2009 IEEE International Conference on Semantic Computing

Stopword Graphs and Authorship Attribution in Text Corpora

R. Arun, V. Suresh, C. E. Veni Madhavan
Department of Computer Science and Automation
Indian Institute of Science
Bangalore, INDIA
Email: {arun_r, vsuresh, cevm}@csa.iisc.ernet.in

***Abstract*—In this work we identify interactions of stopwords –noisewords– in text corpora as a fundamental feature to effect author classification. It is convenient to view such interactions as graphs wherein nodes are stopwords and the interaction between a pair of stopwords are represented as edge-weights. We define the interaction in terms of the distances between pairs of stopwords in text documents. Given a list of authors, graphs for each author is computed based on their undisputed writings. Authorship of a test document is attributed based on the *closeness* of the graph derived from it to the above graphs. Towards this, we define a closeness measure to compare such graphs based on the Kullback-Leibler divergence. We illustrate the accuracy of our approach by applying it on examples drawn from the Gutenberg archives. Our results show that the proposed approach is effective not only in binary author classification but also performs multiclass author classification for as many as 10 authors at a time and compares favourably with the state-of-the-art in author identification.**

***Keywords*-stylometry; writer invariant; authorship attribution; KL divergence**

I. Introduction

Authorship classification received public notice with the classic work by Mosteller and Wallace [14] that resolved the disputed *Federalist Papers* (http://en.wikipedia.org/wiki/Federalist_Papers) for the first time through scientific means. In general, a situation for authorship attribution or resolution arises when there is uncertainty about a document's authorship and there is list of potential authors among whom one is its rightful author. In addition, for each author in the list, undisputed documents produced by the rightful author is available for the resolution mechanism.

A. Stylometry

The usual approach to resolution is by studying the available document for each author and deciding if the style of the writing matches with the style seen in the disputed document. Such author resolution methods employ techniques that are collectively known as stylometry or the study of linguistic style (http://en.wikipedia.org/wiki/Stylometry) whose objective is to establish a *Writer invariant* – http://en.wikipedia.org/wiki/Writer_invariant – property of text which is uniquely tagged to its author and is similar for all the texts produced by the same author. Such techniques are also used to capture *Writeprints* (http://en.wikipedia.org/wiki/Writeprint) analogous to finger print for forensic applications involving cases of plagiarism. Such techniques can help in better understanding of historical figures. A study of Reagan's radio speeches [1] is one such interesting instance.

Table I
Some function words and their grammatical categories

Function Words	Examples
Prepositions	of, at, in, without, between
Pronouns	he, they, anybody, it, one
Determiners	the, a, that, my, more, much, either, neither
Conjunctions	and, that, when, while, although, or
Modal verbs	can, must, will, should, ought, need, used
Auxilliary verbs	be (is, am, are), have, got, do

B. Function Words

In this work we present a scheme to resolve authorship for the binary and multi class problems. As with any scientific approach to stylometry, the basic inspiration for our approach is drawn from the work that essentially founded the modern science of stylometry — namely the [14] classic result on Federalist papers. Their approach was based on the distribution of function words in the text corpus and building a Naive Bayes classifier to place the distribution resulting from the test document. The term *function word* is a misnomer as these words do not perform any 'function' as the name suggests. In fact they are usually ambiguous and have minimal meaning outside the context of the sentence. One is referred to *http://en.wikipedia.org/wiki/Function_word* for a more detailed description of function words. Though the approach based on function words is proven to be effective, it is surprising as one would rather expect an author's style to captured in terms of grammatical constructs that convey syntactic and semantic meaning. A tentative list of function words are given in Table I.

