CWAAP: An Authorship Attribution Forensic Platform for Chinese Web Information

Jianbin Ma[1], Ying Li[2], Guifa Teng[1]
[1]College of Information Science and Technology, Agricultural University of Hebei, Baoding, China
[2]College of Economic and Trade, Agricultural University of Hebei, Baoding, China
Email: majianbin@hebau.edu.cn, sxliying@hebau.edu.cn, tguifa@hebau.edu.cn

***Abstract*— Illegal web information is common on the Internet. To prevent phenomena of illegal web information from happening, providing effective evidence for court to punish the criminals by means of law is one effective method. In this paper, an authorship attribution platform for Chinese web information, CWAAP, is described. Based on the language characteristics of Chinese web information, lexical features and structural features which can express the author's writing habit are extracted. Support vector machines (SVM) are used for learning author's writing features. To test the effectiveness of CWAAP, literature, Blog and BBS datasets are used in the experiments on the platform. Five experiments are performed. Experimental results show that lexical features and structural features are effective. The number of words in training samples should exceed 200 at least. By Information Gain feature selection methods, 800 lexical features can express the authors' writing style. There is a small difference between the authors' topics. All the parts of speech reserved are perfect. These results confirm that the platform is effective and feasible for cybercrime forensic.**

***Index Terms*—CWAAP, Authorship attribution, Forensic, Support Vector Machine, Chinese, Web information**

I. INTRODUCTION

Various Internet service such as E-mail, BBS, Blog, Microblog has been widely applied to people's daily life. While Internet provides convenient to people, a lot of problems appear at the same time. Some illegal web information, such as antisocial information, fraud information, pornographic information, terroristic threatening information, gambling information, appears by means of E-mail, BBS or Blogs. The Internet provides criminals new criminous space and means. Illegal web information affects social stabilization and national security seriously. Some measures should be taken urgently. Now, installing filtering software to filter the information containing sensitive words is the main method to prevent these phenomena. However, this passive defensive method cannot stop these phenomena, because criminals can make use of some substitute words to break through the defense of the filtering software. Punishing the criminals by means of law can strike these crimes effectively. Many states have made interrelated laws. However, due to lacking effective evidence, many cases cannot be brought to the court. If web information's authorship is attributed by technical means, criminal's evidence for computer forensic can be collected. This will provide an important application value and practical significance to law enforcement, social safety and stabilization, Internet environments' purification.

Footprint, handwriting, signalment have been used to obtain evidence for courts. But the evidence of criminals via Internet is difficult to collect, because Internet is a free and open place and the messages on the Internet are spread anonymously. Criminals being hidden in any online corner can commit a crime. Though Internet services such as e-mail or BBS require users to fill out their personal information when registering, criminals always forge their real information or log on anonymously. So registering information, IP address, and e-mail's header information cannot provide convincing evidence for the court. However the text's content and structure can be obtained from web information, the same as their handwriting. Authors of web information have their inherent writing habits. The writing habits cannot be changed easily (although the criminals will always try), which embody a writing style such as usage of certain words, the length of sentences and paragraphs, and the format of the text.

Stylometry is the application of the study of linguistic style, usually to written languages. Stylometry is the theoretical basis of authorship attribution which attributes authorship of unidentified writing on the basis of stylistic similarities between the authors' known works and the unidentified piece. Researchers have focused on academic and literary applications ranging from the questions of the authorship of Shakespeare's works to forensic linguistics. The research language of authorship attribution has been mainly English, Arabic, and Japanese etc. However, there were little related authorship attribution researches on the Chinese language. The language characteristics of the Chinese language are very different from other languages such as English and Indo-European languages, where the feature extraction methods for authorship attribution are different. In this

Manuscript received September 1, 2012; revised June 5, 2013; accepted June 20, 2013.
Corresponding author: Guifa Teng

doi:10.4304/jsw.9.1.11-19

paper, an authorship attribution platform for Chinese web information, CWAAP, was introduced and described. Based on the language characteristics of Chinese web information, various authors' writing features including lexical features and structural features which could express the authors' writing habits were extracted. Support vector machines (SVM) were used for learning the writing features.

The remainder of the paper is organized as follows. Section 2 presents a general review of stylometry and previous related work. Section 3 describes the framework of CWAAP. Section 4 is our feature selection and extraction methods. Section 5 provides our experimental methodology and analyses the experimental results. Section 6 draws the conclusions of the paper.

II. Related Work

A. Stylometry and Authorship Attribution

Stylometry is the study of the unique linguistic styles and writing behaviors of individuals in order to determine the authorship. It is an interdisciplinary study of statistics and computer science etc. The research of stylometry is based on the premise of two assumptions. The first assumption is that all authors have distinctive writing habits, which can be captured from a number of quantitative features such as certain vocabulary usage, sentence complexity, and phraseology. The second assumption is that these habits are unconscious. Even if some authors make a conscious effort to disguise one's writing habits, the effect is not obvious. Stylometry focuses on defining authors' subconscious writing features and determining statistical methods to measure these features so that the similarity between two or more pieces of text can be analyzed.

