Regular Paper

Linearization of Zipfian Distribution for Chinese Characters

KIM TENG LUA*

In this paper, we report our results of least-square fittings to 4 sets of data derived from Chinese characters, namely, character strokes, radicals, characters and words. We have found that fitting using a power series, ie f^t versus R^t (f is the frequency of occurrence, R the rank and t is a constant) is better than the use of a logarithm series derived from the original simple Zipf's law, ie fR = constant, or $\log f = c\text{-}\log R$. The dependency of f versus R is found to be of order 5 as we have found that $t=0.2$. We have also discovered a secondary dependency of f on R of lower order. This secondary dependency can be modeled using a cosine function.

1. Introduction

Now it has been established that the Zipfian [1] distribution plays an important role in the modeling of human activities, particularly of the variables studied in bibliometrics and scientometrics. These include the appearance of characters, words, productivity of researchers etc [2-5]. It has been established by Zipf in his pioneer work done in 1948 [1] that the frequency of occurrence (f) of a symbol (a word in his original study) is inversely proportional to its rank (R), ie

$$f=\frac{c}{R} \qquad (1)$$

However, Zipf's formulation which was derived from his restricted observations based on the occurrence of English words is at best only a very rough estimation. Deviations has been reported widely. Likewise, we have found significant deviations when we applied the formula to the occurrence of Chinese characters and words, the world's only ideographical language [5].

In this paper, our objective is to search for a general empirical formula that will accurately describe the behavior of Zipfian distributions for Chinese language. Such a formula will have useful applications in data base design, information storage and retrieval systems under the Chinese language environment.

Four sets of data from Chinese ideographic characters are used in the present study: (i) occurrence of character strokes (32), (ii) occurrence of character radicals (623) (iii) occurrence of characters (5584) and (iv) occurrence of words (46,520). The numbers in the parentheses indicate the number of unique symbols in the set. This is to say, we have 32 unique character strokes, 623 radical components, 5584 Chinese characters and 46,520 words in the current exercise.

It has to be noted that Chinese words are formed by characters; and characters are formed by radical components and finally, radicals are formed by character strokes. The selection of strokes, radicals, characters and words in our study is to provide a coherent test set for the least-square fittings from which the empirical formulae are derived. Some inner regularities of the Zipfian law might thus be discovered.

Four mathematical models are considered in the least-squares fittings, namely (i) extended Zipf's law in hyperbola form (f versus $1/R$), (ii) extended Zipf's law in logarithm form ($\log f$ versus $\log R$) (iii) polynomial series of f^t versus R^t (t is a constant) and (iv) Mandelbrot's model, ie $f=a/(R+C)^b$.

Our conclusion is that the third model provides the best result in terms of simplicity and accuracy. It has the least number of parameters. The value of t is found to be very close to 0.2. We thus conclude that the dependency of f on R is of order 5.

We have also discovered that there is a secondary dependency of f on R at even higher orders. We can obtain a numerical fitting to within the limit of experimental errors by fitting the observed data to a high order polynomial of order 10 to 20. But this results in empirical formulae with large number of parameters whose significance are hard to interpret.

An alternative is to

(1) fit the observed data to a polynomial of lower order (order 2-3);

(2) compute the residual errors by subtracting the calculated frequencies of occurrence with the observed frequencies of occurrence;

$$\begin{aligned}\Delta &= \textit{residual error}\\ &= \textit{calculated frequency} - \textit{observed frequency}\end{aligned}$$

(3) model the residual errors using a cosine function

[*] National University of Singapore, Kent Ridge, Singapore 0511.

of the form:

$$\Delta=\cos\left(\sum_{j=0}^{j=m} b_j x^j\right)\exp\sum_{i=0}^{i=2} a_i x^i \qquad (2)$$

where $x=\log R$ or $x=R^t$, depends on the formula chosen. t is a constant.

The current exercise has two major drawbacks. Firstly, we are unable to obtain a formula that fits the entire range of distribution. Often, the first few (or up to 10) pairs of data are abandoned as they do not seem to fall onto the same line. Second, the treatment of the residual errors using a cosine function has no theoretical support.

