DOI: 10.5769/C2012004 or http://dx.doi.org/10.5769/C2012004

Authorship of Electronic Documents

Use Data Compression to Attribute Authorship in Brazilian Portuguese Electronic Documents

Walter Oliveira[1], Edson Justino[1], Luiz Oliveira[1] and Alexandre Brondani[2]
(1) Pontifícia Universidade Católica do Paraná
(2) Instituto de Criminalistica do Paraná

Abstract—Discovering the author of electronic documents can be a challenging task, since some elements that are found in traditional documents aren't present. In such task, the expert can use tools that help him finding out the writing style of the probable authors. One possible method is the use of data compression to identify which authors have a similar stylistic. In this work, 300 documents written in Brazilian Portuguese from 20 different authors are analyzed, and results are compared with previous techniques.

Keywords- Authorship Attribution; Electronic Documents; Normalized Compression Distance; Conditional Complexity of Compression.

I. Introduction

Authorship attribution is the task of verifying who, from a set of possible authors, is the probable author of a given document. In electronic documents some elements that could be used to accomplish this task aren't available, for example the handwriting style of the authors and ink and papers elements. Other elements have to be used, and sometimes the only available characteristic is the writing style of the author.

Many techniques are used to identify this stylistics, and computational tools can help providing faster and less subjective ways of identifying and analyzing the profile of each author. The profile is defined on how the author uses the many writing elements to compose his documents.

According to Stamatatos [1], those elements can be divided according to some characteristics in four categories: character-based characteristics (e.g. symbols used, punctuation marks), word-based characteristics (e.g. word frequency, vocabulary richness), lexical features (e.g. how the sentences are structured) and content-specific characteristics (e.g. how the HTML formatting is applied in a Web page).

Methods that use data compressors to classify documents have been proposed. Those methods rely on character characteristics, since the compressor uses the symbols from an archive to generate better models of representation of the archive contents and achieve a better compression. Among those methods, two will be studied: the Normalized Compression Distance (NCD) and the Conditional Complexity of Compression (CCC).

Those methods are better explained in the second section of this paper, after an introduction to the Kolmogorov theory that embases them. In the third section, the database used in the experiments is described. The fourth section explains the methods that were used, and in the fifth section the results are presented and analyzed. In the last section conclusions are made.

Since the documents used in this work were previously used in other works, results can be compared with another technique based on Support Vector Machine (SVM) classifiers. In that technique, statistics from known documents were extracted to create an author profile. These statistics were related to verbs, pronouns, adverbs and conjunctions frequencies. The same characteristics were extracted from the questioned documents and then analyzed with the use of a SVM classifier. This method is better described in [2].

II. Theory

A. Kolmogorov Complexity

The NCD and the CCC methods are based on the Kolmogorov theory on information complexity, so a brief introduction to this theory is necessary.

Kolmogorov proposed, in 1965, a theory to analyze how complex some information is [3]. According to this theory, an information complexity can be measured by the amount of symbols required to represent that information in some universal language. So, the Kolmogorov Complexity of some information x, $K(x)$, is the length of that information in the universal language. Intuitively, information like "01010101010101010101" seems to be less complex than "01101010110100101010", since it's possible to easily extract some pattern from it.

It's also possible to establish the conditional Kolmogorov complexity of some information. The conditional complexity

K(x|y) of an information *x* is the length of the program *y* that is processed through a Turing machine and gives *x* as the output.".

But the Kolmogorov complexity is incomputable, and so are its upper and lower bounds [4]. So, it's not possible to be sure that a given *y* is the smallest program to output *x* and halt the Turing machine.

B. Normalized Compression Distance

Based on the Kolmogorov theory, Li and Vitányi [4] affirmed that it's possible to measure the similarity of two information using the Kolmogorov complexity. The Normalized Information Distance (NID) is expressed in (1),

$$NID(x, y) = \frac{max\{K(x|y), K(y|x)\}}{max\{K(x), K(y)\}} \qquad (1)$$

where *x* and *y* are two information, *K(x|y)* is the conditional Kolmogorov complexity of *x* given *y*, *K(x)* is the Kolmogorov complexity of *x* and *max{ . , . }* is a function that returns the biggest of two values. But, since the Kolmogorov complexity is incomputable, so is the NID. But these same authors proposed that the Kolmogorov complexity can be approximated using a compressor, since a compressor tries to output the information *x* given an input *y* [4].

