NEW APPROACH FOR PLAGIARISM DETECTION

Dan Lemberg^{1 §}, Avi Soffer², Zeev Volkovich³

^{1,2,3}Software Engineering Department
ORT Braude College of Engineering
Karmiel 21982, ISRAEL

Abstract: The paper proposes a new approach for intrinsic plagiarism detection, based on a new unique method, which enables identifying style changes in a text using novel chronology-based similarity measures. A model for finding significant deviations in the style across a given document is constructed aiming to indicate text parts which are suspected to be written by co-authors, or to be devoted to a different thematic, or to be a plagiarism. We consider each segment as “result of the text evolution” provided by its predecessors in the text. Resting upon this evolution standpoint, the metric evaluating dissimilarity between two given segments is introduced, and a text is clustered using this measure aiming to turn out disparity of the text. We also propose a new clustering procedure involving an embedding of data in an Euclidean space with subsequent clustering using the *K-means* approach. The obtained results demonstrate high ability of the method.

AMS Subject Classification: 94A13, 62H30, 68T10

Key Words: plagiarism detection, evolution metric, n-gram model

1. Introduction

Plagiarism has become a widespread phenomenon with the proliferation of electronic publishing and availability of on-line sources. Plagiarism detection is the process of locating instances of plagiarism within a document or a book. Various methods can be used for Computer-assisted Plagiarism Detection including

Received: March 22, 2016

© 2016 Academic Publications

[§]Correspondence author

document-comparison algorithms and approaches that utilize language-specific features, for example authorship attribution. A comparative review of these methods can be found in [3]. Our work proposes a new approach for intrinsic plagiarism detection, based on a new unique method, which enables identifying style changes in a text using novel chronology-based similarity measures. A model for finding significant deviations in the style across a given document is constructed aiming like [3] to indicate text parts which are suspected to be written by co-authors, or to be devoted to a different thematic, or to be a plagiarism.

In our methodology, the examined text is divided in a chain of sequential equal sized pieces such that each one of them is represented as a histogram of *3-grams*. After that, the *3-grams* in a texts segment are ranked according to their frequencies in a histogram. Then, a connection between the segments is characterized by means of the Spearman rank correlation coefficient averaged during numerous segment for-goings. Namely, we consider each segment as “result of the text evolution” provided by its predecessors in the text. Resting upon this evolution standpoint, the metric evaluating dissimilarity between two given segments is introduced, and a text is clustered using this measure aiming to turn out disparity of the text. We also propose a new clustering procedure involving an embedding of data in an Euclidean space with subsequent clustering using the *K-means* approach. The methodology described in detail in Section 2, and in Section 3 we present several results of applying our method on real data. The obtained results demonstrate high ability of the proposed method.

2. Method

As was mentioned earlier, our purpose is to divide a text into segments written by different styles. To implement it we introduce the following model. Let us introduce a set $D_{j,T} = \{x_{j-T}, \dots, x_{j-1}\}$ of T sequential chunks of a document $\mathbf{X} = \{x_1, \dots, x_m\}$ divided in m sequential chunks. The *Mean Rank Dependency* of item x_i with a set $D_{j,T}$ is defined as

$$ZV(x_i, D_{j,T}) = \frac{1}{T} \sum_{x \in D_{j,T}} \rho(x_i, x), \quad i = T + 1, \dots, m,$$

where ρ is the famous Pearson’s correlation coefficient (see, e.g. [4]) calculated between the ranks of the *3-grams* within the histograms, which are obtained via the ordering of the *3-grams* according to with their frequencies in the chunks.

Targeting to distinguish writing styles in a document a new function quantifying difference amid texts pieces is proposed by the following way:

$$DZV_T(x_i, x_j) = |ZV(x_i, D_{i,T}) + ZV(x_j, D_{j,T}) - ZV(x_i, D_{j,T}) - ZV(x_j, D_{i,T})|.$$

It is easy to see that it is a semi-metric. So, $DZV_T(x_i, x_j) \geq 0$ and $DZV_T(x_i, x_i) = 0$ for every x_i , but $DZV_T(x_i, x_j) = 0$ does not formally imply that $x_i = x_j$. In the last case a sub-document exhibits close relationships with own foregoing neighbors and the foregoing neighbors of another sub-document. From the writing style standpoint the sub-documents appear to be identical. The fact, that DZV_T is actually not a metric, might lead to ambiguity of a clustering process. To overcome this obstacle we provide a metrication of $(\mathbf{X}, DZV_T)$ by means of the Fréchet-Kuratowski embedding [5], [2] into the Euclidean space $\mathbb{R}^m$ equipped by the standard Euclidean distance induced by norm $\|\cdot\|$:

$$\begin{aligned} \pi : (\mathbf{X}, DZV_T) &\rightarrow (\mathbb{R}^m, \|\cdot\|) \\ \pi(x) &= (DZV_T(x, x_1), \dots, DZV_T(x, x_m)), \end{aligned} \quad (1)$$

which induce a new metric on $\mathbf{X}$:

$$DZVE_T(x, y) = \|\pi(x) - \pi(y)\|. \quad (2)$$

Now, the well-known K -means algorithm [6], [8] can be fairly applied. The proposed methodology for plagiarism detecting is as follows:

