1-26-2024

Content analysis and applicability of Zipf’s Law in technical writing in the Domain of Library & Information Science

Pallavi Ramkrushna Dhoke Miss
Ramrao Adik Institute of Technology, Nerul, Navi Mumbai, India, pallavi.dhoke@rait.ac.in

Mohan Kherde Dr
Retired Director Knowledge Resource Center , Sant Gadge Baba Amravati University , Amravati, mohan_kherde2@rediffmail.com

Content analysis and applicability of Zipf’s Law in technical writing in the Domain of Library & Information Science

Abstract

Through content analysis of articles published in library and information science journals, this work aims to analyse, determine, and apply Zipf's Law in the technical writing of library and information science. DESIDOC Journal of Library and Information Technology, Annals of Library and Information Studies, and Library Herald are the journals that were picked for this research. The examination of the 573 articles that were published in the three mentioned journals during the years 2013 to 2017 is the primary focus of the study. If we assume that an abstract should be between 100 and 200 words in length, then 342 articles fall within that range as per the result of this analysis. There are one to five keywords found in 390 articles. In percent, this is 68.06%. The set hypothesis is rejected in the present study which conclude that Zipf's Law in Library and Information Science technical writing is applicable.

Keywords

Abstract, Keywords, Zipf’s Law.

Introduction

Alan Pritchard, who came up with the term in 1969, defined "bibliometrics" as the "application of mathematical and statistical methods to books and other media of communication." (Rao, 1983). Bibliometrics is quantitative in nature, it can be used successfully in the field of library and information science to research literature. In the subject of library and information science, it has been observed that this kind of research is more effective at spotting literary trends. Bibliometric studies are particularly useful for examining library facilities, resource sharing, resource allocation, collection building, making decisions, and weeding.

The bibliometric laws, which highlight specific basic patterns and relationships governing information items and activities, are useful for determining various data phenomena and may aid in the planning of many library activities. Even though bibliometric laws are empirical laws, they only apply when the sociology and framework of information creation, transmission, and dissemination are constant. The fundamental organisational structure may no longer be relevant once it is altered by social and technical advancements.

Three universal laws are regarded as foundations in bibliometrics and are listed below.

1. Frequency of occurrence of words in a text (Zipf’s Law)
2. Productivity of authors in terms of scientific papers (Lotka’s Law)
3. Scattering of articles from different journals (Bradford’s Law)

American Linguist George Kingsley Zipf first introduced Zipf’s Law in 1949. Zipf’s Law is a statistical observation that describes the relationship between the frequency of a word in a text corpus and its rank in that corpus. According to the law, a word's frequency is inversely correlated to its position in the frequency. A well-known linguist named George Kingsley Zipf attempted to explore the topic of linguistics from a scientific perspective and discovered that a term's length and frequency of a word are closely associated - the more frequently a word is used, the shorter it is.

It has been observed that a variety of natural language texts including English, French, German, and Japanese found to follow Zipf's law (Kanwal, 2017). In non-linguistic circumstances, such as the frequency of cities in a nation by population or the frequency of tags on social media platforms, it has also been noticed (Ioannides, 2000).

In disciplines including linguistics, information retrieval, and natural language processing, Zipf's law has proven helpful. This law is also applicable to literary works such as fiction, poetry, and short stories (Sen, 1998). In technical writings, the concern is whether this law is applicable or not, this statement we have to check. Each phrase in technical writing typically refers to a certain concept, which can be utilized repeatedly by the author whenever they discuss that topic, increasing the frequency of words with which they are used. On the other hand, literary writing concerns the poet or author using different words with the same meaning concepts (Sen, 1998). Aside from this, technical writing may contain elements that are typically absent from literary writing, such as tables, charts, formulas, symbols, etc. It is planned to investigate the applicability of Zipf's result in the field of library and information science (LIS), which falls under the umbrella of technical writing.

Review of Literature

Numerous studies have been conducted on the study of Zipf's law, many of which have attempted to explain the fundamental mechanics of the law. One of the most popular explanations for Zipf's law is the concept of least effort, which contends that writers and speakers typically use the most prominent terms in a language to protect cognitive resources. As a result, there is a power-law distribution of word frequencies, with a small number of terms being used very often and most of them being used rarely. Several investigations have looked into how Zipf's law applies in different fields.