Li, Chen, Li, Ma and Vitányi [5] suggested that it's possible to approximate the NID using compressors and called this similarity measure as the NCD, expressed in (2),

$$NCD(x, y) = \frac{C(xy) - min\{\, C(x),\ C(y)\}}{max\{\, C(x),\ C(y)\}} \qquad (2)$$

where *x* and *y* are the information which NCD will be measured, *xy* is the concatenation of *x* and *y*, *C(x)* is the size of the compressed *x*, *max{ . , . }* is a function that returns the biggest of two values and *min{ . , . }* is a function that returns the smallest of two values.

NCD is a normalized distance and the result should always be in the range [0,1], where a value closer to 0 indicates a bigger similarity among the information.

C. Conditional Complexity of Compression

The CCC was proposed by Malyutov, Wickramasinghe and Li [6] and is also based in the Kolmogorov complexity. The CCC is expressed in (3),

$$CCC(x|y) = C(yx) – C(y) \qquad (3)$$

where CCC(x|y) is the conditional complexity of compression of x, given y, and yx is the concatenation of information y and x.

The authors also define the relative CCC, expressed in (4) as,

$$CCCr(x|y) = CCC(x|y) / x \qquad (4)$$

where *x* is the size of the information *x* and everything else was specified earlier.

III. Database

The database used in the tests consisted in 2 groups, each group having 10 authors and 15 documents for each author. All documents were extracted from online blogs and newspaper, written in Brazilian Portuguese. Information that could reveal the authorship (like name, email) was removed. All hyphenation was also removed, since it was just caused by the newspaper formatting, and was not a relevant content to author identification.

The documents had an average size of 2979 bytes, with a standard deviation of 713. The fig. 1 illustrates the file size distribution.

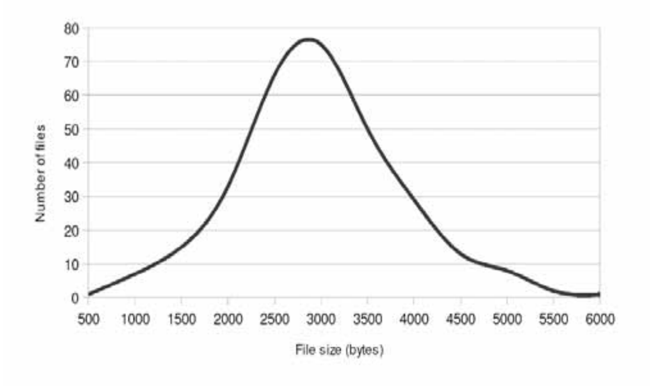

Figure 1. File size distribution of all documents

Tests were performed on both groups, and a third test was executed considering all the 20 authors.

This database was previously used by Pavelec [2], and we used the same separation for training and testing documents, so that it's possible to compare results.

For each test (two tests with 10 authors, one with 20 authors) the documents were separated in two groups: one group was the training set, and was composed by 5 documents for each author, and the testing set, composed by the remaining 10 documents. It was possible to make three rounds of tests, each one using a different subset as the training group.

IV. Method

For each test, documents were separated in training and testing sets, as stated earlier. The evaluation of the NCD and CCC / CCCr was made for each tested document, as described below.

Each tested document had its NCD and CCC / CCCr calculated against each training document. Since there were 5 training documents for each author, each document was tested 5 times for each author, for all the 10 or 20 possible authors. For each tested document 50 (or 100) NCD or CCC

/ CCCr measures were obtained.

Three methods of choice of the best result were studied. In the first method, the authorship attribution was made to the author whose training document had the best NCD or CCC or CCCr. For NCD, the best result is the smallest NCD, since a value closer to 0 indicates that the documents are very similar. For CCC and CCCr, the best value is the also the lower, closer to 0.

The method can be represented by the following algorithm.

```
for each document x in “testing set” (
        C(x) = size of x after compression
        for each document y in “training set” (
C(y) = size of y after compression
C(xy) = size of (x+y) after compression
calculate NCD
        )
        verify which y had the smaller NCD
        attribute authorship to the author of y
)
```

The second method of choice was a voting system. The five best results were considered, and a vote was given to the author of the chosen training document. The author who got most votes was considered as the probable author. For example, if for some test document the best NCD or CCC was given to documents from authors “A, B, A, C, D”, the authorship was attributed to the author “A”.

The third method of choice was the average result of each method. Instead of considering just the best value, it was calculated the average result of the NCD or CCC of the tested document with all the training documents.

