# Authorship Analysis based on Data Compression

Daniele Cerra, Mihai Datcu, and Peter Reinartz

*German Aerospace Center (DLR), Muenchner str. 20, 82234 Wessling, Germany*
*Corresponding author's email: daniele.cerra@dlr.de,*
*phone: +49 8153 28-1496, fax: +49 8153 28-1444.*

## Abstract

This paper proposes to perform authorship analysis using the Fast Compression Distance (FCD), a similarity measure based on compression with dictionaries directly extracted from the written texts. The FCD computes a similarity between two documents through an effective binary search on the intersection set between the two related dictionaries. In the reported experiments the proposed method is applied to documents which are heterogeneous in style, written in five different languages and coming from different historical periods. Results are comparable to the state of the art and outperform traditional compression-based methods.

*Keywords:* Authorship Analysis, Data Compression, Similarity Measure

## 1. Introduction

The task of automatically recognizing the author of a given text finds several uses in practical applications, ranging from authorship attribution to plagiarism detection, and it is a challenging one (Stamatatos, 2009). While the structure of a document can be easily interpreted by a machine, the description of the style of each author is in general subjective, and therefore hard to derive in natural language; it is even harder to find a description which enables a machine to automatically tell one author from the other. A literature review on modern authorship attribution methods, usually coming from the fields of machine learning and statistical analysis, is reported in Stamatatos (2009); Jockers and Witten (2010); Koppel et al. (2009); Grieve (2007); Juola (2006). Among these, algorithms based on similarity measures such as Benedetto et al. (2002) and Koppel et al. (2011) are widely employed and usually assign an anonymous text to the author of the most similar document in the training data.

During the last decade, compression-based distance measures have been effectively applied to cluster texts written by different authors (Cilibrasi and Vitányi, 2005) and to perform plagiarism detection (Chen et al., 2004). Such universal similarity measures, of which the most well-known is the Normalized Compression Distance (NCD), employ general compressors to estimate the amount of shared information between two objects. Similar concepts are also used by methods using runlength histograms to retrieve and classify documents (Gordo et al., 2013). Experiments carried out in Oliveira et al. (2013) conclude that NCD-based methods for authorship analysis outperform state-of-the-art classification methodologies such as Support Vector Machines. A study on larger and more statistically meaningful datasets shows NCD-methods to be competitive with respect to the state of the art (de Graaff, 2012), while Stamatatos (2009) reports that compression-based methods are effective but hard to use in practice as they are very slow.

Indeed the universality of these measures comes at a price, as the compression algorithm must be run at least $n^2$ times on $n$ objects to derive a distance matrix, slowing down the analysis. Furthermore, as these methods are applied to raw data they cannot be tuned to increase their performance on a given data type. We propose then to perform these tasks using the Fast Compression Distance (FCD) recently defined in Cerra and Datcu (2012), which provides superior performances with a reduced computational complexity with respect to the NCD, and can be tuned according to the kind of data at hand. In the case of natural texts, only FCD's general settings should be adjusted according to the language of the dataset, thus keeping the desirable parameter-free approach typical of NCD. Applications to authorship and plagiarism analysis are derived by extracting meaningful dictionaries directly from the strings representing the data instances and matching them. The reported experiments show that improvements over traditional compression-based analysis can be dramatic, and that the FCD could become an important tool of easy usage for the automated analysis of texts, as satisfactory results are achieved skipping any parameters setting step. The only exception is an optional text preprocessing step which only needs to be set once for documents of a given language, and does not depend on the specific dataset.

The paper is structured as follows. Section 2 introduces compression-based similarity measures and the FCD, which will be validated in an array of experiments reported in Section 3. We conclude in Section 4.

## 2. Fast Compression Distance

Compression-based similarity measures exploit general off-the-shelf compressors to estimate the amount of information shared by any two objects. They have been employed for clustering and classification on diverse data types such as texts and images (Watanabe et al., 2002), with Keogh et al. (2004) reporting that they outperform general distance measures. The most widely known and used of such notions is the Normalized Compression Distance (NCD), defined for any two objects $x$ and $y$ as:

$$NCD(x,y) = \frac{C(x,y) - \min C(x), C(y)}{\max C(x), C(y)} \tag{1}$$

where $C(x)$ represents the size of $x$ after being compressed by a compressor (such as Gzip), and $C(x, y)$ is the size of the compressed version of $x$ appended to $y$. If $x = y$, the NCD is approximately 0, as the full string $y$ can be described in terms of previous strings found in $x$; if $x$ and $y$ share no common information the NCD is $1 + e$, where $e$ is a small quantity (usually $e < 0.1$) due to imperfections characterizing real compressors. The idea is that if $x$ and $y$ share common information they will compress better together than separately, as the compressor will be able to reuse recurring patterns found in one of them to more efficiently compress the other. The generality of NCD allows applying it to diverse datatypes, including natural texts. Applications to authorship categorization have been presented by

Cilibrasi and Vitányi (2005), while plagiarism detection of students assignments has been succesfully carried out by Chen et al. (2004).

