ISSN 0005-1055, Automatic Documentation and Mathematical Linguistics, 2015, Vol. 49, No. 3, pp. 83–88. © Allerton Press, Inc., 2015.
Original Russian Text © V.A. Yatsko, 2015, published in Nauchno-Tekhnicheskaya Informatsiya, Seriya 2, 2015, No. 5, pp. 7–18.

Automatic Text Classification Method Based on Zipf's Law

V. A. Yatsko

Katanov Khakass State University, Abakan, Russia
e-mail: viatcheslav-yatsko@rambler.ru

Received February 17, 2015

Abstract—This paper describes a method for automatic text classification based on analysing the deviation of the word distribution from Zipf's law, combined with the zonal data-processing approach. Deviation is understood as the difference between the actual numerical score of a word and its score according to Zipf's law. The proposed method involves the division of input and reference texts into J_0, J_1, and J_2 zones, and the creation of a numerical series using the words that are contained in the J_0 zone. The constructed numerical series shows the difference between the real scores of words and the scores calculated according to Zipf's law. The proposed method can significantly reduce text dimensionality and thus improve the running speed of automatic text classification.

Keywords: Zipf's law, zonal text processing, automatic classification of text documents, efficiency improvement

DOI: 10.3103/S0005105515030048

In our previous work [1], we introduced the concept of linguistic informatics to refer to a discipline that studies the distribution patterns of textual information, and investigates problems, principles, methods, and algorithms of linguistic software and hardware development. Our interpretation differs from the Anglo–American scientific tradition, where applied research on linguistic software development is considered part of *computational linguistics*, whereas the laws and patterns of textual information distribution are studied under *quantitative linguistics* [2]. The proposed concept carries on the Soviet scientific tradition given the fact that the term *informatics* was introduced by the Soviet scientists [3] in order to refer to a discipline that combines both theoretical and applied aspects of research.

Zipf's law, which establishes the relationship between the frequency of any word in the text and its rank [4], is one of the most well–known and widely used laws in the domain of linguistic informatics. The established relationship can be presented in the form of the following equation:

$$F_r \propto 1/r^a, \quad (1)$$

where F is the frequency of a word and r is its rank in the ranked list. The a exponent is approximately equal to 1.

Zipf's law has predictive power: if the frequency and the rank of a given word are known, it is possible to establish the frequencies and ranks of all other words in the text. For example, if the tenth-ranking word occurs in the text 36 times, the frequency of the fifth-ranking word can be calculated as follows: $36 \times 10/5 = 72$.

Traditionally, Zipf's law has been used to reduce the index size and improve the processing speed of data-retrieval systems [5]. It is also used for text stratification [6] and typological classification of languages [7]. In addition, it was found that the distribution of urban population complies with Zipf's law [8].

In this paper, we show how Zipf's law, combined with the method of zonal processing, can be applied for the purpose of automatic classification and author attribution of texts. In this context, we discuss the main problems that are related to the automatic classification of text documents and describe the method of zonal text processing.

The objective of automatic text classification is to identify the class that incorporates a given text based on the analysis of the text content. Text documents serve as the input for the classifier. The output consists in the title of the class to which these documents belong [9]. Automatic text classification is widely used in various linguistic applications and systems, including data-retrieval systems, spam filtering software, and digital libraries, as well as plagiarism detection and author attribution tools. To date, two methods of automatic text classification, so-called dictionary and distant approaches, have emerged. The dictionary approach relies on the development of a reference dictionary for a particular class. Each unit in this dictionary has strong discriminatory power, i.e., the ability

to uniquely identify a given class C and, thus, distinguish it from all other classes. We can talk about binary classification when a given class is compared with only one other class, denoted as $\sim C$. Spam filtering is a typical example of binary classification, as its goal is to classify the input text either as "spam" or "not spam."

Reference dictionaries are constructed using fairly complex metrics, such as chi-square, odds ratio, and information gain [10], which allow one to identify the discriminating power of lexical units. In addition to the reference dictionary designed for a given class, a negative dictionary, which includes units that represent the class $\sim C$, is constructed. If the input (test) document contains a word from the reference dictionary, it is assigned a positive coefficient (e.g., 1), whereas the words that belong to the negative dictionary are assigned negative coefficients (e.g., -1). Next, the sum of positive and negative coefficients is calculated. If this sum exceeds a given threshold, the input document is correlated with the class C. If the threshold is not exceeded, the text is either sorted into the class $\sim C$ (in the case of binary classification) or ignored [11].

