Journal of Mathematics Research; Vol. 10, No. 4; August 2018
ISSN 1916-9795 E-ISSN 1916-9809
Published by Canadian Center of Science and Education

Handwriting Detection Model Based on Four-Dimensional Vector Space Model

Lin Li[1], Xiuteng Duan[2] & Yutong Li[1]

[1] School of Translation Studies, Jinan University, China

[2] School of Electric & Information, Jinan University, China

Correspondence: Lin Li, School of Translation Studies, Jinan University, Zhuhai, China.

Received: April 4, 2018 Accepted: April 17, 2018 Online Published: May 28, 2018

doi:10.5539/jmr.v10n4p32 URL: https://doi.org/10.5539/jmr.v10n4p32

Abstract

Handwriting detection is mainly used in the criminal investigation. We can use four-dimensional vector space model to build a model for handwriting detection. This article selects feature quantities such as word frequency, language style, average word length, and sentence structure from the texts and quantizes them, transforming them into relations between vectors. After quantifying and normalizing the features in an author's article in advance, we can obtain a standard reference vector. Then we do the same processing on the target text database, and compare it with the standard reference vector in terms of the modulus value and the included angle. Then we could estimate whether the author is the owner of database value. The simulation result shows that the model is more accurate and the author of particular texts can be obtained.

Keywords: Vector Space Model, handwriting detection, normalized processing

1. Introduction

With the popularity of e-mails, mobile phone messages, and social networking messages, more and more anonymous texts information have become involved materials. Identifying the origin of the texts is crucial in the detection of criminal cases. From view of functional linguistic, language reflects the inner world of human beings at all times. The world we know is realized through language. Language identification has also become a highly regarded detective method. This gave birth to handwriting detection, that is, matching articles and their authors through specific features in the article.

At present, there are few researches on handwriting detection models. Most of them focus on single-field research and are usually difficult to apply universally. Vector Space Model(VSM), proposed by Gerard Salton and McGill in 1988, using vectors to represent texts, and take weights of the features of texts as components. Through the calculation of word frequency and dimension reduction of the vectors, this model can compute the similarity of texts. Li Xuelei&Zhang Dongmo established a model which correlates the text characters, the term frequency, the hypertext markup language tag information in the web, and semantic analysis for the question sentences to calculate an adjustable term frequency weighting parameter and to increase the separability of feature words vector. (Xuelei Li & Dongmo Zhang, 2003) Turney also used a vector-based model of semantic relations to attain a score of 56% on some multiple-choice questions which are from SAT test. (Turney, Peter D., 2006) This paper proposes a handwriting detection model based on four-dimensional vector space model, which can be applied to many types of text to solve the problem of author identification in criminal investigation.

2. Handwriting Detection Model Based on Four-dimensional Vector Space Model

The vector space model is an algebraic model that applies to information filtering, information selection, indexing, and assessment of relevance. It simplifies the processing of text content into operations in vector space, and expresses the similarity of language features with spatial similarity. We measure the similarity of texts by calculating the similarity of the vectors. In this article, we also select the feature quantities in the texts and quantify them.

2.1 Selecting Feature

(1) Selecting the word with highest frequency in each text as one of the recognition criteria. Because the word with highest frequency of each message also reflects a person's writing habits, in addition to representing the subject of this article. Even the word frequency of one person's several articles settles in a range of frequency fluctuations. (Huiling Wang & et al., 2001)

(2) Language style is also an indispensable criteria of great importance. It is undeniable that each person's writing style is very different. What we choose here is the expression of exclamatory sentences. We calculate the frequency of sentences

that express the strong emotion of persons as a criterion.

(3) The average word length is also an important evidence on whether an article is written by a particular people. Some people prefer to use short, informal words, while others prefer the long ones. At the same time, we also tend to use long and formal words while writing some formal documents. Therefore, the average word length is a very good measure to judge the author's identity, educational level, etc. (Qiang Li & Jianhua Li, 2006)

(4) Sentence structure is also selected as one of the criteria for judgment. Here we use the frequency of link verbs as the research object. This is because the sophistication and difficulty of writing sentences in the articles not only represents a person's educational level, but also deeply reflects one's writing habits. Even everyone has his habitual structure of sentences, and there is a proportion of each structure type.