Stylometry is the basis of authorship analysis. Authorship analysis can be divided into three distinct problems, namely, authorship attribution, authorship characterization, and plagiarism detection. The aim of authorship attribution is to determine the author of a piece of text by comparing the similarity of writing style between the author's known works and unknown ones. Authorship characterization attempts to formulate author's sociolinguistic profile by making inferences about gender, educational, and cultural background on the basis of writing style. The purpose of plagiarism detection is to calculate the similarity of two or more pieces of text and to determine if a piece of text has been plagiarized.

The following are several typical authorship attribution studies. "The Federalist Papers" are a series of 85 articles or essays serially in The Independent Journal and The New York Packet between October 1787 and August 1788 with the aim of advocating the ratification of the United States Constitution. 12 articles have disputed authorship between Hamilton and Madison. Pioneered authorship attribution methods were famously used by Mosteller and Wallace in the early 1960s to attempt to answer this question. Frequencies of a set of function words selected from articles were compared. Mosteller and Wallace(1964)[1] came to the conclusion that the 12 disputed articles were written by Madison. Another well-known study is the attribution of disputed Shakespeare works. Elliot (1991)[2] compared the writing style of Shakespeare's work "Earl of Oxford". The writing style included unusual diction, frequency of certain words, choice of rhymes, and habits of hyphenation. "And Quite flows the Don" was written by Sholokhov between 1928 and 1940. Sholokhov was accused of plagiarizing from Kryukov. Kjetsaa(1979)[3] draw the conclusion that Sholokhov was the true author of "And Quiet Flows the Don" by comparing the statistical features of Sholokhov and Kryukov. The features included the length of sentences, part of the speech, sentence structure etc. "Dream of the red chamber" is a masterpiece of Chinese literature and is generally acknowledged to be the pinnacle of classical Chinese novels. For a long time, the first 80 chapters written by Cao Xueqin and the 40 additional chapters written by Gao E were recognized universally. Professor Chen Bingzao at university of Wisconsin researched on the authorship of "Dream of the red chamber" for the first time. Computers were used to calculate and analyze the frequency of words occurring in the masterpiece. He came to the conclusion that all the 120 chapters were written by Cao Xueqin.

B. Authorship Attribution Features

The frequency of certain word-usage, the length of sentences etc can be used to attribute authorship. The former researchers have focused on what the features could represent the writing style of authors. However, no fixed features set were agreed on. The following is several types of features.

(1) Word-length and Sentence-length

The origins of stylometry might be traced back to the work of Mendenhall (1887)[4] on word-lengths. Morton (1968)[5] used sentence-lengths for tests of authorship of Greek prose.

(2) Function Words

Word-usage was usually used for discrimination in authorship of texts. In the same author's work, some words vary considerably in their rate of use, while other words show remarkable stability. Function words were used to attribute the author of "The Federalist Papers" by Mosteller and Wallace (1964)[1]. Morton (1978)[6] developed techniques of studying the position and immediate context of individual word-occurrences. However the method had come under much criticism and Smith (1985)[7] had demonstrated that it could not reliably distinguish between the works of Elizabethan and Jacobean playwrights. Burrows (1987)[8] proposed the common high-frequency words (at least 50 strong). Holmes and Forsyth (1995)[9] had successfully applied the technique to the classic "The Federalist Papers" problem.

(3) Vocabulary Distributions

One of the fundamental notions in authorship attribution is the measurement of richness of an author's vocabulary. The frequency of word-usage can be estimated by analyzing a text produced by a writer. Mathematical models for the frequency distributions of

the number of vocabulary items appearing exactly r times (r=1,2,3....) have aroused the interest of statisticians ever since the work of Zipf (1932)[10]. The best fitting model attributed to Sichel (1975)[11], and the Sichel model in addition to the once-occurring words (hapax legomena) and twice-occurring words (hapax dislegomena) were useful stylometric tools.

C. Authorship Attribution Methods

Some technical means have been used to analyze the writing features to arrive at the purpose of attributing a text's authorship. Mathematical methods and intelligent algorithms were adopted. The techniques vary with different periods. The following is summary of three common authorship analysis approaches.

(1) Probabilistic and Statistical Approaches

Efron and Thisted (1976)[12] considered how many words Shakespeare knew. Probabilistic techniques were used to study the number of words used once, twice in the Shakespeare canon. A parametric empirical Bayes model and a nonparametric model were examined. The models supposed that Shakespeare knew at least 3,5000 more words, which could be regarded as evidences of Shakespeare's authorship. Smith (1983)[13] selected the average word-length, the average sentence-length, collocations, and measures of words in certain positions in sentences as features. Chi squared statistic methods were used to detect differences between Shakespeare and Marlowe. Farringdon (1996)[14] used the Cusum technique to test authorship of a small number of text samples.

