Forensic Science International 228 (2013) 100–104

Contents lists available at SciVerse ScienceDirect

Forensic Science International

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

Comparing compression models for authorship attribution

W. Oliveira Jr.[a], E. Justino[a], L.S. Oliveira[b,*]

[a] Pontifical Catholic University of Parana (PUCPR), R. Imaculada Conceição, 1155 Curitiba, PR, Brazil
[b] Federal University of Paraná (UFPR), R. Rua Cel. Francisco H. dos Santos, 100 Curitiba, PR 81531-990, Brazil

ARTICLE INFO

Article history:
Received 16 April 2012
Received in revised form 12 February 2013
Accepted 13 February 2013
Available online

Keywords:
Compression models
Authorship attribution
Questioned documents

ABSTRACT

In this paper we compare different compression models for authorship attribution. To this end, three different types of compressors, Lempel-Ziv type (GZip), block sorting type (BZip) and statistical type (PPM), along with two different similarity measures were considered in our experiments. Besides, two different attribution methods are analyzed in this paper. Through a series of experiments performed on two different databases, we were able to show that all the compressors behave similarly, but the similarity measures can vary considerably depending on the strategy used for authorship attribution. Our results corroborate with the literature in the sense that compression models are a good alternative for authorship attribution surpassing traditional pattern recognition systems based on classifiers and feature extraction.

1. Introduction

Authorship attribution can be defined as the task of inferring characteristics of a document's author from the textual characteristics of the document itself. The use of electronic documents like e-mails continue to grow exponentially and in spite of the fact that reliable technology is available to trace a particular computer/or IP address where the document has been produced, the fundamental problem is to identify who was behind the keyboard when the document was produced.

With this impressive growth of the information technology there is also a growth in the volume with which lawyers and courts have called upon the expertise of linguists in cases of disputed authorship. Hence, practical applications for author identification have grown in several different areas such as, criminal law (identifying writers of ransom notes and harassment letters), civil law (copyright and estate disputes), and computer security (mining email content).

In general, the problem of authorship attribution can be formulated as a typical classification problem which depends on discriminant features to represent the style of an author. In this context, stylometric features which include various measures of vocabulary richness and lexical repetition based on word frequency distributions, play an important role. As observed by Madigan et al. [15], most of these measures, however, are strongly dependent on the length of the text being studied and, hence, are difficult to apply reliably. Many other types of features have been investigated, including word class frequencies, syntactic analysis, word collocations, grammatical errors, number of words, sentences, clauses, and paragraph lengths [1,8,11,12,20]. Along with these features, several different machine learning models, such as Support Vector Machines [7,19], Neural Networks [26] and Decision Trees [24] have been used for authorship attribution.

In order to avoid defining features, some authors have proposed the use of compression models for authorship attribution [13,16] and several other problems such as text categorization [17], language recognition [2], genome categorization, and clustering [4].

The rationale behind this is that compression algorithms are able to build a model or dictionary of the files they process. Therefore, they can be used to train classifiers on the labeled documents for each class. An unknown document (testing sample) can be assigned to a class by compressing it multiple times, each time using a different class model or dictionary obtained during training. The testing sample is assigned to the class that produced the highest compression rate. Roughly speaking, two documents are deemed close if we can significantly compress one given the information in the other, the idea being that if two pieces are more similar, then we can more succinctly describe one given the other [4]. In order to measure such a similarity, different metrics have been proposed in the literature [2,14,16]

In every authorship identification problem, there is a set of candidate authors, a set of text samples of known authorship covering all the candidate authors (training corpus), and a set of text samples of unknown authorship (test corpus), each one of which should be attributed to a candidate author. According to

* Corresponding author. Tel.: +55 41 33613655.
E-mail address: lesoliveira@inf.ufpr.br (L.S. Oliveira).

0379-0738/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.forsciint.2013.02.025

Stamatatos [20] authorship attribution methods can be distinguished according to whether they treat each training text individually (instance-based) or cumulatively (profile-based). The literature has shown that the most successful approaches reported in the literature follow the profile-based methodology [17].