2.2 Establishing Feature Database

2.2.1 Extracting the Highest Frequency and Establishing a Database

In the processing of word frequency feature, only the value of the highest word frequency is considered. In this way, we can obtain the writer's writing habits (some people have wordy writing style, while others cherish words like gold). Therefore, the highest word frequency can be used as indicator of observation and analysis.

(1) The initial data is simply a bunch of text database, we wrote the program through JAVA, and analyze the highest word frequency in each article, establishing a new database.

(2) We compute the arithmetic average of each highest word frequency in the above new database, and finally obtain the arithmetic average as a component of the standard parameter vector.

The formula is as follows:

$$f_{av} = \frac{\sum_{i=1}^{m} f_i}{m} \tag{1}$$

(3) We extract the f_{max} and f_{min} in many articles,then subtract from f_{av}, comparing two absolute values. The result is as follows:

If $| f_{max} - f_{av} | > | f_{min} - f_{av} |$

As a result, f_{max} deviates from the highest average frequency;

If not, f_{min} deviates from the highest average frequency.

The results above can be used as the calculation of the dynamic range in the model.

2.2.2 Analyzing the Style of Article Sentences and Building a Quantitative Database

Language styles are various. Usually the number of exclamations and the frequency of modal particles can represent the writing style of a person best. When quantifying, we deal with them mainly by looking for the number and frequency of exclamatory sentences.

(1) The same as the quantification of the highest word frequency database above, the initial data is also just a simple text library. We analyzed the frequency of exclamations and modal particles in each text through JAVA programming to form a new database.

(2) We compute the arithmetic average of each frequency figure in the new database above, and finally obtain the arithmetic average as a component of the standard parameter vector. Assuming the total number of texts is m, then the formula is as follows:

$$y_{av} = \frac{\sum_{i=1}^{m} y_i}{m} \tag{2}$$

(3) We extract y_{max} and y_{min} from the database, and then subtract from y_{av}, comparing two absolute values. The result is as follows:

If $| y_{max} - y_{av} | > | y_{min} - y_{av} |$

As a result, y_{max} deviates from the highest average frequency;

If not, y_{min} deviates from the highest average frequency.

The results above can be used as the calculation of the dynamic range in the model.

2.2.3 The Statistics of Average Word Length and Its Quantification

The average word length often reflects a person's educational level (the formal words are usually longer). Therefore, the average word length can also be a indicator to identify a person's identity.

(1) We use JAVA to write the program based on the initial corpus, analyzing the average word length in each text to form a new database.

(2) We compute the arithmetic average of the average length of each word in the new database above, and finally obtain the arithmetic average as a component of the standard parameter vector.

The formula is as follows:

$$L_{av} = \frac{\sum_{i=1}^{m} L_i}{m} \tag{3}$$

(3) We extract the L_{max} and L_{min} in many emails , then subtract from L_{av}, comparing two absolute values. The result is as follows:

If $| L_{max} - L_{av} | > | L_{min} - L_{av} |$

As a result, L_{max} deviates from the highest average frequency;

If not, L_{min} deviates from the highest average frequency.

The results above can be used as the calculation of the dynamic range in the model.

2.2.4 The Analysis and Quantification of Sentence Structure

Sentence structure is also a tool to reflect a person's writing habits, many people like to use the subject-linking verb-predicative structure or subject-verb-object structure. Here we choose the the subject-linking verb-predicative structure to quantify its proportion.

(1) The original data is still the text library, we use the program written by JAVA programming language to obtain linking verbs in each article, so as to get the proportion of the subject-linking verb-predicative system structure statements, constituting a new database.

(2) We compute the arithmetic average of the proportion in the new database above, and finally obtain the arithmetic average as a component of the standard parameter vector. Assuming the total number of texts is m, then the formula is as follows:

$$S_{av} = \frac{\sum_{i=1}^{m} S_i}{m} \tag{4}$$

(3) We extract $S_{max}$and $S_{min}$from the database, and then subtract from S_{av}, comparing two absolute values. The result is as follows:

If $| S_{max} - S_{av} | > | S_{min} - S_{av} |$

As a result, S_{max} deviates from the highest average frequency;

If not, S_{min} deviates from the highest average frequency.

The results above can be used as the calculation of the dynamic range in the model.

