Number Frequency on the Web

Willem Robert van Hage
SynerScope B.V.
Horsten 1, 5612 AX
Eindhoven, The Netherlands

Thomas Ploeger
SynerScope B.V.
Horsten 1, 5612 AX
Eindhoven, The Netherlands

{willem.van.hage|thomas.ploeger}@synerscope.com

Jesper Hoeksema
Network Institute
VU University Amsterdam
de Boelelaan 1081a, 1108 HV
Amsterdam, The Netherlands
J.E.Hoeksema@vu.nl

Categories and Subject Descriptors

I.5.4 [**Pattern Recognition**]: Applications—*Text Processing*; J.5 [**Arts and Humanities**]: Linguistics

Keywords

Web Science; Corpus Statistics; Numbers; Zipf's law; Benford's law; Power Law; Wikipedia; Common Crawl

1. INTRODUCTION

In this article we investigate the properties of the frequency distribution of numbers on the Web. We show that, like words, numbers on the Web follow a Power law distribution, and obey Benford's law of first-digits. We show and explain regularities in the distribution, and compare the regularities in Common Crawl to those in Wikipedia. The comparison stresses which patterns in the frequency distributions follow from human thought. We take a simple approach: We simply count how often all the numbers occur in these corpora, assume these counts to be good estimates of the distribution of the numbers in general, and draw conclusions from the properties of the observed distribution. We use Common Crawl as a general example of Web data, and Wikipedia as a general example of encyclopedic Web data. Comparing the two highlights the influence of human thought, cq. cultural influence, on the distribution of numbers on the Web. The goal of our study is to answer the following research questions:

1. Does the Power Law that governs the frequency distribution of general terms on the Web also hold for numbers?
2. What is the relation between the ordinal scale of the numbers and their frequency distribution?
3. Does Benford's law that governs the first-digit distribution of many numeric data sources hold for numbers on the Web?
4. What can be said about the relation between the frequency of numbers on Wikipedia and on the Web in general?

We consider a number to be any real number in either normal (arabic) notation or scientific (exponent) notation. Both negative and positive numbers are included and the extracted numbers are assumed to be base ten. Specifically, we use the following regular expression to detect number notations: `^[-+]?[0-9]*.?[0-9]+([eE][-+]?[0-9]+)?$` We limit ourselves to numerical notation, disregarding counting words like *"one"*, because they are relatively rare and are only used for a small number of integers close to zero. Our goal is to describe the actual frequency distribution of such numbers on the Web. To make this feasible we take a sampling approach. We use two large data sets of Web content: Common Crawl and Wikipedia. We do not do any crawling of our own for the sake of reproducibility of our results. We used the SURFsara Common Crawl data set of end 2012, containing about 3.8 billion Web pages,[1] and the english Wikipedia dump of october 2012.[2] For the processing and statistical tests we used Pig and R scripts available on GitHub.[3]

2. RESULTS

Power law: We show that research question 1: *"Does the Power Law that governs the frequency distribution of general terms on the Web also hold for numbers?"* can be answered positively. The number frequency distributions of Common Crawl and Wikipedia follow a Power Law. We base this on the observation that, as defined by Newman [1], $\alpha \approx 2$. Specifically, for Common Crawl $\alpha = 1.878912$ and for Wikipedia $\alpha = 2.201795$, so Common Crawl follows a Power law more closely than Wikipedia. Wikipedia has relatively many highly frequent numbers.

Frequency counts: The Zipf curve of the number frequencies consists of a number of separate Zipf curves, one for the integers, one per negative exponent, i.e. one starting from 1.0, one from 0.1, one from 0.01, etc. This goes for both datasets, as can be seen in Figure 1. We also see a very strong bias towards highly frequent numbers just below 2012. This peak is much stronger and wider in Wikipedia than in Common Crawl. This is due to the focus on historical dates in Wikipedia. The more recent the year, the higher the frequency. The peak starts after the first millennium. Both datasets have a very long and sparse tail of extremely high numbers ending near the maximum value of floating point numbers, which is $2^{2^7} \approx 3.4 \times 10^{38}$. Numbers

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576962.

