Expert Systems with Applications 40 (2013) 3756–3763

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style

Gabriel Oberreuter, Juan D. Velásquez *

Web Intelligence Consortium Chile Research Centre, Department of Industrial Engineering, Universidad de Chile, Av. República 701, P.O. Box 8370439, Chile

ARTICLE INFO

Keywords:
Text mining
Text classification
Plagiarism
Copy detection
Intrinsic plagiarism detection

ABSTRACT

Plagiarism detection is of special interest to educational institutions, and with the proliferation of digital documents on the Web the use of computational systems for such a task has become important. While traditional methods for automatic detection of plagiarism compute the similarity measures on a document-to-document basis, this is not always possible since the potential source documents are not always available. We do text mining, exploring the use of words as a linguistic feature for analyzing a document by modeling the writing style present in it. The main goal is to discover deviations in the style, looking for segments of the document that could have been written by another person. This can be considered as a classification problem using self-based information where paragraphs with significant deviations in style are treated as outliers. This so-called intrinsic plagiarism detection approach does not need comparison against possible sources at all, and our model relies only on the use of words, so it is not language specific. We demonstrate that this feature shows promise in this area, achieving reasonable results compared to benchmark models.

1. Introduction

Plagiarism cases are an everyday topic, for example, in academics, journalism, scientific research and even in politics. The recent case where the Hungarian President had to quit over a plagiarism scandal in April of 2012 is only one of the examples where copying and plagiarism can become a real problem. With the explosive growth of content found throughout the Web, people can find nearly everything they need for their written work, but detection of such cases can become a tedious task. For these reasons society needs to tackle this problem with computer-assisted approaches, and consequently, multiple studies in the field are being conducted. Various methods can be implemented, ranging from document-comparison algorithms and systems to scan the Web, to approaches that utilize language-specific features, for example for the authorship-attribution task.

Generally speaking, the task of plagiarism detection from an algorithmic point of view can be divided into two main strategies (Meyer zu Eißen & Stein, 2006; Meyer zu Eißen, Stein, & Kulig, 2007; Potthast, Stein, Eiselt, Barrón-Cedeño, & Rosso, 2009; Potthast, Barrón-Cedeño, Eiselt, Stein, & Rosso, 2010); those that utilize only information within the suspected document, denominated intrinsic plagiarism detection, and those that compare the suspected document against a set of possible sources (ideally, but unrealistically, the entire Web). Intrinsic plagiarism detection aims at discovering plagiarism by analyzing only the suspicious document, trying to identify those segments that are potentially written by another person. For this, current algorithms usually use writing style modeling techniques, searching for meaningful variations. On the other hand, external plagiarism detection refers to the task of comparing the suspected document against possible sources. This is the classical approach, in which the systems usually begin with some kind of tokenization or indexing of the documents, and then look at coincidences and document features before generating the detected copied passages as output. From exact document copying to paraphrasing, several levels of plagiarism techniques can be used in different contexts, according to Meyer zu Eißen et al. (2007).

The contribution of this work relates to the modeling of writing style. We explore a model for writing style quantification, aimed at finding significant deviations in a document's writing style; these differing segments could have been plagiarized, and are probably useful as a starting point to search for possible source candidates. For example, Bravo-Marquez, L'Huillier, Ríos, and Velásquez (2011) introduce a way of searching over the Internet for sources by taking information from the document; one could use these deviated segments as inputs for the query construction.

This paper is structured as follows. First, in Section 2, the definition of the problem is presented. Second, in Section 3 we review

* Corresponding author. Tel.: +562 978 4834.
E-mail addresses: goberreu@ing.uchile.cl (G. Oberreuter), jvelasqu@dii.uchile.cl (J.D. Velásquez).
URL: http://wi.dii.uchile.cl/ (J.D. Velásquez).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.12.082

plagiarism detection research. Afterwards, in Section 4, the proposed intrinsic plagiarism detection method is described. In Section 5 the experiments and results are presented. Finally, in Section 6 the conclusions are discussed.

2. Problem definition

One can consider various types of plagiarism in written documents. The most commonly criticized and easiest to detect is the infamous "copy and paste", or literal copying, found especially in students' work in academic institutions. The different kinds of plagiarism are:

- **Exact Copy**. The passage is copied word for word, without citation.
- **Paraphrasing**. The text is modified, but the idea and part of the words remain the same.
- **Plagiarism of ideas**. The idea is copied using different words and language resources.

Verbatim or exact copying can be easily found if the source documents are present, as shown in Potthast, Eiselt, Barrón-Cedeño, Stein, and Rosso (2011) and Oberreuter, L'Huillier, Ríos, and Velásquez (2011). As the kind of plagiarism becomes more sophisticated, the difficulty of detecting cases of copying increases, usually because modifications to the copied text become more complex. This is because automatic methods and algorithms for plagiarism detection utilize language resources that are easy to handle in terms of computer science, for example words, synonyms, structures like word *n*-grams, sentences and so on. But when the idea is copied using a completely different vocabulary, or even when paraphrasing, the task becomes harder. It is very difficult to normalize the text in order to capture the words that represent the idea behind it using algorithms, and as such, there is still a long road ahead in terms of research. Nevertheless, improvements to algorithms in recent years are advancing in that direction.

The problem of plagiarism can be found in many areas and thus affects us in multiple ways. These areas include:

- Academia.
- Scientific research.
- Journalism.
- Patents.
- Literature.

