文章编号：1003-0077（2011）00-0000-00

微博语言的复杂网络特征研究*

马宏炜，陆蓓，谌志群，黄孝喜，王荣波

（杭州电子科技大学 计算机学院 认知与智能计算研究所，浙江省 杭州市 310018）

摘要：基于大规模微博语料库，构建了 3 个词同现语言网络，并采用复杂网络分析工具对这些语言网络进行分析。主要目的是探索复杂网络分析方法应用于微博文本的可行性，进而研究微博语言网络的个性特征。研究结果表明，复杂网络分析方法在微博文本上是可行的，在复杂网络的相关参数，如度分布、聚类系数、平均最短路径等方面反映了微博语言的语体特征。本研究不仅拓展了复杂网络方法在语言学领域的应用，而且为基于复杂网络的微博内容挖掘提供了可行途径。

关键词：微博；语言特征；语言网络；复杂网络

中图分类号：TP391　　**文献标识码：**A

Research on MicroBlog Language Characteristics Based on Complex Network

MA Hongwei, LU Bei, CHEN Zhiqun, HUANG Xiaoxi, WANG Rongbo

(Institute of Cognitive and Intelligent Computing, Hangzhou Dianzi University, Hangzhou, Zhejiang 310018, China)

Abstract: Based on the large-scale MicroBlog text corpus, three different Microblog word co-occurrence language networks are constructed in this paper, and their network characteristics are analyzed by using complex network analysis tools. The main purpose of this paper is to explore the feasibility of applying complex network analysis methods to the MicroBlog text for studying MicroBlog language network's special characteristics. The experimental results show that the complex network methods are feasible for MicroBlog text. MicroBlog text characteristics are described by the complex network's parameters, such as degree distribution, clustering coefficient, average shortest path, etc.. This research extends the applications of complex network methods into linguistics domain, and provides an effective data mining method on MicroBlog text based on complex network.

Key words: MicroBlog；language characteristics；language network；complex network

1 引言

2006 年诞生的微博，相比于传统媒体虽然还是一种新鲜事物，但由于其独树一帜的简短性（每条微博不超过 140 字）和普及性（人人都可发微博），近几年得到了很大的发展。微博的出现极大地促进了信息的传播和共享，并日益显现出其巨大商业价值。

早期的微博文本相关研究工作主要集中在语言表面特征分析上。Java 等[1]对微博的概念和作用进行了总结，分析了微博及时、快速传播的特点，统计了微博使用增长情况，并根据用户之间的关系，发现了分享相同微博信息的用户之间的共同点。Kwak 等[2]讨论了微博的出现对世界的影响，并全面统计分析了 Twitter 出现三年来相关的数据，包括日发布数、总

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金青年项目（61202281,61103101）；教育部人文社会科学研究青年基金项目（12YJCZH201）

作者简介：马宏炜（1988—），男，硕士研究生，主要研究方向为自然语言处理；陆蓓（1960—），女，教授，主要研究方向为自然语言处理；谌志群（1973—），男，副教授，主要研究方向自然语言处理；黄孝喜（1979—），男，讲师，主要研究方向语言认知计算；王荣波（1978—），男，副教授，主要研究方向自然语言处理。

发布数、总使用人数等。邹艳菁[3]通过使用较大规模的新浪微博语料库分析了微博的话语特征，以定量分析的形式指出了微博语言在话语表达倾向上的一些特点。邬智慧[4]同样通过分析新浪微博语料，统计分析了微博中的字、词、句的使用情况，并将微博语言与手机短信语言、博客语言做了对比，总结出中文微博具有开放性、精炼性、随意性、独特性等特征。

在研究微博文本语言特征的同时，研究者们开始尝试对微博文本进行处理。对微博文本的传统处理主要包括文本分类和聚类、信息抽取、话题检测和情感分析等，主要采用特征提取、分类及聚类算法等文本挖掘方法。这些方法大部分是基于向量空间模型的，其不足之处在于采用传统文本处理方法处理微博文本的时候，忽视了微博的独特特征。由于一条微博的文本限定在 140 个字以内，很多微博文本只是一个句子甚至一个短语，这给传统的文本数据挖掘带来严重的数据稀疏问题。本文尝试使用复杂网络的方法来分析微博文本。

