Writing with Style: Venue Classification

Zaihan Yang Brian D. Davison
Department of Computer Science and Engineering, Lehigh University
Bethlehem, PA 18015
Email: {zay206,davison}@cse.lehigh.edu

***Abstract*—As early as the late nineteenth century, scientists began research in author attribution, mostly by identifying the writing styles of authors. Following research over centuries has repeatedly demonstrated that people tend to have distinguishable writing styles. Today we not only have more authors, but we also have all different kinds of publications: journals, conferences, workshops, etc., covering different topics and requiring different writing formats. In spite of successful research in author attribution, no work has been carried out to find out whether publication venues are similarly distinguishable by their writing styles. Our work takes the first step into exploring this problem. By approaching the problem using a traditional classification method, we extract three types of writing style-based features and carry out detailed experiments in examining the different impacts among features, and classification techniques, as well as the influence of venue content, topics and genres. Experiments on real data from ACM and CiteSeer digital libraries demonstrate our approach to be an effective method in distinguishing venues in terms of their writing styles.**

I. Introduction

As early as the late nineteenth century, the research scientist T.C. Mendenhall conducted his pioneering studies in authorship attribution among Bacon, Marlowe, and Shakespeare. More than half a century later, another two scientists, Mosteller and Wallace, carried out their famous study on the mystery of the authorship of the Federalist papers [1]. They examined 146 political essays from the late eighteenth century, of which most are acknowledged to have been written by John Jay, Alexander Hamilton and James Madison; however, twelve of them are claimed to be co-authored by Hamilton and Madison. By extracting function words as one of the most important stylometric features and making use of Bayesian statistical analysis, Mosteller and Wallace assigned all twelve disputed papers only to Madison.

These early studies initiated research in author attribution, also known as author verification or identification, and demonstrated that writing style is a key feature in distinguishing among authors. Today we not only have many more authors writing and publishing papers, but also have many different kinds of publications, covering different topics, with different genres and requiring different writing formats. In this paper, we regard the publishing venues of all kinds of publications as venues. We have different venues for different research domains; for example, the ‘SIGIR’ conference for Information Retrieval (IR) research, and the ‘VLDB’ conference for database research. Moreover, even in one research domain, we also have multiple venues. To take the ‘IR’ research domain as an example, we have journals such as *Information Retrieval* and *J.ASIST*, as well as conferences, such as SIGIR, JCDL, WWW, CIKM, etc. We also have posters, workshops, technical reports, patents, etc. With so many different kinds of venues provided, a straightforward question may arise: how can they be distinguished from each other? Besides their topic-related differences, are they also distinguishable in writing styles?

A writing style, according to Karlgren [2], is a consistent and distinguishable tendency in making some linguistic choices. Compared to the content of a paper, writing style more reflects the preferences of authors in organizing sentences and choosing words. Identifying distinguished features of venues in terms of their writing styles can provide us another point of view or additional information in evaluating the differences among multiple venues, and such a differentiation can provide benefit to some applications, one of which is venue recommendation, whose main task is to generate a list of venues to which a given paper may be submitted. It is sometimes difficult to choose a proper venue since there are many available choices and some of them even focus on similar topics. If additional hints can be provided from venues’ writing styles, researchers may find it easier to make their choices. In a similar way, distinguishing venues by their writing styles may also help us determine the real publishing venue of a given paper whose venue information is missing, yet that information is needed, for example, in digital libraries. These potential applications stimulate our research into the task of venue classification, which has not been explored before.

In summary, our paper makes the following contributions: 1) the first exploration into distinguishing venues by their writing styles; 2) detailed experiments examining different impact of several factors for classification on two real data sets.

II. Related Work

There is a lack of prior work exploring the problem of classifying venues by their writing styles. However, there has been a long history in the research of author attribution, also known as author identification or verification, whose main task is to determine the author of a piece of work, mostly by identifying the unique writing styles of authors. Author attribution has been used in a small yet diverse number of applications, such as authorship verification for literature and published articles, for online messages [3], [4], plagiarism detection and forensic analysis for criminal cases.

