Authorship Attribution of Documents Using Data Compression as a Classifier

W. R. Oliveira Jr, E. J. R. Justino and L. E. S. Oliveira

Abstract— **Automatic attribution of text subject and even authorship attribution is possible with the use of classifiers. A possible approach is the use of the Normalized Compression Distance. This approach uses previously classified documents as a paradigm but doesn't require previous training. Tests were performed in a dataset made of 3,000 documents from online newspapers and blogs, from 100 authors, in 10 distinct subjects. The correct classification rate was 79,61%on average and the areas which had better classification were Sports (96,96% correct categorization) and Technology (91,74%).**

Index Terms— **authorship attribution, data compression, document classification.**

I. Introduction

Data classification is an important task and automatizing such task with correct attribution to previously specified categories can improve the user experience, for example for information retrieval or content-providing.

There are many methods to categorize documents. Many researches have used Bayesian statistical methods [1] and Support Vector Machine [2]. Those methods achieve good results but depend on previous training or previous selection of relevant words.

One appealing alternative is the use of the Normalized Compression Distance (NCD) to perform such task. The NCD is computed based on the archive size after compression and indicates how similar two documents are. In this method no previous characteristic selection or model training is made and very little effort need to be done to change the classification model.

NCD has being used before to classify different kinds of information. For example, music [3], protein sequence [4] and even fetal heart rate [5].

To analyze the performance of this measure to classify documents into predefined categories, a dataset with 3,000 documents, from 100 authors and 10 categories was used. Tests were performed and the results were analyzed, especially to verify which categories showed most confusion.

This document is divided as follow: the first section is this introduction, the second section explains the NCD, the third section details the dataset used and the forth section explains the method. The fifth section provides the results and analyzes them and the sixth section gives the conclusions.

II. Normalized Compression Distance

The NCD was proposed by [6] and is based in the Kolmogorov theory of information complexity. According to this theory, the complexity of information can be measured by the amount of symbols required to represent it in some specific language. For example, the Kolmogorov complexity *K(x)* of the information *x* is the amount of symbols required to represent it. Since it's not possible to determine if *K(x)* is actually the smaller representation of *x*, the Kolmogorov complexity *K(x)* is incomputable, and so are its lower and upper bounds [7].

It's also possible to define the Kolmogorov conditional complexity of some information. The Kolmogorov conditional complexity *K(x|y)* is the smallest program *y* that outputs the result *x* when processed through a universal fixed Turing machine. This conditional complexity is also incomputable.

Reference [8] proposed that when two documents have a small Kolmogorov conditional complexity, they are more similar. This similarity is called Normalized Information Distance (NID) and is defined as

$$NID(x,y) = \frac{\max\{K(x \mid y), K(y \mid x)\}}{\max\{K(x), K(y)\}} \quad (1)$$

where *K(x|y)* is the Kolmogorov conditional complexity of document *x*, given a document *y, K(x)* is the Kolmogorov complexity of document *x* and *max{x,y}* is the function that returns the biggest of two values. But NID is also incomputable.

According to [6], it's possible to approximate the Kolmogorov conditional complexity with the use of compressor. Given that a compressor can have a input **y** and act like a Turing machine, processing that input and then outputting *x*, it can be used to approximate *K(x|y)* or *K(x)*. Thus, using a compressor *C* to calculate *C(x)*, the equation (1) would become

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

where *xy* is the concatenation of the documents *x* and *y* and *C(x)* is the size in bytes of the compression of the document *x*.

The NCD is a normalized distance and its result will be a value in the [0,1] range, where values close to 0 indicate a similarity among the documents and a value close to 1

Manuscript received July 23, 2012; revised August 14, 2012. This research has been partly supported by The National Council for Scientific and Technological Development (CNPq) grant 301653/2011-9.

W. R. Oliveira Jr. is with the Pontificia Universidade Católica do Paraná – Brazil (corresponding author to provide phone: +55-41-3206-3586; e-mail: w.oliveira.jr@gmail.com).

E. J. R. Justino is with Pontificia Universidade Católica do Paraná – Brazil (e-mail: edson.justion@pucpr.br).

L. E. S. Oliveira is with the Universidade Federal do Paraná, Brazil, (e-mail: lesoliveira@inf.ufpr.br).

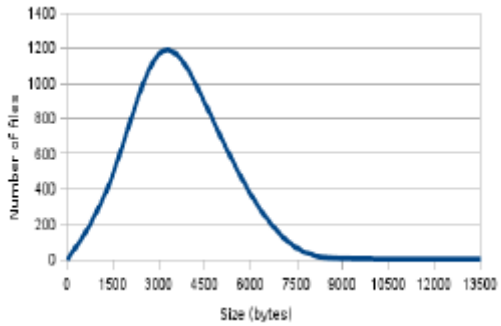
indicate that the documents have very few similarity.

