International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

Influence of compression distance measures on Authorship Attribution

P.Vijayapal Reddy
Department of Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
vijayapalreddy76@gmail.com

Abstract

Authorship attribution (AA) can be defined as the task of inferring characteristics of a document's author from the textual characteristics of the document itself. In this paper, it is evaluated the compression model for AA on Telugu text. It considered LZW compression model with three different compression distance measures such as Normalized Compressor Distance (NCD), Compression Dissimilarity Measure (CDM) and Conditional Complexity of Compression (CCC) . The results shows that the compression models are good alternatives for Authorship attribution. The model is evaluated using micro-average F1, macro-average F1 and accuracy measures.

Keywords: Authorship attribution, Compression distance measures, Macro-average, Micro-average, Accuracy.

1. Introduction

Authorship attribution research can be broadly categorized in two ways. A set of features with machine learning algorithms and using compression algorithms. From the last decade, compression algorithms were effectively applied to group different text documents [2]. In compression algorithms, similarity between two documents can be calculated using compression distance. As the distance between documents is small, then the two documents are more similar to each other, while a large distance indicates a dissimilarity between test and training set. Data compression algorithms are best alternative approach for authorship attribution compared with the text classification model. File compression algorithms find duplicated strings in the text and checks for the longest matching strings. More frequent text sequences are coded with less bytes where as rare sequences will be coded with more bytes [5].

Compression is the process of encoding original document using fewer number of bits. The process of authorship attribution using data compression is as follows. Given an unknown document d_i, and a set of training documents A_j of author j then compression algorithm S is applied to the original document set A_j and also to the concatenated of documents A_j and d_i such as $A_j + d_i$. The relative size after compression ΔS is then calculated as $S(A_j+d_i) - S(A_j)$ where $S(A_j+d_i)$ is the size of concatenated document after compression and $S(A_j)$ is the size of A_j after compression. The test document d_i is assigned to the author j if the smallest ΔS is computed with A_j. This difference is the cross-entropy between the two text documents.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

The advantages data compression algorithms [3] for authorship attribution compared with classification model is that it avoids the word ambiguities, it considers only phrasal effects other than word boundaries, it deals with different types of documents uniformly. In this paper an attempt is made for authorship attribution using different different distance measures with the LZW as a compression model.

2. Related Work

There is an extensive research carried on authorship attribution using various features and with various classifiers. In [1], word length is used as a feature for authorship attribution. In [17], sentence lengths are used to judge authorship. The function words for authorship attribution is considered in [15]. The authors in [8] conducted experiments with support vector machine classifiers with various features. In [4], the study for authorship recognition implements multiple regression and discriminant analysis. In [7], a function is generated to co-relate the word frequency and the text length. Karlgren-Cutting in [15], considered various style markers of the text for authorship attribution. Biber in [6] considered the syntactic and lexical style markers. Burrows in [3,13] used principal components analysis (PCA) to combine various style markers which can discriminate among set of authors. In [15] machine learning algorithms such as naive bayes, decision tree and support vector machines were used to design discrimination models on large number of documents and features. In [17], author considered syllables per word using ngrams for authorship attribution. Stylometric features such as vocabulary richness and lexical repetition based on Zipf's [18] were studied on word frequency. Features such as word class frequencies, syntactic analysis, word collocations, grammatical errors, word, sentence, clause, and paragraph lengths for authorship attribution were applied in [16].

There are two types of approaches for authorship attribution namely instance-based approach and profile-based approach. Compression algorithms are used to compress test documents and compare these compressed test documents with author profiles which is author wise compressed training documents sets. A high compression rate of test document with a particular author profile shows attribution towards that particular author. Many compression approaches were proposed for authorship attribution to assign test documents to corresponding author [5,6]. Compression rate between documents, compression distances and other approaches are used to attribute a text. Preprocessing is not required for input documents while using compression algorithms for authorship attribution. Many compression methods have been used to attribute and categorize texts such as LZ76, LZ77, LZW, RAR, gzip, PPM. The method proposed in [7], shows good results with LZ76 where as other methods supports PPM family over LZ variants [9].