自然界中存在的大量复杂系统都可以用网络来描述，其中具有自组织、自相似、小世界、无标度等特性的网络称为复杂网络。复杂网络的小世界（small world）现象和无标度（scare free）特性是二十世纪末的两个重大发现，奠定了复杂网络的理论基础。1998 年 Watts 和 Strogatz[5]将小世界模型引入到了复杂网络的研究当中，建立了 WS 小世界模型。1999 年 Barabási 和 Albert[6]揭示了复杂网络中的无标度性质，并建立了相应的模型阐述了这些特性的产生机理。这两篇文章的出现，标志着网络科学的兴起。

语言系统是一种复杂的网络结构体，其在词语、语法、语义各个层面上都显示出极其复杂的网络结构。复杂网络理论的兴起，提供了新的视角来研究人类语言的本质。通过用计算复杂网络参数的方法来分析语言网络的特性，可以研究其整体特征，发现人类语言与认知之间的关系。语言网络的研究课题涉及到了复杂系统、语言学、自然语言处理、统计学等多个学科，具有重要的科学意义。

本文基于微博语料库，构建对应的语言网络，并采用复杂网络分析方法对该语言网络进行分析，得到其整体特性，并且运用可视化分析方法，对其特征进行研究。本文第 2 节介绍基于语言网络的相关研究。第 3 节介绍微博语言网络模型及其构建方法。第 4 节给出了数据处理方案、实验流程和结果分析。最后一节是本文的结论和展望。

2 基于复杂网络的语言研究

各国学者已在语言复杂网络研究方面做了很多的研究。这些研究涉及到了多种人类语言，其构造原则也多种多样，包括字同现、词同现、句法依存关系、语义关系等。英语语言网络的研究已经取得了很多成果。Cancho 和 Solé[7]在一千万个词的英语国家语料库基础上建立了词同现网络和句法网络。Motter 和 Moura 等[8]基于 3000 多个英语单词之间的概念相似性构建了英语的概念网；Sigman 等[9]基于 Wordnet 上 66025 个名词之间的语义关系构建了英语的语义网。研究表明这些网络都表现出复杂网络的基本特征：小世界特性和无标度特性。

汉语语言网络的研究也已经取得了一些成果。韦落霞等[10]根据一个基本词集构建了汉语词网络及词组网络；刘知远等[11]在《人民日报》1300 万字的人工分词语料库和国语委 5000 万字人工分词语料库基础上构建了汉语词同现网络；刘海涛[12]基于“实话实说”和“新闻联播”构建了词共现和句法依存网络。对不同语体的字、词同现网络的研究表明，这些网络同样都具有复杂网络的小世界和无标度特性。

语言网络只是研究语言的手段，并不是研究的目的。除了用复杂网络的理论模型来分析语言网络的各项参数之外，更重要的是挖掘其在语言研究中的应用。微博作为语言载体之一，可以通过对微博文本构建语言网络，来对其复杂网路参数进行分析。复杂网络分析技术可以在大规模真实语料的基础上，通过实证方法来研究微博语言网络的特征，加深对微博这种新兴语言形式的了解。复杂网络方法有益于对以下问题的了解：微博语言网络的特征；不同文体网络结构的特征；复杂网络作为语言研究手段的可能性；语言网络作为微博信息挖掘手段

的可行性。

3 微博语言网络模型

基于复杂网络的微博语言特征研究主要分为以下几个步骤：微博数据获取和预处理；词同现网络构建；复杂网络整体参数分析；结果可视化等。下面对各个步骤进行说明。

（1）微博数据获取和预处理。根据任务需求，获取相应的微博数据，并对数据进行预处理，去掉其中的冗余数据和结构，得到结构相对简单的文本待进一步利用。主要任务：①去掉其中的用户名、@用户名、转发关系和网络链接地址，提取出需要的微博内容部分；②剔除处理后长度过短的文本。

