International Archive of Applied Sciences and Technology
Int. Arch. App. Sci. Technol; Vol 9 [2] June 2018: 36-40
© 2018 Society of Education, India
[ISO9001: 2008 Certified Organization]
www.soeagra.com/iaast.html

IAAST
ONLINE ISSN 2277- 1565
PRINT ISSN 0976 - 4828

CODEN: IAASCA ORIGINAL ARTICLE

DOI: .10.15515/iaast.0976-4828.9.2.3640

Impact of Zipf's Law in Information Retrieval for Gujarati Language

Rajnish M. Rakholia and Jatinderkumar R. Saini
[1]MCA Department, S. S. Agrawal Institute of Computer Science, Navsari, Gujarat
[2] MCA Department, Narmada College of Computer Application, Bharuch, Gujarat
Email: rajnish.rakholia@gmail.com, saini_expert@yahoo.com

ABSTRACT
This paper present Zipf's law distribution for the information retrieval. Based on large corpus of Gujarati written texts the distribution of term frequency is much skewed. Very small number of terms occurs more often than a large number of terms occur with low frequency. This distribution impact the performance of Information Retrieval in which only medium terms frequency can be considered for indexing. Zipf's law holds that the frequency of the term is inversely proportional to the rank. Taking the log-log plot for frequency and rank, the graph will be straight line with slop close to -1. If the linear fit is not closer to the Zipfian ideal of -1 than those terms are discarded to improve the efficiency of information retrieval. The Zipf's law provides a reasonable summary of the data for the very frequent and less frequently occurs terms but predict the few different terms which is more relevant to the content and user query.
Keywords: *Zipf's law, Information Retrieval, Gujarati language*

Received 22/02/2018 Revised 20/03/2018 Accepted 09/05/2018

Citation of this article
Rajnish R, Jatinderkumar S. Impact of Zipf's Law in Information Retrieval for Gujarati Language. Int. Arch. App. Sci. Technol; Vol 9 [2] June 2018. 36-40.

Nomenclature:

c	:	Constant
r	:	Rank
f	:	Frequency

INTRODUCTION

To retrieve the most relevant documents from the web is a significant task to satisfy the demands of different users. It is more difficult for the resource poor language like Gujarati, Panjabi, Marathi and other Indian languages. Main objective of this research is to enhance the performance of Information Retrieval (IR) and other Natural Language Processing (NLP) applications such as library system, mail classification, sentiment analysis and survey classification etc., for Gujarati language by discarding terms from the documents [2,3].

Gujarati Language

Gujarati is an official and regional language of Gujarat state in India. It is 23rd most widely spoken language in the world today, which is spoken by more than 46 million people. Approximately 45.5 million people speak Gujarati language in India and half million speakers are from Outside of India that includes Tanzania, Uganda, Pakistan, Kenya and Zambia. Gujarati language is belongs to Indo-Aryan language of Indo-European language family and it is also closely related to Indian Hindi language.

Zipf's law distribution

It is a distribution model to predict the term which is most relevant to the user query and expected to occur in corpus. According to the Zip's law frequency of term in the corpus is inversely proportion to the rank of term. Zipf's law is mostly used for large corpus and very

small number of terms occurs more often than a large number of terms occur with low frequency. The Zipf's law provides a reasonable summary of the data for the very frequent and less frequently occurs terms but predict the few different terms which is more relevant to the content and user query

MATERIALS AND METHODS

Zipf [1] found that, the frequency of term in data is inversely proportional to the rank of that term in frequency list.
Rank (r): the numerical position of the term in frequency list sorted by decreasing order.
According to ths Zipf's law:

$$\text{Frequency (f)} \times \text{Rank (r)} = \text{Constant (c)} \quad (1)$$

A Zipf's law is power law in the form of $y = kx^c$, where c is constant and c= -1

A power law gives a straight line which is closer fit to the constant c= -1, by taking the logarithms of each side:

$$\log (y) = \log(kx^c) = \log (k) + c \log(x) \quad (2)$$

If Zipf's law holds a collection then graph of log (f) against log(r) gives a roughly leaner with slop close to -1. Slop for head and tail may not be linear which will be discarded for further process.