The compression algorithms builds a dictionary or a model using training text documents set. These generated models are used to the train classifiers. Test document can be assigned to a particular author by compressing this test document for each author specific model or dictionary which is generated during training phase. The test document is attributed to an author which is produced the highest compression rate [2]. In order to measure the compressed distance similarity many metrics were proposed in the literature [5].

3. Compression Model for Authorship Attribution

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

Let C be a set of n authors, L is a set of training documents of all the known authors and T is a set of test documents. Authorship attribution method assigns each document from test set T to a candidate author from set C. In the first step, all the training texts of each author are concatenated and saved in one file. Concatenated training document per author is compressed using any compression algorithm which results to a author's profile, represents author's style. In the second step, the similarity between compressed test document t and the author profiles is computed. Then, the test document is assigned to one of the authors that minimizes the similarity distance as shown in Figure 3.1.

3.1 Flowchart

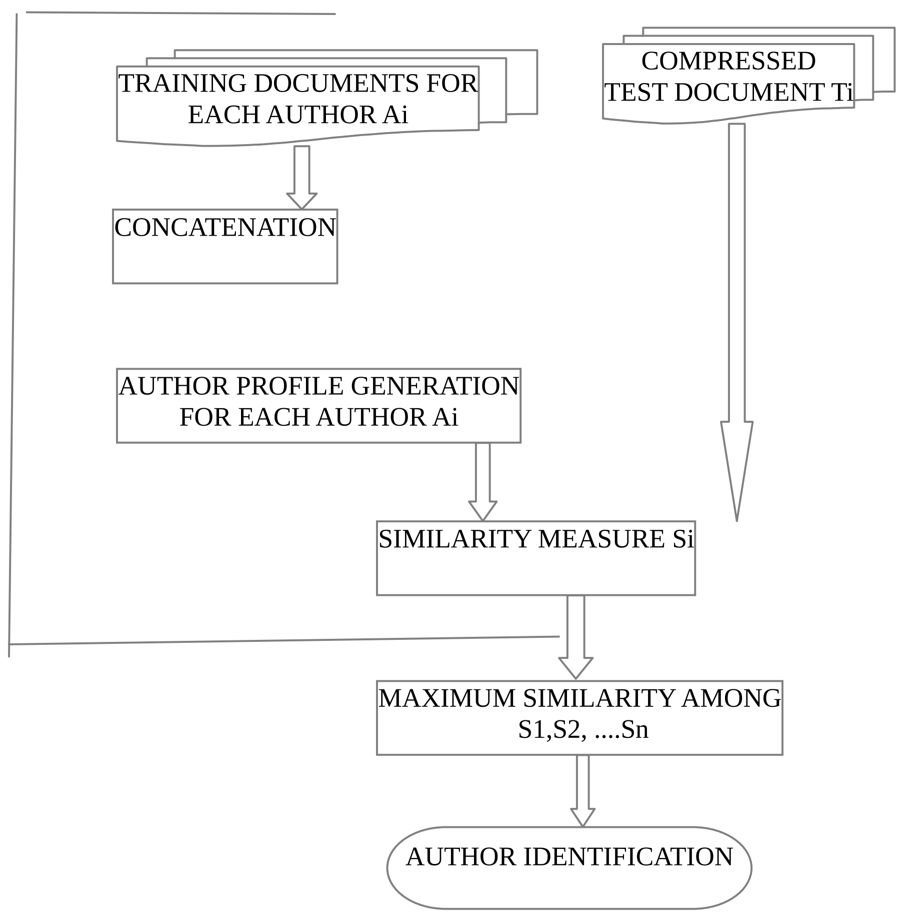

Figure 3.1: Flowchart for profile-based Authorship Attribution using data compression

Data compression is the process of reducing the size of the data file. Compressors are to find the shortest sequence of bits needed to represent a text. There are two ways of compressing data namely lossless data compression and lossy data compression. For authorship attribution lossless

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

data compression techniques are used. All data compression algorithms consist of two parts, a model which estimates the probability distribution and a coder which assigns the shortest codes to the most likely character. In Lempel Ziv Welch algorithm (**LZW**) compression algorithm the input file is read character by character and they are combined to form a string. The process continues till it reaches the end of file. Every new string is assigned some code and stored in Code table.