The only merit of this paper is the discovery of high order dependency between f and R. We are still far from being able to work out a simple enough formula for the Zipfian distribution. We are even more remoted from being able to provide a satisfactory theoretical explanation for the observed distribution.

2. Sources of Data

The occurrence of character strokes and radicals are obtained from *A Dictionary of Chinese Character Information* [6]. The data on the occurrence of character strokes are obtained from author's own work [7]. The occurrence of characters and words are derived from a *Electronic Word Frequency Dictionary* [8].

From [6], there are actually two sets of data for character strokes and radicals, ie a dynamic and a static distribution. For dynamic distribution, the frequency of usage of a symbol is considered and for the latter, the occurrence is counted by the number of times it appears in a list, with regardless of its frequency of usage. For example, *kuo* [□] occurs 2,532,281 (5.6368%) times in a text sample of 44.9123 million characters when the frequency of usage is considered. But it appears only 1321 (5.94777%) times when we consider only its occurrence in a character list of totally 23,2102 radical components (from 7785 unique characters).

Likewise, the distribution of the character strokes are derived from the analysis of the first 1000 most frequently used characters. A total of 8037 strokes were obtained by decomposing the characters [7]. The frequencies of occurrence of the characters in a text sample are not considered. Therefore each character is only counted once.

Our selection is based on the fact that the usage of a symbol and its appearance as a component of a symbol list (ie radicals in a character list) are two different entities. The first is a combined function of the second and the usage function. Thus our selection is concurrent with our desire to know how frequent a symbol occurs when it is used to produce another set of symbol. This assumption has important bearing in human psychology.

When we read a text, we focus our attention on the words that make up a sentence. We seldom pay attention to the individual characters forming a word or the character strokes forming a character.

Psychologically, we normally focus our attention on one level. We switch our attention to a lower layer only when we encounter some difficulty in an upper layer. For example, when we read, we only attempt to guess the meaning of a character from its constituent radicals and strokes if we cannot recognise the character. Otherwise, the character will be recognised as a complete symbol without our further attention to its detailed structure.

The frequency-rank curves for strokes, radicals, characters and words are shown in Fig. 1-4. Note that f are in fractions.

3. Conformation to Zipf's Law

Our item sizes vary from 32 to 46,520. In one of our earlier work [7], we have found that the conformation to the Zipf's law depends on a ratio r which can be written as:

$$r=\frac{\textit{number of possible symbols}}{\textit{number of unique symbols}}$$

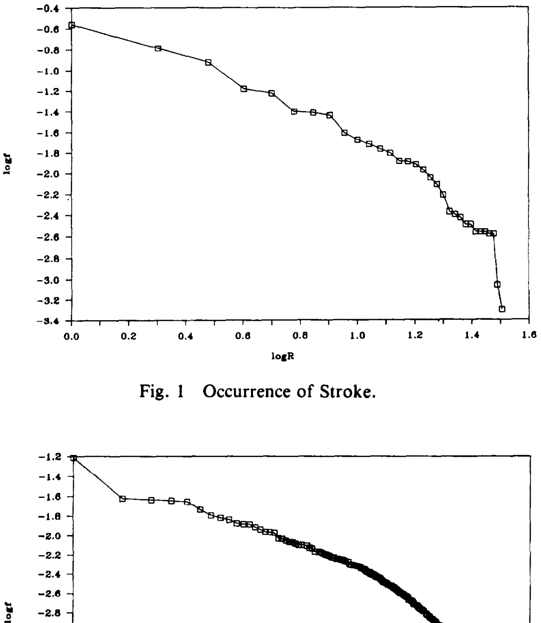

Fig. 1 Occurrence of Stroke.

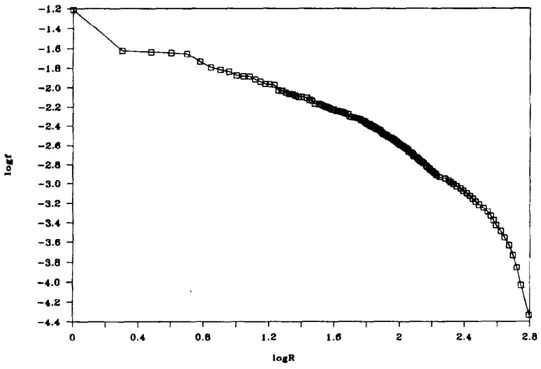

Fig. 2 Occurrence of Radical.