The algorithm for the second and third methods is very similar to the only presented previously, with just the step before the actual attribution being made in a different way, as described above.

Then the number of correct attributions was calculated, and is expressed in a percentage.

Since the NCD, CCC and CCCr methods use data compressors to approximate the Kolmogorov complexity but doesn’t have special requirements about the compressors, three different compressors were used. The first compressor was PPM-D (prediction by partial match – escape D), the second was Bzip, and the third was Zip.

The previous work of Pavelec [2] used SVM classifiers. Documents were analyzed and statistical features of adverbs and conjunctions frequencies were extracted. Those statistics were used to train the SVM classifier, and then the tested documents were classified according to similarities in the selected words frequencies. A more detailed explanation can be found in his work.

V. Results and Discussion

A. Choice by best result

The first test was conducted with the first 10 authors and the attribution being made according to the best result. The first tested compressor was Bzip. Results are presented in table 1. The SVM results are from a previous work using a SVM classifier [2] and are independent of the compressor.

TABLE I. Bzip Compressor

Training Set	Method			
	SVM	*CCC*	*CCCr*	*NCD*
1 – 5	80 %	95 %	26 %	97 %
6 – 10	80 %	88%	30 %	100 %
11 – 15	72 %	92 %	25 %	94 %
Average	77,33 %	91,67 %	27 %	97 %

The NCD method presented the best results, with a correct attribution of authorship in 97 % of the documents. It’s possible to observe that the CCCr presents the worst result for these documents.

In table 2 are shown the results for the PPM-D compressor.

TABLE II. PPM-D Compressor

Training Set	Method			
	SVM	*CCC*	*CCCr*	*NCD*
1 – 5	80 %	90 %	28 %	98 %
6 – 10	80 %	87%	34 %	99 %
11 – 15	72 %	95 %	29 %	95 %
Average	77,33 %	90,67 %	30,33 %	97,33 %

The NCD method presented the best results again, but with a very small difference to the Bzip compressor. The CCCr method showed a better performance than when the Bzip compressor was used, but still very inferior to the CCC and NCD methods.

In table 3 are shown the results for the Zip compressor.

TABLE III. Zip Compressor

Training Set	Method			
	SVM	*CCC*	*CCCr*	*NCD*
1 – 5	80 %	94 %	25 %	100 %
6 – 10	80 %	90 %	29 %	99 %
11 – 15	72 %	93 %	29 %	98 %
Average	77,33 %	92,33 %	27,67 %	99 %

The NCD method exhibit the best results again and they were superior to the ones obtained with Bzip and PPM-D compressors.

Comparing all results from the first set of documents, it's possible to observe that the compressor had a small impact on the result, and the best compressor to the CCC and NCD methods is the Zip compressor.

The fig. 2 shows the average results of the three compressors.

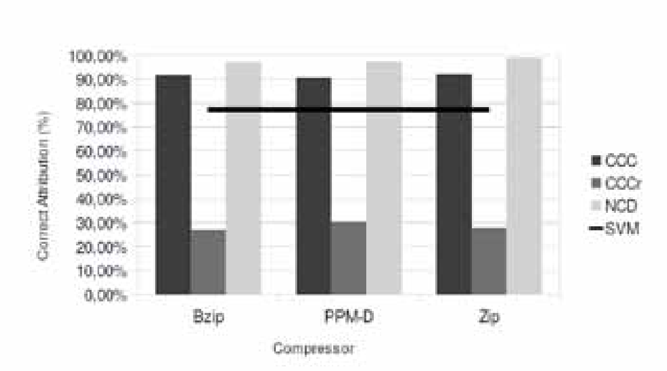

Figure 2. Results comparison

It's possible to observe that both CCC and NCD method had a better correct attribution rate than the SVM classifier, and that the NCD outperformed the CCC in all compressors.

Since the CCCr method produces results that are consistently worse than the other methods, it won't be considered in the next results.

B. *Choice by vote*

The second test was made considering the voting system. In table 4 are presented results for the Bzip compressor.

TABLE IV. Bzip Compressor

Training Set	**Method**		
	SVM	*CCC*	*NCD*
1 – 5	80 %	86 %	93 %
6 – 10	80 %	89%	97 %
11 – 15	72 %	90 %	92 %
Average	77,33 %	88,33 %	94 %

It's possible to observe that, although still superior to the SVM method, CCC and NCD had a worse result compared to the "best value" choosing method, with NCD having a better result than CCC.