Sen, B. K. et. Al (1998) researched to evaluate Zipf's Law's applicability to word length and usage patterns in the library and information fields. They chose six samples for this purpose. According to a comparison of samples of about 5,800 words, the study's findings show that Zip's Law is reliable with one deviation. This low may not be suitable for words with more than one letter. This research also discovered that Zipf's Law when only the textual portion of writings—avoiding alphanumeric and alphanumeric symbolic expressions, titles, abbreviations, references, figures, and formulae—is executed into account in LIS writings.

Rajneesh and M. S. Rana (2015) applied Zipf's law in the area of computer science literature that was published in ACM journals. Out of 107,467 total keywords gathered from publications published between 1954 to 2008, the research focused on 13, 053 unique terms. The occurrence of keywords suggests that a variety of computer science books were searched to conduct the experiments. However, Zipf's Law is only consistently correct in small areas, not over the full data set.

Clark, J. I.; Lua, K T. and McCallum (1990) applied Zipf’s Law on 20,22,604 Chinese ideograms in their article. It was determined that the law is consistent with the data. The outcome of this investigation demonstrates that while this law does not apply to a single Chinese character it does apply to compound phrases that contain it.

Corral, Alvaro, Boleda, Gemma and Ferrer-i-Cancho, Ramon (2015) investigated the morphological complexity of four distinct languages over a range of time. Zipf's law is a key framework in the statistics of spoken and written natural language, as well as other communication systems. Zipf's law holds in any situation.

Kanwal, Jasmeen et al. (2017) described the relationship between word length and frequency. It is concluded that more frequently a word is used, the shorter it appears to be in their study. Language users optimise a miniature lexicon in this investigation, they employ a scaled-down artificial language learning model. This study supports Zipf's theory that the Principle of Least Effort may explain this common feature of word length distributions by proving that language users only optimise form-meaning mappings when accuracy and efficiency pressures are applied concurrently during a communicative task.

Cancho, Raman Ferrer I (2010), in his study observed that, by Zipf's law of word frequencies, the most frequent word in a text follows most of the order at frequency. It demonstrates the universal applicability of Zipf's Law.

Objectives of the study

The present study is having following objectives.

1) To figure out the proportion of keywords and the length of the abstract in technical writing in the domain of library and information science.

2) To identify the length of sentences used in the articles.

3) To identify the frequency distribution of words in the text in order of rank.

4) To verify Zipf's Law in Library and Information Science Technical Writing

Scope

"Content analysis and applicability of Zipf’s Law in technical writing in the Domain of Library & Information Science" is the topic for the current article. A few Journals published in Library and Information Science in India served as the main information source for the current study. The following three journals, published from 2013 to 2017 are taken into consideration while compiling the necessary information. These journals are mentioned below which are available online.

1) DESIDOC Journal of Library and Information Technology.

2) The Annals of Library and Information Studies and

3) Library Herald

Methodology

A total of 573 published articles in the above-mentioned journals were chosen for this study. All the words in these published articles were counted. While counting the words, the names of the authors of articles, author affiliations, alpha-numeric words, alpha-symbolic expressions, abbreviations, the punctuation mark, numbers printed in digits, periodic numbers, equations, references, tables, figures, appendices and single alphabets excepting ‘a’ were removed. For this study, only textual matter was considered. A text analyser tool used to analyse textual matter.

The references include certain fixed information, such as the author, year, article title, and other bibliographical information, which is not the author's original work. Formulas, abbreviations, numbers written with digits, alphanumeric expressions, and alpha-symbolic expressions are not considered words and are therefore not accepted. Appendices, Figures and Tables are not considered as text and therefore excluded. With the previously mentioned exceptions, only the text of the article was taken into account because it sounded to be the most pertinent for evaluating the word use.

The attempt has also been made to count number of keywords are there in every published articles in three source journals. Moreover, it also being identified that how many keywords are there in the respective titles. Similarly, number of sentences in each articles were also counted.

In this way after collecting required data, it has been analysed in consideration with the set objectives. While testing the set hypothesis the KS statistical test has been applied.

Hypothesis

The following is the hypothesis formulated for the study.

The chosen dataset from the source Journals of Library & Information Science is not suitable for Zipf's Law.

