A Weighted Profile Intersection Measure for Profile-based Authorship Attribution

Hugo Jair Escalante[1], Manuel Montes-y-Gómez[2,3], Thamar Solorio[3]
hugojair@gmail.com, mmontesg@inaoep.mx, solorio@cis.uab.edu

[1] Universidad Autónoma de Nuevo León,
San Nicolas de los Garza, 66450, N. L., Mexico
[2] Instituto Nacional de Astrofísica, Óptica y Electrónica,
Tonantzintla, 72840, Puebla, Mexico
[3] University of Alabama at Birmingham,
Birmingham, AL, 35294, USA

Abstract. This paper introduces a new similarity measure called weighted profile intersection (WPI) for profile-based authorship attribution (PBAA). Authorship attribution (AA) is the task of determining which, from a set of candidate authors, wrote a given document. Under PBAA an author's profile is created by combining information extracted from sample documents written by the author of interest. An unseen document is associated with the author whose profile is most similar to the document. Although competitive performance has been obtained with PBAA, the method is limited in that the most used similarity measure only accounts for the number of overlapping terms among test documents and authors' profiles. We propose a new measure for PBAA, WPI, which takes into account an inter-author term penalization factor, besides the number of overlapping terms. Intuitively, in WPI we rely more on those terms that are (frequently) used by the author of interest and not (frequently) used by other authors when computing the similarity of the author's profile and a test document. We evaluate the proposed method in several AA data sets, including many data subsets from Twitter. Experimental results show that the proposed technique outperforms the standard PBAA method in all of the considered data sets; although the baseline method resulted very effective. Further, the proposed method achieves performance comparable to classifier-based AA methods (e.g., methods based on SVMs), which often obtain better classification results at the expense of limited interpretability and a higher computational cost.

1 Introduction

Recent advances on information technology have motivated the generation of huge amounts of data. For example, users of social networks (e.g., Facebook[1]) and microblogging websites (e.g., Twitter[2]) generate millions of texts every day.

[1] http://www.facebook.com/
[2] http://www.twitter.com/

Analyzing such information has important benefits (e.g., anticipation of terrorist attacks, cyber-crime detection, tracking of marketing trends, and opinion mining), thus posing a major challenge to the Natural Language Processing and Artificial Intelligence communities, in terms of both efficiency and performance.

Authorship attribution (AA) is the task of identifying whom, from a set of candidates, is the author of a given document [19]. AA applications include spam filtering [2], fraud detection, computer forensics [10], cyber bullying [14] and plagiarism detection [16]. Because of its wide applicability, mainly in security aspects, the development of automated AA techniques has received much attention recently [19]. Many AA methods have been proposed so far, some more complex than others. One of the most used approaches nowadays is that based on author profiles [5, 11].

In profile-based authorship attribution (PBAA) an author's profile is created by combining information extracted from sample documents written by the author of interest [5, 19]. An unseen document is associated with the author whose profile is most similar to the document. Figure 1 depicts the PBAA formulation. The PBAA approach is highly efficient, and it has the additional benefit that the generated profiles can reveal helpful and interpretable information about the writing style of authors. Competitive performance has been obtained with PBAA [5, 7–9, 11, 19], however, the method is limited in that the most used similarity measure only accounts for the number overlapping terms among test documents and authors' profiles.