A modified version of NCD based on the extraction of dictionaries has been first defined by Macedonas et al. (2008). The advantages of using dictionary-based methods have been then studied by Cerra and Datcu (2012), in which the authors define a Fast Compression Distance (FCD), and successfully apply it to image analysis. The algorithm can be used for texts analysis as follows.

First of all, all special characters such as punctuation marks are removed from a string $x$, which is subsequently tokenized in a set of words $W_x$. The sequence of tokens is analysed by the encoding algorithm of the Lempel-Ziv-Welch (LZW) compressor (Welch, 1984), with the difference that words rather than characters are taken into account. The algorithm initializes the dictionary $D(x)$ with all the words $W_x$. Then the string $x$ is scanned for successively longer sequences of words in $D(x)$ until a mismatch in $D(x)$ takes place; at this point the code for the longest pattern $p$ in the dictionary is sent to output, and the new string ($p$ + the last word which caused a mismatch) is added to $D(x)$. The last input word is then used as the next starting point: in this way, successively longer sequences of words are registered in the dictionary and made available for subsequent encoding, with no repeated entries in $D(x)$. An example for the encoding of the string "TO BE OR NOT TO BE OR NOT TO BE OR WHAT" after tokenization is reported in Table 1. It helps to remark that the output of the simulated compression

Table 1: LZW encoding of the tokens composing the string ”TO BE OR NOT TO BE OR NOT TO BE OR WHAT”. The compressor tries to substitute pattern codes referring to sequences of words which occurred previously in the text.

| Current token | Next token | Output | Added to Dictionary |
|---|---|---|---|
| Null | TO | | |
| TO | BE | $TO$ | TO BE=$< 1 >$ |
| BE | OR | $BE$ | BE OR=$< 2 >$ |
| OR | NOT | $OR$ | OR NOT=$< 3 >$ |
| NOT | TO | $NOT$ | NOT TO=$< 4 >$ |
| TO BE | OR | $< 1 >$ | TO BE OR=$< 5 >$ |
| OR NOT | TO | $< 3 >$ | OR NOT TO=$< 6 >$ |
| TO BE OR | WHAT | $< 5 >$ | TO BE OR WHAT=$< 7 >$ |
| WHAT | $\sharp$ | $WHAT$ | |

process is not of interest for us, as the only thing that will be used is the dictionary.

The patterns contained in the dictionary $D(x)$ are then sorted in ascending alphabetical order to enable the binary search of each pattern in time $O(logN)$, where $N$ is the number of entries in $D(x)$. The dictionary is finally stored for future use: this procedure may be carried out offline and has to be performed only once for each data instance. Whenever a string $x$ is checked against a database containing $n$ dictionaries, a dictionary $D(x)$ is extracted from $x$ as described and matched against each of the $n$ dictionaries. The FCD between $x$ and an object $y$ represented by $D(y)$ is defined as:

$$FCD(x, y) = \frac{|D(x)| - \cap(D(x), D(y))}{|D(x)|} \tag{2}$$

where $|D(x)|$ and $|D(y)|$ are the sizes of the relative dictionaries, represented by the number of entries they contain, and $\cap(D(x), D(y))$ is the number of patterns which are found in both dictionaries. We have $FCD(x, y) = 0$ iff all patterns in $D(x)$ are contained also in $D(y)$, and $FCD(x, y) = 1$ if no single pattern is shared between the two objects.

The FCD allows computing a compression-based distance between two objects in a faster way with respect to NCD (up to one order of magnitude), as the dictionary for each object must be extracted only once and computing the intersection between two dictionaries $D(x)$ and $D(y)$ is faster than compressing the concatenation of $x$ appended to $y$ (Cerra and Datcu, 2012). The FCD is also more accurate, as it overcomes drawbacks such as the limited size of the lookup tables, which are employed by real compressors for efficiency constraints: this allows exploiting all the patterns contained in a string. Furthermore, while the NCD is totally data-driven, the FCD enables a token-based analysis which allows preprocessing the data, by decomposing the objects into fragments which are semantically relevant for a given data type or application. This constitutes a great advantage in the case of plain texts, as the direct analysis of words contained in a document and their concatenations allows focusing on the relevant informational content. In plain English, this means that the matching of substrings in words which may have no semantic relation between them (e.g. 'butter' and 'butterfly') is prevented. Additional improvements can be made depending on the texts language. For the case of English texts, the subfix 's' can be removed from each token, while from documents in Italian it helps to remove the last vowel from each word: this avoids considering semantically different plurals and some verbal forms.