In table 5 are presented results for the PPM-D compressor and in table 6 the results of the Zip compressor.

TABLE V. PPM-D Compressor

Training Set	**Method**		
	SVM	*CCC*	*NCD*
1 – 5	80 %	86 %	96 %
6 – 10	80 %	90%	98 %
11 – 15	72 %	90 %	91 %
Average	77,33 %	88,67 %	95 %

TABLE VI. Zip Compressor

Training Set	**Method**		
	SVM	*CCC*	*NCD*
1 – 5	80 %	90 %	97 %
6 – 10	80 %	90%	99 %
11 – 15	72 %	88 %	95 %
Average	77,33 %	89,33 %	97 %

Again it's possible to observe that the ZIP compressor have the best results and NCD is superior to the CCC method. The fig. 3 shows the average result of the three compressors.

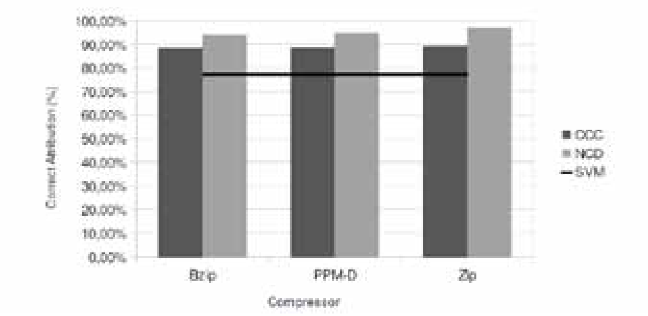

Figure 3. Results comparison

C. *Choice by average result*

In the last attribution method the results for the Bzip, PPM-D and Zip compressors are shown in tables 7, 8 and 9, respectively.

TABLE VII. Bzip Compressor

Training Set	**Method**		
	SVM	*CCC*	*NCD*
1 – 5	80 %	90 %	91 %
6 – 10	80 %	86%	97 %
11 – 15	72 %	89 %	91 %
Average	77,33 %	88,33 %	93 %

TABLE VIII. PPMD Compressor

Training Set	Method		
	SVM	*CCC*	*NCD*
1 – 5	80 %	86 %	92 %
6 – 10	80 %	84%	97 %
11 – 15	72 %	89 %	90 %
Average	77,33 %	86,33 %	93 %

TABLE IX. Zip Compressor

Training Set	Method		
	SVM	*CCC*	*NCD*
1 – 5	80 %	89 %	97 %
6 – 10	80 %	85%	98 %
11 – 15	72 %	90 %	97 %
Average	77,33 %	88 %	97,33 %

It’s possible to observe that the NCD method gives better results than the CCC and SVM methods. The Zip compressor also gives the best results.

Comparing with the other choosing methods for the authorship attribution, the attribution made using the “best result” gives the best result, followed by the “average result” and then by the voting system.

Since NCD is a measure of similarity among information and documents, this result might be caused by the fact that having only one similar document is enough to characterize the author profile. But, with this technique, problems might occur when due to an intentional or unintentional choice of words, outliers aren’t detected and results are influenced by a single document. It might be necessary to consider this hypothesis and, consequently, verify if results of the other choosing methods, like the average result, aren’t accurate enough to help the authorship attribution task with smaller false positives results.

D. Tests with documents from another authors

The second set of documents, from 10 different authors, was submitted to the same tests. The first test was made with the attribution being made for the best result of each method. For simplicity, only the average results are shown in table 10. The correct attribution rate of the SVM method was 88,67 % for this set of documents.

TABLE X. Attribution with Best Results

Compressor	Method	
	CCC	*NCD*
Bzip	79,33 %	92,67 %
PPM-D	91,33 %	97,33 %
Zip	94 %	97 %

For these documents, the NCD method has better results than the CCC method, and also outperformed the SVM method. For the Bzip compressor, the CCC method had a worse correct attribution ratio than the SVM method. And the PPM-D compressor had a slightly better result in the NCD method than the Zip compressor.

Results of the voting system are presented in table 11 and results of the attribution made by the average results are presented in table 12.

TABLE XI. Attribution with Voting System

Compressor	Method	
	CCC	*NCD*
Bzip	80 %	95 %
PPM-D	89,33 %	97,33 %
Zip	93,33 %	97,67 %

TABLE XII. Attribution with Average Results

Compressor	Method	
	CCC	*NCD*
Bzip	80,67 %	94,33 %
PPM-D	93 %	95,67 %
Zip	95,33 %	97 %

For this set of documents the voting system provides equal or better results for the NCD method for all the compressors. The Zip compressor also shows the best results in almost all methods of choice of the authorship. .

