Physica A 468 (2017) 802–808

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Spectra of English evolving word co-occurrence networks

Wei Liang

School of Mathematics and Information Science, Henan Polytechnic University, Jiaozuo, Henan 454000, China

HIGHLIGHTS

- A total of 205 English word co-occurrence networks' spectra are computed.
- The largest eigenvalue increases as $\propto N^{0.66}$ and $\propto E^{0.54}$ respectively.
- The number of different eigenvalues increases as $\propto N^{0.58}$ and $\propto E^{0.47}$, respectively.
- The single network's spectral distribution can be fitted by two segments.
- An "M" shape appears in each of the large networks' spectral densities.

ARTICLE INFO

Article history:
Received 20 January 2016
Received in revised form 16 May 2016
Available online 23 November 2016

Keywords:
English language
Word co-occurrence network
Spectra
Adjacency matrix

ABSTRACT

Spectral analysis is a powerful tool that provides global measures of the network properties. In this paper, 200 English articles are collected. A word co-occurrence network is constructed from each single article (denoted by single network). Furthermore, 5 large English word co-occurrence networks are constructed (denoted by large network). Spectra of their adjacency matrices are computed. The largest eigenvalue, λ_1, depends on the network size N and the number of edges E as $\lambda_1 \propto N^{0.66}$ and $\lambda_1 \propto E^{0.54}$, respectively. The number of different eigenvalues, N_λ, increase in the manner of $N_\lambda \propto N^{0.58}$ and $N_\lambda \propto E^{0.47}$. The middle part of the spectral distribution can be fitted by a line with slope -0.01 in each of the large networks, whereas two segments with the same slope -0.03 for $0 \ll N < 260$ and -0.02 for $260 < N < 2800$ are needed for the single networks. An "M"-shape distribution appears in each of the spectral densities of the large networks. These and other results can provide useful insight into the structural properties of English linguistic networks.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

There are at least 6800 different languages in the world [1]. The English language is one of the mostly spoken one, which is a complex system emerged from a long-time evolution [2]. Most of the existing studies on linguistic networks, for example co-occurrence and syntax frameworks [3,4], capture their static structural properties, such as degree distributions and clustering coefficients. In contrast, spectral analysis is a powerful tool, that provides global measures of the network properties [5]. It refers to the systematic study of eigenvalues and eigenvectors of the adjacency matrix of a network [6].

Although spectral properties have not been widely investigated for linguistic networks, there are some reports in the literature. Eigenvectors of the Laplace matrices are analyzed to obtain a two-dimensional visualization of network models for English and French languages [7]. Spectral methods can cluster words of the same class in a syntactic dependency network [8]. Spectral densities of the word co-occurrence networks for seven different languages are studied, and it is

E-mail address: wliang@hpu.edu.cn.

http://dx.doi.org/10.1016/j.physa.2016.11.096
0378-4371/© 2016 Elsevier B.V. All rights reserved.

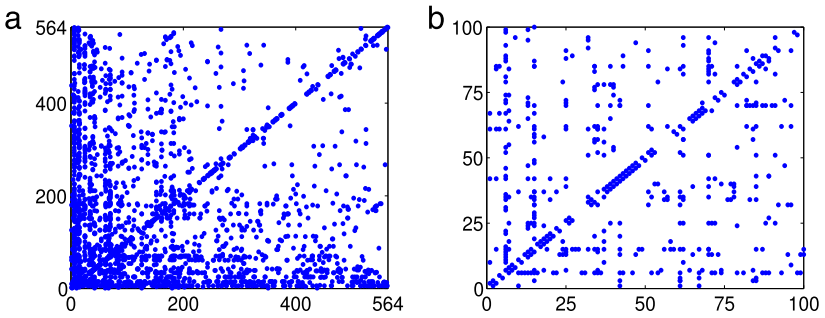

Fig. 1. The adjacency matrix of a network with $N = 564$ and $E = 1178$. (b) Zoom-in the beginning part of (a). There is a dot in position (i, j) if and only if there is an edge between nodes i and j.