The advantages of the compression models for authorship attribution are (i) they are extremely easy to apply, (ii) they are parameter-free in that they do not use any feature or background knowledge about the data and (iii) they yields an overall judgment on the document as a whole, rather than discarding information by pre-selecting features while avoiding the messy and rather artificial problem of defining word boundaries [9]. One can see compression models as black boxes whose inner workings are unclear, but in fact they are well grounded in information theory [22]. The compression rate measures the cross-entropy between the training text and the new document, and the new document is assigned to the class whose training text minimizes that cross-entropy [17].

The main contribution of this work lies in comparing three different types of compressors, Lempel-Ziv type (GZip), block sorting type (BZip) and statistical type (PPM) along with two different compression-based similarity measures, NCD (Normalized Compression Distance) [14] and CCC (Conditional Complexity of Compression). These six possible combinations were tested using both instance- and profile-based attribution methods. Unlike the literature that states that the profile-based approach achieves better results, our experiments on two databases show a strong correlation between the attribution method and the similarity measure. Besides, our experimental results also show that the compression algorithms are an interesting alternative for authorship identification comparing favorably to traditional strategies based on feature extraction and classification.

This paper is organized as follows. Section 2 introduces the similarity measures tested in this work. Section 3 describes both instance- and profile-based attribution methods. Section 4 presents both databases used in all experiments. Section 5 reports all the experiments we have performed and also presents some discussion. Finally, Section 6 concludes this work.

2. Compression-based similarity measures

Text categorization, language recognition, DNA classification, and author attribution belong to a class of problems that are intrinsically described by a string of characters. When analyzing a string of characters the main concern is to extract somehow the information it brings. In the case of author attribution, we want to know who is the author.

Recently some researchers demonstrated that the information theory can be useful to extract the information encoded in the strings of characters, more specifically by measuring its entropy. According to Benedetto et al. [2], probably the best definition in this context is the Chaitin–Kolmogorov entropy: the entropy of a string of characters is the length (in bits) of the smallest program which produces the string as output. In fact, it is impossible to find such a program. However, there are algorithms explicitly conceived to approach this theoretical limit. These are the file compressors, which take a file and try to transform it into the shortest possible file. It is clear that this is not the best way to encode the file but it represents a good approximation of it [2].

Based on that, some authors proposed exploiting compression algorithms to define measures of similarity or remoteness between pairs of sequences of characters [2,14,16]. Li et al. [14] used the theory of Kolmogorov conditional complexity to develop the concept of information distance and introduced what they call Normalized Compression Distance (NCD) (Eq. (1))

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

where $C(xy)$ denotes the compressed size of the concatenation of x and y, $C(x)$ denotes the compressed size of x, and $C(y)$ denotes the compressed size of y. The NCD is a nonnegative number in the interval $[0.1 + \varepsilon]$ representing how different the two files are. Smaller numbers represent more similar files. The ε in the upper bound is due to imperfections of the compression techniques, but for most standard compression algorithms one is unlikely to see an ε above 0.1. More details about the theory of NCD is also available in [4].

Another measure used very often in the literature is the Conditional Complexity of Compression (CCC) [2,16]. The CCC of text y given text x is calculated by Eq. (2).

$$\mathrm{CCC}(y|x) = |S_c| - |x_c| \tag{2}$$

where $|x_c|$ is the length of the compressed text x. The concatenated text $S = xy$ is the text starting with x and proceeding to text y without stop. According to Malyutov et al. [16], CCC approximates a more abstract Kolmogorov conditional complexity concept and measures how well the compressor adapts to patterns in the training text for better compressing the questioned text.

Other measures such as Compression Ratio (Eq. (3)) and Relative CCC (Eq. (4)) are also found in the literature. In our experiments, though, NCD and CCC produced much better results and for this reason we report only the results with these two measures.

$$\mathrm{CR} = |x_c| - |x| \tag{3}$$

$$\mathrm{CCC}r(y|x) = \frac{\mathrm{CCC}(y|x)}{|y|} \tag{4}$$

3. Attribution methods

Let C be a set of n candidate authors, L a set of text samples of known authorship covering all the candidate authors (training set), and T a set of text samples of unknown authorship (testing set). The task of an authorship attribution method consists in assigning each T_i to a candidate author from C. There are two different approaches to accomplish this task according to whether they treat each training text individually or cumulatively [20].

