Chinese language processing with complex network theory

You-Yang Yu Zhi-Qing Wang Wei-Nan Gao Guo-Qing Gu
School of Information Science and Technology, East China Normal University,
Shanghai, China
yyy82@163.com

***Abstract*—We defined two kinds of Chinese words network (CWN) in this paper. The nodes in the network are composed by the Chinese characters, phrases or classical idioms from authority dictionaries. Studying the network characters and giving a new evolution model to simulate the CWN, we found that the phrase construction follows random and preferential choosing method and we found 55 Chinese characters used by ten thousand distinct phrases. Finally, the Chinese words network exhibiting Small-world character will make it easy to search information as fast as English language.**

Keyword: complex networ; Chinese language processing

I. Introduction

Natural language is an important tool in our daily life. There are more than five thousand different languages in the world [1]. English as the most widely used language has been investigated for many years. However, using complex network theory in language study is a new method in recent years.

Since the seminal papers by Watts and Strogatz [2] on the small-world character and by Barabasi and Albert on Scale-free feature [3] were published in Nature and Science, the structure and the character of the complex network have become one of the hottest topics in physics and other relation fields. The small-world concept means two nodes connect each other in very short path but the clustering coefficient is higher than random network which has the same number of nodes. Studies have shown that in nature there are so many networks representing the feature of Small-world, such as food webs [4], human sexual contacts [5] and so on. The scale-free property, on the other hand, is defined by the power law behavior in the probability distribution of degree. The real big Scale-free network is WWW [6]. The dynamic model for Scale-free network had been proposed by Barabasi and Albert [3] which demonstrates growth and preferential attachment as two basic factors for the evolution of the Scale-free network. Although the theory is perfect and so many real networks exist, the complex network method had not been used to study the human language until 2001.

Ferrer I cancho built the first English word network in 2001[7]. He defined the node of network as the word and link as significative co-occurrences between words in the same sentence. According to the simulation, he found that the English word network is a small world network, to be specific, the average shortest path between any two words is 3. This character makes the English have higher hunt speed, which is more important for its using in World Wide Web. Conceptual network is another efficiency way to study the language using complex network. It was proposed by A.E. Motter in 2002 [8]. He constructed a conceptual network from the entries in a Thesaurus dictionary and considered two words connected if they express similar concepts. The conceptual network presenting a Small-world property indicates the information saved in our mind is efficient and is picked up rapidly.

The above successful experience proofs that the complex network theory is a good way to study natural language. In recent years, more and more people pay much attention to the network construct of Chinese language [9,10,11]. How it develops and how to use it more efficiently are the useful question deserving us to investigate.

II. Model

Chinese language has a long history. There are so many kinds of word classes, such as special words, spoken language or classical language and so on. In 2005, Li-Yong with his associates made the first Chinese phrase network [9,10] based on three different sources: daily used phrases including 10746 nodes, Internet phrase including 47951 nodes and other 96234 nodes provided by the Chinese information research group. CPN exhibits Small-word effect and power law distribution of degree as English network. But in CPN there are not Chinese characters which are more important elements in Chinese language. So we think it is useful to build a new complete network including Chinese characters and phrases.

The first important thing is to choose an authority vocabulary to make the network more practical. In our country, there are so many dictionaries. Here, we choose XinHua dictionary published by the Commercial Press in 2003 which is the first modern Chinese dictionary issued in 1953. Revised many times in large scale by hundreds of experts, this book includes almost every field terms, such as politics, economy, culture, technology and so on. Widely used in our every day life, it is a perfect word sources for Chinese language research. On the other hand, we not only study the Chinese characters and phrases but also research classical idioms, a special part in Chinese language. The classical idiom is not the same as daily idioms or proverbs. Their form is to some extent unchanged. Normally, they are composed by four Chinese characters. Classical idiom is wisdom of ancient people. The differences between the classical idioms and normal Chinese words construction may reflect

This work was supported in part by the NNSF of China under Grant No. 10635040.

978-0-7695-3336-0/08 $25.00 © 2008 IEEE
DOI 10.1109/CSSE.2008.986

language development. The source here is also XinHua idioms dictionary.

