Zipf's Law

Viswanath Poosala
900 839 0750
poosala@cs.wisc.edu

Abstract

Many naturally occurring phenomena exhibit regularity to some extent. G. K. Zipf noticed that subjects as diverse as income distribution, word frequency and genera-species distributions exhibited a common regularity. A common empirical regularity suggests some universal principle behind the phenomena. He proposed a law to model this behavior, which is essentially an algebraically decaying function describing the probability distribution. This report describes various formulations of the law and concentrates on a few attempts by statisticians and linguists to model the causality behind such phenomena which would explain when and why Zipf's law should hold. It also attempts to consolidate various phenomena which have been empirically proven to obey Zipf's law. Finally, verification of the law using some real-life information reinforces its validity.

1 Introduction

Nature is full of phenomena which seem to obey a few laws. Some, such as falling apples, can be explained satisfactorily based on certain laws of physics or other mathematical sciences. On the other hand, there are some events, such as those occurring in chaotic systems, which do not exhibit any apparent regularity. There is another type of phenomena which exhibit empirically observable regularities, but do not directly yield to explanation using simple laws of nature. Once the laws are established, (which can be physical, mathematical or statistical etc.) these phenomena transgress to the first class. This scientific process usually proceeds by proposing a small set of laws which can model the phenomena and establishing the necessary and sufficient conditions for the occurrence of this apparent regularity.

Whenever such apparently regular phenomena are observed over large samples of data, it becomes an important task for the statisticians to detect the statistical properties of the system which cause such regularities. As an example, if we collect data regarding the relative frequency of getting r *head*s in n tosses of an unbiased coin, after many such experiments, a pattern will emerge for the relative frequencies which depends on r and n. This statistical phenomenon can be explained once it is established that each toss of a coin is a Bernaulli experiment and n such tosses will follow binomial distribution *Bin(n, 0.5)*. Then the probability of tossing exactly r heads in n attempts is simply $(nCr)(0.5)^n$. This is an example of a regular phenomena occurring in nature which can be satisfactorily explained using statistical principles.

This report deals with another such empirical phenomenon which has been observed in fields as diverse as population distribution, word usage and biological genera and species. G. K. Zipf first proposed a law (named Zipf's law) which he observed to be approximately obeyed in many of these domains [Zipf 49]. This ubiquitous empirical regularity suggests the presence of a universal principle. This report mainly concentrates on various formulations of the law and describes a few attempts at statistically explaining its theoretical underpinnings. In particular, the work relating to frequency of usage of words is presented in most detail.

A more practical goal of this project is to consolidate various cases in which Zipf's law has been empirically shown to hold. This could be valuable, for example, in designing databases involving certain statistical assumptions about the distribution of the underlying data. As an independent verification of the law, some real data distributions were also investigated.

The report is organized as follows. In section 2, we present the various formulations of Zipf's law. In section 3, we present a detailed overview of various approaches to explain the theoretical foundation of the law. Next, we brifely state the assumptions made about the underlying system's behavior by one of the derivations for Zipf's law to hold good. After that, we present a set of phenomena obeying the Zipf's law and the results of investigating

data distribution in a real life database (NBA statitistics), which essentially verified the Zipf's law. Finally, we summarize the conclusions drawn from this survey.

2 Formulation of Zipf's Law

Different versions of Zipf's law exist which vary in their generality. As explained below, the simplest form of Zipf's law has been criticized on multiple aspects and later generalized to a more complex form.

2.1 Simple form of Zipf's law

Consider a set of data values, ranked by their value such that

$$x_1 >= x_2 >= ..x_n,$$

r being the *rank* of x_r in this order. x_r can be thought of as the size of the r'th data value in the ordered set. Zipf's law is a relation between the rank of a data value and its actual value which has been empirically noted to be as follows (in a non-general form):

$$rx_r = constant \tag{1}$$

Zipf and others verified that this law holds for various kinds of domains as listed in the introduction and in chapter 5. This *rank-size relation* is known as Zipf's law and its graph is a rectangular hyperbola (fig 1).