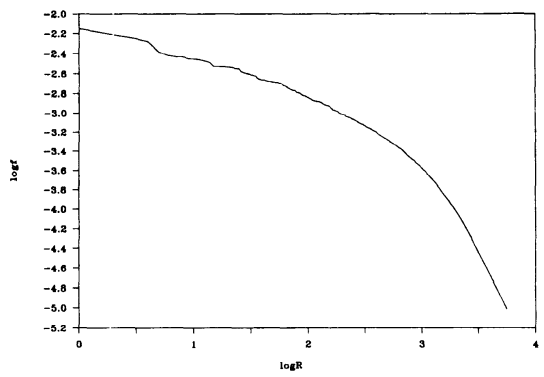

Fig. 3 Occurrence of Character.

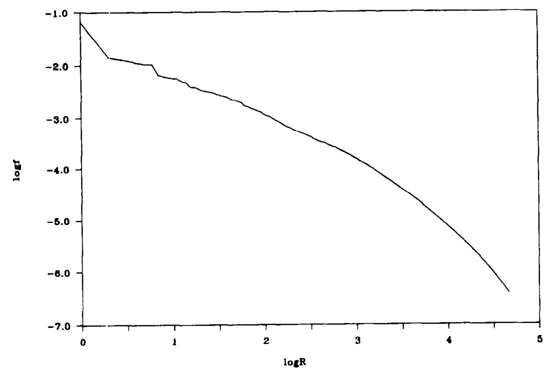

Fig. 4 Occurrence of Word.

We have found that the larger the value of r, the more the distribution moves closer to equation (1). For the present study, the values of r are 1 for all but the last category, ie occurrence of words. For example, there are 32 unique strokes in Chinese characters. In our sample, all the 32 strokes appear. Likewise, there are 623 unique radical components and 5584 unique characters. All radical components and characters occur in the samples. But this is not the case for words. In a word base derived from news items, Beijing Information Technology Institute collected 140,000 unique word items (words and phrasal words) [9]. Thus $r=3.0$ or greater. Thus the Zipfian curve of words behave differently from the other three because of the larger r. The first three distributions have sharp cutoffs at the ends whereas the cut-off for the word distribution curve is much slower.

We will therefore like to infer that only the first three curves demonstrate a complete Zipfian distribution. If we are able to compile our frequency data from a list of more words, say 200,000 words, we properly will also observe a complete Zipfian curve. [See Figs. 1-4 for Zipfian distribution]. This observation is also confirmed by the fact that the parameters derived from word distribution differ greatly from those obtained from strokes, radicals and characters (See Section 5).

4. Empirical Formulations

As we know that simple Zipf's law of $Rf=constant$ does not hold, we started from more general formulations which were extensions of the original Zipf's law, equation (1) or,

$$\log f = a - \log R \tag{3}$$

where $a=\log c$. Equation [3] is being extended to:

$$f=\sum_{i=0}^{i=n} a_i R^{-ik} \tag{4}$$

and

$$\log f=\sum_{i=0}^{i=n} a_i (\log R)^i \tag{5}$$

where a_i and k are constants.

However, careful observation also shows the following characteristics of the Zipfian distribution of Chinese linguistic symbols, ie:

(1) There exist some types of symmetry between f and R. The distribution is symmetrical with respect to some axis starting from point (a, b) and inclined to the x-axis at angle θ.

(2) The symmetric property of (1) breaks down when R is small. This is due to fact that R must be an integer while f is a real number.

Observations and experimenting again show that the distributions can be linearized to straight lines of the form (See [12], chapter XIII for description on linearization):

$$f^t = b_0 + b_1 R^t \tag{6}$$

where b_0, b_1 and t are constants.

Equations (4)-(6) are the three mathematical models used in this study.