A drawback of the proposed method is that it cannot be applied effectively to very short texts. The algorithm needs to find reoccurring word sequences in order to extract dictionaries of a relevant size, which are needed in order to find patterns shared with other dictionaries. Therefore, the compression of the initial part of a string is not effective: we estimated empirically 1000 tokens or words to be a reasonable size for learning the model of a document and to be effective in its compression.

## 3. Experimental Results

The FCD as described in the previous section can be effectively employed in tasks like authorship and plagiarism analysis. We report in this section experiments on four datasets written in English, Italian, and German.

### *3.1. The Federalist Papers*

We consider a dataset of English texts known as Federalist Papers, a collection of 85 political articles written by Alexander Hamilton, James Madison and John Jay, published in 1787-88 under the anonymous pseudonym 'Publius'. This corpus is particularly interesting, as Hamilton and Madison claimed later the authorship of their texts, but a number of essays (the ones numbered 49-58 and 62-63) have been claimed by both of them. This is a classical dataset employed in the early days of authorship attribution literature, as the candidate authors are well-defined and the texts are uniform

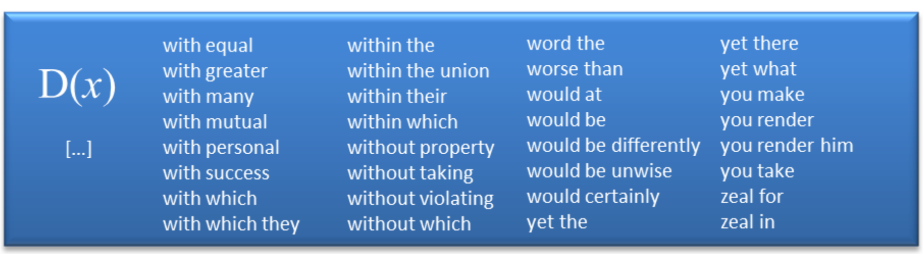


Figure 1: Subset from a dictionary $D(x)$ extracted from a sample text $x$ belonging to the Federalist dataset.

in thematics (Stamatatos, 2009). Several studies agreed on assigning the disputed works in their entirety to Madison, while Papers 18-20 have generally been found to be written jointly by Hamilton and Madison as Hamilton claimed, even though some researchers tend to attribute them to Madison alone (Jockers and Witten, 2010; Meyerson, 2008; Adair, 1974).

We analyzed a dataset composed of a randomly selected number of texts of certain attribution by Hamilton and Madison, plus all the disputed and jointly written essays. We then computed a distance matrix related to the described dataset according to the FCD distance, and performed on the matrix a hierarchical clustering which is by definition unsupervised. A dendrogram (binary tree) is heuristically derived to represent the distance matrix in 2 dimensions through the application of genetic algorithms (Cilibrasi, 2007; Cilibrasi and Vitányi, 2005). Results are reported in Fig. 2, and have been obtained using the freely available tool CompLearn available at Cilibrasi et al. (2002). Each leaf represents a text, with the documents which behave more similarly in terms of distances from all the others appearing as siblings. The evaluation is done by visually inspecting if texts written by the same authors are correctly clustered in some branch of the tree, i.e. by checking how well the texts by the two authors can be isolated by 'cutting' the tree at a convenient point. The clustering agrees with the general interpretation of the texts: all the disputed texts are clearly placed in the section of the tree containing Madison's works. Furthermore, the three jointly written works are clustered together and placed exactly between Hamilton and Madison's essays. We compare results with the hierarchical clustering derived from the distance matrix obtained on the basis of NCD distances (Fig. 3), run with the default blocksort compression algorithm provided by CompLearn: in this case the misplacements of the documents is evident, as disputed works are in general closer to Madison texts but are scattered throughout the tree.

