Network Properties of Social Tags and Collocates

ABSTRACT

In this poster, we apply network analysis methods in two information networks: a social tagging network and a collocation network. We try to identify their distinct network patterns under different language environments and their relationship with IDF.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *discourse, language modeling.*

General Terms

Measurement, Documentation, Languages, Theory, Verification.

Keywords

Social tagging, Collocation Network, Chinese

1. INTRODUCTION

In recent years, large-scale network analysis has developed rapidly in various disciplines in order to understand the complex behavior of human societies. Current large-scale networks studied in previous research mainly include social networks, information networks, technology networks and biological networks[1]. Despite differences in the nature of these networks, they exhibit similar structural features, namely, small-world and scale-free properties.

In this poster, we apply network analysis methods in two information networks: a social tagging network and a collocation network. We try to identify their distinct network patterns under different language environments. In addition, since the datasets are comprised of the behavioral dynamics of hundreds of millions of people, we attempt to discover new insights about human language that will enable a better understanding of it at both the individual and societal level.

2. SOCIAL TAGS AND COLLOCATES

Social tagging is the practice of collaboratively creating and managing tags to annotate and categorize content. Recently,

social tagging has become very popular in Web 2.0 sites, such as Del.icio.us and Flickr, allowing users to store, organize, and share Web pages conveniently by means of informal tags instead of the traditional formal metadata.

As a whole, current research in social tagging can be classified into four categories. The first category is to construct a network of tags and demonstrate its network properties. The second category is to apply semantic relationships to reorganize tags. The third category is to exploit social tags to improve information retrieval, such as query formulation or snippet extraction. The last is to find social trends from the popular tags.

In particular, a closer examination of the properties of tag networks can lead to a better understanding of how language is used to describe and classify information. However, relatively few studies have compared the practice of tagging in English with that of Chinese sites and users. This raises both technical and linguistic questions. What is the significance of scale-free and small-world properties of tag networks across multiple languages? What is the relationship between tagging and collocation?

Collocation can be defined statistically as words co-occurring within a certain distance of each other or it can be defined linguistically on syntactic/semantic bases[1]. Although common collocations (ex. "partisan politics") do not intuitively seem to be useful for distinguishing documents, they are nonetheless what many people turn to in social tagging. Collocation also has relevance for computerized extraction of keywords for automated indexing. In this poster, we utilize the Sogou corpus[3] to construct the Chinese collocation network. An illustration of the subnetwork is as shown in figure 1.
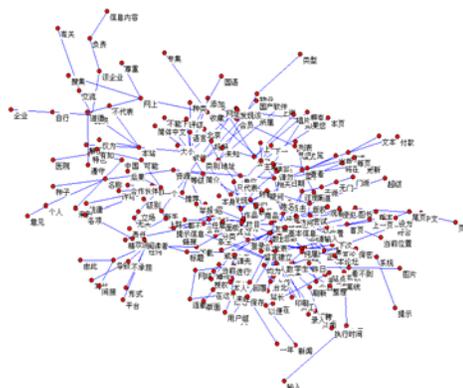

Figure 1. Subnetwork of a Chinese lexical collocation network (Pajek)

3. HYPOTHESIS
3.1 Statement of Hypothesis
In this poster, we will demonstrate the scale-free characteristic of a Chinese lexical collocation network, and we propose the hypothesis that:

There is negative correlation between terms' degree and their inverse document frequency (IDF).

3.2 Testing
To test this hypothesis, we can use a dataset based on more than one hundred million Web pages. First, we construct the collocation network with the dataset and calculate the network degree of each term. Then, we build three samples, each of which is consists of five hundred terms with their degrees randomly selected from the collocation network. At the same time, we fetch the document number N for each term in the samples from the Chinese search engine "Baidu." Therefore, we can test the hypothesis with these datasets as follows.

$$Log(K) \sim Log(N)$$, K is the degree and N is the document number.

In fact, the terms' degree in network can show their collocation ability, which may affect their occurrence in documents. That is, terms can occur in more documents if they have stronger collocation ability, as indicated by their higher degree. This hypothesis could also provide a potential explanation for Inverse Document Frequency (IDF), and indicates that the terms' degree in the collocation network could be used in terms' weighting and keyword extraction for automated indexing.

4. RESULTS
4.1 Small-world pattern [4]
Table 1. Small-world pattern in social tagging network

Network	C	C(Random)	L	L(Random)
Social tagging	0.792	0.000394	3.99483	5.562321

From table 1, we can conclude that $C \gg C_{random}$, and $L \simeq L_{random}$, therefore, this network is a small-world network.