3.2 Compression distance measures

Compression distances measures are used to compute a distance between two compressed text files. Compression based measures are used to estimate the amount of information shared by any two text documents. They can be utilized for clustering and classification on different types of data such as texts and images [20, 1].

Compression Dissimilarity Measure (CDM)

Compression Dissimilarity Measure (CDM) proposed in [1]. For documents x and y, the compression dissimilarity measure is defined as:

$$CDM(x,y) = \frac{C(xy)}{C(x)+C(y)} \quad (1)$$

where C (x) is the size of the compressed object x, C (y) is the size of the compressed object y, xy is the concatenation of x and y and C(xy) is the size of the compressed object xy.

Normalized Compressor Distance (NCD)

Normalized Compression Distance (NCD) proposed in [2] uses general compressors to estimate the amount of shared information between two objects. The Normalized Compression Distance is defined as:

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

where C(x) is the size of the compressed object x. 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 due to imperfections characterizing real compressors.

Conditional Complexity of Compression (CCC)

Conditional Complexity of Compression (CCC) proposed in [5,27]. The CCC of text y given text x is calculated by

$$CCC(y/x) = (S_c) - (x_c) \quad (3)$$

where $|x_c|$ is the length of the compressed text x. The S is the concatenated text of xy. CCC approximates a more abstract Kolmogorov conditional complexity and measures adapts to patterns in the training text for better compressing the unknown text.

4. Results and Discussions

In order to evaluate the performance of the proposed model, standard information retrieval metrics such as precision, recall and F1 -measure are used. Precision P_A, for author A, is defined as:

$$P_A = \frac{correct(A)}{totalretrieveddocuments(A)} = \frac{TP_A}{TP_A + FP_A} \quad (4)$$

where TP_A (True Positive) is the number of documents that are correctly attributed to author A and FP_A (False Positive) is the number of documents that are incorrectly attributed to author A.

Recall R_A, for author A, is defined as :

$$R_A = \frac{correct(A)}{totalrelevantdocuments(A)} = \frac{TP_A}{TP_A + FN_A} \quad (5)$$

where FN_A (False Negative) is the number of missed attributions for author A.

F1 -measure, which is defined as the harmonic mean of recall and precision as:

$$F_1 = \frac{2 * P_A * R_A}{P_A + R_A} \quad (6)$$

F1 depends on author A. In order to aggregate these measures over all different authors micro-average and macro-average were defined as follows.

Given a metric M (precision, recall or F1), for a set of n authors, these measures are defined as:

$$macro-average_M = 1/n \sum_{i=1}^{n} M_{Ai} \quad (7)$$

$$micro-average_M = 1/k \sum_{i=1}^{n} (D_{Ai}) M_{Ai} \quad (8)$$

where k is the total number of test documents and $|D_{Ai}|$ is the number of documents in the test set for author A_i.

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

Accuracy is another measure and defined as:

$$Accuracy = \frac{Numberofdocumentsthatarecorrectlyassigned}{Totalnumberoftestdocuments} \quad (9)$$

Experiments using the data set with various distance measures were conducted. All the documents are compressed together to create the author's profile. The similarity is then computed between the compressed test document and the compressed author specific documents that contains author profiles. The obtained results are presented in Table 4.1

Measure	**Macro_average F1 measure**	**Micro_average F1 measure**	**Accuracy**
Compression distance			
NCD	**0.55**	**0.50**	**0.68**
CDM	**0.51**	**0.51**	**0.66**
CCC	**0.57**	**0.62**	**0.71**

Table 4.1: Macro, Micro-F1 measures and accuracy for various compression distance measures

The compression method LWZ with three distance measures CDM, NCD and CCC are used to test the performance. The compressor is performing well with Conditional Complexity of Compression (CCC) distance measure.