[1] Common Crawl sara, `https://github.com/norvigaward/CommonCrawl-examples`

[2] Wikipedia dumps, `http://dumps.wikimedia.org/`

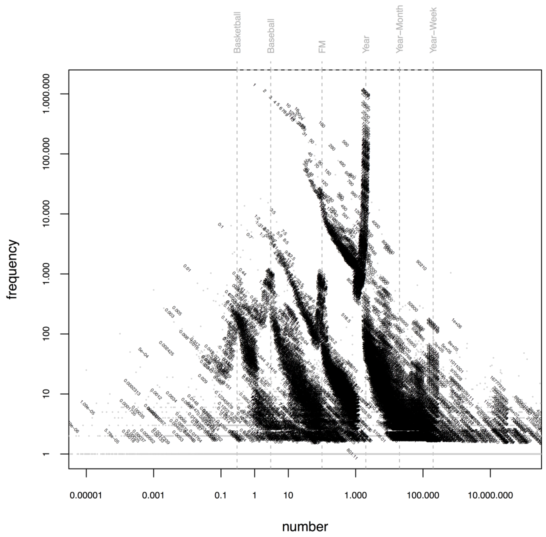
[3] Numbers on the Web, `https://github.com/wrvhage/numbers`

greater than that are mapped to infinity. The majority of the numbers is positive, in Common Crawl this is around 99% of the numbers. In Wikipedia it is 93%. This number is lower due to a relatively high number of (negative) geocoordinates west of Greenwich in Wikipedia. There is a clear predominance of round numbers in both corpora. This can be attributed to the visual beauty of the number, which leads these numbers to be used as names, such as the Fiat 500 or the movie 300. Numbers with a specific significance, such as 360 (degrees), 90210 (Beverly Hills ZIP code), 106.1 (KISS FM), 737 (Boeing), 802.1 (WiFi), and 12345678 (integer sequence), are much more common than can be expected by the average incline of the number beams.

Correlation between frequency and magnitude: It is clearly the case that low numbers are more frequent than high numbers, at least within their own power of ten. For both data sets the Pearson's product-moment correlation between the frequency and the magnitude of the numbers lies in the order of magnitude -10^{-5}, so approximately zero. This low correlation can be attributed to the distorting influence of the layered beams of the various decimal representations. If we linearize the data (take out the exponential factor) and only consider integer numbers we see a small correlation ($r \approx -0.25$) between frequency and magnitude for both datasets. This correlation is not higher due to the spread of the heavy tail of the distribution. This adds noise. For research question 2: *How does the frequency distribution of numbers relate to the ordinal number sequence?* we conclude that there is a correlation between frequency and magnitude, especially at the top of the frequency ranking. The tail of the distribution, containing mostly high numbers, is less correlated.

Most Significant Digit Distribution: Using the two-sided Kolmogorov-Smirnov test we test whether Benford's law applies by checking if the observed most significant digit probabilities could have been drawn from the expected distribution of most significant digits according to Benford's law at a confidence level of 95%. For Wikipedia $p = 0.07463$ and for Common Crawl $p = 0.8737$, so we conclude there is no significant difference between the expected distribution and the observed distributions. However, for Wikipedia the likelihood is a close call. This is most likely due to the high frequency of historical years that tend to start with 1 or 2. We conclude that research question 3: *Does Benford's law that governs the first-digit distribution of many numeric data sources hold for numbers on the Web?* can be answered positively.

Common Crawl versus Wikipedia: The final research question, 4, we want to answer is: *What can be said about the relation between the occurrence of numbers on Wikipedia in comparison to the occurrence on the Web?* In short we can conclude that the main difference between the two datasets is that Wikipedia contains more symbolic number such as model type numbers (737, 501, 360) and "beautiful" numbers (100, 66666), while Common Crawl contains more household numbers, like prices, coordinates, ZIP codes, telephone numbers, and counting numbers. It is also clear that there is some delay before years become prolific in Wikipedia, while they rise in popularity much faster on the rest of the Web. Wikipedia contains an unexpectedly high number of sports statistics pages, that completely dominate some regions of the plot. For instance, earned run averages from baseball statistics pages.



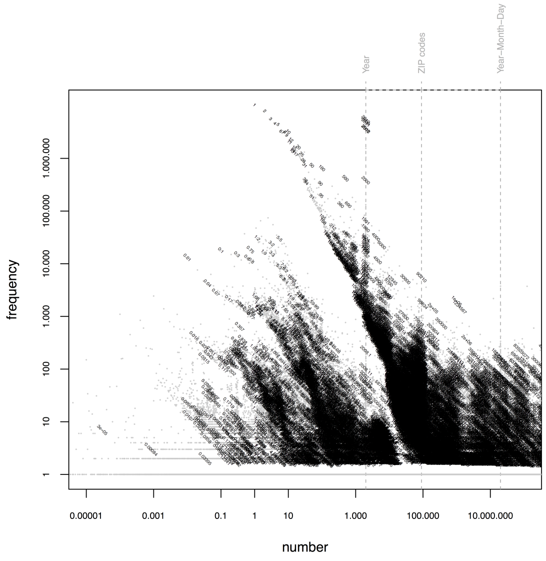


Figure 1: Wikipedia (above) and Common Crawl (below) number frequency distribution.

Acknowledgements

This research was performed in the context of the COMBINE project supported by the ONR Global NICOP grant N62909-11-1-7060.

References

[1] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005. `http://www.auditnet.org/articles/JFA-V-1-17-34.pdf`.