Only equation (6) has the required property of symmetry that agrees with the previously stated observations. It is a straight line with a +1 gradient when we plot f^t against $-bR^t$.

Neither equation (4) nor (5) provide the required symmetry property for the distributions. We finally selected (5) but abandoned (4) due to the following reasons:

(1) Least-squares fitting to equation (4) yield results that over emphasis low R data due to the nature of $1/R$ function. This is undesirable because it is the occurrence of high R symbols are more important in practice. We normally wish to predict occurrence of high R symbols from data of low R symbols.

(2) Equation (3) has a relatively higher degree of symmetry due to the property of $\log f$ and $\log R$ (in comparison to that of f and $1/R$).

Finally, we performed least-square fittings using equations (5) and (6). Equation (6) is further extended to a polynomial of power series of:

$$f'=\sum_{i=0}^{i=n} b_i R^{ib} \qquad (7)$$

In selecting R value when several symbols have same frequency of occurrence, we followed the approach of Tangue and Nicholls [13]. Thus the maximum value of R is selected (see also [14]).

5. Results of Least-Square Fittings

5.1 Logarithm Polynomial

It is important to know the magnitude of experimental errors for our 4 data sets [11, 12]. This can be obtained by fitting the observed data with equation (5) from order 1 to 15 [See Fig. 5–8]. The first order fitting with equation (5) always results in the largest error. It varies from 26% (word) to almost 60% (stroke). This again confirms our earlier conclusion (in [5]) that the simple Zipf's law does not hold for Chinese characters. However, the magnitude of error of fitting drops exponentially as the order increases. It levels off at two occasions. The first occurs at relatively low orders (from 3 to 5) but the later at much higher orders (from 7 to 9). We call them the first and second minimal errors (See Table 1).

Table 1 First and Second Minimal (Logarithm Fitting).

Type	First minimal		Second minimal		No of data
	order	error	order	error	
Stroke	4	28.0%	9	13.2%	32
Radical	5	9.4%	7	2.8%	135
Character	4	3.7%	7	1.8%	222
Word	3	3.2%	8	1.4%	2416

We have to note that the first data pair for stroke distribution and the first 10 data pairs for word distribution are excluded in the least-square fitting as they do not seem to fall onto the same smooth curves. The numbers of data pairs involved in the least-square fittings are also listed in Table 1. These numbers are different from those indicated in parathesis in section 1 as we only use one data point with the highest rank when there are more than one data point having equal

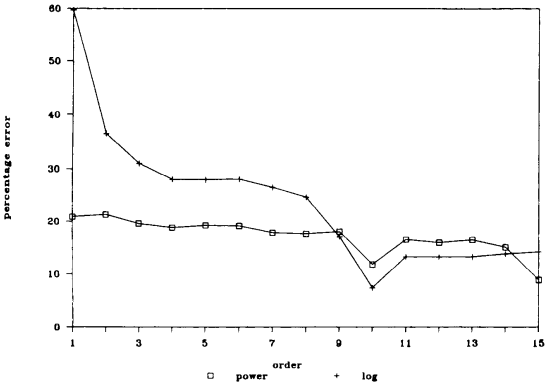

Fig. 5 Error Curves (Stroke).

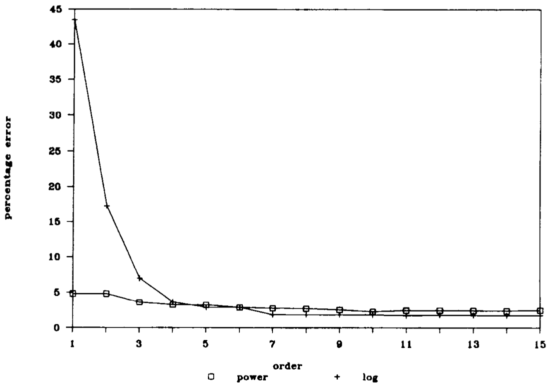

Fig. 7 Error Curves (Character).