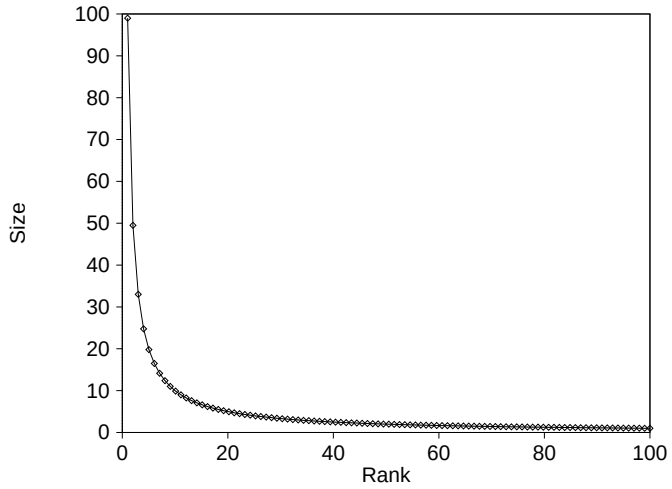

Figure 1: Size vs Rank for the simple version of Zipf's law

Let x be the size of an object and $f(x)$ be its relative frequency of occurrence, where $\int_0^\infty f(x)dx = 1$. If n is the number of objects in the data set and $N(x)$ the number of objects

with size greater than x then,

$$N(x) = \int_x^\infty nf(u)du \tag{2}$$

is the rank of the object of size x. Under Zipf's law (1), $xN(x) = K$ (some constant). Hence,

$$f(x) = -N'(x)/n = K'/x^2 \tag{3}$$

where $K' = K/n$. Equation (3) is the *size-frequency relation* corresponding to (1). G. K. Zipf attempted to explain the origins of the law in the nature of human behavior, through the principle of least effort.

The above formulation has the following deficiencies:

1. Zipf's explanation [Zipf 49] in terms of human behaviour does not explain the underlying statistical process.
2. The value of the constant K' in (3) depends on the number of objects n.
3. As discussed below, a statistical explanation for the phenomena observed by Zipf leads to a family of distributions and (3) is just a special case of them.

2.2 Generalized Zipf's law

As explained in the above section, a main drawback of the Zipf's law is that the phenomena observed by Zipf and justified by statistical rationale lead to a family of distributions, namely,

$$r^a x_r = constant, a > 0 \tag{4}$$

After some analysis, it leads to the following size-frequency relation:

$$f(r) = Ar^{-(1+a)}, r = 1, 2, ..., \tag{5}$$

where $a > 0$, and

$$A = \zeta(1 + a) = \sum_{r=1}^{\infty} r^{-(1+a)} \tag{6}$$

is the zeta function. The above equation defines the discrete Pareto distribution [John 69], which includes (3) as a special case when $a = 1$.

Figure 2 plots the above rank-frequency function for various values of a ranging from 0 to 4 in intervals of 0.5. Note that when $a = 0$, the distribution is uniform and as a increases, the skew of the function increases.

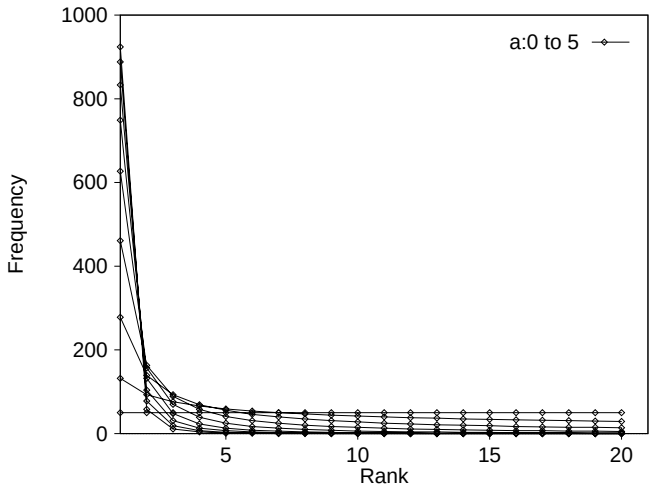

Figure 2: Frequency vs Rank for the general version of Zipf's law

3 Theoretical foundation of Zipf's law

There are at least three major schools of thought on the theoretical underpinnings of Zipf's law. Hill and Woodroofe [Hill 75, Wood 75] and others show that Zipf's law can be derived from various stochastic processes, including Bose-Einstein model and Fisher's logarithmic series distribution. Another approach, proposed by [Price 76], has been to manipulate classical occupancy models to yield hyperbolic distributions. The third approach due to MandelBrot [Mandelbrot 53] takes an information theoretic approach to studying the statistical structure of language, thus leading to the Zipf's law. The following subsection deals with Price's derivation in detail. After that, a brief description of Mandelbrot's approach is given.