One of the most important components for author attribution is to identify representative stylometric features, which compared to the features used in text content classification, are assumed to be topic-independent and context-free. Stylometric features used in early author attribute studies are lexical [1] (i.e., character and word) based, such as number of words and characters, word length, vocabulary richness [5], [6]. Further study then began to make use of syntactic features [7]. The three most representative syntactic features are function words [8], [9], punctuation [10] and part-of-speech tags [7]. More recently, structural features [11], such as number of paragraphs, use of indentation, use of signature, have attracted attention, especially for online message authorship identification. Other useful stylometric features include character-based n-grams [12] and POS-based n-grams [13]. However, due to different applications, no set of significant stylometric features have been identified to be the most discriminative.

Just as there are a range of stylometric features, there are also many techniques for author attribution. In most cases, this task has been treated as a single-label multi-class classification task, and therefore many classification techniques have been considered [3]. Besides that, there are other techniques such as statistical approaches [14], neural networks [15], genetic algorithms [9], and principle component analysis approaches [8]. Most recently, researchers have started to use latent factor models into author attribution task [16], [17]. However, there is no consensus on which particular approach can perform the best due to different applications.

In this paper we conduct a detailed study of venue classification by adopting a set of stylometric features that have been demonstrated useful in author attribution. Unlike most author attribution experiments, we test large numbers of classes (venues). We work on real data sets, collecting paper instances according to the actual distributions of venues in the data corpus. Moreover, we compare classification results using different feature sets and classifiers and further examine the distinguishing power between creating style-based classifiers and content-based classifiers. We further explore the relationship between writing styles and topics and genres respectively.

III. Problem Identification

Given a set of papers, with their full or partial content provided, the task of venue classification is to determine the likelihood of a paper to be published in a particular venue. We can approach the task using traditional classification techniques, where a set of papers with known venue information are used for training, and the ultimate goal is to automatically determine the corresponding publishing venue of a paper whose venue information is missing. In particular, we are interested in exploring the following research questions:

- How well can venues be distinguishable from each other in terms of writing styles?
- What are the valuable features to represent writing styles?
- How sensitive is venue classification to classifier choice?
- Compared with using content-based features, can we improve classification results using stylometric features?
- Are topically-similar venues distinguishable by writing styles?
- Are venues of different genres distinguishable by writing styles?

IV. Features

Since we focus on writing-style based venue classification, one of the main concerns is to define an appropriate quantitative text representation that captures the writing style of scientific papers. To avoid the influence from paper content, the features we employed need to be unrelated to topic and context-free. Based on previous studies and analyses in the task of author attribution, we incorporated three types of features into the feature set: lexical features, syntactic features and structural features. The entire set of features is listed in Table I.

TABLE I
Features

Type	Features	Description
Lexical	TokenNum	Total number of words
	TypeNum	Total number of distict words
	CharNum	Total number of characters
	SentenceNum	Total number of sentences
	AvgSenLen	Average sentence length
	AvgWordLen	Average word length
	ShortWordNum	Total number of short words (less than 3 characters) normalized by TokenNum
	HapaxVSToken	Frequency of once-occuring words normalized by TokenNum
	HapaxVSType	Frequency of once-occuring words normalized by TypeNum
	ValidCharNum	Total number of characters excluding the non-digital, non-alphabetical and non-white-space characters
	AlphaCharNum	Total number of alphabetic characters normalized by CharNum
	DigitalCharNum	Total number of digital characters normalized by CharNum
	UpperCaseNum	Total number of characters in upper-case normalized by CharNum
	WhiteSpaceNum	Total number of white-space characters normalized by CharNum
	SpaceNum	Total number of space characters normalized by CharNum
	TabSpaceNum	Total number tab spaces normalized by CharNum
	Vocabulary Richness	A vocabulary richness measure defined by Zipf
Syntactic	FuncWordNum	Total number of function words
	PunctuationNum	Total number of punctuation characters (‘.’, ‘?’, ‘!’, ‘,’, ‘:’, ‘;’, ‘”, ‘ / ’)
	FuncWordFreq	Frequency of function words normalied by FuncWordNum (298 features)
Structural	SectionNum	Total number of sections
	FigureNum	Total number of figures
	EquationNum	Total number of equations
	TableNum	Toatl number of tables
	ReferenceNum	Total number of references

Lexical Features: Lexical features can be further divided into character-based or word-based features. It reflects a paper's preference for particular character or word usage. In our work, we included character-based features like number of terms, number of distinct terms, etc. The number of Hapax terms, one of the features we used, is defined to be the number of distinct terms that appear only once in the paper. We also used vocabulary richness as defined in [18]. In total, we have 66 lexical features.