The distant approach involves calculating the distance between the input text and the reference text (or its certain model) parameters. Parameters are text units with attributed weights. Vector modelling is one of the most common methods that are applied to calculate the distances between texts. Each text parameter is represented as a point in a multidimensional space. The relationship between the respective parameters of the input text and the reference text is interpreted as the direction (typically from the input to the reference text) and length of a vector, which is calculated as the ratio of weighting scores. The distance between the input and the reference texts is calculated as the sum of absolute differences between the vector values. The shorter is the distance, the greater the probability is that the input text belongs to the class represented by the reference text. Thresholds are also used in the distant approach. If the distance between the input text and the reference text is smaller than a certain threshold level, a decision is made about the attribution of the input text to the class C.

The distant approach significantly differs from the dictionary approach. While the former can use stop words as the parameters, the latter filters them out. Articles, prepositions, pronouns, and conjunctions occur in all kinds of texts regardless of their genre or stylistic features. Differences in the distribution of such words can be used to calculate the distances between texts.

In this paper, we use the distant approach, combined with the method of zonal correlation text analysis, which was described in detail in our previous works [12, 13]. The reference text is compiled out for five novels by Theodore Dreiser (*The Genius, The Financier, The Titan, Sister Carrie*, and *Genie Gerhardt*). These novels were downloaded from the Gutenberg[1] project website. The texts were edited and the details about the project were removed. As a result, a text, which contained 23591 unique words and 1003944 total tokens, was obtained. Statistical data on the distribution of tokens were generated by using AntConc 3.1.3 concordancer[2]. The works of Theodore Dreiser were chosen for the analysis for the following reasons. Unlike other literary styles, fiction is typically characterized by a rich vocabulary. In comparison, academic literature usually applies a standardized terminology that is specific to a particular subject area. Dreiser's works belong to the classical literature, so they contain fewer neologisms, jargon, or rare words, than, for example, science fiction. The selected artworks are associated with the naturalistic direction in the literature[3], which distinguishes them from the works of other authors. This fact allows one to obtain more illustrative results while comparing Dreiser's texts with the works of other writers. The novels by Theodore Dreiser are sufficient in volume, which makes it possible to compose a reference text file with a million common tokens. This size satisfies the criteria of representativeness given that Zipf's law holds on the Brown Corpus, which contains one million tokens [14].

Dreiser's novel *The Stoic* and Charles Dickens's *David Copperfield* were selected as test documents. Dickens's works also belong to the classical literature masterpieces. However, Charles Dickens was a British and not an American writer. He also lived before Theodore Dreiser. In our opinion, this circumstance allows us not only to identify a variety of common words and similar parameters, but also demonstrate the differences in the distribution parameters. The distance (*Ds*) between the Dickens's text (file *Di*) and the reference text that includes the above-mentioned works of Dreiser (*Dr1*) must be greater than the distance between the reference text and the another novel by Dreiser, *The Stoic* (file *Dr2*). Therefore, it is necessary to find the following:

$$\mathrm{Ds}(Dr1, Di) = |P(Dr1 - P(Di)| \quad (2)$$

$$\mathrm{Ds}(Dr1, Dr2) = |P(Dr1 - P(Dr2)| \quad (3)$$

where P is a given parameter. Clearly, $\mathrm{Ds}(Dr1, Di)$ must be greater than $\mathrm{Ds}(Dr1, Dr2)$. It is expected that

[1] https://www.gutenberg.org/. The Gutenberg project promotes digitalization and editing of classical literary works that are exempted from copyright.

[2] http://www.laurenceanthony.net/software.html.

[3] http://www.chitai.kraslib.ru/28.html.