Data Analysis

Table 1: Year-wise Number of Articles Published in the Journals under Study

Sr. No.	Year	No. of Articles Published in			Total
		Annals of Library and Information Studies (ALIS)	DESIDOC Journal of Library and Information Technology	Library Herald	
1	2013	35	60	22	117
2	2014	34	59	24	117
3	2015	37	51	24	112
4	2016	31	50	32	113
5	2017	28	56	30	114
Total		**165**	**276**	**132**	**573**

Table 1 shows the year-wise number of articles published in three journals viz. Annals of Library and Information Studies, DESIDOC Journal of Library and Information Technology, and Library Herald published in the domain of Library and Information Science considered for the present study. The data for the study was collected from these journals published during the year 2013 to 2017. A total of 573 articles were published in five years from 2013 to 2017. Out of 573 articles, the highest number of articles were published in DESIDOC Journal of Library and Information Technology i.e. 276. Table 1 also shows the lowest number of articles published in the Library Herald i.e. 132. The highest number of articles were published in the year 2013 in the DESIDOC Journal of Library and Information Technology i.e. 60 and compared to others the lowest number of articles were published in Library Herald in 2013 i.e. 22. In the years 2013 and 2014, the highest number of articles were published i.e. 117 respectively.

The abstract is the most essential part of any research article. The abstract includes the basic background of the study, methods used, results and conclusions. After reading the abstract, users can understand whether the article is useful for their study or not. Generally, the abstract should be between 100 to 200 words in length. The policy regarding it may vary from journal to journal. But it should not be too long. In all three source journals, the Author's Guidelines are given in which the required length of abstract is given. In the DESIDIC Journal of Library and Information Technology, it is mentioned as 200 words. In Annals of Library and Information Studies, it is 150 to 200 words whereas in Library Herald it is mentioned that the length of the abstract should be 300 words.

In this study, tried to check the length of abstracts of articles published in source journals by considering their words and it is displayed in Table 2. From Table 1 it shows that a total of 573 articles were published in the selected three journals from 2013 to 2017. Table 2 shows that out of 573 articles, the length of the abstract of 188 articles has ranged between 101 to 150 words. The percentage of which is 32.81 %. The next lowest figure is 154 articles. The number of words in the abstract of these articles is 151-200. The percentage of it is 26.88 %. If we consider that the standard length of the abstract should be 100 to 200 words then 342 (188 + 154) articles are coming in the range of it. The percentage of it is 59.69 %. From Table 2 it is also observed that 566 articles (98.78 %) have an abstract length of up to 300 words.

Table 2; Number of articles in terms of length of abstract considering the number of words

Sr. No.	Range of Words in Abstract	No. of Articles	Percentag e	Cumulative No. of Articles	Cumulative Percentage
1	00-50	11	1.92 %	11	1.92 %
2	51-100	111	19.37 %	122	21.29 %
3	101-150	188	32.81 %	310	54.10 %
4	151-200	154	26.88 %	464	80.98 %
5	201-250	78	13.61 %	542	94.59 %
6	251-300	24	4.19 %	566	98.78 %
7	301-350	6	1.05 %	572	99.83 %
8	351-400	0	0	572	99.83 %
9	401-450	1	0.17 %	573	100 %
Total		**573**	**100 %**		

Table 3: Number of Articles in terms of Number of Keywords

Sr. No.	Range of Keywords	No. of Articles	Percentage
1	0	14	2.44 %
2	01-5	390	68.06 %
3	06-10	161	28.10 %
4	11-15	8	1.4 %
Total		**573**	100 %

Keywords indicate what concepts are discussed in the article. It helps in searching the relevant information or articles. Therefore, in searching the information keywords play a vital role. Because of this every research article must have keywords. As a standard nothing is mentioned elsewhere about how many keywords should be given to any research article. But generally, it should be up to 10 keywords. Among, three source journals in the Author’s Guidelines of DESIDOC Journal of Library and Information Technology, it is mentioned that authors should put a maximum of 6 to 10 keywords in their article. In the present study, out of 573 articles published in three source journals during the years 2013 to 2017, in 390 articles, 1 to 5 keywords are observed. The percentage of which is 68.06 %. Likewise, in 161 articles 6 to 10 keywords are observed. The percentage of which is 28.10 %. It comes to the notice that out of 573 articles, a total of 551 articles have keywords up to 10 excluding 0 keyword articles. Only 14 articles have no keywords.