Secondly, the nodes in our Chinese words network (CWN) are Chinese characters or phrases. For any two different nodes, if one phrase contains the other phrase or Chinese character, they are linked together. For example, in figure 1, the node "计算机" includes the other four phrases or characters, i.e. "计" "算" "机 "and "计算", so it has four edges in our network.

We use $G(V,L)$ to describe Chinese words network (CWN), $V=\{v_i\}$, $L=\{l_{ij}\,|\,v_i \subset v_j \ or\ v_i \supset v_j\}$, where v_i denotes individual node i , l_{ij} denotes the connection between nodes v_i and v_j if node i contain the node j 's characters or node j contains the node i 's characters.

The difference between our Chinese words network (CWN) and CPN [9,10] is not only word source but also construction method. In CPN, the link between nodes was defined by the co-presences of the same character in the two phrases. For example, there are links between the nodes "计算机" and "机器" in CPN, but no link in CWN.

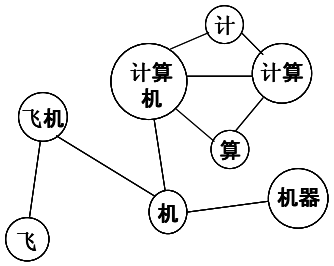

Figure 1. Network construction for CWN

Unlike English language whose basic elements for words are 26 letters, Chinese phrases are composed by Chinese characters. In Xinhua dictionary, there are 39991 distinct phrases, but only 9620 Chinese characters. How the small number of characters composes so many different phrases is an interesting question. In this paper, we try to give the answer.

III. Results and Discussion

Let us consider three important parameters in complex network, probability distribution of degree, clustering coefficient and average shortest path. Every node in network has a number of edges, which is defined as the degree k of this node. The probability of this degree k happens in the whole network is the probability distribution of the degree. For the CWN we draw the probability distribution of degree in figure 2. Figure 2(a) is for Chinese characters and phrases from Xinhua dictionary and figure 2(b) is for classical idioms. When k is small in the figure, there are the biggest value for $k \approx 7$. This phenomenon is obvious especially in figure 2(b). That is to say, when k is small, the random links in network construction is as dominant factor, but with the increase of degree k , probability distribution exhibit power law behavior satisfying $P(k)=ck^{-\gamma}$ with $\gamma=1.9$. The similitude exponent displayed by modern Chinese words network and classical idioms network shows that there exist some principle for language evolution. In order to explore the principle, we give a new evolutionary model for CWN.

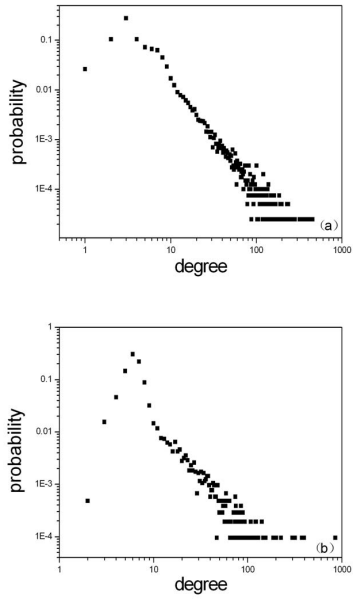

Figure 2. probability distribution of degree for CWN
(a) Chinese characters and phrases (b) classical idioms

The evolution network model (ENM) include $N+M$ nodes. There are N initialized nodes and M new nodes added to the system one by one in every time step. The generation algorithm is as follow:

(1)Initialization : N isolated points whose initialized degree is zero. The initial attract probability for every node $i \in N$ is α_0 .

