Article

Construction and analysis of the word network based on the Random Reading Frame (RRF) method

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 9 August 2018; *Accepted 15 January 2021*; *Published 1 September 2021*
(cc) BY

Abstract
In present study, a method was developed to construct and analyze the word network. The core of the method is Random Reading Frame (RRF) method. First, download or collect word files (in various formats, e.g., pdf, txt, doc, docx, rtf, html, etc.) from internet or local machine in terms of the concerned topics. All files were then combined in a final text file. Excepting for splitting words and stop words, all words were arranged in a word vector following their orders in the combined text file. In the RRF method, for a given pair of unique words (x, y), x, $y \in \{u_1,u_2,\ldots,u_m\}$, a reading frame with randomly changeable width is randomly placed on the vector to count the respective number of the two words in the frame. Randomly repeating the procedure p times, the paired numbers are thus achieved: (x_1, y_1), (x_2, y_2), …, (x_p, y_p). In such a way, the paired numbers for all pairs of unique words are achieved. Thereafter, for a given pair of unique words (x, y), Pearson correlation and Pearson partial correlation, Spearman rank correlation, or point correlation is used to calculate their correlation value according to their paired numbers (x_1, y_1), (x_2, y_2), …, (x_p, y_p), and the statistically significance can be determined by t-test (Pearson correlation, Pearson partial correlation, Spearman rank correlation) or χ^2-test (point correlation). In such a way, all statistically significant word pairs are achieved in terms of the correlation measure chosen by user. Finally, the word network, in terms of the correlation measure chosen, can be constructed based on these word pairs, and no links between statistically insignificant word pairs. Network analysis is conducted for the word network constructed from significant between-word positive correlations among all unique words. Word centrality measures, word tree, word chains, word modules, etc., can be calculated in the method. The Matlab software, wordNetwork for the method was given also.

Keywords word association; association rules; correlation measures; Random Reading Frame; network construction; network analysis; algorithm; text mining.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Word statistics in terms of internet resources and multiple files is of significant in analyzing word association, word clustering (Qi et al., 2018), etc. Furthermore, the construction of a word network based on internet

community. Selforganizology, 1(2): 89-129

Zhang WJ, Wang R, Zhang DL, et al. 2014. Interspecific associations of weed species around rice fields in Pearl River Delta, China: A regional survey. Selforganizology, 1(3-4): 143-205

Zhang WJ, Jiang LQ, Chen WJ. 2014. Effect of parasitism on food webs: Topological analysis and goodness test of cascade model. Network Biology, 4(4): 170-178

Zhang WJ. 2015. Calculation and statistic test of partial correlation of general correlation measures. Selforganizology, 2(4): 65-77

Zhang WJ. 2016a. A Matlab program for finding shortest paths in the network: Application in the tumor pathway. Network Pharmacology, 1(1): 42-53

Zhang WJ. 2016b. A method for identifying hierarchical sub-networks / modules and weighting network links based on their similarity in sub-network / module affiliation. Network Pharmacology, 1(2): 54-65

Zhang WJ. 2016c. Finding trees in the network: Some Matlab programs and application in tumor pathways. Network Pharmacology, 1(2): 66-73

Zhang WJ. 2016d. Selforganizology: The Science of Self-Organization. World Scientific, Singapore

Zhang WJ. 2017a. Finding the shortest tree in the network: A Matlab program and application in tumor pathway. Network Pharmacology, 2(1): 13-16

Zhang WJ. 2017b. Network pharmacology of medicinal attributes and functions of Chinese herbal medicines: (II) Relational networks and pharmacological mechanisms of medicinal attributes and functions of Chinese herbal medicines. Network Pharmacology, 2(2): 38-66

Zhang WJ. 2017c. Network pharmacology of medicinal attributes and functions of Chinese herbal medicines: (IV) Classification and network analysis of medicinal functions of Chinese herbal medicines. Network Pharmacology, 2(3): 82-104

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific, London, UK

Zhang GL, Zhang WJ. 2019. Protein–protein interaction network analysis of insecticide resistance molecular mechanism in Drosophila melanogaster. Archives of Insect Biochemistry and Physiology, 100(1): e21523

Zhang WJ, Li X. 2015a. General correlation and partial correlation analysis in finding interactions: with Spearman rank correlation and proportion correlation as correlation measures. Network Biology, 5(4): 163-168

Zhang WJ, Li X. 2015b. Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed. Selforganizology, 2(3): 39-45

Zhang WJ, Qi YH. 2020. Matlab algorithm to generate adjacency matrix from connection pairs that nodes are represented by strings. Selforganizology, 7(3-4): 8-14

Zhang WJ, Zhan CY. 2011. An algorithm for calculation of degree distribution and detection of network type: with application in food webs. Network Biology, 1(3-4): 159-170