Authorship Identification Using Dynamic Selection of Features from Probabilistic Feature Set[⋆]

Hamed Zamani, Hossein Nasr Esfahani, Pariya Babaie, Samira Abnar,
Mostafa Dehghani, and Azadeh Shakery

School of Electrical and Computer Engineering, College of Engineering,
University of Tehran, Tehran, Iran
{h.zamani,h_nasr,pariya.babaie,s.abnar,mo.dehghani,shakery}@ut.ac.ir

Abstract. Authorship identification was introduced as one of the important problems in the law and journalism fields and it is one of the major techniques in plagiarism detection. In this paper, to tackle the authorship verification problem, we propose a probabilistic distribution model to represent each document as a feature set to increase the interpretability of the results and features. We also introduce a distance measure to compute the distance between two feature sets. Finally, we exploit a KNN-based approach and a dynamic feature selection method to detect the features which discriminate the author's writing style.

The experimental results on PAN at CLEF 2013 dataset show the effectiveness of the proposed method. We also show that feature selection is necessary to achieve an outstanding performance. In addition, we conduct a comprehensive analysis on our proposed dynamic feature selection method which shows that discriminative features are different for different authors.

Keywords: authorship identification, dynamic feature selection, k-nearest neighbors, probabilistic feature set.

1 Introduction

Authorship identification is an important problem in many fields such as law and journalism. During the last decade, automatic authorship identification was considered as an applicable problem in Computer Science. As a result, many approaches related to machine learning, information retrieval, and natural language processing have been proposed for this purpose [15]. Authorship identification includes two separate problems: authorship attribution and authorship verification, where the latter is the most realistic interpretation of the authorship

[⋆] A simplified version of the approach proposed in this paper participated in PAN at CLEF 2014 Authorship Identification competition. In PAN 2014, we did not consider knee detection technique for feature selection and only selected the best two features. It is worth mentioning that the achieved results on English Novels and Dutch Reviews datasets were promising.

E. Kanoulas et al. (Eds.): CLEF 2014, LNCS 8685, pp. 128–140, 2014.

identification task [9]. Authorship verification has a considerable overlap with plagiarism detection, especially in intrinsic plagiarism detection, where the goal is to determine whether a given paragraph (or a part of a document) is written by a given author or not [6,16].

Many research studies have been conducted on authorship identification and plagiarism detection until now, especially in the "evaluation labs on uncovering plagiarism, authorship, and social software misuse" (PAN)[1]. The authorship identification task of PAN at CLEF 2013 and 2014 focused on authorship verification [2]. The authorship verification task is to verify whether an unknown authorship document is written by a given author or not, when there are only a limited number of documents written by this author are available.

Authorship verification can be simply considered as a binary classification task. However, there are many challenges in this task, such as limited number of positive documents, unbalanced number of negative and positive documents, and existence of many features for each document. Moreover, each author has his/her own writing style and detecting the discriminative features of the writing style of each author is a challenge.

In most of the previous studies in this field, a vector of features is defined for each document [15]. This approach brings two problems. First of all, different features are put together in one vector, and thus the values in a vector are not interpretable and are meaningless in comparison with each other (e.g., a vector containing frequency of stopwords and punctuations). The second problem is about the dimensionality of the feature vector, which may be quite high. In order to solve this problem, feature selection algorithms are used. The output of most of the feature selection algorithms, such as principal component analysis, is a new vector that is the original vector with some eliminated or/and combined cells. Although this solution may decrease the input vector's dimensionality, the result vector is less interpretable for further analysis. The interpretability is essential in this task, especially when the authorship verification problem is designed as a human-in-the-loop system. To tackle these problems, we propose to define a feature set, instead of using a global vector including all feature values. Each element of the feature set is a probabilistic distribution (e.g., the probabilistic distribution of stopwords in the document). Since all features are extracted and stored in a probabilistic format, we do not have to worry about the variability of document lengths. Additionally, there are some well-defined mathematical comparison measures for probabilistic distributions which may perform better than the heuristic ones used extensively in the existing authorship identification methods.

To verify whether an unknown authorship document is written by a given author or not, we use k-nearest neighbors (KNN) technique which can outperform learning-based classification methods, when the amount of training data is limited. In addition, KNN is fast and we can use it repeatedly in our algorithm without worrying about efficiency.

