ISBN: 978-981-18-3959-7
2022 the 12th International Workshop on Computer Science and Engineering (WCSE 2022)
doi: 10.18178/wcse.2022.06.037

Research on Topic Concentration of a Text from the Perspective of Readability Research

Yongquan Li [1 +]

[1] College of Chinese Language and Culture, Jinan University, Guangzhou 510610, China

Abstract. In the time of information explosion, how to find the information which we need quickly, high quality and accurately is very pivotal. Scientific and efficient reading is particularly important in the era of information explosion. How to measure readers' reading level and find the text suitable for their reading level is a difficult problem to solve. In order to solve these problems, the primary problem that must be solved is how to scientifically measure the difficulty of a text. Based on the research perspective of readability, this paper attempts to mine and calculate the difficulty of a text from the topic concentration of the text. This paper extracts the topic words from the topic concentration formula in Quantitative Index Text Analyzer (QUITA), calculates the "topic concentration" of each topic word (its value is the "topic concentration" of a single word multiplied by the frequency of the word) according to the "topic degree" formula of Hua Liu, and then sums these values to obtain the final value of "topic concentration". The constructed topic concentration can distinguish the differences of versions and grades to a certain extent. It has the feasibility of automatic extraction by computer and reduces the degree of manual intervention.

Keywords: Chinese, frequency, topic concentration, circulation degree

1. Introduction

Reading is an activity in which readers make use of language symbols to obtain aesthetic experience and knowledge. Let those who seek knowledge learn from it, and let the ignorant become knowledgeable. In 1931, Waples and Tyler published *What adults want to read about* [1]. In 1935, Gray and Leary published *What makes a book readable* [2]. These kinds of research on adult reading shows that many readers lack appropriate reading materials. The impact of these research and the great impact of the great depression in the 1930s prompted the USA government to increase investment in adult education. In the time of information explosion, how to find the information which we need quickly, high quality and accurately is very pivotal. Therefore, scientific and efficient reading is particularly important. In order to improve people's cultural quality, the first thing is to improve their reading ability.

Improving the reading ability will involve a series of problems, such as how to investigate the reading grades of readers, how to measure the difficulty of a text, and how to provide reading materials for readers with specific reading grades. In order to solve these problems, the primary problem that must be solved is how to scientifically measure the difficulty of a text. Studying the proposition of the difficulty of a text, it is one of the focuses of readability research.

2. Linguistic Background

Firstly, the concept of readability is discussed. Generally speaking, readability is generally the difficulty of studying the text. In my doctoral thesis [3], I divide the concept of readability into two levels: broad sense and narrow sense. The research on readability in broad sense mainly includes literature, printing, language and content, diversity, interest, etc., which is similar to the concept of readability that proposed by [4-5], etc. Generally speaking, the factors can be divided into objective factors and subjective factors. Objective factors refer to all objectively existing factors affecting readability according to the true face of a text, including text factors and environmental factors. Subjective factors refer to the readers' own conditions for reading, such as reading motivation, interest, intelligence, language level. The narrow research on readability is based on the

[+] Corresponding author. Tel.: (+86) 17372099793.
E-mail address: lyq333@stu2016.jnu.edu.cn.

texts, carries out the research on the superficial language features of a text, and gradually explores the deeper language features such as text artistry, structure and discourse cohesion, which is similar to the concept of readability that proposed by [6-7], etc.

As the saying goes, “下笔千言, 离题万里 (Write a thousand words and leave the topic thousands of miles away)”. This sentence means that writing an article or speaking is far away from the topic to be talked about and has nothing to do with it. It vividly illustrates the importance of the article around the topic. There is a big gap between the connotation and extension of the word topic in different disciplines, so it is necessary to define the concept of topic used in this paper. This paper understands that topic is the main object of discourse statement, which objectively exists in the world and should belong to the category of conceptual entity, and has the characteristics of discourse, information, consistency and cohesion [8].