3. Related work

The plagiarism issue can be treated from two perspectives, prevention and detection. As Schleimer, Wilkerson, and Aiken (2003) states, both can be combined to effectively reduce it. While copy detection methods can only help after the plagiarism has been committed, prevention methods can and should educate and encourage people not to do it, further decreasing its level. Notwithstanding this fact, prevention methods need the participation of society as a whole, thus its solution is not trivial. Copy plagiarism detection methods, on the other hand, are easier to implement, and tackle the problem at different levels, from simple manual comparison to complex automatic algorithms (Potthast et al., 2010, 2009).

A short overview of plagiarism detection approaches is presented.

3.1. Intrinsic plagiarism detection

Intrinsic detection of plagiarism refers to the analysis of a document, in which one tries to infer if a portion of the text has potentially been plagiarized. The concept was recently introduced by Meyer zu Eißen and Stein (2006), and is close related to authorship attribution, where one analyzes the writing style of the text identifying segments written differently.

This approach to plagiarism detection is especially useful when no reference collection is available or not all the possible copy sources are present, thus document-to-document comparison algorithms cannot be used. In their first work, Meyer zu Eissen et al. studied the use of style aspects and experimented with an approach in which they observe that this kind of plagiarism detection is possible.

In the following years, more studies have been published in which the intrinsic plagiarism detection problem is further investigated (Meyer zu Eißen et al., 2007; Oberreuter et al., 2011; Rao, Gupta, Singhal, & Majumder, 2011; Seaward & Matwin, 2009; Stamatatos, 2009; Stein, Lipka, & Prettenhofer, 2011;).

Stamatatos (2009) presented a method for intrinsic plagiarism detection. As described by its author, this approach 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. Style profiles are first constructed using a sliding window. For the construction of those profiles the author proposed the use of character *n*-grams. These *n*-grams are used for getting information about the writer's style. The method then analyzes changes in the profiles to determine if a change is significant enough to indicate another author style.

Seaward and Matwin (2009) introduced Kolmogorov complexity measures as a way of extracting structural information from texts for intrinsic plagiarism detection. They experimented with complexity features based on the Lempel–Ziv compression algorithm for detecting style shifts within a single document, thus revealing possible plagiarized passages.

Stein et al. (2011) presented in their work a description of features used for writing style modeling, aimed at intrinsic plagiarism detection. Their analysis also included results obtained when using such features to discover plagiarism cases, in which the top three performers were the Flesch Reading Ease Score, the average number of syllables per word and the frequency of the term "of".

Oberreuter et al. (2011) reported acceptable results by quantifying and analyzing variations in the use of words.

3.2. Authorship attribution

Authorship attribution is the task of characterizing the writing style of a document, aiming at recognizing the style of a particular author. As Juola (2006) in his extensive review explains, authorship attribution has been important in historic cases where confusion regarding documents and their authors must be clarified. It is important in the field of plagiarism detection, as authorship attribution is closely related to intrinsic plagiarism detection, where the writing style present in the document is analyzed.

In automatic authorship attribution it is important to define and select linguistic features that can represent the writing style of authors. In this regard, nearly all studies conducted in automatic authorship attribution face and treat this fundamental problem. These features include syntactical and lexical measures with analysis at the character, word, sentences and whole-text level. Grieve (2007) studies and compares thirty-nine different measures; Baayen, van Halteren, and Tweedie (1996) studies the use of words and the use of syntax-based measures and in van Halteren (2004) a "linguistic profile" is constructed based on multiple linguistic measures.

Other studies that utilize syntactic and lexical measures include Kern, Seifert, Zechner, and Granitzer (2011); Koppel, Schler, and Argamon (2009, 2011); Kourtis and Stamatatos (2011); Stamatatos

(2009); Stamatatos, Fakotakis, and Kokkinakis (1999, 2001) and Tanguy, Urieli, Calderone, Hathout, and Sajous (2011).

3.3. External plagiarism detection

External plagiarism detection refers to the task of comparing a suspicious document against the possible sources. Therefore, if a plagiarism case is found, one has the proof, the passage in the suspicious document and its homologue in a particular source document. In the case of a comparison between documents two factors become important. First, the comparison between documents should be made quickly, potentially considering a large collection of possible sources. Second, the comparison should be effective, detecting slightly modified copied passages as well as non-obfuscated ones.

One can separate the multiple approaches proposed so far into two categories. In the first category, the models compare documents and their output to determine whether or not the pair of documents contains plagiarized passages, but provide no detailed information on which paragraphs are copied. These models can be considered as one-class classification models, determining whether a pair of documents is flagged or not.

In this category we can find the approaches from the machine learning community (,) namely Bao, Shen, Liu, Liu, and Zhang (2004); Chow and Rahman (2009) and Jun-Peng, Jun-Yi, Xiao-Dong, Hai-Yan, and Xiao-Di (2003). Bao et al. in Jun-Peng et al. (2003) and then in Bao et al. (2004) proposed using a semantic sequence kernel (SSK), and then inserting it into a traditional support vector machines (SVMs) formulation based on the structural risk minimization (SRM) (Boser, Guyon, & Vapnik, 1992; Vapnik, 1999) principle from statistical learning theory, where the general objective is finding out the optimal classification hyper plane for the binary classification problem (plagiarized, not plagiarized.) Likewise, other approaches solve the same classification problem by using self-organizing feature maps (SOFM) (Kohonen, 2001), with promising results in classification performance. In terms of using latent semantic analysis (LSA) for the plagiarism detection task, Ceska in Ceska (2008) proposed a method using singular value decomposition (SVD) (Berry, Dumais, & O'Brien, 1995) for finding associated phrases from a given pair of documents. This approach, as noted by the author, uses the technique to infer the latent semantic associations and subsequently determine the document similarity.