（2）词同现语言网络的构建。所谓词同现，是指在一个句子中间隔距离小于某个 n 值的两个词语，在该距离内的词可以称为共词关系。

（3）语言网络复杂网络参数定量分析。将构建出的词同现网络导入到复杂网络分析软件中计算得到复杂网络参数。

（4）结果可视化。将词同现网络的复杂网络分析结果以可视化的形式直观地展示出来。

（5）最后，利用网络分析的结果和其他语体的类似网络进行对比，得出微博文本网络的特性。

3.1 微博词同现网络的构建

对于语言网络的构建，首先要解决的问题是网络中的节点和边代表什么。对词法网络来说，语料库中的每个词，对应着同现网络中的一个节点。如果在一个句子中，两个词之间在小于 n 的邻间距离条件下存在同现关系，则认为网络中相应节点之间存在一个链接。依次对语料库中的所有句子进行上述处理，便构建出词同现网络。刘知远等构建的词同现网络表明，邻间距离的 n 取 2 比较合适，一方面可以真实反映上下文之间的约束关系，另一方面可有效降低网络的复杂程度[11]。

对于微博作为语料库的词同现网络的构建，首先要解决的就是每条微博内容的分词问题。微博文本中充斥着语言的不规则使用现象，并且有大量的新词出现。在选择分词工具的时候考虑到要有新词发现的能力，并且支持自定义词库，对于少数不能通过分词工具得到的词语，可以将其添加到自定义词库中，通过人工干预得到准确度相对高的分词结果。实验中采用了 Python 作为文本处理工具，对微博语料库进行预处理和清理工作，为了便于处理，分词工具选择了 Python 中的中文分词组件结巴分词。该分词组件采用基于图的动态规划查找最大概率算法，从所有可能成词情况所构成的有向无环图中找出基于词频的最大切分组合。对于未登录词，采用了 HMM 模型和 Viterbi 算法。实验结果表明，该分词组件对微博的分词基本可以满足需求。

一个词同现网络可以抽象成为一个无向图 G，顶点 V 代表词集，边 E 代表两个词之间的同现关系。当构建出网络 G(V,E)之后，可以对该网络进行分析。在语言网络中，网络的平均最短路径代表网络中任意两个词之间有联系的最短距离，聚类系数代表与该词有联系的词之间的聚集倾向，度分布代表该词与其他词的结合能力。

下面给出由一条真实微博文本生成汉语词同现网络的简单示例。

这条微博的原始内容为：“我好象不太喜欢听上海话…我喜欢听粤语”。先将微博中的内容根据标点符号分句，得到“我好象不太喜欢听上海话”和“我喜欢听粤语”两个句子。然后分别对这两个句子分词得到的词同现网络如图 1 所示。

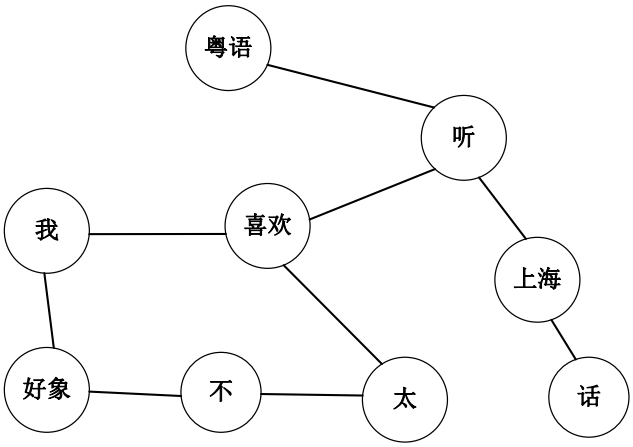

图 1 由一条微博生成的词同现网络

3.2 网络参数

在复杂网络上，通常可以通过以下几个参数来分析其网络的复杂性：

（1）小世界特性：平均最短路径长度和聚类系数

网络中两个节点 i 和 j 之间的最短路径是 d_{ij} 指链接这两个节点的边数最少的路径。无向网络的平均最短路径长度 L 是任意两个节点之间距离的平均值，见公式（1）：

$$L = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij} \tag{1}$$

其中 N 是网络中的节点数。设网络节点的平均度为<k>，对“小世界网络”，则有：

$$L \approx \ln(N) / \ln(< k >) \tag{2}$$

聚类系数是用来衡量网络聚类倾向的指标，反映了其相邻节点构成集合的聚集程度。设网络节点 i 有 k 个节点与它相连，E_i 是其 k 个邻接点之间实际存在的边数，那么 E_i 与这 k 个节点之间最多可有的边数 k(k-1)之比就成为该节点 i 的聚类系数 C：