Only by making comprehensive use of multi-disciplinary knowledge such as computer science, linguistics and psychology can the topic of a text be scientifically represented. As far as the field of natural language processing technology is concerned, at present, academic circles have proposed a series of statistical models called topic model for the research of topic. They try to find the deep semantic relationships such as polysemy and synonymy behind words from the text, which avoids the defect that the vector space model can’t deeply represent the semantic relationship and separate the relationship between words. According to the difference of mathematical statistical characteristics, topic model can be divided into: latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA), latent Dirichlet allocation (LDA) and hierarchical Dirichlet process (HDP).

In the field of linguistics, [8-10] respectively studied the topic expressiveness of a text from the perspective of subject, predicate and object. The research appropriately revealed the topic expressiveness of different components at different levels of syntax, but the degree of manual intervention is too high, and it is difficult to automatically extract the topic expressiveness parameters.

In the meanwhile, topic concentration is an index to measure the concentration of the main object of an article. If there were reasonable language symbols showing the topic in the title, opening and ending of the article, its topic concentration is higher; If the article was far from the topic, its topic concentration is not high.

3. Experimental Design and Result Analysis

In order to successfully carry out the research on the topic concentration of Chinese, we selected the texts of Chinese in primary school and Chinese in junior middle school from People’s Education Publishing House (rjb), Jiangsu Education Publishing House (sjb) and Language and Culture Press (ywb). We eliminated classical Chinese, ancient vernacular, Chinese phonetic alphabets (We call it *pinyin*) annotation, repeated text, introductions, footnotes, poetry and dramas, and finally we get a small textbook corpus of 984 texts. For text segmentation, we use the natural language processing toolkit of Stanford University (https://nlp.stanford.edu/software/).

3.1. Research on Topic Concentration of QUITA in Chinese Textbooks Corpus

The text analysis software QUITA is an analysis software developed by teachers and students of the Department of linguistics of Palacký University in the Czech Republic, which mainly calculates the relevant characteristics of text frequency. It can calculate 22 indicators such as vocabulary richness (R_1), repetition rate (RR), type symbol ratio (TTR), entropy and topic concentration (TC). It has two indicators for topic: topic concentration (TC) and secondary topic concentration (Secondary TC). They give the degree of text concentration on a topic based on word frequency distribution. It defines the topic concentration formula as follows:

$$TC = 2\sum_{r'=1}^{T}\frac{(h-r')f(r')}{h(h-1)f(1)} \tag{1}$$

Among them, h is the “H-point value”, which is a critical point in the rank frequency distribution of words in the text. The text is arranged in descending order according to the frequency of words, so that each

frequency has a frequency order value to match it. H-point value is the point where the frequency order is equal to the frequency. If the frequency sequence and frequency were not equal, the point where the straight line formed by two adjacent points close to H-point intersects with the straight line y = x can be taken as H-point value. r' is the frequency order of any notional word in front of point h, $f(r')$ is its frequency, and $f(1)$ is the maximum frequency value of a text [11].

Secondary TC calculates the sum of the topic concentration of notional words between H and 2H. The calculation formula is similar to TC, just replace the corresponding value. Only when there are all function words before point H, this makes the value of TC 0. The calculation of sub topic concentration is constructed for this situation. This chapter does not take it into account, because the investigation of the textbook corpus found that only: rjb second grade volume 1 “欢庆(celebration)”, sjb first grade volume 2 “草原的早晨 (morning on the grassland)”, sjb second grade volume 1 “夕阳真美 (beautiful sunset)”, sjb second grade volume 2 “欢乐的泼水节 (happy water sprinkling festival)”, ywb first grade volume 1 “聪明的小白兔 (smart little white rabbit)” and ywb first grade volume 2 “小树谣 (the ballad of little tree)”. The TC of these six articles were calculated as 0. The length of their texts is very short. At this stage, students’ learning task is mainly to learn and understand basic words. Therefore, it is quite normal to calculate the value of 0 with TC formula. However, 0 does not mean that these texts have no topic, but TC formula can’t be calculated well.