Table 1. The results of the zonal text processin

Text	Zone	Quantitative values	Range of words	Number of words in the zone	Number of common tokens	Number of unique words
Di	S(J0)	253073	1–228	228	365542	14131
	S(J1)	84359	229–2725	2497		
	S(J2)	28110	2726–14131	11406		
*Dr*2	S(J0)	88253	1–278	278	127463	9390
	S(J1)	29413	279–2793	2515		
	S(J2)	9797	2794–9390	6597		
*Dr*1	S(J0)	694910	1–287	287	1003944	23591
	S(J1)	231681	288–3472	3185		
	S(J2)	77353	3473–23591	20119		

using Zipf's Law in combination with the method of zonal correlation analysis will expose the considerable difference between *Ds*(*Dr*1, *Di*) and Ds(*Dr*1, *Dr*2). To verify this hypothesis, the following procedures are performed.

1. Zonal text processing. The method of zonal processing involves dividing each text into three zones: J_0, J_1 and J_2. The J_0 zone includes stop words. Since it has the smallest number of elements with the highest frequency, it can be called a zone of information concentration. The J_1 zone consists of notional words that reflect the content of the text. The J_2 zone contains rarely used words, abbreviations, and neologisms coined by the author. It is the area of the largest information dissemination, as it contains the largest number of elements with the lowest frequency.

The division of texts into the three zones is performed on the basis of the following system of equations:

$$\begin{cases} S(J2) = C/K \\ K = (q^n - 1)/(q - 1) \\ S(J1) = S(J2)q \\ S(J0) = S(J1)q \end{cases}, \quad (4)$$

where C is the sum of numerical series, whose elements are the word frequencies in the text; $S(J_1)$, $S(J_2)$, and $S(J_0)$ are numerical values of the respective zones; $n = 3$ is a constant that is equal to the number of zones; q is the zonal coefficient established empirically, whose value depends on a specific subject area. Based on the earlier analysis of the distribution of stop words, it was shown that $q = 3$ is the optimal value for literary texts [12]. First, we calculate the abstract threshold, which represents the numerical value of each zone, by using the formulas (4). Next, we calculate the real threshold, which is as close as possible to the abstract threshold and equal to the sum of frequencies of words contained in a given zone.

Table 1 presents the results of the zonal processing step. The column "Quantitative values" refers to the real threshold levels.

2. Weighing the words and establishing the parameters. As a result of weighing each word in the text is assigned a probability factor calculated on the basis of the following equation:

$$P = \frac{f(w_{ij})}{\sum_{j=1}^{n} f(w_j)}, \quad (5)$$

where f is the frequency of the word w_i in the j-th text. The coefficients are rounded to seven decimal digits.

The standard deviation of Zipf distribution in the zones J_0 of the three texts is used as the parameter. It is calculated as follows:

$$\sigma(Rx) = \sqrt{Var(Rx)}, \quad (6)$$

where *Var* is the dispersion, and (*Rx*) is the numerical series:

$$R(w_i \ldots w_n) = |P(w_i \ldots w_n) - Z(w_i \ldots w_n)|, \quad (7)$$

where P is the probability factor that is assigned to each word in the zones of J_0, and Z is the word value corresponding to Zipf's law (Zipf distribution). Zipf distribution is calculated by using the following equation:

$$P(w_{ij}) = P(w1_j)/R(w_{ij}), \quad (8)$$

where $P(w1_j)$ is the probability factor of the first-ranking word, and R is the rank of a given word. If $P(w1) = 0.0376728$, and the score of the second-ranking word is $P(w2) = 0.0368959$, Zipf distribution is $Z(w2) = 0.0188364$, and the deviation from this distribution is

Table 2. The distance between the reference text *Dr*1 and the test documents *Dr*2, *Di*, as calculated by two parameter

Parameter	*Dr*2	*Dr*1	*Di*	*Ds*(*Dr*1, *Dr*2)	*Ds*(*Dr*1, *Di*)	Distance difference (%)
Ms	0.0015843	0.0015262	0.00204367	0.0000582	0.0005175	789.70%
σ(*Rx*)	0.0027441	0.0025749	0.0031659	0.0001692	0.0005910	249.21%

Table 3. The intersection of the J_1 zones in the three texts

Area of intersection	Number of words	Five randomly selected words	*P*(*Dr*1)	*P*(*Dr*2)	*P*(*Di*)
A	1818	Returned	0.0003277	0.0004158	
		Our	0.0002341	0.0004080	
		Understand	0.0002988	0.0004080	
		Able	0.0002829	0.0004001	
		Believe	0.0003556	0.0004001	
B	1639	Master	0.0000588		0.0005307
		Child	0.0002839		0.0005170
		Boy	0.0002161		0.0005116
		Cried	0.0000647		0.0005116
		Name	0.0003217		0.0005116

$R(w2) = |0.0368959–0.0188364| = 0.018059$. According to the equation (8), $P(w1) = Z(w1)$.

