# Authorship Attribution for Forensic Investigation with Thousands of Authors

Min Yang, Kam-Pui Chow

# Authorship Attribution for Forensic Investigation with Thousands of Authors

Min Yang, Kam-Pei Chow

Department of Computer Science, Faculty of Engineering
The University of Hong Kong, Hong Kong
{myang,chow}@cs.hku.hk

**Abstract.** With the popularity of computer and Internet, a growing number of criminals have been using the Internet to distribute a wide range of illegal materials and false information globally in an anonymous manner, making criminal identity tracing difficult in the cybercrime investigation process. Consequently, automatic authorship attribution of online messages becomes increasingly crucial for forensic investigation. Although researchers have got many achievements, the accuracies of authorship attribution with tens or thousands of candidate are still relatively poor which is generally among 20%~40%, and cannot be used as evidence in forensic investigation. Instead of asserting that a given text was written by a given user, this paper proposes a novel authorship attribution model combining both profile-based and instance-based approaches to reduce the size of the candidate authors to a small number and narrow the scope of investigation with a high level of accuracy. To evaluate the effectiveness of our model, we conduct extensive experiments on a blog corpus with thousands of candidate authors. The experimental results show that our algorithm can successfully output a small number of candidate authors with high accuracy.

## 1 Introduction

With the development of Internet technologies and online social network, web services (e.g., emails, blogs, forums and micro-blogs) become a means by which new ideas and information spread rapidly. However, since people on the virtual space do not need to provide their real identities, accurate automatic authorship attribution of anonymous documents is increasingly requisite in various cybercriminal scenarios, including online fraud detection, terror message origination, article counterfeit and plagiarism detection. Authorship attribution techniques can assist law enforcement to discover criminals who supply false information in their virtual identities, and collect digital evidence for cybercrime investigation.

Automatic authorship attribution is a problem of computationally inferring the author of an anonymous text or text whose authorship is in doubt[15]. Generally, authorship attribution approaches fall into two major categories: the profile-based approaches and the instance-based approaches[3,4,14,22]. For the profile-based approach, a single representation (i.e., profile) is produced for each author

using training data. Each text of unknown authorship is then compared with the profile of each author and is assigned to the most likely one. The profile-based approach is able to handle very short texts since they concatenate all the texts by the same author. For the instance-based approach, it produces one representation per training text, and a classification model is built to estimate the most likely author of a anonymous text. Compared with the profile-based approach, the instance-based approach is easier to combine different text representation features, and it is more robust when the size of candidate authors set is large. Further study[12] shows that profile-based and instance-based approaches could be complementary to each other, and combining these two approaches can significantly improve the performance of author attribution.

Most of the previous works on authorship attribution focus on formal texts with only a few possible authors by applying statistics [3] and machine learning approaches [4,14,22]. Unfortunately, this version of the authorship attribution problem dose not often arise in the real world. Recently, researchers have turned their attention to informal texts (e.g., emails and social blogs) and tens to thousands of authors [17,11,19,16,12,10,20]. For example, Koppel et al. [10] use similarity-based methods along with multiple randomized feature set to achieve high precision when the set of known candidates is extremely large (many thousands). To conquer the challenge that there are not enough labeled examples to construct an accurate classifier, some semi-supervised learning approaches for authorship attribution have been proposed. These approaches usually use the test data as unlabeled examples to improve the classification model [7,12]. Considering the content of documents and the interests of authors, the author-topic model has been used for authorship attribution [20], which yield a state-of-the-art performance in terms of classification accuracy when tens or thousands of candidate authors are taken into account.

Although researchers have got many achievements, the accuracies of authorship attribution with tens or thousands of candidate are still relatively poor, which are generally among 20%~40% [16,10,20]. The traditional authorship attribution approaches cannot be used as evidence in forensic investigation since a quite accurate prediction is required for digital forensics, and it is useless in the sense that we could never confidently assert that a given text was written by a given user. In this paper, we propose a novel hierarchical authorship attribution algorithm combining both profile-based and instance-based approaches to reduce the size of the candidate authors and narrow the scope of investigation with a high level of accuracy. First, we apply profile-based paradigm to build two classifiers with different feature sets for gender and age attribution respectively. And then, a authorship attribution classifier is built using the probabilities distribution of gender and age attribution obtained in the previous step as prior. With this classifier, we can obtain a small number of the most possible authors from thousands of candidate authors with a high level of accuracy that is higher than a threshold (e.g., 95%). At last, we output a set of most possible authors with probabilities. Extensive experiments have been conducted to verify the proposed approach on real-world blog dataset.