(2) Computational Approaches

With the development of computer technology, sensitive classification techniques rather than simple count statistics have been applied to authorship attribution. Burrows (1992)[15] analyzed the frequency of words. The Pearson product-moment method correlated each word with all others. Principal component analysis methods were used to transform the original variables to a set of new uncorrelated variables. Holmes (2001)[16] described how traditional and non-traditional methods were used to identify seventeen previously unknown articles that were believed to be written by Stephen Crane. 3000 word samples of text were analyzed for frequencies of 50 common words. Principal component analysis was used as the method of discrimination.

(3) Machine Learning Approaches

Machine learning is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. In recent years, machine learning approaches have been applied to stylometry. Neural network classifiers were employed for stylometry by Merriam and Matthews (1994)[17]. Kjell (1994)[18] used neural networks and Bayesian as classifiers. Hoorn et al. (1999)[19] used neural network with letter sequences as the feature set for authorship analysis of three Dutch poets. Holmes (1998)[20] compared the effects of vocabulary richness, and word frequency analysis with a genetic rule based learner on the problem of attributing "The Federalist papers".

D. Web Information Authorship Attribution for Forensic Investigation

Authorship analysis has been widely used in resolving authorship attribution of literary and conventional writing. With increasing cybercrime arising in Internet, web information authorship attribution began to draw researchers' attention.

E-mail is a special type of web information. With rapid growth of e-mail misuse phenomena, E-mail authorship analysis has been researched for forensic investigation. De Vel (2000, 2001)[21-23] applied the support vector machine classification model over a set of linguistic and structural features for e-mail authorship attribution for the forensic purpose. Tsuboi (2002)[24] studied authorship attribution of e-mail messages and World Wide Web documents written in Japanese. The sequential word patterns or word n-grams with n=2 and 3 from each sentence in the documents was used as features set. Zheng (2003, 2006)[25-26] analyzed the authorship of web-forum, using a comprehensive set of lexical, syntactical, structural features, and content-specific features. Abbasi (2005, 2006, 2008)[27-29] analyzed the authorship identification and similarity detection of web information. Iqbal (2008)[30] mined write-prints called frequent patterns for authorship attribution in e-mail forensic.

The above researches are for English, Japanese, and Arabic documents' authorship analysis. However, techniques of authorship analysis used for feature extraction are dependent on languages, and in fact differ dramatically from one language to another. For example, Chinese does not have word boundaries explicitly in texts. In fact, word segmentation itself is a difficult problem in the Chinese-like languages. So feature extraction methods for Chinese documents are different from other languages such as English and other Indo-European languages.

III. THE FRAMEWORK OF CWAAP

Figure 1 presents the framework of CWAAP (Chinese web information authorship attribution platform). According to the process of Chinese web information authorship attribution, there are six steps, namely information collection, information pre-processing, Chinese word segmentation, feature selection and extraction, authorship training, and authorship attribution.

The precondition of authorship attribution is that the web information of suspected authors can be obtained. We assume that there is enough web information of suspected authors. By analyzing the known author's web information, the author's writing style is gained. Then the author of unidentified information can be attributed. So the first step of authorship attribution is to collect web information of suspected authors as much as possible.

There are many categories of web information, such as e-mail, BBS, and Blog. The object of authorship attribution is the text of web information. Disorderly information such as photos, sound, and advertising

information should be removed. So it is necessary to pre-process the web information and leave the useful texts of web information to be analyzed.