2.3 Weight Calculation of Feature Quantities

Since the four characteristic quantities have different effects on identifying the author of the text, we need to quantify the proportion of the effect each characteristic quantity. The weight calculation formula is as follows:

$$W_{ik} = \frac{tf_{ik}[\log(\frac{m}{n_{k+0.01}})]}{\sqrt{\sum_{i=1}^{m} (tf_{ik})^2 [\log(\frac{m}{n_{k+0.01}})]^2}} \tag{5}$$

tf_{ik} is the frequency at which the feature appears in the text;

And the denominator is the normalization factor.

2.4 Four-dimensional Space Vector Formula

Reference Vector:

$$\vec{I} = (W_{fk} \cdot \overrightarrow{f_{avi}}, W_{yk} \cdot \overrightarrow{y_{avi}}, W_{Lk} \cdot \overrightarrow{L_{avi}}, W_{Sk} \cdot \overrightarrow{S_{avi}}) \tag{6}$$

tolerance scope:

$$\Delta\vec{I} = (W_{fk} \cdot \Delta\overrightarrow{f_{avi}}, W_{yk} \cdot \Delta\overrightarrow{y_{avi}}, W_{Lk} \cdot \Delta\overrightarrow{L_{avi}}, W_{Sk} \cdot \Delta\overrightarrow{S_{avi}}) \tag{7}$$

The result of vector quantification is:

$$\vec{I_s} = (W_{fk} \cdot \overrightarrow{f_{avi}}, W_{yk} \cdot \overrightarrow{y_{avi}}, W_{Lk} \cdot \overrightarrow{L_{avi}}, W_{Sk} \cdot \overrightarrow{S_{avi}}) \tag{8}$$

After that, we proceed as follow. Firstly, we calculate the difference between modules:

If $0 <| \mid \vec{I} \mid - \mid \vec{I_s} \mid |<| \Delta\vec{I} |$, the modulo value satisfies the requirement.

Secondly, we calculate the values of included angle:

$$0 < a \leq \frac{\vec{I} \cdot \vec{I_s}}{\mid \vec{I} \mid \cdot \mid \vec{I_s} \mid} < \frac{\vec{I} \cdot (\overrightarrow{f_{av} + \Delta f_i}, \overrightarrow{y_{av} + \Delta y_i}, \overrightarrow{L_{av} + \Delta L_i}, \overrightarrow{S_{av} + \Delta S_i})}{\mid \vec{I} \mid \cdot \mid \overrightarrow{f_{av} + \Delta f_i}, \overrightarrow{y_{av} + \Delta y_i}, \overrightarrow{L_{av} + \Delta L_i}, \overrightarrow{S_{av} + \Delta S_i} \mid} \tag{9}$$

$$\vec{P} = (\overrightarrow{f_{av} + \Delta f_i}, \overrightarrow{y_{av} + \Delta y_i}, \overrightarrow{L_{av} + \Delta L_i}, \overrightarrow{S_{av} + \Delta S_i}) \tag{10}$$

3. Simulation Results

We use texts with more than 1000 words for simulation to ensure accuracy of the model. Therefore, we selected three sets of sample, each with a sample size of 15. (Yiping Zeng & Xiaowen Zhu, 2006)

3.1 Data Results

The digital data obtained after processing is shown in the following table (sheets 4-1 to 4-3):

Author 1:

sample	Average word length	The highest frequency	Exclamation sentence proportion	The proportion of subject-linking verb-predicative structure
1 text 3111	5.98	18/385=0.0468	0/46=0	20/46=0.43
2 text 7664	6.11	14/393=0.0356	0/89=0	14/89=0.16
3 text 7932	4.61	52/1433=0.0363	0/138=0	64/138=0.46
4 text 7972	4.29	199/4863=0.0410	0/370=0	237/370=0.64
5 text 7992	3.67	22/1515=0.0145	0/128=0	36/128=0.28
6 text 8018	3.74	63/2328=0.0271	1/192=0.0052	70/192=0.36
7 text 8627	4.03	67/2417=0.0265	2/175=0.0114	96/175=0.55
8 text 9083	5.88	31/633=0.7	0/107=0	36/107=0.34
9 text 9167	6.00	10/241=0.0415	0/41=0	13/41=0.32
10 text 9175	5.08	19/690=0.0275	0/89=0	22/89=0.25
11 text 9234	5.12	40/1133=0.0353	0/152=0	48/152=0.32
12 text 9237	5.62	15/429=0.0350	0/44=0	14/44=0.32
13 text 10425	6.36	5/227=0.0220	1/39=0.0256	11/39=0.28
14 text 12447	6.11	8/196=0.0408	0/30=0	8/30=0.27
15 text 19961	5.86	10/329=0.0304	0/53=0	19/53=0.36
average	5.23	0.0341	0.0028	0.36