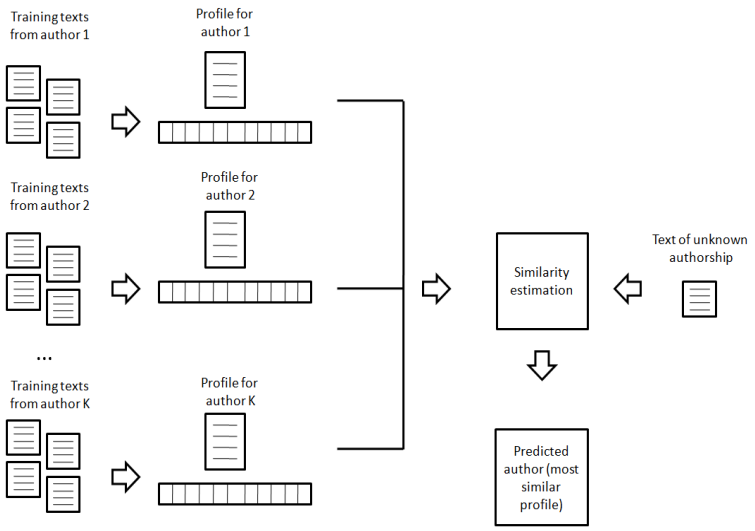

Fig. 1. Diagram of the profile-based authorship attribution approach. Documents from each author are combined to obtain a profile (e.g., a prototypical vector or document). Unseen documents are compared to every profile. The document is associated with the author of the most similar profile. Figure inspired from [19].

In this paper we propose a new similarity measure for PBAA, the weighted profile intersection (WPI) measure, which takes into account an inter-author term penalization factor, besides the number of overlapping terms. Intuitively, the WPI measure gives a high weight to terms that are used by less authors and it assigns a low weight to terms that are shared across profiles for different authors. The term weights are considered for weighting the number of overlapping terms among profiles and documents.

We evaluate the proposed method in several AA data sets, including many subsets from Twitter (a challenging corpus because of the short length of its texts). Experimental results confirm that the baseline PBAA method resulted very effective for AA. However, the proposed technique outperforms the standard PBAA method in all of the considered data sets. Also, we compare the performance of the proposed method to that obtained by several classifiers (under classifier-based AA). We found that the proposed method achieves comparable performance to that of classifier-based AA methods. However, one should note that PBAA methods are advantageous over classifier-based methods because they are based on more interpretable representations and are more efficient.

The rest of the paper is organized as follows. The next section reviews related work on AA. Section 3 presents the proposed PBAA method. Section 4 describes the experimental evaluation of the proposed method. Section 5 presents conclusions and outlines future work.

2 Related work

The AA task can be faced as one of single-label multiclass classification, with as many classes as candidate authors. However, unlike usual text categorization tasks, where the core problem is modeling the thematic content of documents [18], the goal in AA is modeling authors writing style [19]. Hence, most of the work in AA proposes the use of document representations that are believed to be topic-free and geared towards revealing the writeprint of authors [7, 9, 12, 19]. These features have been loosely called stylometric features, due to the early work on stylometry. Stylometric features typically include character, lexical, syntactical, grammatical and semantic features [7, 9, 19]. Nevertheless, despite the fact that elaborated stylometric features have been used for AA, representations based on character n-grams or words are predominant [1, 4, 6–9, 12, 19]. In the classifier-based approach several standard learning algorithms have been evaluated, including the popular support vector machine (SVM) classifiers [4, 6] neural networks [20], Bayesian classifiers [1] and decision trees [9].

Another popular formulation for AA is the so called profile-based AA (PBAA) approach [7–9, 11, 19]. Under this approach the information of sample documents from each author is combined for building author profiles (i.e., a prototypical document or vector). An unseen document is associated with the author whose profile is most similar to the document. PBAA methods are advantageous over classifier-based approaches in that they are easy to implement, efficient to apply, scale well to large numbers of documents and authors, and they have attractive

interpretability properties (e.g., we can know what terms are more important for each author, and we can also compare profiles of different authors).

In the most used PBAA method called *common n-grams* (CNG) a profile is the set of the top$-L$ more frequent words used by the author in their sample documents [8, 19]. A test document and a profile are compared by a normalized frequency of overlapping terms. CNG, introduced by Keselj et al., is perhaps the most used PBAA method [7–9, 11, 19]. Similar profiles are adopted by Frantzeskou et al. although they propose a simplified similarity measure called: simplified profile intersection (SPI) [5]. SPI is simply the un-normalized number of overlapping terms among profiles and test documents. The SPI similarity measure outperformed the similarity measure adopted by Keselj et al. [8], which is slightly more complex. Therefore, in recent AA studies the SPI similarity measure is preferred [11]. This paper proposes an extension to the SPI measure that accounts for an inter-author term-weighting factor besides considering the number of overlapping terms among profiles and test documents. In Section 4 we show that our method outperforms the standard PBAA that uses the SPI measure in a suite of AA data sets, and that it compares favorably with classification-based methods.