Data sets

The travelling corpus was created for experiment purpose. The data was collected from multiple free Gujarati websites to avoid the bias of a single website. The corpus contained total 132,509,08 words in which 53,566 were unique words. From this corpus we have created 272 documents and each document that contained almost 48000 words. Experiment was carryout for each document.

EMPIRICAL SETUP

Table-1 presents head terms (frequency based sort in decreasing order) from document which is randomly selected from travelling corpus. Figure-1 presents a log-log plotting of randomly selected document from corpus. Selected document contained 48709 words in which 9958 were unique words.

Table-1 Sample of head terms from randomly selected document

Rank (R)	Words	Frequency(F)	Log10 (R)	Log10 (F)	Rank (R)	Words	Frequency (F)	Log10 (R)	Log10 (F)
1	છે	1377	0.00000	3.13893	50	મંદિર	68	1.69897	1.83251
2	આ	659	0.30103	2.81889	51	પાણી	65	1.70757	1.81291
3	પણ	633	0.47712	2.80140	52	થાય	65	1.71600	1.81291
4	જે	602	0.60206	2.77960	53	ના	63	1.72428	1.79934
5	અને	578	0.69897	2.76193	54	માટે	63	1.73239	1.79934
6	એક	565	0.77815	2.75205	55	ઉપર	61	1.74036	1.78533
7	તો	549	0.84510	2.73957	56	ગયા	61	1.74819	1.78533
8	હતી	350	0.90309	2.54407	57	વાર	61	1.75587	1.78533
9	એ	330	0.95424	2.51851	58	સામે	60	1.76343	1.77815
10	પર	324	1.00000	2.51055	59	બહાર	60	1.77085	1.77815
11	હતો	296	1.04139	2.47129	60	આગળ	59	1.77815	1.77085
12	હતું	285	1.07918	2.45484	61	તેની	58	1.78533	1.76343
13	પછી	262	1.11394	2.41830	62	જતી	58	1.79239	1.76343
14	અહીં	245	1.14613	2.38917	63	મને	57	1.79934	1.75587
15	તે	241	1.17609	2.38202	64	શરૂ	57	1.80618	1.75587
16	હોય	178	1.20412	2.25042	65	આવ્યો	55	1.81291	1.74036
17	હતા	165	1.23045	2.21748	66	નજર	55	1.81954	1.74036
18	થઈ	164	1.25527	2.21484	67	બસ	55	1.82607	1.74036
19	સાથે	163	1.27875	2.21219	68	પાસે	54	1.83251	1.73239
20	કરી	153	1.30103	2.18469	69	લઈ	54	1.83885	1.73239
21	કોઈ	153	1.32222	2.18469	70	છુ	53	1.84510	1.72428
22	આવી	149	1.34242	2.17319	71	ગયું	53	1.85126	1.72428
23	જાય	141	1.36173	2.14922	72	ગયાં	53	1.85733	1.72428