sheet 4-1

Author 2:

sample	Average word length	The highest frequency	Exclamation sentence proportion	The proportion of subject-linking verb-predicative structure
1 text 1825	4.31	13/458=0.0284	1/65=0.0154	26/65=0.40
2 text 7725	5.59	4/262=0.0153	1/48=0.0208	14/48=0.29
3 text 9176	4.91	29/710=0.0408	0/88=0	41/88=0.47
4 text 54544	5.41	5/133=0.0376	0/32=0	8/32=0.25
5 text 54545	6.26	5/232=0.0216	0/51=0	26/51=0.51
6 text 54604	5.35	8/279=0.0287	0/33=0	9/33=0.27
7 text 54633	4.93	12/408=0.0294	0/62=0	23/62=0.37
8 text 54634	5.00	12/395=0.0304	0/53=0	17/53=0.32
9 text 54659	4.70	13/329=0.0395	3/49=0.0612	20/49=0.41
10 text 173410	5.95	3/913=0.0033	0/24=0	5/24=0.21
11 text 173470	5.90	25/458=0.0546	2/136=0.0147	12/136=0.09
12 text 173949	5.09	18/387=0.0465	0/35=0	19/35=0.54
13 text 193986	6.17	3/96=0.0313	0/16=0	2/16=0.13
14 text 173997	4.41	41/1364=0.0301	3/160=0.0188	80/160=0.50
15 text 175318	5.71	6/145=0.0414	0/19=0	3/19=0.16
average	5.31	0.0319	0.0087	0.328

sheet 4-2

Author 3:

sample	Average word length	The highest frequency	Exclamation sentence proportion	The proportion of subject-linking verb-predicative structure
1 text 9085	5.48	7/276=0.0254	0/39=0	15/39=0.38
2 text 9159	6.60	7/286=0.0245	0/34=0	19/34=0.56
3 text 9191	6.81	4/141=0.0284	0/24=0	4/24=0.17
4 text 12030	5.47	12/264=0.0455	0/23=0	7/23=0.30
5 text 12174	6.94	6/179=0.0335	1/29=0.0345	5/29=0.17
6 text 12176	4.45	56/1671=0.0335	0/185=0	91/185=0.49
7 text 50307	5.37	11/246=0.0447	1/39=0.0256	15/39=0.38
8 text 52201	6.55	5/152=0.0329	0/23=0	8/23=0.35
9 text 53536	4.08	11/727=0.0151	0/77=0	36/77=0.47
10 text 54263	5.56	10/381=0.0262	0/53=0	39/53=0.74
11 text 54536	5.79	28/639=0.0438	0/77=0	72/77=0.94
12 text 54537	7.27	4/116=0.0345	0/25=0	10/25=0.4
13 text 54540	4.60	4/207=0.0193	0/46=0	11/46=0.24
14 text 54577	5.88	7/380=0.0184	0/66=0	26/66=0.39
15 text 54582	5.73	6/210=0.0286	0/32=0	13/32=0.41
average	5.77	0.0303	0.0040	0.426
average	5.31	0.0319	0.0087	0.328

sheet 4-3

3.2 The Average Value of the Characteristic Quantities and the Maximum Offset

The first set: $f_{av1} = 0.0341; L_{av1} = 5.23; y_{av1} = 0.0028; S_{av1} = 0.36$
$\Delta f_1 = 0.0196; \Delta L_1 = 1.56; \Delta y_1 = 0.0028; \Delta S_1 = 0.20$

$$\vec{I} = (W_{fk} \cdot \overrightarrow{0.0341}, W_{Lk} \cdot \overrightarrow{5.23}, W_{yk} \cdot \overrightarrow{0.0028}, W_{Sk} \cdot \overrightarrow{0.36}) \tag{11}$$