The first approach, known as profile-based, concatenates all the training samples per author (references) in just one file and extracts a cumulative representation of that author's style (usually called the author's profile) from this concatenated text. In this approach, the differences between texts of the same writer are disregarded. In the case of compression models, all the documents of a given author C_i are compressed into a single file A_i. During the testing phase, the similarity of a questioned sample t and the author profiles (A_i) are computed. Then, the questioned sample is assigned to the author that maximizes the similarity measure (Eq. (5)).

$$\mathrm{Max\,Rule}(t) = \max_{i=1}^{C}(\mathrm{similarity}(t, A_i)) \tag{5}$$

Fig. 1 depicts a diagram of the profile-based approach used in this work.

The second approach, known as instance-based, requires multiple training text samples per author in order to develop an accurate attribution model. That is, each training text is individually represented as a separate instance of authorial style [20]. The difference from the previous approach is that the reference files are not concatenated. Therefore, to classify a given text t we compute the similarity measure k times, where k is the number of references per author.

As one can notice, in the profile-based approach the references are combined beforehand by concatenating all of them into a single file. In the instance-based approach we need a fusion rule to combine the k similarity measures before using the decision rule described by Eq. (5). This could be performed in several different way, e.g., by voting, averaging, or assuming the maximum similarity. In our experiments we have tried out all these fusion strategies but the best results were always produced by the Max and Majority Voting rules. Fig. 2 shows the diagram of the instance-based approach.

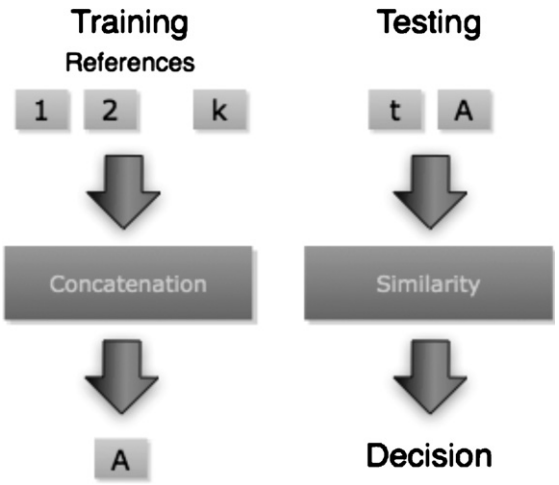

Fig. 1. Diagram of the profile-based approach used in this work.

The main criticism of the instance-based approach is that it is too slow since it has to call the compression algorithm so many times (as many as the training texts). In fact the compression algorithm is called as many times as the candidate authors. Hence, the running time will be significantly lower for the profile-based compression-based method.

4. Database

In this work two different databases were considered. The first one, proposed by Pavelec et al. [19], contains 20 different authors with profiles in Economics (7), Politics (4), Sports (2), Literature (3), Miscellaneous (3), Gossip (1), and Wine (1). The articles were extracted from two different different Brazilian newspapers, *Gazeta do Povo* and *Tribuna do Paraná*. Each author has 30 short articles which usually deal with polemic subjects and express the author's personal opinion. On average, the articles have 691 tokens and 412 hapaxes. The option for short articles was made because in real life forensic experts can count only on short pieces of texts to identify a given writer. Another aspect worth noting is that this kind of article can go through some revision process, which may remove some personal characteristics of the texts. Fig. 3 depicts an example an article in our database.