Another parameter is the average sum of differences in the numerical series Rx:

$$\mathrm{Ms} = \frac{\sum R(w_i \ldots w_n)}{x}, \quad (9)$$

where x is the number of words in *Rx*.

3. Comparing the parameter distribution in the zones J_0 of the three texts and calculating the distances between the texts. The results of the parameter distribution analysis are presented in Table 2. They confirm the earlier formulated hypothesis that stipulates that the use of the proposed method makes it possible to establish that the distance between the zones J_0 of the texts that were written by one author is significantly smaller (by 789.70% and 249.21%) than the distance between the same zones of the texts that were written by different authors.

To confirm the efficient use of the zonal analysis for text classification, an additional correlation analysis is performed for the distribution of words in the zones J_1 of the three texts. As mentioned above, these zones contain notional words that reflect the content of the texts. As expected, the J_1 zones in the works of Dreiser $J_1(Dr1)$ and $J_2(Dr2)$ contain more identical words than the corresponding zones in the reference text of Dreiser $J_1(Dr1)$ and the Dickens text $J_1(Di)$. Accordingly, the parameter sum of these zones must be a large numerical value. To confirm this hypothesis, the following operations are performed.

1. The zones J_1 in the three texts are intersected. The identical words contained in these zones are identified.

$$A = J_1(Dr1) \cap (J_1(Dr2) \quad (10)$$

$$B = J_1(Dr1) \cap (J_1(Di). \quad (11)$$

The intersection was performed using standard functions of MS Excel 2010, such as IFERROR, VLOOKUP, FALSE. Table 3 shows the results of the intersection. As expected, the number of words in *A* is greater than in *B*.

2. Calculating the sum of the word probability values in *A* and *B*. Each word in the area of intersection has two coefficients that are summed up: $\sum P(Dr1)$ and $\sum P(Dr2)$; $\sum P(Dr1)$ and $\sum P(Di)$.

3. Calculating the average probability value $Mp = \sum P(T_i)/x$, where x is the number of words in a given zone, J_1.

4. Calculating the sum of the average probability values.

$$Mp(A) = \frac{\sum P(Dr1)}{x} + \frac{\sum P(Dr2)}{x} = 0.0001228. \quad (12)$$

Table 4. The frequencies of the words that are contained in the zone J_2 of the reference document

Word frequency	Range	Number of words	The percentage of the total number of words	Examples of words
1	15535–23591	8057	34.15%	Ziner zithers zouave
2	12163–15534	3372	14.29%	Sapient sappho sarah
3	10179–12162	1984	8.41%	Inanimate incarnation incense
4	8822–10178	1357	5.75%	Tolerate tom tongues
5	7848–8821	974	4.13%	Tinder tinsel touring
6	7135–7847	713	3.02%	Pitied planet planted
7	6510–7134	625	2.65%	Circular civic clay
8	6059–6509	451	1.91%	Excuses exercised existing
9	5683–6058	376	1.59%	Debt defeated defend
10	5335–5682	348	1.48%	Foul freight frock

$$Mp(B) = \frac{\sum P(Dr1)}{x} + \frac{\sum P(Di)}{x} = 0.0001145. \quad (13)$$

Thus, $Mp(A)$ is by 7.21% higher than $Mp(B)$, which confirms the above hypothesis.