- Any stop-words, punctuation and spaces are removed from the text.
- The text is divided into non overlapping sequential segments $x_1, \dots, x_m$ of length l .
- Set $\mathbf{X}$ endowed with the distance $DZVE_T$ (eq.2) is clustered using the K -means algorithm for the different number of clusters $2 \leq k \leq k_{\max}$, where $k_{\max}$ is predefined parameter presenting the maximal number of clusters to be tested.
- The optimal number of clusters is chosen according to the silhouette criterion [7].
- Texts segments belonging to different clusters in the partition obtained for the optimal number of clusters are interpreted as ones written by different styles. (So, the clusters are the parts of the text, written in the same style.)

Description	Symbol	Value
Maximal clusters number	$K_{\max}$	10
Size of segments	l	1000
Length of evolution	T	10

Table 1: Used parameters.

3. Numerical Experiment

We provide several experiments using the following values of the algorithm parameters. Below are the results of plagiarism detecting:

Figures 1, 2 show analysis results of the union from three books: “2010: Odyssey Two” A. Clarke, “Fahrenheit 451” R. Bradbury and “The Da Vinci Code” D. Brown. Figure 2 shows the results of calculating the average value of silhouettes, from which it is clear that the optimal number of clusters is $k = 3$. Figure 1 shows the results of clustering. Namely, the segments from first to 234 form the first continuous cluster, segments from 235 to 355 form the second continuous cluster and 356 to 627 form the third. What exactly corresponds to the first, second and third book.

Figures 3, 4 show analysis results of the file document 00002 from database PAN-PC-09, [10]. Figure 4 shows the results of calculating the average value of silhouettes, from which it is clear that the optimal number of clusters is $k = 2$. Figure 3 shows the results of clustering. Namely, the segments from first to 114, from 121 to 258 and from 265 to the end of the file form one cluster (base text). Segments from 115 to 120 and from 258 to 264 form two continuous clusters (plagiarism).

Figures 5, 6 show analysis results of “2010: Odyssey Two” A. Clarke. Figure 4 shows the results of calculating the average value of silhouettes, from which it is clear that the optimal number of clusters is $k = 2$. On the other hand, as can be seen from the diagram in Figure 5, the number of continuous clusters greatly exceeds the $K_{\max}$. Thus the text is considered plagiarism-free.

References

- [1] E. Stamatatos, Intrinsic plagiarism detection using character n-gram profiles, In: B. Stein, P. Rosso, E. Stamatatos, M. Koppel, E. Agirre (Eds.), *Proc. of the SEPLN '09 Workshop on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN 09)*, CEUR-WS.org (2009), 38-46.
- [2] M. Fréchet, Sur quelques points du calcul fonctionnel, *Rendiconti del Circolo Matematico di Palermo*, **22** (1906), 1-74.

Figure 1: 2010: Odyssey Two, Fahrenheit 451 and The Da Vinci Code.

Figure 2: Silhouettes mens of 2010: Odyssey Two, Fahrenheit 451 and The Da Vinci Code.

Figure 3: Document 00002.

Figure 4: Silhouettes mens of document 00002.

- [3] G. Oberreuter, J.D. Velasquez, Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style, *Expert Systems with Applications*, **40** (2013), 3756-3763.
- [4] M.G. Kendall, J.D. Gibbons, *Rank Correlation Methods* (5th ed.), Edward Arnold, London (1990).
- [5] C. Kuratowski, Quelques problèmes concernant les espaces métriques non-séparables (Some problems concerning non-separable metric spaces), *Fundamenta Mathematicae*, **25** (1935), 534-545.
- [6] J.B. MacQueen, Some methods for classification and analysis of multivari-

Figure 5: 2010: Odyssey Two.
(A plagiarism-free document.)

Figure 6: Silhouettes mens of
2010: Odyssey Two.

ate observations, In: *Proc. of 5th Berkeley Symposium on Mathematical Statistics and Probability 1*, Univ. of California Press (1967), 281-297.

- [7] J. Peter, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Rousseeuw J. of Computational and Applied Mathematics*, **20** (1987), 53-65.
- [8] H. Steinhaus, Sur la division des corps matériels en parties, *Bull. Acad. Polon. Sci.* **4**, No 12 (1957), 801-804 (in French); MR 0090073, Zbl 0079.16403.
- [9] T. Dunning, Statistical identification of language, *Technical Report MCCS*, New Mexico State University (1994), 94-273.
- [10] <http://www.uni-weimar.de/en/media/chairs/webis/corpora/pan-pc-09/#webis-download>