3 Proposed method

The PBAA approach to AA is depicted in Figure 1. In agreement with related work we adopt the CNG approach as base model for developing our PBAA method. Under the CNG method with SPI similarity measure (hereafter CNG-SPI), a profile for an author is the set of the $L-$most frequent words in the sample documents from that author. Consider a scenario where we have $K-$candidate authors, let $P_1, \ldots, P_K$ denote the profiles for each of the $K-$authors. When a test document needs to be classified, a profile is obtained for the test document as well. We consider a profile for a test document to be the set of all of the terms appearing in that document. Let T_j denote the j^{th} test document and let $I_j^i = P_i \cap L_j$ be the set of terms that overlap between the profile of the i^{th} author, P_i, and the j^{th} test document, T_j, then the SPI similarity measure is defined as [5]:

$$S_{spi}(P_i, T_j) = |I_j^i| \tag{1}$$

The test document will be assigned to the author's profile with the largest SPI measure.

Formula (1) above only takes into account the raw number of terms in the intersection between profiles. Despite its simplicity, very good results have been reported with the CNG-SPI technique [5, 11, 19]. However, we think that the CNG method can be improved by adopting a weighted similarity measure.

We propose the weighted profile intersection (WPI) measure, which incorporates a term-weighting factor proportional to the usage of the term across the profiles of the candidate authors. The intuition behind CNG-WPI, is that terms in the intersection I_j^i that are used by a single author (or by a small number of authors) must receive a higher weight, as it is more likely that the intersection

of these terms is indicative of the agreement between profile and test document. On the other hand, terms in I_j^i that appear in the profiles of most of the candidate authors should receive a lower weight, as these terms are likely to appear in many profiles. In agreement with the above arguments we consider the following weighting factor for each term that appears in at least one author profile:

$$w_l = \frac{1}{\sum_{k=1}^{K} \mathbf{1}_{t_l \in P_k}} \tag{2}$$

where w_l is the weight associated with term t_l and $\mathbf{1}_{t_l \in P_k}$ is an indicator function taking the value 1 when $t_l \in P_k$ is true and 0 otherwise. Therefore, the weight associated to term t_l is inversely proportional to the number of profiles in which the term occurs. This weighting scheme was partially inspired by the well known tf-idf weighting scheme in information retrieval, where the idea is to sink the relevance of terms that occur frequently in most of the documents in the collection while boosting the weight of terms that are rare in the document collection. The motivation behind both weighting schemes is very similar in spirit with the exception that in our approach we do not account for the frequency of the terms explicitly and we penalize the terms according to their presence in the profiles of the authors, not the entire collection of sample documents. One should note that term frequency is considered implicitly in the proposed approach, as we use it as the only criterion to select terms for building the profile for an author.

Besides including a term weighting factor we wanted to reduce the impact of test documents with a large number of terms, which are more likely to have a larger number of overlapping terms in the intersection with the authors' profiles just by chance. Therefore, we define the WPI similarity measure between an author profile P_i and the test document T_j as follows:

$$S_{wspi}(P_i, T_j) = \frac{1}{|I_j^i|} \times \sum_{k=1}^{|I_j^i|} w_{I_j^i} \tag{3}$$

Formula (3) is just the average weight of overlapping terms between the test document and the author profile. By using the average instead of the sum we reduce the influence of the number of terms in the test documents.

The next section reports experimental results in several AA data sets that have been used in previous AA studies. From Formulas (1) and (3) we can see that the only parameter of CNG-SPI and CNG-WPI is L, the number of more frequent terms to consider for building the authors' profiles. We evaluate the performance of both methods with respect to this parameter in the next section as well.