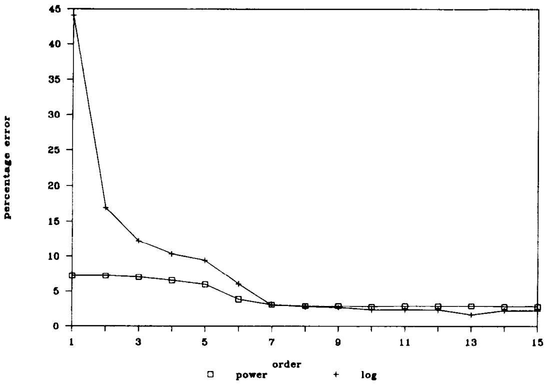

Fig. 6 Error Curves (Radical).

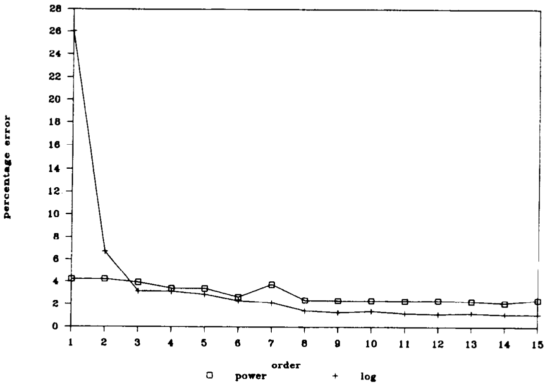

Fig. 8 Error Curves (Word).

frequency of occurrence.

The error of fitting is computed as,

$$error\ in\ f = f_{measure} - f_{computed} \qquad (8)$$

and

$$percentage\ error\ in\ f = \frac{error\ in\ f}{f} \qquad (9)$$

It is also observed that the error of fitting drops as the maximum rank increases [See Fig. 17]. This can be due to the fact that quantization error of R has maximum effect when R is small.

5.2 Linearization

Linearization (see chapter XIII of [12]) can be achieved by observing the value of b_2 during the fitting. This is the value of t when $b_2=0$ or to the nearest of it.

The t values and the errors are given in Table 2.

It is again observed that error of fitting drops as R increases. But comparing to the logarithm fitting, errors from power polynomials are almost ,doubled.

Another significant result obtained from this fitting is that $t=0.1973$ or close to 0.2 in the first three types where $r=1$. But $t=0.08$ when $r>1$. We may therefore

Table 2 Linearization Factors and Errors.

Type	t	b_0	b_1	Error of fitting
Stroke	0.197335	1.2708	−0.50396	20.9%
Radical	0.199088	0.6350	−0.13196	7.2%
Character	0.194546	0.4283	−0.05980	4.8%
Word	0.080245	1.0108	−0.29729	4.2%

Table 3 Comparing Logarithm and Power Fitting.

Type	Error of power fitting	Order of logarithm fitting
Stroke	20.9%	9
Radical	7.2%	6
Character	4.8%	5
Word	4.2%	3

Table 4 Fitting to Mandelbrot's Model.

Type	a	b	C	Error of fitting
Stroke	1.1634×10^{6}	5.2349	19	30.4%
Radical	65.403	2.0023	60	17.2%
Character	4.8354×10^{29}	9.1280	3600	16.6%
Word	1.9991	1.3463	151	13.7%

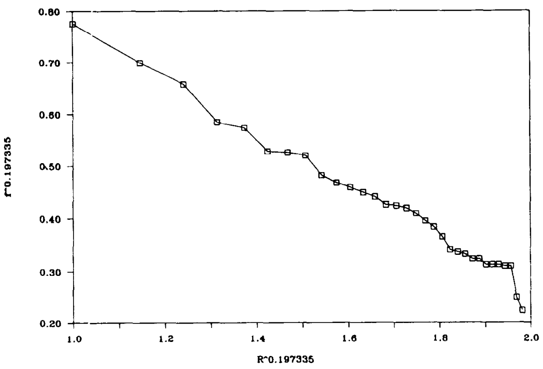

Fig. 9 Linearization (Stroke).