According to QUITA’s TC formula, the paper calculated the topic concentration of the 984 textbooks. The following table shows the data.

Table 1: Topic concentration of textbooks corpus calculated by QUITA’s formula

grade	quantity of articles	mean	standard deviation
1	115	0.60449	0.27148518
2	133	0.4163449	0.24277536
3	140	0.3280127	0.16728525
4	131	0.2661453	0.12566076
5	127	0.2800936	0.12496811
6	107	0.2390765	0.09527681
7	88	0.2118358	0.08830337
8	83	0.2048695	0.08756234
9	60	0.2319501	0.08454924

From the data in the table 1, it can be found that the mean and standard deviation of TC are inversely proportional to the grades. The smaller the grade is, the greater the mean and standard deviation are; The higher the grade is, the smaller the mean and standard deviation are. This phenomenon is caused by TC formula. TC formula is derived from the rank frequency distribution point of words in the text -- H-point. Therefore, if a notional word appeared frequently in the text, it obviously would be very important. Its frequency and frequency order will significantly affect the result of TC formula.

According to the statistics of these words before the H-point value, the average occurrence times of topic words in Grade 1-9 are 16.83333, 23.18045, 37.67857, 49.47328, 71.63780, 74.22430, 158.07955, 158.36145 and 247.13333 respectively. Their values are gradually increasing, but their proportion in the text is decreasing. The average number of topic words at all grades in each text is 0.0082205836, 0.0011101387, 0.0006943849, 0.0006170769, 0.0005309859, 0.0004599104, 0.0007599510, 0.0006268836 and 0.0007398824 respectively. The proportion of topic words in grade 1 and 2 is much greater than that in other grades. The reason is that in the early stage of language teaching, there will be a large number of repeated words in the texts, most of which are notional words.

We calculated the Spearman correlation coefficient between TC value and text grades and tested its significance. P value is less than 2.2e-16. Through the significance test, the correlation coefficient is negative: - 0.449612. This proves once again that there is a negative correlation between TC value and text grade. Therefore, using QUITA formula to calculate the topic concentration of a text has a certain statistical significance.

3.2. The Construction of Topic Concentration in Chinese Textbook Corpus

In different grades, there are many differences in the number of words at all grades. Commonly used function words such as “得 (de)” and “的 (de)” will appear in a large number of texts at all grades, while notional words will not. For example, the words “我(I)”, “田野 (field)” and “天空 (sky)” in the lower grades can appear in the higher grade text, while the words “格物致知 (learn from things)” and “任劳任怨 (bear hardship without complaint)” in the higher grade text will not appear in the lower grade text. Because language learning is a step-by-step process, the acquisition of lower grade vocabulary is the basis of higher-grade learning. Therefore, the words of different grades are unevenly distributed in the texts of each grade. Therefore, it is necessary to study the importance of words in different text grades, which is also conducive to the study of readability.

Term Frequency - Inverse Document Frequency (TF-IDF) method is often used to evaluate the importance of a word to a document set or one of the documents in a corpus. TF-IDF is a commonly used weighting technology for information retrieval and data mining. The importance of a word increases in proportion to the number of times it appears in the document, but decreases in inverse proportion to the frequency of its appearance in the corpus. Simply think that the words with low frequency are more important, and the words with high frequency are more useless. When there is a large gap in the number of texts at all grades of the corpus, if the number of texts in some categories was too large, the IDF value of words in the text would be too small; If the quantity of texts in some categories was too small, the IDF value of words in the texts would be too large.

