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

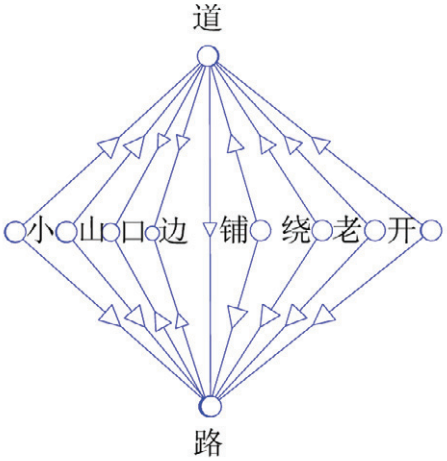


Figure 1: Part of the complex networks

MotifID	38	46	102	108	110	238	all motifs
Z_{score}	45.46	34.29	18.57	32.45	27.83	14.74	
characters	3,307	1,877	1940	2,073	1,047	263	3,361
phrases	59,678	21,704	22,806	23,162	8,110	1,108	65,933
occurrences	171,433	14,080	16,166	15,664	4,851	274	222,468

Table 1: Z_{score} of each motif and the number of characters and words covered by each motif.

networks are discussed in the accompanied material of [1] and in [5]. Then, if certain n-node subgraph patterns are significantly more prevalent than in the random case, these are considered motifs of the network.

The quantity we used to measure a subgraph pattern to be a motif is Z_{score}. Z_{score} is a general variable to measure the statistical significance of certain quantity, in our case the number of occurrence of certain subgraph. It indicates how many standard deviations a quantity is above the mean value. So to calculate the Z_{score} a comparison with the corresponding randomized networks is necessary. Generally speaking, the higher Z_{score} a motif has, the more significant for it to present typical characteristics in a network. For a given network N, The Z_{score} is defined as in equation 1. Generally speaking, if Z_{score} higher than 2, we may treat this pattern is statistically significant and define it as a motif. In [14] the author proposes to use normalized Z_{score} to measure the importance of motifs.

$$Z_{score} = \frac{N_{real} - N_{rand}}{std(N_{rand})}, \tag{1}$$

Several software packages for detecting motifs exist. We used both $Mfinder$ and $Manmod$ to detect the three nodes network motifs as they both provide useful information. The difference is that $Manmod$ is much more efficient and robust than $Mfinder$ for the sampling scheme it introduced [15].

4. Motif Types, Frequency, Distribution

Table 1 described the distribution of three node network motifs, and the number of characters and words which they covered. Fig. 4 illustrates these motifs. From the table we can see that motif with ID 38 appears most frequently among the motifs and has the highest Z_{score}. From the illustration we can see that motif 38 has feed-forward loop structure (FFL for short). Also we can generalize that most motifs are simple generalization of FFL structure. For instance, motif 108, 46 and 102 are FFL structure with an extension of an edge. Motif 110 is the extension of two edges, while motif 238 is three edges. Triad 98 is not a motif as its Z_{score} is 1.87, which means it is not statistically significant compared with random graphs. From the table we also found that motifs covered most of the characters (nodes) and words (phrases), they served as the basic building blocks of the network.

Compared with other types of networks, Chinese words display some distinct organization principle. In [2] word-adjacency networks were studied in which each node represents a word and a directed connection occurs when one word directly follows the other in the text. Five languages including English, French, Spanish, and Japanese were analyzed. The author argues that words belong to categories and a word from one category tends to be followed by one from a different category. So the fully connected or loop structure are not likely to be present in the word-adjacency networks.

Our network is different from the word-adjacency networks. In our networks one edge represents a word or phrase, that is, conventionally fixed. But in the word-adjacency networks, edges may exist between non-phrases. Also the word-adjacency networks are extracted from texts, so the distribution is weighted by the occurrence of the words in texts. And the most important distinction is that they are constructing a phrase or sentence network, and we are constructing a word network. So the level of abstraction is different. Hence the motif structure and distribution may display a huge difference.

We also carried out a four node motif detection task. For four node motifs, we used the *Fanmod* program [5] as it is more efficient. For directional networks there are 199 types of subgraphs patterns. And we checked the “bifan” motifs which are found to be quite significant in biological and electronic circuits networks [1]. The frequency in the real network is 0.273% while in random network this frequency is 0.28736%, result in a Z_{score} of -1.5603. So this “bifan” subgraph pattern is not motif in our network. We think this is a difference with networks studied in [1]. Fig.3 illustrates the motifs with both high frequency and high Z_{score}. We found that in our network, the original motif finding algorithm is biased towards fully connected patterns. For instance, all these motifs can be seen as an extension of less than two edges from the three node motif with ID 238. This may result from that Chinese language is quite densely connected compared to biology and electronic networks [1]. And we think this is the cumulative effort to express more idea with commonly used characters. For instance, when new Chinese words are coined from western science, not very many new characters are created while the words for whole science concept are stacked into this network. We think it would be interesting to study the evolution of the word network of Chinese language in the future.

