do with plagiarism. To take such cases into account, an additional criterion was used to detect plagiarized passages:

Special section criterion: $l'<t_2$

This criterion is combined with the plagiarized passage criterion. Based on empirical evaluation, the value of the threshold t_2 was estimated at 1,500 (or 1.5l). Note that this criterion excludes text sections with overly real window length. However, one can take advantage of this criterion and disguise plagiarism by inserting many formatting characters to a text section so that l' is considerably increased. Moreover, a plagiarized section within a special section (e.g. table of contents) that resembles the style of that section will not be detected.

4 Evaluation

In the framework of the 1st International competition on plagiarism detection a large corpus has been released for the Intrinsic Plagiarism Analysis Task (Potthast et al., 2009). This corpus is segmented into a development part (IPAT-DC) and a competition part (IPAT-CC) each one comprising 3,091 documents. An artificial plagiarism tool has been used to automatically insert plagiarized passages within the documents. The following evaluation results are mainly based on IPAT-DC since this corpus also provides ground truth data. IPAT-DC comprises a wide variety of texts covering many genres and topics. The text length varies from (roughly) 3,000 characters to 2.5 million characters. Interestingly, the plagiarized passages begin in randomly selected positions covering arbitrary combinations of words, sentences, and paragraphs.

4.1 Evaluation on the document level

First, we evaluate the plagiarism-free criterion that operates on the document level. Table 2 shows the confusion matrix of IPAT-DC after the application of this criterion. It is important that over 70% of the plagiarism-free documents were correctly classified. This is crucial to keep the overall precision on reasonable level. On the other hand, false positives (see Figure 4) harm the precision while false negatives (see Figure 3) harm the recall.

	Actual	
Guess	Plagiarism-free	Plagiarized
Plagiarism-free	1102	545 (22%)
Plagiarized	443	1001 (78%)

Table 2: Confusion matrix (on the document level) after the application of the plagiarism-free criterion. The percentage of plagiarized passages included in the documents are inside parentheses.

As can be seen, roughly 1/3 of the plagiarized documents are considered plagiarism-free. However, taking into account the number of plagiarized passages within each document (indicated inside parentheses in the table), we see that 22% of the plagiarized passages is missed. So, the upper bound for the recall on the passage level will be 78%. A closer look to the false negatives shows that text-length is a crucial factor. Figure 6 depicts the distribution of false negatives over text-length of documents. As can be seen, the majority of false negatives are relatively short documents (<30K chars). Moreover, the shorter a document, the more likely to be false negative.

Corpus	IPAT-DC	IPAT-CC
Recall	0.4552	0.4607
Precision	0.2183	0.2321
F-score	0.2876	0.3086
Granularity	1.22	1.25
Overall score	0.2358	0.2462

Table 3: Performance of the plagiarism passage criterion on two corpora (development and competition corpus).

4.2 Evaluation on the passage level

To evaluate the plagiarism detection method, we should first define appropriate measures. In particular, we used the performance measures defined in the framework of the 1st int. competition on plagiarism detection: recall, precision, granularity, and overall score. Let r denote a plagiarized passage and $|R|$ be the set of all plagiarized passages in the corpus. Moreover, let p be a detected passage by the proposed method, $|P|$ be the set of all detected passages, and $|R_p|$ be the subset of R that overlap with at least one member of $|P|$. Finally, let $|r|$ and $|\hat{r}|$ be the length of a plagiarized passage and the sum of its detected characters by the plagiarism detection method, respectively. Similarly, $|p|$ and $|\hat{p}|$ are the length of a detected passage and the sum of their chars that belong to any plagiarized passage. Then, recall, precision, and granularity can be defined as follows:

$$recall = \frac{1}{|R|}\sum_{i=1}^{|R|}\frac{|\hat{r}_i|}{|r_i|}$$

$$precision = \frac{1}{|P|}\sum_{i=1}^{|P|}\frac{|\hat{p}_i|}{|p_i|}$$

$$granularity = \log_2(1+\frac{1}{|R_P|}\sum_{i=1}^{|R_P|}|r_i \cap P|$$

$$overall = \frac{F}{granularity}$$

where $|r_i \cap P|$ denotes the number of different detections that overlap with the plagiarized passage r_i, and F is the harmonic mean of recall and precision. Essentially, the granularity measure indicates the fragmentation of the detected passages. A granularity value of 1 means that at most one detected section overlaps with a plagiarized passage.

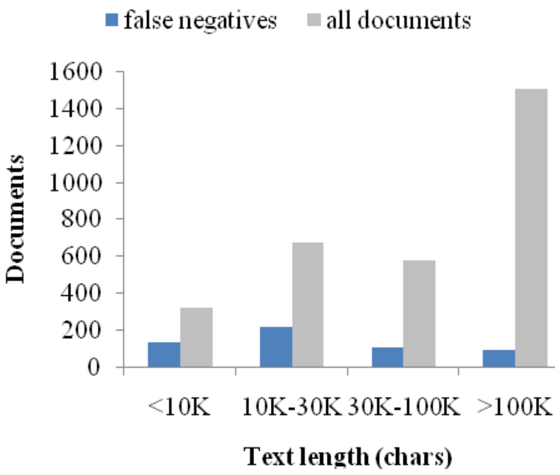

Figure 6: Distribution of false negatives over text length.