3.1 Cumulative Advantage Distribution (Price's approach)

Price [Price 76] presents the cumulative advantage distribution, which can be derived as a stochastic birth process. Consider a population of n_T individuals. Let r_i be the total number of "successes" achieved by the i'th individual. Let $f(r)$ be the fraction of individuals with r successes,

$\sum_1^\infty f(r) = 1.$

The mean number of previous successes is,

$R = \sum_1^\infty r f(r).$

An individual with r successes is considered to be in state r. Transitions occur only by the incidence of a further success on an individual, which transforms the individual from state r to $r+1$. Note that this is a strictly birth-only process because no transitions occur in the reverse direction.

Now, suppose that a small number (dn_T) of new individuals are added to the populations. And with them Rdn_T new successes are also sprinkled evenly at random over *all* the population. Note that the total number of successes is proportional to the population size. Hence, the number of new successes per single previous success is,

$$dn_T/n_T. \tag{7}$$

By definition, there are $f(r)n_T$ individuals in state r (before sprinkling). Hence, the number of previous successes in state r is $rf(r)n_T$. By equation (7), the number of new successes sprinkled in the state r is,

$$rf(r)dn_T. \tag{8}$$

This is also the number of individuals transferring from state r to state $r+1$. Similarly, the number of individuals transferring from state $r-1$ to state r is,

$$(r-1)f(r-1)dn_T. \tag{9}$$

From equations 8 and 9, it follows that:

$$\begin{aligned} \frac{d}{dn_T} n_T f(r) &= -rf(r) + (r-1)f(r-1), \qquad r > 1 \\ &= -f(1) + 1, \qquad\qquad r = 1. \end{aligned}$$

Hence it follows:

$$\begin{aligned} n_T \frac{d}{dn_T} f(r) &= -(r-1)f(r) + (r-1)f(r-1), \qquad r > 1 \\ &= -2f(1) + 1, \qquad\qquad r = 1. \end{aligned}$$

The distribution over the states is defined by this series of difference equations. It can be seen that for a stable distribution, for which $f(r)$ is independent of n_T, the left-hand side of the above equation becomes zero and solving recursively:

$$\begin{aligned} f(r) &= \frac{r-1}{r+1} f(r-1) \\ &= \frac{r-1}{r+1} \cdot \frac{r-2}{r} \cdots \frac{1}{3} \cdot \frac{1}{2} \\ &= \frac{1}{r(r+1)}. \end{aligned}$$

This is a special and important form of the Zipf relationship [Ijiri 77].

3.2 Mandelbrot's derivation

Mandelbrot's work in information theory and linguistics led to another derivation of Zipf's law. He assumed that the aim of language is to transmit the most information per symbol, in the information theoretical sense of Shannon, with the least effort. He obtained the following relationship in [Mandelbrot 53],

$$f(r) = K(r + c)^{-\theta}, \tag{10}$$

where $f(r)$ is the word frequency and r is the rank of the word. The constant c improves the fit for small r and the exponent improves the fir for large r. Through a series of substitutions into a more complex argument of Mandelbrot [Mandelbrot 57] Booth demonstrated ([Booth 67]) that Zipf's law and Mandelbrot's revision are equivalent.