5. Motif Generalizations

In the network motif detection task, it is rather slow to detect motifs larger than four nodes. And the combination of the motifs is good demonstration of the importance of the pattern in the networks. So we carried out analysis based on the topological generalization of network motifs [9]. We focused on the generalization of FFL motif as it is both significant in frequency and in Z_{score}. There are three kinds of motifs generalization, each corresponds to a role in the motif (see Fig.5). We carried out the experiment in the following way. First we list all those nodes and edges covered by FFL motif. Then we will search the neighboring instances of each appearance of FFL motif. Two appearances are neighbors if they differ in only one node. We will store all these occurrences of generalizations. The capacity of the generalizations is measured by the appearances a generalization which comprises (see Fig.5).

In our experiment, we recorded those generalizations which comprise of the most appearances of certain motif. We found that for the multi input generalization, the biggest generalizations are (We use X, Y, Z to denote indefinite words) “X出了”[2], “X出入”, and “X到了”, which roughly corresponds to “X, out, done”, “X, (get) out, into”, “X, to, done”. There are 83 characters can act as “X” in “X出了”, 78 characters for “X出入”, and 78 character for “X到了”. Note that there may be intersections between these 3 generalizations. For example, “写” appears in all the 3 generalizations as instance of “X”. “写出” means write down or write out, “写了” means written, “写入” means write into, “写到” means write to something or write to someplace. The character “了” is an auxiliary character for expressing that something is done, work finished or goal reached. And the antonym and auxiliary characters appear very frequently in the biggest generalizations. For the multi-y generalization, the biggest generalizations are “不Y了”, “小Y子”, which roughly corresponds to “no, Y, done”, “small, Y, child”. In this generalization we can see that the negatives and auxiliary characters frequently emerge. For the multi-output generalization, the biggest generalizations

[2] There are Simplified Chinese Characters in this paper, better viewed with Adobe Reader Simplified Chinese Addon [16].

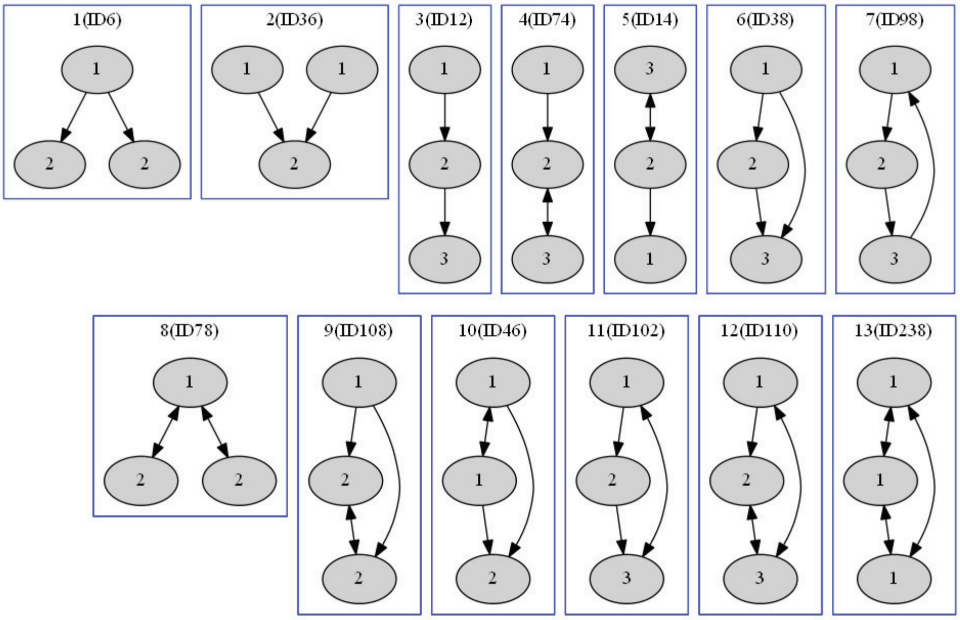
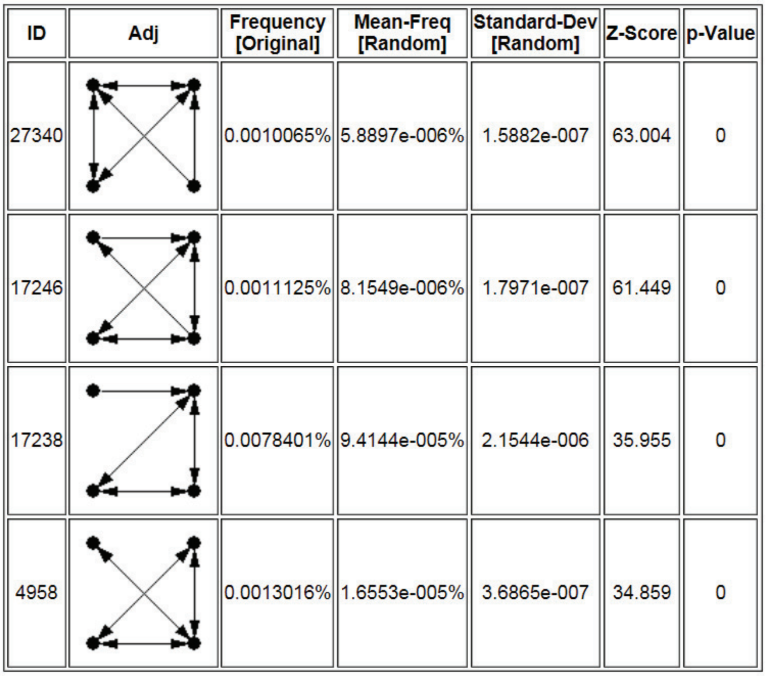