4 Experimental evaluation

This section reports an experimental evaluation of the proposed CNG-WPI method. We first describe the considered data sets and then we present the experimental results.

4.1 Authorship attribution data sets

We consider several AA data sets that have been used in previous studies [11, 15, 17]. Table 1 shows some statistics about these data sets[3]. Our inclusion criteria was to provide a good variety of genre and domains, as well as corpora sizes and number of candidate authors. Five data sets are due to Raghavan et al. (rows 2-6), these documents were collected from the web [17]. The CCAT data set was first used by Stamatatos et al. [6] and then by Plakias et al. [15]. This data set contains documents from news about the same topic written by different authors. Finally, the Twitter data set is a subset of the data set collected by Layton et al. [11]. It contains around $5,000$ documents written by 50 authors. Note that this data set is particularly challenging as each document is a tweet of 140 characters or less. With the exception of Twitter, in each of the considered data sets the authors wrote documents in the same topic. Hence, it is expected that the theme of documents does not have an impact in the performance of the considered methods.

Table 1. Data sets considered for experimentation. We show the number of authors, terms, training and test documents for each data set.

Data set	Authors	Terms	Train	Test	Reference
Football	3	8620	52	45	[17]
Business	6	10550	85	90	[17]
Travel	4	11581	112	60	[17]
Cricket	4	10044	98	60	[17]
Poetry	6	8016	145	55	[17]
CCAT	10	15587	500	500	[6]
Twitter	50	26156	4500	500	[11]

Besides the Twitter subset described in Table 1, we performed experiments by generating subsets of different sizes taken from the original corpus of Layton et al. [11]. Specifically, we considered documents in the folder *raw-depth_sample_II*, where there are documents written by 100 different authors with an average of 177 documents per author. Different evaluation subsets were randomly generated with replacement. In each subset 70% of the documents are used for training and the rest for testing.

Based on previous studies we used as terms character n-grams with $n = 3$, where spaces and punctuation marks are considered terms. Character $3-grams$ have proved to be very effective, this is the most used representation for the AA task [4, 6, 8, 13, 15, 19]. For the evaluation of the different methods we consider accuracy (percentage of documents assigned to the correct author) as leading evaluation measure.

[3] Each data set can be obtained by contacting the authors of the respective references.

4.2 CNG-SPI vs CNG-WPI

In this section we compare the performance of the CNG-SPI and CNG-WPI methods. For these experiments we consider subsets of the Twitter collection, our choice is justified by the fact that this is the more challenging collection in terms of sparsity (each document is a tweet of 140 characters or less [11]) and number of candidate authors (50). Also, this collection contains many documents per author, therefore we can control the number of documents per author and generate multiple subsets for the evaluation of different aspects of CNG-WPI. Furthermore, the generation of multiple subsets allows us to determine whether differences in performance are statistically significant. For this experiment we randomly generated subsets of documents from the Twitter data set. For each subset we randomly chose 50 authors and a different number of documents per author: 20, 50, 100, 150 and 200.

Figure 2 shows the performance obtained with the CNG-WPI method for different values of L (the number of most frequent terms to consider for the profiles). Each result is the average over five runs (using a different subset randomly generated for each run) of the application of CNG-WPI. Rather than using specific values for L we defined it in terms of the size of the vocabulary for each collection. We tried the values of L that represent 10%, 20%, 30%, 40% and 50% of the number of terms in each data set. Then a total of $5 \times 5 \times 5 = 125$ runs were performed, (5-runs, 5-author sizes and 5-values of L).

From Figure 2 we can see that the accuracy is low for all of the settings, although it is much higher than the random-guessing baseline (expected performance of 2%). As expected, the performance of the CNG-WPI method improves as the number of documents per author increases. Regarding the value of L, using the 10% of the vocabulary size seems to be the best option. This can be due to the fact that as more terms are considered, the number of terms shared by different profiles increases, thus reducing the impact of the proposed term-weighting.

