2010 International Conference on Multimedia Information Networking and Security

Variable Length Character N-gram Approach for Online Writeprint Identification

Jianwen Sun, Zongkai Yang, Pei Wang, Sanya Liu*
National Engineering Research Center for E-learning
Huazhong Normal University
Wuhan 430079, China
e-mail: lsy5918@mail.ccnu.edu.cn

Abstract—**The Internet's numerous benefits have always been coupled with shortcomings due to the abuses of online anonymity. Writeprint identification is a technique to identify individuals based on textual identity cues people leave behind online messages. Character n-gram is one of the most effective approaches to identify writeprint according to previous research. In this study, we propose a variable length character n-gram based writeprint identification framework to address the identity tracing problem, integrating a genetic algorithm (GA) based feature selection component to solve the definition problem of n. To examine the approach, experiments are conducted on a test bed encompassing hundreds of reviews posted by 20 Amazon customers. The experimental results show the proposed approach is effective, obtaining a considerable improvement in identification accuracy and a heavy reduction of feature dimensionality.**

Keywords-writeprint identification; character n-gram; feature selection; genetic algorithm

I. INTRODUCTION

In spite of the numerous benefits of Internet, the abuses of online anonymity and its ubiquitous nature make it an ideal place to distribute inappropriate or illegal information which has resulted in many trust-related and criminal problems. For example, electronic commerce is susceptible to deception from easy identity change and reputation manipulation which have facilitated numerous forms of fraud. Criminals often use online messages to distribute illegal materials, including pirated software, child pornography, and stolen properties. Moreover, terrorist and extremist organizations such as Osama bin Laden and Al Qaeda are using online forum as one of their major communication channels for supporting psychological warfare, fundraising, recruitment, coordination, and distribution of propaganda materials [1].

It is necessary to counter anonymity abuses and strengthen the social accountability in cyberspace. As one of the major communication media involved in all the above-mentioned forms of Internet misuse, online messages have been used extensively on web-based channels such as email, forum, blog, internet newsgroup, and online review. Hence, we can utilize the writeprint hidden in people's writings as the potential traces of their identities [2]. Similar to human fingerprint, writeprint is composed of multiple features such as frequency of word, vocabulary richness, use of punctuation, length of sentence, structural information, etc. These features can represent an author's unique, immutable writing style and further become the basis of writeprint analysis [3].

From a machine learning point of view, writeprint identification can be seen as a single-label multi-class text categorization problem. One major subtask of the writeprint identification is to extract the most appropriate features for representing the writeprint. Previous studies proposed taxonomies of features under different labels and criteria [2] [4] [5]. Among all the measures, extracting frequencies of character n-grams is a more effective approach which is able to capture nuances of higher level and tolerate the noises such as grammatical errors or misuse of punctuations. Moreover, the procedure of extracting n-grams is language-independent and requires no special tools, especially for oriental languages such as Chinese where the tokenization procedure is not trivial [6].

The character n-gram approach has been proven to be quite useful to quantify the writing style [7]. Kjell [8] first used character bigrams and trigrams to discriminate the Federalist Papers. Keselj [9] and Stamatatos [10] reported very good results using character n-gram information. Moreover, one of the best performing algorithms in an authorship attribution competition organized in 2004 was also based on a character n-gram representation [11].

An important issue of the character n-gram approach is the definition of n. A large n would better capture lexical and contextual information but it would also capture thematic information and increase the dimensionality of the representation, while a small n would not be adequate for representing the contextual information. The drawbacks of defining a fixed value for n can be avoided by extracting variable length n-grams [12].

This paper is aimed at introducing a variable length character n-gram approach for writeprint identification of online messages. We first propose an overall research framework, and then develop a GA based feature selection method to identify the key writeprint features while solving the definition problem of n simultaneously. Finally, experiments are conducted to evaluate the effectiveness of the proposed approach.

The remainder of the paper is organized as follows. Section 2 presents the variable length character n-gram approach for online writeprint identification. Section 3 includes the performed experiments. The conclusions and future work are given in section 4.