The second category of approaches are the ones that go further and provide detailed information, specifically indicating the copied passages found. These models can be considered as the "second generation" of algorithms for automatic plagiarism detection. Their complexity is generally greater than the models of the first category, and the computation time required is also often greater, as they must analyze the documents in detail. Approaches that fit into this category generally use some kinds of string-matching algorithms. In particular, the use of word n-grams as comparison tokens have been shown to give some flexibility to the detection task, as reworded text fragments could still be detected when using $n = 3$, as studied by Lyon, Malcolm, and Dickerson (2001, 2004); Barrón-Cedeño, Basile, Degli Esposti, and Rosso (2010); Oberreuter, L'Huillier, Ríos, and Velásquez (2010).

Kasprzak and Brandejs (2010) introduced their model for automatic external plagiarism detection. It consists of two main phases (,) building the document index and computing the similarities. This approach uses word n-grams, with n ranging from 4 to 6, and takes into account the number of matches of those n-grams between the suspicious documents and the source documents for computing the detections. The algorithm has won the authors first place at the PAN@2010 competition (Potthast et al., 2010).

In 2011, within the PAN evaluation framework (Potthast et al., 2011), new studies have been published. Grman and Ravas (2011) compares the passages in terms of word coincidences, and based on this approach they achieved the best results in terms of precision and recall in the evaluation. Grozea and Popescu (2011) and Oberreuter et al. (2011) present their studies, in which both use n-grams as tokens.

4. Proposed method for intrinsic plagiarism detection

For intrinsic plagiarism detection, first we considered some other investigators' studies regarding the characterization of the writing style of an author. As Stein et al. (2011) investigated, multiple writing style characteristics were tested in order to determine plagiarism, for example lexical character features, lexical word features and syntactical features. Likewise, Stamatatos (2009) experimented with character tri-grams in combination with "n-gram profiles" for the same purpose. For this, it is fundamental to carefully choose one or a set of language resources an author utilizes for his writing to be able to differentiate it from others.

In the following, some of the core ideas developed in this research are presented:

- To be able to distinguish different authors within the same document, one must characterize the writing style present in the text.
- The use of "n-gram profiles" compares segments of the document against the whole document. This approach works based on the assumption that the document has a main author, who wrote the majority, if not all, of the text. Therefore, it is logical that the comparison between the style of a particular segment with the whole document style could lead to detections of important variations, meaning that other authors are involved.
- Based on reading and contemplation, one of the characteristics that was shown to be of interest is the author's use of words. Different authors tend to use different words to write their ideas, whether on the same topic or not.

These ideas lead to the following intuition for the development of the algorithm: If some of the words used in the document are author-specific, one can think that those words could be concentrated in the paragraphs (or more generally, in the segments) that the mentioned author wrote.

4.1. The method

First, the document is preprocessed by removing numbers and all other characters that do not belong to the a–z group. All characters are considered lowercase. Second, the method uses word unigrams and considers all words; stop-words are not removed. Next, a word-frequency-based algorithm to test the self-similarity of a document is proposed. A hard (not normalized) frequency vector $\mathbf{v}$ is built for all words in the given document. Then, the complete document is clustered creating groups $\mathcal{C}$. As a first approach, these groups or segments $c \in \mathcal{C}$ are created using a sliding window of length m over the complete document. Afterwards, for each segment $c \in \mathcal{C}$, a new frequency vector v_c is computed, which is used in further steps to compare whether a segment deviates with respect to the footprint of the complete document.

Let $\mathcal{V}$ be a vector of words that defines the vocabulary to be used. We will refer to a word w, as a basic unit of discrete data, indexed by $\{1,\ldots,|\mathcal{V}|\}$. A document d is a sequence of S words ($|d| = S$) defined by $\mathbf{w} = (w^1,\ldots,w^S)$, where w^s represents the sth word in the message. Finally, a corpus is defined by a collection of $\mathcal{D}$ documents denoted by $\mathcal{C} = (\mathbf{w}_1,\ldots,\mathbf{w}_{|\mathcal{D}|})$.

As presented in Algorithm 1, the general footprint or style of the document is represented by the average of all differences computed for each segment and the complete document. Note that

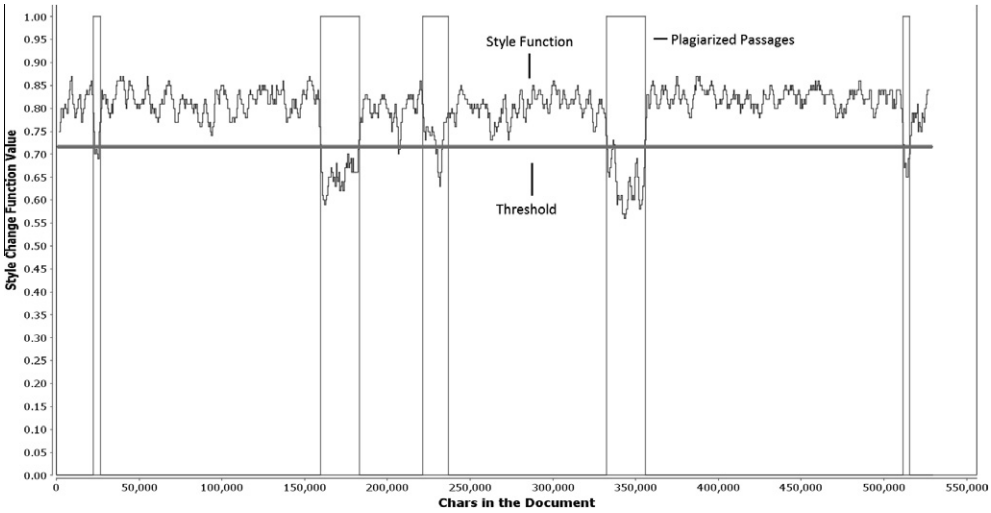

Fig. 1. Intrinsic plagiarism detection example. One document is being analyzed, its style function changing as the sliding window moves forward.

every segment is compared against the whole document only in terms of the words present in the segment. Also, this algorithm takes into account the above intuition,that if certain words are only used in a certain segment, the comparison of that segment against the whole document would lead to a low value, because the frequency of those words would be the same in both the whole document and in the segment. Finally, all segments are classified according to their distance with respect to the document's style. As an example, in Fig. 1, a graphical representation of this evaluation is presented.

