

Scientometrics by Zipf-Mandelbrot distribution on Wikipedia

Luc E. Brunet

DOI:10.17601/RD_Mediation2016:2

Abstract

This paper presents an application of the zip-Mandelbrot law to a large (655000) set of scientific publications in Wikipedia. The Zipf-Mandelbrot distribution shows a very good fit to the number of citations versus rank curve.

Background

Wikipedia articles often cite scientific publications as a reference in various scientific domains. At the beginning of the encyclopedia, citations were unformatted and the ontology is still on-going under dppedia for future SPARQL queries. More and more publications are described by a digital object index (DOI). The idea was to extract all the publications in Wikipedia (English version) then build a database of the number of times each of them is cited.

Previous studies sometimes based on a reduced text corpus due to the calculation resources available at the beginning of the millennial show clues that a set of publication could follow a Zipf-Mandelbrot distribution[1][2][3]. These researches make it possible to implement impact and notoriety for scientometrics applications like author assessment[4].

Methodology

We started from the data dump of Wikipedia English streamed on 6th july 2016. From this archive, we extract DOI by a regular expression 629570 references.

```
/([10].[0-9]{4,}){?:[.][0-9]+)* /(?:{?!["&\'<>]}\S)+)/
```

REGEX USED TO EXTRACT DOI

The resulting file contains some DOI to be cleaned manually. The DOI are sorted to create a database containing the DOI vs number of citations. These DOI are converted in a full structured database by querying DOI.ORG API.

The Zipf-Mandelbrot distribution is calculated through SciPY Python library :

```
import numpy as np
from scipy.optimize import curve_fit

from numpy import loadtxt

print "Zipf Mandelbrot"

lines = loadtxt("Freq.csv", comments="#", delimiter=" ",
unpack=False)

xdata = np.array(lines[:,0])
ydata = np.array(lines[:,1])

#calcul frequences
maxi=ydata[0]
ydata=ydata/maxi

# ZM(x)=a/((b+x)^m)
def ZM(x, a,b,m):
    return a/((b+x)**m)

popt, pcov = curve_fit(ZM, xdata, ydata,p0=(1,0.0,2.0))

print popt

p1 = popt[0]
p2 = popt[1]
p3 = popt[2]
simul = ZM(xdata,p1,p2,p3)
coe=np.corrcoef(ydata,simul)

print coe[0,1]
```

We correlate the number of times the citation is at least cited to the rank.

Results

The following table shows the tabulated distribution :