In this paper, we discuss a method for the classification of text documents based on the analysis of Zipf distribution in combination with zonal data processing. This method involves dividing the input documents and the reference text into three zones and comparing the distribution of words in the zones J_0. Four numerical series are designed: the first includes the words and raw frequencies; the second contains the words and probability values that were calculated on the basis of the frequencies; the third includes the words and probability values that were calculated on the basis of Zipf's law; the fourth contains the absolute differences between the values in the third and the second numerical series. Thus, the fourth row includes the numerical values that show the extent to which the distribution of words in the zone J_0 deviates from Zipf's law. Next, we compare the distribution of values in the fourth numerical series for the input texts and the reference text, and calculate their distance from each other. This calculation supports the decision made about the attribution of the input texts to the class represented by the reference text.

One of the major problems that has been highlighted in the research studies on automatic classification of text documents [15] is associated with the large number of such text documents. Text processing may include thousands or even tens of thousands of words, thus, their distribution analysis has a negative impact on the operational speed of a system. The proposed approach makes it possible to significantly reduce text dimensionality by limiting text analysis to zones J_0, which only include a few hundred words. In addition, the suggested approach allows one to simplify the mathematical apparatus by using simple probability values.

It should be noted that the analysis of word distribution deviation from Zipf's Law can only be applied to the zone J_0. Multiple words in the two other zones have the same frequency and an insufficient range of values. Table 4 shows that the words that appear once or twice in the text represent nearly half of all words in the text.

In the zone J_0, the frequency dispersion is much higher, which allows one to explain the fact that the values of Ms and $\sigma(Rx)$ are much higher than those of $Mp(A)$ $Mp(B)$ that were calculated on the basis of the parameter distribution in the zone J_1.

Overall, the obtained results indicate that the proposed method can be applied effectively for automatic text classification.

REFERENCES

1. Yatsko, V.A., Computational linguistics or linguistic informatics? *Autom. Doc. Math. Linguist.*, 2014, vol. 48, no. 3, pp. 149–157.
2. Köhler, R. and Rieger, B.B., Preface, in *Contributions to quantitative linguistics. Proc. 1st Int. Conf. on Quantitative Linguistics*, Dordrecht, 1993, pp. i–ix.
3. Mikhailov, A.I., Chernyi, A.I., and Gilyarevskii, R.S., Informatics is the new name of the theory of scientific information, *Nauchn.-Tekhn. Inform.*, 1966, no. 12, pp. 35–39.
4. Piantadosi, S.T., Zipf's word frequency law in natural language: A critical review and future directions. http://colala.bcs.rochester.edu/papers/piantadosi2014-zipfs.pdf.
5. Manning, C.D., Raghavan, P., and Schutze, H., An Introduction to Information Retrieval. Online Edition, Cambridge (UK), 2009. http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf
6. Altmann, G., Popescu, I.-I., and Zotta, D., Stratification in texts, *Glottometrics*, 2013, no. 25, pp. 85–93.

7. Popescu, I.-I., Mačutek, J., and Altmann, G., *Aspects of Word Frequencies*, Ludenscheid: RAM-Verlag, 2009.
8. Gabaix, X., Zipf's law for cities: An explanation, *Q. J. Econ.*, 1999, vol. 114, no. 3, pp. 739–767.
9. Novovičová, J. and Malik, A., Information-theoretic feature selection algorithms for text classification, *Proc. Int. Joint Conf. on Neural Networks,* Montreal, 2005. http://staff.utia.cas.cz/novovic/files/1483.pdf
10. Nicolosi, N., Feature selection methods for text classification. http://www.cs.rit.edu/~nan2563/feature_selection.pdf
11. Oakes, M.P., Gaizauskas, R., and Fowkes, H., A method based on the chi-square test for document classification, *SIGIR '01 Proc. 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, 2001. http://pers-www.wlv.ac.uk/~in4326/old/2001_Oakes_SIGIR.pdf
12. Yatsko, V.A., The method of zonal text analysis, *V Mire Nauchn. Otkryt.*, 2013, no. 6.1, pp. 166–182.
13. Yatsko, V.A., The method of zonal correlation text analysis, *Autom. Doc. Math. Linguist.*, 2014, vol. 48, no. 5, pp. 259–263.
14. West, M., The mystery of Zipf. http://plus.maths.org/content/mystery-zipf
15. Ahlgren, O., Malo, P., Sinha, A., et al. A dimensionality reduction approach for semantic document classification. http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/SPIM/spim2011_paper6.pdf

Translated by V. Kupriyanova-Ashina