3.3 Simon's approach

Simon [Simon 55] expanded on Zipf's work by describing a set of empirically derived skew distribution functions. His model is also presented in terms of word frequencies. He shows that the distribution of words in a text behaves according to the following equation:

$$f(r) = A\beta(r, \rho + 1), \qquad \sum_{r=1}^{\infty} f(r) = 1, \tag{11}$$

where A and ρ are constants and $\beta(r, \rho + 1)$ is the beta function of r and $\rho + 1$ given by:

$$\begin{aligned} \beta(r, \rho + 1) &= \int_0^1 \lambda^{(r-1)}(1 - \lambda)^{\rho} d\lambda \\ &= \frac{\Gamma(r)\Gamma(\rho + 1)}{\Gamma(r + \rho + 1)}. \end{aligned}$$

As r goes to infinity and for any constant $\rho + 1$,

$$\frac{\Gamma(r)}{\Gamma(r + \rho + 1)} \rightarrow r^{-(\rho+1)} \tag{12}$$

So, from the above equation,

$$f(r) = A\Gamma(\rho + 1)r^{-(\rho+1)} \tag{13}$$

This is equivalent to Zipf's law, $f(r) = c/r^{\alpha}$, for α equal to $\rho + 1$ and c equal to $A\Gamma(\alpha)$.

3.4 Rationale behind Zipf's law

In the previous section we presented some of the statistical derivations of the Zipf's law, with tacit assumptions about the properties of the underlying system. In this section, we briefly

summarize some of the assumptions made about the system being studied. For the sake of brevity, the system being studied is limited to the usage frequency of words in literature. These assumptions are used in Simon's derivation presented above.

It has been observed [Simon 55] that the stochastic process by which words are chosen to be included in written text follows two steps:

- By processes of association, i.e., sampling earlier segments of his/her word sequences.
- By imitation, i.e., sampling from other works by self or other authors.

The assumptions made in deriving Simon's formula (11) are:

1. The probability that the $(T+1)$st word has appeared exactly r times is proportional to $rf(r,T)$, that is, to the the total number of occurrences of all words that have appeared exactly r times.
2. For large T, there is a constant probability ω that the $(T+1)$st word is a new word (hasn't appeared in the first T words).

Words chosen by association can only be the results of assumption 1 whereas words chosen by imitation can also be new. Note that these assumptions are quite valid in practice.

By assigning probabilities for imitation and association, these assumptions can be shown to lead to (13). The derivation is not complicated but is too long to be included here. It is presented in [Fedo 81].

4 Domains in which Zipf's law holds

Often in the design of various systems some assumptions need to be made about the underlying domain. For most systems dealing with non-deterministic data, even small amount of correct knowledge of the underlying data distribution can be highly beneficial. Various researchers have analyzed different data domains and identified a few of them which empirically obey the Zipf's law. It is the aim of this chapter to consolidate some of this work and present the domains in which the law has been verified to hold. The table below makes it apparent that the Zipf's law holds on vastly diverse domains and phenomena which do not have any relation.

Domain	Examples
Bibliography	Frequency of occurrence of words in an article
	Number of publications of authors
	Books by number of pages
	Citation Frequency of an article
Geography	Length of a rugged coastline
	Cities by population
Biology	Genera by number of species
Computer Sci	Distribution of data in a database
People	Income distribution of employees in a firm
	First letters of people's names
	Last names of people

5 Verification of Zipf's law on real distributions

So far in the report we have concentrated on the theoretical verification of Zipf's law and also some of the documented observations. In order to independently verify if the law holds for some of the real data distributions, a database of statistics of some NBA basketball players for the years 1991-92 was obtained [NBA 92]. These statistics include the number of goals scored by a player in a season, number of blocked shots etc. To verify Zipf's law on the number of shots blocked, we obtained the frequencies of various number of such shots. i.e, for a given number of blocked shots n, the number of players with n such shots (f_n) is obtained and plotted. If Zipf's law holds, we expect the distribution to be roughly hyperbolic, similar to one of the curves in figure 2. The frequency distribution is plotted in figure 3, for number of blocked shots.

As can be observed from the above figure, the distribution is roughly hyperbolic. It was also verified that many of the statistics yielded similar graphs, thus empirically verifying Zipf's law for a real life database.