### *3.2. The Liber Liber dataset*

The rise of interest in compression-based methods is in part due to the concept of relative entropy as described in Benedetto et al. (2002), which quantifies a distance between two isolated strings relying on information theoretical notions. In this work the authors succesfully perform clustering and classification of documents: one of the considered problems is to automatically recognize the authors of a collection comprising 90 texts of 11 known Italian authors spanning the centuries XIII-XX, available at Onlus (2003). Each text $x$ was used as a query against the rest of the database, its closest object $y$ minimizing the relative entropy $D(x, y)$ was retrieved, and $x$ was then assigned to the author of $y$. In the following experiment the same procedure as Benedetto et al. (2002) and a dataset as close as possible have been adopted, with each text $x$ assigned to the author of the text $y$ which minimizes $FCD(x, y)$. We compare our results with the ones obtained by the Common N-grams (CNG) method proposed by Kešelj et al. (2003) using the most relevant 500, 1000 and 1500 3-grams in Table 2. The FCD finds the correct author in 97.8% of the cases, while the best n-grams setting

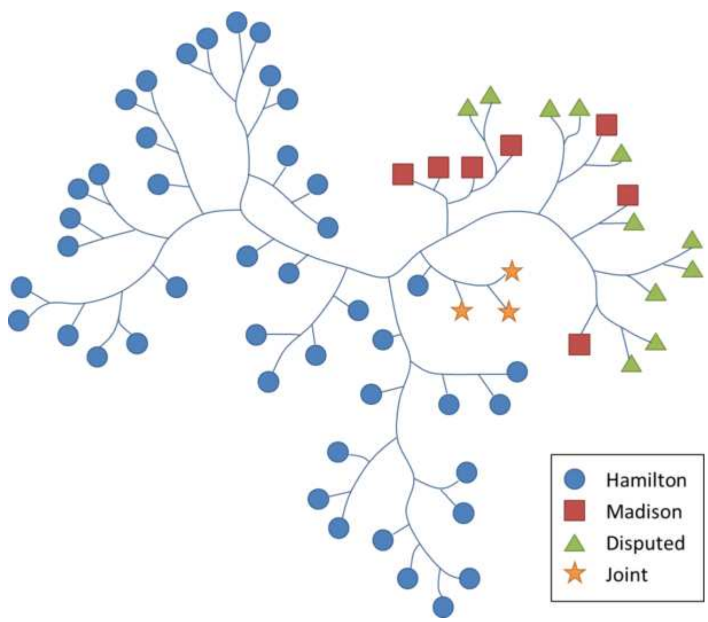


Figure 2: Hierarchical clustering of the Federalist dataset, derived by a full distance matrix obtained on the basis of the FCD distance.

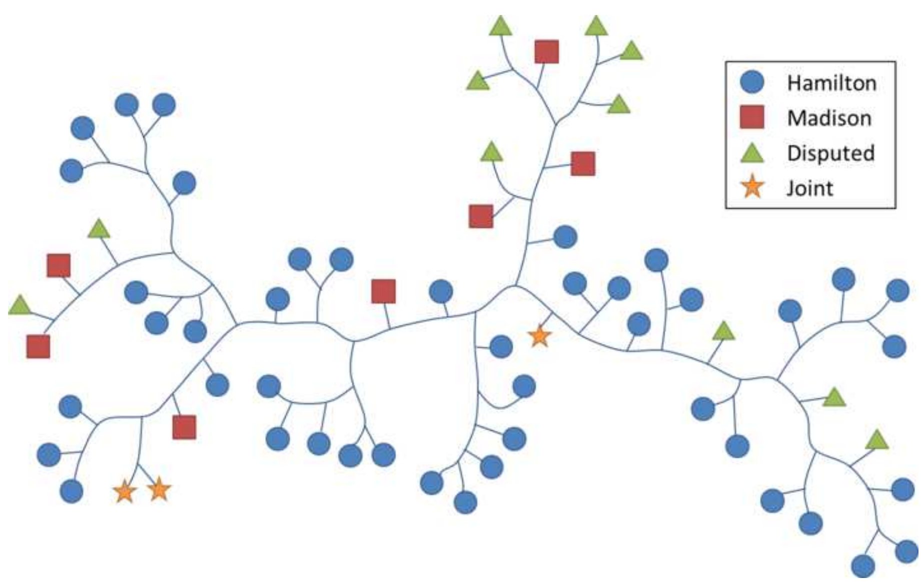


Figure 3: Hierarchical clustering of the Federalist dataset obtained on the basis of the NCD distance.

yields an accuracy of 90%. For FCD only two texts, *L'Asino* and *Discorsi sopra la prima deca di Tito Livio*, both by Niccoló Machiavelli, are incorrectly assigned respectively to Dante and Guicciardini, but these errors may be justified: the former is a poem strongly influenced by Dante (Caesar, 1989), while the latter was found similar to a collection of critical notes on the very *Discorsi* compiled by Guicciardini, who was Machiavelli's friend (Machiavelli et al., 2002). The N-grams-based method also assigns incorrectly Guicciardini's notes and a Dante's poem to Machiavelli, among others misclassifications.