978-0-7695-4258-4/10 $26.00 © 2010 IEEE
DOI 10.1109/MINES.2010.109

II. VARIABLE LENGTH CHARACTER N-GRAM APPROACH FOR WRITEPRINT IDENTIFICATION

Base on the previous review, the key of writeprint identification based on variable length character n-gram approach is to extract a set of representative and discriminative character n-grams with variable length instead of fixed length. Since we can theoretically enumerate all the possible character n-grams with length from 1 to n-1 in the whole space for a given dataset, the essence of this question is a feature selection problem. Therefore, we propose a variable length character n-gram based framework for writeprint identification as shown in Fig.1.

A. Data Collection

The first step of the framework is to collect a set of online messages written by a certain amount of authors for further writeprint profiling.

B. Initial Dominant Character N-grams Extraction

The purpose of this step is to pre-select an initial set consisting of dominant character n-grams with variable length which are extracted from the training dataset. Intuitively, all the possible length (from 1 to n-1) character n-grams should be included in the search space as the candidate initial set. But the set is too large that a pre-process must be done to eliminate the totally insignificant elements.

According to previous research, frequency is an important criterion to select the valuable writeprint features. Therefore, we use frequency as the measure to rank all the n-grams and then filter the trivial ones. Given certain value of M and N, the elements with length from 1 to N which are not part of the M most frequent n-grams would be excluded in the initial set. For example, if M=1,000 and N=4, only the 1,000 most frequent 1-gram, the 1,000 most frequent 2-grams, the 1,000 most frequent 3-grams, and the 1,000 most frequent 4-grams need to be extracted to compose the initial dominant character n-grams set.

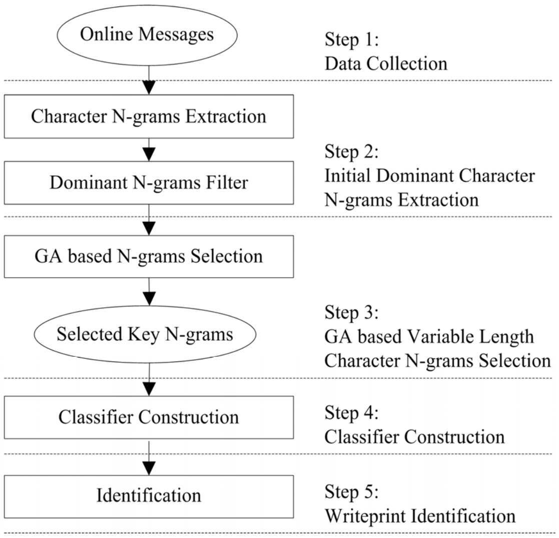

Figure 1. The proposed framework for writeprint identification

C. GA based Variable Length Character N-grams Selection

As aforementioned, the process of variable length character n-grams extraction is essentially a feature selection problem which can be viewed as a search problem here. GA is a heuristic search algorithm that can deal with large search spaces efficiently and has less chance to get local optima. Hence, we employ a GA based feature selection methodology to identify the most representative character n-grams with variable length.

1) Chromosome Decoding and Initial Population

The chromosome in this study is decoded using a binary string with the length equal to the number of a candidate feature subset. Each gene in a chromosome may have values “0” or “1” which indicates whether a feature is selected or not. For instance, a chromosome represented as “10000001” means the first and the last one are selected while others are discarded.

Since the variable length character n-grams selection is a huge-scale feature selection problem, some heuristics need to be provided in order to accelerate the convergence speed. Here, the frequency of character n-grams is incorporated into the GA’s initial population as the heuristic. First, all the genes are ranked according to the frequency of the corresponding character n-grams. Then the chromosomes are randomly generated by using only the top 500 genes to form the initial population. It means that all the genes of the chromosomes in the initial population always have value “0” except the first 500 ones.