C. Previous Works

We now give a overview of the classification literature. Authorship classification is presented under three different categories. Binary class — Given two authors attribute the disputed document to one of them. Such classification techniques are presented in [3], [7] and [2], [5]. Multi class — Given more than two authors, assign authorship

978-0-7695-3800-6/09 $26.00 © 2009 IEEE
DOI 10.1109/ICSC.2009.101

for the disputed document. Refer to [8], [10], [6] for multi author classification techniques. [11] presents classification for Single class — Of the available documents, some are written by one author and the task is to verify if the remainder of the unattributed documents are authored by the same author. This is also known as authorship verification.

Apart from function words, classification techniques have tapped into other features. Martindale and McKenzie [13] use tone of words for classification. Idiosyncrasies of word usage and bigrams of POS tagged text are used by Koppel and Schler in [9], n-grams are considered by Frantzeskou *et al* [4] for classification. Among classification techniques Support Vector Machines are popular in the literature [17], [12], [5]. We come across one instance of KL divergence in Zhao *et al* [18], a measure which the present work also utilizes. Principle Components Analysis is used for classification in [15].

D. Organization of the Work

This paper is organized as follows. In Section II we introduce our classification approach. This is followed by a description of our experiments and results in Section III. Subsequent to this we present our conclusions in Section IV.

II. Stopword Graphs

A. Stopwords

Our approach is based on a what are known as stopwords or noise words – words that convey very little semantic meaning in a sentence but serve to add details to it. "Children *are* playing *in the* garden". Without stopwords: Children – playing – garden. In this spirit they are like function words but stopwords lists also include words that are outside function word lists. For more pointers and lists of stopwords consult *http://en.wikipedia.org/wiki/Stop_words*. Stopwords are usually identified based on their prevalence in text. In general they occupy a significant percentage of the text – A first cut approximation would be 50%. This is far in excess of its proportion in the English lexicon.

Success of [14] and many subsequent successful author identification approaches based on stopwords frequencies is a vindication of their resolving capacities. In the following we conjecture a short account on why this might be the case. Though the meaning of a sentence is not captured by stopwords, their sheer numbers ensure that their distributions are to be taken seriously for stylometric purposes. At hind sight this makes sense: words that occur infrequently may indeed lead to ad-hoc techniques for author identification but may not lend to any generalization. For example, a phrase like, *"She gave a quizzical glance"*, would help one conclude that it could not have been written by Shakespeare; word quiz was not in vogue during the late 16^{th} and early 17^{th} century England (http://www.askoxford.com/asktheexperts/faq/aboutwordorigins/quiz). Stopwords on the other hand are inevitable in the output of any author and hence a generalizable technique cannot but tap their properties. More than that, stopwords are result of an unconscious process of forming sentences and hence may serve as writeprint of authors. By this we mean that authors do not plan in advance the number of stopwords they are going to use once they conceive a broad idea and start penning the narrative. Though one might rack their brains to retrieve an appropriate word to be used in a sentence, one seldom gives a thought before using a stopword – they come spontaneously[1]. So however common might be their occurrence in texts, stopwords seem to provide enough room for authors to embed their style through its use in quantifiable ways. We consider a list of 571 stopwords used in [12] for the present work.

B. Viewing Stopwords as a Graph

Unlike previous approaches that have considered function words for classification, we not only consider the occurrence of stopwords but also take into account the distances between them. An apt way to view inter-stopword distances is by considering stopwords as nodes of a stopword graph with the distances between them captured by the edge weights. Our heuristic assigns more weight to edges that correspond to stopwords that have smaller distances between them, while stopword pairs that are far removed in the corpus get lower edge weights. Distance here refers to the number of words that separate a stopword pair.

1) Construction: Initially we start with a collection of isolated nodes, each one corresponding to a stopword in the list. As a stopword w_s is encountered in the corpus, we measure its distance to recent occurrences of all other stopwords w_i. We then update the edge weights of all the edges leading to other stopwords from w_s as follows. If w_s has occurred in position p_s in the corpus and w_i recent occurrence was in p_i, we add $e^{-|p_i - p_s|}$ the weight of the edge (w_s, w_i). We update the incoming edges to (w_i, w_s) in a similar manner as we do not distinguish between edges (w_i, w_s) and (w_s, w_i). Note that the negative exponential function for determining the edge weight is in accordance with our objective to give more weights for closely occurring stopword pairs and penalizing stopword pairs that occur far apart.