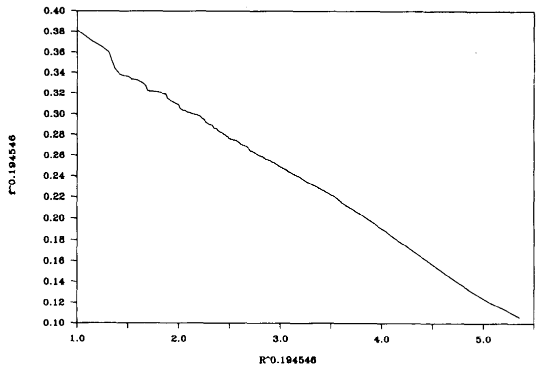

Fig. 11 Linearization (Character).

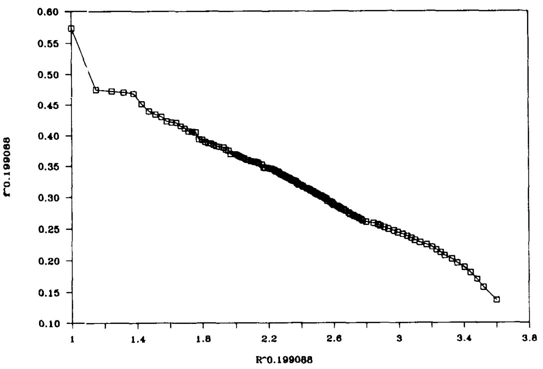

Fig. 10 Linearization (Radical).

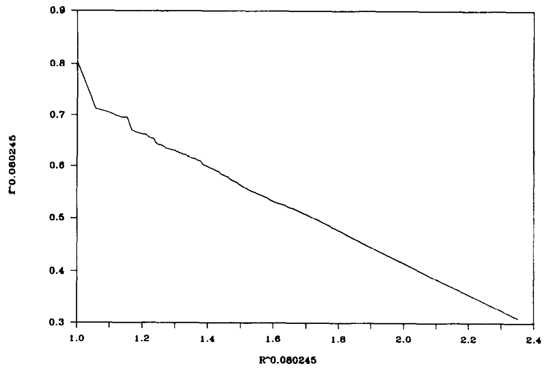

Fig. 12 Linearization (Word).

conclude that the dependency between f and R is of order 5. But we cannot derive dependency from the word distribution as it is considered as an incomplete Zipfian distribution.

We consider equation (6) a more appropriate representation for the Zipfian distribution. This is because for the same error of fitting, a much smaller set of parameters are involved.

It is seen that only for the last type, a logarithm fitting can perform as well as a power fitting.

One disadvantage of the power fitting is that it does not provide the smallest errors when we extend the fitting to higher orders. May be this is because the secondary dependency is of a lower order.

5.3 Mandelbrot's Model

We have noted that Mandelbrot [15] also provided a modified Zipf's equation with three parameters, ie

$$f=\frac{a}{(R+C)^b} \tag{10}$$

where a, b and C are constants. C must not be a negative value as it will cause problem during numercial fittings. To compare this model with our linearization function, ie equation [6], we perform least-squares fittings on the 4 data sets with various values of C. The value C is selected when equation (10) yields the minimum error of fitting (See Table 4).

The Mandelbrot's model does not provide better fitting than our equation (6) although the same number of parameters are used.

6. Residual Errors

Normally, by using equation (3) to fit the data to high order, ie $N=15$ and above, we are able to obtain an equation that fits the observed data to within their experimental fluctuations. But such an equation suffers from the disadvantage of having too many parameters. Moreover, the series obtained is not converging.

Thus we chose to fit the observed data to a lower order polynomial and do the final correction by treating their residual errors.

The residual errors are modeled by equation (2). The final representation of a Zipfian distribution is thus a sum of equation (5) or equation (6) and equation (2). Using this approach, although we are able to fit the distribution to within its limits of experimental errors,

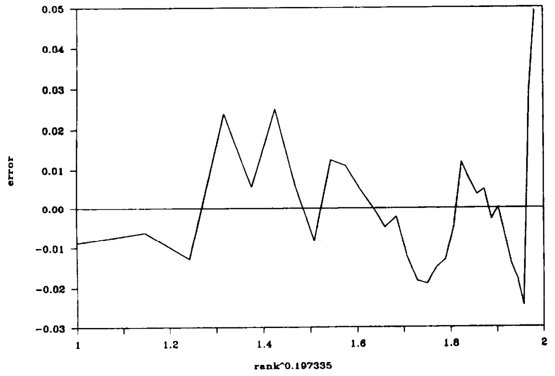

Fig. 13 Residual Error (Stroke).