Nc	p
1	629570
2	134113
3	47054
4	23570
5	13237

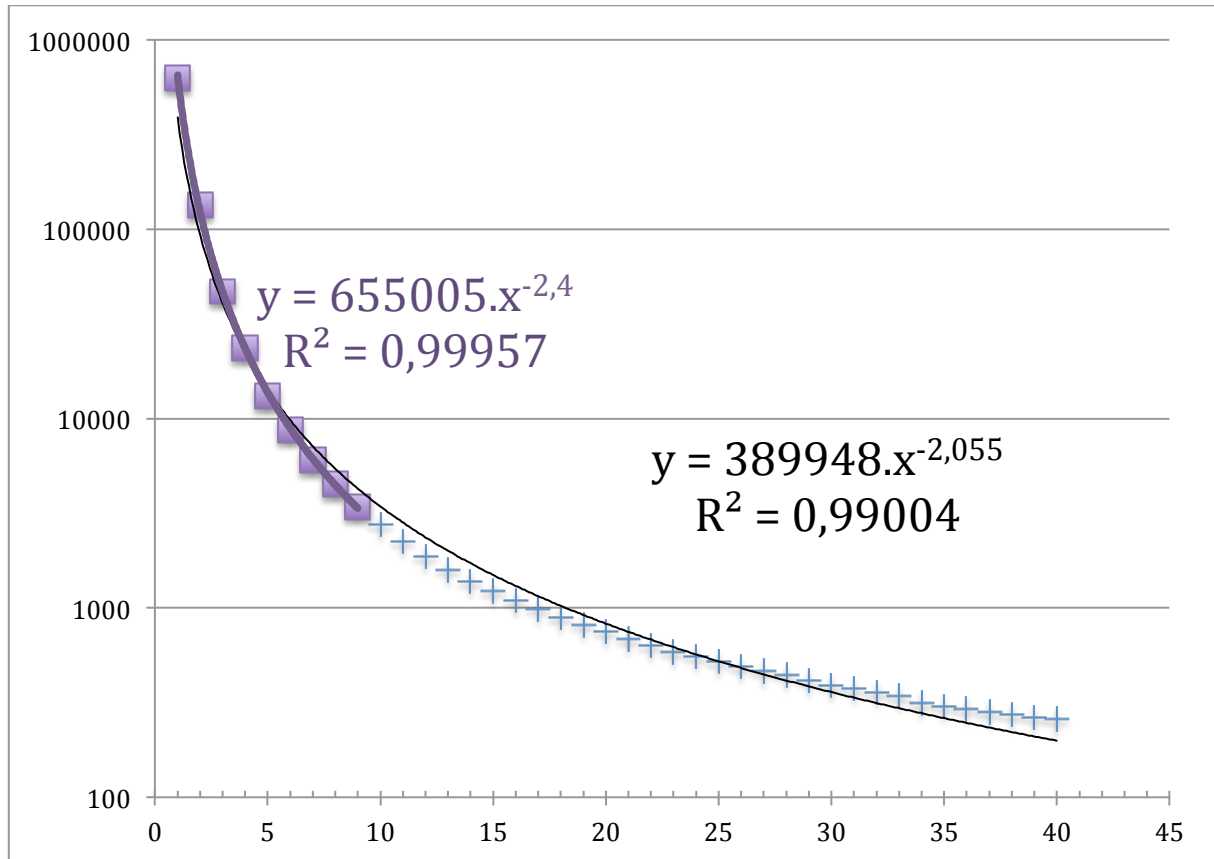
6	8730
7	6045
8	4503
9	3437
10	2752
11	2245
12	1870
13	1594
14	1377
15	1225
16	1097
17	989
18	896
19	811
20	750
21	683
22	632
23	588
24	554
25	523
26	489
27	464
28	442
29	412
30	391
31	375
32	360
33	342
34	315
35	302
36	294
37	283
38	273
39	265
40	259

Nc represents the number of times the set of p publications is cited at least. Two form were tested :

Zipf form	$f = \frac{a}{x^m}$
Zipf-Mandelbrot form	$f = \frac{a}{(b + x^m)}$

Zipf law

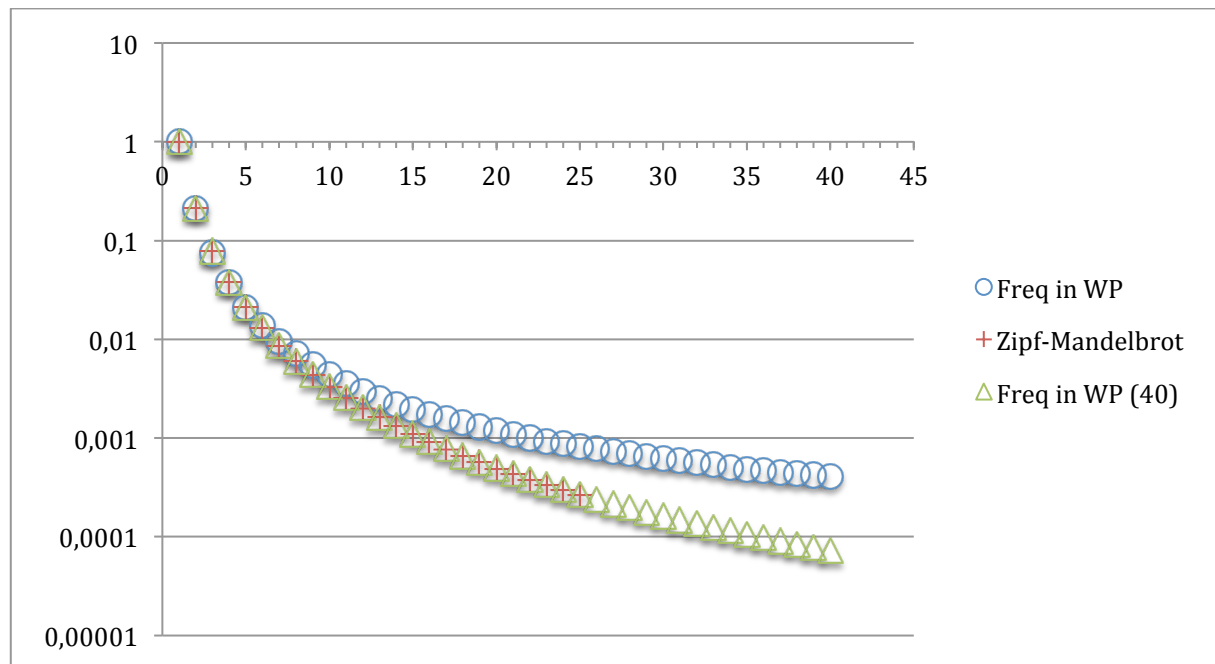
The Zipf law is often stated in word count in large corpora of texts of automated language treatment.