2) Classification Heuristic: We describe the classification for the binary class problem, this is generalizable to the multi class situation. We construct two stopword graphs one each for the text corpus corresponding to the two authors. We will call these as the training graphs. Another stopword graph is constructed for the document in contention. We will call this as the test graph. The document is awarded to one of the authors based on how similar the test graph is to one of these two training graphs. This is done by comparing

[1]As an illustration, the reader could try writing a two 200 word abstracts on a same topic: one without using more than 20 stopwords and another without any constraints on the number of stopwords and compare the efforts in either case.

the neighborhood of each stopword node from the test to the neighbourhood of the corresponding node in the two training graphs. A measure quantifying the similarity of the neighborhoods is computed for each node and the sum total of this measure for all the nodes of the test graph versus each of the training graph is computed. The training graph closest to the test graph in terms of the cumulative sum total of the similarity measure determines the author to which the document is attributed.

3) Kullback-Leibler(KL) Divergence: For comparing the neighbourhoods of two stopwords in the stopword graph, we make use of Kullback-Leibler divergence. $\mathcal{KL}$ divergence is a non commutative measure of similarity of discrete probability distributions P and Q defined as $\mathcal{KL}(P\|Q) = \sum_i P(i)\log(P(i)/Q(i))$, where $P(i)$ and $Q(i)$ are the i_{th} point in the respective distributions. $\mathcal{KL}(P\|Q) \geq 0, 0 \iff P = Q$. P and Q are similar for smaller values of divergence and dissimilar for higher values. The above measure is non commutative – usually $\mathcal{KL}(P\|Q) \neq \mathcal{KL}(Q\|P)$. Hence we opt for the symmetric variant $\mathcal{KL}(P\|Q) = \sum_i (P(i)\log(P(i)/Q(i)) + Q(i)\log(Q(i)/P(i))/2$ as it is commutative and suits our approach.

Each node gives rise to a probability distribution in terms of its neighborhood edge weights. Authorship is attributed in terms of the similarity of these distributions for corresponding stopword nodes from the training and testing graphs based on the above $\mathcal{KL}$ divergence measure.

In this case the similarity is greater if the divergence is smaller. Hence smaller the divergence of the test graph to a training graph, closer they are. The smallest such cumulative sum total of $\mathcal{KL}$ divergences determines the author attribution to the document. If such cumulative sums happen to be the same for both the training graphs, then the classification is undecided. We consider this as a remote possibility and have not encountered this situation so far.

We mention two cautionary approaches we take while computing $\mathcal{KL}$ divergence. We convert the weight distribution of edges of each node into a probability distribution by normalizing the weights to lie in $[0-1]$. It is possible that a stopword does not occur in the test corpus. In such cases we set the weights of corresponding edges in the training graphs as zero. The above mentioned normalization is done after that. In addition a residue ϵ is added to all the weights so that there is no inadvertent computation of $log(0)$ due to the zeros introduced from the above step. Note that this takes care of situations wherein a stopword present in the test corpus is absent in the training corpora.

4) Pseudocode: The pseudocode of our classification heuristic is presented below for the binary case.