Syntactic Features: Compared to lexical features, the discriminating power of syntactic features is derived from different formats and patterns in which sentences of a paper are organized. They are more likely to be content-independent. One of the most important syntactic features is the set of short yet all-purpose words, which are often referred to as function words, such as ‘the’, ‘a’, ‘and’, ‘to’, etc. Research in author attribution demonstrated that function words play an important role in identifying authors, since their frequency of usage are often unaffected by papers' subjective topics. We adopted a set of 298 function words. Another example of a syntactic feature is punctuation. We count the sum of appearances of eight predefined punctuation symbols that appear in the paper.

Structural Features: Structural features represent the layout of a piece of writing. De Vel [11] introduced several structural features specifically for email. In our work, we adopted five structural features specifically for scientific papers: the number of sections, figures, equations, tables, and bibliographic references. Due to the fact that the original paper content available is in raw text format, in order to retrieve the number of figures in one specific paper, we simply count the number of times the word ‘figure’ or ‘Figure’ appears in the paper. We did the same for number of sections, number of tables and number of equations. We add number of references as an extra feature, not only because it is available in our data set, but also because this kind of feature is important for scientific papers. We can retrieve all of these five features for the papers in the CiteSeer data set, where the full paper content is available. For papers in the ACM data set, we can only retrieve the number of references feature.

In summary, we have 371 features for papers in the CiteSeer data set, and 367 features for papers in the ACM data set. The data sets are described below.

V. Experimental Evaluation

A. Data Collection

In order to test whether we can successfully classify venues by their writing styles, we perform experiments on two real world data sets. The first data set is a subset of the **ACM Digital Library**, from which we crawled one descriptive web page for each of 172,890 distinct papers having both title and abstract information.

For each published paper, we extract its publishing venue and citation references. Due to possible venue name ambiguity, we first convert all upper-case characters into lower-case, and remove all non-alphabetical symbols. We further removed all digits as well as the ordinal numbers, such as the 1st, the 2nd, and applied Jaccard similarity match to merge duplicate venue names. We finally obtained 2,197 distinct venues.

The second data set we utilize is the **CiteSeer digital library** scientific literature distributed by the 2011 HCIR challenge workshop[1]. The whole data corpus is divided into two parts. Meta-data about a paper, such as its title, publishing venue, publishing year, abstract, and information about citation references are kept in XML format; the full content of that paper is in plain text. We collected 119,727 papers published between 1949 and 2010 that have both abstracts and full content information. We applied the same working process as we did for the ACM data set to merge ambiguous venue names, and finally obtained 48,797 venues.

B. Overall Classification Results

In a first analysis, we determine whether venues are distinguishable by their writing styles under general circumstances, regardless of content, topic and genre effects.

For all experiment settings, we make use of 10-fold cross validation, and adopt Accuracy and F_1 score, the two traditional classification metrics for performance evaluation.

1) Multi-Class Classification Results: To examine multi-class classification results, we randomly choose K venues, where K indicates the number of venues on which we tested. In our experiments, we change the value of K among 2, 5, 10, 30, 50, 100 and 150. For each value of K, we randomly choose K venues that have at least 100 papers for the ACM data set (at least 50 papers for the CiteSeer data set). We collect all the papers published in those chosen venues to construct the training/testing sets. The same process is repeated ten times for each particular K, and the results are an average of all the iterations.