24	હશે	138	1.38021	2.13988	73	જરા	52	1.86332	1.71600
25	નથી	133	1.39794	2.12385	74	રહ્યો	52	1.86923	1.71600
26	ન	132	1.41497	2.12057	75	જતાં	52	1.87506	1.71600
27	હતાં	126	1.43136	2.10037	76	જવું	51	1.88081	1.70757
28	ત્યાં	121	1.44716	2.08279	77	ફરી	51	1.88649	1.70757
29	નહીં	120	1.46240	2.07918	78	જોઈએ	50	1.89209	1.69897
30	અમે	115	1.47712	2.06070	79	નીકળી	50	1.89763	1.69897
31	હવે	115	1.49136	2.06070	80	જવા	50	1.90309	1.69897
32	એટલે	114	1.50515	2.05690	81	જઈ	50	1.90849	1.69897
33	જાણે	103	1.51851	2.01284	82	કરતાં	49	1.91381	1.69020
34	ગયો	103	1.53148	2.01284	83	કદાચ	49	1.91908	1.69020
35	આજ	102	1.54407	2.00860	84	તરફ	49	1.92428	1.69020
36	દૂર	100	1.55630	2.00000	85	તેમ	49	1.92942	1.69020
37	ગઈ	96	1.56820	1.98227	86	વળી	49	1.93450	1.69020
38	રહી	90	1.57978	1.95424	87	એવું	48	1.93952	1.68124
39	વાત	84	1.59106	1.92428	88	નદી	48	1.94448	1.68124
40	જોઇ	83	1.60206	1.91908	89	અનેક	48	1.94939	1.68124
41	વચ્ચે	80	1.61278	1.90309	90	શું	48	1.95424	1.68124
42	ત્યારે	79	1.62325	1.89763	91	આવ્યા	47	1.95904	1.67210
43	બે	79	1.63347	1.89763	92	કહ્યું	47	1.96379	1.67210
44	આવે	72	1.64345	1.85733	93	રીતે	47	1.96848	1.67210
45	સુધી	71	1.65321	1.85126	94	સાંજ	46	1.97313	1.66276
46	લાગે	70	1.66276	1.84510	95	કરે	45	1.97772	1.65321
47	નામ	69	1.67210	1.83885	96	બાજુ	45	1.98227	1.65321
48	બની	69	1.68124	1.83885	97	આવ્યું	44	1.98677	1.64345
49	નીચે	69	1.69020	1.83885	98	ઊભી	43	1.99123	1.63347

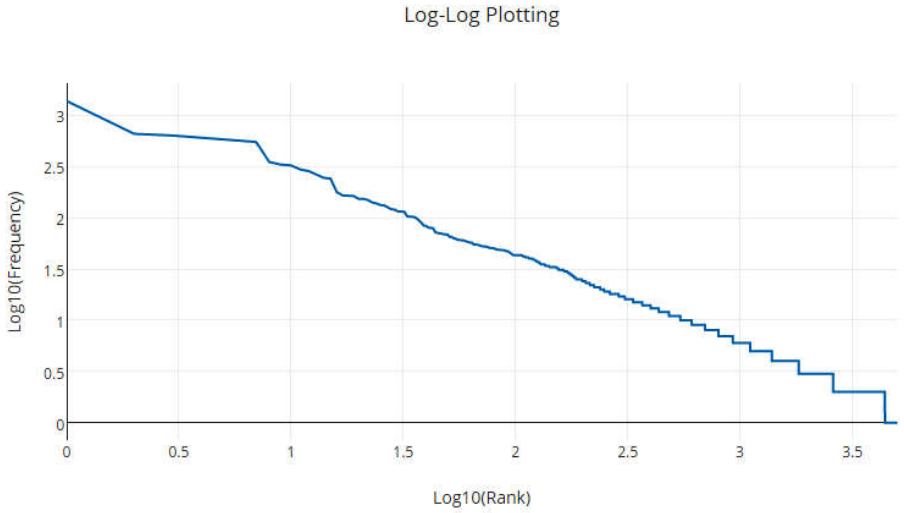

Fig. – (1) Fit to Zipf's for travelling corpus

Based on Figure-1, log-log plotting for the rank and frequency is roughly linear but slop of this graph is -0.87271222 which is little larger than the Zipfian ideal of -1.

Luhn [4] suggested that extremely common and uncommon terms were not useful for indexing**.**

In natural language, very few frequent terms including stop words and many rare terms occurs. By removing 25% of unique terms from tail which is rare terms and almost it's half number of terms from head we got slop of graph is -0.997295926 which is linear fit and closer to -1.

Zipf's law is quite accurate except for head and tail terms which have very high and low rank in corpus. To improve the performance of information retrieval, only medium term frequency can be considered instead of tail and head frequent terms.

The following exemptions are made to carry out this research:

Exemption-1: A number of new words will increase with the size of corpus, but for the very large dataset new words will decrease when corpus size will increase more.
Exemption-2: By removing 25% of unique terms from tail which is rare terms and almost it's half number of terms from head it is possible to get slop of graph closer to -1 and linear fit.