Algorithm 1. Intrinsic plagiarism evaluation

Require: $\mathcal{C}$, $\mathbf{v}$, m, δ
1: **for** $c \in \mathcal{C}$ **do**
2: $d_c \leftarrow 0$
3: build v_c using term frequencies on segment c
4: **for** word $w \in v_c$ **do**
5: $d_c \leftarrow d_c + \frac{|freq(w,\mathbf{v}) - freq(w,v_c)|}{freq(w,\mathbf{v}) + freq(w,v_c)}$
6: **end for**
7: **end for**
8: style $\leftarrow \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} d_c$
9: **for** $c \in \mathcal{C}$ **do**
10: **if** $d_c <$ *style* $- \delta$ **then**
11: Mark segment c as outlier and potential plagiarized passage.
12: **end if**
13: **end for**

The main function in Algorithm 1, fifth line, computes the differences in the use of words of two segments. The function is constructed so segments of the document that have many words that are exclusively in that segment will have a low value. The idea is that the use of words should be stable, with at least a high proportion of the words used throughout the document. If a portion of the text has a high proportion of its words isolated in that portion, the function value will be below the average, hinting at a possible change in writing style.

Since the algorithm considers the information of each document to construct and evaluate variations in style, the function remains somewhat stable over varying document lengths. The strong assumption here is that the majority of the text was written with the same writing style, otherwise no reliable information could be extracted from this model.

In this case, the average value of the comparison of all segments with the whole document represents the document "main" style. This value is roughly computed by the difference in the frequency of words between vectors $\mathbf{v}$ and v_c, $\forall c \in \mathcal{C}$. If the variation is significant, the style function will be lower than the average value minus δ (the threshold), so the segment is classified as suspicious. In this example, real plagiarized annotations are presented along with the style function value of each segment. Five cases of plagiarism could be discovered; the value of the style function in those cases is lower than the threshold.

As a preview of the algorithm's output, we analyze four different books: "Don Quijote de la Mancha" from Miguel de Cervantes Saavedra (Spanish), "El Sombrero de Tres Picos" from Pedro Antonio de Alarcón (Spanish), "Les Misèrables from Victor Hugo (French) and the English translation of "War and Peace" from Leo Tolstoy.

In Fig. 2 the model's representation of the writing style remains stable throughout the text. Variations can be observed at the beginning and at the end of the document; these zones are typically the colophon (technical info.), content index and preface, so variations in style can be expected.

In Fig. 3 an errant evolution of the style's value can be observed. No conclusion can be stated here; either the model does not work well in this particular example or different styles are present in the book, rendering the purpose of this exercise useless.

Fig. 4 shows the style of the book *Les Misèrables*. The function remains stable as a whole, with a few segments whose values are under the threshold.

The analysis of the translated book of Leo Tolstoy, shown in Fig. 5, indicates a mainly stable writing style, with slight variations.

5. Evaluation

We evaluate the presented approach utilizing the PAN corpora (Potthast et al., 2011), which is publicly available.[1] The metrics are described in detail in Potthast et al. (2009), and are common information retrieval measures, which are adapted to be applied to cases of copying.

- *Precision*, which states the degree to which a pair of passages identified as a plagiarism case indeed have copying between them, and *recall*, which states the percentage of plagiarized passages that the classifier manages to classify correctly. These

[1] http://www.pan.webis.de/

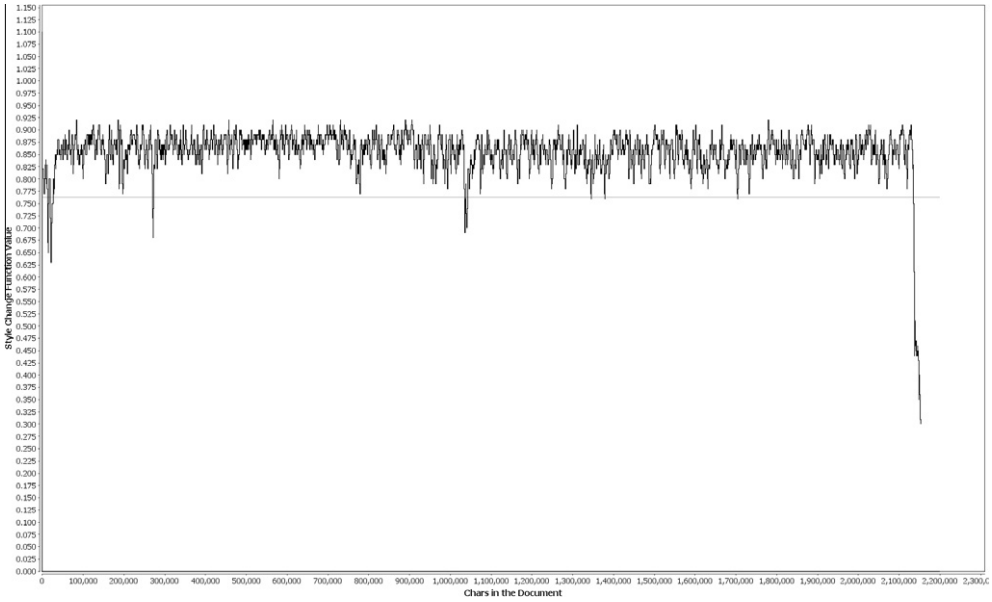

Fig. 2. Don Quijote de la Mancha.