Table 4; Percentage occurrence of Keywords in the title

Sr. No.	Range of Percentage of Keywords	No. of title
1	00-20	148
2	21-40	183
3	41-60	122
4	61-80	93
5	81-100	27
Total		**573**

Keywords are the words used to search the concerned articles. They are important to many facets of content optimisation, information retrieval, and natural language processing. “Keyword presence in titles is a critical aspect of various content-related strategies. Keyword-rich titles can aid readers in recognising a document's theme or subject matter more quickly. Some of the journals’, keyword criteria are fixed but not all journals follow the same rule. Out of three journals selected for the study, in only one journal, the keyword criteria is fixed which 6 to 10 keywords per article is and this journal is DESIDOC Journal of Library and Information Technology.

Table no. 4 shows that out of 573 articles the highest no. of occurrence of keywords in the title i.e. 183 numbers covered 21-40 percent. Followed by 148 keywords coming in at 00-20 percent. The lowest number of keywords occurred in the title is 81 to 100 percent.

While analysing 573 articles published during the years 2013 and 2017, this study tried to determine the number of sentences in each article. The result of this analysis is displayed in Table 5. It can be seen that the highest number of articles i.e. 332 articles are in the range of 101 to 200 sentences. Followed by 129 articles, in the range of 201 to 300 sentences. However in the range of 601 to 700, 701 to 800 and 801 to 900 only one article was published each. It means that a high sentence range of articles published in journals is very rare.

Table 5. Length of Articles in Terms of Number of Sentences

Sr. No.	Range of Sentences	No. of Articles	Percentage
1	001-100	64	11.17 %
2	101-200	332	57.95 %
3	201-300	129	22.51 %
4	301-400	35	6.12 %
5	401-500	7	1.22 %
6	501-600	3	0.52 %
7	601-700	1	0.17 %
8	701-800	1	0.17 %
9	801-900	1	0.17 %
	Total	**573**	**100 %**

Table 6. Length of Articles in Terms of number of words

Sr. No.	No. of Words	No. of Articles	Percentage
1	0001-1500	19	3.31 %
2	1501-3000	286	49.91 %
3	3001-4500	196	34.20 %
4	4501-6000	50	8.72 %
5	6001-7500	17	2.95 %
6	7501-9000	3	0.52 %
7	9001-10500	2	0.39 %
	Total	**573**	**100 %**

When we are talking about the length of research articles, generally it should be 2500 to 3000 words. Given this, the study analysed 573 articles published during the years 2013 to 2017 in three source journals. This analysis is presented in Table 6. From this table, it is observed that out of 573 articles, the length of 286 articles ranges between 1501 to 3000 words. The length of 196 articles ranges between 3001 to 4500 words. It is also observed that 3.31 % of articles (i.e. 19 articles) are short in length having the range between 0001 to 1500 words. On the contrary, few articles mean 3 published articles had a word length between7501 to 9000. The percentage of which is only 0.52 %. Similarly, 0.39 % i.e. only two articles are with the words in between 9001 to 10500.

Testing of Hypothesis

As mentioned above the following hypothesis was formulated to fulfil the forth objective of the study i.e**.** ‘to verify Zipf's Law in Library and Information Science Technical Writing’.

The chosen dataset from the source Journals of Library & Information Science is not suitable for Zipf's Law.

In the study 573 articles published in three source journals were analysed. While analysing the data, in each 573 articles rank of words and their frequencies were identified that are termed as observed rank and observed frequencies of words respectively. It has been done because the statement of Zipf’s Law is mathematically written as –

$$r * f = c$$

Where r is the rank of the words, f is the frequency of each words and c is constant.

With this mathematical formula, on the basis of observed ranks and observed frequencies of words of each 573 articles the observed constant has been calculated. Later on for fulfilling the above forth objective of the study the Kolmogorov-Smirnov Test (KS Test) was applied to all of these sample articles. The outcome of this test is that, in all 573 samples, the Asymptotic Significance 2 tailed value for all variables is less than our value of significance (i.e., 0.000<0.05). As a result, we were unable to accept the null hypothesis. This result of is displayed in table 7 as null hypothesis accepted and null hypothesis rejected. From table 7 it is observed that in all 573 cases the null hypothesis has been rejected and hence it conclude that Zipf's Law in Library and Information Science technical writing is applicable.