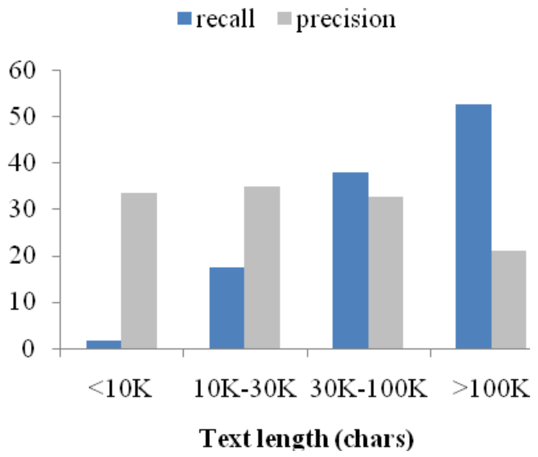

Figure 7: Recall and precision for varying text length.

The results of the evaluation of the plagiarized passage criterion are included in table 3 on the development and competition corpus of the intrinsic plagiarism analysis task (taken by the official results of the competition). The parameter values shown in table 1 have been used to produce these results. As can be seen, the performance of the proposed method remains stable for both corpora. Actually, the performance on IPAT-CC is better than on IPAT-DC that was used for estimating the values of parameters. This indicates that the proposed settings are quite general and robust.

Figure 7 provides a closer look in the recall-precision results on IPAT-DC with respect to text-length of documents. It is obvious that recall is dramatically affected by decreasing text length. The distribution of false negatives showed in figure 6 offers a reasonable explanation for this. Precision is more stable. However, it tends to decrease while text length increases.

5 Discussion

In this paper a new method for intrinsic plagiarism detection has been presented. The proposed approach is based on character *n*-gram profiles, a style change function using an appropriate dissimilarity measure as well as a set of heuristic rules to detect plagiarized passages. The evaluation results demonstrate that it is able to detect roughly half of the plagiarized sections. On the other hand, the precision remains low. An important factor for improving precision is the development of more sophisticated and accurate plagiarism-free criteria on the document level. The precision can also be improved by increasing the sensitivity parameter *a*. However, this will harm recall.

The proposed method is easy to follow and requires no language-dependent resources. Moreover, it requires no text segmentation or preprocessing. The proposed parameter settings proved to be effective when the approach was evaluated on the IPAT-CC. Note that the parameter values of table 1 were not optimized for IPAT-DC. However, the application of machine learning algorithms could improve the estimation of these parameters. Especially, the definition of the window length is crucial since it determines the shortest plagiarized passage that can be detected. On the other hand, a very short window would not adequately capture the stylistic properties of the text.

Another future work direction is to examine different schemes for comparing a text window with the whole document. The approach followed in this paper is fast since it requires the calculation of only one profile for the whole document. Alternative approaches include the comparison of the text window with the window complement (the document without the window) and the comparison of a text window with all the other text windows.

Finally, character *n*-grams of higher order could be used. Preliminary experiments using character 4-grams and 5-grams did not show significant improvement on the performance of the method. However, this remains to be carefully examined.

References

Graham, N. Hirst, G. and Marthi, B. (2005). Segmenting Documents by Stylistic Character. *Natural Language Engineering*, 11(4): 397-415.

Hoad, T.C. and J. Zobel. 2003. Methods for Identifying Versioned and Plagiarised Documents. *Journal of the American Society for Information Science and Technology*, 54(3):203–215.

Holmes, D.I. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, 13(3): 111-117.

Kanaris, I. and E. Stamatatos. 2007. Webpage Genre Identification Using Variable-length Character *n*-grams, In *Proc. of the 19th IEEE Int. Conf. on Tools with Artificial Intelligence*, v.2, pp. 3-10.

Keselj, V., F. Peng, N. Cercone, and C. Thomas. 2003.. N-gram-based Author Profiles for Authorship Attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pp. 255-264.

Koppel, M., J. Schler, and S. Argamon. 2009. Computational Methods in Authorship Attribution, *Journal of the American Society for information Science and Technology*, 60(1): 9-26.

Maurer, H., F. Kappe, and B. Zaka. 2006. Plagiarism - A Survey. *Journal of Universal Computer Science*, 12(8): 1050-1084.

Meyer zu Eissen, S., B. Stein, and M. Kulig. 2007. Plagiarism Detection without Reference Collections. *Advances in Data Analysis*, pp. 359-366, Springer.

Potthast, M., A. Eiselt, B. Stein, A. Barron, and P. Rosso. 2009. Plagiarism Corpus PAN-PC-09. Webis at Bauhaus-Universitaet Weimar and NLEL at Universidad Polytecnica de Valencia. (http://www.webis.de/research/corpora)

Stamatatos, E. 2009. A Survey of Modern Authorship Attribution Methods, *Journal of* *the American Society for information Science and Technology*, 60(3): 538-556.

Stamatatos, E. 2007. Author Identification Using Imbalanced and Limited Training Texts. In *Proceedings of the 4th International Workshop on Text-based Information Retrieval*, pp. 237-241.

Stamatatos, E. 2006. Ensemble-based Author Identification Using Character N-grams, In *Proc. of the 3rd Int. Workshop on Text-based Information Retrieval*, pp. 41-46.

Stein, B., and S. Meyer zu Eissen. 2007. Intrinsic Plagiarism Analysis with Meta Learning. In *Proceedings of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection*, pp.45-50.

Stein, B. 2005. Fuzzy-Fingerprints for Text-Based Information Retrieval. In Proceedings of the 5th International Conference on Knowledge Management, J.UCS: 572–579.