The textbooks corpus contains 984 articles in total. There are many differences in the number of texts between versions and grades. Therefore, if TF-IDF was selected to represent words and texts, it would not appropriate. For this reason, we refer to the new formula of topic degree developed by Liu [12] to correct the defect of TF-IDF.

$$\text{Topic Degree} = \sqrt{\sqrt{\sum_j \left(P_{ij} - \overline{P_i}\right)^2} \Big/ \sum_j P_{ij}} \times [log(N(w_i)/N]^2 \tag{2}$$

Based on the comprehensive consideration of the domain heterogeneity of words and the adjacent domain, the formula corrects the important disadvantage of TF-IDF representing words. For example, in the textbook corpus, the topic degree of “我 (I)” is 7.566321541579605, “的 (de)” is 0.27819566830483267, “自然 (nature)” is 6.293751792567146. The frequently used function words have low topic expressiveness, while nouns have high topic expressiveness. Obviously, the topic degree here is the importance of a word to the category of training corpus. The higher the topic degree, the more it can express the characteristics of the category.

The concept of topic in formula (2) is different from the topic in TC and Secondary TC of QUITA. The topic of QUITA is the object that the author should focus on when writing or speaking, which can’t be function words, and the TC and Secondary TC depend on word frequency. The topic in formula (2) is a category, which clarifies the contribution of words to distinguishing categories. The former is internal to a text, while the latter focuses on highlighting the external contribution of words. If we could combine the two, we could not only investigate the topic concentration within a text, but also highlight the contribution of each topic word to the readability grade classification. The following section study the topic concentration from the combination of formula (1) and formula (2).

Firstly, this paper extracted the topic words with TC formula, then calculate the topic degree of each topic word according to formula (2) (its value is the topic degree of a single word is multiplied by the frequency of the word), and then sum these values to obtain the final topic concentration result. The topic calculation of the combination of the two formulas hopes to scientifically construct the calculation method of topic concentration of teaching material corpus. The calculation results of this research method are summarized in Table 2. And we drew the mean value of topic concentration in Figure 1.

Table 2: Topic concentration of textbook corpus calculated by QUITA's united formulas

grade	number of articles	mean	standard deviation
1	115	124.1286	96.21474
2	133	159.4116	121.1538
3	140	230.4675	165.8794
4	131	233.5857	190.2756
5	127	385.9534	312.7656
6	107	313.9831	242.0316
7	88	436.6825	425.5319
8	83	498.6195	493.5012
9	60	675.8646	733.0662

Fig. 1: The mean value of topic concentration.

The figure above shows that the value of topic concentration increases with the increase of grades, and the two values are in positive proportion. The value of grade 1-6 is less than that of grade 7-9. In grade 1-3 of the lower grades, the growth rate of topic concentration is not large, and its value does not exceed 300. In the senior of grade 7-9, the increasing trend of topic concentration is more and more obvious, and the numerical gap between the three versions of teaching materials is increasing. Among them, the value of ywb corpus has been higher than that of sjb and rjb since the fourth grade. Therefore, from the figure, the constructed topic concentration can distinguish the differences of versions and grades to a certain extent. Calculate the Spearman correlation coefficient between the topic concentration value of each text and the text grade, and its value is 0.422998. P value is less than 2.2e-16. Through significance test, it shows that there is a certain degree of positive correlation between them.

Secondly, in my last paper Li [3] proposed the evaluation formula of Chinese readability by means of stepwise regression:

Grade Y = - 0.0017036 * quantity of words in the draft + 0.0205674 * quantity of paragraphs + 0.118785 5 * average quantity of sentences in paragraphs - 0.1939400 * average quantity of characters in paragraphs + 8.3354991 * The ratio of the quantity of word species to the quantity of words + 1.8289492 * entropy (words) + 0.0197803 * quantity of word with single character - 0.0286672 * quantity of grade 1 words - 0.0078154 * quantity of words not in the draft + 0.2137848 * average punctuation spacing + 2.4467443 * cube root of total characters + 0.1017426 * quantity of conjunctions + 0.0904817 * average size of phrase structure grammar tree - 0.0010551 * topic concentration - 8.8872316 (3)

In the formula, the draft means *Common vocabulary of compulsory education (Draft)* written by the Language and Text Information Management Department of the Ministry of education of China. And the topic concentration calculated directly from QUITA formula does not appear in the formula, indicating that it does not occupy an important position in the Chinese readability formula. The feature of topic concentration constructed by the combination of QUITA and Liu [12] appears in the formula, which is a very significant

feature in statistical significance. Therefore, it is an important parameter to calculate Chinese readability. This also shows that the topic concentration we construct is scientific and reasonable.