TABLE II
Statistics over Chosen Venues

	Avg. No. of Papers Per Venue	Avg. length of Papers per Venue (Abstract)	Avg. length of Papers per Venue (Full Paper)
ACM	415	105 words	N/A
CiteSeer	98	140 words	6490 words

TABLE III
Multi-Class Venue Classification for ACM Data Set. Value* is significantly better than the Baseline Classifier

		Accuracy	F_1 Score
2-Venue	Baseline	0.503	0.481
	Stylometric	**0.806***	**0.713***
5-Venue	Baseline	0.195	0.177
	Stylometric	**0.584***	**0.454***
10-Venue	Baseline	0.099	0.085
	Stylometric	**0.434***	**0.309***
30-Venue	Baseline	0.033	0.027
	Stylometric	**0.267***	**0.118***
50-Venue	Baseline	0.020	0.015
	Stylometric	**0.207***	**0.077***
100-Venue	Baseline	0.010	0.008
	Stylometric	**0.113***	**0.050***
150-Venue	Baseline	0.007	0.005
	Stylometric	**0.099***	**0.040***

We construct RandomForest classifiers **Stylometric(A)** and **Stylometric(F)** for the CiteSeer data set, since we have both abstract and full content information for papers in this data set. Stylometric features are extracted from either abstract content or paper full content respectively. For the ACM data set where the full content of papers is missing, we work only on papers’ abstracts to generate the stylometric features. Table II shows some brief statistics over the randomly chosen venues we tested. In order to demonstrate the effectiveness of the classification results, we further construct a **Baseline Classifier** for comparison, which randomly guesses the venue label for paper instances in the testing set.

As shown in Table III and Table IV, our stylometric classifier can outperform the baseline classifier under all

TABLE IV
Multi-Class Venue Classification for CiteSeer Data Set. Value * is significantly better than the Baseline Classifier. Value † is significantly better than the Stylometric(A) Classifier

		Accuracy	F_1 Score
2-Venue	Baseline	0.498	0.485
	Stylometric(A)	0.707*	0.658*
	Stylometric(F)	**0.847***	**0.828***
5-Venue	Baseline	0.206	0.197
	Stylometric(A)	0.413*	0.342*
	Stylometric(F)	**0.625***	**0.570***
10-Venue	Baseline	0.101	0.095
	Stylometric(A)	0.254*	0.196*
	Stylometric(F)	**0.450***	**0.391***
30-Venue	Baseline	0.033	0.031
	Stylometric(A)	0.106*	0.079*
	Stylometric(F)	**0.246***†	**0.188***†
50-Venue	Baseline	0.019	0.017
	Stylometric(A)	0.066*	0.051*
	Stylometric(F)	**0.156***†	**0.116***†
100-Venue	Baseline	0.010	0.009
	Stylometric(A)	0.034*	0.028*
	Stylometric(F)	**0.094***†	**0.044***†
150-Venue	Baseline	0.007	0.007
	Stylometric(A)	0.022*	0.018*
	Stylometric(F)	**0.062***†	**0.044***†

[1] http://hcir.info/hcir-2011

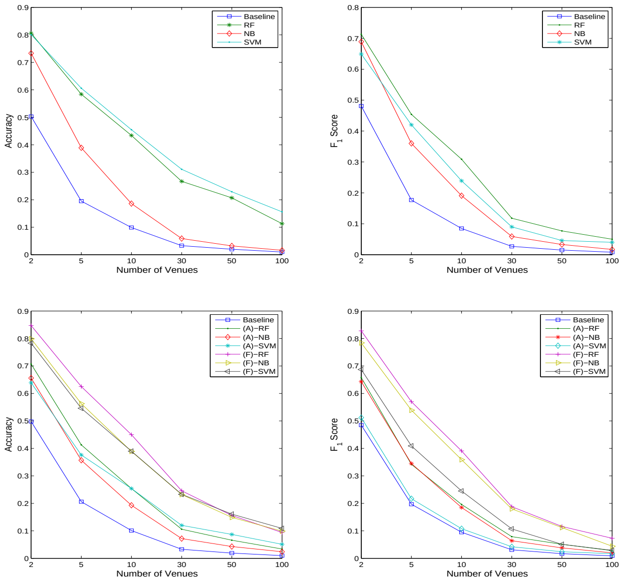

Fig. 1. Comparison of Classifiers: Accuracy and F_1 Score for ACM data (above) and CiteSeer (below).