The rest of this paper is organized as follows. In Section 2, we review related work in authorship attribution. In Section 3, we introduce our model and algorithm used in this paper. In Section 4, we presents the experiment data and experimental setting. In Section 5, the experiment results are showed and discussed. Finally, Section 6 concludes the paper and indicates some future works.

## 2 Related Work

In resent years, plenty of statistical and machine learning techniques have been proposed for authorship attribution using different kinds of features [13,1]. Stamatatos details most of the existing techniques for automatic authorship attribution in [21]. Generally speaking, authorship attribution approaches can be classified into two groups: profile-based approaches and instance-based approaches. Profile-based approaches concatenate training texts per author in one single text file. An unseen text is, then, compared with each author file, and the most likely author is estimated based on a distance measure. For example, Keselj, et al.[9] propose a widely used n-grams profile-based method, which is based on building a byte-level n-gram author profile of an authors writing. Frantzeskou et al. [6] proposed a novel and simple distance, called simplified profile intersection (SPI), which simply counts the amount of common n-grams of the two author profiles. This approach to authorship identification of source code provide better results than other distances. In order to utilize the differences between the training texts by the same author, the majority of the modern authorship attribution approaches are instance-based, which consider each training text sample as a unit that contributes separately to the authorship attribution. Such as, Burrows [2] presents principal components analysis with word frequencies to analyze authorship, and the results show a high level accuracy. Peng et al. [18] extend the naïve Bayes algorithm for authorship attribution with statistical language models. Halteren [8] proposes a method that borrows some elements from both profile-based and instance-based approaches.

To conquer the challenges that there are not enough labeled examples to construct an accurate classifier, some semi-supervised learning approaches have been proposed since it is possible to use the test sets as unlabeled examples and use some information from them to improve the classification model [7,12]. Guzman et al. [7] propose a self-learning method that is specially suited to work with just a few training examples to tackle the problem that lacks of training data with the same writing style. That method considers the automatic extraction of the unlabeled examples from the web and its iterative integration into the training data set. Kourtis and Stamatatos [12] apply a co-training learning approach for authorship attribution by combining the Common N-Grams (CNG) [9] model and a Support Vector Machine classifier based on character n-grams. Its main idea is to combine the outputs of these classifiers in the test set and augment the training set with additional document.

Most of the previous works consider the simple version of the authorship attribution problems, and focus on formal texts with only a small, closed set

of candidate authors. In order to apply authorship attribution in real life data, some large candidate sets with informal texts have been considered by researchers recently. Luyckx and Daelemans [16] propose a memory-based learning approach in doing authorship attribution with many authors and limited training data, and the results show the robustness of the memory-based learning approach when compared to eager learning methods such as SVMs and maximum entropy learning. Madigan et al. [17] conduct experiments on a collection of data released by Reuters consisting of 114 authors using sparse Bayesian logistic regression. This proposed algorithm shows promising performance as a tool for authorship attribution with high-dimensional document representations. Different from the approaches mentioned above, our algorithm take authors profile information (e.g., age and gender) into consideration, which is motivated by the analysis conducted by Schler et al. in [19] indicating significant differences in writing style and content between male and female bloggers as well as among authors of different ages. In addition, instead of asserting that a anonymous text was written by a given user, our authorship attribution algorithm proposes to reduce the size of the candidate authors and narrow the scope of investigation with a high level of accuracy.

## 3 The Proposed Method

In this section, we describe a novel authorship attribution algorithm combining both profile-based and instance-based approaches to reduce the size of the candidate authors and narrow the scope of investigation with a high level of accuracy. Generally, the proposed algorithm consists of two phase, as shown in Figure 1. In the first phase, we apply profile-based paradigm to build two classifiers with different feature sets for gender and age attribution, respectively. The probability distribution of gender and age can be used as a prior for the next phase. In the second phase, a logistic regression classifier is built using the probability distribution of gender and age as prior based on instance-based paradigms. With this classifier, we can obtain a small number of the most possible authors from thousands of candidate authors with a high level of accuracy that is higher than a threshold.