For our corpus, Zipf law presents already a very good agreement, especially for publications having fewer than 10 citations.

Zipf-Mandelbrot law

Zipf-Mandelbrot law fits with a correlation coefficient of 1 the frequencies from 1 to 25.



The coefficients are :

Cited at least 1 to 25 times	$\alpha=2.3558590$ $b=0.3562435$ $m=2.8118950$ $R^2=1.00000$
Cited at least 1 to 40 times	$A=2.33488803$ $b=0.352709$ $m=2.8065778$ $R^2=0.99999$

Remarks about publications

Most cited publishers (top 100 among 956)

129908 Elsevier BV
 36423 Nature Publishing Group
 26876 Informa UK Limited
 19815 JSTOR
 14521 Proceedings of the National Academy of Sciences
 11761 SAGE Publications
 9800 IOP Publishing
 8042 BMJ
 7416 The Royal Society
 6523 American Society for Microbiology
 4617 University of Chicago Press
 4346 AIP Publishing
 4227 Annual Reviews
 3053 Rockefeller University Press
 2782 Microbiology Society
 2583 Cold Spring Harbor Laboratory Press
 2506 The Endocrine Society
 2406 EDP Sciences

2297 Mary Ann Liebert Inc
 2103 Walter de Gruyter GmbH
 2001 Canadian Science Publishing
 1999 Johns Hopkins University Press
 1674 CSIRO Publishing
 1590 Thieme Publishing Group
 1570 American Physiological Society
 1478 American Society for Clinical Investigation
 1460 The Company of Biologists
 1445 University of California Press
 1435 Brill Academic Publishers
 1417 Hindawi Publishing Corporation
 1402 American Society of Hematology
 1147 Society for Neuroscience
 1105 The American Association of Immunologists
 1068 Informa Healthcare
 913 American Psychiatric Publishing
 898 American Meteorological Society

884 Emerald
855 Pleiades Publishing Ltd
854 American College of Physicians
846 World Scientific Pub Co Pte Lt
797 American Diabetes Association
764 Medknow
758 Botanical Society of America
747 Maney Publishing
738 MDPI AG
732 Frontiers Media SA
717 Pensoft Publishers
678 Duke University Press
675 Environmental Health Perspectives
660 Institute of Mathematical Statistics
621 Geological Society of London
609 Spandidos Publications
605 Royal College of Psychiatrists
597 American Thoracic Society
595 American Public Health Association
576 BioScientifica
574 Society for the Experimental Analysis of Behavior
534 Geological Society of America
506 American Economic Association
503 Scientific Societies
488 American College of Chest Physicians
458 Magnolia Press
451 Smithsonian Institution
451 FASEB
401 Pharmaceutical Society of Japan
386 Springer Berlin Heidelberg
376 PERSEE Program
364 Edinburgh University Press