Advantages:

- No need to identify stop words in pre-processing steps of information retrieval.
- Reduce the size of corpus in terms of dimensionality reduction and storage cost.
- Fast retrieval: Information retrieval process will be done on medium frequency term, thus it will retrieve most relevant information.
- Independent of corpus and its size

Disadvantages:

- Difficult to determine rare terms in corpus.
- Sometimes reduce the efficiency of some meaningful statistical analysis

RESULTS AND DISCUSSION

The experiment was conducted for 272 different documents which are prepared from travelling corpus and each document that contained minimum 48000 words. Table-2 presents slop in log-log plotting for randomly selected documents from corpus.

Table-2 log-log plotting slop for randomly selected documents

# Document	Total no of terms	Total no of unique terms	Slop before removing head and tail words	Slop after removing head and tail
1	49800	10700	-0.8827527	-0. 99729592
2	48990	99900	-0.8658020	-0.99994059
3	48210	97401	-0.8738900	-0.99996050
4	50101	11020	-0.8640302	-0.99805033
5	49020	98002	-0.8799802	-0.99796050
6	48320	85072	-0.9065030	-0.99903204
7	49400	95040	-0.8940200	-0.99897890
8	49800	98053	-0.8709403	-0.99909695
9	48710	87340	-0.8904030	-0.99850490
10	49106	90302	-0.8789706	-0.99304050

Based on our experiment we conclude that Zipf's law is quite accurate except only most frequently occurs (head terms) and rare terms (tail terms). For all document approximate we got -0.99976848 which is very closer to the Zipfian ideal of -1.

CONCLUSION AND FUTURE WORK

Zipf's law distribution has been applied to carry out this research for Gujarati information retrieval. Based on Zipf's law distribution only medium terms frequency has considered for further processing because extremely frequent and rare terms will not used for indexing. Zipf's law is quite accurate except for head and tail terms which have very high and low rank in corpus. Thus, Zipf's law has been applied in information retrieval instead of stop words removal pre-processing step. Travelling corpus has used for experiment, log-log plotting for the rank and frequency is roughly linear but slop of this graph is -0.87271222 which is little larger than the Zipfian ideal of -1. But by removing head and tail terms from corpus we got slop of graph is -0.997295926 which is linear fit and closer to -1. This approach could not be applied for semantic relationship between two terms in document. In future we will use Ontology based hybrid approach for information retrieval and document classification for Gujarati language.

REFERENCES

1. Zipf, George K. (1949). Human behavior and the principle of least effort. Cambridge, (Mass.): Addison-Wesley, pp. 573
2. Ridley, D. R. (1982). Zipf's law in transcribed speech. Psychological research, 44(1), 97-103.
3. Basu, B., & Bandyapadhyay, S. (2009). Zipf's law and distribution of population in Indian cities. Indian Journal of Physics, 83(11), 1575-1582.
4. Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. IBM Journal of research and development, 1(4), 309-317.

5. Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future
6. directions. Psychonomic bulletin & review, 21(5), 1112-1130.
7. Chen, Y. (2016). The evolution of Zipf's law indicative of city development. Physica A: Statistical Mechanics and its Applications, 443, 555-567.
8. Rana, M. S. (2015, January). Content analysis and application of Zipf's Law in Computer Science literature. In Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 4th International Symposium on (pp. 223-227). IEEE.
9. Al-Talib, G. A., & Hassan, H. S.(2013). A Study on Analysis of SMS Classification Using TF-IDF Weighting. International Journal of Computer Networks and Communications Security (IJCNCS)**, 1(5),** 189–194.
10. Kumar, M., & Vig, R. (2013). e-Library Content Generation Using WorldNet Tf-Idf Semantics. In Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) (pp. 221-227). Springer Berlin Heidelberg
11. Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. In Intelligent Data Engineering and Automated Learning–IDEAL 2013 (pp. 611-618). Springer Berlin Heidelberg.
12. Sartori, C. (2016, March). A Comparison of Term Weighting Schemes for Text Classification and Sentiment Analysis with a Supervised Variant of tf. idf. In Data Management Technologies and Applications: 4th International Conference, DATA 2015, Colmar, France, July 20-22, 2015, Revised Selected Papers (Vol. 584, p. 39). Springer
13. Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765