[1] http://pan.webis.de/

Each author has his/her own writing style; hence, in authorship verification, we should focus on the discriminative features for each author. To detect the writing styles, we propose a dynamic feature selection method which uses leave-one-out technique to determine the discriminative features of each author.

We use the dataset of PAN at CLEF 2013 authorship identification task in our evaluations. The experimental results indicate that the proposed method outperforms all the methods presented in PAN at CLEF 2013.

The remaining of this paper is structured as follows: Section 2 reviews the related research studies and Section 3 includes our proposed methodology to verify the authorship of documents. We evaluate our methodology and discuss the results in Section 4, and finally we conclude our paper and illustrate future works in Section 5.

2 Related Work

Two general approaches are mainly used for author verification task: profile-based and instance-based approaches[15]. In the profile-based methods, documents of each author are considered as a single document. Concatenation of each author's documents results in a profile which represents the author's general writing style. Therefore, the concatenated document is used for extracting the writing style features. The likelihood of the unseen document being written by the given author is determined by a defined distance measure [12].

In the instance-based methods, documents are considered as independent samples. The known authorship instances, represented by a feature vector, are fed to a classifier as the training data. In order to achieve a general and accurate model, each author should have sufficient numbers of sample instances. Hence, in the case of availability of a limited number of instances, possibly long ones, the idea is to segment each document to shorter ones with equal size. Nevertheless, the limited amount of training data continues to be a challenge [11]. It has been proposed that with multiple training instances, each having a different length per author, the documents length must be normalized [13].

Instance-based methods generally include a classifier. The feature selection method, the classification algorithm, and the comparison method affect the performance of the model. Numerosity of attributes within a document's content further add to the importance of both feature extraction and selection. Since feature extraction and selection can solve the problem of overfitness on the training data, considering them potentially can improve the performance significantly [15].

The classification algorithm is chosen based on the application of author identification. Two types of algorithms are generally used in this task, learning-based algorithms, such as neural networks [4,8], SVM [11,13], and Bayesian regression [1,3] and memory-based algorithms, such as k-nearest neighbors [5,17]. Learning-based algorithms require sufficient amount of training data, which is not always available. K-nearest neighbors is used to calculate style deviation score between documents of a known authorship and an unseen document. Based on the calculated score and the defined threshold, the unseen document may belong to this class [5]. In this paper, we also use a KNN-based approach to cope with the authorship verification problem.

3 Methodology

In this section, we first introduce our probabilistic feature set for each document. Then, we propose a distance measure to compute the distance of two different feature sets. After that, we introduce the proposed KNN-based approach for classification and finally we state our feature selection technique which selects the most discriminant features of each author's documents dynamically. Dynamic feature selection is used to improve the performance of authorship verification.

3.1 Probabilistic Feature Set

In the author verification task, features should be defined such that they can discriminate authors' writing styles. Several previous studies have introduced features which represent the authors' writing styles using a single number (e.g., number of different words); however, these features are not highly effective. It is notable that the previous methods store all of the features in a single vector. This kind of feature gathering suffers from lack of interpretation. In other words, when all features (e.g., lexical, stylish, and content-based features) are stored in a single vector, analysing the features is difficult. In addition, all the features are counted as equally important. Another point is that when feature selection techniques are applied on a single feature vector, the result may be meaningless; because some features (e.g., stopwords) may contain more than one value in the feature vector and feature selection techniques may omit some of them.

To tackle the mentioned problems, we store a set of features where each of the elements is a probabilistic distribution of one of the defined features. In other words, we define a feature set for each document, which includes probabilistic distributions. The probabilistic distribution of feature F are estimated using Maximum Likelihood Estimation (MLE) for each feature element f, such that the probability of each feature element is calculated as follows:

$$p(f|d) = \frac{count(f,d)}{\sum_{f' \in F} count(f',d)} \tag{1}$$

where $count(f, d)$ indicates the frequency of feature element f in document d. In other words, f is one of the elements of probabilistic distribution F. These features are defined below:

1. **Probabilistic distribution of stopword usage**: Each element of this feature indicates the percentage of using a specific stopword. This feature is almost independent of the topic the author is writing about. Hence, it could show the writing style of the authors.