found that they have triangle-like shapes [9]. We investigated the spectral densities of 606 Chinese character and word co-occurrence networks and 404 English word co-occurrence networks, constructed from Chinese and English poems, respectively [10,11]. We found that "M"-shape eigenvalue distributions appear in the spectral densities of 1007 networks, while the other 3 spectral densities are triangle-like shapes, which are similar to that of the BA network [10,11]. Nevertheless, our works are just simple study the spectral densities of the Chinese and English poem networks. Recently, we studied the spectra of the networks constructed from other Chinese literary genres [12]. What are spectra properties of English networks constructed from other literary genres? Is the spectral behavior of the English linguistic topology consistent over time? What are commonalities and differences in the Chinese and English languages from a network perspective? Can insightful conclusions be made by studying such networks? The present paper attempts to address such interesting questions.

The present work is to in-depth study spectral properties of English evolving word co-occurrence networks in the framework of complex networks. This study, in turn, provides us with interesting insights into the nature of English language. Here, a total of 205 English word co-occurrence networks are constructed. Their spectra, especially the largest eigenvalues, the second largest eigenvalues, the smallest eigenvalues, the number of different eigenvalues, the spectral distributions, and the spectral densities of the above-mentioned networks are studied.

The rest of the paper is laid out as follows. Section 2 introduces the construction of networks and some concepts involved in spectral analysis. Section 3 contains mathematical analysis and numerical simulations on the spectra: the largest eigenvalue in Section 3.1, the spectral distribution in Section 3.2, and the spectral density in Section 3.3. Finally, some conclusions are drawn from the above investigations.

2. Preliminary

My analysis was based on 200 English articles including four literary genres: essays, novels, popular science articles, and news reports. 50 articles in each genre are randomly selected from Ref. [13], respectively. Therefore, most of the articles are originally written in English, and no changes are made to the texts. The length of an article is the number of words, including repeated words. The lengths fall in the ranges 454–4542 (essays), 610–13 260 (novels), 459–2980 (popular science articles), and 450–1941 (news), and their average lengths are 1142, 5224, 961, and 748. In a word co-occurrence network based on a given article, nodes denote words; two words are linked by an edge if they occur consecutively within at least one sentence. In this manner, a word co-occurrence network is constructed from each single article (denoted by S-network). Furthermore, 25, 50, 100, 150, and 200 articles are put together, respectively. The 25 articles are randomly selected from above 50 essays, the 50 articles are the 50 essays, the 100 articles contain the 50 essays and the 50 popular science articles, the 150 articles include the 50 essays, the 50 popular science articles and the 50 news reports, and the 200 articles are all the texts. Therefore, 5 large word co-occurrence networks are also constructed, denoted by Net25, Net50, Net100, Net150, and Net200, respectively. These networks can be viewed as English evolving networks.

For an undirected network consisting of N nodes (i.e., network size is N), its *adjacency matrix* A is defined by $(a_{ij})_{N\times N}$, where $a_{ij} = 1$ if nodes i and j are connected via an edge and $a_{ij} = 0$ otherwise. An adjacency matrix is illustrated by Fig. 1. Fig. 1(a) shows the adjacency matrix of a network with 564 nodes and 1178 edges (i.e., the number of edges, E, is 1178), which is constructed from a novel named "A Bad Business" written by Anton Chekhov, and Fig. 1(b) enlarges the beginning part of Fig. 1(a), where a dot in position (i, j) corresponds to an edge between nodes i and j. It can be seen from Fig. 1(b) that two intermittently parallel lines lie on the diagonal (i.e., $a_{i,i+1} = a_{i+1,i} = 1$ for most of the nodes i of a network), due to the fact that the constructed networks are co-occurrence; that is, two words are linked by an edge if they occur consecutively within at least one sentence in an article.

λ is an *eigenvalue* of A if there is a nonzero vector x such that $Ax = \lambda x$. The *spectrum* of a network is the set of all the eigenvalues of its adjacency matrix.

Download English Version:

https://daneshyari.com/en/article/5103535

Download Persian Version:

https://daneshyari.com/article/5103535

Daneshyari.com