TABLE V
ACCURACY FOR DIFFERENT FEATURE SETS AND TECHNIQUES

	ACM			CiteSeer		
	RF	NB	SVM	RF	NB	SVM
Lexical	**0.425**	0.170	0.403	**0.435**	0.315	0.355
Syntactic	0.382	0.165	**0.402**	**0.416**	0.366	0.267
Structural	**0.304**	0.131	0.291	**0.294**	0.265	0.221
Lexi+Syn	0.429	0.177	**0.433**	**0.447**	0.383	0.388
Lexi+Str	**0.423**	0.173	0.414	**0.441**	0.329	0.357
Syn+Str	0.386	0.165	**0.410**	**0.436**	0.372	0.269
Lexi+Syn+Str	0.434	0.186	**0.455**	**0.450**	0.389	0.390

TABLE VI
F_1 SCORE FOR DIFFERENT FEATURE SETS AND TECHNIQUES

	ACM			CiteSeer		
	RF	NB	SVM	RF	NB	SVM
Lexical	**0.273**	0.132	0.146	**0.382**	0.257	0.203
Syntactic	**0.224**	0.158	0.151	**0.354**	0.339	0.076
Structural	**0.109**	0.105	0.100	**0.247**	0.199	0.038
Lexi+Syn	**0.298**	0.182	0.224	**0.389**	0.349	0.240
Lexi+Str	**0.285**	0.173	0.147	**0.376**	0.274	0.207
Syn+Str	**0.247**	0.165	0.149	**0.373**	0.347	0.089
Lexi+Syn+Str	**0.309**	0.191	0.239	**0.391**	0.359	0.245

circumstances. Based on the p value computed from the students' t test, all improvement over the Baseline classifier is statistically significant ($p \leq 0.05$), which confirms that venues are distinguishable by their writing styles. Moreover, there exists a tendency to achieve greater improvement over the random guessing baseline as the number of venues tested increased. Working on CiteSeer data with paper full content, there is a 70.25% improvement for 2-venue classification, and the performance is 7.45 times over random guessing for 30-venue and 8.86 times for 150-venue respectively. We also notice from the experiment results in CiteSeer data that we can achieve better performance working on the full paper content to retrieve the stylometric features than just from paper abstracts. The improvement is statistically significant when 30 or more venues are taken as testing venues.

2) Comparison of Classification Techniques: To evaluate the classification results of different classifiers, we repeat the same experimental process as described above using three state-of-the-art classifiers: RandomForest (RF), NaiveBayes (NB), and Support Vector Machines (SVM). For the CiteSeer data set, experiments were carried out for both paper abstract (A) and full content (F) separately. We report experimental results in Figure 1. We can see that all classifiers achieve better performance than random guessing; however, different classifiers have different impacts on the performance over the two data sets.

For ACM data set, RandomForest and SVM work better than NaiveBayes for both Accuracy and F_1 Score. SVM outperforms RandomForest in terms of Accuracy, however, RandomForest can achieve higher F_1 Score than SVM.

For CiteSeer data set, all three classifiers can achieve better performance working with paper full content than paper abstract. For both working with paper abstract and full content, RandomForest performs the best with small number of testing venues, and is then outperformed by SVM when the number of venues exceeds 30 and 50 respectively. NaiveBayes is the worst in general in terms of Accuracy, however, it gradually catches up with the performance of RandomForest and SVM when the number of venues tested is increased. In terms of F_1 Score, RandomForest is the best classifier working on both data sets. NaiveBayes shows comparable performance as RandomForest. SVM turns out to be the worst of the three, whose performance is only slightly better than random guessing when working on paper abstracts.

3) Comparison of Feature Types: As introduced in previous sections, we have three groups of stylometric features: lexical, syntactic and structural. To examine the contribution of different feature sets, we first test the performance on each individual group, and then add them one by one to test the changes in performance. We fix the number of venues tested to be 10. Performance in terms of Accuracy and F_1 Score are summarized in Tables V and VI respectively.

We can see that lexical features still play the most important role in venue classification. Structural features are the least useful, probably due to our rough calculation method for collecting number of sections, number of figures, etc. However, we can also find that each group of features contributes positively to the overall performance, since when we add them together, performance is better than each individualy.