Table 7. Hypothetical Result of Zipf's Equations considering all Samples

Sr. No.	Year	Null hypothesis		Total	Percentage (Accepted)
		Accepted in case of No. of Articles	Rejected in case of No. of Articles		
1	2013	0	117	117	0%
2	2014	0	117	117	0%
3	2015	0	112	112	0%
4	2016	0	113	113	0%
5	2017	0	114	114	0%
Total			573	573	0%

Conclusion

This present study not only discussed the Zipf's law but also focuses on other features of the source articles, such as length of abstracts, number of Keywords, proportion of keywords in the title of articles, sentences, and word rankings. The standard length of abstract is generally in between 100 to 200 words. However, in this study it reveals that near about 60% articles are having there abstract in between 100 to 200 words. So far as number of keywords are concern it is expected that this number should be in the range of 6 to 10 words. However, in the present study out of 573 articles, 161 articles having the keywords in this range. The percentage of which is only 28.1 %. It is noticed that, generally, length of research articles should be between 2500 to 3000 words long. But in the present study it is observed that near about 47% articles having more than 3000 words. The study investigated whether Zipf's Law could be applied to a corpus of technical writing in the field of library and information science or not. The findings showed that Zipf's law is appropriate for the chosen dataset of source journals in the library and information science fields when only the textual portion of the writing is taken into consideration.

References

1) Booth, Andrew D. (1967). A Law of Occurrences for Words of Low Frequency. *Information and Control*, 10, 386-393.(2020, March 9) Retrieved fromhttps://www.sciencedirect.com/science/article/pii/S001999586790201X
2) Ferrer-i-Cancho Ramon and Sile, Richard V. (2002). Zipf's Law and Random Texts. *Advances in Complex Systems, 5*(1), 1-6.

3) Clark,J. L.; Lua, K. T.; McCallum, J. (1990) Conformance of Chinese text to Zipf's law. *International Conference of Chinese Computing*, USA. Retrieved from https://ieeexplore.ieee.org/document/77200.

4) Corral, Alvaro; Boleda, Gemma, & Ferrer-i-Cancho, Ramon (2015). Zipf's law for word frequencies: Word forms versus lemmas in long texts. *PLoS One*, *10*(7) doi:http://dx.doi.org/10.1371/journal.pone.0129031

5) Devarajan, G. (1997). *Bibliometric Studies.* New Delhi, Ess Ess Publications.

6) Ferrer-i-Cancho Ramon, Elvevag, Brita (2010) Random Texts Do Not Exhibit the Real Zipf's Law-Like Rank Distribution. *PLoS ONE* 5(3); doi:10.1371/ journal.pone.0009411

7) Hill, Bruce M. (1974). The Rank-Frequency Form of Zipf's Law. *Journal of the American Statistical Association*, *69* (348), 1017-1026.

8) Kanwal, Jasmeen et.al. (2017) Zipf's Law of Abbreviation and the Principle of Least Effort: Language users optimise a miniature lexicon for efficient communication, *Cognition*,165,45-52.Retrived from *(https://www.sciencedirect.com/science/article/pii/S0010027717301166)*

9) Piantadosi, Steven T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychan Bull Rev. 21* (5), 1112-1130. (2020, March 12). Retrieved from https://link.springer.com/article/10.3758/s13423-014-0585-6

10) Rajneesh and Rana, M. S. (2020), Content Analysis and application of Zipf's Law in Computer Science Literature. Retrieved from https://ieeexplore.ieee.org/document/7048202

11) Rao, I. K. Ravichandra. (1983). *Quantitative Methods for* Library *and Information Science.* New Delhi, Wiley Eastern Limited.

12) Sen, B. K. et.al. (1998). Zipf's Law and Writings on LIS. *Malaysian Journal of Library & Information Science*, *3*(2), 93-98.

13) Su, Kuichun (2001). Comparing Frequency of Word Occurrence in Abstracts and Texts Using Two Stop Word Lists. (2020, March 12). Retrieved from https://www.researchgate.net/publication/11535557_Comparing_frequency_of_word_occurrences_in_abstracts_and_texts_using_two_stop_word_lists