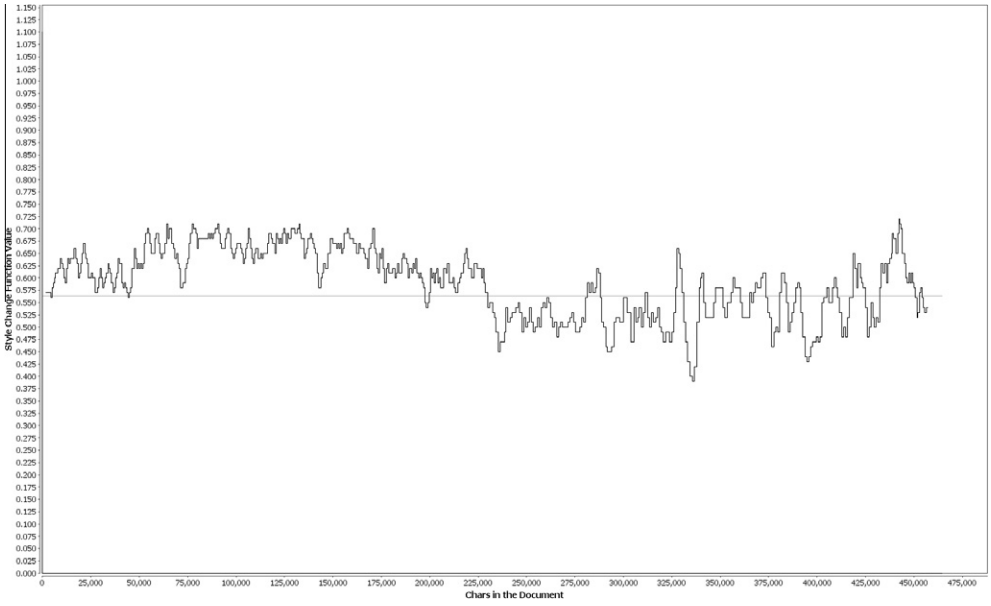

Fig. 3. El Sombrero de Tres Picos.

measures can be interpreted in conjunction with the classifier's effectiveness. *TP* means "True Positive" – documents found to be plagiarized – a detection which is correct. *FP* means "False Positive", that a document that should have been identified as plagiarized, was not. *TN* means that a document was classified as plagiarized, which is incorrect. And finally, *FN* indicates a document that was not identified as plagiarized, when the correct decision would have been the opposite.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{1}$$

- *F-measure*, the harmonic mean between precision and recall.

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2}$$

- *Granularity*, is a metric introduced by Potthast et al. (2009) to evaluate the usability of the algorithm; each real case of plagiarism should be reported once. If the algorithm report is fragmented (the detections are multiple for only one real case), the granularity increases. The desirable granularity is 1.

$$\text{Granularity} = \frac{1}{|S_R|} \sum_{s \in S_R} |C_S| \tag{3}$$

S_R correspond to the correctly detected cases, and $|C_S|$ correspond to the number of detections for case s.

- *Overall Score*. The overall score is calculated as follows:

$$\text{Overall} = \frac{\text{F-measure}}{\log_2(1 + \text{granularity})} \tag{4}$$

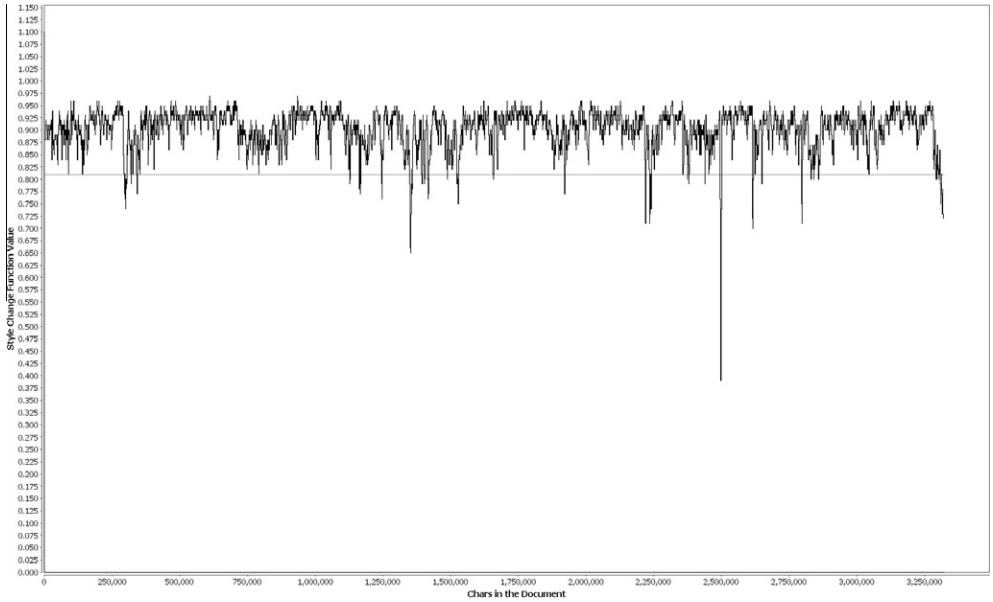

Fig. 4. Les Misèrables.

