Using Unlabeled Data to Improve Author Identification

R. Guzmán Cabrera, J. R. Guzmán Sepúlveda, J. A. Gordillo Sosa , M. Torres Cisneros, and J. Herrera Cabral

***Abstract*—Authorship attribution may be considered as a text categorization problem. Text categorization requires a large number of training examples which are particularly difficult to obtain in the case of authorship attribution task. In this paper, we investigate the possibility of using Web-based text-mining methods for the identification of the author of a given poem. In particular, we propose a semi-supervised method that is specially suited to work with just few training examples in order to tackle the problem of the lack of data with the same writing style. The results obtained on poem categorization show that this method may significantly improve the classification accuracy and it is appropriate to handle the attribution of short documents.**

***Index Terms*—Authorship attribution, text classification, machine learning.**

I. Introduction

Authorship attribution is the task of identifying the author of a given text. It can be considered as a typical text categorization problem [1], where a set of documents with known authorship are used for training and the aim is to automatically determine the corresponding author of an anonymous text.

Determining the author of a particular piece of text has raised methodological questions for centuries. Questions of authorship can be of interest not only to humanities scholars, but in a much more practical sense to politicians, journalists, and lawyers as in the examples above. Investigative journalism, combined with scientific analysis of documents and simple close reading by experts has traditionally given good results.

But recent developments of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora have made the automatic and objective inference of authorship a practical option.

Applications of authorship attribution include plagiarism detection (i.e. college essays), deducing the writer of inappropriate communications that were sent anonymously or under a pseudonym (i.e. threatening or harassing e-mails), as well as resolving historical questions of unclear or disputed authorship. Specific examples are the Federalist papers [2] and the forensic analysis of the Unabomber manifesto [3].

Within the area of automatic author attribution, recently it has been shown that encouraging performance can be achieved via the use of probabilistic models based on n-grams [4] and Markov chains of characters and words [5]. In [6] SVMs with syntactic and semantic features are used to obtain relatively (minor accuracy) improvements over the use of function word frequencies and part-of-speech trigrams.

The analysis of style for authorship attribution is mainly based on the assumption that each author has habits in wording (i.e., in the use of words) that makes her/his writing unique. However, this assumption is not completely true, since the style of an author may be variable depending on the target audience, or may change because of differences in topics or genre. For this reason, it is difficult to determine a stable set of features to these variations but adequate to distinguish between the writing style of different authors. There are several methods for authorship attribution. These methods may be clustered in the following three main approaches based on:

Stylistic measures. This approach takes into consideration the length of words and sentences as well as the richness of the vocabulary [7,8]. Its results are not conclusive, but have demonstrated that the features taken into account are not sufficient for the task. It seems that they vary depending on the genre of the text, and that they lost most of their meaning when dealing with short texts. In order to measure the quality level of a text in [9] three formulae are introduced to calculate: complexity (which is much related to the nature of a document), variety (which depends mostly on the author style and gives an idea about the variety of expressions), and correctness (which is related to the distribution frequency of the words) of a given text.

In [10] the authors employ the above approach on different kinds of texts (poems, technical reports etc.).

Syntactic cues. This approach uses a set of style markers. These markers go beyond the stylistic measures by integrating information related to the structure of the language, which is obtained by an in-depth syntactic analysis of documents [3]-[11]. Basically, texts are characterized by the presence and frequency of certain syntactic structures. This characterization is very detailed and relevant; unfortunately, it is computationally expensive and even impossible to build for languages lacking of text-processing resources (e.g. POS tagger, syntactic parser, etc.). Besides, it is also clearly influenced by the length of documents.

Words of a document. This approach includes at least three different kinds of methods. The first one characterizes documents using a set of functional words, ignoring the

Manuscript received Ocotber 29, 2012; revised January 28, 2013.

R. Guzmán-Cabrera and M. Torres Cisneros are with the Grupo de NanoBioFotónica, DICIS, Universidad de Guanajuato, Salamanca, Gto., México (e-mail: guzmanc@ugto.mx, mtorres@ugto.mx).