The second database, proposed by Varela et al. [25] contains 100 different authors whose texts were uniformly distributed over 10 different subjects: Miscellaneous, Law, Economics, Sports, Gastronomy, Literature, Politics, Health, Technology, and Tourism. The sources were 15 Brazilian newspapers located all over the country. Unlike the previous dataset where for some classes we have few authors writing about the same subject, in this database all the subjects have ten different authors. Besides, the articles are shorter, with 486 tokens and 296 hapaxes on average. Therefore, it

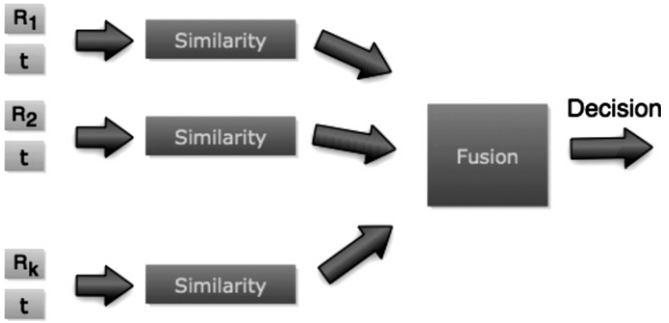

Fig. 2. Diagram of the instance-based approach used in this work.

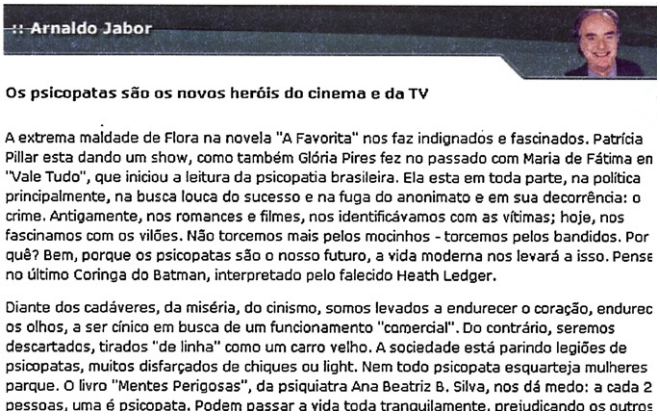

Fig. 3. An example of an article used in this work.

is a more difficult database to perform authorship attribution. Table 1 summarizes both databases.

5. Experiments and discussion

As stated before, three different types of compressors were used in this work, Lempel-Ziv, block sorting, and statistical. The Lempel-Zip [27] is a dictionary encoding technique that attempts to replace a string of symbols with a reference to a dictionary location for the same string. In this work, GZip represent this class of compressors. The block sorting is represented by the BZip algorithm [3]. It compresses data in blocks of size between 100 and 900 Kb using a Burrows–Wheeler transform (BTW) block sorting text compression algorithm and Huffman coding. Compression is generally considerably better than that achieved by more conventional Lempel-Ziv based compressors, and approaches the performance of the PPM family of statistical compressors, which are our third class of compressors. The PPM (Prediction by Partial Matching) [18] is the algorithm used in this type of compressor where the compression is achieved by context modeling and prediction.

5.1. Experiments on Database I

Following the same protocol proposed by Pavelec et al. in [19], the documents of the 20 authors were randomly divided into training (5 documents) and testing (15 documents). All the experiments were performed three times. The recognition rate is the average of these three runs.

Our first strand of experiments was designed to compare the three different compressors and the two similarity measures using the instance-based approach. As discussed before, in this approach each testing document is compressed with an instance of the reference documents (training) to compute the similarity. Then, as depicted in Fig. 2 all these similarities are combined into a fusion rule, which produces a final decision. Table 2 summarizes the results for the instance-based approach.

Table 2 shows us that for the instance-based approach the best results were achieved using the Zip compressor, NCD as similarity measure, and Max Rule as fusion rule. As one can notice, the choice

Table 1
Description of the database.

Database	Number of authors	Articles per author	Average tokens	Average hapaxes
I	30	20	691 (σ = 197)	412 (σ = 103)
II	100	30	386 (σ = 197)	296 (σ = 131)

Table 2
Results of the instance-based approach on Database I.

Fusion rule	BZip		PPM		GZip	
	CCC	NCD	CCC	NCD	CCC	NCD
Max	91.6	97.0	90.6	97.3	92.3	99.0
Voting	88.3	84.0	88.6	95.0	89.3	97.0

Table 3
Results of the profile-based approach on Database I.

