Zipf’s Law and the Frequency of Characters or Words of Oracles

Wang Xiuli
Anhui University
Hefei, Anhui 230039 China
Email: wangxiuli@ahu.edu.cn

Abstract—The article discusses the frequency of characters of Oracle, concluding that the frequency and the rank of a word or character is fit to Zipf-Mandelboit Law or Zipf’s law with three parameters, and figuring out the parameters based on the frequency, and pointing out that what some researchers of Oracle call the assembling on the two ends is just a description by their impression about the Oracle data.

I. INTRODUCTION

It is known that the American linguist George Kingsley Zipf presented the law about frequency of word and the rank of word according to its frequency, the Zipf’s law has experienced much intensive study in different domains by different researchers such as Levyand Mandelbroit, and hence has been modified to fit into data more precisely, therefore there are varieties of it, one variety of the law is

\[ f_i = \frac{c}{r^\alpha} \]

where \( f \) is the frequency, \( r \) is the rank, and \( \alpha \) is a parameter, and \( c \) is a constant. Another variety is

\[ f_i = \frac{c}{(r + a)^\alpha} \]

where \( a \) is another parameter.

George Kingsley Zipf argues that the law is valid because the least effort principle in human behavior [2]. A recent research shows that Simultaneous minimization in the effort of both hearer and speaker based on the game-theoretic evolution will lead to Zipf’s law in the transition between referentially useless systems and indexical reference systems [1].

Chinese characters of Oracles are Chinese archaic characters that are different from modern Chinese characters and inscribed on animal bones and tortoise shell. Since archaic Chinese is single-syllable language that is almost one syllable corresponds to at least one word, and one character represents one word usually. Hence we can know the frequency of words by just counting the frequency of characters. And scholars of Oracle have made some ambiguous or even wrong claim about the distribution of characters of Oracle such as “assembling on two ends”.

In this paper, we will first list some varieties of Zipf’s law, and then give the distribution of words of oracles based on the data collected from transliteration of oracles. We show the distribution is Zipf’s law with three parameters, give the value of the parameters, and explain why these parameters take such values.

II. ZIPF’S LAW

Zipf’s Law is an empirical law, there are varieties of it, here we discuss some varieties that are relevant to our study.

A. Zipf’s law with \( \alpha = 1 \)

One variety of Zipf Law or the original one is the following

\[ f_i = \frac{c}{r} \]

The following example is list about frequency and rank of words of Mandarin example 1, see the figure 2 in the bottom.

B. Zipf law with \( \alpha = 1 \)

If series

\[ F(\alpha) = \sum_{i=1}^{\infty} f_i = \sum_{i=1}^{\infty} \frac{c}{r^\alpha} \]

is defined on the complex plane, that is \( \alpha \in \mathbb{C} \), it is what mathematicians call Riemann \( \zeta \) function which has a lot of results to be applicable to areas relevant to Zipf law or the like.

There is another variety called Zipf-Mandelbroit formula

\[ f_i = \frac{c}{(a + r)^\alpha} \]

where \( 0 \leq a < 1 \) these can be regarded as terms of the generalized harmonic series:

\[ H(\alpha) = \sum_{i=1}^{\infty} f_i = \sum_{i=1}^{\infty} \frac{c}{(a + r)^\alpha} \]

III. THE WORD FREQUENCY OF ARCHAIC CHINESE AND ZIPF LAW

A. Word frequency of archaic Chinese of the classical Chinese documents and Zipf Law

There are some corpus which are in time near to oracle in history, we have counted the frequency of words of a work called ShiJi which means it is the record of the history before it’s completion, it was written in about B.C. the time of it’s writing is near Shang dynasty relatively in the term of inherited documents. In order to make comparison with word frequency...
of oracles, we list the frequency and the rank of words of ShiJi in the following:

**Example 2**

Word frequency of ShiJi

Part, see the figures 3-5 in the bottom.

Obviously, Zipf law is valid for ShiJi.

B. The Frequency of Characters or Words of Oracles and Zipf's Law

The language used in oracles is archaic Chinese, but the oracles that scholars had found and collected are in about several hundred years long time from Shang dynasty (from 1600 BC to 1046 BC) if we do not consider about the oracle bones of Zhou dynasty. It can be regarded as corpus of archaic Chinese of that period, so we expect that the corpus is fit into Zipf law.

The following is data of frequency of characters of Oracle by The Center for Studying and Application of Chinese Character of Huadong Normal University:

Based on the data above, we make them fit to Zipf-Mandelbroit formula and by calculating, get the following:

\[
f_i = \frac{c}{(a + r)^\alpha}
\]

\(\alpha \approx 1, a \approx 1\) where its parameters are \(\alpha \approx 1, a \approx 1\). We can conclude from Zipf-Mandelbroit formula that is fit to the data above the corpus of oracle or the language is different from Mandarin or ShiJi, \(\alpha \approx 1, a \approx 1\) the parameter reveals that there should be a word with higher frequency than the word with the highest frequency in the corpus. It seems the word should be a commas, but we know varieties of Zipf law are controversial over the explanation of its physics meaning.

IV. Conclusion

Word frequency or character frequency of oracle is fit to Zipf-Mandelbroit formula as following:

\[
f_i = \frac{c}{(a + r)^\alpha}
\]

where \(\alpha \approx 1, a \approx 1\). And some researches' conclusion that "the distribution of characters or character frequency of oracle assembles on the two end" is just a description by impression.

---

**References**

[1] Ramon Ferrer i Cancho and Ricard V. Sol. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791, 2003.

[2] G.K. Zipf. Human behavior and principle of least effort: an introduction to human ecology. Addison Wesley, Cambridge, Massachusetts, 1949.

**Appendix**

Received date Jun. 2014
example 2  word frequency of Shi Ji (Part)

word frequency

, 56752
, 30756
之13511
10065
工8911
不7907
以7512
為7386
子6619
而6503
曰6037
其5532
* 5387
; 5330

Fig. 3. ExampleTwo.

* 3228
于3161
人1807
公1709
也4554
將4301
; 3901
十3498
年2352
有3282
大3180
秦3114
下2879
候2823
與2660
夫2525

Fig. 4. ExampleTwo.

Fig. 5. ExampleTwo.