J. R. Guzmán-Sepulveda is with the Departamento de Electrónica, UAM Reynosa-Rodhe, Universidad Autónoma de Tamaulipas, Carr. Reynosa-San Fernando S/N, Reynosa, Tamaulipas 88779, México (e-mail: jrafael_guzmans@yahoo.com.mx)

J. A. Gordillo-Sosa and Joel Herrera Cabral are with the Depto. de TIC. Univ. Tecnológica del Suroeste de Gto. Carr. Valle-Huanímaro km.1.2, Valle de Santiago, Gto. México (e-mail: antgor@antoniogordillo.com, jherrera@utsoe.edu.mx).

A. González Parada is with Universidad de Guanajuato, DICIS, Salamanca, Gto., México (e-mail: gonzaleza@ugto.mx).

content words since they tend to be highly correlated with the document topics [12], [13]. This method works properly, but it is also affected by the size of documents. In this case, the document length not only influences the frequency of occurrence of the functional words but also their sole presence. The second method applies the traditional bag-of-words representation and uses single content-words as document features [2], [7]. It is very robust and produces excellent results when there is a noticeable relation between authors and topics. Finally, a third method considers word n-gram features, i.e., features consisting of sequences of n consecutive words. This method attempts to capture the language structure of texts by simple word sequences instead of by complex syntactic structures [14]. Somehow, its purpose is to obtain a rich characterization of texts without performing an expensive syntactic analysis. Nevertheless, due to the feature explosion, it tends to use only n-grams up to three words.

A major difficulty with this kind of supervised techniques is that they commonly require a great number of labeled examples (training instances) in order to construct an accurate classifier. Unfortunately, because a human expert must manually label these examples, the training data sets are extremely small for many application domains. Therefore, it seems not possible to apply traditional semi-supervised learning for the task of authorship attribution because most of the times it is not easy to obtain texts of the desired writing style (in our case poems).

In order to deal with the problem of the lack of training data, recently in many natural language processing tasks the Web, which has begun a huge repository of information, has been used as a lexical corpus. For instance, in [15] the authors present a headline emotion classification approach based on frequency and co-occurrence information collected from the Web. In [16] Zelikovitz and Kogan proposed a method for mining the Web to improve text categorization by creating a background text set.

The use of information extracted from the Web seems more intuitive in the case of a topic-based text categorization (e.g. news about natural disasters [17]).

For a task such as authorship attribution, which depends on the writing style of an author, the use of information on the Web seems quite inappropriate. In fact, we do not look for snippets of a given author but passages, which have written more or less using the same style.

In this paper, we propose a new method for semi-supervised authorship attribution. This method differs from previous approaches in three main concerns. First, it is specially suited to work with very few training examples.

Whereas previous methods consider groups of hundreds of training examples, our method allows working with just few labeled examples per class. Second, it does not require a predefined set of unlabeled examples: it considers the automatic extraction of related untagged data from the Web. Initially the method employs very few training examples and it does not aim to include a lot of additional information in the training phase: on the contrary, it only incorporates a small group of examples that considerably augment the dissimilarities among classes.

The rest of the paper is organized as follows. Section II shows the proposed method for authorship attribution. Section III presents some evaluation results on a corpus of Mexican poets. Finally, in Section VI we draw some conclusions.

II. Proposed Method

Fig. 1 shows the general scheme of the proposed method. It consists of two main processes. The first one deals with the corpora acquisition from the Web, whereas the second one focuses on the semi-supervised learning problem. The following subsections describe into detail these two processes.

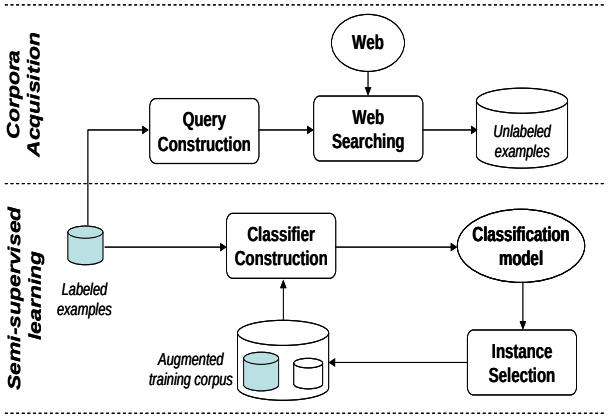


Fig. 1. General overview of the proposed method.