BZip		PPM		GZip	
CCC	NCD	CCC	NCD	CCC	NCD
98.0	62.3	96.0	61.6	94.6	77.0

of the compressor does not have much impact on the final results. The choice of the similarity measure, on the other hand, makes an enormous difference. The NCD exploits better the similarities between the references and questioned samples than the CCC. Max Rule fits well in this architecture since it takes only one good reference to provide the correct results. Since in this database few authors write about the same theme, this strategy seems very suitable in this context. The other two rules, Voting and Average, are penalized by those references that do not have a high degree of similarity.

The second experiment using Database I was set up to explore the profile-based strategy. As depicted in Fig. 1, in this approach the fusion rule does not exist since all the documents are compressed together to create the author's profile. The similarity is then computed between the questioned sample and the compressed file that contains all the references. Table 3 shows the results achieved using this approach.

In contrast to the previous experiment, the NCD has a poor performance. Compressing all the references before computing the similarity makes the NCD not suitable at all for the profile-based approach. The CCC, on the other hand, took advantage of this early compression and yielded a recognition rate of 98% using the BZip compressor. It is worth noting that the choice of the compressor matters when using the profile-based approach. As one can see, GZip is considerably worse than PPM and BZip. GZip is a dictionary-based compression algorithm and uses a sliding window of 32 K to build the dictionary. This means that if a training text is long enough the beginning of that document will be ignored when GZIp attempts to compress the concatenation of that file with the unseen text.

These results can be compared directly with the results published by Pavelec et al. in [19], since they use exactly the same experimental protocol. In their work, a stylometry-based approach was employed along with a SVM classifier. The best result reported in [19] was a recognition rate of 83.2%. Table 4 summarizes the best results achieved on Database I.

5.2. Experiments on Database II

The second batch of experiments was performed on a more complex database. As stated before, this database was proposed by Varela et al. [25] and contains 100 writers and 30 documents per writer. In their experiments they used 7 documents for training and the remaining 23 for testing. In order to be able to compare the results, we adopted the same protocol. The documents were randomly divided into training and testing and the results are the average of three runs.

Table 4
Summary of the best results achieved on Database I.

Strategy	Performance (%)
Instance-base (PPM + NCD)	99.0
Profile-based (BZip + CCC)	98.0
Stylometric features + SVM [19]	83.2

Table 5
Results of the instance-based approach on Database I using NCD.

Fusion	BZip	PPM	Zip
Max	73.26	74.04	73.96
Voting	70.83	70.43	72.35

Table 6
Results of the profile-based approach on Database I using CCC.

BZip	PPM	GZip
75.6	77.0	58.7

Following the same protocol applied to Database I, the first experiment with Database II used the instance-based approach applying the same fusion rules. Taking into account what we have learnt from the previous experiment, i.e., that the NCD is more suitable for the instance-based approach, Table 5 reports only the results using NCD.

In the second experiment we have used the profile-based approach. The results are reported in Table 6. Following the same idea of the previous experiment, only the results with CCC are reported. As in the experiments in Database I, here GZip also achieved a lower performance. Unlike in the previous experiments on Database I, here with a bigger database we were able to observe that. The best performance of the profile-based approach was 77% against 74% of the instance-based approach. Besides, the instance-based approach has to deal with a large number of calls to the compression algorithm.

The experiments reported in this section can be compared directly with the results published by Varela et al. [25], since they use exactly the same experimental protocol. In their work, a stylometry-based approach was employed along with a SVM classifier and a feature selection using a multi-objective algorithm. In their work they report the results before feature selection (58% using 408 features) and after feature selection (74% using 48 features). Table 7 summarizes the results achieved on Database II.

Since we have 100 authors in the database, analyzing the confusion matrix would be complicated. However, we were able to get some insight into the problem by analyzing the confusion matrix grouped by subject. Such a matrix can be visualized in Table 6 and it shows that the recognition rate in terms of subjects is about 80%.

The lowest performance was found for those authors writing about Gastronomy and Literature. Between them, these two classes have the greatest richness of vocabulary. Other confusions occurring very often are Politics and Economics, Literature and Misc, and Health and Gastronomy. In all these cases, authors share the same vocabulary when writing about these subjects. An example of this confusion is an author writing about health and giving some hints about healthy food (Gastronomy, Table 8).

Table 9 lists some works on authorship attribution published in the literature. Comparing different studies is not a straightforward

Table 7
Summary of the best results achieved on Database II.