We also compared our results with an array of other compression-based similarity measures (Table 3): our results outperform both the Ziv-Merhav distance (Pereira Coutinho and Figueiredo, 2005) and the relative entropy as described in Benedetto et al. (2002), while the algorithmic Kullback-Leibler divergence (Cerra and Datcu, 2011) obtains the same results in a considerably higher running time. Accuracy for the NCD method using an array of linear compressors ranged from the 93.3% obtained using the bzip2 compressor to the 96.6% obtained with the blocksort compressor. Even though accuracies are comparable and the dataset may be small to be statistically meaningful, another advantage of FCD over NCD is the decrease in computational complexity. While for NCD it took 202 seconds to build a distance matrix for the 90 pre-formatted texts using the zlib compressor (with no appreciable variation when using other compressors), just 35 seconds were needed on the same machine for the FCD: 10 to extract the dictionaries and

Table 2: Classification results on the Liber Liber dataset. Each text from the 11 authors is used to query the database, and it is considered to be written by the author of the most similar retrieved work. The authors' full names: Dante Alighieri, Gabriele D'Annunzio, Grazia Deledda, Antonio Fogazzaro, Francesco Guicciardini, Niccoló Machiavelli, Alessandro Manzoni, Luigi Pirandello, Emilio Salgari, Italo Svevo, Giovanni Verga. The CNG method has been tested using the reported amounts of n-grams.

| Author | Texts | FCD | CNG-500 | CNG-1000 | CNG-1500 |
|---|---|---|---|---|---|
| Dante Alighieri | 8 | 8 | 6 | 5 | 7 |
| D'Annunzio | 4 | 4 | 4 | 3 | 4 |
| Deledda | 15 | 15 | 15 | 15 | 14 |
| Fogazzaro | 5 | 5 | 4 | 5 | 5 |
| Guicciardini | 6 | 6 | 5 | 5 | 5 |
| Machiavelli | 12 | 10 | 8 | 10 | 9 |
| Manzoni | 4 | 4 | 4 | 4 | 4 |
| Pirandello | 11 | 11 | 5 | 10 | 8 |
| Salgari | 11 | 11 | 10 | 10 | 9 |
| Svevo | 5 | 5 | 4 | 5 | 5 |
| Verga | 9 | 9 | 6 | 9 | 8 |
| Total | 90 | 88 | 71 | 81 | 78 |
| Accuracy (%) | 100 | 97.8 | 78.9 | 90 | 86.7 |

the rest to build the full distance matrix.

### *3.3. The PAN Benchmark Dataset*

We tested our algorithm on datasets from the two most recent PAN CLEF (2013) competitions, which provide benchmark datasets for authorship attribution. From PAN 2013 we selected the author identification task described in Juola and Stamatatos (2013). In this task 349 training texts are provided, divided in 85 problems out of which 30 are in English, 30 in Greek and 25 in Spanish. For each set of documents written by a single author it must be determined if a questioned document was written by the same author or not. Each text is approximately 1000 words long, which is close to our empirical

Table 3: Accuracy and running time for different compression-based methods applied to the Liber Liber dataset.

| Method | Accuracy (%) | Running Time (sec) |
|---|---|---|
| FCD | 97.8 | 35 |
| Relative Entropy | 95.4 | NA |
| Ziv-Merhav | 95.4 | NA |
| NCD (zlib) | 94.4 | 202 |
| NCD (bzip2) | 93.3 | 198 |
| NCD (blocksort) | 96.7 | 208 |
| Algorithmic KL | 97.8 | 450 |

estimation of the minimum size for FCD to find relevant patterns in a data instance (Section 2). For each problem, we consider an unknown text to be written by the same author of a given set of documents if the average FCD distance to the latter is smaller than the mean distance from all documents of a given language. Compared to the performance of the 18 methods reported in Juola and Stamatatos (2013), the FCD finds the correct solution in 72.9% of the cases and yields the second best results, ranking first for the set of English problems and fifth for both the Greek and Spanish sets (Table 4), outperforming among others two compression-based and several n-grams-based methods. It must be stressed that the FCD took approximately 38 seconds to process the whole dataset, while the imposters method by Seidman (2013), which ranked first in the competition for all problems excluded the ones in Spanish, took more than 18 hours. Furthermore, the latter method requires the setting of a threshold, while the FCD skips this step. On the other hand, the contest participants had only a small subset of the available ground truth to test their algorithms.

