Distinguishing Venues by Writing Styles

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

ABSTRACT

A principal goal for most research scientists is to publish. There are different kinds of publications covering different topics and requiring different writing formats. While authors tend to have unique personal writing styles, no work has been carried out to find out whether publication venues are distinguishable by their writing styles. Our work takes the first step into exploring this problem. Using the traditional classification approach and carrying out experiments on real data from the CiteSeer digital library, we demonstrate that venues are also distinguished by their writing styles.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithms, Performance

Keywords: classification, features

1. INTRODUCTION

For research scientists, a fundamental task is to publish their work. There are many different kinds of publications 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 journal publications such as J.ASIST, as well as conferences, such as SIGIR, JCDL, WWW, CIKM, WSDM, 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 [5], 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. It has long been recognized in author attribution (verification or identification) that writing styles are one of the key features in distinguishing among authors. The earliest work was conducted by Mendenhall [6] in the nineteenth century who studied authorship attribution among Bacon, Marlowe and Shakespeare. Much subsequent work has demonstrated that authors tend to have personal writing styles [7, 8].

However, no work has been carried out, to the best of our knowledge, investigating whether *venues* are also distinguishable by their writing styles. This task is actually equivalent to the question as to whether the papers published in one specific venue share common characteristics in writing styles, and how are they distinguishable from papers published in other venues. We approach this problem by using classification, in which 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. Specifically, we are interested in determining the extent to which venues can be distinguished from each other in terms of writing styles and what features are valuable for that purpose.

JCDL'12, June 10–14, 2012, Washington, DC, USA.

2. FEATURES

Since we focus on writing-style based venue classification, one of the main concerns is to extract features that are unrelated to topic and context-free. Based on the review of previous studies in the task of author attribution, we incorporate three types of features: lexical features, syntactic features and structural features. The entire set of features are listed in Table 1.

Lexical Features: Lexical features can be further divided into character-based or word-based features. It reflects a paper's preference for particular types of characters or words. In our work, we include features like number of terms, number of distinct terms, vocabulary richness [3], Hapax terms, etc., resulting in a total of 66 lexical features.

Syntactic Features: The discriminating power of syntactic features is derived from different formats and patterns in which sentences of a paper are organized. One of the most important syntactic features is the short yet all-purpose words, which are often referred to as the function words [4], such as 'the', 'a', 'and', 'to' etc. Another example syntactic feature is punctuation which is considered as the graphical correlation of intonation that is the phonetic correlation of syntactic structure [2]. We adopt a set of 298 function words, and compute the count of eight predefined punctuation symbols that appear in the paper.

Structural Features: Structural features represent the layout of a piece of writing. De Vel [1] introduced several structural features specifically for email. We adopt five structural features specifically for scientific papers: the number of sections, figures, equations, tables, and bibliographic references. Since the original paper content

Table 1: Features

Type	Features		
Lexical	TokenNum	TypeNum	CharNum
	SentenceNum	AvgSenLen	AvgWordLen
	ShortWordNum	HapaxVSToken	HapaxVSType
	ValidCharNum	AlphaCharNum	DigitalCharNum
	UpperCaseNum	WhiteSpaceNum	SpaceNum
	TabSpaceNum	VocabularyRichness	
Syntactic	FuncWordNum	PunctuationNum	FuncWordFreq
Structural	SectionNum	FigureNum	EquationNum
	TableNum	ReferenceNum	

available is in raw text format, we approximate the values by counting the number of times the word 'figure' or 'Figure' appears in the paper. We do the same for number of sections, number of tables and number of equations.

In summary, we have collected a total of 371 features.

3. EXPERIMENTS

3.1 Data Corpus

We carried out experiments on the CiteSeer digital library of scientific literature, which was distributed by the 2011 HCIR challenge workshop[1]. The corpus is divided into two parts. Meta-data about a paper, such as its title, publishing venue, publishing year, abstract, citation references are kept in XML format; the full content of that paper is in pure text format. We extract 119,727 papers published between 1949 and 2010, which have abstract, full content and venue information, and 48,797 venues that have at least one paper with full content provided.