Table 2. Improvement of CNG-WPI over CNG-SPI for Twitter subsets. Each result is the average of five runs of the methods in different data subsets. Columns show the percentage of terms from the vocabulary that was used for building profiles (L) and rows show the number of documents per author considered. Light grayed cells show differences that were statistically significant at the 95% level.

	Accuracy Improvement over CNG-SPI				
Terms	**10**%	**20**%	**30**%	**40**%	**50**%
20 Docs.	11.25%	7.50%	9.60%	3.80%	5.00%
50 Docs.	5.68%	7.36%	6.56%	4.16%	5.60%
100 Docs.	2.56%	3.35%	2.53%	0.47%	1.90%
150 Docs.	2.47%	1.93%	2.20%	1.53%	1.87%
200 Docs.	−0.20%	1.10%	−0.20%	−1.10%	−0.05%

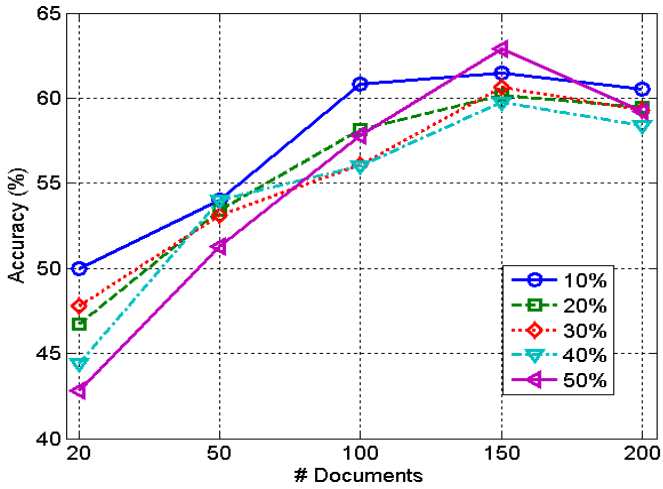

Fig. 2. Performance of CNG-WPI method in the Twitter subsets with 50 authors. Each result is the average of five runs of the method in different data subsets. We show accuracy curves for different values of L, the percentage of most frequent terms (3-grams) that was used for building profiles.

Table 2 shows the average (over five runs in different data sets) improvement offered by CNG-WPI over CNG-SPI, for different values of L and different numbers of training documents per author. The differences that were statistically significant at the 95% level are colored in light gray. We used a Wilcoxon signed-rank test as it is the recommended test when comparing classification methods over different data sets [3]. We can see that CNG-WPI outperforms CNG-SPI for most of the considered settings. The improvements are more important for Twitter subsets with a small number of documents per author (e.g., see columns 10 Docs. and 20 Docs. columns). This result is interesting since in real scenarios one usually deals with data sets with a limited number of documents per author. The largest improvement is obtained when L is 10% of the vocabulary. This result is somewhat expected as when less terms are considered, the weighting factor is more important. When 200 documents per author were used, the CNG-SPI method outperformed CNG-WPI for 4 out of the 5 values of L tested. This can be due to the fact that for this number of documents more words tend to co-occur in profiles of many authors, which causes some words to be underestimated by the weight factor in Formula (2). However, the differences in performance in Table 2, row 7 (200 Docs.) are only statistically significant for the columns 20% and 40%.

Summarizing the results from this section, we showed that the proposed method outperforms the CNG-SPI technique. Recall that the experimental results reported in Table 2 comprise a total of 125 runs using different subsets for

CNG-SPI and CNG-WPI. While the improvement is not dramatic it is important and evidences the benefits of our approach.

4.3 CNG-SPI vs Classification-based methods

This section reports experimental results of the comparison of CNG-WPI and classification-based approaches. Different classifiers were evaluated using the bag-of-words[4] representation for documents (using character 3-grams as terms). The considered classifiers are naive Bayes, support vector machine (SVM), neural network (neural), random forest (RF) and 1-nearest neighbor (KNN). Table 3 shows the results of the comparison for the data sets described in Section 4.1.