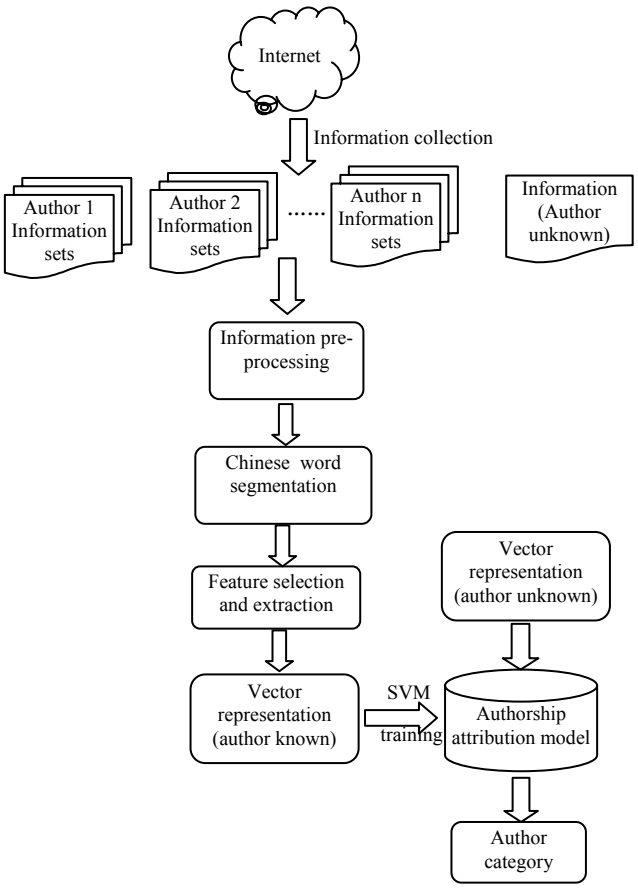


Figure 1. The framework of CWAAP

Different from the English language, Chinese does not have clear natural word segmentation markers. The lexical features are main writing features to extract. The precision of word segmentation relates to the effect of feature extraction. Now a lot of Chinese word segmentation software packages are available for use. However, the latest appearing words such as newbie are difficult to segment correctly. In CWAAP, word segmentation software named segtag developed by Professor Xiaodong Shi at Xiamen University was used for word segmentation and part of speech tagging. An additional dictionary was used to supply the new appearing words. In the case of incorrect word segmentation, the platform provides adjustment functions manually.

What features set can represent web information authors' writing style is the next step. In the feature extraction step, the extensive stylometric features including lexical features, structural features were extracted. The writing features were represented by the vector space model (VSM). Thus one web information document was denoted as a dot in the high dimensional space.

Machine learning techniques including decision trees, neural networks, and support vector machines(SVM) are the most common analytical approaches used for authorship attribution in recent years. The distinctive advantage of the SVM is its ability to process many high-dimensional applications such as text classification and authorship attribution. Zheng (2006)[25] and Abbasi (2005)[27] have drawn the conclusion that SVM significantly outperform neural networks and decision trees in authorship analysis. In our study, support vector machines were used for learning the authors' writing features, and authorship attribution model was gained.

The unknown authorship of web information could be attributed automatically by the authorship attribution model that was trained in authorship training step.

IV. Feature Selection and Extraction

A. The Characteristics of Chinese Web Information

As the particular pictograph in the world, Chinese language is highly uniform and canonical. Compared with English and other European languages, Chinese has the following characteristics.

(1) Form: blank spaces are regarded as the delimiters of words in English texts. Chinese texts don't have natural delimiter between words.

(2) Syntax: the components of the sentence depend on word order and empty word. Maybe the same of word order has different meanings.

(3) Glossary: In English language, words are composed of 26 letters. Sentences consist of several words. In Chinese language, there are above 90,000 Chinese characters totally. Just the commonly used Chinese characters amount to above 7,000. There are many more words than the characters, because words are composed of several characters.

With the popularization of Internet, cyber-language begins to spread, which has struck the criterion of the traditional languages. Cyber-language is free in use and not restricted with grammar. The style of cyber-language has the following characteristics.

(1) The words are typed into the computer's screen by keyboard. So to save typing time, users do not obey the rules of the usual writing. Elliptical sentences and incomplete sentences are common in web information. Furthermore, the writing is free in the Internet. A lot of blank line and blank spaces are inputted at will. The sentences are brief. Sentences usually consisting of two or three words are common.

(2) Writing in the Internet doesn't obey the rules of punctuation. Interrogation marks, exclamatory marks, and suspension points are used frequently. Authors input a succession of exclamatory mark when they approve others viewpoint and input several suspension points when they do not understand others viewpoint.

(3) New words appear in the Internet frequently. They are spread by egregious speed. For example, the word geili has been popular in Internet and everyday communication since 2010.

Authors writing in the Internet have formed fixed writing styles. Grasping web information authors' writing style is easier than literary writing. Based on analyzing the characteristics of Chinese web information, the

author's writing features were divided into two types, namely, lexical features and structural features.

B. Lexical Features Extraction and Selection Methods

The frequency of certain words reflects author's preference or habit for usage of some specific words. In our study, the frequency of certain words was expressed as lexical features.

Lexical features could be extracted by tf-idf techniques which had been used in the research of text classification. The weight of lexical features was calculated as formula 1.

$$W(t,\vec{d}) = tf(t,\vec{d}) \times \log(N / n_t + 0.01) \quad (1)$$

where $W(t,\vec{d})$ is the weight of term t in document d, $tf(t,\vec{d})$ is the frequency of term t in document d, N is the total number of documents, n_t is the number of documents that contain term t.

