Paper Review - Zipf's Law

Anastasios Antoniadis

November 15, 2014

Abstract

Whenever phenomena with apparent irregularity are observed over large samples of data, it is important to detect the statistical properties of the system which cause such regularities. This paper presents Zipf's Law, a law proposed by G.K. Zipf to model the regularity behind some naturally occurring phenomena. This is, in essence, an algebraically decaying function describing the probability distribution.

1 Introduction

Zipf's Law is a law first formulated by G. K. Zipf using mathematical statistics. This paper presents on various formulations of Zipf's law and describes a few attempts at statistically explaining its theoretical underpinnings. Zipf's Law models an empirical phenomenon which has been observed in data studied in several physical and social sciences and the author attempts to consolidate various cases in which Zipf's law has been empirically shown to hold and concludes with the investigation of real data distributions, as an independent verification of the law.

2 Formulation of Zipf's Law

The simple form of Zipf's suggests that $rxr = constant$ where r is the rank of xr is the size of the rth data value in an ordered set. This rank-size relation is known as Zipf's Law and its graph is a rectangular hyperbola.

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 described by the zeta function.

3 Theoretical foundation of Zipf's Law

3.1 Cumulative Advantage Distribution

Price presented the cumulative advantage distribution, which can be derived as a stochastic birth process.

3.2 Mandelbrot's derivation

By assuming that the aim of language is to transmit the most information per symbol with the least effort Mandelbrot obtained the following relationship:

$$f(r) = K(r + c)^{-\theta} \tag{1}$$

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 fit for large r.

3.3 Simon's approach

Simon expanded on Zipf's work by describing a set of empirically derived skew distribution functions.

3.4 Rationale behind Zipf's law

The empowerment of Zipf's law assumes some properties about the system being studied. In the case of a system limited to the usage frequency of words in literature, Simon observed that the stochastic process by which words are chosen to be included in written text follows two steps:

- By process 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 Simon's formula are:

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

The process of association produces words which can only be the results of the first assumption, while the process of imitation can also produce words which are the results of assumption 2.

4 Verification of Zipf's law on real distributions

The author attempted to verify Zipf's law using real life data, in particular using a database of statistics of some NBA players for the years 1991-92. Many of the statistics demonstrated roughly hyperbolic graphs, which lead to the claim that Zipf's law was empirically verified for this particular real life database.

5 Paper Evaluation

5.1 Strengths

- Some categories of phenomena display some apparent regularities but are not easy to explain using simple laws of nature. However, it is essential to be able to model such categories of phenomena even if it is not (yet) possible to fully model establish a mathematical which reasons them. Having some information about a set of data can make a lot of difference compared to having no information at all. This is why Zipf's law has so many applications in spite of its empirical nature which is a limiting factor to its application.

5.2 Weaknesses

- As already mentioned the empirical nature of Zipf's law is itself limiting. A series of unbiased experiments is required in order to ascertain that the behaviour described by Zipf's law is present. Moreover, while in some cases it is possible to explain the behaviour observed (for instance in the case of Big cities vs Small cities) in other cases it is possible that the behaviour observed is accidental or the subset of a behaviour described by a more generalized mathematical model.

- The verification method used in Section 5: "Verification of Zipf's law on real distributions" is an anti-paradigm of how scientific evaluation experiments should be demonstrated in scientific work. The selective presentation of a plot and the statement that the expected behaviour was observed in other statistics is by no means a scientifically approved method of verification[1].

[1] While I am aware of the nature of this paper which was an introductory article regarding Zipf's law, its generality and its derivations and the properties of systems obeying Zipf's law rather than a full work intended for peer review, for the purpose of this homework I consider my my argument valid. In fact this paper reminded of an article by Jeffrey D. Ullman on his website: Experiments as Research Validation Have We Gone too Far?