347 Ecological Society of America
344 Genetics Society of America
339 Mycological Society of America
313 Copernicus GmbH
283 Ammons Scientific
278 Society for the Study of Reproduction
275 Future Medicine Ltd
274 Social Science Electronic Publishing
270 American Roentgen Ray Society
268 Schweizerbart
262 American Society for Nutrition
217 IBM
216 Guilford Publications
215 American Dairy Science Association
213 Canadian Mathematical Society
195 Mathematical Sciences Publishers
195 Frontiers in Bioscience
187 Southern Medical Association
184 ASME International
181 Indiana University Press
177 The Academy of Management
166 Wildlife Disease Association
163 Japan Antibiotics Research Association
159 GeoScienceWorld
159 Chemical Society of Japan
158 Japanese Pharmacological Society
157 CrossRef Test Account
155 American Speech Language Hearing Association
150 The Electrochemical Society
145 Sigma Xi
145 American Society of Parasitologists
141 Society for Sedimentary Geology

Most cited journals (top 100 among 18647)

15706 Nature
14657 Journal of Biological Chemistry
14250 Proceedings of the National Academy of Sciences
12901 Science
5352 PLoS ONE
5322 Journal of the American Chemical Society
4960 The Astrophysical Journal
4309 N Engl J Med
4128 Biochemical and Biophysical Research Communications
3664 Cell
3582 BMJ
3548 Nucleic Acids Research
3462 The Lancet
3215 Physical Review Letters
3018 Genomics
2987 Biochemistry
2844 The Oxford Dictionary of National Biography
2601 FEBS Letters
2020 Biographical Memoirs of Fellows of the Royal Society
1997 Journal of Molecular Biology
1917 The Journal of Cell Biology
1913 Reviews
1874 Molecular and Cellular Biology
1832 Human Molecular Genetics

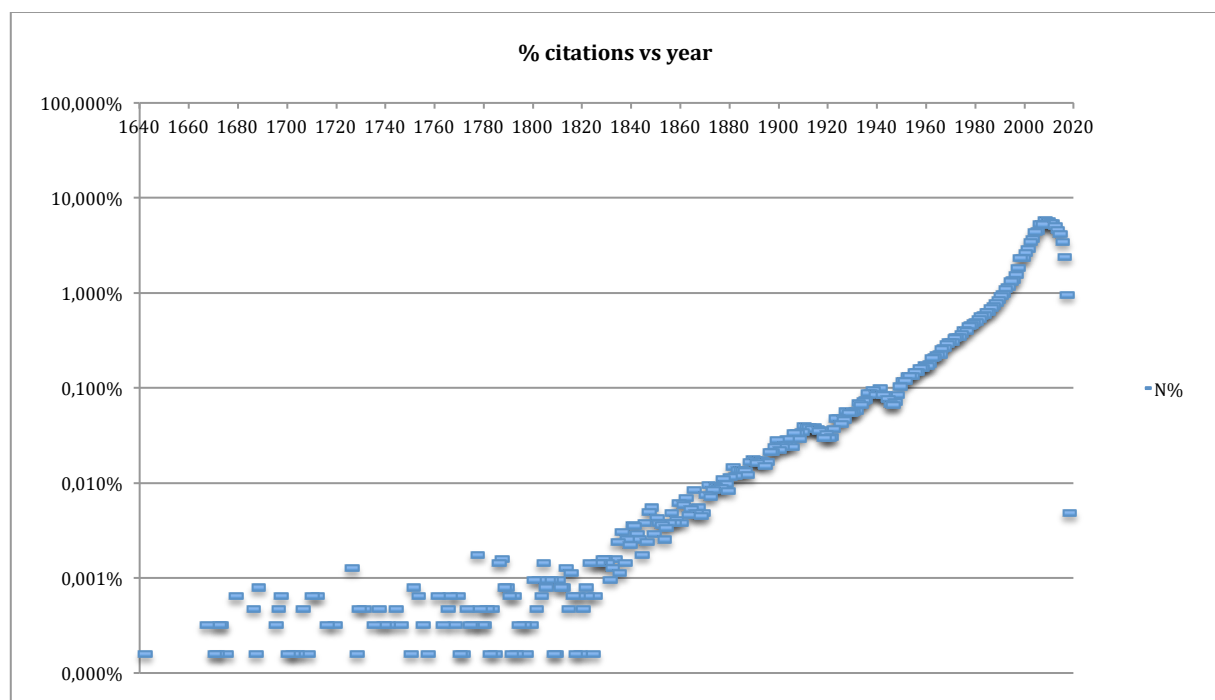
1831 Nat Genet
1789 Oncogene
1777 The EMBO Journal
1773 The American Journal of Human Genetics
1734 Monthly Notices of the Royal Astronomical Society
1717 Bulletin of the American Mathematical Society
1580 Journal of Geophysical Research
1571 Current Biology
1554 Gene
1529 European Journal of Biochemistry
1387 Journal of Medicinal Chemistry
1356 Blood
1340 INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY
1321 JAMA
1245 Molecular Cell
1224 The Astronomical Journal
1224 Lecture Notes in Computer Science
1212 Journal of Personality and Social Psychology
1194 Neurology
1186 Physical Review
1147 Journal of Neuroscience
1140 The American Mathematical Monthly
1130 Neuron