If all the words were treated as lexical features, the number of features can reach thousands of the dimensions. But some features are useless, which can waste storage space and result in system degradation. In our study, information gain (IG) feature selection method was adopted to select effective features. The information gain of lexical features was calculated as formula 2.

$$Gain(w) = -\sum_{i=1}^{m} p(c_i) \log P(C_i) + p(w) \sum_{i=1}^{m} p(c_i / w) \log P(c_i / w) + P(\overline{w}) \sum_{i=1}^{m} p(c_i / \overline{w}) \log p(c_i / \overline{w}) \quad (2)$$

where w denotes a certain feature. m is the number of classes. c_i denotes one certain class. $\overline{w}$ denotes that the feature w doesn't appear. $P(w)$ denotes the probability that the feature w appears. $P(\overline{w})$ denotes the probability that the feature w doesn't appear. $P(c_i \mid w)$ denotes the probability that the document belongs to class c_i on condition that the document contains feature w. $P(c_i \mid \overline{w})$ denotes the probability that the document belongs to class c_i on condition that the document does not contain feature w.

C. Structural Features Extraction Methods

In web texts, authors always ignore some punctuations or use incorrect punctuations. The authors can write freely on the premise of expressing the author's meaning. So the structure of web texts is loose. Furthermore, the authors have a preference for part of speech usage which can reflect the authors' degree of education. So we extracted three aspects of structural features, namely, punctuations features, structural characteristics, and part of speech features. Table Ⅰ shows the structural features.

The web text should be inputted by keyboard. At the same time, authors always ignore the difference of Chinese and English punctuation, which can treated as writing habits to extract. Table Ⅱ is the punctuation features. The weight of punctuation features is the ratio of the number of a particular punctuation in the document to the total number of punctuations in document.

The rate of parts of speech can reflect the preference for word class usage. For example, some authors always use exclamation, however some authors hardly ever. The usage of parts of speech can reflect the authors' degree of education. Chinese has 12 categories parts of speech in common use which are listed in table Ⅲ. The weight of parts of speech is the ratio of number of the parts of speech in the document to total number of parts of speech in the document.

TABLE I.
STRUCTURAL FEATURES

Features
Number of distinct punctuations/total number of punctuations
Number of distinct words/total number of words
Mean sentence length
Mean paragraph length
Number of digital characters/total number of words
Number of lowercase letters/total number of words
Number of uppercase letters/total number of words
Number of space/total number of words
Number of blank lines/total number of lines
Number of indents/total number of words

TABLE Ⅱ.
THE PUNCTUATION FEATURES

Chinese Punctuations			English Punctuations		
——	…	。	,	.	,
、	；	：	?	:	?
!	“	”	〔	(	)
〕	《	》	•	“	!
.	‘	’	–	;	‘

TABLE Ⅲ.
THE PUNCTUATION FEATURES

Number	Features	Number	Features
1	noun	7	adverb
2	verb	8	preposition
3	adjective	9	conjunction
4	numeral	10	auxiliary
5	quantity	11	exclamation
6	pronoun	12	onomatopoeia

V. EXPERIMENTS ON CWAAP

Figure 2. The main interface of CWAAP

CWAAP was developed in Visual c++ development environment. The operating system was Windows XP. Other software tools including segtag(Chinese word segmentation software package), libsvm-2.9(support vector machine software package) were used. The system was composed of four modules, which were web information's content extraction, web information's features extraction, web information's authorship training, and web information's authorship attribution. Figure 2 shows the main interface of CWAAP.

A. Datasets and Experimental Methods

To test the effectiveness of CWAAP, three datasets including literature, BBS, and Blog were collected, and several experiments were made. The detail information of the three datasets was showed in table Ⅳ.

TABLE IV.
THE INFORMATION OF THREE DATASETS

Dataset	Number of authors	Average number of documents	Document size(words)	
			Min	Max
literature	9	51	12	21334
blog	7	198	11	4046
BBS	6	72	3	489

The literature dataset was collected from one online books library. The blog dataset came from the website http://blog.sina.com.cn/. We gained the BBS dataset from one web forum. Every dataset's author was different from others.

A linear kernel function was used as the kernel function of support vector machine. Since there were only a small amount of data to produce a model of authorship attribution, the experiments results were measured by k-fold cross-validation to provide a more meaningful results. Accuracy was used to evaluate the experimental results. Five experiments were performed. The first experiment is to test the validity of two types of features. The second experiment is to test the effect of document size on experimental results. The third experiment is to test the effect of number of lexical features on experimental results. The effect of author's topics was tested in the fourth experiment. Different parts of speech were tested in the fifth experiment.

B. Experimental Results and Discussions

(1) The first experiment

To test whether the two types of features extracted in our study are effective, the first experiment was made. Different features and features combination on different dataset were tested. 1000 lexical features were selected by the IG features selection method in formula 2. 5-fold cross-validation was used to validate the experimental results. The experimental results are showed in table V.