2) Fitness Evaluation

The fitness function for evaluation of a chromosome combines two criteria: the number of selected features and the accuracy of the classification. The 2-criteria fitness formula is given below.

$$Fitness(x) = \alpha \times \exp(-\frac{|x|}{n}) + (1-\alpha) \times \exp(\frac{A(x) - \beta A_0}{(1-\beta) A_0}) \qquad (1)$$

Here, $Fitness(x)$ is the fitness of the feature subset denoted by x. $|x|$ indicates the size of x and n is the size of all the candidate features. $A(x)$ represents the classification accuracy by using the subset x, and A_0 represents the accuracy by using the full set. α and β are two constants specified according to the problem. α is used to control the contributes to the fitness function from two parts: the number of selected features and the accuracy of the classification by using that subset. β is used to set the threshold of the classification accuracy by using x (i.e. βA_0). When $A(x) < \beta A_0$ the exponential function penalizes heavily, which means if the classification accuracy by using x is not compatible with β of the accuracy by using the whole set, the value of the second part will drop sharply and result in x not selected. Hence, this fitness function requires the answer subset x to satisfy $A(x) > \beta A_0$ first and then to be as small as possible.

3) Genetic Operators

The genetic operators include selection, crossover, and mutation. In this research, we adopt the traditional operators including the roulette-wheel selection scheme, single-point crossover, and simple random mutation.

By applying genetic operators in the successive generations iteratively, the output of GA will achieve the steady fitness value. The feature subset with the highest classification accuracy among all the generations is regarded as the optimum.

D. *Classifier Construction*

Based on the optimal feature subset, we employ a classifier to build a model for online writeprint identification. As in a typical classifier learning process, iterative training and testing procedures are needed to develop a performing model.

E. *Writeprint recognition*

If the performance of the classifier is verified by the testing set, it can be used to identify the writeprint of new online messages.

III. EXPERIMENTAL EVALUATION

A. *Test bed*

The top 20 customer's online reviews from Amazon's Top Customer Reviewers were collected as our test bed, of which the main characteristics are listed in TABLE I.

B. *Experimental design*

Two experimental tasks were conducted to examine the effectiveness of the proposed approach. The first task was aimed at examining the performance of variable length character n-gram approach in comparison with fixed length approach. The second task was concerning the effectiveness evaluation of the GA based feature selection technology.

The classifier we employed in this study was realized by Support Vector Machine (SVM) which outperformed other classifiers in most cases for this context according to previous studies [2] [5]. We used the LIBLINEAR algorithm [13] included in WEKA [14]. In all experiments, linear kernels were used with C=1.

The first experiment was based on Information Gain (IG) measure to examine the performance of variable length character n-gram approach. Given an initial set of 18,167 character n-grams (including all the 91 1-gram, 3,076 2-grams, the 5,000 most frequent 3-grams, the 5,000 most frequent 4-grams, and the 5,000 most frequent 5-grams), IG was used to select the most significant 1,000 to 10,000 n-grams with a step of 1,000. The same approach was followed by fixed length n-grams. For example, using an initial set of 15,000 most frequent 3-grams, we also used IG to select the best 1,000 to 10,000 3-grams with a step of 1,000.

In the second experiment, we compared the proposed GA based feature selection method with no feature selection (as Baseline) and other two typical feature selection methods: IG and RelieF (RF). We used the standard 10-fold cross-validation to validate these techniques. For each fold of 10, feature selection was performed on the 540 training samples, while the remaining 60 were used to evaluate the classification accuracy for that particular fold. For GA, it was run using SVM with 5-fold cross-validation on the training samples for each fold. For IG and RF, we applied them to training samples for feature ranking and selection in each fold. The features selected by all techniques were then used on testing samples to evaluate the accuracy for that fold. The overall accuracy and number of selected features were computed as the average value across all 10 folds.

TABLE I. THE CHARACTERISTICS OF THE TEST BED

No. of authors	Average no. of messages per author	Average length of message	Time span
20	30	1,383 characters	1 year

GA was run for 1000 iterations, with a population size of 50 for each generation, using a crossover probability of 0.7 and a mutation probability of 0.03. The α and β in fitness function were set to 0.2 and 0.9 respectively. For RF, the number of nearest neighbors was used with k=10.