$$C_i = \frac{2E_i}{k(k-1)} \tag{3}$$

整个网络的聚类系数 C 为所有节点聚类系数 C_i 的平均值：

$$C = \frac{1}{N} \sum_{i=1}^{N} C_i \tag{4}$$

其中 N 为网络的节点数。

利用网络的平均最短路径长度和聚类系数，可以来衡量网络是否具有小世界特性。小世界指的是这样一种网络：虽然网络很庞大，但网络中任意两个节点间都存在一条较短的路径相互连接，聚类系数要比随机网络大的多，即 $L \approx L_r$，$C \gg C_r$。L_r 和 C_r 代表用同样边数和节点数构建的随机复杂网络的平均最短路径和聚类系数。

（2）无标度特性：度分布

度数即网络中某个节点 i 拥有相邻节点的数目，对于有向图来说，度数又分为入度与出

度。网络中度为 k 的节点所占的比列称为度分布，用度分布函数 P(k)来描述。P(k)的期望<k>称为网络的平均度分布。度分布服从幂律分布的网络叫做无标度网络。

4 实证分析

4.1 实验描述

实验利用了从爬盟中国上下载的 2012 年 5 月 25 日起一周内加 V 用户发表的 2 万条微博作为实验语料。应用本文提到的理论和方法，考查微博词同现网络的复杂网络性质。每条微博包括消息原始 ID、微博内容、转发数和评论数、用户名称、发布时间等字段。

为了考察不同网络规模下微博文本的网络特性，设计了三组实验，分别构建 3 个不同规模的词同现网络。第一个网络简称为 CW1，是从两万条微博中随机选取了 2000 条构建的网络；第二个网络简称为 CW2，是从 2 万条中随机选取了 6300 条构建的网络；第三个网络简称为 CW3，采用全部的 2 万条微博来构建网络。构建出词同现网络之后，再分别对网络进行复杂网络参数分析，计算其复杂网络参数。

4.2 微博词同现网络的特征

3 个不同规模微博网络的节点度分布情况见表 1。通过表 1 可以发现：CW1，CW2 和 CW3 的度分布最小值、四分之一分位数、中位数都相同。不同之处在于度分布的平均值、四分之三分位数和最大值。CW1 度分布最大值为 3479，平均值为 4.912，四分之三分位数为 3。CW2 度分布的最大值为 8901，平均值为 6.621，四分之三分位数为 4。CW3 度分布的最大值为 17575，平均值为 9.128，四分之三分位数为 5。由此可见，微博中只出现过一次或两次的词汇占到了一半。这与微博的语体特征是符合的，一方面词语使用不规范的现象在微博中普遍存在，会出现大量的新词。另一方面微博内容覆盖面广，内容多来源于微博用户的日常生活。所以一段时间内的微博可能涉及到生活中的各个方面，其词汇的重复率相比小说、新闻稿等规范文本要低的多。不仅如此，通过比较这 3 个规模由小到大的网络的度分布情况可以发现，随着网络规模的增大，新加入的节点会与已经存在的节点相连接，这就会导致度分布最大值增大，度分布平均值增大。这与实际生活中词语的使用情况是相符的，不断会产生词语的新用法，出现新的词语搭配使用情况。

表 1 节点的度分布情况

网络	最小值	四分之一分位数	中位数	平均数	四分之三分位数	最大值
CW1	1.000	1.000	2.000	4.912	3.000	3479.000
CW2	1.000	1.000	2.000	6.621	4.000	8901.000
CW3	1.000	1.000	2.000	9.128	5.000	17575.000

表 2 给出的是 3 个网络中度数排名前 10 的词语。观察表 2 发现，虽然 3 个网络的规模不同，但网络度数前 10 的节点基本是一致的。不同之处在于 CW1 节点度数前 10 的词其度数并不像 CW2 和 CW3 一样是严格递减分布的。可以认为这是由于 CW1 的规模小，低于能正常反应词语使用情况的阈值，因为部分常用词语还未得到充分使用。在语言网络中，节点的度是由词节点本身所具有的配价能力决定的[13]。通过分析发现，这 10 个词主要是虚词和指示代词，前者有着重要的粘着成句作用，而后者具有指示作用。陈芯莹等[14]通过用“实话实说”和“新闻联播”两种不同语体的语料库构建的依存句法网络研究了词频最高的虚词“的”、“了”和“在”这 3 个节点的网络特性，得出虚词是网络中的中心节点的结论。一旦去除这些词，会影响网络整体结构。同样在微博文本构建的词同现网络中，这 3 个虚词也是

