International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1653-1661
© Research India Publications
http://www.ripublication.com

Document Weighted Approach for Authorship Attribution

P. Jeevan Kumar[1], G. Srikanth Reddy [2], T. Raghunadha Reddy [3],

[1,2,3] Dept. of IT, Vardhaman College of Engineering, Hyderabad, India.

Abstract

Authorship Attribution is a text classification technique, which is used to find the author of an anonymous document by analyzing the documents of multiple authors. The accuracy of author identification mainly depends on the writing styles of the authors. Various researchers proposed several features such as linguistic and content based features to differentiate the writing style of the authors and some of them used different classifiers to classify the documents. The existing approaches in authorship attribution used the bag of words approach to represent the document vectors. In this work, a new approach namely author specific document weighted approach is proposed, where in the document weight is used to represent the document vector instead of using features or terms in the document. In the proposed approach, term weight measure is used to assign appropriate weight to the terms. These term weights are used to compute the document weight against the documents of the authors. Different classifications algorithms are experimented with these document vectors to generate the classification model. The experimentation is carried on review corpus of various authors and the results achieved for author prediction is prominent than most of the existing approaches.

Keywords: Authorship Attribution, Author Prediction, Term Weight Measure, BOW approach, ADW approach.

1. INTRODUCTION

Authorship analysis is a procedure of finding the authorship of a text by inspecting its characteristics. Authorship Analysis is classified into two categories such as Authorship Attribution and Authorship Verification. Authorship Attribution predicts the author of a given anonymous document by analyzing the documents of multiple authors [1]. Authorship verification finds whether the given document is written by a particular author or not by analyzing the documents of a single author [2]. Authorship Attribution is used in several applications such as forensic analysis, security and literary research.

In forensic analysis, the suicide notes and property wills are analyzed whether the note or will is written by a correct person or not by analyzing the writing styles of suspected authors. The terrorist organizations send threatening mails, Authorship Attribution techniques are used to identify which terrorist organization send the mail or to conform the mail whether it came from correct source or not. In literary research, some researchers try to claim the innovations of others without proper acknowledgement. The Authorship Attribution is used to identify author of a document by analyzing the writing styles of the various authors.

This paper is structured as follows. Section 2 explains the existing work of Authorship Attribution. The existing model used by several researchers to represent a document and dataset characteristics are described in section 3. In section 4, the proposed model is explained with a term weight measure and document weight measure. The experimental results are analyzed in Section 5. Sections 6 conclude this work and suggest the future work in Authorship Attribution.

2. LITERATURE SURVEY

Authorship Attribution is divided in to two different subtasks. First, identification of most discriminative features to differentiate the writing styles of the authors. Second, determining the appropriate classification algorithm in order to detect the most probable author of a test document [4].

Ludovic Tanguy et al., experimented [5] with linguistic features such as sub-word level features, word-level features, sentence-level features, message-level features to represent the document vectors. They used maximum entropy technique and machine learning algorithms such as rule based learners and decision trees to evaluate the accuracy of author prediction. It was observed that maximum entropy technique achieved good accuracy than machine learning algorithms.

N. Akiva used [6] single vector representation that captures presence or absence of common words in a text. They used SVM-Light classification algorithm to generate the classification model. They observed that the accuracy of author prediction is increased when binary BOW representation is used to represent the document vector and also observed that the accuracy is increased when the number of authors is increased in the training data.

Stefan Ruseti et al., [7] extracted character trigrams, POS bigrams and trigrams, suffixes, word length, syntactic complexity and structure, percentage of direct speech from the documents to represent the document vector. They experimented with Sequential Minimal Optimization (SMO) algorithm and obtained an accuracy of overall 77% in Authorship identification. It was observed that the accuracy is increased when the application specific features are added.

3. TRADITIONAL APPROACH

Most of the Authorship Attribution approaches used the Bag Of Words approach to represent the document vector.

3.1 Bag Of Words (BOW) Approach

In this approach, first the preprocessing techniques are applied on the collected dataset. Extract the most frequent terms or features that are important to discriminate the writing styles of the authors from the modified dataset. Consider these terms or features as bag of words. Every document in the dataset is represented with this bag of words. Each value in the document vector is the weights of the bag of words. Finally, the document vectors are used to generate classification model.