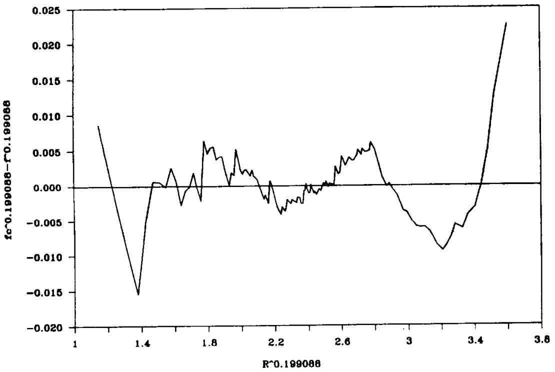

Fig. 14 Residual Error (Radical).

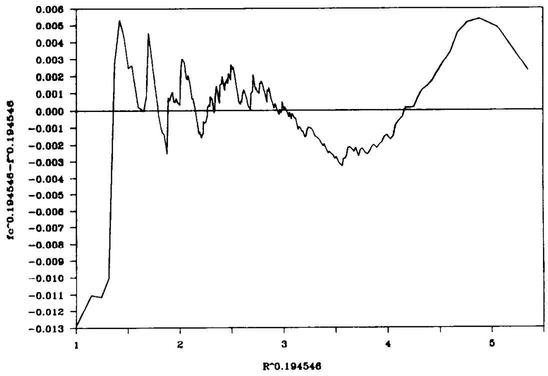

Fig. 15 Residual Error (Character).

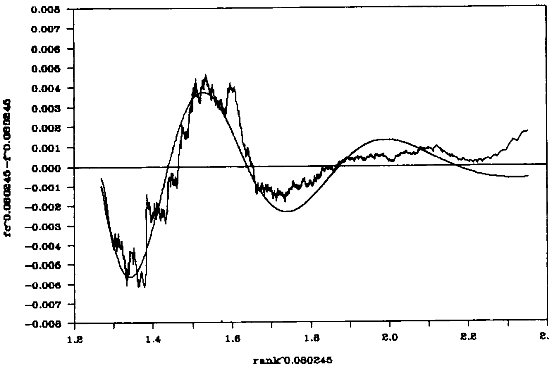

Fig. 16 Residual Error (Word).

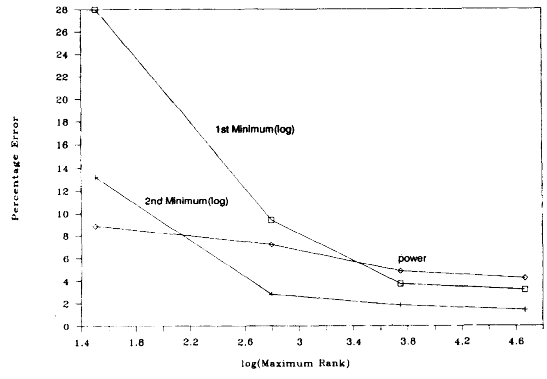

Fig. 17 Error and Maximum Rank.

we are still unable to explain its theoretical significance.

The fitting of the this equation involves two steps:

(1) obtain the envelope function, ie $\exp(a_0+a_1x+...)$

(2) obtain the period function, ie $b_0+b_1x+...$

The result of fitting is shown in Fig. 16 for the residual errors of words. It is computed that after this correction, the average error of fitting is reduced to 0.000596 or 1.6%. This last figure is comparable to a minimum value of 1.4% obtained from the logarithm fitting.

The residual error curves for other types are shown in Figs. 13 to 15. It is seen that the errors are more randomly distributed than that is shown for words in Fig. 16. However, the cosine nature of the curve can still be seen.