We further conducted five individual pairwise t tests in order to examine the significance of improvement. Table VII shows the p value of the t tests for feature comparison for both ACM and CiteSeer data sets. Both lexical and syntactic features work significantly better than structural features. Combining lexical

TABLE VII
P-VALUES OF PAIRWISE T TESTS ON ACCURACY FOR DIFFERENT TYPES. SYMBOL * INDICATES STATISTICAL SIGNIFICANCE

Feature Sets	ACM	CiteSeer
Lexical vs. Syntactic	0.2179	0.1264
Lexical vs. Structural	0.0018*	0.0005*
Syntactic vs. Structural	0.0035*	0.0012*
Lex vs. Lex+Syn	0.0482*	0.0407*
Lex+Syn vs. Lex+Syn+Stru	0.2210	0.1987

TABLE VIII
CONTENT VS. WRITING STYLE: ACM DATA SET. VALUE* IS SIGNIFICANTLY BETTER THAN STYLOMETRIC CLASSIFIER

		Accuracy	F_1 Score
2-Venue	Stylometric	0.806	0.713
	Content	**0.916**	**0.888**
	Combine	0.884	0.836
5-Venue	Stylometric	0.584	0.454
	Content	**0.798**	**0.706**
	Combine	0.742	0. 636
10-Venue	Stylometric	0.434	0.309
	Content	**0.657**	**0.528**
	Combine	0.595	0.444
30-Venue	Stylometric	0.267	0.118
	Content	**0.491***	**0.302***
	Combine	0.419*	0.227*
50-Venue	Stylometric	0.207	0.077
	Content	**0.407***	**0.216***
	Combine	0.342*	0.155*
100-Venue	Stylomeric	0.113	0.050
	Content	**0.280***	**0.141***
	Combine	0.217*	0.101*
150-Venue	Stylometric	0.099	0.040
	Content	0.135*	**0.085***
	Combine	**0.179***	0.074*

TABLE IX
CONTENT VS. WRITING STYLE: CITESEER DATA SET. VALUE* IS SIGNIFICANTLY BETTER THAN STYLOMETRIC CLASSIFIER. VALUE† INDICATES THAT 'COMBINE' CLASSIFIER IS SIGNIFICANTLY BETTER THAN 'CONTENT' CLASSIFIER

		Accuracy	F_1 Score
2-Venue	Stylometric(F)	0.847	0.828
	Content	0.885	**0.868**
	Combine	**0.886**	0.866
5-Venue	Stylometric(F)	0.625	0.570
	Content	0.687	0.638
	Combine	**0.691**	**0.645**
10-Venue	Stylometric(F)	0.450	0.391
	Content	0.504	0.442
	Combine	**0.516†**	**0.458†**
30-Venue	Stylometric(F)	0.246	0.188
	Content	0.270	0.211
	Combine	**0.286**	**0.225**
50-Venue	Stylometric(F)	0.156	0.116
	Content	0.187*	0.141*
	Combine	**0.191***	**0.145***
100-Venue	Stylometric(F)	0.094	0.044
	Content	0.111*	0.086*
	Combine	**0.116*†**	**0.087*†**
150-Venue	Stylometric(F)	0.062	0.044
	Content	0.075*	0.059*
	Combine	**0.079***	**0.060***

and syntactic features can provide significant improvement over pure lexical features, however, the improvement is not significant when we further add structural features. The results are consistent across the two data sets.

C. Content vs. Writing Styles

Under all experimental settings mentioned in previous sections, we work on pure stylometric features. Besides the difference in writing styles, venues also differ in their content. In order to compare the classification performance between writing-style based features and topic/content based features, we further construct the RandomForest-based **Content Classifier**, in which we represent each paper by the TF-IDF scores of the Top 500 most frequent appearing terms in the whole corpus, and the **Combine Classifier**, where we combine both stylometric and content-based features.