(2)Grow: add a new node m to the system in each time step, until the whole new node number is M . A new node's initialized degree is k_0 , $2 \le k_0 < \infty$. The frequency for k_0 is $(1/2)^{k_0-1}$, recorded as $P(k_0)=(1/2)^{k_0-1}$. For example, the probability is ½ for adding a new node with degree 2 to the system, but the probability is ¼ for adding a new node with degree 3. Anyway, the bigger degree the node is with, the little probability it is added. The sum for probability satisfies formula (1). On the other hand, the initial attract probability is given for every new node $\alpha_m=\alpha_0$.

$$\sum_{M} P(k_0) = \sum_{k_0=2}^{\infty} \left(\frac{1}{2}\right)^{k_0-1} = 1 \quad (1)$$

(3) Preferential: probability Π_i that the new node will be connected to node i depends on the attract probability α_i of node i.

$$\Pi_i = \frac{\alpha_i}{\sum_j \alpha_j} \quad (2)$$

Denominator in (2) represents the sum of every node's attract probability in the system. With the increase of the degree in every node, the attract probability changed according to the following formula (3).

$$\alpha_i = \alpha_0 (1 + k_{ti}) \quad (3)$$

If the node $i \in N$, then $k_{ti} = k_i$, else if node $i \in M$, then $k_{ti} = k_i - k_0$. k_i is the degree of node i at this time and k_0 is the initialized degree for node i when it was first added to the system.

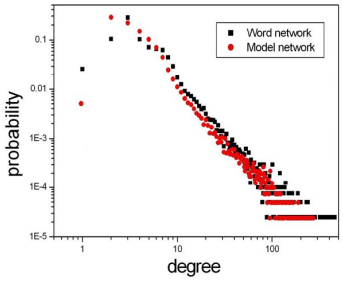

Figure 3. comparison between CWN and ENM

In the evolution rule, we found that when the node number is small, random choosing is mainly available method, but with the increase of the node, the old node with higher degree has more and more bigger attract probability, so they may become hubs nodes (with many links) in the future. The whole network's probability distribution exhibits power law behavior. Figure 3 is the comparison between the evolution network model ENM and the CWN. The experience is executed with $N = 1000$ $M = 40000$ and $\alpha_0 = 2$. The abscissa is degree of nodes, and y-axis is probability distribution. When k is big enough, the simulation error is small. So we get the conclusion that the Chinese phrases are built with preferential attachment, i.e. more widely used Chinese characters easily compose the other new phrases. We called those Chinese characters the core characters. When k is small, because of the random factors, the fitting presented relatively bigger error.

In the following, we attempt to uncover the other two statistical properties of CWN, clustering coefficient and average shortest path. In a network, let i be a node with degree k_i, it is easy to verify that there are at most $k_i(k_i - 1)/2$ edges among its k_i neighbor nodes. Let E_i be the number of the real edges existing among those k_i nodes. Then the ratio between the real and possible numbers of the edges in the sub graph of the k_i nodes is defined to be the clustering coefficient of node i, denoted C_i .namely: $C_i = 2E_i / k_i(k_i - 1)$. The average clustering coefficient for the network is $C = \sum_{i=1}^{N} C_i / N$.

The shortest path between every two nodes v_i and v_j is defined as d_{ij}. Then the average shortest path for the whole network can be written as:

$$d = \frac{1}{C_N^2} \sum_{i,j \in N}^{i \neq j} d_{ij} \quad (4)$$

By the above method, calculate the two statistical properties for CWN seeing TABLE Ⅰ. Group 1 represents Chinese characters and phrases from Xinhua dictionary. Group 2 represents classical idioms from Xinhua idiom dictionary. Because the whole network is not connective, we choose the biggest sub web to calculate those two parameters.

TABLEI. STATISTICAL PROPERTIES FOR CWN

Network	Node	Edges	C	d
Group1	35164	2059403264	0.54	2.82
Group2	10225	257642420	0.63	2.46

Obviously, CWN exhibits Small-world character, its clustering coefficient is 100 times more than the random network with the same number of nodes, but the average shortest path is the same as the random network. This result is consistent with CPN [9] although the construction method is very different from each other. Our results proved again that Chinese language has a high hunting speed as fast as English language. However, in information processing, coding high frequency using Chinese characters with relatively shorter code will save more space and raise processing speed which can not be achieved by English.