3.2 Evaluation Measures

Various measures are used such as precision, recall, F1 measure and accuracy by the researchers in Authorship Attribution to test the accuracy of author prediction. In this work, accuracy measure is used to evaluate the performance of the author prediction. Accuracy is the ratio of number of test documents correctly predicted their author and the number of test documents considered. Accuracy measure is represented as

$$Accuracy = \frac{Number\ of\ \text{documents predicted their author correctly}}{Total\ number\ of\ documents}$$

3.3 Dataset Characteristics

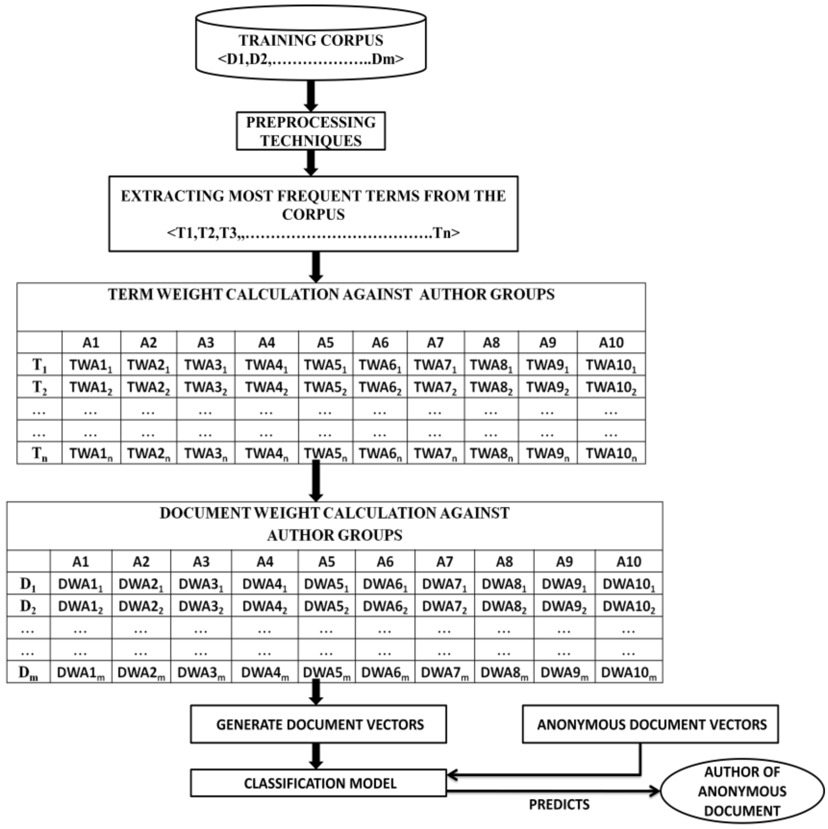
The dataset was collected from amazon.com and it contains 10 different authors reviews on different products. The corpus is balanced in terms of number of documents in each author group and each author group contains 400 reviews of each.

Table 1. Dataset characteristics of reviews

S No	Author	Number of documents
1	**A1**	**400**
2	**A2**	**400**
3	**A3**	**400**
4	**A4**	**400**
5	**A5**	**400**
6	**A6**	**400**
7	**A7**	**400**
8	**A8**	**400**
9	**A9**	**400**
10	**A10**	**400**

4. PROPOSED AUTHOR SPECIFIC DOCUMENT WEIGHTED (ADW) MODEL

Fig. 3 represents the architecture of proposed author specific document weighted model.


TRAINING CORPUS
<D1,D2,………………..Dm>
PREPROCESSING TECHNIQUES
EXTRACTING MOST FREQUENT TERMS FROM THE CORPUS
<T1,T2,T3,,……………………………….Tn>
TERM WEIGHT CALCULATION AGAINST AUTHOR GROUPS
DOCUMENT WEIGHT CALCULATION AGAINST AUTHOR GROUPS
GENERATE DOCUMENT VECTORS
ANONYMOUS DOCUMENT VECTORS
CLASSIFICATION MODEL
PREDICTS
AUTHOR OF ANONYMOUS DOCUMENT

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
T_1	$TWA1_1$	$TWA2_1$	$TWA3_1$	$TWA4_1$	$TWA5_1$	$TWA6_1$	$TWA7_1$	$TWA8_1$	$TWA9_1$	$TWA10_1$
T_2	$TWA1_2$	$TWA2_2$	$TWA3_2$	$TWA4_2$	$TWA5_2$	$TWA6_2$	$TWA7_2$	$TWA8_2$	$TWA9_2$	$TWA10_2$
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
T_n	$TWA1_n$	$TWA2_n$	$TWA3_n$	$TWA4_n$	$TWA5_n$	$TWA6_n$	$TWA7_n$	$TWA8_n$	$TWA9_n$	$TWA10_n$