E. Tests with documents from all authors

In this test all 20 authors were considered as possible authors, increasing the complexity of the task. The testing procedure was the same of the previous sections, with the same compressors and same methods of choice of the probable author being tested

Results of attribution made with the best result are presented in table 13, attribution made by the voting system are presented in table 14 and results of the attribution made by the average results are presented in table 15. The average result of the SVM classifier is 83,67% of correct attributions.

TABLE XIII. Attribution with Best Results

Compressor	Method	
	CCC	*NCD*
Bzip	81,67 %	94 %
PPM-D	88,83 %	95,83 %
Zip	90,17 %	96,83 %

TABLE XIV. Attribution with Voting System

Compressor	Method	
	CCC	*NCD*
Bzip	80 %	93,83 %
PPM-D	85 %	95,67 %
Zip	87,17 %	97 %

TABLE XV. Attribution with Average Results

Compressor	Method	
	CCC	*NCD*
Bzip	82,33 %	93,33 %
PPM-D	88 %	94 %
Zip	90,67 %	97,17 %

It's possible to observe that, despite the increased number of probable authors, the performance of both methods is not very different of the results from the previous tests, when there were only 10 probable authors.

The NCD method, in special, had very similar results, with an approximate 70% rate of correct attributions, independently of the choosing method of the correct result. In all cases, the Zip compressor with the NCD method had the better overall performance.

VI. Conclusion

The authorship attribution task can be benefited from researches of new techniques of automation of this task or from tools that help the expert giving objective results of probable authors.

One viable technique is the use of data compressors to verify how similar are documents from the same author. Based on the theory of information complexity, the methods analyzed presented good results with a correct authorship attribution rate superior to that obtained in previous work. According to Stamatatos [1], it's important to have database of documents that can be used in different tests, making results of various tests comparable. The amount of possible authors and available documents are reasonable to permit a relevant number of tests and results.

The NCD method with the Zip compressor presented the best results in almost all tests, with the PPM-D compressor being responsible for some good results too. This independence of the compressor might be interesting, especially because different compressors have different requirements of available memory and computational capability.

The use of compressor-based methods also presented the advantage of not requiring a previous training of statistical models. When using classifiers like SVM, the characteristics that will be considered have to be chosen previously and a statistical model have to be generated from the training documents, creating a profile that will be used afterwards in the classification task. The NCD and CCC methods, on the other hand, only require that the tested document is compressed with the training documents, and no previous selection of characteristics is necessary. Each compressor, according to its technology, will analyze the document to generate the best possible model for data compression, and by doing such activity it'll be creating a statistic model of the document, and this is used by the mentioned methods to calculate how similar two documents are.

As the result of all the methods of choosing the probable author were similar, it's important to ponder if some method is more susceptible to the presence of badly chosen documents, that could influence the result because it ~~would~~ could contain words that are common to the subject of the document, and not for the author stylistic.

In future works more test should be done with more documents and more possible authors, inclusive separating documents about common themes to verify whether the subject of the document is relevant or leads to confusion in the attribution.

Acknowledgements

To be done in the final version, after acceptance.

References

[1] Stamatatos, E. (2009) "A survey of modern authorship attribution methods" In Journal of the American Society for Information Science and Technology, Volume 60, Issue 3, pp. 538–556

[2] Pavelec, D. F. (2007) "Identificação da Autoria de Documentos: Análise Estilométrica da Língua Portuguesa usando SVM". Dissertação de Mestrado, Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná, Brasil.

[3] Kolmogorov, A. N. (1965) "Three approaches to the quantitative definition of information". In Problems Information Transmission, 1, pp. 1–7

[4] Li, M. e Vitányi, P. M. B. (1997) "An Introduction to Kolmogorov Complexity and Its Applications", Springer, 2nd edition.

[5] Li, M., Chen, X., Li, X., Ma, B. and Vitányi, P.M.B. "The similarity metric", In IEEE Trans. Inform. Theory 50 (12) pp. 3250–3264, 2004.

[6] Malyutov, M.B., Wickramasinghe, C.I. and Li, S. (2007) "Conditional Complexity of Compression for Authorship Attribution", In SFB 649 Discussion Paper No. 57, Humboldt University, Berlin.