2. **Probabilistic distribution of punctuation usage**: Each element of this distribution shows the percentage of using a given punctuation mark. It is obvious that this feature is almost independent of the context.
3. **N-gram probabilistic distribution**: This feature shows the usage frequency of N-grams in the content. This feature can help us detect the phrases which are frequently used by an author. In addition, this feature can be effective when an author always writes about a few subjects.
4. **Probabilistic distribution of sentence length**: Sentence length is referred to the number of words used in a given sentence. Since, complex sentences are longer, this feature shows the writing complexity of each author.
5. **Probabilistic distribution of paragraph length**: Paragraph length can be considered as one of the discriminative features of writing styles. Thus, this feature is defined as the number of sentences which are used in a paragraph to obtain the distribution of the paragraph length.
6. **Part of Speech (POS) tag probabilistic distribution**: POS tags distribution shows how much a given author uses each POS tag. The intuition behind this feature is that the pattern of using POS tags is a good representation of the grammatical manner in writing. Since grammar is one of the key features of the writing styles, exploiting this feature could be beneficial.
7. **Word length probabilistic distribution**: The i^{th} element of this feature shows how many of the words in a document contains exactly i characters. Long words are commonly more complex than the short ones. Therefore, this feature may show the expertise of the author in vocabulary knowledge.

The difference between the size of the documents may cause some problems during the authorship verification process. For instance, long documents contain more stopwords. Since all of the mentioned features are probabilistic distributions, they do not depend on the document's length.

It is worth mentioning that we will exclude some of the features during the feature selection phase. To avoid the zero probabilities and to solve the sparseness problem (for further usages in calculating the distance between two feature sets), smoothing methods can be used. In this paper, we use Laplace smoothing approach [10].

3.2 Distance Measure

To measure the distance of two documents, we compare their probabilistic feature sets. This comparison is based on the divergence of two corresponding feature distributions. In other words, if FS_i and FS_j are the feature sets of documents i and j, respectively, the distance of these two feature sets is calculated as:

$$distance(i,j) = \mathcal{F}\ (Dist(FS_i||FS_j)) \tag{2}$$

where $\mathcal{F}$ is a function whose input is inputs a list of numbers and it outputs a single number representing the distance between documents i and j. In Equation 2, $Dist(FS_i||FS_j)$ is a vector whose k^{th} element shows the divergence of

k^{th} probabilistic distributions of feature sets FS_i and FS_j. The k^{th} element of $Dist(FS_i||FS_j)$ is computed as:

$$Dist_k(FS_i||FS_j) = JSD(FS_{ik}||FS_{jk}) \tag{3}$$

where $JSD(FS_{ik}||FS_{jk})$ is the Jensen-Shannon divergence (JS-divergence) of k^{th} distributions of feature sets FS_i and FS_j, which is given by:

$$JSD(FS_{ik}||FS_{jk}) = \frac{1}{2} * D(FS_{ik}||\overline{FS_k}) + \frac{1}{2} * D(FS_{jk}||\overline{FS_k}) \tag{4}$$

where $\overline{FS_k}$ is the average of FS_{ik} and FS_{jk} distributions and D demonstrates the Kullback-Leibler divergence (KL-divergence) [7] between two distributions, that is calculated as:

$$D(P||Q) = \sum_{i=1}^{m} P_i \log \frac{P_i}{Q_i} \tag{5}$$

where P_i and Q_i are the ith elements of distributions P and Q respectively and m is the number of elements in both distributions.

JS-divergence is one of the metrics used to compare two distributions which have commutative property. In this context, it means that the distance of documents A and B is equal to the distance of B and A, which is reasonable and is one of our reasons to use this measure, instead of KL-divergence.

In our experiments, we consider the sum operation as the function $\mathcal{F}$ in Equation 2. In other words, we calculate the divergence of each pair of the corresponding features in two feature sets and consider the summation of them as the distance measure.

3.3 Authorship Verification Using a KNN-Based Approach

The most trivial solution to determined whether a document is written by a given author or not, is using a binary classifier, in which the positive class means that the given author is the writer of the document and the negative class means not. However, there are two main challenges in this solution:

- In most of the cases, the number of features in author identification task is large and classifiers require huge amounts of training data for their learning phase. However, this amount of training data is not always available.
- The number of documents which are not written by an author (negative documents) is extremely larger than the number of documents written by him/her. If we choose all of the negative documents to learn a classifier, unbalance in training data leads to learning a biased model. Although random sampling of negative documents could be a naive solution, it may eventuate wrong results in some cases.