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
D_1	$DWA1_1$	$DWA2_1$	$DWA3_1$	$DWA4_1$	$DWA5_1$	$DWA6_1$	$DWA7_1$	$DWA8_1$	$DWA9_1$	$DWA10_1$
D_2	$DWA1_2$	$DWA2_2$	$DWA3_2$	$DWA4_2$	$DWA5_2$	$DWA6_2$	$DWA7_2$	$DWA8_2$	$DWA9_2$	$DWA10_2$
…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…
D_m	$DWA1_m$	$DWA2_m$	$DWA3_m$	$DWA4_m$	$DWA5_m$	$DWA6_m$	$DWA7_m$	$DWA8_m$	$DWA9_m$	$DWA10_m$

Fig. 3. Architecture of proposed ADW model

The procedure in the proposed approach

1 Collect the corpus.
2 Apply preprocessing techniques to the corpus of documents such as stop words removal and stemming.
3 Extract frequent terms that occur at least 5 times in the total corpus.
4 Compute term weights in each author group of documents using term weight measures.
5 Document weights are determined for each author group by aggregating the weights of the terms in a document using document weight measure.
6 Generate document vectors with document weights to build a classification model.
7 The classification model is used to predict the author of a unknown document.

In this model, $\{D_1,D_2,\ldots\ldots D_m\}$is a list of documents in the corpus, $\{T_1,T_2,\ldots..T_n\}$ is a list of vocabulary terms, $\{A_1,A_2,\ldots..A_q\}$ is a set of author groups. $\{TWA1_n, TWA2_n, TWA3_n, TWA4_n, TWA5_n, TWA6_n, TWA7_n, TWA8_n, TWA9_n, TWA10_n\}$ is a vector of term T_n weight in the author groups A1, A2, A3, A4, A5, A6, A7, A8, A9, A10 respectively. $\{DWA1_m, DWA2_m, DWA3_m, DWA4_m, DWA5_m, DWA6_m, DWA7_m, DWA8_m, DWA9_m, DWA10_m\}$ is a vector of document D_m weight in the author groups A1, A2, A3, A4, A5, A6, A7, A8, A9, A10 respectively.

The profiles of an anonymous document are predicted using classification model. In this approach, identification of suitable weight measures for calculating term weight and document weight is important. The following sub section 4.1 and 4.2 discuss about the weight measures used in this approach.

4.1 Term weight specific to author group

Term weighting is an important concept in the modern information analysis. Different terms have different importance in a text. The term weight measure is used to find the importance of a term in a text. In general Author Attribution techniques easily analyze and predict the author of a document when the document contains large amount of text. For small documents, it is difficult to predict the author. In this paper pivoted Document length normalization technique is used to remove the difficulty of analyzing small sized texts. The document length normalization technique [8] maintains the term weights for a document in accordance with its weight.

Let A = $\{A_1,A_2,\ldots.A_q\}$ is the set of author groups, $\{D_1,D_2,\ldots.,D_m\}$is a collection of documents in the corpus, V = $\{t_1,t_2,\ldots.,t_n\}$ is a collection of vocabulary terms for analysis. Each term $t_i \in V$ is represented as a vector t_{ij}, i.e., $t_{ij} = \{t_{i1},t_{i2},\ldots.,t_{iq}\}$, where the dimension t_{ij} represents the term t_i weight on the author group A_j. Equation (1) is used to calculate the term weight in a specific author group.

$$W_{t_{ij}} = W(t_i, A_j) = \sum_{k=1}^{m} \frac{(1+\log(TF_i))/(1+\log(AVGTF_i))}{(1-slope)*AVGUT_k + slope*UT_k} \tag{1}$$

Where, $W(t_i, A_j)$ is the weight of i^th^ term in j^th^ author. TF$_i$ (Term Frequency) is the number of times the term t$_i$ is occurred in a document k, AVGTF$_i$ is a ratio of the term frequency t$_i$ to the total number of terms in kth document. As the experiment performed in pivoted unique normalization, the constant value 0.2 for slope is effective. UT$_k$ is a number of unique terms in kth document, and AVGUT$_k$ is a ratio of number of unique terms to total number of terms in kth document.

4.2 Document Weight Measure

The proposed document weight measure as in equation (2) is used to calculate the weight of a document on corpus of each author group. This measure used the combination of term weights that are specific to document and specific to author group. The TFIDF measure used to compute term weights specific to a document and term weight measures are used to determine the term weights specific to author group. The document weight computation considers the correlation between the terms in that document. The document weight computation is expressed as below

$$W_{d_{kj}} = \sum_{t_i \in d_k, d_k \in A_j} TFIDF(t_i, d_k) * W_{t_{ij}} \tag{2}$$

Where, W_{dkj} is the weight of document d_k in the author group A$_j$, Wt_{ij} is the weight of a term t_i in the corpus of author group A$_j$.