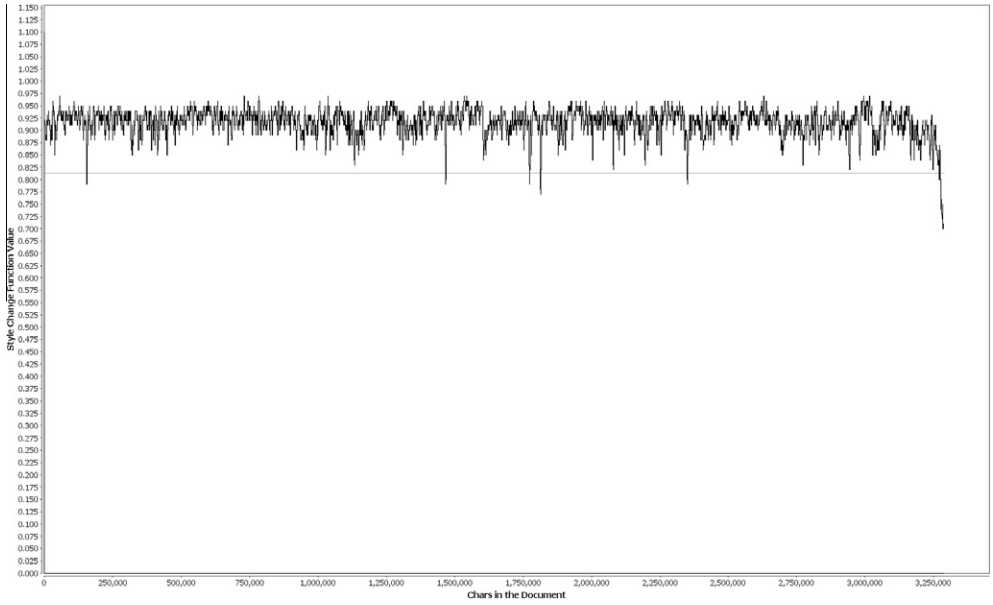

Fig. 5. War and Peace.

5.1. Parameters analysis

The experimentation was conducted using a sliding window of 400 words.

In Fig. 6 the results of using different thresholds are presented to explore the effects on precision and recall. As one can expect, with a tight threshold the model achieves a better recall at the cost of precision. The best results in terms of f-Measure are obtained when both metrics, recall and precision, are balanced. This parameter can be adjusted and calibrated depending on the sensitivity desired.

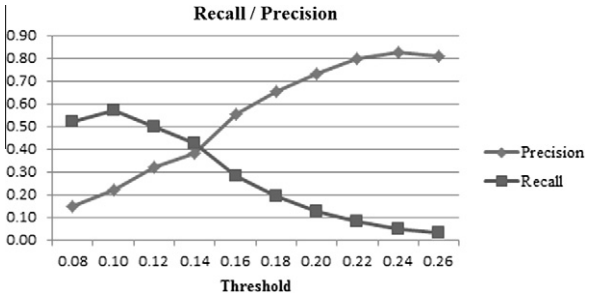

Fig. 6. Precision and Recall when adjusting the threshold.

5.2. Intrinsic plagiarism detection

The results for the intrinsic task at PAN@2009 are shown in Table 1 and for PAN@2011 in Table 2. The results are based on the quality of the detection, which only considers the information in each document itself.

Table 1
Results for the intrinsic proposed model using the corpus PAN2009.

Rank	Overall	Precision	Recall	Granularity	Authors
1	0.25	0.23	0.46	1.38	Stamatatos
2	0.20	0.11	0.94	1.00	Hagbi
3	0.18	0.20	0.27	1.45	Muhr
4	0.12	0.10	0.56	1.70	Seaward
**	**0.35**	**0.39**	**0.31**	**1.00**	**Oberreuter**

Table 2
Official results for the intrinsic proposed model using the corpus PAN2011.

Rank	Overall	Precision	Recall	Granularity	Authors
1	**0.33**	**0.34**	**0.31**	**1.00**	**Oberreuter**
2	0.17	0.43	0.11	1.03	Luyckx
3	0.08	0.13	0.07	1.05	Akiva
4	0.07	0.19	0.08	1.48	Gupta

Table 3
Detailed results for the intrinsic proposed model using the corpus PAN2011.

Corpus subset	Overall	Precision	Recall	Granularity
entire	0.33	0.31	0.34	1.00
Case length				
short	0.03	0.02	0.16	1.00
medium	0.26	0.19	0.45	1.00
long	0.36	0.26	0.57	1.00
Translation				
automatic	0.31	0.34	0.29	1.00
manual	0.14	0.10	0.22	1.00
Plagiarism per document				
hardly	0.37	0.45	0.32	1.00
medium	0.35	0.33	0.36	1.00
Document length				
short	0.38	0.37	0.38	1.00
medium	0.40	0.44	0.37	1.00
long	0.28	0.32	0.25	1.00

In the official workshop the winner was a Stamatatos (2009) approach, with a recall of 0.4607, precision of 0.2321 and granularity of 1.3839. This method achieved a good combination of precision and recall, but was not a top performer in granularity.

The proposed method, in the same workshop, managed to get an overall score of 0.3457, greater than any other approach, with a positive difference of 0.0995 compared with the winner's approach. Our model gets the best results in F-measure, precision and granularity.