4.2 Scale-free pattern [5]
According to the network analysis results, we find the power-law degree distribution, that is, the frequency, P (k), of a degree, K, is proportional to the degree to the power of a constant, -1.46 as follows:

$$p(k) \sim k^{-1.46}$$

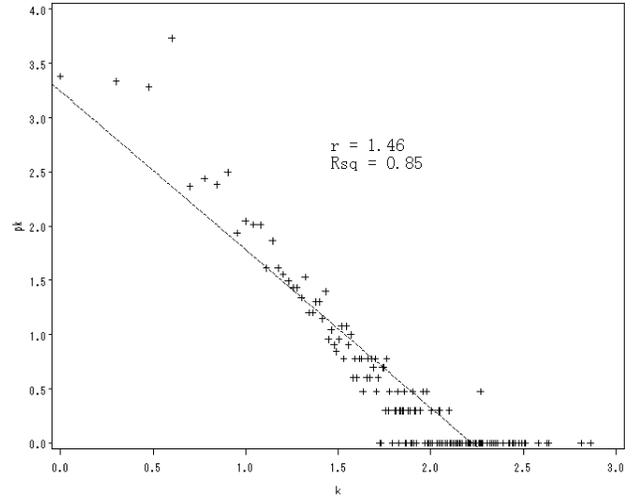The scatter diagram is as shown in Figure 2 with logarithmic scale.

Figure 2. Degree distribution of social tagging network (in logarithmic scale)

In the meantime, in the collocation network, assuming that the frequency, P (k), of a degree, K, is the number of nodes with degree K, we find that the frequency, P (k), of a term's degree, K, is proportional to the term's degree to the power of a constant, -1.171 (λ =1.5, 2.75[6]) as below:

$$p(k) \sim k^{-1.171}$$

Thus, the scale-free feature in the l Collocation Network is attested. The scatter diagram is as shown in Figure 3 with logarithmic scale.
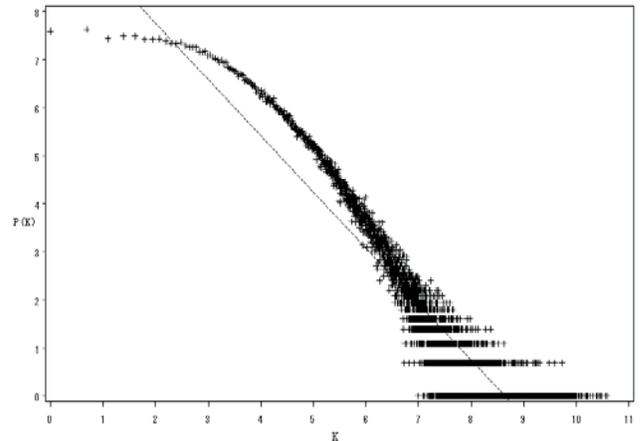

Figure 3. Chinese Lexical Collocation Network Degree Distributions. (Log-log plot of frequency P (K) versus the degree K)

4.3 Analysis of three samples
After analyzing the three samples, each of which is constituted by five hundred terms with their degrees randomly selected from the collocation network, we find persuasive evidence to support hypothesis in all these three samples. The scatter plot of these three samples is indicated in Figures 4, 5, and 6.
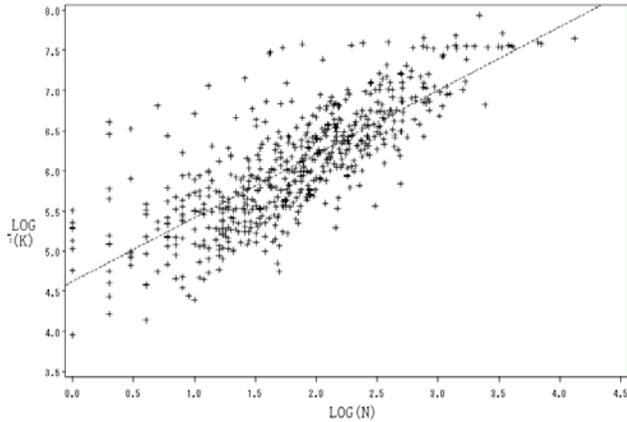
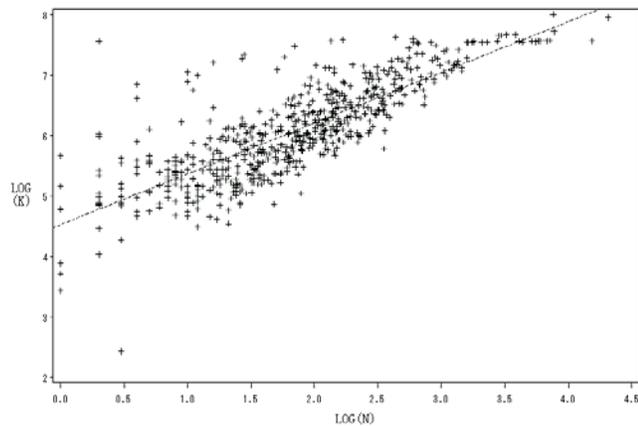**Figure 4. Log-log plot of degree K versus document number N in the first sample**