TABLE X
WRITING STYLES VS. TOPICS

		Accuracy	F_1 Score
SIGIR	WWW	0.730	0.729
SIGIR	CIKM	0.660	0.659
SIGIR	SIGKDD	0.755	0.755
SIGIR	JCDL	0.690	0.688
SIGIR	computer architecture	0.855	0.855
SIGIR	parallel computing	0.895	0.895
SIGIR	graphics	0.845	0.844

As shown in Tables VIII and IX, the Content Classifier works better than the Stylometric Classifier. It indicates that topic-related difference is more distinguishable than writing styles for venues. When combining both stylometric and content features, the performance is not improved on the ACM data set; however, we can get improved performance on CiteSeer data set when features over full content are integrated.

D. Topics vs. Writing Styles

Working on CiteSeer data set, we randomly select 100 papers published in the venue 'SIGIR'. We would like to test whether papers in this venue can be successfully distinguished from papers published in other venues, either with more or less similarity with the venue 'SIGIR' in terms of venue topics. We select six other venues, and randomly select 100 papers for each of them. RandomForest is used as the classifier.

We can find that papers published in similar venues can also be successfully distinguished with high probability (e.g., 73% for papers in SIGIR and WWW) based on writing style features. There shows an increase in classification accuracy when venues are talking about different topics than similar topics.

E. Genres vs. Writing Styles

We are also interested in discovering the impact of different genres of venues on similar topics in terms of their writing styles. As we already know, there exist many different genres of venues even for the same topic. For example, the journal of SIGMOD Record compared with the conference of SIGMOD in database research domain. In this group of experiments, we collect papers published in journals and conferences, and show their classification results. RandomForest is used as the classifier. As shown in Table XI, we first test on the overall performance for all journals and conferences regardless of topic difference. For doing this, we randomly select 1000 journal venues and 1000 conferences venues, collect all their published papers, and carry out the classification. As indicated, we can retrieve an accuracy over 76%. We further choose three different research domains; for each of them, we collected 100 papers published in their corresponding journal venues and conference venues respectively. Results show that in database and computer architecture domain, the classification results are better than that in the graphics domain. Even though we cannot determine exactly the effect of research topics on the classification results between journals and conferences, we can still see that on a general basis, these two are distinguishable.

F. Improving Classification Results

We have also experimented with techniques to further improve the accuracy of our classifier. Two popular techniques,

TABLE XI
WRITING STYLES VS. GENRES

Conference vs. Journal	Accuracy	F_1 Score
Overall	0.7680	0.7679
Database	0.7965	0.7949
Computer Graphics	0.5887	0.5885
Computer Architecture	0.7670	0.7668

Boosting and Bagging (Bootstrap aggregating), have been adopted, both of which essentially construct a set of classifiers which are then combined to form a composite classifier. The composite classifier is generally believed to perform better than the individual classifiers.

We apply both Bagging and Adaboost, provided by WEKA, on both ACM and CiteSeer data sets. We experimented on different numbers of venues (2, 5, 10, 30 and 50). For venues in CiteSeer data set, we also test the performance by either using only paper abstract or full content respectively. RandomForest is used as the basic classifier, and the results are also evaluated using 10-fold cross validation. We report results in terms of accuracy and F_1 score in Figure 2.

Both Bagging and Boosting provide significant improvement over the original classification results. Bagging shows better ability in improving accuracy. The improvement increases when more venues are tested. Working on 10-venue task, the improvement of Bagging is 12.44% for ACM data set, 27.56% for CiteSeer abstract and 16.4% for CiteSeer full paper content. AdaBoost, however, works better for improving the performance in terms of F_1 Score: it improves performance by 10.36% for ACM, 10.71% for CiteSeer abstract and 15.09% for CiteSeer full paper content.

VI. CONCLUSION

We addressed in this paper the task of venue classification, for which we tested whether venues are distinguishable by the writing styles of papers published in them. We applied the traditional classification approach for this task, and identified over 300 stylometric features for representing papers' writing styles. Experiments on both ACM and CiteSeer data sets demonstrated that venues can be distinguished by their writing styles. By combining both stylometric features with traditional content-based features using papers' full content, we can get improved performance for venue classification. We examined the impact of three different classifiers: RandomForest, NaiveBayes and SVM. Even though they perform differently on different experimental settings, RandomForest, however, turns out to work the best in general. We further examined the contribution of different feature sets in which lexical features were found to be the most valuable. Moreover, we carried out experiments to test the relationship between venues topics and writing styles as well as venue genres and writing styles, both of which achieved positive results on the tested venues.