From table V., we can see that accuracy of lexical features on literature, BBS and Blog dataset were 77.02%, 56.26% and 80.51% respectively. Accuracy of structural features was 94.97%, 62.86% and 84.35% respectively. Accuracy of combination of lexical and structural features was 95.62%, 70.99% and 89.06% respectively. Accuracy of structural features on all dataset was higher than lexical features, which proves that structural features were one effective feature. Accuracy of combination of lexical and structural features was higher than structural features, which shows that the combination of lexical and structural features was more effective than lexical features or structural features singly. The accuracy exceeded 80% by experimenting on Blog datasets. The accuracy of BBS dataset was low, which might be caused by too few words in BBS document.

TABLE V..
EXPERIMENTAL RESULTS OF DIFFERENT FEATURE COMBINATION ON DIFFERENT DATASET

Dataset	Feature type	Accuracy(%)
literature	T_L	77.02
	T_S	94.97
	T_{L+S}	95.62
BBS	T_L	56.26
	T_S	62.86
	T_{L+S}	70.99
Blog	T_L	80.51
	T_S	84.35
	T_{L+S}	89.06

T_L：lexical feature T_S：structural feature
T_{L+S}：lexical+ structural feature

(2) The second experiment

The former study on authorship attribution needs 1000 words in one sample at least, which can express author's writing style better. However, the number of words in web information is small. Two or three words in BBS or E-mail texts are common. How many words in one document can be used to attribute authorship reliably? We made experiments on the literature dataset. Three authors' samples were experimented. Every author had 30 samples. The number of words in samples was 50, 100, 200, 500, and 1000. 5-fold cross-validation was used to validate the experimental results. The experimental results that were measured as table VI.

TABLE VI
THE EXPERIMENT RESULTS OF DIFFERENT NUMBER OF WORDS

Number of words	Accuracy(%)
50	85.78
100	88.89
200	95.58
500	97.53
1000	98.82

From table VI, we could see that the accuracy increased with the increase of words in samples. That was because the more words in samples, the writing style could be expressed better. The experimental results showed that the accuracy did not have distinct change when the number of words exceeded 200. Conclusion could be draw that words in samples reached 200 could be used to attribute web information's authorship.

(3) The third experiment

To test IG feature selection method, the number of lexical features from 100 to 2000 was tested on Blog dataset. 5-fold cross-validation was used to validate the experimental results. Table VII and figure 3 were the experimental results.

TABLE VII.
THE EXPERIMENTAL RESULTS OF DIFFERENT NUMBER OF LEXICAL FEATURES

Features	100	200	300	400	500	600	700	800	900	1000
Accuracy	60.14	71.16	76.38	76.09	79.49	78.7	80.36	79.57	80.65	80.51
Features	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000
Accuracy	80.36	80.50	79.71	82.03	82.10	81.09	82.75	80.94	81.88	82.75

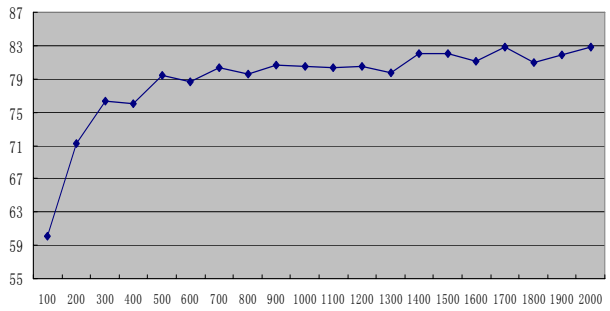


Figure 3. The experimental results of different number of lexical features

The table Ⅶ and figure 3 show that the overall trend of results was ascending in the rough, though waves occurred in the course. The accuracy of 100 lexical features was low, which proved that 100 lexical features could not express authors' lexical writing style adequately. There was not distinct change in experimental results when the number of lexical features reached 800. Too few lexical features could not express the author's writing style adequately. Too many lexical features might result in storage space wasting and system performance degradation, and improve little on the results.

(4) The fourth experiment

If authors' writing topic is same, their lexical usage is similar. Whether their writings are not easy to differentiate is concerned about. The experimental results of lexical features in Blog dataset comparing four authors remarking on the entertainment topic with two authors remarking on the entertainment topic and two authors remarking on the sports topic were given. The results were validated by 5-fold cross-validation. 1000 lexical features selected by IG methods were extracted. The experimental results were showed in table VIII.

TABLE VIII
THE EXPERIMENTAL RESULTS OF DIFFERENT AUTHOR'S TOPIC

Authors' topic	Accuracy(%)
two authors(entertainment topic) two authors(sports topic)	86.38
four authors(entertainment topic)	84.19

Table VIII showed that there were a small difference between the same authors' topic and different authors' topic. That was because that authors' topic was embodied in noun or verb. Some other parts of speech could express authors' writing style well.