IV. Application

The degree of word in TABLE Ⅱ indicates the frequency in which it appeared in phases. As we know, the language is accumulated and innovated all the time. The more we are familiar with a word, the more probability we will use it make up new phases. Therefore, a word with high frequently appeared

in phases will be more important than others in daily life. There are various kinds of character sets for Chinese character, for example, GBK, GB2312, Unicode and so on. Among these sets, each character is expressed and stored by two bytes. It's really hard to modify big changes for those sets, however, on the certain condition, it can be more effective if we encode the core Chinese characters by Huffman coding. For example, in some environment, only Chinese character is accepted both in input or output, short coding length for high frequency Chinese characters will save more space and enhance processing speed.

TABLEII. CORE CHARACTER WITH ITS DEGREE IN CWN

不(454)	人(418)	大(382)	国(359)	子(332)
生(319)	地(282)	一(272)	物(271)	行(267)
心(265)	中(262)	动(261)	法(256)	义(254)
主(252)	文(248)	化(239)	无(237)	天(236)
会(235)	电(230)	体(229)	水(229)	中(262)
主(252)	学(223)	制(217)	风(215)	山(209)
海(208)	经(207)	本(200)	气(199)	马(199)
战(198)	金(198)	之(197)	分(192)	自(192)
机(186)	合(186)	合(186)	事(185)	然(184)
工(182)	业(181)	性(180)	公(177)	民(170)
成(169)	流(169)	道(168)	力(164)	平(164)

On the other hand, from the TABLE Ⅱ we can see that only one word may appear in 454 distinct phases, collecting all 55 widely used characters we will get more than ten thousand different phrases. As a new language learner, it will be efficient if he masters the core characters first. In other words, TABLE Ⅱ illustrates that a small group of Chinese characters can make up the overwhelming phases. Thus, to people who want to study Chinese fast, they can make great progress in short term if they begin from these core characters.

V. Conclusion

In our paper, we built a new Chinese words network including Chinese characters, phrases and classical idioms. This CWN exhibits Scale-free character and Small-world property. In order to explore the development of Chinese language and find out the relationship between the Chinese character and the phrase, we proposed a new evolution model which exhibits a very similar probability distribution of degree as the CWN. According to this evolution model, we found that random and preferential attachment are the most important factors in phrase construction. Although Chinese language is more complex than English, its Small-world character make it high search speed. So with the development of computer, Chinese will play an important role in information communion.

REFERENCES

[1] Hugh.Kisholl,etal., Encyclopedia Britannica, Encyclopedia Britannica, Inc, London, 1911.

[2] D.J. Watts, S.H. Strogatz, “Collective dynamics of 'small-world' networks”, Nature Vol 393, 1998, pp 440-442.

[3] A.L. Barabási, R. Albert, “Emergence of scaling in random networks”, Science, Vol 286,1999,pp.509-512

[4] Montoya. J.M.Sole, R.V., “Small world patterns in food webs”, J. Theor. Biol., Vol 214, 2002, pp:405-412

[5] Liljeros, F., Edling, C.R..Amaral, L.A.N. et al., “The web of human sexual contacts”, Nature, Vol 411, 2001,pp:907-908

[6] R. Albert, H.Jeong and A.L. Barabási, “Diameter of the world-wide web”, Nature ,Vol 401, 1999,pp.130-131

[7] Ramon Ferrer i Cancho ,R.V. Sole, “The small world of human language”, Proc, R, Soc. Lond. B, Vol 268,2001,pp. 2261-2265.

[8] A.E. Motter, Alessandro P.S. de Moura, Y.C. Lai, and P.Dasgupta, “Topology of the conceptual network of language”, Physical Review E, Vol 65,2002,pp.065102

[9] Y.Li, L.X. Wei, W.Li. Y.Niu and S.Y. Lou, “Small-world patterns in Chinese phrase networks”, Chinese Science Bulletin, Vol 50, 2005, pp 286-288.

[10] Y.Li, L.X. Wei, Y.Niu and J.X. Yin, “Structural organization and scale-free properties in Chinese Phrase Networks”, Chinese Science Bulletin, Vol 50, 2005, pp 1304-1308

[11] H.T. Liu, “The complexity of Chinese syntactic dependency networks”, Physica A, Vol 387, 2008, pp.3048-3058.