A. Corpora Acquisition

This process considers the automatic extraction of unlabeled examples from the Web. In order to do this, it first constructs a number of queries by combining the most significant words for each class; then, using these queries it looks on the Web for some additional training examples related to the given classes.

Query Construction. In order to form queries for searching the Web, it is necessary to previously determine the set of relevant words for each class in the training corpus. The criterion used for this purpose is based on a combination of the frequency of occurrence and the Information Gain (IG) [1] of words. We consider that a word wi is relevant for class C if:

1) The frequency of occurrence of wi in C is greater than the average occurrence of all words (happening more than once) in that class. That is:

$$f_{w_i}^{C} > \frac{1}{|C'|} \sum_{\forall w \in C'} f_{w}^{C}, \qquad (1)$$

where $C' = \{ w \in C \mid f_{w}^{C} > 1 \}$

2) The information gain of wi with respect to C is positive. That is, if $IG_{w_i}^{C} > 0$. Once obtained the set of relevant words per class, it is possible to construct the corresponding set of queries. Founded on the method by Zelikovitz and Kogan [16], we decide to construct queries of three words. This way, we create as many

queries per class as all three-word combinations of its relevant words. We measure the significance of a query $q = \{w1, w2, w3\}$ to the class C as indicated below

$$\Gamma_C(q) = \sum_{i=1}^{3} f_{w_i}^{C} \times IG_{w_i}^{C} \quad (2)$$

Web Searching. The next action is using the defined queries to extract from the Web a set of additional unlabeled text examples. Based on the observation that most significant queries tend to retrieve the most relevant web pages, our method for searching the Web determines the number of downloaded examples per query in a direct proportion to its Γ-value. Therefore, given a set of M queries $\{q1,\ldots, qM\}$ for a class C, and considering that we want to download a total of N additional examples per class, the number of examples to be extracted by a query $\boldsymbol{q_i}$ is determined as follows

$$\Psi_C(q_i) = \frac{N}{\sum_{k=1}^{M} \Gamma_C(q_k)} \times \Gamma_C(q_i) \quad (3)$$

B. Semi-Supervised Learning

As we previously mentioned, the purpose of this process is to increase the classification accuracy by gradually augmenting the originally small training set with the examples downloaded from the Web. Our semi-supervised learning algorithm is based on the method proposed in [18]. The difference consists in the way the new information is added to the training set at each iteration. More precisely, the information is selected through an array type stacking which is composed of two classifiers (Naïve Bayes and support vector machine) which allow selecting only the best snippets at each iteration of the method.

It is important to point out that the proposed algorithm could be applied in combination with several different classification techniques (e.g., Naïve Bayes, support vector machines, nearest-neighbour, etc.). It mainly considers the following steps:

1) Build a weak classifier (C_l) using a specified learning method (l) and the training set available (T)[1].
2) Classify the downloaded examples (E) using the constructed classifier (C_l). In other words, estimate the class for all downloaded examples.
3) Select the best m examples (Em $\subseteq$ E) based on the following two conditions:
   a) The estimate class of the example corresponds to the class of the query used to download it. In some way, this filter works as an ensemble of two classifiers: C_l and the Web (expressed by the set of queries).
   b) The example has one of the m-highest confidence predictions.
4) Combine the selected examples with the original training set (T $\leftarrow$ T $\cup$ Em) in order to form a new training set. At the same time, eliminate these examples from the set of downloaded instances (E $\leftarrow$ E – Em).
5) Iterate σ times over steps 1 to 4 or repeat until Em = ∅. In this case σ is a user specified threshold.
6) Construct the final classifier using the enriched training set2.

III. Experimental Evaluation

A. Experimental Setup

Corpus. We use the corpus of contemporary Mexican poets used in [19]. This corpus was gathered from the Web. It consists of 353 poems written by five different authors. For the experimental evaluation, we organized the corpus in the training and test sets with 80% and 20% of files by category, respectively. Table I shows some numbers about this collection.

TABLE I: Training/Test Data Sets

Poets	Training Set	Test Set	Vocabulary Size
Efrain Huerta	38	10	2827
Jaime Sabines	64	16	2749
Octavio Paz	60	15	2431
Rosario Castellanos	64	16	3280
Rubén Bonifaz	56	14	3552
Total	282	71	8377

Searching the Web. We used Google as search engine. We downloaded 2,400 additional examples (snippets for these experiments) per class.