The second set: $f_{av2} = 0.0319; L_{av2} = 5.31; y_{av2} = 0.0087; S_{av2} = 0.328$
$\Delta f_2 = 0.0166; \Delta L_2 = 1.00; \Delta y_2 = 0.0087; \Delta S_2 = 0.198$

$$\vec{I} = (W_{fk} \cdot \overrightarrow{0.0319}, W_{Lk} \cdot \overrightarrow{5.31}, W_{yk} \cdot \overrightarrow{0.0087}, W_{Sk} \cdot \overrightarrow{0.328}) \tag{12}$$

Third set: $f_{av3} = 0.0303; L_{av3} = 5.77; y_{av3} = 0.0040; S_{av3} = 0.426$
$\Delta f_3 = 0.0144; \Delta L_3 = 1.69; \Delta y_3 = 0.0040; \Delta S_3 = 0.414$

$$\vec{I} = (W_{fk} \cdot \overrightarrow{0.0303}, W_{Lk} \cdot \overrightarrow{5.77}, W_{yk} \cdot \overrightarrow{0.0040}, W_{Sk} \cdot \overrightarrow{0.426}) \tag{13}$$

3.3 Model Testing

At first, we need to compute the weight;

The result of normalization is as follow

Word frequency weight = 0.323

Sentence structure weight = 0.323

Average word length weight = 0.323

Language style weight = 0.032

There are three databases above, we use the first database to validate the model:

We extract a text randomly from the author's article database to calculate the vector:

$$\vec{I}_s = (\overrightarrow{0.0143}, \overrightarrow{2.038}, \overrightarrow{0}, \overrightarrow{0.0023}) \tag{14}$$

The first step, validate the modulus:

$$\vec{I} = (\overrightarrow{0.011}, \overrightarrow{1.67}, \overrightarrow{0.00009}, \overrightarrow{0.116}) \tag{15}$$

Then:

$$| \vec{I} | = 1.674 \tag{16}$$

$$\Delta\vec{I} = (\overrightarrow{0.0064}, \overrightarrow{0.5039}, \overrightarrow{0.00009}, \overrightarrow{0.0646}) \tag{17}$$

$$| \Delta\vec{I} | = 0.5081 \tag{18}$$

$$\Delta\vec{I}_s = (\overrightarrow{0.0143}, \overrightarrow{2.038}, \overrightarrow{0}, \overrightarrow{0.0023}) \tag{19}$$

$$| \vec{I}_s | = 2.038 \tag{20}$$

Therefore:$0 < 2.038 - 1.674 = 0.364 < 0.5081$

So the modulus fit the model

The second step: the comparison of the angles

Maximum deviation:

$$\vec{P} = (\overrightarrow{0.0173}, \overrightarrow{2.193}, \overrightarrow{0.0001792}, \overrightarrow{0.1809}) \tag{21}$$

$$| \vec{P} | = 2.2005 \tag{22}$$

We calculate the included angle between $\vec{I}$ and $\vec{P}$.

The result is as follows: 0.99995 rad.

We calculate the included angle between $\vec{I}$ and $\vec{I}_s$.

The result is as follows: 0.99815 rad.

It satisfies $0 < 0.99815 < 0.99995$

Therefore, we get the result. This article is written by author 1.

The model validity is completed. this model can be used for handwriting analysis.

4. Conclusion

The vector space model turns the articles into vectors and the concept of our model is relatively simple. When the number of texts is high and the data is complete, the results of model are more accurate. Our model can be used to analyze the handwriting of different people, rather than just be used under a certain situation. By establishing a specific database, we can apply it in every situation, which shows that the model is flexible and reliable. We need to compute a lot while using this model, so there are still some disadvantages needed to be optimized, but the model is of great significance and it provides realistic guidance for criminal cases bases on e-mail.

References

Xuelei Li, & Dongmo Zhang (2003). A Text Categorization Method Based on VSM. *Computer Engineering, 29*(17), 90–92.

Turney Peter, D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416, MIT Press.

Huiling Wang, Rou Song, & Weichang Dai (2001). *The Research of Text Categorization on Style*. The 6th national conference on computational linguistics.

Qiang Li, & Jianhua Li (2006). Method of Filting Reactionary Text Based on Vector Space Model. *Computer Engineering, 32*(10), 4–5.

Yiping Zeng, & Xiaowen Zhu (2006). Application of Computational Methods to the Study of Stylistics in China. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition)*, *1*, 14–17.

Copyrights