6 Conclusions

In this report, we presented various formulations of the Zipf's law and concentrated on some of its theoretical derivations. The multiple derivations leading to the same law strongly hint at the universality of the principle. We also presented the assumptions that must hold good in the underlying systems for one of the derivations to be valid. The naturality of assumptions reinforces validity of the derivation. We have also presented a set of diverse

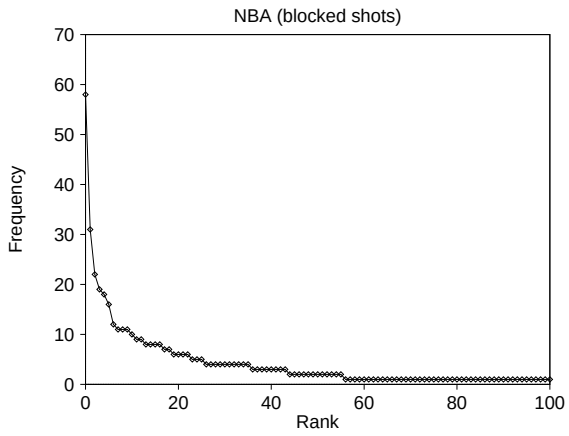

Figure 3: Frequency vs Rank for NBA statistics

domains and phenomena where Zipf's law had been empirically verified to hold. The vast disparity between these domains speaks for the generality of Zipf's law. Finally we observed that Zipf's law holds for a real-life database as well.

Postscript (or how the report *literally* verified Zipf's law)

This report dealt with Zipf's law mostly in the domain of word usage, and is probably incomplete without verifying the law on the report itself. So, as an afterthought, I measured the frequency of words used in this report (ignoring case and excluding the postscript) and the resulting plot is shown in figure 4. Clearly, the distribution is very close to being a highly skewed Zipf-ian distribution (compare with the curves in figure (2)). Not suprisingly, "zipf" is among the most frequent words, used 61 times.

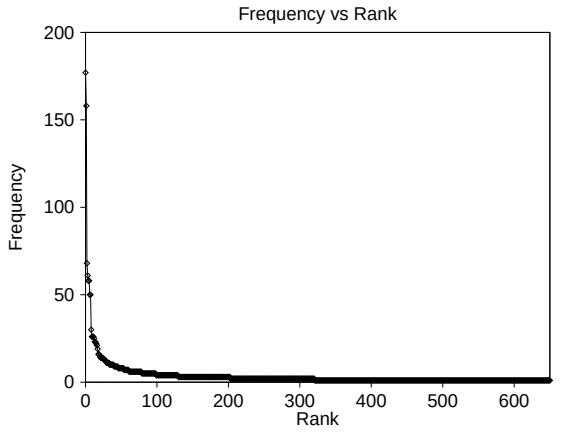

Figure 4: Frequency vs Rank for word usage in this report

References

[Booth 67] A. D. Booth, "A law of occurrences for words of low frequency", Information and control, 10(4), April 1967.

[Chen 80] Wen-Chen Chen,"On the weak form of Zipf's law", J. Applied Probability, 17, 1980.

[Fedo 81] Jane Fedorowicz, "The theoretical foundations of Zipf's law and its application to the bibliographic database environment", Journal of the american society for information science.

[Hill 74] B. Hill, "Rank frequency forms of Zipf's law", Journal of the american statistical association, 1974.

[Hill 75] Bruce Hill, Michael Woodroofe, "Stronger forms of Zipf's law", Journal of American Statistical Association, 1975.

[Ijiri 77] Y. Ijiri, H. A. Simon, "Skew distribution functions and the size of business firms", New Yotk, North Holland, 1977.

[John 69] N. L. Johnson, S. Kotz, "Distributions in statistics: Discrete distributions", Vol. 1, Wiley, New York.

[Mandelbrot 53] B. Mandelbrot, "An information theory of the statistical structure of language", Proc. symposium on applications of communication theory, Sept 152.

[Mandelbrot 57] B. Mandelbrot, "Theorie mathematique de la loi d'estoupZipf", Paris, Insitut de statistique de l'universite, 1957.

[Price 76] Price, D. De Solla, "A general theory of bibliometric and other cumulative advantage processes", Journal of American Statistical Association, 1976.

[NBA 92] "nba9192.txt", Internet (ftp: olympos.cs.umd.edu, Univ of Maryland).

[Simon 55] H. A. Simon, "On a class of skew distribution functions", Biometrika, 42, 1955.

[Wood 75] Michael Woodroofe, Bruce Hill, "On Zipf's law", J. Applied Probability, 12, 1975.

[Zipf 49] G. K. Zipf, "Human behavior and the principle of least effort", 1949, Addison-Wesley, Reading MA.