Learning methods. We selected two state-of-the-art methods for text classification, namely, Support Vector Machines (SVM) and Naïve Bayes [6], [20].

Evaluation measure. The effectiveness of the method is measured by the classification accuracy, which indicates the percentage of documents that have been correctly classified from the entire document set.

Baseline. Baseline results correspond to the direct application of the selected classifiers on the test data. Table 2 shows these results for different training conditions. They mainly evidence that traditional classification approaches achieve poor performance levels when dealing with few training examples.

B. Experimental Setup

This section presents some results related to the main processes of the proposed method, namely, the corpora acquisition from the Web and the semi-supervised learning approach. The central task for corpora acquisition is the automatic construction of a set of queries which express the relevant content of each class.

Using the automatically constructed queries, we collected from the Web a set of 2,400 snippets per class, obtaining a total of 12,000 additional unlabeled examples. Then, we

[1] Any classification algorithm may be used (in our case we employed an ensemble of Bayes and Support Vector Machine).

Any classification algorithm may be used.

added some of these examples to the original training set. Mainly, we performed three different experiments by varying the size of n-grams which was used as a parameter of classification (1, 2 or 3: respectively, single words, bigrams or trigrams).

It is interesting to point out that thanks to the snippet's small size (that only considers the immediate context of the query's words), the additional examples tend to be lees ambiguous and contain several valuable words that are highly related with the topic at hand.

In the experiments, the downloaded snippets were classified using the original document collection as training set. The best ten examples per class, i.e., those with more confidence predictions, were selected at each iteration and were incorporated to the original training set in order to form a new training collection. In the experiments, we carried out five iterations. Table II shows the accuracy results for all iterations of the experiments, obtained employing both Naïve Bayes and SVM.

TABLE II: ACCURACY PERCENTAGE AFTER THE TRAINING CORPUS ENRICHMENT

value n-grams	Baseline Accuracy		Iteration				
			1	2	3	4	5
1	Bayes	78.87	77.46	80.28	78.87	78.87	74.64
	SVM	56.34	64.78	64.78	64.78	64.78	66.19
2	Bayes	78.87	80.28	82.87	80.28	78.87	78.87
	SVM	66.20	66.20	74.64	66.20	68.29	68.29
3	Bayes	74.64	74.65	78.80	80.28	80.28	78.68
	SVM	64.80	64.80	66.20	64.78	64.78	68.29
Vocabulary Size		8377	8732	9019	9319	9676	9915

The integration of new information allowed improving the baseline since the first iteration. The best results are obtained during the first iterations when the most relevant snippets are added. The results show a different behaviour with respect to the size of n-gram (used as a parameter of classification).

The best results are obtained with a value equal to 2. This is due to the fact that bigrams are better suited to look for the most used collocations of an author. With n-gram whose size is greater than 2, we have a great increase in attributes. This aspect together with the small training data which is available do not allow for obtaining better results.

It is also interesting to notice that the behaviour of both classifiers is different. Thanks to the new information added by the proposed method, both classifiers improve their accuracy with respect to the baseline even if not exactly in the same way. In the case of SVM, its accuracy is always less than the one obtained with Naïve Bayes. Probably this is due to the small size of the training data.

In [19] some experiments were carried out on the same collection of poems. This method considers the use of four different kinds of word-based features: functional words, content words, and the combination of functional words, content words and n-grams. The authors used in their experiments a 10-fold cross-validation obtaining a precision of 78.8% with n-grams (unigrams plus bigrams). With the proposed semi-supervised learning method, we obtained an improvement in the accuracy percentage of 4%. Moreover, it is important to remark that we did not use cross-validation where a priori knowledge of all the vocabulary is considered. This is a very important aspect to take into account because in the poem classification task the variety of the vocabulary employed by the poets is huge.

IV. CONCLUSION

This paper proposed a novel approach for authorship attribution based on a semi-supervised learning method. We explored the use of text-mining methods for the identification of the author of a text. In particular, we proposed a semi-supervised method that is specially suited to work with few training examples.