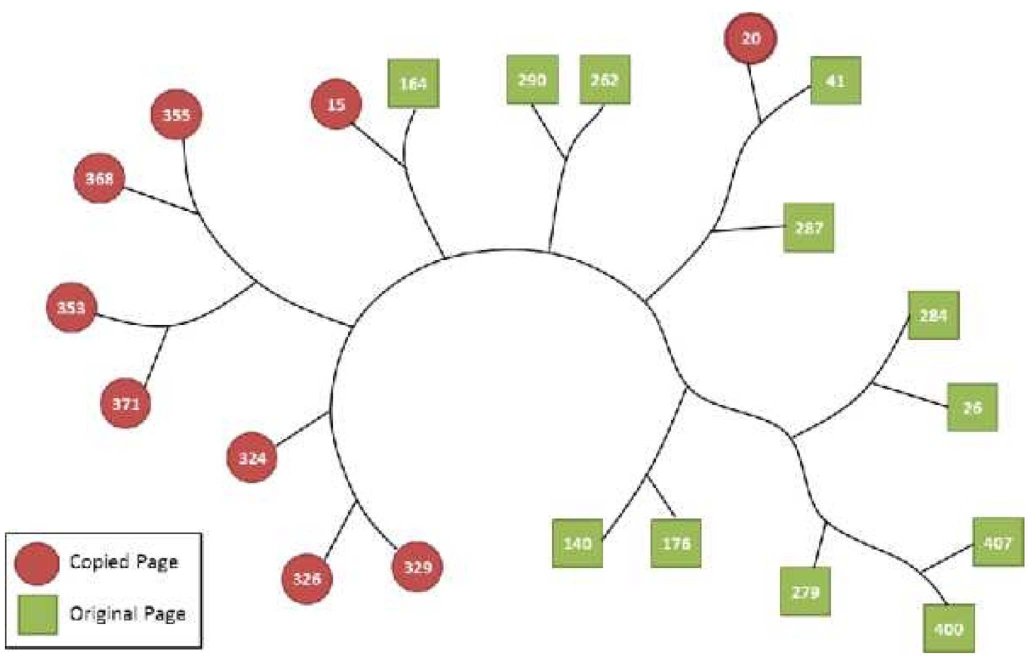


Figure 4: Hierarchical clustering of pages extracted from Guttenberg PhD thesis.

Table 4: Author identification task of the CLEF PAN 2013 dataset. The dataset contains 349 training texts plus 85 test documents of questioned authorship, with problems given in English, Greek and Spanish. The table reports how the FCD ranks compared to 18 participants to the PAN 2013 contest. The first ranked submission for each problem is reported as 'Best PAN'.

| Task | FCD | Best PAN | Rank |
|---|---|---|---|
| Overall | 72.9 % | 75.3 % | 2 |
| English | 83 % | 80 % | 1 |
| Greek | 63 % | 83 % | 5 |
| Spanish | 72 % | 84 % | 5 |

We tested FCD also on the largest closed-class classification problem (task I) from the 2012 PAN competition: open-class problems were not considered as the simple classification algorithm adopted does not allow a rejection class. Using a corpus of 14 test and 28 training texts belonging to 14 different authors, the FCD (using a simple nearest neighbour classification criterion) assigns correctly 12 out of 14 documents to their correct authors. Out of the 25 which took part to the competition, only 4 methods submitted by three groups (Sapkota and Solorio, 2012; Tanguy et al., 2012; Popescu and Grozea, 2012) outperformed our method (all of them with 13 documents correctly recognized). As a comparison, the NCD and trigrams-based CNG (using the most meaningful 1000 trigrams per document, as this setting yields the best results in Table 2) assigned 2 and 9 documents out of 14 to the correct author, respectively. The results in Tables 4 and 5 are encouraging, specially if we consider that the FCD is a general method which is not specific for the described tasks.