(5) The fifth experiment

For text classification, the useless empty words are removed. The substantives such as nouns, verbs, and adjectives are accepted. However, conjunctions, prepositions, and adverbs are useful for attributing authorship. The fifth experiment was concerned about whether all the parts of speech should be reserved to attribute authorship. 1000 features of the Blog dataset selected by the IG method were extracted as features. The results were validated by 5-fold cross- validation. Table Ⅸ showed the experimental results of different parts of speech.

TABLE IX
THE EXPERIMENTAL RESULTS OF DIFFERENT PARTS OF SPEECH

Part of speech	Accuracy(%)	Part of speech	Accuracy(%)
noun	67.25	preposition	40.58
verb	65.72	quantity	40.00
adverb	60.07	auxiliary	35.29
conjunction	53.70	modal particle	27.10
adjective	48.33	n+ v + adv + c + adj + p	76.81
pronoun	55.72	rest	64.64

n+ v + adv + c + adj + p: noun, verb, adverb, conjunction, adjective and pronoun reserved
rest: the rest part of speech reserved except noun, verb, adverb, conjunction, adjective and pronoun

From tableIX, we could see that nouns, verbs, adverbs, conjunctions, adjectives and pronouns tested solely were better. The accuracy of preposition, quantity, auxiliary, modal particle was lower. However, the accuracy that every part of speech tested solely did not exceed the accuracy that all part of speeches reserved. Except for the above six types of part of speech, the accuracy of the rest part of speech was 64.64%, which showed that the rest part of speech had discrimination ability. The fifth experiment showed that the results of all the part of speech reserved were perfect. Some empty words had discrimination ability for attributing authorship, which should be reserved, differing from text classification.

VI. CONCLUSIONS

The crimes utilizing Internet increase rapidly. For the purpose of providing evidences for the court, CWAAP, an authorship attributing platform for Chinese web information was developed. In this paper, the framework of the system was provided. Two types of features including lexical features and structural features were extracted. To test the effect of CWAAP, three datasets were collected. Five experiments were designed and performed. Experimental results proved that the two features extraction methods were effective. The number of words in samples used for authorship attribution exceeded 200 at least. By IG feature selection methods, 800 lexical features could express the authors' writing style. There was a small difference between the authors' topics. All the parts of speech reserved were perfect. The accuracy exceeded 80% by experimenting on the Blog datasets. The experimental results suggest that the platform is effective and feasible to apply for cybercrime forensic.

REFENRENCES

[1] F. Mosteller and D.L.Wallace, *Inference and Disputed Authorship:The Federalist*, In:behavioral science: quantitative methods edition, Massachusetts:Addison-Wesley, 1964.

[2] W. Elliot and R. Valenza, "Was the Earl of Oxford the true Shakespeare?," *Notes and Queries*,vol.38,pp. 501-506, 1991.

[3] G. Kjetsaa, "And Quiet Flows the Don Through the Computer," *Association for Literary and Linguistic Computing Bulletin*,vol.7, pp.248-256, 1979.

[4] T.C.Mendenhall, "The Characteristic Curves of Composition, "*Science*, vol.IX, pp.237-249, 1887.

[5] A. Q. Morton, "The Authorship of Greek Prose," *Journal of the Royal Statistical Society (A)*, vol.128, pp.169-233, 1968.

[6] A. Q. Morton, *Literary Detection*, Scribners New York, 1978.

[7] M. W. A. Smith, "An Investigation of Morton's Method to Distinguish Elizabethan Playwrights," *Computers and the Humanities*,vol.19, pp.3-21, 1985.

[8] J. F. Burrows, "Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing*,vol.2, no.4, pp.61-70, 1987.

[9] D. I. Holmes and R. S. Forsyth, "The 'Federalist' Revisited: New Directions in Authorship Attribution," *Literary and Linguistic Computing*,vol.10,pp.111-127,1995.

[10] G. K. Zipf, *Selected Studies of the Principle of Relative Frequency in Language*, Harvard University Press, 1932.

[11] H. S. Sichel, "On a Distribution Law for Word Frequencies," *Journal of the American Statistical Association*,vol.70, pp. 542-547,1975.

[12] R. Efron and B. Thisted, "Estimating the number of unseen species: How many words did Shakespeare know?, "*Biometrika*, vol.63, no.3, pp. 435–447,1976.

[13] M.W.A.Smith, "Recent experience and new developments of methods for the determination of authorship," *ALLC Bulletin*,vol.11, pp.73-82, 1983.

[14] J. M. Farringdon, A. Q. Morton and M. G. Farringdon, *Analysing for Authorship: A Guide to the Cusum Technique*, Cardiff, University of Wales Press,1996.

[15] J. F. Burrows," Computers and the study of literature," *In C. Butler, editor, Computers and Written Text, Applied Language Studies*, Blackwell, Oxford, pp.167–204, 1992.

[16] D. I. Holmes, M. Robertson and R. Paez, "Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution," *Computers and the Humanities*, vol.35, no.3, pp.315-331, 2001.