### 3.1 Features Selection

In this work, we consider differences in male and female authors and differences among authors of different ages. Broadly speaking, two different kinds of potential distinguishing features can be considered: content-based features and style-based features . This is motivated by the observation that different people might tend to write about different topics as well as to express themselves differently about the same topic. For style-based features, we consider individual parts-of-speech and function words, which is described by Eggins in [5]. Content-based features are simple content words, and we apply unigrams in this work. For gender, we choose 2000 features with greatest information gain for gender.

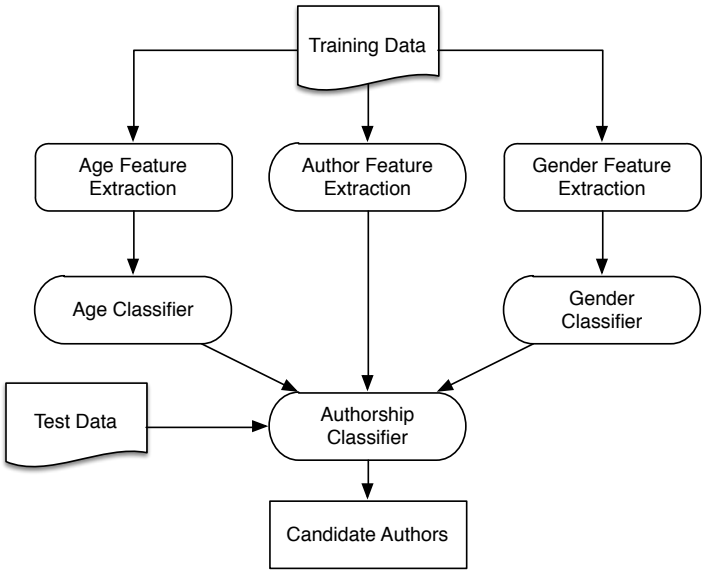


**Fig. 1.** Algorithm Overview

And for age, we choose 2000 features with greatest information gain for age. Similarly, 2000 features with greatest information gain for author are selected to train the classifier for authorship attribution.

### 3.2 Age and Gender Attribution

For each author, we concatenate all his documents and extract both content-based features and style-based features from them. To present the document with extracted features, we represent a document as a numerical vector $X = (x_1, \ldots, x_i, \ldots, x_n)$, where n is the number of features and $x_i$ is the relative frequency of feature $i$ in the document. Once labeled training documents have been represented in this way, we can apply machine-learning algorithms to learn classifiers that assign new documents to categories. In this paper, we use logistic regression [13] to learn two classifier that classify texts according to author's gender and age, respectively. Logistic regression is widely used for classification problems recently, and the independent variables do not have to be normally distributed, or have equal variance in each group. Whats more, logistic regression is more than just a classifier. Instead, it can make stronger and more detailed predictions such as the predicted probabilities. Consider a binary supervised leaning problem, where were given a set of instance-label pairs $\{(x_i, y_i) \,|\, i = 1, \ldots, n\}$. Here, $x_i \in R^M$ is an $M$-dimensional feature vector, and $y_i \in \{0, 1\}$ is the class label. Formally, the logistic regression model which predicts the conditional probability distribution of the class label y given the input feature

vector x is that

$$\Pr(Y = 1|X; \beta_0, \beta) = \pi(x; \beta_0, \beta) = \frac{1}{1 + \exp(-(\beta_0 + x \cdot \beta))} \tag{1}$$

Where $\beta \in R^M$ is the coefficient vector of $X$, and $\beta_0$ is the intercept term. $\pi(x; \beta_0, \beta)$ is the probability of the outcome of interest. For multi-class classification problem, we can still use logistic model. Assume $Y$ can take $k$ values, instead of having one set of parameters $\beta_0$ and $\beta$ , each class $c$ will have its own parameters $\beta_0^{(c)}$ and $\beta^{(c)}$ , then the predicted probability distribution will be

$$\Pr(Y = 1|X; \beta_0^{(c)}, \beta^{(c)}) = \pi(x; \beta_0^{(c)}, \beta^{(c)}) = \frac{1}{1 + \exp(-(\beta_0^{(c)} + x \cdot \beta^{(c)}))} \tag{2}$$

### 3.3 Authorship Attribution

After we obtained the estimator for gender and age, we can use the predicted probabilities for gender and sex as a prior for authorship attribution.