Considering the aforementioned facts, we use an algorithm that is described in Algorithm 1. In this algorithm, we use the k-nearest neighbors (KNN) method.

Algorithm 1: KNN-based Authorship Classification Algorithm

```
Input: The set of documents $D$ written by a given author, the set $L$ including
       the sets of documents of all authors having the same language as the
       given author, and an unknown document $d_u$
Output: The estimated probability that the document $d_u$ is written by the
        given author
$p \leftarrow 0$
foreach $D' \in L$ & $D' \neq D$ do
    $k \leftarrow \min(|D|, |D'|)$
    $C \leftarrow \{k \text{ nearest documents of } \{D \cup D'\} \text{ to } d_u\}$
    if $|C \cap D| > k/2$ then
        $p = p + 1/(|L| - 1)$
    end
end
return $p$
```

To verify that a document is written by author A or not, we take all of the other authors into account. In each step, we consider the documents of author A as instances of the positive class and the documents of one of the other authors (i.e. author B) as instances of the negative class and determine the class of the unknown document using KNN. We set the parameter k of KNN algorithm as the minimum of number of documents written by A and number of documents written by B. We repeat this procedure with other authors as class B and calculate in how many of them, the unknown document is assigned to A.

The output of Algorithm 1 is the fraction of times the unknown authorship document is assigned to the questioned author. If the output of Algorithm 1 is greater than a threshold, we decide that the unknown document is written by A.

3.4 Dynamic Feature Selection

The idea behind dynamic feature selection is that discriminative features could be different for each author. For example, there could be an author that uses stopwords with a special writing style, but uses punctuation like other authors. Thus, for this author we need to emphasize stopword feature instead of considering all features similarly. Therefore, unlike previous methods that use all features to verify authors, we try to select discriminative features for each author. Dynamic feature selection consists of two main parts. First, we assign a score to each feature for each author and then we decide how many features we should use for each author and select the high score features. In the following, we describe these two parts in detail.

Assigning Score to Each Feature for Each Author. In feature selection, we consider each element of the probabilistic feature set individually. In other words, we assume that the features are independent. Although, this assumption is not always true, but it helps us have a faster algorithm; since, without this assumption, we should consider all subsets of the feature sets and select the most discriminant one, which is extremely time-consuming and costly. Assume that C_j is a classifier that only uses the j^{th} feature. Using a single feature can help us understand the effectiveness of each feature individually, considering the independence assumption between features.

To select the top most discriminative features, we try to assign a score S_j to each feature. In order to calculate this score we apply the leave-one-out technique on the documents of each author (the known authorship documents). As an example, suppose that an author has k known documents $\{d_1, d_2, ..., d_k\}$. For each document d_i, we exclude d_i from the known document set and consider it as an unknown authorship document. Next we apply every C_j on the new unknown document (d_i). Each C_j will return a score between zero and one, indicating the probability of considering this document as a relative document to the corresponding author. Score zero means that this unknown document is completely irrelevant to this author and score one means that this new unknown authorship document is definitely relevant to the author. As we know that this document was written by this author, we expect that the score would be close to one. Therefore, the higher the score returned by the classifier C_j, the more effective and discriminative the jth feature will be. Hence, we add the score returned by C_j to S_j. S_j is calculated as:

$$S_j = \sum_{i=1}^{k} C_j(d_i) \tag{6}$$

$C_j(d_i)$ donates applying our classification using only j^{th} feature where the unknown authorship document is d_i. Hence, we assign a score to each document for each author. Note that since, S_js are independents, we can parallelize the calculation of these scores to increase the efficiency.

Selecting Effective Features. To select the effective features, we use knee detection as described as follows: First we sort all S_js in descending order so that $S_o(l)$ means the l^{th} greatest element of all scores. Then, we find the l^* as follow:

$$l^* = \arg\max_{1 \leq l < k} \{S_o(l)/S_o(l+1)\} \tag{7}$$

This means that if we order features by their scores, the distance between l^{*th} feature and $(l^*+1)^{th}$ feature is greater than any other adjacent scores. We select the first l^* features as the most effective features and use them in classification.