网络的中心节点。不仅如此，经过对比，可以发现“是”这个词在微博中词频排名要比“实话实说”和“新闻联播”要高的多，其度数比“在”这个字要高。“是”在句子中主要起肯定和联系的作用，并可以表示多种关系。由此可以推断微博和“实话实说”与“新闻联播”这两种语体相比，微博的内容更多的跟发布者自身相关，多用来表达自己的认知，更加愿意分享自身的活动。这与发布微博的目的也是相符的。可以认为，“是”是对语体敏感的词语。

表 2　CW1、CW2 和 CW3 度数前 10 的词语

网络	的	了	是	在	和	我	都	有	就	你
CW1	3479	1172	895	654	509	570	458	374	288	398
CW2	8901	2929	2095	1896	1497	1411	1082	1034	1019	1011
CW3	17575	6461	4850	4083	3481	3182	2623	2413	2364	2302

表 3 中 E 代表复杂网络的边数，在构建网络过程中，将多重边合成为一条边，多重边的数量作为边的属性存储。N 代表复杂网络的节点数，<k>代表平均度分布，C 代表聚类系数，L 代表平均最短路径长度，D 代表网络直径，L_{random} 代表相同边数和节点数的随机网络的平均最短路径，C_{random} 代表相同边数和节点数的随机网络的聚类系数。通过观察发现，词同现网络 CW1 的直径为 15，CW2 和 CW3 的直径均为 13，平均最短路径 CW1 为 3.78，CW2 为 3.54，CW3 为 3.34，聚类系数 CW1 为 9,79*10^{-3},CW2 为 9.6*10^{-3}，CW3 为 1.195*10^{-2}。虽然得到的网络很庞大，但其平均最短路径都很小，并且满足 $D \approx D_{random}$ 和 $C \gg C_{random}$，由此可以得出结论：这 3 个词同现网络符合复杂网络的小世界特性。

表 3 其它复杂网络参数

网络	E	N	<K>	C	C_{random}	L	L_{random}	D
CW1	48401	16492	4.912	9.79*10^{-3}	4.2*10^{-4}	3.78	5.69	15
CW2	168129	37438	6.621	9.6*10^{-3}	2*10^{-4}	3.54	5.04	13
CW3	294709	64569	9.128	1.195*10^{-2}	1.283*10^{-4}	3.34	5.26	13