Given documents $\boldsymbol{S}$ for an unknown author, given the information of all candidate authors' gender and age information, the probability distribution for authorship attribution is given by

$$p\left(author = i|\boldsymbol{S}, gender, age\_class\right) \propto p\left(author = i|\boldsymbol{S}\right) \cdot p\left(gender(i)|\boldsymbol{S}\right) \cdot p\left(age\_class(i)|\boldsymbol{S}\right) \tag{3}$$

where the first term acts like a likelihood, and the second and third terms served as priors.

$p\left(author = i|\boldsymbol{S}\right)$ is estimated with an instance-based approach, i.e. we estimate the likelihood for each instance documents separately, and multiply the probabilities together to get the combined probability

$$p(author|\boldsymbol{S}) \propto \prod_i p(author|S_i) \tag{4}$$

## 4 Experiments

### 4.1 Dataset description

The Blog Authorship Corpus consists of 678,161 blog posts of 19,320 bloggers gathered by Schler et al. [19] from blogger.com in August 2004. There are approximately 35 posts and 7250 words per person. The blog posts can be about any topic, but the large number of authors ensures that every topic is likely to interest at least some authors. The (self-reported) age and gender of each author is known and for each age interval the corpus includes an equal number of male and female author based on the authors reported age, we label each blog in our corpus as belonging to one of three age groups: 13-17 (42.7%), 23-27 (41.9%) and 33-47 (15.5%).

**Algorithm 1** Hierarchical Author Identification

- Given $N$ documents written by $K$ authors $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ and their authorship labels $\boldsymbol{y} = \{y_1, \ldots, y_N\}$, where $y_i \in \{1, \ldots, K\}$.
- For each author $i \in \{1, \ldots, K\}$, we have $gender(i) \in \{M, F\}$ and $age\_class(i) \in$ {teenagers, young adults, middle-aged}.
- Gender attribution
  - Extract features $\boldsymbol{X}^{(g)}$ for gender classifier using stylistic features and content-based features selected by information gain for gender
  - Use $\boldsymbol{X}^{(g)}$ and $\boldsymbol{y}^{(g)} = \{gender(y_1), \ldots, gender(y_N)\}$ to train a gender classifier with profile-based paradigm
  - Given documents $\boldsymbol{S}$ from an unknown author, the gender classifier can estimate the probability distribution $p(gender|\boldsymbol{S})$ for $gender \in \{M, F\}$.
- Age class attribution
  - Extract features $\boldsymbol{X}^{(a)}$ for age classifier using stylitic features and content-based features selected by information gain for age class
  - Use $\boldsymbol{X}^{(a)}$ and $\boldsymbol{y}^{(a)} = \{age\_class(y_1), \ldots, age\_class(y_N)\}$ to train an age classifier with profile-based paradigm
  - Given documents $\boldsymbol{S}$ from an unknown author, the gender classifier can estimate the probability distribution $p(age\_class|\boldsymbol{S})$ for $age\_class \in$ {teenagers, young adults, middle-aged}.
- Authorship attribution
  - Use $\boldsymbol{X}$ and $\boldsymbol{y}$ to train an authorship classifier using instance-based pradigm
  - Given documents $\boldsymbol{S}$ from an unknown author, the authorship classifier can estimate the probability distribution $p(author|\boldsymbol{S})$ using 4.
  - For any author $i$, calculate thhe posterior probability with author's gender and age class distribution as prior using 3.
  - Output $M$ authors $\{a_1, \ldots, a_M\}$ who have highest posterior probability. The number $M$ of authors to output is chosen to ensure $\sum p(author = a_i) > threshold$.

Since some blogs in this corpus are meaningless for authorship attribution, such as advertisements and lyrics of songs, we first remove the authors who wrote smaller than 20 blogs. And then, we use Akismet[1] to remove the potential spams among these authors. Finally, we obtain 2,077 prolific authors with their full posts as our training data. The statistics of the dataset is shown in Table 1.

### 4.2 Experimental Setup

In the experiments, data preprocessing was performed on both data sets. First, the texts are tokenized with a natural language toolkit NLTK[2]. Then, we remove non-alphabet characters, numbers, pronoun, punctuation and stop words from

[1] http://akismet.com/
[2] http://www.nltk.org

| Age | Gender | | |
|---|---|---|---|
| | Female | Male | Total |
| 13~17 | 335 | 348 | 683 |
| 23~27 | 543 | 503 | 1046 |
| 33~47 | 170 | 178 | 348 |
| Total | 1048 | 1029 | 2077 |

**Table 1.** Dataset statistics

the texts. Finally, WordNet stemmer[3] is applied so as to reduce the vocabulary size and settle the issue of data spareness.