---

module: **getStopWordGraph**
Input: Corpus, $stop\_word$ list
Output: Stopword network, $G = (V, E)$
init: $V = \{w_1, w_2, \ldots, w_n\}, E = \phi$
for every occurrence of w_s, at position p_s
  (note: p_i=0 until w_i appears in the corpus)
  $\forall i = 1 \ldots n$
  update weight of edges $(w_i, w_s), (w_s, w_i)$
  if($p_i \neq 0$) $W_{i,s} = W_{s,i} \leftarrow W_{s,i} + e^{-|p_i - p_s|}$
    (p_i : most recent occurrence of w_i)
return $G = (V, E)$

module: **authorKLdiv**
Input: $G_{trn_1}, G_{trn_2}, G_{tst}$
Output: 1, if $G_{tst} \simeq G_{trn_1}; -1$ if $G_{tst} \simeq G_{trn_2}$ note: if $w_s \in V_{trn_1}, V_{trn_2}$ and $\notin V_{tst}$:
  set $(w_i, w_s) = (w_s, w_i) = 0; \forall i = 1 \ldots n$
Normalize Edge Weights of $G_{trn_1}, G_{trn_2}, G_{tst}$
replace all $(w_i, w_j) = 0$ with $(w_i, w_j) = \epsilon$
$KL_1 = 0, KL_2 = 0$
for each stop_word w_i
  $P_1 = \{W_{i,s}\} : W_{i,s} \in E_{trn_1}, \forall s = 1 \ldots n$
  $P_2 = \{W_{i,s}\} : W_{i,s} \in E_{trn_2}, \forall s = 1 \ldots n$
  $Q = \{W_{i,s}\} : W_{i,s} \in E_{tst}, \forall s = 1 \ldots n$
  $kl_1 = (\mathcal{KL}(P_1\|Q) + \mathcal{KL}(Q\|P_1))/2$
  $kl_2 = (\mathcal{KL}(P_2\|Q) + \mathcal{KL}(Q\|P_2))/2$
  $\mathcal{KL}_1 \leftarrow \mathcal{KL}_1 + kl_1$
  $\mathcal{KL}_2 \leftarrow \mathcal{KL}_2 + kl_2$

if($KL_1 < KL_2$)
  return 1
else
  return -1

module: **main**
Input: Corpus with class labels, test corpus
Output: Class of test corpus
G_1 = getStopWordGraph($Corpus_1$)
G_{-1} = getStopWordGraph($Corpus_{-1}$)
G_X = getStopWordGraph$Corpus_X$)
X = authKLDiv($G_{trn_1}, G_{trn_{-1}}, G_{tst}$)
return X

Though we have presented the heuristic for the binary case alone, it is easy to extend it to the multi-class case. Given N authors and a disputed document, one needs to compute training graphs for the corpus corresponding to each author and the one corresponding to the disputed document and compare the similarity of test graph with each of the training graph. The authorship of the disputed document is awarded to the one with the least cumulative divergence from the test graph.

III. Experiments and Results

We have drawn text documents from Project Gutenberg, *www.gutenberg.org*. For computing training graphs we have taken documents that contain at least $50,000$ words. The test

graphs are derived from documents that are at least $10,000$ words long. In the following we present the results for binary and the multi author cases.

Results of our classification are presented in Table II.

The top four entries in the table correspond to multi classification among 10 authors — *Thomas Hardy, Henry Haggard, Anthony Trollope, Mark Twain, P. G. Wodehouse, Conan Doyle, Somerset Maugham, Agatha Christie, Charles Dickens and Leo Tolstoy*.

The number of documents considered as test documents for each author is known from the fifth column in the table. Eg: Hardy: $9/10$ implies, 10 novels by Hardy are considered out of which nine were identified correctly as Hardy's. A single document is used for each author for training. These are outside the set mentioned in column five.

We have not come across a multi-author classifier in the literature that performs for these many authors. The results corresponding to rows inclusive of and below Wodehouse, are our early results with lesser classes. Wodehouse was classified against four authors below him in the table. From Doyle downwards, this classification was among four authors. The table shows that classification accuracy does not vary much with increase in the number of authors. These results are better than [18] which considers at most five authors at one time.

As an interesting observation, Dickens' novel *The Life of our Lord* which was published in 1934[2] (long after his death in 1870) was correctly attributed by our classifier.

We now discuss the effects of translation and the change of author-style over a period on our measure. This is in conjunction with the absence of Tolstoy in Table II and also the relatively poorer performance of our measure for Mark Twain.