接下来计算网络节点累积度分布，以度为 x 轴，累积度分布为 y 轴，得到其累积度分布曲线见图 2。累积度分布是度不少于 k 的节点的分布概率：

$$P(k) = \sum_{j=k}^{\infty} \Pr(j) \qquad (5)$$

可以看到三组实验结果都服从幂律分布，显示了其无标度特性。

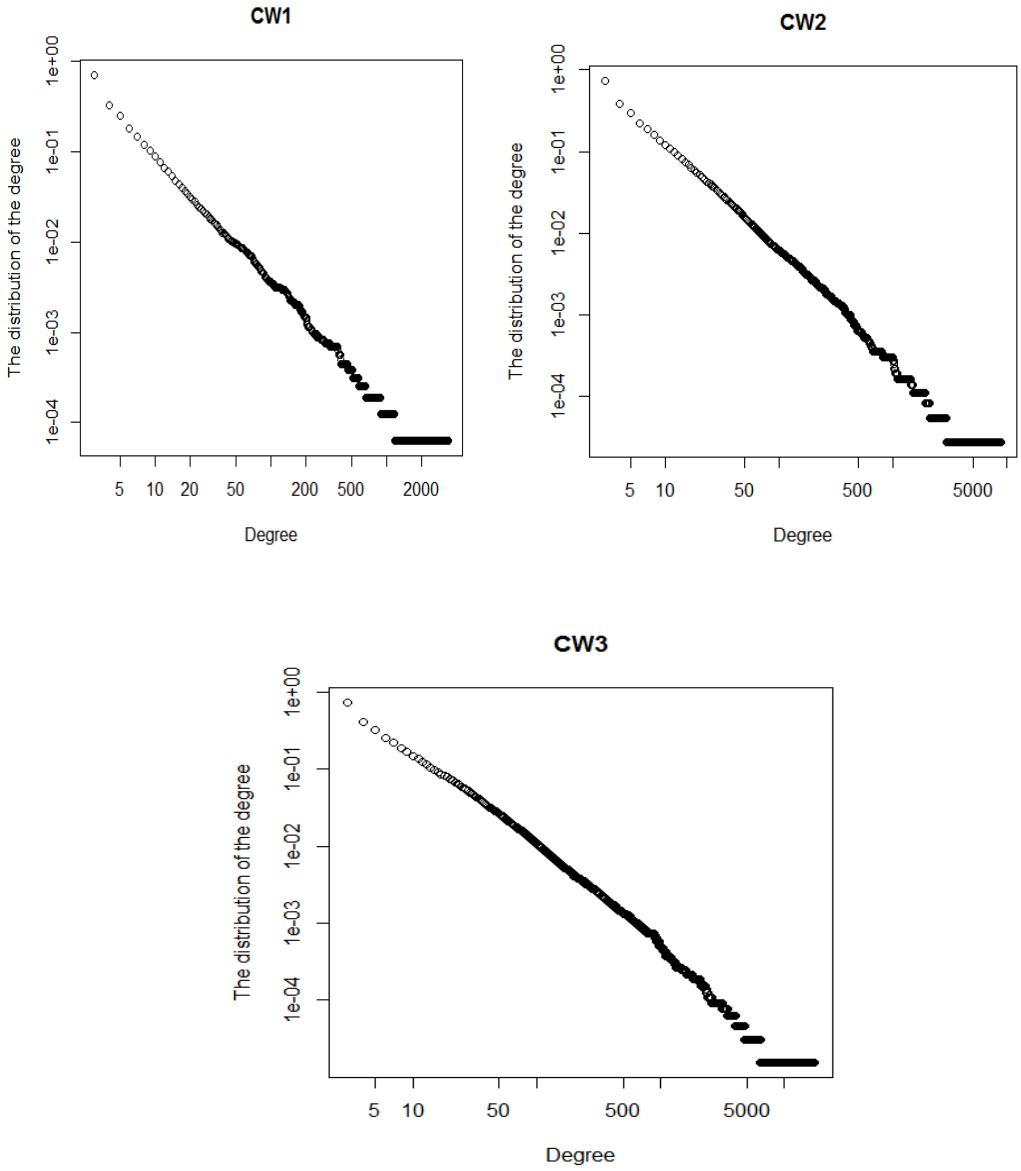

图 2 累积度分布

在计算 CW1，CW2 和 CW3 的小世界和无标度特性参数的基础上，还计算了 CW3 全部两万条微博构成的词同现网络中每个节点的介数（betweennes），紧密度（closeness），聚类系数（cluster coefficient）和 PageRank 值，并且分别以节点的度为 x 轴，这 4 个参数为 y 轴作图。可以得到 4 副图，可以直观看到这 4 个参数与度之间的相关性关系。分布图见图 3。简单来讲，一个节点的介数等于网络中的所有节点对之间经过该节点的最短路径条数。

$$betweenness(v) = sum(\frac{g_{ivj}}{g_{ij}}, i \neq j, i \neq v, j \neq v) \quad (6)$$