For all experiments, we perform ten-fold cross validation, and the results are evaluated using classification accuracy, i.e., the percentage of test documents that were correctly assigned the author. What's more, L1-regularized logistic regression is used to train the classifiers, which is well-suited for large-scale text classification. Here, LIBLINEAR[4] is used as the implementation of logistic regression classifier. We experiment with cost parameter valued from the set $\{0.01, 0.1, 1, 10\}$, until no accuracy improvement was obtained.

## 5 Experimental Results

In this section, we present and discuss the experimental results in details.

### 5.1 Gender and age attribution results

For gender attribution, we consider it as a binary classification problem with the class label female and male. Accuracies of gender attribution are shown in the first line of Table 2. From Table 2, we observe that the proposed algorithm can obtain 85.1% accuracy with the combined features, which is 3.2% higher than that using style-based features and 5.9% higher than that using content-based features.

For age attribution, we label each blog in our corpus as belonging to one of three age groups {teenagers, young adults, middle-aged}, based on author's reported age. The age attribution problem in this paper can be treated as a multiclass L1-regularised logistic regression problem. Results for age attribution are shown in the second line of Table 2. Similar to gender attribution, we can gain the best performance with 80.1% accuracy using both style and content features.

[3] http://wordnet.princeton.edu
[4] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

Based on the results in Table 2, we can summarize that authors' gender and age information might be highly helpful for authorship attribution because of the high accuracies of gender and age attribution on the blog dataset. With the probability distribution gained from gender and age attribution, most of the candidate authors who have different profiles (i.e., gender and age) with the author of anonymous documents, can be filtered in authorship attribution.

| Task | Style features | Content features | Style and content features |
|---|---|---|---|
| Gender attribution | 81.9% | 79.2% | **85.1%** |
| Age attribution | 70.5% | 78.3% | **80.1%** |

**Table 2.** Accuracies of gender and age attribution using various feature sets

### 5.2 Authorship attribution results

To evaluate the effectiveness of our approach, we compared the proposed algorithm with standard logistic regression and Disjoint Author-Document Topic Model (DADT) used in [20], utilizing the same features. We report the classification accuracies in Table 3. As an easy observation, the proposed hierarchical authorship attribution algorithm significantly outperforms the other algorithms that are widely used for authorship attribution. For example, The accuracy of our algorithm is 3.7% higher than DADT model and 7.4% higher than standard logistic regression, validating the effectiveness of the proposed algorithm.

| Algorithm | Style features | Content features | Style and content feature |
|---|---|---|---|
| Logistic regression | 23.6% | 27.4% | 32.1% |
| DADT | **30.5%** | 31.3% | 35.8% |
| Proposed algorithm | 29.2% | **32.6%** | **39.5%** |

**Table 3.** Accuracies of three authorship attribution algorithm

Although our algotithm achieves state-of-the-art performance compared to the existing model for authorship attribution, it cannot produce a practicable and authentic result for forensics in real life. Hence, instead of classifying a given text to a specific author with relatively low accuracy, we seek to output a small number of potential candidate authors with a high level of accuracy. The experiment result of our proposed hierarchical authorship attribution is

described in Figure 2. With the accuracy of 96.5%, we can narrow down the size of candidates authors from 2,000 to 20. Furthermore, we can get 100% accuracy when choosing 35 candidate authors from the original experimental dataset. These results show that our model can help investigator to narrow the scope of investigation with a high level of accuracy. The advantages of our approach comes from its capability of taking the probability distribution gained from gender and age attribution, as prior knowledge to filter many easily confused candidate authors.