Table 3. Experimental results of different AA methods for the data sets described in Table 1.

	CNG-SPI	**CNG-WPI**	**Naive**	**SVM**	**Neural**	**RF**	**KNN**
Football	91.11%	**93.34%**	88.89%	91.11%	91.11%	84.44%	77.78%
Business	77.78%	80.00%	**82.22%**	81.11%	77.78%	71.11%	50.00%
Travel	71.67%	73.33%	70.00%	**75.00%**	74.00%	**75.00%**	55.00%
Cricket	88.33%	90.00%	93.33%	**98.33%**	90.00%	83.33%	70.00%
Poetry	78.18%	**85.45%**	74.55%	60.00%	65.45%	40.00%	27.27%
CCAT	64.00%	73.60%	73.60%	**79.00%**	76.80%	73.00%	63.60%
Twitter	53.20%	58.20%	**60.80%**	52.60%	58.20%	N/A	26.40%
Avg.	74.90%	**79.13%**	77.63%	76.74%	76.19%	71.15%	52.86%

We can see from this table that the CNG-SPI method is a very effective AA method when compared with classification-based approaches, however, CNG-WPI is more effective. It is worth mentioning that the performances achieved by the different classifiers are comparable to that reported in related works that have used the same data sets [6, 11, 15, 17]. The proposed method outperforms the different classification based approaches in different data sets. CNG-WPI outperforms the naive Bayes classifier in 3 out of 7 data sets and these methods tie in one data set. The proposed method outperforms the neural net classifier in a similar way. CNG-WPI outperforms KNN in all of the data sets, and it outperforms random forest in 6 out of the 7 data sets. It is interesting that CNG-WPI even outperforms an SVM (the most used classifier in classification-based AA [4, 15, 6]) in 3 out of the 7 data sets. On average (last row in Table 3) the CNG-WPI method obtained the best performance among the considered methods. Giving evidence of its suitability for the AA task.

One should note that though CNG-WPI did not outperform the other methods in all of the considered data sets, PBAA methods are advantageous over

[4] We also performed preliminary experiments with the tf-idf representation for documents, although we found that the performance of most of the considered classifiers was worse than that obtained when the boolean bag-of-words was used.

classification-based approaches as they are more informative in terms of interpretability. More important, PBAA methods are much more efficient. Also, one should note that we have used only lexical features for representing documents with CNG-WPI. We would like to evaluate the performance of the proposed formulation using other types of features.

5 Conclusions

We have described CNG-WPI[5], a prototype-based authorship attribution method based on a new similarity measure, the weighted profile intersection (WPI). Under PBAA, an author's profile is created by combining information extracted from sample documents written by the author of interest. An unseen document is associated with the author whose profile is most similar to the document. Traditional PBAA methods consider the number of overlapping terms as similarity measure. The proposed similarity measure incorporates a term-weighting factor that accounts for the usage of terms across profiles for different authors. Terms shared by several authors receive a lower weight than those terms that are used by one (or a few authors).

We performed experiments with several data sets previously used for the evaluation of AA methods, including multiple subsets of a Twitter corpus. Experimental results show that the proposed approach outperforms the standard PBAA method for most of the considered settings. The improvement was consistent for several values of L and different data set sizes, evidencing the effectiveness of the proposed method. Furthermore, we compared the performance of the proposed method to that obtained by classification-based AA methods. We found that the proposed method outperforms classifier-based techniques for different data sets. Besides achieving comparable performance, PBAA methods are advantageous over classification-based methods in terms of interpretability and efficiency. Thus showing the suitability of the proposed approach to AA.