C. *Result discussion*

The results of the first experiment are shown in Fig.2. As can be seen, the variable length n-gram approach outperformed fixed length approaches when the selected features were less than 6,000. But for a higher dimensionality, the improvement on performance began to decline. It could be observed that the variable length n-gram approach even failed to compete with fixed length approach, especially for 3-grams, when the number of features was over 6,000. These results were generally consistent with previous study [12].

Fig.3 indicates the change of accuracy and the number of selected features across 1000 generations in the second experiment. The evolutionary process of GA converged after about 350 generations. Among all the chromosomes in the 1000 generations, the one with highest accuracy was regarded as the optimum.

Fig.4 and Fig.5 shows the performance of no feature selection and other two feature selection techniques using the features at a lower and higher dimensionality respectively. As can be seen, both of the two methods substantially enhanced the classification accuracy over the baseline.

Figure 2. The results of the first experiment

Figure 3. The process of GA based character n-grams selection

TABLE II shows the best performance of the baseline and all the three feature selection methods adopted in this study. All of the three techniques significantly outperformed the baseline. GA obtained a considerable performance improvement, resulting in a 3% improvement over the no feature selection baseline and a 1% improvement over IG. Although the mean accuracy of GA was inferior to RF, the most important was that GA heavily reduced the feature dimensionality, identifying a more concise set of key features that was about 90% smaller than IG and RF, and over 96% smaller than the baseline.

Figure 4. Performance of Base, IG and RF by using features from 100 to 1,000

Figure 5. Performance of Base, IG and RF by using features from 1,000 to 10,000

The performance of all the methods by using the same number of features as GA can be seen in TABLE III. For IG and RF, the top 645 features were selected according to their ranking information. For baseline, the 645 most frequent features were used to compare with GA. As shown in TABLE III, GA obtained a great performance improvement, resulting in a 15% improvement over the no feature selection baseline and a 7%-9% improvement over RF and IG.

TABLE IV illustrates the top 50 most discriminative character n-grams selected by the three feature selection methods. The character "_" in the table denoted as a single space. It can be seen that too much redundant information was contained in the set, especially for IG and RF. For example, the n-grams "_boo", "_book", "book", and "boo" were all selected by IG in the top 50 n-grams set. This indicated IG and RF could not efficiently detect multiple redundant features. On the other hand, GA was more efficient to eliminate redundant features. Hence, compared with IG and RF, the feature set selected by GA will be richer in different character n-grams corresponding to different kind of stylometric information when the same amount of character n-grams were selected.

TABLE II. COMPARISON OF THE BEST PERFORMANCE

Methods	**Performance**		
	Mean Accuracy	***Std.Dev***	***No. of Features***
Base	90.17%	0.024	18,167
IG	92.96%	0.029	5,000
RF	94.50%	0.015	9,000
GA	93.67%	0.016	645

TABLE III. COMPARISON OF THE PERFORMANCE BY USING THE SAME NUMBER OF FEATURES AS GA

Methods	**Performance**		
	Mean Accuracy	***Std.Dev***	***No. of Features***
Base	78.17%	0.051	645
IG	84.33%	0.031	645
RF	86.17%	0.036	645
GA	93.67%	0.016	645

TABLE IV. ILLUSTRATION OF THE KEY CHARACTER N-GRAMS

Methods	The top 50 key character n-grams
IG	"_-_", "-_", "_-", "__", "s_boo", "mende", "-", """, "_boo", "..", "t's_", "t's", "_book", ".__", ",_and", "t'", "...", "is_bo", "'_'", |book", "boo", " __-", "ended", """", "_(", "(", ""_", "I_", ",_an", "__-_", "_I_", "lanke", "_and_", "book_", "_"", "anke", "_and", "it's_", "it's", "it'", "_Oz", "Oz", "arks_", "on_B", ")", "Oza", ",_a", "The_O", "_Blan", "n_Bla"
RF	"Oza", "n_Bla", "p__", "__Don", "mende", "ended", "nded.", "n", "Highl", "__Hig", "__Hi", "_I_", "I_", "d!_", "-", "__", "i", "_", "a", "on_B", "zar", "e", "5/5", "_5/", ".__", "e_", "r", "ecomm", "mmend", "mend", "_I", "o", "s", "'s", "_a", "'s_", "_Blan", "The_0", "_Don_", "0z", "_0z", "_and_", "_and", "t", ",", "nship", "nshi", "nsh", ",_", "__Do"
GA	"i", "n", "r", "e_", "d", "_t", "c", "u", "s_", "m", "g", "_th", "in", "d_", "w", "er", "b", "an", "_i", ".", "_s", "_o", "y_", "he_", "._", "is", ",", "nd", "v", "_b", "the_", "at", "it", "'_the_", "es", "'en", "k", "ha", "ng", "_f", "st", "_an", "and", "ve", "_h", "_p", "le", "ti", "ed", "se"