5. Conclusions and Future Scope

This paper evaluates the performance of compression-based similarity measures on authorship analysis on natural texts. 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. In this study in order to compute the similarity between Author profile and test document, three different compression-based similarity measures were used in the experiments such as NCD (Normalized Compression Distance) and CCC (Conditional Complexity of Compression) and CDM (Compression Dissimilarity Measure). Our experimental results shows that the compression algorithms are an interesting alternative for authorship identification comparing favourably to traditional strategies based on feature extraction and classification. CCC seems more suitable for the profile-based approach compared with NCD and CDM.

References

1. E. Keogh, S. Lonardi, and C. A. Ratanamahatana, "Towards parameter-free data mining," in

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 02 Issue 11
November 2015

Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD ’04. New York, NY, USA: ACM, 2004, pp. 206–215.

2. R. Cilibrasi and P. M. B. Vit, “Clustering by Compression,” IEEE Transactions on Information Theory, vol. 51, no. 4, pp. 1523–1545, 2005.

3. C. J. V. Rijsbergen, Information Retrieval, 2nd ed. Butterworth-Heinemann, 1979.

4. Y. Yang, “An Evaluation of Statistical Approaches to Text Categorization,” Information Retrieval, vol. 1, no. 1-2, pp. 69–90, 1999.

5. D. Benedetto, E. Caglioti, and V. Loreto, “Language Trees and Zipping,” Physical Review Letters, vol. 88, p. 048702, 2002.

6. D. V. Khmelev and W. J. Teahan, “A repetition based measure for verification of text collections and for text categorization,” in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, ser. SIGIR ’03. New York, NY, USA: ACM, 2003, pp. 104–110.

7. M. Lambers and C. Veenman, “Forensic Authorship Attribution Using Compression Distances to Prototypes,” in Proceedings of the Third International Workshop on Computational Forensics. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 13–24.

8. Y. Marton, N. Wu, and L. Hellerstein, “On compression-based text classification,” In Proceedings of the European Conference on Information Retrieval, pp. 300–314, 2005.

9. D. Sculley and C. E. Brodley, “Compression and Machine Learning: A New Perspective on Feature Space Vectors,” in Proceedings of the Data Compression Conference, ser. DCC ’06. Washington, DC, USA: IEEE Computer Society, 2006, p. 332.

10. V. Bobicev, “Text Classification Using Word-Based PPM Models,” The Computer Science Journal of Moldova, vol. 14, no. 2, pp. 183–201, 2006. 11. E. Frank, C. Chui, and I. H. Witten, “Text Categorization Using Compression Models,” in Proceedings of the Conference on Data Compression, ser. DCC ’00. Washington, DC, USA: IEEE Computer Society, 2000, p. 555.

12. F. Peng, D. Schuurmans, and S. Wang, “Language and task independent text categorization with simple language models,” in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, ser. NAACL ’03. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003, pp. 110–117.

13. C. Shannon, “A Mathematical Theory of Communication,” The Bell System Technical Journal, vol. 27, pp. 379–423 & 623–656, 1948.

14. A. Lempel and J. Ziv, “On the Complexity of Finite Sequences,” IEEE Transactions on Information

Theory, vol. 22, no. 1, pp. 75–81, 1976.

15. J. Cleary and I. Witten, “Data compression using adaptive coding and partial string matching,” IEEE Transactions on Communications, vol. 32, no. 4, pp. 396–402, 1984.

16. “PPM: One Step to Practicality,” in Proceedings of the Data Compression Conference, ser. DCC ’02. Washington, DC, USA: IEEE Computer Society, 2002, p. 202.

17. P. G. Howard, “The Design and Analysis of Efficient Lossless Data Compression Systems,” Providence, RI, USA, Tech. Rep., 1993.

18. Bratko, A., Filipic, B.: Spam Filtering Using Compression Models. Department of Intelligent Systems, Jozef Stefan Institute, Ljubljana, Slovenia, IJS-DP-9227 (2005)

19. Cerra, D., Datcu, M., 2012. A fast compression-based similarity measure with applications to