Future work directions include the use of the proposed method with other type of features (e.g., syntactical, grammatical or semantic), as well as studying the relationship between L and the number of training documents and between L and the length of documents. Also, we would like to evaluate other term-weighting factors (e.g., inverse-document frequency and related weighting factors) for weighting the intersection of terms in PBAA. Additionally, it would be interesting to evaluate the impact of feature (term) selection in the construction of author profiles for AA.

Acknowledgements. The authors are grateful with the reviewers for providing helpful suggestions to improve the paper as well as interesting ideas that we will pursue as future work. The first author acknowledges the support of PROMEP under grant 103.5/11/4330. This work was supported in part by the CONACYT-Mexico (project

[5] The code with the implementation of the proposed method, as well as the preprocessed data used for experimentation are available on demand by contacting the authors.

no. 134186) and by the European Commission as part of the WIQ-EI project (project no. 269180) within the FP7 People Programme.

References

1. R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, and P. Rosso. Authorship attribution using word sequences. In *Proceedings of 11th Iberoamerican Congress on Pattern Recognition*, volume 4225 of *LNCS*, pages 844–852, Cancun, Mexico, 2006. Springer.
2. O. de Vel, A. Anderson, M. Corney, and G. Mohay. Multitopic email authorship attribution forensics. In *Proceedings of the ACM Conference on Computer Security - Workshop on Data Mining for Security Applications*, Philadelphia, PA, USA., 2001.
3. J. Demsar. Statistical comparisons of classfifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, January 2006.
4. H. J. Escalante, T. Solorio, and M. Montes. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 288–298. Association for Computational Linguistics (ACL), 2011.
5. G. Frantzeskou, E. Stamatatos, S. Gritzalis, and C. E. Chaski. Identifying authorship by byte-level n-grams: The source code author profile (SCAP) method. *International Journal of Digital Evidence*, 6(1):1–18, 2007.
6. J. Houvardas and E. Stamatatos. N-gram feature selection for author identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, volume 4183 of *LNCS*, pages 77–86, Varna, Bulgaria, 2006. Springer.
7. P. Joula. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233334, 2006.
8. V. Keselj, F. Peng, N. Cercone, and C. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Association for Computational Linguistics*, pages 255–264, Halifax, Canada, 2003.
9. M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60:9–26, 2009.
10. M. Lambers and C. J. Veenman. Forensic authorship attribution using compression distances to prototypes. In *Computational Forensics, Lecture Notes in Computer Science, Volume 5718. ISBN 978-3-642-03520-3. Springer Berlin Heidelberg, 2009, p. 13*, volume 5718 of *LNCS*, pages 13–24. Springer, 2009.
11. R. Layton, P. Watters, and R. Dazeley. Authorship attribution for twitter in 140 characters or less. In *Second Cybercrime and Trustworthy Computing Workshop (CTC)*, pages 1 – 8, Ballarat, Vic. Australia, 2010. IEEE.
12. K. Luyckx and W. Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 513–520, Manchester, UK, 2008. ACM Press.
13. K. Luyckx and W. Daelemans. The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, pages 1–21, August 2010.

14. S. R. Pillay and T. Solorio. Authorship attribution of web forum posts. In *Proceedings of the eCrime Researchers Summit (eCrime), 2010*, pages 1–7, Dallas, TX, USA, 2010. IEEE.
15. S. Plakias and E. Stamatatos. Author identification using a tensor space representation. In *Proceedings of the 18th European Conference on Artificial Intelligence*, volume 178, pages 833–834, Patras, Greece, 2008. IOS Press.
16. M. Potthast, B. Stein, A. Barrón, and P. Rosso. An evaluation framework for plagiarism detection. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 997–1005. Association for Computational Linguistics, August 2010.
17. S. Raghavan, A. Kovashka, and R. Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference, Short Papers*, pages 38–42, Uppsala, Sweden, 2010. Association for Computational Linguistics (ACL).
18. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
19. E. Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.
20. M. Tearle, K. Taylor, and H. Demuth. An algorithm for automated authorship attribution using neural networks. *Literary and Linguist Computing*, 23(4):425–442, 2008.