IV. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a variable length character n-gram based framework for online writeprint identification and integrated a GA based feature selection component to solve the definition problem of n. The experimental results indicated the effectiveness of the proposed framework and the potential of GA based key character n-grams selection. The GA based method not only obtained a considerable improvement in identification accuracy, but also got a heavy dimensionality reduction. Moreover, GA was more efficient to eliminate the redundant character n-grams compared with IG or RF methods.

In the future we would continue to improve the GA based feature selection technology for dealing with large solution spaces. We believe this technology could improve the performance and scalability of online writeprint identification. We also intend to explore the hybrid GA by incorporating more efficient heuristics to facilitate convergence and improve accuracy.

ACKNOWLEDGMENT

This work was supported by National High-tech R&D Program of China (863 Program) (No.2008AA01Z131) and self-determined research funds of CCNU from the colleges'basic research and operation of MOE (No.CCNU09A02006). We are also grateful for the research assistance provided by fellow members including Zhi Liu, Yuying Wei and others who wish to remain anonymous.

REFERENCES

[1] Jialun Qin, Yilu Zhou, Edna Reid, Guanpi Lai, and Hsinchun Chen, "Analyzing terror campaigns on the internet: Technical sophistication, content richness, and Web interactivity," International Journal of Human-Computer Studies, v.65, n.1, p.71-84, January, 2007.

[2] Zheng, R., Li, J., Chen, H., and Huang, Z., "A framework for authorship identification of online messages: Writing style features and classification techniques," Journal of the American Society of Information Science and Technology, 57(3), 378-393, 2006.

[3] Li, J., Zheng, R., and Chen, H., "From fingerprint to writeprint," Communications of the ACM, 49(4), 76–82, 2006.

[4] Stamatatos, E., Fakotakis, N., and Kokkinakis, G., "Automatic text categorization in terms of genre and author," Computational Linguistics, 26(4), 471–495, 2000.

[5] Abbasi, A., and Chen, H., "Applying authorship analysis to extremist-group web forum messages," IEEE Intelligent Systems, 20(5), pp. 67-75, 2005.

[6] Stamatatos, E., "A survey of modern authorship attribution methods," Journal of the American Society of Information Science and Technology, 60(3), 538–556, 2009.

[7] Grieve, J., "Quantitative authorship attribution: An evaluation of techniques," Literary and Linguistic Computing, 22(3), 251-270, 2007.

[8] Kjell, B., "Discrimination of authorship using visualization," Information Processing and Management, 30(1), 141-150, 1994.

[9] Keselj, V., Peng, F., Cercone, N., and Thomas, C., "N-gram-based author profiles for authorship attribution," In Proceedings of the Pacific Association for Computational Linguistics, pp. 255-264, 2003.

[10] Stamatatos, E., "Ensemble-based author identification using character n-grams," In Proceedings of the 3rd International Workshop on Text-based Information Retrieval, pp. 41-46, 2006.

[11] Juola, P., "Ad-hoc authorship attribution competition," In Proceedings of the Joint Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, pp. 175-176, 2004.

[12] Houvardas, J., and Stamatatos E., "N-gram feature selection for authorship identification," In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications, pp. 77-86, Springer, 2006.

[13] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin, "LIBLINEAR: A library for large linear classication," The Journal of Machine Learning Research, 9:1871–1874, 2008.

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, "The WEKA data mining software: An update," SIGKDD Explorations, 11(1), 2009.