Figure 2: Three node motifs, and the motif ID is the same with [14] and noted in brackets. The number in the circle denotes its role in the motif [9]. Note that not all these triads are motifs, only those with the Z_{score} higher than 2 are motifs.



ID	Adj	Frequency [Original]	Mean-Freq [Random]	Standard-Dev [Random]	Z-Score	p-Value
27340		0.0010065%	5.8897e-006%	1.5882e-007	63.004	0
17246		0.0011125%	8.1549e-006%	1.7971e-007	61.449	0
17238		0.0078401%	9.4144e-005%	2.1544e-006	35.955	0
4958		0.0013016%	1.6553e-005%	3.6865e-007	34.859	0

Figure 3: Four node motifs, extracted from Fanmod output, note that we only displayed the most representational motifs

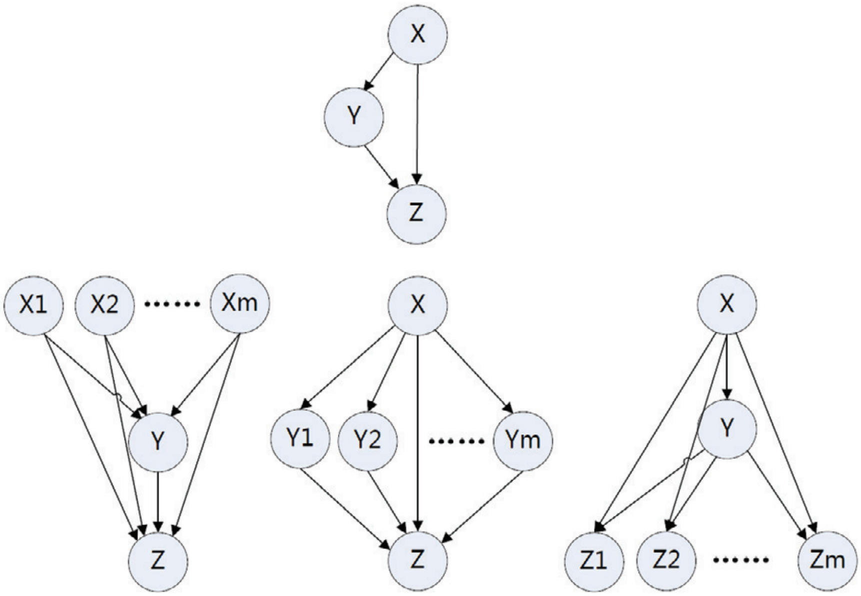


Figure 4: Different generalizations of FFL motif. The above is the original FFL motif. The below row from left to right multi-input generalization, multi-y generalization, multi-output generalization. The size of the generalization is measured by the number of appearances which it comprises (“m” for these three generalizations).

are “大小Z”, “有无Z”, which roughly corresponds to “big, small, Z”, “have, haven’t, Z”. In this generalization we can see that antonym and adjectives frequently emerge.