4. Conclusions

In the time of information explosion, how to find the information which we need quickly, high quality and accurately is very pivotal. Scientific and efficient reading is particularly important in the era of information explosion. How to measure readers' reading level and find the text suitable for their reading level is a very difficult problem to solve. In order to solve these problems, the primary problem that must be solved is how to scientifically measure the difficulty of the text. Based on the research perspective of readability, the paper attempts to mine and calculate the difficulty of the text from the topic concentration of a text. This paper extracts the topic words from the TC formula in QUITA, calculates the topic concentration of each topic word according to the topic degree formula of Hua Liu, and then sums these values to obtain the final value of topic concentration. The constructed topic concentration can distinguish the differences of versions and grades to a certain extent. It has the feasibility of automatic extraction by computer and reduces the degree of manual intervention.

5. References

[1] D. Waples, R. Tyler. *What adults want to read about*. The University of Chicago Press, Chicago, USA, 1931.

[2] W. Gray, B. Leary. *What makes a book readable*. The University of Chicago Press, Chicago, USA, 1935.

[3] Y. Li. *An Empirical Study on the Readability of Chinese Text Based on Chinese Textbook Corpus*. Ph.D. Dissertation. Jinan University, Guangzhou, China, 2020. (in Chinese)

[4] Y. Song, R. Chen, Y. Li, et al. 2013. Investigating Chinese Text Readability: Linguistic Features, Modeling, and Validation. *Chinese Journal of Psychology*, 2013, 55(01): 75-106. https://doi.org/10.6129/CJP.20120621 (in Chinese)

[5] R. Chen, X. Cai, Y. Song, Y. Li. 2015. The Development of a Text Leveling Framework. *Journal of Research in Education Sciences*, 2015, (01): 1-32. (in Chinese)

[6] L. Wang. Some Concepts of Readability Formula and Relevant Research Paradigm as well as the Research Tasks of Formula in TCFL. *Language Teaching and Linguistic Studies*, 2008, (06): 46-53. (in Chinese)

[7] G. Sun. 2015. Research on readability prediction methods Based on linear regression for Chinese documents. MA. Dissertation. Nanjing University, Nanjing, China, 2015. (in Chinese)

[8] J. Gao, W. Zhang, K. Zhang. A Study on the Expressiveness of the Theme from the Perspective of the Predicate. *Applied Linguistics*. 2018, (01): 71-80. (in Chinese) https://doi.org/10.16499/j.cnki.1003-5397.2018.01.010

[9] J. Zhou, Y. Luo, B. Chen, A Study on the Expressiveness of the Theme from the Perspective of the Subject. *Applied Linguistics*. 2018, (01): 61-70. (in Chinese) https://doi.org/10.16499/j.cnki.1003-5397.2018.01.009

[10] Y. Yu, Y. Tong, Y. Wen, X. Liu, A Study on the Expressiveness of the Theme from the Perspective of the Object. *Applied Linguistics*. 2018, (01): 61-70. (in Chinese) https://doi.org/10.16499/j.cnki.1003-5397.2018.01.012

[11] H, Liu. *An Introduction to Quantitative Linguistics*. The Commercial Press, Beijing, China, 2017, pp.136.

[12] H. Liu. *Word calculation and Application*. Jinan University Press, Guangzhou, China, 2010. (in Chinese)