These numbers are confirmed with similar results in the PAN@2011 competition presented in Table 2; the proposed model gets roughly the same overall score, 0.3254, with comparable precision (0.34) and worse but not significantly different recall (0.31). We get the best results in the official competition, followed by Luyckx et al., with an overall score of 0.17, almost doubling their score.

The detailed results are presented in Table 3. The complete results including the other models that were evaluated with the corpus can be found in Potthast et al. (2011).

The short cases range from 50 to 150 words, medium cases range from 300 to 500 words and long cases range from 3000 to 5000 words. As shown in Table 3, the model works best with long cases, achieving an overall score of 0.36. With short cases the model shows its weaknesses – its recall in this case is 0.16 but it is highly unreliable, only achieving a precision of 0.02.

In the case of translated cases, the model gets an overall score of 0.34 in the automatic translation, mirroring the score of the entire corpus. In manual translation however its performance is not as high, with a score of 0.14.

The approach achieves an overall score of 0.37 in the document with only a small percentage of plagiarism (between 5 and 20 %). The most remarkable metric is the precision, in this case of 0.45, the best for this set of parameters. This can be explained by the fact that the model builds itself with information taken from each document. If only a small portion of the document has been plagiarized, and thus only a small portion of the document writing style is different, the model can better quantify this variation.

In the case of documents with medium-sized portions of plagiarism (between 20 and 50%), the model achieves an overall score of 0.35.

Other interesting results are the performance of the model with short documents (from 1 to 10 pages) in which the algorithm achieves an overall score of 0.38, and with medium documents (10 to 100 pages) a score of 0.40. In the case of long documents (100 to 1000 pages) its score is 0.28. The better results with short and medium-sized documents can be explained by the assumption that with this size it can be expected that the main author wrote almost if not all the text without much intervention, so its style remains as is. In the case of long documents one can expect that the author wrote with the help of other people, therefore intervention in the document main style can be observed, and as a consequence is more difficult to analyze and identify.

6. Conclusions

In this study we explore the problem of text plagiarism and the possibility of its detection by the use of computer algorithms. With the rising utilization of digital documents and the Web, plagiarism is increasing as well. In view of this, a proliferation of techniques and approaches to detect digital plagiarism have been introduced, and as seen, huge progress is being made in the field of automatic plagiarism detection.

One of the first problems the systems face is the collection of possible sources to compare the suspected documents with. This represent an entire problem in itself, and it is common that the ideal and real sources are not always available, limiting the potential of algorithms that compute similarity document-to-document.

Considering the latter issue, algorithms that do not rely on the available sources are being studied. This is why the so-called intrinsic plagiarism detection concept was introduced. The idea, to analyze the document looking for variations that could hint at plagiarized passages, was recently tested and studies utilizing different writing style markers are being introduced. The study of linguistic features for the data mining process here is crucial, therefore the exploration of different approaches and writing style characteristics is welcomed.

We study a self-based information algorithm, whose basic idea is the use of a function to quantify the writing style based solely on the use of words. Our experiments show that this simple and easily computable idea can be used for this purpose with the best results available so far, when compared using the benchmark workshop and competition PAN (Potthast et al., 2011).

The results in terms of precision are low (0.3) indicating the still immature and unreliable nature of the approach, but so far it has been the most accurate model in this regard. The main experiments were conducted using documents written in English, but the method does not utilize language-dependent features such as verbs or stop-words, thus providing a starting point to experiment with other languages.

Future work will be necessary to study the impact and implications of different sets of parameters on the behavior of the model, especially with short cases of plagiarism (50–150 words) where the model shows its weaknesses by obtaining very low precision.

In this study experiments in varying the threshold were conducted, obtaining a reasonable trade-off between precision and recall, showing that the usage of words can be analyzed and utilized to detect variations in style with great accuracy at the cost of detecting fewer cases.

Acknowledgments

The authors would like to acknowledge the continuous support of the Chilean Millennium Institute of Complex Engineering Systems (ICM: P-05-004-F, CONICYT: FBO16); the INNOVA CORFO project (11DL2-10399) entitled, DOCODE: Document Copy Detection (www.docode.cl). Gabriel Oberreuter is currently "Becario CONICYT".

References

Baayen, H., van Halteren, H., & Tweedie, F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*, 121–132.

Bao, J.-P., Shen, J.-Y., Liu, X.-D., Liu, H.-Y., & Zhang, X.-D. (2004). Semantic sequence kin: A method of document copy detection. In H. Dai, R. Srikant, & C. Zhang (Eds.), *Advances in knowledge discovery and data mining. Lecture notes in computer science* (Vol. 3056, pp. 529–538). Berlin/Heidelberg: Springer.

Barrón-Cedeño, A., Basile, C., Degli Esposti, M., & Rosso, P. (2010). Word length *n*-grams for text re-use detection. In A. Gelbukh (Ed.). *Computational linguistics and intelligent text processing. lecture notes in computer science* (Vol. 6008, pp. 687–699). Berlin/Heidelberg: Springer.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*, 573–595.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on computational learning theory COLT '92* (pp. 144–152). New York, NY, USA: ACM.

Bravo-Marquez, F., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2011). A text similarity meta-search engine based on document fingerprints and search results records. *Proceedings of the 2011 IEEE/WIC/ACM international conferences on web intelligence and intelligent agent technology – WI-IAT '11* (Vol. 01, pp. 146–153). Washington, DC, USA: IEEE Computer Society.

Ceska, Z. (2008). Plagiarism detection based on singular value decomposition. In *GoTAL '08: Proceedings of the sixth international conference on advances in natural language processing* (pp. 108–119). Berlin/Heidelberg: Springer.