After we investigate the three kinds of generalizations, we can extract several compositional rules for the words. First we can see that the most frequently emerged generalizations correspond to prefix, suffix and tense transform in English language. “不” corresponds to negative prefix in English as “in-”, “un-” or “a-”, “了” corresponds to past participle in English. Secondly many Chinese words were combinations of main character with an auxiliary character for defining its meaning or just to discern it from other words. For instance, “子” is the most frequently used second character. The words which they link mostly denote the same or similar meaning as the first character. It is mostly used to compose nouns from single character nouns in traditional Chinese. “被子” (quilt) is a modern Chinese substitution for “被”, “靶子” (target) for “靶” and “豹子” (leopard) for “豹”. Thirdly, many antonyms can combine with the same character, resulting in the difference of level or contrary meaning. For instance, “大事” (major event),“小事” (trivial event),“有事” (to do something or something happens),“无事” (to do nothing or nothing happens), etc. This makes the composition of new words much more flexible.

6. Conclusion and Discussion

In this paper we analyzed the motif structure in Chinese word network. We found that the FFL motif is an important motif in the network. This property is like the biological network and social network, rather different from the language network generated from texts [2]. This may be due to the addition of many compositions which are not words from the real world texts. And also we can see that the Chinese word network is quite different from other languages. We carried on analyzing the FFL motifs. And we found that the biggest generalizations of FFL motifs are those with two antonym characters or an auxiliary character. We found that those auxiliary characters are often used to denote the category of some words or state of something or someone. “子” and “了” which we discussed in the above sections are two examples. These words appear more frequently in modern Chinese while traditional Chinese uses a different method to express these concepts.

Chinese is distinct from other languages in that its characters reached maturity two thousand years ago. During that period, new characters were seldom created throughout two millennium period. But the changing environment requires new constructions to represent the newly founded and developed objects and ideas. Therefore more and more words are created with characters. The composition rule plays a very important rule in forming new Chinese words. Different characters are possible to construct new words which have same, similar or opposite meanings. For instance, the words used to denote the sun is “太阳”, and the moon is “太阴”, and “阴阳” is used to denote the uncertain variation between bright and dim. So this is an example to use antonyms to construct new words. Such examples abound in Chinese language, which we analyzed earlier (FFL motif). So motif analysis of this complex network is very useful for understanding the development and evolution of Chinese language.

In the future, other than the aforementioned work, we will use motif analysis to model the composition of longer words. We think this would result in a hierarchical representation of the motifs, and it would give us more interesting discovery.

References

1. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Network motifs: simple building blocks of complex networks, Science 298 (5594) (2002) 824.
2. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Superfamilies of evolved and designed networks, Science 303 (5663) (2004) 1538.
3. C. Echtermeyer, L. da Fontoura Costa, F. Rodrigues, M. Kaiser, Automatic network fingerprinting through single-node motifs, PloS one 6 (1) (2011) e15765.
4. N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs, Bioinformatics 20 (11) (2004) 1746–1758.
5. S. Wernicke, Efficient detection of network motifs, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2006) 347–359.
6. S. Wernicke, F. Rasche, Fanmod: a tool for fast network motif detection, Bioinformatics 22 (9) (2006) 1152.
7. J. Li, J. Zhou, Chinese character structure analysis based on complex networks, Physica A: Statistical Mechanics and its Applications 380 (2007) 629–638.
8. S. Valverde, R. Solé, Network motifs in computational graphs: A case study in software architecture, Physical Review E 72 (2) (2005) 026107.
9. N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, Topological generalizations of network motifs, Physical Review E 70 (3) (2004) 031909.
10. L. Sheng, C. Li, English and chinese languages as weighted complex networks, Physica A: Statistical Mechanics and its Applications 388 (12) (2009) 2561–2570.
11. Chinese Academy of Social Science, Modern Chinese Dictionary, The Commercial Press PRC, 2008.
12. Editorial Board for Contemporary Dictionary, Contemporary Chinese Dictionary, Zhonghua Book Company, 2011.
13. The Commercial Press, Xinhua Dictionary, The Commercial Press,PRC, 2011.
14. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, U. Alon, Response to comment on” network motifs: Simple building blocks of complex networks” and” superfamilies of evolved and designed networks”, Science 305 (5687) (2004) 1107.
15. Z. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, A. Masoudi-Nejad, Kavosh: a new algorithm for finding network motifs, BMC bioinformatics 10 (1) (2009) 318.
16. Adobe reader simplified chinese addon, `http://www.adobe.com/support/downloads/product.jsp?platform=windows&product=10`.