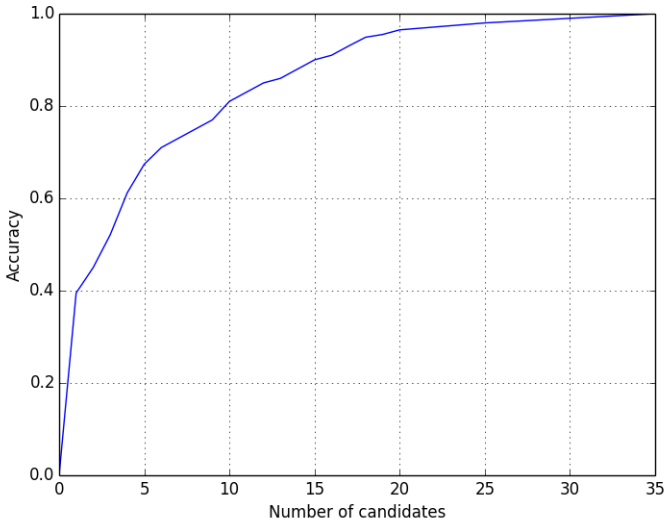


**Fig. 2.** Accuracy with different number of candidate authors

## 6 Conclusions and Future Work

In this paper, we introduce a hierarchical classifier which combines profile-based and instance-based paradigms to automatic authorship attribution in cases with a large number of informal texts from thousands of authors. Instead of predicting a author with relatively low accuracy, we seek to reduce the size of the candidate authors and narrow the scope of investigation with a high level of accuracy. The mainly advantage of our approach comes from its capability of taking the probability distribution gained from gender and age attribution, as prior knowledge to filter many easily confused candidate authors. Extensive experimental results indicate that our algorithm can successfully output a small number of candidate authors with high accuracy. For example, we can choose 20 potential candidate

authors from 2000 candidates with the accuracy of 96.5%, and the probability of each candidate authors can be output to assist the investigators.

Our approach performed well for closed-set tasks, while it cannot be applied to open-set tasks in which the true author of an anonymous text might not be one of the known candidates. In the future, we will devote our effort to extend the proposed hierarchical classifier to handle the open-set tasks by assessing classification confidence. Another possible research direction is to deal with the imbalance problem where a relatively small number of text at least for some candidate authors compared to other candidates.

## References


1. Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE*, 20(5):67–75, 2005.
2. John F Burrows. An ocean where each kind: Statistical analysis and some major determinants of literary style. *Computers and the Humanities*, 23(4-5):309–321, 1989.
3. Ronald R Coifman and M Victor Wickerhauser. Entropy-based algorithms for best basis selection. *Information Theory, IEEE Transactions on*, 38(2):713–718, 1992.
4. Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.
5. Suzanne Eggins. *Introduction to systemic functional linguistics*. Continuum International Publishing Group, 2004.
6. Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, and Sokratis Katsikas. Effective identification of source code authors using byte-level information. In *Proceedings of the 28th international conference on Software engineering*, pages 893–896. ACM, 2006.
7. Rafael Guzmán-Cabrera, Manuel Montes-y Gómez, Paolo Rosso, and Luis Villaseñor-Pineda. A web-based self-training approach for authorship attribution. In *Advances in Natural Language Processing*, pages 160–168. Springer, 2008.
8. Hans Van Halteren. Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1, 2007.
9. Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.
10. Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1):83–94, 2011.
11. Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 659–660. ACM, 2006.
12. Ioannis Kourtis and Efstathios Stamatatos. Author identification using semi-supervised learning. In *CLEF 2011: Proceedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation (Lab and Workshop Notebook Papers), Amsterdam, The Netherlands*, 2011.

13. Su-In Lee, Honglak Lee, Pieter Abbeel, and Andrew Y Ng. Efficient l˜ 1 regularized logistic regression. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 401. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
14. Jiexun Li, Rong Zheng, and Hsinchun Chen. From fingerprint to writeprint. *Communications of the ACM*, 49(4):76–82, 2006.
15. Harold Love. *Attributing authorship: An introduction*. Cambridge University Press, 2002.
16. Kim Luyckx and Walter 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. Association for Computational Linguistics, 2008.
17. David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. Author identification on the large scale. In *Proc. of the Meeting of the Classification Society of North America*, 2005.
18. Fuchun Peng, Dale Schuurmans, and Shaojun Wang. Augmenting naive bayes classifiers with statistical language models. *Information Retrieval*, 7(3-4):317–345, 2004.
19. Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 199–205, 2006.
20. Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, pages 264–269. Association for Computational Linguistics, 2012.
21. Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.
22. Rong Zheng, Yi Qin, Zan Huang, and Hsinchun Chen. Authorship analysis in cybercrime investigation. In *Intelligence and Security Informatics*, pages 59–73. Springer, 2003.