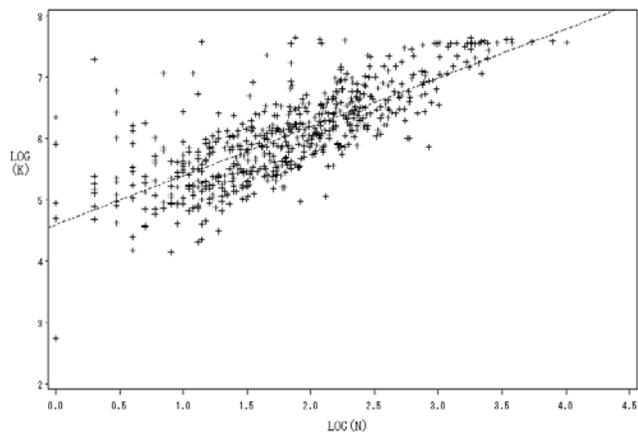**Figure 5. Log-log plot of degree K versus document number N in the second sample**

Figure 6. Log-log plot of degree K versus document number N in the third sample

5. CONCLUSION

5.1 Significance

One question that arises from the preliminary results in this poster: What implications might these properties have for the dynamic behavior of human language systems? In fact, it may be that the language networks combine significant elements of both order and randomness with such properties as small world or scale-free. Thus, the development of language, which is illustrated by language network growth, may also be influenced by both random and ordered factors. In addition, there may be some significant new dynamic phenomena in language development when it is a small-world network. However, since research into large-scale networks is relatively new, more studies in this field are needed to obtain more useful results. As for the hypothesis, it can provide a potential explanation for the term frequency (TF) and IDF term weighting method from a network perspective. Moreover, this hypothesis may also be utilized as a new potential method to weigh terms in the field of information retrieval and natural language processing.

5.2 Future Research

As more and more of the Chinese population gains Internet access, it will be increasingly important for researchers to contextualize their practice through a more global perspective. This is especially true for those pursuing scholarship in the areas of information organization and retrieval, as the Internet user experience begins to shift from an English-centric to a more multilingual one. In fact, by the Chinese government's own estimation, the number of Internet users in China has multiplied by a factor of 28 since the year 2000, with the rate of market penetration still only at approximately 19 percent[7].

In the future, additional empirical studies should be conducted using larger datasets to verify the observation that social tag networks have small world and scale-free properties. A greater emphasis on semantic relationships between tag words will also improve the quality of future findings. Moreover, it is worthwhile to connect the network analysis techniques with computational linguistics, because it is helpful to indicate the potential relationship between terms in the process of language evolution and human mind evolution. This kind of research could also facilitate artificial intelligence, automatic text processing and information retrieval.

In addition, tagging, as a way of organizing information, has implications not only for information professionals and computational linguists, but for cognitive scientists as well. That is, social tagging practices offer insight into ontological modeling and the social nature of epistemology. In this regard, the topic seems an especially productive one for the convergence of multiple disciplines.

6. REFERENCES

[1] Newman, M. E. J. 2003. The structure and function of complex networks. SIAM Review, 45, 2 (2003), 167-256.

[2] Gledhill, C. 2000. Collocations in science writing. Tuebingen, Germany: Gunter Narr Verlag.

[3] Sogou Lab. http://www.sogou.com/labs/resources.html, 2008-11-29.

[4] Watts, D. J. and Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks. Nature 393 (1998), 440-442.

[5] Albert, R. and Barabási, A. L. 2002. Statistical mechanics of complex networks. Rev Mod Phys 74 (2002), 47-97.

[6] Ferrer-i-Cancho, R. and Solé, R. V. 2001. The small world of human language. P Roy Soc Lond B Bio (2001), 268, 7, 2261-2265.

[7] China Internet Network Information Center, Ministry of Information Industry of the People's Republic of China.