Strategy	Performance (%)
Instance-based (PPM + NCD)	74.0
Profile-based (PPM + CCC)	77.0
408 Stylometric features + SVM [25]	58.0
58 Stylometric features (+feature selection) + SVM [25]	74.0

Table 8
Confusion matrix by subjects in %.

	a. Misc	b. Law	c. Economics	d. Sports	e. Gastronomy	f. Literature	g. Politics	h. Health	i. Technology	j. Tourism
a.	87.0	1.7	3.0	1.3	0.0	3.9	1.3	0.9	0.4	0.4
b.	5.0	83.5	3.9	0.9	2.2	0.8	0.4	0.4	1.7	0.8
c.	4.7	7.8	74.4	0.0	1.3	0.0	6.1	2.2	2.1	1.3
d.	0.4	0.4	0.4	97.0	0.4	0.0	0.4	0.4	0.0	0.4
e.	1.7	7.0	6.7	2.2	65.6	2.2	0.0	8.2	4.3	1.7
f.	6.1	9.5	4.3	2.2	3.0	57.4	4.3	4.7	6.1	2.2
g.	2.2	0.8	7.4	0.0	0.0	0.4	83.9	0.8	4.3	0.0
h.	1.7	4.3	7.0	0.9	2.1	0.0	0.0	81.7	2.1	0.0
i.	0.4	1.3	3.5	0.9	0.0	0.4	0.0	0.8	91.8	0.9
j.	3.5	2.2	7.4	1.7	0.4	0.8	2.2	3.5	4.3	73.9

Table 9
Published works on authorship attribution.

Ref.	Classifier	Database	Rec. rate (%)
[23]	SVM	Web pages	66–80
[7]	SVM	German newspaper	80
[10]	SVM	3 sister's letters	75
[24]	kNN	Novels	66–76
[5]	Distance	Brazilian novels	78
[19]	SVM	Brazilian newspaper	72
[6]	Bayes	Mexican poems	60–80
[21]	Bayes	Turkish newspaper	80
[25]	SVM	Brazilian newspaper	74

task since most of the studies use different databases and classifiers. However, by analyzing Table 9 we can see that the results achieved in this study compare to the state of the art.

6. Conclusion

In this paper we discussed the use of compression algorithms for authorship identification. In this study we have selected three different types of compressors: Lempel-Ziv type (GZip), block sorting type (BZip) and statistical type (PPM). In order to compute the dissimilarity between two documents, two different compression-based similarity measures were assessed in our tests, the NCD (Normalized Compression Distance) and CCC (Conditional Complexity of Compression). These six possible combinations were tested using both instance- and profile-based attribution methods.

A series of comprehensive experiments on two different databases show a strong correlation between the attribution method and the similarity measures tested, i.e., NCD seems more suitable for the instance-based approach while CCC always produce better results when used with the profile-based approach. In addition, the experiments also show that the compression algorithms are an interesting alternative for authorship identification comparing favorably to traditional strategies based on feature extraction and classification. In the case of the Database I, the tuple GZip-NCD achieved 99% against 82% of the traditional approach based on stylometric features and a SVM classifier. In the case of the Database II, the compression-based approach brought an improvement of about 3%.

By analyzing the confusion matrices, we have observed that different compression algorithms generate different confusions. With this in mind, in future studies we plan to combine the results of different compressors to increase the final recognition rate.

Acknowledgements

This research has been supported by The National Council for Scientific and Technological Development (CNPq) grant 301653/2011-9 and Araucaria Foundation grant 21277.