It is reasonable to assume that a translator's free will and spontaneity are in a state of tension with the constraint imposed by the primary purpose — that of capturing the original author's sentiments. It is illustrative to observe the results for translated documents in order to check if the translated texts are attributed correctly. If so, one would conclude that the original author's sentiments prevail inspite of translation. If not the conclusion is to the contrary – translator has sufficient 'space' to operate upon while translating and yet could preserve the style component of the original author.

We conducted our experiments on Tolstoy's works to observe the effect of translations by different authors on the original Russian+French texts. None of the works, eg: War and Peace and Anna Karena were identified as Tolstoy in the multi-author case. In most cases, the similarity ranks are far below in the table suggesting that the writer invariant is not captured in translated texts by the present approach. One could say that the writer invariant is occluded by the effect of translation – at least within the confines of our approach. Comparable results are not known from the literature regarding translated texts.

As with translation affecting the identification, it is possible that changes in the author's style would affect the quality of inference through our measure if the training graph is not representative of the author's other writings. This might be the case if the training corpus comes from the early writings of the author and the testing corpus comprises documents that were produced much later in the author's literary career. One might wish to call this as drift in style or simply *author-drift*. We suspect that the relatively poorer results for Twain's writing are due to this effect. It is highly probable – even inevitable – that an author's writing continues to evolve during the writing career. This may have a bearing on the measure we compute. This aspect requires further study. That our results are highly accurate for other authors who were perhaps not immune to such effects exemplifies the robustness of our approach.

This also presents another interesting prospect for stylometric analysis – that of predicting the time of writing of documents and in the process computing a time-line for the author's writing. Also this would enable choosing of more representative documents for building training graphs for authors with diverging styles.

IV. Conclusions

We have presented a scalable – works for multiple authors – and *non-adhoc* – it does not depend on idiosyncrasies of word usage – approach for author classification by modelling stopword co-occurrences in text as graphs and have shown the efficacy of such a representation in effective author classification both in binary and multi-author classification. Our method establishes stopword pairs and their inter-distances as an effective writer invariant. Thus instead of just looking at frequencies of occurrences of stopwords as in earlier approaches, we have also looked at their 'flow' in the document. In this process, our approach integrates the frequency and the interaction of stopwords into one unified structure in the form of stopword graphs. Since the use of stopwords and their inter distances are usually unplanned and involuntary (unlike content words that carry the narration forward), this unification might correspond to a stronger writer invariant in comparison to the conventional ones obtained by a mere frequency based approach.

For the present approach we have used 571 stopwords, and $50,000$ words for training –for building stopword graphs for authors, and $10,000$ words for test documents. The exact number of stopwords required to calibrate an author's style, the lower bounds on the size of the training corpus and the test documents need to be identified and their impact on the classification accuracy requires further studies.

[2]The reason for the delayed publication can be seen in *http://en.wikipedia.org/wiki/The_Life_of_Our_Lord*.

Table II
A TABULATION OF EXPERIMENTAL RESULTS: RESULTS IN THE TOP FOUR ROWS CORRESPOND TO MULTI-CLASS RESOLUTION WITH 10 AUTHORS. FIFTH ROW CORRESPONDING TO WODEHOUSE CORRESPONDS TO MULTI-CLASS RESOLUTION WITH FIVE AUTHORS THAT ARE LISTED BELOW WODEHOUSE. RESULTS FROM THE SIXTH ROW DOWNWARDS CORRESPONDS TO MULTI-CLASS RESOLUTION WITH FOUR AUTHORS, NAMELY: DOYLE, MAUGHAM, CHRISTIE AND DICKENS. THE DIFFERENCES IN THE NUMBER OF CLASSES FROM THE BOTTOM ROW TO THE TOP INDICATES THE PROGRESSION OF OUR EXPERIMENTS WITH INCREASING NUMBER OF AUTHORS.