The collections of training documents are finally represented using equation (3)

$$Z = \bigcup_{d_k \in A_j} (z_k, c_j) \tag{3}$$

Where, z$_k$ = {W$_{dk1}$, W$_{dk2,....,}$W$_{dkq}$ } and c$_j$ is a class label of author group A$_j$.

The vector Z contains weights of a document specific to each author group with document author label.

5 EMPIRICAL EVALUATIONS

5.1 Results of BOW approach

Table 2. Accuracies of author prediction when BOW approach is used

CLASSIFIER/ NUMBER OF TERMS	NAIVEBAYES MULTINOMIAL	LOGISTIC	RANDOM FOREST
1000	72.09	69.67	68.21
2000	74.29	70.71	69.45
3000	75.48	72.23	71.01
4000	76.31	74.49	71.13
5000	78.17	76.97	71.87
6000	79.87	77.02	73.51
7000	81.39	77.11	73.92
8000	82.16	78.82	75.83

Table 2 represents the accuracies of author prediction in BOW approach using various classifiers. In the BOW approach, unlike other classifiers, the Naive Bayes Multinomial classifier achieved an accuracy of 82.16% for the most frequent 8000 terms.

5.2 Results of ADW approach

The accuracies of author prediction in ADW approach using various classification algorithms are represented in Table 3. With the PDW approach, the Naïve Bayes Multinomial classifier obtained accuracy with 97.71% for the most frequent 8000 terms. When the number of terms is increased, the accuracy is also increased in all the classifiers.

Table 3. Accuracies of author prediction when ADW approach is used

CLASSIFIER/ NUMBER OF TERMS	NAIVEBAYES MULTINOMIAL	LOGISTIC	RANDOM FOREST
1000	87.97	84.22	81.27
2000	88.33	86.38	82.33
3000	90.37	87.63	83.39
4000	91.43	88.69	85.86
5000	92.07	89.91	86.77
6000	94.87	91.67	88.69
7000	95.97	93.08	90.71
8000	97.71	94.89	91.53

The proposed PDW approach is evaluated and compared against the traditional BOW approach. This approach achieved better results than most of the existing approaches in Authorship Attribution. The Naïve Bayes Multinomial classifier produced more accurate results for author prediction with BOW and PDW approach. It is also notable that the most frequent 8000 terms as feature set, the proposed model generated good accuracies when compared with existing approaches. The Logistic, and RandomForest classifiers achieves a good accuracy with 8000 frequent terms in both approaches.

6. CONCLUSIONS AND FUTURE SCOPE

In this work, The proposed model achieved an accuracy of 97.71% for author prediction when Naïve Bayes Multinomial classifier is used. The BOW approach obtained an accuracy of 81.16% for author prediction when Naïve Bayes Multinomial classifier is used. In BOW approach the terms are independently participated in the classification process, but in proposed ADW model the terms are collaboratively in the form of document weight participated in the classification process. This is the main reason for obtaining good accuracies in the proposed model.

In our future work, it is planned to consider the domain characteristics, categorical features, usage of semantic and syntactic structure of the language to compute the weight of a document.

REFERENCES

[1] M. Sudheep Elayidom1, Chinchu Jose2, Anitta Puthussery3, Neenu K Sasi4 “TEXT CLASSIFICATION FOR AUTHORSHIP ATTRIBUTION ANALYSIS”, Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September 2013.

[2] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. J. Mach. Learn. Res. 8, 1261–1276 (Dec 2007).

[3] Efstathios Stamatatos. "A survey of modern authorship attribution methods", Journal of the American Society for Information Science and Technology, 03/2009

[4] Juola, P.: Authorship attribution. Found. Trends Inf. Retr. 1 (2006) 233-334

[5] Ludovic Tanguy, Assaf Urieli, Basilio Calderone, Nabil Hathout, and Franck Sajous. A Multitude of Linguistically-rich Features for Authorship Attribution. *CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, Netherlands*, September 2011. ISBN 978-88-904810-1-7. ISSN 2038-4963. .

[6] Navot Akiva. Authorship and Plagiarism Detection Using Binary BOW

Features, CLEF 2012 Evaluation Labs and Workshop, 17-20 September, Rome, Italy, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

[7] Stefan Ruseti and Traian Rebedea. Authorship Identification Using a Reduced Set of Linguistic Features—Notebook for PAN at CLEF 2012. CLEF 2012 Evaluation Labs and Workshop, 17-20 September, Rome, Italy, September 2012. ISBN 978-88-904810-3-1. ISSN 2038-4963.

[8] Amit S, Chris B, Mandar M. Pivoted document length normalization. In SIGIR ’96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, New York, USA, pp. 21–29