References

[1] S. Argamon, M. Saric, S.S. Stein, Style Mining of Electronic Messages for Multiple Author Discrimination, in: ACM Conference on Knowledge Discovery and Data Mining, 2003.
[2] D. Benedetto, E. Caglioti, V. Loreto, Language trees and zipping, Phys. Rev. Lett. 88 (January (4)) (2002) 2–5.
[3] M. Burrows, D.J. Wheeler, A block-sorting lossless data compression algorithm. Technical Report 124, Digital SRC Research, 1994.
[4] R. Cilibrasi, P. Vitanyi, Clustering by compression, IEEE Trans. Inf. Theory 51 (4) (2005) 1523–1545.
[5] B.C. Coutinho, L.M. Macedo, A. Rique-JR, L.V. Batista, Atribuição de autoria usando PPM, in: XXV Congress of the SBC, 2004, pp. 2208–2217 (in Portuguese).
[6] R.M. Coyotl-Morales, L. Villasenor-Pineda, M.M. Gomez, P. Rosso, Authorship attribution Using Word Sequences, in: Iberoamerican Congress on Pattern Recognition, 2006, 844–853.
[7] J. Diederich, J. Kindermann, J. Leopods, G. Paass, Authorship attribution with support vector machines, Appl. Intell. 1 (2003).
[8] R.S. Forsyth, D.I. Holmes, Feature finding for text classification, Literary Linguist. Comput. 11 (4) (1996) 163–174.
[9] E. Frank, C. Chui, I.H. Witten, Text Categorization Using Compression Models, in: Data Compression Conference, 2000.
[10] M. Gamon, Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features, in: 20th International Conference on Computational Linguistics, 2004, 611–617.
[11] P. Juola, Future trends in authorship attribution, in: Advances in Digital Forensics III, Springer, Boston, 2007, pp. 119–132.
[12] M. Koppel, J. Schler, Exploiting Stylistic Idiosyncrasies for Authorship Attribution, in: Workshop on Computational Approaches to Style Analysis and Synthesis, 2003.
[13] O.V. Kukushkina, A.A. Polikarpov, D.V. Khmelev, Using literal and grammatical statistics for authorship attribution, Probl. Inf. Trans. 37 (2) (2001) 172–184.
[14] M. Li, X. Chen, X. Li, B. Ma, P. Vitanyi, The similarity metric, IEEE Trans. Inf. Theory 50 (December (12)) (2004) 3250–3264.
[15] D. Madigan, A. Genkin, D.D. Lewis, S. Argamon, D. Fradkin, L. Ye, Author Identification on the Large Scale, in: Joint Annual Meeting of the Interface and the Classification Society of North America (CSNA), 2005.
[16] M. Malyutov, Authorship attribution of texts: a review, Electron. Notes Discrete Math. 21 (August) (2005) 353–357.
[17] Y. Marton, N. Wu, L. Hellerstein, On Compression-based Text Classification, in: European Conference on Information Retrieval, 2005, 300–314.
[18] A. Moffat, Implementing the PPM data compression scheme, IEEE Trans. Commun. 38 (11) (1990) 1917–1921.
[19] D. Pavelec, L.S. Oliveira, E. Justino, L.V. Batista, Using conjunctions and adverbs for author verification, J. Univers. Comput. Sci. 14 (2008) 2967–2981.
[20] E. Stamatatos, A survey of modern authorship attribution methods, J. Am. Soc. Inf. Sci. Technol. 60 (3) (2009) 538–556.
[21] T. Tas, A.K. Gorur, Author identification for Turkish texts, J. Arts Sci. 7 (2007) 151–161.
[22] W. Teahan, D. Harper, Using Compression-based Language Models for Text Categorization, in: Workshop on Language Modeling and Information Rerieval, 2001.
[23] Y. Tsuboi, Y. Matsumoto, Authorship identification for heterogeneous documents, IPSJ SIG Notes, 2002, pp. 17–24.
[24] O. Uzuner, B. Katz, A Comparative Study of Language Models for Book and Author Recognition, in: 2nd International Joint Conference on Natural Language Processing, 2005, 969–980.
[25] P.J. Varela, E. Justino, L.S. Oliveira, Selecting Syntactic Attributes For Authorship Attribution, in: IEEE International Joint Conference on Neural Networks, 2011, 161–172.
[26] R. Zheng, J. Li, H. Chen, Z. Huang, A framework for authorship identification of online messages: writing style features and classification techniques, J. Am. Soc. Inf. Sci. Technol. 57 (3) (2006) 378–393.
[27] J. Ziv, A. Lempel, A universal algorithm for sequential data compression, IEEE Trans. Inf. Theory 23 (3) (1977) 337–343.