[17] T. Merriam and R. Matthews, "Neural compuation in stylometry II: An application to the works of Shakespeare and Marlowe," *Literary and Linguistic Computing*,vol.9, pp.1-6, 1994.

[18] B. Kjell, "Authorship attribution of text samples using neural networks and Bayesian classifiers," *In IEEE International Conference on Systems, Man and Cybernetics,* San Antonio, USA; 1994.

[19] J. F. Hoorn, S. L. Frank, W. Kowalczyk, and F. Van Der Ham, "Neural network identification of poets using letter sequences," *Literary and Linguistic Computing*, vol.14, no.3, pp.311-338,1999.

[20] D. I. Holmes, "The evolution of stylometry in humanities scholarship," *Literary and Linguistic Computing*,vol.13,no.3,pp.111-117, 1998.

[21] O. De. Vel, "Mining e-mail authorship," *Proceedings of workshop on text mining. In: ACM international conference on knowledge discovery and data mining (KDD)*, Boston, MA, USA, 2000.

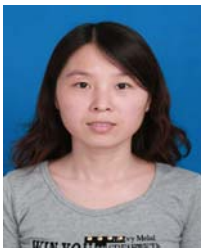
[22] O. De. Vel, A. Anderson, M. Corney and G. Mohay, "Mining e-mail content for author identification forensics," *SIGMOD Record*, vol.30, no.4, pp.55–64, 2001.

[23] O. De. Vel, A. Anderson, M. Corney and G. Mohay, "Multi-topic e-mail authorship attribution forensics," *Proceedings of ACM conference on computer security –*

Workshop on data mining for security applications; Philadelphia, PA, 2001.
[24] Y. Tsuboi, *Authorship Identification for Heterogeneous Documents*, Japanese: Nara Institute of Science and Technology, University of Information Science, 2002.
[25] R. Zheng, J. Li, H. Chen and Z. Huang, "A framework for authorshipidentification of online messages: writing-style features and classification techniques, "*Journal of the American Society for Information Science and Technology.* vol.57, no.3, pp. 378–393, 2006.
[26] R. Zheng, Y. Qin, Z. Huang and H. Chen, "Authorship analysis in cybercrime investigation," *Proceedings of the first international symposium on intelligence and security informatics (ISI),* Seattle Washington, USA , 2003.
[27] A. Abbasi and H. Chen, "Applying Authorship Analysis to Extremist- Group Web Forum Messages," *IEEE Intelligence System,* vol.20, no.5, pp.67-75, 2005.
[28] A. Abbasi and H. Chen, "Visualizing authorship for identification," *Proceeding of IEEE International Conference on Intelligence and Security Informatics*, San Diego, USA, 2006.
[29] A. Abbasi and H. Chen, "Writeprints: a stylometric approach to identitylevel identification and similarity detection in cyberspace," *ACM Transactions on Information Systems,*vol.26,no.2, pp.1-29, 2008.
[30] F. Iqbal, R. Hadjidj, B. C. M. Fung and M. Debbabi, "A novel approach of mining write-prints for authorship attribution in e-mail forensics," *Digital Investigation*, vol.5, supplement, pp.S42-S51, 2008.
[31] M. Asif and N. Tripathi, "Evaluation of OpenID-Based Double-Factor Authentication for Preventing Session Hijacking in Web Applications," *Journal of Computers*, vol.7,no.11, pp.2623-2628, November 2012.
[32] A. Ezzouhairi, A. Quintero, S. Pierre, "Adaptive Decision Making Strategy for Handoff triggering and Network Selection," *Journal of Computers*, vol.6,no.11,pp.2255-2266,November 2011.
[33] Z. Liu, Z. K. Yang and S. Y. Liu, "A Novel Random Subspace Method for Online Writeprint Identification," *Journal of Computers*, vol.7,no.12,pp.2997-3004, December 2012.
[34] X. Q. Yu, "Internal P-set and Security Transmission-identification of Information," *Journal of Computers*, vol.6,no.10, pp.2249-2254, October 2011.

Jianbin Ma received his MSc and PhD degree from agricultural university of Hebei in 2004 and 2010 respectively. Since 2004, he is a lecturer at college of information science and technology, agricultural university of Hebei. His research interests include Software Engineering, Computing, Information Retrieval and Management, Artificial Intelligence.

Ying Li received her MSc degree from agricultural university of Hebei in 2006. Since 2001 to 2008, she is a lecturer at college of science, agricultural university of Hebei. Now, she is teaching and researching in college of economics and trade. Her research interests include Artificial Intelligence, Intelligent Algorithm, Agricultural Economics and Asset Appraisal.



Guifa Teng received his phD degree from peking university in 2005. Since 2001, he is a professor at college of information science and technology, agricultural university of hebei. His research interests include Machine Learning, Software Engineering, Computer Applications.