This semi-supervised authorship attribution method differs from others in that: (i) it is specially suited to work with very few training examples, (ii) it automatically collects from the Web the unlabeled data and, (iii) it only incorporates into the training phase a small group of highly discriminative unlabeled examples.

In general, the achieved results allow us to formulate the following conclusions. On the one hand, the proposed combined approach can be a practical solution for the problem of authorship attribution. On the other hand, our Web-based semi-supervised learning method seems to be quite portable to other text categorization tasks, since it allows achieving very good results using very small training sets (e.g. for the categorization of news on natural disasters and the authorship attribution).

The experimental results on a set of contemporary Mexican poets showed the viability of the Web-based method even in a difficult task like this in which what we needed were not snippets of a certain author but passages written with, more or less, the same style.. In some way, they confirm our hypothesis that when dealing with very few training instances it is better to add a selected set of unlabeled examples (those that considerably augment the dissimilarity among classes) than incorporate a lot of doubtable-quality information. In particular, our method obtained the best results when we added to the training set ten unlabeled examples by iteration. It was also noticeable that our method achieved the best results only after two or three iterations. As future work, we plan to apply the proposed method to other text categorization problems. In particular, we would like to employ it for named entity recognition and word sense disambiguation.

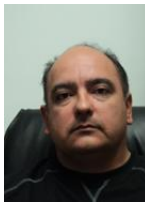
REFERENCES

[1] K. Aas and L. Eikvil, *Text Categorization: A survey, Technical Report, number 941*, Norwegian Computing Center, 1999.

[2] C. Chaski, "Who's at the Keyword? Authorship Attribution in Digital Evidence Investigations," *International Journal of Digital Evidence*, vol. 4, issue 1, 2005.

[3] A. Kaster, S. Siersdorfer, and G. Weikum, "Combining Text and Linguistic Document Representations for Authorship Attribution,"

Workshop Stylistic Analysis of Text for Information Access, 28th Int. SIGIR 1. MPI, Saarbrücken 2005, pp. 27-35, 2005.
[4] R. Bekkerman and J. Allan, "Using Bigrams in Text Categorization," CIIR Technical Report IR-408 Center for Intelligent Information Retrieval, University of Massachusetts Amherst, USA, 2004.
[5] E. Stamatatos, N. Fakotakis, and G.. Kokkinakis, "Computer-Based Authorship Attribution Without Lexical Measures," *Computers and the Humanities*, Kluwer Academic Publishers, vol. 35, pp. 193-214, 2001.
[6] J. Diederich, J. Kindermann, E. Leopold, and G. Paas, "Authorship Attribution with Support Vector Machines," *Applied Intelligence*, vol. 19, no. 1, pp. 109-123, 2003.
[7] R. García-Hernández, F. Martínez-Trinidad, and A. Carrasco-Ochoa, "A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection," *Lecture Notes in Computer Science*, vol. 3878, pp. 514-523, Springer, 2006.
[8] D. Holmes, *Authorship Attribution. Computers and the Humanities*, Kluwer Academic Publishers, 1995, vol. 28, pp. 87-106.
[9] P. P. Makagonov and M. A. Alexandrov, "Some Statistical Characteristics for Formal Evaluation off the Quality of Textbooks and Manuals," in A. Guzman and R. Menchaca (Eds.), *Computing Research Journal*, Mexico, pp. 99-103, 2000.
[10] Y. Benajiba and P. Rosso, "Towards a measure for Arabic corpora quality," in *Proc. Int. Colloquium on Arabic Language Processing, CITALA-2007*, Rabat, Morroco, pp. 213-221, June 18-19, 2007.
[11] S. Argamon and S. Levitan, *Measuring the Usefulness of Function Words for Authorship Attribution*, Association for Literary and Linguistic Computing/ Association Computer Humanities, University of Victoria, Canada, 2005.
[12] H. Ahonen-Myka, "Discovery of Frequent Word Sequences in Text Source," in *Proc. the ESF Exploratory Workshop on Pattern Detection and Discovery*, UK: London, 2002.
[13] Y. Zhao and J. Zobel, "Effective and Scalable Authorship Attribution Using Function Words," *Lecture Notes in Computer Science*, vol. 3689, pp. 174-189, Springer, 2005.
[14] M. B. Malyutov, "Authorship Attribution of Texts: a Review," in *Proc. the program Information transfer*, University of Bielefeld, Germany, pp. 17, 2004.
[15] Z. Kozareva, B. Navarro, S. Vazquez, and A. Montoyo, "UA-ZBSA: A Headline Emotion Classification through Web Information," in *Proc. the 4th International Workshop on Semantic Evaluations (SemEval)*, Prague, Czech Republic, 2007.
[16] S. Zelikovitz and M. Kogan, "Using Web Searches on Important Words to Create Background Sets for LSI Classification," presented at 19th International FLAIRS conference, Melbourne Beach, Florida, 2006.
[17] R. Guzmán-Cabrera, M. Montes-y-Gómez, P. Rosso, and L. Villaseñor-Pineda, "Improving Text Classification by Web Corpora," *Advances in Soft Computing*, no. 43, Springer, pp. 154-159.
[18] T. Solorio, "Using unlabeled data to improve classifier accuracy," M. Sc. Degree Thesis, Computer Science Department, Inaoe, Mexico, 2002.
[19] R. M. Coyotl-Morales, L. Villaseñor-Pineda, M. Montes-y-Gómez, and P. Rosso, "Authorship Attribution using Word Sequences," *Lecture Notes in Computer Science*, vol. 4225, pp. 844-853, Springer, 2006.
[20] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, "Augmenting Naïve Bayes Classifiers with Statistical Languages Models," *Information Retrieval*, vol. 7, pp. 317-345, Kluwer Academic Publishers, 2004.



