Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 550 – 556

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Motifs and motif generalization in Chinese Word Networks

Jianyu Li[a,1], Feng Xiao[b], Jie Zhou[b], Zhanxin Yang[a]

[a]*Engineering center of Digital Audio and Video, Communication University of China, Beijing, China 100024*
[b]*Dept. of Automation, Tsinghua University, Beijing, China 100084*

Abstract

The most significant semantic unit of Chinese language is words composed of individual characters. This compositional structure produces great variability and representability compared to individual characters, which is quite distinct from other languages. In this paper we utilized complex networks to model the composition of words from characters. We focus on network motifs, the local pattern which appears more often in a statistically significant sense. Network motifs describe the most significant connection pattern between these nodes. We investigated their functions and semantical relationship. We also investigated different generalizations of network motifs and analyzed the larger pattern in the complex networks. As the word network is quite huge and the motif detection is very slow when motifs are much larger, for larger pattern in the network we used topology generalization of simple motifs rather than carry out a thorough motif detection task. The results on motifs and motif generalization in this paper not only offer us a big picture how Chinese words are formed, but also support the conclusion that motifs play a very important role in research of complex systems.

Keywords: complex networks; motif; motif generalization

1. Introduction

Complex systems are ubiquitous in nature surrounding us. Examples of complex systems include ecosystems, financial markets, neural system, road traffic and the Internet, and even entire human societies. Such systems contain a number of elements which interact with each other and function as a whole, thus exhibiting some kinds of similar structural features, and imply some kinds of function-related similarities among them. Therefore detecting the regular patterns and their interaction, and analyzing their function is very important to explore and understand complex systems.

Complex networks are an abstraction of complex systems based on small world property and scale free property. Recently, to describe the complex interactions, many local patterns are also studied, like community, hierarchy and especially motif structures [1, 2, 3]. The concept of "network motifs" was first proposed by Uri Alon's group [1]. Network Motifs are defined as patterns of interconnections that occur in many different parts of a network at frequencies much higher than those found in randomized networks. Recent work on network motifs includes the development

Email addresses: lijianyu@tsinghua.edu.cn (Jianyu Li), xiaof99@mails.tsinghua.edu.cn (Feng Xiao), jzhou@tsinghua.edu.cn (Jie Zhou), yangzx@cuc.edu.cn (Zhanxin Yang)
[1]Corresponding author

doi:10.1016/j.procs.2012.04.059

of efficient methods of motif identification and understanding the distribution of motifs imposed by underlying network geometry [4, 5, 6]. Certain motifs exhibiting dynamical behavior have been identified as essential ingredients of specific biological processes.

The research in natural system like biology, ecology, languages and brain greatly enhanced the understanding of these complex systems [7]. Some human based systems such as software and electric networks are also found to comply with the complex network property [8, 9]. Some researchers propose that English and Chinese languages are weighted complex networks [10].

Compared to English, Chinese has two levels of compositions, from character to words and from words to phrases. However, the gap between words and phrases are quite vague. In this paper we will use *words* to denote commonly used words and phrases and select a generally used Chinese words database as the source of analysis. The quantity of Chinese characters is small compared to English words, daily usage requires only 4,000 characters, and these characters can form at least 100,000 words and phrases. So the compositional structure from characters to words is very important in Chinese language analysis. In this paper we used motifs and motif generalization as a medium to understand the formation and function of words and phrases. This research will promote the understanding of complex system as well as linguistics.

The paper is organized as follows: in section 2 we will introduce the data source and the construction of the word networks; in section 3 we will check the network comply to the property of complex network and introduce the procedure of motif detection; in section 4 we will analyze motif types, frequency and distribution; in section 5 we will analyze the topological generalization of detected motifs; In section 6 we will make a conclusion and discuss some idea on future work.

2. Data and Construction of the network

The purpose of the research is to study the formation and organization of Chinese words. The data were mainly collected from several popular middle-sized Chinese dictionaries such as Modern Chinese Dictionary [11], Contemporary Chinese Dictionary [12] and Xinhua Dictionary [13], which all contain over 50,000 entries including characters, words, phrases, colloquialisms and idioms. More specific or comprehensive dictionaries are not considered, as they contain a great deal of special nouns and classical characters which are unrelated to the analysis of modern Chinese.

Our data set contains 72,923 two-character words, 11,581 three-character words and 28,533 four-character words. In this paper we will only carry out analysis on two character words. There are several reasons for this. First, longer words will introduce difficulty into the definition of adjacency. In fact, we postulate that there are hierarchical structures in longer words. So the model of the networks will be much more complicated and we remain it to be future work. Secondly, longer words are more probable to be phrases rather than words. As we explained earlier, it is not able to distinct phrases from words due to the ambiguity in semantics and context. However, it would be better to exclude those are very likely to affect our analysis. We also excluded those words composed of the same characters, and after these preprocessing, our data set contains 72,217 two-character words. And we construct a directed word network based on this data set.

The directed word network is constructed in the following ways: (1) Each node of the network denotes a single character; (2) connections are established between two characters if they form a word (see Fig. 1). For example, if AB is a phrase which consists of character A and character B in the order of A and B, there is a directed edge from A to B.

3. Network Motif Detection

Before we carry out the motif detection work, we will first check that this network comply with the properties of common complex network. And a preliminary experiment shows the networks display high clustering and short averaged path, and their degree follows power law distribution. In summary, the networks are small world and scale free (The power law exponent, clustering coefficient and the averaged path length are 3.32, 0.4548 and 3.04). So it is reasonable to model this network as a complex network.

Motifs of length n will be detected in the following procedure: For each subgraph pattern count all n-node subgraphs that conform to this pattern in the real network, then compare those counts against randomly generated networks with the same $n-1$th subgraph connection with the real network. The detailed procedure for generating random