author	binary accuracy(%)	multi-class accuracy(%)	binary correct/total	multi-class correct/total	classes considered
Hardy	96.67	90	87/90	9/10	10
Haggard	98.89	90	89/90	9/10	10
Trollope	100	100	90/90	10/10	10
Twain	83.3	30	75/90	3/10	10
Wodehouse	97.22	88.9	128/144	32/36	5
Doyle	90.3	80.9	118/126	34/42	4
Maugham	88.89	67	16/18	4/6	4
Christie	100	100	3/3	1/1	4
Dickens	97.22	91.7	188/192	44/48	4
average accuracy	binary 94.72%	multi-class 82.05%			

Though unstated, it is implicit that all our results pertain to writing in English. It would be illustrative to check the efficacy of stopword distances for author classification in other languages. In addition, exploring the framework's potential to identify gender, genre and mix of styles —author collaborations, are also some of the future possibilities.

Though we have modelled the stop-word interactions as graphs, we have barely utilized the rich properties that stem from the current understanding of such graphs discussed under the generic header known as complex networks — real-world graphs such as social networks, WWW and the like. For a survey of this area refer to Dorogovtsev and Mendes [16]. Studying stopword graphs under the scanner of complex networks would be an interesting prospect.

REFERENCES

[1] E. Airoldi, S. Fienberg, and K. Skinner. Whose ideas? Whose words? Authorship of Ronald Reagan's radio addresses. *Political Science and Politics*, 40(03):501–506, 2007.

[2] J. N. G. Binongo. Who wrote the 15th book of Oz? An application of multivariate statistics to authorship attribution. *Literary and Linguistic Computing*, 16(2):9–17, 2003.

[3] M. R. D. I. Holmes and R. Paez. Stephen crane and the new-york tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3):315–331, 2001.

[4] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. K. Katsikas. *Source Code Author Identification Based on N-gram Author Profiles*, pages 508–515, 2006.

[5] G. Fung. The disputed federalist papers: SVM feature selection via concave minimization. *In Proceedings of the 2003 Conference on Diversity in Computing*, pages 42–46, 2003.

[6] A. N. H. Baayen, H. van Halteren and F. Tweedie. An experiment in authorship attribution. *Proceedings of JADT*, pages 29–37, 2002.

[7] D. I. Holmes. The evolution of stylometry in humanities scholarship. *Literary and Linguistic Computing*, 13(3):111–17, 1998.

[8] E. L. J. Diederich, J. Kindermann and G. Paass. Authorship attribution with support vector machines. *Applied Intelligence*, 19:109–123, 2003.

[9] M. K. Jonathan and J. Schler. Exploiting stylistic idiosyncrasies for authorship attribution. *In IJCAI03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 69–72, 2003.

[10] P. Juola and H. Baayen. A controlled-corpus experiment in authorship identification by cross-entropy. *Literary and Linguistic Computing*, 20:59–67, 2003.

[11] M. Koppel and J. Schler. Authorship verification as a one-class classification problem. *ICML*, pages 489–495, 2004.

[12] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[13] C. Martindale and D. McKenzie. On the utility of content analysis in author attribution: The federalist. *Computers and Humanities*, 29:259–270, 1995.

[14] Mosteller and D. L. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, 1984.

[15] R. Peng and H. Hengartner. Quantitative analysis of literary styles. *The American Statistician*, 56(3):175–185, 2002.

[16] J. F. F. M. S. N. Dorogovtsev. *Evolution of Networks: From Biological Nets to the Internet and WWW*. 2003.

[17] Y. Zhao and J. Zobel. Effective authorship attribution using function word. *Proc. 2nd AIRS Asian Information Retrieval Symposium*, pages 174–190, 2005.

[18] Y. Zhao, J. Zobel, and P. Vines. Using relative entropy for authorship attribution. *Proc. 3rd AIRS Asian Information Retrieval Symposium*, pages 92–105, 2006.