Table 5: Classification results on task I of the CLEF PAN 2012 dataset. The dataset contains 28 texts belonging to 14 different authors for training and 14 for testing. The best results obtained in the PAN 2012 contest are reported as 'Best'.

| Method | FCD | NCD | CNG | Best |
|---|---|---|---|---|
| Correct (out of 14) | 12 | 2 | 9 | 13 |

### *3.4. The Guttenberg Case*

In February 2011, evidence was made public that the former German minister Karl-Theodor zu Guttenberg had violated the academic code by copying several passages of his PhD thesis without properly referencing them. This eventually led to Guttenberg losing his PhD title, resigning from being minister, and being nicknamed Baron Cut-and-Paste, Zu Copyberg and Zu Googleberg by the German media (BBC, 2011). Evidence of the plagiarism and a detailed list of the copied sections and of the different sources used by the minister is available at GuttenPlag (2011).

We selected randomly two sets of pages from this controversial dissertation, with the first containing plagiarism instances, and the second material originally written by the ex-minister. Then we performed an unsupervised hierarchical clustering on the distance matrix derived from FCD distances as described in Section 3.1. First attempts made by analyzing single pages failed at separating the original pages in a satisfactory way, as the compressor needs a reasonable amount of data to be able to correctly identify shared patterns between the texts. We selected then two-pages long excerpts from the thesis, with the resulting clustering reported in Fig. 4 showing a good separation of the texts containing plagiarism instances (in red in the picture). The only confusion comes from pages starting at 41 with pages starting at 20, in the bottom-left part of the clustering. This is justified by the fact that page 41 refers to the works of Loewenstein, who happens to be the same author from which part of page 20 was plagiarized (Loewenstein, 1959). Therefore, the system considers page 20 to be similar to the original style of the author at page 41.

Even though the described procedure is not able to detect plagiarism, it can find excerpts in a text which are similar to a given one. If instances of plagiarized text can be identified, objects close to them in the hierarchical clustering will be characterized by a similar style: therefore, this tool could be helpful in identifying texts which are most likely to have been copied from similar sources.

## 4. Conclusions

This paper evaluates the performance of compression-based similarity measures on authorship and plagiarism analysis on natural texts. Instead of the well-known Normalized Compression Distance (NCD), we propose using the dictionary-based Fast Compression Distance (FCD), which decomposes the texts in sets of reoccurring combinations of words captured in a dictionary, which describe the text regularities, and are compared to estimate the shared information between any two documents. The reported experiments show the universality and adaptability of these methods, which can be applied without altering the general workflow to documents written in English, Italian, Greek, Spanish and German. The main advantage of the FCD with respect to traditional compression-based methods, apart from the reduced computational complexity, is that it yields more accurate results. We can justify this with two remarks: firstly, the FCD should be more robust since it performs a word-based analysis, focusing exclusively on meaningful patterns which better capture the information contained in the documents; secondly, the use of a full dictionary allows discarding any limitation that real compressors have concerning the size of buffers and lookup tables employed, being the size of the dictionaries bounded only by the number of relevant patterns contained in the objects. At the same time, the data-driven approach typical of NCD is maintained. This allows keeping an objective, parameter-free workflow for all the problems considered in the applications section, in which promising results are presented on collections of texts in Italian, English, and German.

## References

Adair, D., 1974. Fame and the Founding Fathers. Indianapolis: Liberty Fund.

BBC, 2011. Germany's Guttenberg 'deliberately' plagiarised. URL `http://www.bbc.co.uk/news/world-europe-13310042`

Benedetto, D., Caglioti, E., Loreto, V., 2002. Language trees and zipping. Physical Review Letters 88 (4), 48702.

Caesar, M., 1989. Dante, the critical heritage, 1314-1870. Routledge.

Cerra, D., Datcu, M., 2011. Algorithmic relative complexity. Entropy 13 (4), 902–914.

Cerra, D., Datcu, M., 2012. A fast compression-based similarity measure with applications to content-based image retrieval. Journal of Visual Communication and Image Representation 23 (2), 293 – 302.

Chen, X., Francia, B., Li, M., McKinnon, B., Seker, A., 2004. Shared information and program plagiarism detection. IEEE Transactions on Information Theory 50 (7), 1545–1551.

Cilibrasi, R., 2007. Statistical inference through data compression. Lulu.com Press.

Cilibrasi, R., Cruz, A., de Rooij, S., Keijzer, M., 2002. CompLearn. URL http://www.complearn.org

Cilibrasi, R., Vitányi, P. M. B., 2005. Clustering by compression. IEEE Transactions on Information Theory 51 (4), 1523–1545.