There's no restriction about the compressors that can be used. With text documents in general just lossless compressors are used. Tests were made using the Zip, Bzip and PPM-D data compressors. The Zip compressor is a dictionary-based compressor, widely used to compress documents [9]. In a brief explanation, data is compressed by verifying information that is repeated along the document and then representing it with a reference to the previously seem information, instead of repeating it again. The Bzip compressor uses the Burrows-Wheeler algorithm, with compression being done in blocks, and the PPM-D (prediction by partial match – escape D) is a statistical compressor.

III. Database

The dataset used was composed of documents extracted from online newspapers and blogs, all written in Brazilian Portuguese. There were 3,000 documents, from 100 different authors, categorized in 10 categories according to the subject: Unspecified Subject, Law, Economy, Sports, Gastronomy, Literature, Politics, Health, Technology and Tourism. Those categories were chosen because they are most used in daily newspapers, with an "Unspecified Subject" left to documents that wouldn't fit in other category. This dataset has being used elsewhere in authorship attribution works [10].

The documents have an average size of 2989 bytes, with a 1531 standard deviation. Each document has, in average, 486 tokens and 286 Hapax (words occurring just once). The figure 1 illustrates the size distribution of the documents.



Documents were separated in two groups. The first group was the paradigm set and was composed with 7 documents of each author (70 documents for each category), in a total of 700 documents. The second group was the testing set and was composed with the 2,300 reminding documents.

Since the NCD doesn't allow specifying which words will be considered to the classification task, all documents were processed so that information that was particular to the author of each document was removed For example the author name or the author email was removed from the documents, and only content without direct author indication was considered.

IV. Method

Each document of the testing set was tested according to the following pseudocode:

```
For each document x in the testing set (
    For each document y in the paradigm set (
        Calculate the NCD(x,y)
        Verify which y document gives the smaller NCD
        Attribute the document x to the author of the y
            document
    )
)
```

So, each tested document had 7 documents of the correct author as an example and 693 documents of different authors.

V. Results

After proceeding with the method described in section 4, the results presented in Table 1 were obtained for the Zip compressor.

Table 1
Correct Categorizations

Category	% correct categorizations
Unspecified Subject	83,04%
Law	65,65%
Economy	79,57%
Sports	87,39%
Gastronomy	53,04%
Literature	61,74%
Politics	83,04%
Health	63,91%
Technology	79,13%
Tourism	83,04%
Average	73,96%

The average of correct authorship attribution was 73,96%. Some categories (sports, politics and tourism) had a performance above 80% of correct categorizations, where the correct author was identified. This could indicate that the information in such documents have a very specific writing style or very specific words, allowing the compressor to use information seem in the first document (the paradigm one) to better compress the second (tested) document.

On the other hand, some categories had a bad performance, with Literature and Gastronomy having less than 65% of correct authorship attribution.

For the Bzip compressor, results are presented in Table 2.

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 2
Correct Categorizations

Category	% correct categorizations
Unspecified Subject	79,57%
Law	63,91%
Economy	77,83%
Sports	82,61%
Gastronomy	44,78%
Literature	59,13%
Politics	81,74%
Health	58,26%
Technology	74,78%
Tourism	80,00%
Average	70,26%

And, for the PPM-D compressor, results are presented in Table 3.

Table 3
Correct Categorizations

Category	% correct categorizations
Unspecified Subject	81,74%
Law	68,26%
Economy	78,70%
Sports	85,65%
Gastronomy	54,35%
Literature	66,96%
Politics	83,91%
Health	61,30%
Technology	77,39%
Tourism	82,17%
Average	74,04%

It's possible to observe that the PPM-D had a better average result than the others compressors. Statistical compressors of the PPM family are known for better compressing texts and that fact might have helped the performance of this compressor in this task.

The Bzip compressor had the worst result, with an average of 70% of the documents being correctly attributed. In the Gastronomy category the result was below 50%.

It's important to note that there were always 100 possible authors, so the attribution of one document, if made randomly, would have only 1% chance of being correctly attributed.

To verify how the authorship attribution was influenced by the presence of other categories, a confusion matrix was elaborated and analyzed. It is presented in Table 4. In these results, only confusion among categories is considered. The results are a bit better than the ones shown in the previous tables because, in Table 4, the attribution to a wrong author, from the correct category, is considered correct.