1110 Annals of the New York Academy of Sciences
 1105 The Journal of Immunology
 1083 Journal of Virology
 1074 Biochemical Journal
 1072 Circulation
 1059 Icarus
 1056 Physical Review B
 1020 Journal of Vertebrate Paleontology
 1012 Journal of Agricultural and Food Chemistry
 996 Mycologia
 994 Cancer Research
 987 Journal of Experimental Medicine
 974 Physics Today
 965 Geophysical Research Letters
 960 Molecular Phylogenetics and Evolution
 917 Bioinformatics
 915 Archives of Biochemistry and Biophysics
 862 Molecular Biology of the Cell
 854 Ann Intern Med
 852 Human Genetics
 844 PLoS Biol
 841 PEDIATRICS
 835 Applied Physics Letters
 821 Journal of Bacteriology
 791 Journal of Chemical Education
 784 Animal Behaviour
 768 Phytochemistry

757 Tetrahedron Letters
 756 American Journal of Botany
 736 Environmental Health Perspectives
 731 Applied and Environmental Microbiology
 723 European Journal of Pharmacology
 713 Journal of Mammalogy
 697 Brain
 686 Brain Research
 680 Obituary Notices of Fellows of the Royal Society
 675 Nature Biotechnology
 672 Journal of Medical Genetics
 650 British Journal of Pharmacology
 649 Molecular Biology and Evolution
 643 Ibis
 638 Endocrinology
 625 Genome Research
 619 International Journal of Systematic Bacteriology
 617 The Annals of Mathematics
 617 Journal of Investigative Dermatology
 614 The Auk
 602 PLANT PHYSIOLOGY
 601 Emerging Infectious Diseases
 598 Nat Med
 595 Evolution
 587 Journal of Cell Science
 582 American Psychologist

Timeline

The following figure shows the frequency of cited publication versus year.



- [1] L. A. Adamic and B. A. Huberman, "Zipf's law and the Internet," *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.
- [2] L. Dębowski, "Zipf's Law: What and Why?," URL [Httpwww lpipan Waw Pldebowsk](http://www.lpipan.waw.pl/debowski), 2000.
- [3] Z. K. Silagadze, "Citations and the Zipf-Mandelbrot's law," *arXiv:physics/9901035*, Jan. 1999.
- [4] M. Ausloos, "Zipf-Mandelbrot-Pareto model for co-authorship popularity," *Scientometrics*, vol. 101, no. 3, pp. 1565–1586, Dec. 2014.