CLEF, 2013. PAN Lab 2012 & 2013: Uncovering Plagiarism, Authorship, and Social Software Misuse.
URL http://pan.webis.de

de Graaff, R., 2012. Authorship attribution using compression distances. Master thesis, Leiden University.
URL http://www.liacs.nl/assets/Bachelorscripties/2012-18RamondeGraaff.pdf

Gordo, A., Perronnin, F., Valveny, E., 2013. Large-scale document image retrieval and classification with runlength histograms and binary embeddings. Pattern Recognition 46 (7), 1898 – 1905.
URL http://www.sciencedirect.com/science/article/pii/S0031320312005304

Grieve, J., 2007. Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing 22 (3), 251–270.
URL http://llc.oxfordjournals.org/cgi/doi/10.1093/llc/fqm020

GuttenPlag, 2011. Collaborative documentation of plagiarism. http://de.guttenplag.wikia.com.

Jockers, M. L., Witten, D. M., 2010. A comparative study of machine learning methods for authorship attribution. Literary and Linguistic Computing 25 (2), 215–223.

Juola, P., Stamatatos, E., 2013. Overview of the Author Identification Task at PAN 2013.
URL http://www.clef-initiative.eu/documents/71612/3095ffc3-376b-40eb-af10-8251c

Juola, P., 2006. Authorship attribution. Found. Trends Inf. Retr. 1 (3), 233–334.
URL http://dx.doi.org/10.1561/1500000005

Keogh, E., Lonardi, S., Ratanamahatana, C., 2004. Towards parameter-free data mining. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 206–215.

Kešelj, V., Peng, F., Cercone, N., Thomas, C., 2003. N-gram-based author profiles for authorship attribution. In: Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING. Vol. 3. pp. 255–264.

Koppel, M., Schler, J., Argamon, S., January 2009. Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. 60 (1), 9–26.
URL http://dx.doi.org/10.1002/asi.v60:1

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

Loewenstein, K., 1959. Verfassungsrecht und verfassungspraxis der vereinigten staaten. Enzyklopaedie der Rechts- und Staatswissenschaft.

Macedonas, A., Besiris, D., Economou, G., Fotopoulos, S., 2008. Dictionary based color image retrieval. Journal of Visual Communication and Image Representation 19 (7), 464–470.

Machiavelli, N., Atkinson, J., Sices, D., 2002. The Sweetness of Power: Machiavelli's Discourses & Guicciardini's Considerations. Northern Illinois University Press.

Meyerson, M., 2008. Liberty's blueprint: how Madison and Hamilton wrote

The Federalist Papers, defined the Constitution, and made democracy safe for the world. Basic Books.

Oliveira, W., Justino, E., Oliveira, L. S., 2013. Comparing compression models for authorship attribution. Forensic Science International 228 (1-3), 100 – 104.
URL http://www.sciencedirect.com/science/article/pii/S0379073813000923

Onlus, L. L., 2003. the Liber Liber dataset. http://www.liberliber.it.

Pereira Coutinho, D., Figueiredo, M., 2005. Information Theoretic Text Classification using the Ziv-Merhav Method. Pattern Recognition and Image Analysis 3523, 355–362.

Popescu, M., Grozea, C., 2012. Kernel methods and string kernels for authorship analysis. In: Forner, P., Karlgren, J., Womser-Hacker, C. (Eds.), CLEF (Online Working Notes/Labs/Workshop).
URL http://dblp.uni-trier.de/db/conf/clef/clef2012w.html

Sapkota, U., Solorio, T., 2012. Sub-profiling by linguistic dimensions to solve the authorship attribution task. In: Forner, P., Karlgren, J., Womser-Hacker, C. (Eds.), CLEF (Online Working Notes/Labs/Workshop).
URL http://dblp.uni-trier.de/db/conf/clef/clef2012w.html

Seidman, U., 2013. Authorship Verification Using the Impostors Method. Notebook for PAN at CLEF 2013
URL http://www.clef-initiative.eu/documents/71612/7a4e6a71-46e9-4bb1-ab66-8ea9c

Stamatatos, E., 2009. A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. 60 (3), 538–556.

Tanguy, L., Sajous, F., Calderone, B., Hathout, N., 2012. Authorship attribution: Using rich linguistic features when training data is scarce. In: Forner, P., Karlgren, J., Womser-Hacker, C. (Eds.), CLEF (Online Working Notes/Labs/Workshop).
URL `http://dblp.uni-trier.de/db/conf/clef/clef2012w.html`

Watanabe, T., Sugawara, K., Sugihara, H., 2002. A new pattern representation scheme using data compression. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5), 579–590.

Welch, T., 1984. Technique for high-performance data compression. IEEE Computer 17 (6), 8–19.