We have also treated the residual errors which arise from logarithm fitting. Same observations are made and our conclusions that the residual error curves are of the cosine functions remain.

7. Conclusion

In this paper, we report results of least-square fittings to 4 sets of data which are derived from Chinese characters, namely, character strokes, radicals, characters and words. We have found that fitting using a power series, ie f^t versus R^t (f is the frequency of occurrence, R the rank and t is a constant) is better than the use of a logarithm series derived from the original simple Zipf's law, ie fR=constant, or $\log f=c-\log R$. The dependency of f versus R is found to be of order 5 as we have found that $t=0.2$. We have also discovered a secondary dependency of f on R of lower order. This secondary dependency can be modeled using a cosine function. Our results also show that the error of observation reduces exponentially as the maximum rank of the data increases.

The current exercise has two major drawbacks. Firstly, we are unable to obtain a formula that fits the entire range of distribution. Often, the first few (or up to 10) pairs of data are abandoned as they do not seem to fall onto the same line. Second, the treatment of the residual errors using a cosine function has no theoretical support.

The only merit of this paper is the discovery of high order dependency between f and R. We are still far from be able to work out simple enough formula for the Zipfian distribution. We are even more remoted from being able to provide a satisfactory theoretical explanation for the observed distribution.

References

1. ZIPF, G. K. Human behavior and the principle of least effort, Addison-Wesley Press, Inc., 1948.

2. BENNETT, J. M. Zipf's law, structured programming and creativity, *Australian Computer Journal*, **16**, 4 (1984), 122-129.

3. BENNET, J. M. Storage design for information retrieval: Scarrott's conjecture and Zipf's law, Basser Dept. of Computer Science, Tech. Rep. No. 106, Oct. 1975, also appeared in Proceedings of International Computing Symposium (ACM European Chapter), Amsterdam, North Holland.

4. PHILIPS, W. J. and SHEPHERD, M. A. Statistical analysis of the rank—frequency distribution of elements in a large database, *Proc. of 13th Annual CAIS Conference: Computer Science and Information Science: at the crossing, Montreal* (Jun. 1985).

5. CLARK, J. L., LUA, K. T. and MCCALLUM, J. Using Zipf's law to analyze the rank frequency distribution elements in Chinese text, *Proc. of the 1986 International Conf. on Chinese Computing, Institute of System Science, National University of Singapore, Singapore*, 321-324.

6. A dictionary of Chinese character information, Farton Science Press Ltd, 1988.

7. LUA, K. T. Analysis of Chinese character stroke sequences, *Computer Processing of Chinese & Oriental Languages*, **6**, 2 (Mar. 1990).

8. Electronic Word Frequency Dictionary, a database of 46,520 Chinese words. The dictionary include word entries in GB2312 codes, hanyu pinyin, frequencies of occurrence and accumulated coverage. Available from Chinese and Oriental Languages Information Processing Society, Singapore (% Department of Information Systems and Computer Science, National University of Singapore, Kent Ridge, Singapore 0511).

9. Private communication with Prof Su Dongzhuang of Beijing Information Technology Technology, Dewai, Beijing, China.

10. SCHIGOLEV, B. M. Mathematical analysis of observations, Elsevier Publishing Co Inc, 1960.

11. GUEST, P. G. Numerical methods of curve fitting, Cambridge University Press, 1961.

12. VELIKANOV, M. A. Measurement errors and empirical relations, Israel Program for Scientific Translation Ltd, IPST Cat No. 1374, 1965.

13. TAGUE, J. and NICHOLLS, P. The maximal value of a Zipf size variable: sampling properties and relationship to other parameters, *Information and Management*, **23**, 3 (1987), 155-170.

14. CHEN, Y. S. and LEIMKUHLER, F. F. Analysis of Zipf's Law: An index approach, *Information Processing and Management*, **23**, 3 (1987), 171-182.

15. BENOIT MANDELBROT. Simple Games of Strategy Occurring in Communication Through Natural Languages, *IRE Transcation on Information Theory*, **IT-3** (1954), 124-137.

(Received October 16, 1990; revised June 25, 1991)