After selecting effective features, we apply our classifier on the unknown authorship document, using only selected features and get the assigned score for the unknown authorship document by the classifier. This score is the final score for relativeness of this document to the corresponding author.

4 Experiments

In this section, we first describe the dataset which is used in our the experiments and then we briefly describe the experimental setup. We report the experimental results and discussions. It should be noted that we use F1-measure as the evaluation metric in our experiments.

4.1 Dataset

We use author identification dataset provided by the 9^{th} evaluation lab on uncovering plagiarism, authorship, and social software misuse (PAN) at CLEF[2] 2013. The dataset includes a number of questions each containing up to 10 documents from an author and exactly one document with unknown authorship. The goal is to determine whether the unknown document is written by the given author or not. All documents of each question are in one of the English, Spanish, and Greek languages. The dataset is separated into two different parts:

- **Training set**: This part of dataset is provided for training the proposed models and parameter tuning. Training set contains 10 questions in English, 5 questions in Spanish, and 20 questions in Greek.
- **Evaluation set**: In the evaluation set, there are 30 questions in English, 25 questions in Spanish, and 30 questions in Greek. This part of the dataset was used in the final evaluations of PAN 2013.

It is notable that we have made the assumption that authors of known authorship documents are not same person in different questions. Therefore, in processing each question, we consider the known authorship documents of other questions as negative instances.

4.2 Experimental Setup

In our experiments, we divide each document to two separate ones to increase the number of documents and to avoid the overfitting problem. To extract the features from documents, we use Apache OpenNLP toolkit[3] in our experiments for sentence detection, tokenization, and also POS tagging. Since we do not have access to POS taggers in Spanish and Greek, we avoid this feature for these two languages. In addition, we do not apply any text normalization technique on the texts and consider them in their original format. It is noteworthy that we have considered $n = 2$ for the N-gram feature and for the stopword distribution feature, we have used a standard set of stopwords for each language, containing around 500 words in that language.

[2] Conference and Labs of the Evaluation Forum.

[3] `https://opennlp.apache.org/`

4.3 Results and Discussion

In this subsection, we first demonstrate the effectiveness of the proposed feature selection technique. Then we discuss about the selected features in different languages on the evaluation set. After that we compare our results with the results of PAN 2013 winners.

Feature Selection. In order to investigate the effectiveness of our proposed feature selection method, we compare different feature selection methods. In Table 1, "w/o FS" refers to using all defined features without any feature selection. "Top1", "Top2", and "Top3" refer to using the one, two, and three features with the highest scores, respectively. "KD" is used for the dynamic feature selection using knee detection technique described in Subsection 3.4. Table 1 reports F1-measure for the mentioned feature selection methods on the evaluation set.

According to Table 1, dynamic feature selection achieves the highest score in all languages and its results are equal to Top3 and Top2 in Spanish and Greek languages, respectively. Table 1 also shows that the feature selection is necessary for author detection when we exploit a KNN-based classifier. Also, the results of knee detection technique shows the effectiveness of this method in selecting the best features among all defined features.

Table 1. Performance of different feature selection methods on the evaluation set in terms of F1-measure

	Overall	English	Spanish	Greek
w/o FS	0.635	0.700	0.760	0.466
Top1	0.611	0.766	0.760	0.333
Top2	0.717	0.766	0.720	**0.666**
Top3	0.705	0.700	**0.840**	0.600
KD	**0.776**	**0.833**	**0.840**	**0.666**

Table 2. Selected features frequency

	English	Spanish	Greek
stopwords	18	15	11
punctuation	7	17	16
N-gram	30	19	30
sentence length	5	8	5
paragraph length	0	8	4
POS tag	5	–	–
word length	7	8	12
average	2.4	3	2.6

Table 3. Comparison with winners of PAN 2013 in terms of F1-measure

	Overall	English	Spanish	Greek
KNN-DFS	**0.776**	**0.833**	**0.840**	0.666
IM	0.753	0.800	0.600	**0.833**
KNNE	0.718	0.700	**0.840**	0.633

Selected Features. In this part, we take a closer look at the selected features in different languages. Table 2 demonstrates how many times each aforementioned feature is selected using dynamic feature selection technique. According to Table 2, N-gram probabilistic distribution is the most selected feature in all languages. This feature is selected in all English and Greek questions. Another discriminant feature for most English documents is stopwords distribution. This feature is also one of the discriminative ones in Spanish.