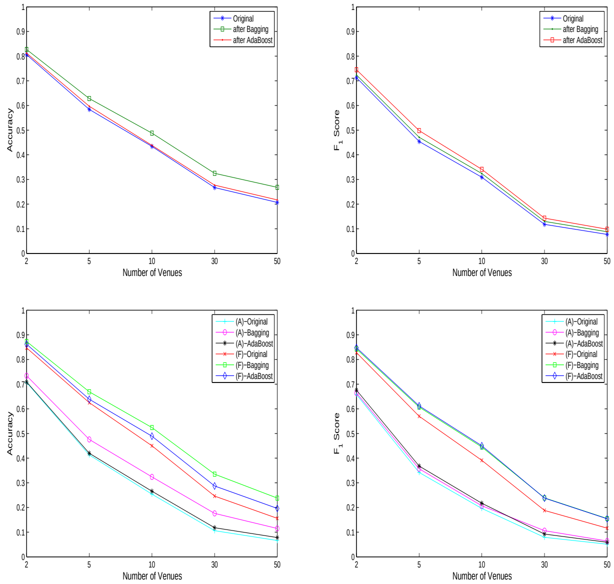

Fig. 2. Bagging and Boosting: Accuracy and F_1 Score for ACM data (above) and CiteSeer data (below).

ACKNOWLEDGMENTS

This work was supported in part by a grant from the National Science Foundation under award IIS-0545875.

REFERENCES

[1] F. Mosteller and D. Wallace, *Inference and Disputed Authorship: The Federalist.* Reading, Mass.: Addison-Wesley, 1964.

[2] J. Karlgren, "The wheres and whyfores for studying text genre computationally." in *Workshop on Style and Meaning in Larguange, Art, Music and Design. National Conference on Artificial Intelligence*, 2004.

[3] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 3, pp. 378–33, 2006.

[4] S. Argamon, M. Saric, and S. Stein, "Style Mining of Electronic Messages for Multi Authorship Discrimintation: First Results," in *KDD*, 2003.

[5] G. Yule, *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.

[6] F. Tweedie and R. Baayen, "How variable may a constant be? Measures of Lexical Richness in Perspective." *Computers and the Humanities*, vol. 32, pp. 323–352, 1998.

[7] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genres and author," *Comp. Ling*, vol. 26, no. 4, pp. 471–495, 2001.

[8] J. Burrows, "'An Ocean where each kind...': Statistical analysis and some major determinants of literay style," *Computers and the Humanities*, vol. 23, no. 4-5, pp. 309–321, 1989.

[9] D. Holmes and R. Forsyth, "The Federalist revisited: New Directions in author attribution," *Literary and Linguistic Computing*, vol. 10, no. 2, pp. 111–127, 1995.

[10] C. Chaski, "Empirical evaluations of language-based author identification techniques," in *Forensic Linguistics*, 2001.

[11] O. de Vel, "Mining email authorship," in *Proc. of the Text Mining Workshop at KDD*, 2000.

[12] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," in *In Proc. of the conference pacific association for computational linguistics*, 2003, pp. 255–264.

[13] O. Feiguina and G. Hirst, "Authorship attribution for small texts: Literary and forensic experiments," in *Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection*, 2007.

[14] J. Farringdon, A. Morton, M. Farringdon, and M. D. Baker, *Analysis for Authorship: A Guide to the Cusum Technique.* University of Wales Press, 1996.

[15] D. Lowe and R. Matthews, "Shakepeare vs. Fletcher: A stylometric analysis by radial basis functions," *Computers and the Humanities*, vol. 29, no. 6, pp. 449–461, 1995.

[16] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship Attribution with Latent Dirichlet Allocation," in *CoNLL*, 2011, pp. 181–189.

[17] R. Arun, R. Saradha, V. Suresh, and M. Murty, "Stopwords and Stylometry: A Latent Dirichlet Allocation Approach," in *NIPS workshop on Application for Topic Models: Text and Beyond*, 2009.

[18] G. Zipf, *Human Behaviour and the principle of least effort. An introduction to human ecology.* Houghton-Mifflin, 1932.