It's possible to verify that documents from Gastronomy were confused with Health, Law and Economy. Analyzing the documents contents, it was possible to verify that many documents have "healthy advices" or "healthy recipes", and that might have caused 18 documents (8,26% of the Gastronomy documents) to be categorized as Health documents.

Documents from Literature were mainly confused with documents of Law (22 documents). This was mainly due to one author of Literature having a writing style very similar to documents that analyzed Law books.

It's also possible to analyze that the use of NCD prevented confusion between Gastronomy and Politics, since no document from one category was attributed to the other. Some categories also showed a very little confusion. For example, only one document from Technology was categorized as Unspecified Subjects, and only one document from Unspecified Subjects was categorized to Technology. The same can be said about other categories, like Sports and Economy or Sports and Politics.

Table 4
Confusion Matrix

	Unspe-cified Subject	Law	Economy	Sports	Gastronomy	Literature	Politics	Health	Technology	Tourism
Unspecified Subject	86.96	1.74	3.04	1.30	0.00	3.91	1.30	0.87	0.43	0.43
Law	5.22	83.48	3.91	0.87	2.17	0.87	0.43	0.43	1.74	0.87
Economy	4.78	7.83	74.35	0.00	1.30	0.00	6.09	2.17	2.17	1.30
Sports	0.43	0.43	0.43	96.96	0.43	0.00	0.43	0.43	0.00	0.43
Gastronomy	1.74	6.96	6.96	2.17	65.65	2.17	0.00	8.26	4.35	1.74
Literature	6.09	9.57	4.35	2.17	3.04	57.39	4.35	4.78	6.09	2.17
Politics	2.17	0.87	7.39	0.00	0.00	0.43	83.91	0.87	4.35	0.00
Health	1.74	4.35	6.96	0.87	2.17	0.00	0.00	81.74	2.17	0.00
Technology	0.43	1.30	3.48	0.87	0.00	0.43	0.00	0.87	91.74	0.87
Tourism	3.48	2.17	7.39	1.74	0.43	0.87	2.17	3.48	4.35	73.91

ISBN: 978-988-19251-6-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VI. Conclusion

The use of data compressors to categorize documents can present some advantages. There is no need to preselect which characteristics will be considered to classify the documents, since the classification is based on the similarity of those documents, measured by a normalized distance. The paradigm documents can also be changed at every evaluation, without the need of special characteristics extraction or previous training.

The use of a large dataset, with more than 2,000 documents being classified, provides a reliable result. The dataset will be made available at http://www.woliveirajr.com so that other works can compare results.

The NCD was sensible enough to classify some Gastronomy documents in the Health category, and that is justified to the fact that the content of the documents were about healthy alimentation. Results showed that the original classification of categories (as identified by the newspapers) can't be considered so strict, and some documents belong to more than one category.

In future work other compressors can be tested to verify if the authorship attribution with NCD can have better results with compressors that have a better compression ratio for text documents.

Acknowledgment

W. R. Oliveira Jr. thanks the Pontifícia Universidade Católica do Paraná – Brazil for the kind support during the research.

References

[1] M. Iwayama and T. Takenobu. Hierarquical Bayesian Clustering for Automatic Text Classification. Natural Language, pp. 1322-1327, 1995.

[2] T. Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. Machine Learning: ECML-98. Lecture Notes in Computer Science, Volume 1398/1998, pp. 137-142, 1998.

[3] R. Cilibrasi, P. M. B. Vitanyi, and R. Wolf. Algorithmic clustering of music based on string compression, Computer Music Journal, vol. 28, no. 4, pp. 49–67, 2004.

[4] A. Kocsor, A. Kertész-Farkas, L. Kaján and S. Pongor. Application of compression-based distance measures to protein sequence classification: A methodology study. Bioinformatics, vol. 22, no. 4, pp. 407–412, 2006.

[5] C. C. Santos, J. Bernardes, P. M. B. Vitanyi and L. Antunes. Clustering fetal heart rate tracings by compression. in Proc. 19th IEEE Symp. Comput.-Based Med. Syst., pp. 685–690, 2006

[6] R. Cilibrasi and P. M. B. Vitányi. Clustering by compression, IEEE Trans. Inform. Theory 51 (4) pp.1523–1545, 2005

[7] M. Li and P. M. B. Vitanyi. An Introduction to Kolmogorov Complexity and Its Applications. Springer. New York, , 2nd edition, 1997.

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

[9] K. Sayood. Introduction to Data Compression. Academic Press. New York, 2nd edition, 2000.

[10] P. J. Varela. O uso de atributos estilométricos na identificação da autoria de textos. Master Degree Dissertation, PUC-PR, Brazil, 2010.