Another point which is shown in Table 2 is that we cannot select a specific set of features for a given language as the discriminative feature set. It demonstrates that each author may have his/her own writing style and the features cannot be selected for each language generally. In addition, the average number of features selected for each question is less than half of all features. This shows the importance of feature selection in authorship verification.

Comparison with Bests of PAN 2013. We compare our results on the evaluation set with the winners of PAN at CLEF 2013 competition in each language. Impostors Method (IM) [14] was the winner of PAN and achieved the best results on English and Greek languages. k-nearest neighbors estimation (KNNE) [5] also had the best results on Spanish language in the final evaluation. Table 3 shows the comparison of the proposed method with IM and KNNE.

As shown in Table 3, KNN-DFS outperforms the best results of PAN on English and also its result in Spanish is equal to the best results of PAN. However, KNN-DFS has lower F1-measure in comparison with the winners of PAN in Greek. The reason may be the lack of text normalization and pre-processing.

5 Conclusions and Future Work

In this paper, we proposed a novel probabilistic feature set to model the features of each document. We further introduced a distance measure to compare two different feature sets and proposed a KNN-based approach to verify the authorship of unknown authorship documents. A dynamic feature selection technique was also used to detect the discriminant features per each author.

We evaluated our approaches on PAN at CLEF 2013 dataset. The experiments showed that the proposed method outperforms the approaches proposed by the winners of PAN at CLEF 2013 in terms of F1-measure. Also, we showed that the proposed feature selection technique can improve the results significantly. In our experiments, N-gram probabilistic distribution was selected as the most discriminant feature, especially in English and Greek languages. We illustrated that each author has his/her own writing style and feature selection should be based on each author, not the languages.

Future research studies can focus on weighting the features for each author. In other words, in addition to selecting some features, a weight can be assigned to each author and these weights can be considered in $\mathcal{F}$ function used in the defined distance measure. Moreover, defining effective features (e.g., some language-dependant features) may improve the performance.

References

1. Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Commun. ACM 52(2), 119–123 (2009)
2. Forner, P., Navigli, R., Tufis, D. (eds.): CLEF 2013 Evaluation Labs and Workshop–Working Notes Papers (2013)
3. Genkin, A., Lewis, D.D., Madigan, D.: Large-scale bayesian logistic regression for text categorization. Technometrics 49, 291–304 (2007)
4. Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Nat. Lang. Eng. 11(4), 397–415 (2005)
5. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship verification via k-nearest neighbor estimation - notebook for pan at clef 2013. In: Forner et al [2]
6. Joula, P., Stamatatos, E.: Overview of the author identification task at pan 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. vol. 8138 (2013)
7. Kullback, S., Leibler, R.A.: On information and sufficiency. Annals of Mathematical Statistics 22, 49–86 (1951)
8. Li, J., Zheng, R., Chen, H.: From fingerprint to writeprint. Commun. ACM 49(4), 76–82 (2006)
9. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics, COLING 2008, pp. 513–520 (2008)
10. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval. Cambridge University Press, New York (2008)
11. Mohtasseb, H., Ahmed, A.: Two-layered blogger identification model integrating profile and instance-based methods. Knowl. Inf. Syst. 31(1), 1–21 (2012)
12. Potha, N., Stamatatos, E.: A profile-based method for authorship verification. In: Likas, A., Blekas, K., Kalles, D. (eds.) SETN 2014. LNCS, vol. 8445, pp. 313–326. Springer, Heidelberg (2014)
13. Sanderson, C., Guenter, S.: Short text authorship attribution via sequence kernels, markov chains, and author unmasking: An investigation. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, EMNLP 2006, pp. 482–491 (2006)
14. Seidman, S.: Authorship verification using the impostors method - notebook for pan at clef 2013. In: Forner et al. [2]
15. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. 60(3), 538–556 (2009)

16. Stamatatos, E., Koppel, M.: Plagiarism and authorship analysis: introduction to the special issue. Language Resources and Evaluation 45(1), 1–4 (2011)
17. Zhao, Y., Zobel, J.: Searching with style: Authorship attribution in classic literature. In: Proceedings of the Thirtieth Australasian Conference on Computer Science, ACSC 2007, pp. 59–68 (2007)