Chow, T. W. S., & Rahman, M. K. M. (2009). Multilayer SOM with tree-structured data for efficient document retrieval and plagiarism detection. *Transactions on Neural Networks, 20*, 1385–1402.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*, 251–270.

Grman, J., & Ravas, R. (2011). Improved implementation for finding text similarities in large sets of data – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Grozea, C., & Popescu, M. (2011). The encoplot similarity measure for automatic detection of plagiarism – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

van Halteren, H. (2004). Linguistic profiling for author recognition and verification. In *Proceedings of the 42nd annual meeting on association for computational linguistics ACL '04*. Stroudsburg, PA, USA: Association for Computational Linguistics.

Jun-Peng, B., Jun-Yi, S., Xiao-Dong, L., Hai-Yan, L., & Xiao-Di, Z. (2003). Document copy detection based on kernel method. In *Proceedings of the 2003 International Conference on Natural Language Processing and Knowledge Engineering* (pp. 250–255).

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval, 1*, 233–334.

Kasprzak, J., & Brandejs, M. (2010). Improving the reliability of the plagiarism detection system – lab report for pan at clef 2010. In M. Braschler, D. Harman, & E. Pianta, (Eds.), *CLEF 2010 labs and workshops, notebook papers*. 22–23 September 2010, Padua, Italy.

Kern, R., Seifert, C., Zechner, M., & Granitzer, M. (2011). Vote/veto meta-classifier for authorship identification – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Kohonen, T. (2001). *Self-organizing maps. Springer series in information sciences*. Springer.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology, 60*, 9–26.

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation, 45*, 83–94. http://dx.doi.org/10.1007/s10579-009-9111-2.

Kourtis, I., & Stamatatos, E. (2011). Author identification using semi-supervised learning – notebook for pan at clef 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Lyon, C., Barrett, R., & 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. In *Proceedings of plagiarism: Prevention, practice and policies conference*. Newcastle, UK.

Lyon, C., Malcolm, J., & Dickerson, B. (2001). Detecting short passages of similar text in large document. In *Proceedings of the 2001 conference on empirical methods in natural language processing* (pp. 118–125). Pennsylvania.

Meyer zu Eißen, S., & Stein, B. (2006). Intrinsic plagiarism detection. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in information retrieval. 28th European conference on IR research (ECIR 06). Lecture notes in computer science* (Vol. 3936, pp. 565–569). Berlin/Heidelberg, New York: Springer.

Meyer zu Eißen, S., Stein, B., & Kulig, M. (2007). Plagiarism detection without reference collections. In R. Decker & H. Lenz (Eds.), *Advances in data analysis. selected papers from the 30th annual conference of the german classification society (GFKL 06). Studies in classification, data analysis, and knowledge organization* (pp. 359–366). Berlin/Heidelberg, New York: Springer.

Oberreuter, G., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2010). Fastdocode: Finding approximated segments of *n*-grams for document copy detection: Lab report for pan at CLEF 2010. In M. Braschler, D. Harman, & E. Pianta, (Eds.), *CLEF 2010 labs and workshops, notebook papers*. 22–23 September 2010, Padua, Italy.

Oberreuter, G., L'Huillier, G., Ríos, S. A., & Velásquez, J. D. (2011). Approaches for intrinsic and external plagiarism detection – notebook for pan at CLEF 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., & Rosso, P. (2010). Overview of the 2nd international competition on plagiarism detection. In M. Braschler, D. Harman, & E. Pianta (Eds.), *Notebook papers of CLEF 10 labs and workshops*.

Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., & Rosso, P. (2011). Overview of the 3rd international competition on plagiarism detection. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., & Rosso, P. (2009). Overview of the 1st international competition on plagiarism detection. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09)* (pp. 1–9). CEUR-WS.org.

Rao, S., Gupta, P., Singhal, K., & Majumder, P. (2011). External & intrinsic plagiarism detection: Vsm & discourse markers based approach – notebook for pan at clef 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Schleimer, S., Wilkerson, D. S., & Aiken, A. (2003). Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on management of data SIGMOD '03* (pp. 76–85). New York, NY, USA: ACM.

Seaward, L., & Matwin, S. (2009). Intrinsic plagiarism detection using complexity analysis. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09)* (pp. 56–61). CEUR-WS.org.

Stamatatos, E. (2009). Intrinsic plagiarism detection using character *n*-gram profiles. In B. Stein, P. Rosso, E. Stamatatos, M. Koppel, & E. Agirre (Eds.), *SEPLN 2009 workshop on uncovering plagiarism, authorship, and social software misuse (PAN 09)* (pp. 38–46). CEUR-WS.org.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*, 538–556.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (1999). Automatic authorship attribution. In *Proceedings of the ninth conference on European chapter of the association for computational linguistics EACL '99* (pp. 158–164). Stroudsburg, PA, USA: Association for computational linguistics.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Computers and the Humanities, 35*, 193–214. http://dx.doi.org/10.1023/A:1002681919510.

Stein, B., Lipka, N., & Prettenhofer, P. (2011). Intrinsic plagiarism analysis. *Language Resources and Evaluation, 45*, 63–82.

Tanguy, L., Urieli, A., Calderone, B., Hathout, N., & Sajous, F. (2011). A multitude of linguistically-rich features for authorship attribution – notebook for pan at clef 2011. In V. Petras, P. Forner, & P. D. Clough, (Eds.), *CLEF 2011 labs and workshop, notebook papers*. 19–22 September 2011, Amsterdam, The Netherlands.

Vapnik, V. N. (1999). *The nature of statistical learning theory (Information science and statistics)*. Berlin/Heidelberg: Springer.