Rafael Guzmán-Cabrera is a full time professor, Faculty of Engineering Mechanics Electrics and Electronics, Guanajuato University. He obtained his PhD in Pattern Recognition and Artificial Intelligence from Polytechnic University of Valencia, Spain. His contributed in research projects in the area of electrical engineering, pattern recognition and artificial intelligence.

José R. Guzmán-Sepúlveda is a student of M.Sc. degree in Electrical and Electronics Engineering - Optoelectronics at the Multidisciplinary Faculty Reynosa-Rodhe (UAMRR), Autonomous University of Tamaulipas, in Mexico. He has been contributor in national research projects in the area of Electrical Engineering related to Pattern Recognition and Artificial Intelligence (University of Guanajuato) as well as contributor in research projects abroad in the area of Optics and Photonics related to the design, fabrication, characterization, and testing of photonics devices, such as fiber optic sensors and microfabricated devices based on semiconductor technology for sensing applications (CREOL, University of Central Florida).

J. A. Gordillo Sosa is a full time professor in Universidad Tecnológica del Suroeste de Guanajuato. He was a PhD candidate Informatics from Polytechnic University of Valencia, Spain. He contributed in research projects in the area of software enginnering, pattern recognition and artificial intelligence.

Miguel Torres Cisneros obtained his Engineering Degree in Electronics at the Universidad de Guanajuato in 1988, his M. Sc. degree from the Centro de Investigaciones en Óptica (CIO) in 1991 and his Ph. D.cum Laude in Sciences from the Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) in 1997, He has been visiting research scientist in Dayton University (2002) and the University of Central Florida (UCF) in 2009. He has been professor at the Tech. of Monterrey and the Universidad de las Américas, and Titular Researcher at the Universidad de Guanajuato since 17 years, where he is involved with the NanoBioPhotonics Group, the Patents Group, and the Design and Manufacture Cell, Electronics & Mechatronics Programs. He has published over hundred scientific papers, holds 5 patents and he has been adviser of 42 graduate and undergraduate students. He holds the status of National Researcher (SNI) in Mexico since 1992, (level 2) and also holds the status of outstanding professor PROMEP since 1999. He was president of the Mexican Academy for Optics in 1999 and become a regular member of the Mexican Academy of Sciences in 2006. From 2006, He is a member of the UNESCO team for *Active Learning for Optics and Photonics* (ALOP), participant as facilitator in several Latin-American countries. From 2011 until now he is the University of Guanajuato Research and Graduate Programs Dean.

Joel Herrera Cabral is with the Communictions and Electronics Engineering and Electric Engineering. He obtained Master Degrees at the Universidad de Guanajuato. He participated as facilitator in many national and international congresses. His research areas include software engineering, software development, mobile computing and high performance computing.