3.2 Overall Performance

We are firstly interested in finding out whether venues are distinguishable by their writing styles in general. For multi-venue classification testing, we randomly choose K venues (where K varies among 2, 5, 10, 30, 50, 100 and 150) that have at least 50 papers in the CiteSeer data set. We collect all the papers published in those chosen venues to construct the training/testing set. The same process is repeated ten times for each particular K, and the result is the average of the ten iterations.

Three state-of-the-art classifiers (SVM, Naive Bayes, RandomForest) provided by WEKA were constructed for classifying, and 10-fold cross validation was used for evaluation. We adopt two traditional IR evaluation metrics, Accuracy and F-1 Score to measure the classification performance.

We compared performance among several classifiers. The **Baseline Classifier** randomly guesses the venue label for each paper instance in the testing set. We extract stylometric features from either the abstract or the full content of papers to construct the **Stylometric (A) Classifier** and **Stylometric (F) Classifier** respectively. Moreover, working on paper full content, we compare the performance under RandomForest, Naive Bayes and SVM classifiers.

As shown in Figure 1, several observations can be found. 1) The classification performance continues to drop as the number of venues considered is increased. However, under all circumstances, our stylometric classifier can retrieve better performance than the baseline classifier. A student's t test indicates that the improvement is statistically significant, which confirms that venues are also distinguishable by their writing styles. 2) There exists a tendency to achieve greater improvement over random guessing as the number of venues tested increased. Working on paper full content with RandomForest, there is a 70.25% improvement for 2-venue classification, and the performance is 7.37 times over random guessing for 10-venue and 8.83 times for 150-venue respectively. 3) We can achieve better performance when working on papers' full content than working only with abstracts. It is reasonable since more signs of writing styles can be presented when more paper content is included. 4) All three classifiers achieve consistent classification results. RandomForest works the best with small number of venues considered, however, SVM outperforms it when the number of venues exceeds 50.

3.3 Importance of Features

To determine the different contributions of lexical, syntactic, and structural groups of features, we first test performance on each individual group, and then add them one by one to determine the changes in performance. Table 2 shows the results when considering on 10 Venues using RandomForest classifier.

We can see that lexical features still play the most important role in venue classification, while structural features are least useful. However, we can also find out that each group of features contributes positively to the overall performance, since when we add them together, the performance is better than each individually. We further investigate the importance of each individual feature by comparing the classification results based on the feature set that leaves out the tested target feature. Working on 5-venue set with RandomForest Classifier, we can find out that the three best features are: FuncWordFreq, TypeNum and TokenNum.

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

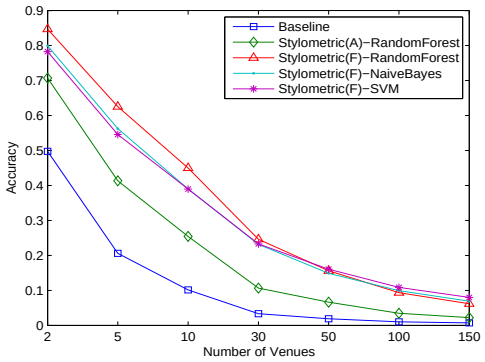

Figure 1: Classification Result: Accuracy

Table 2: CiteSeer Data Set: Feature Contribution

	Accuracy	F-1 Score
Lexical Only	0.4348	0.3822
Syntactic Only	0.4164	0.3543
Structural Only	0.2938	0.2468
Lexi+Syn	0.4474	0.3889
Lexi+Str	0.4405	0.3756
Syn+Str	0.4355	0.3731
Lexi+Syn+Str	0.4502	0.3906

Acknowledgments

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

4. REFERENCES

[1] O. de Vel. Mining Email authorship. In *Text Mining Workshop. KDD*, 2000.

[2] O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4):55–64, 2001.

[3] G.Zipf. *Human Behaviour and the principle of least effort. An introduction to human ecology*. Oxford, England: Addison-Wesley Press, 1949.

[4] D. Holmes and R. Forsyth. The Federalist revisited: New Directions in author attribution. In *Library and Linguistic Computing*, pages 111–127, 1995.

[5] 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.

[6] T. Mendenhall. The characteristics curves of composition. *Science*, 11(11):237–249, 1887.

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

[8] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic text categorization in terms of genres and author. *Comp. Ling*, 26(4):471–495, 2001.