节点紧密度等于该节点到所有其余节点最短路径长度之和的倒数。

$$closeness(v) = \frac{1}{sum(d(v,i), i \neq v)} \quad (7)$$

节点的 PageRank 值是 Google PageRank 算法在语言网络中的应用，Mihalcea 和 Tarau

在其论文中提出了 TextRank 算法，首次将 PageRank 算法应用到了自然语言处理当中，并且验证了在关键字抽取和句子摘要中的有效性[15]。

$$S(v_i)=(1-d)+d*\sum_{j\in In(V_i)}\frac{1}{|Out(V_j)|}S(V_j) \qquad (8)$$

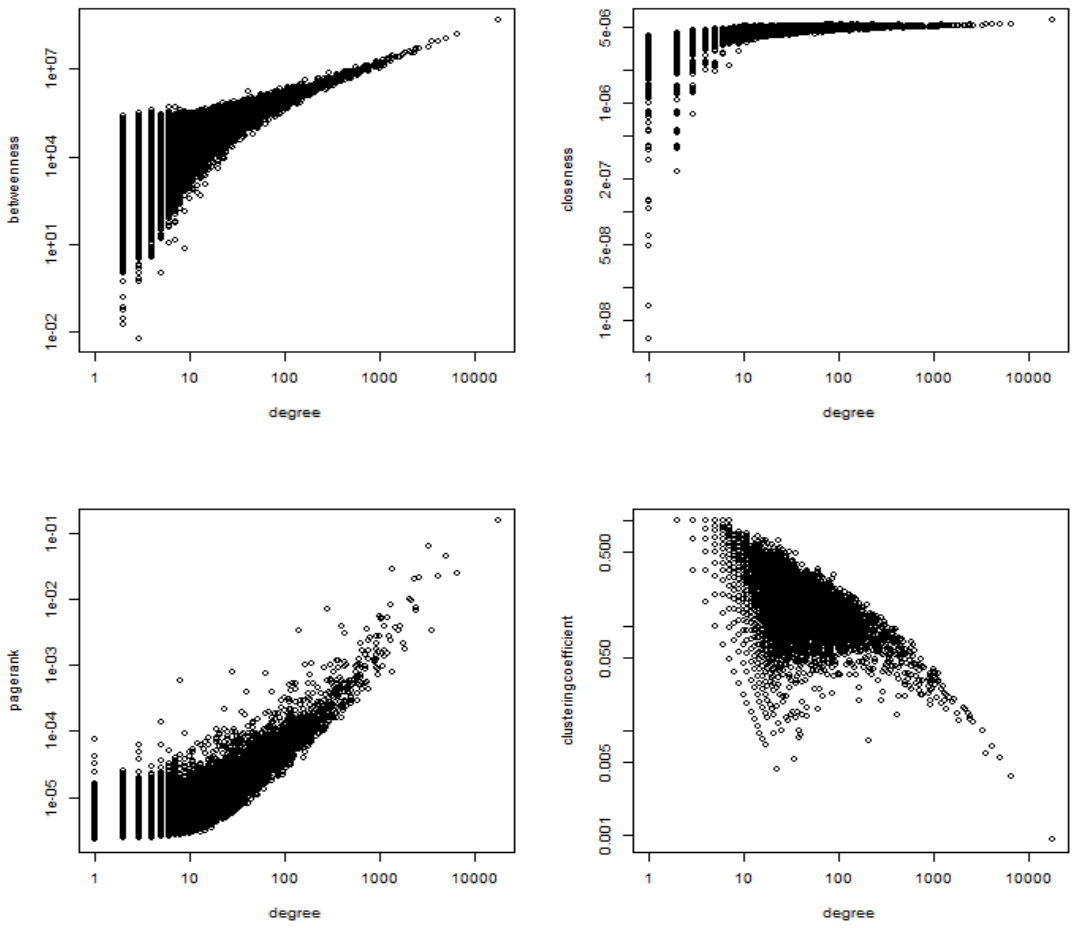

图 3　CW3 的介数、紧密度、PageRank、聚类系数相关性分布

节点的度数反映的是网络中有多少节点与该节点相连，节点的介数和 PageRank 值都可以作为节点在网络中重要性的测量标准，节点的紧密度反映的是节点在网络中的中心性，节点的聚类系数反映其相邻节点的连接情况。通过分析图 3 可以得到，节点的介数、PageRank 和其度之间有着明显的正相关关系，也就是说在该网络中度大的节点，其介数和 PageRank 值也大，节点聚类系数和其度之间基本是负相关关系，对于度大的节点，其聚类系数小。紧密度和节点的度之间不存在明显的正相关关系。

4.3 微博词同现网络和规范文本词同现网络参数对比

刘知远等[11]基于不同规模和类型的语料库，建立了词同现网络。其语料来源是北京大学《人民日报（1998 年上半年）》1300 万字左右的人工分词语料库和国家语委 5000 万字左右的人工分词语料库。前者是新闻语料，后者则包含了各种题材的文本。其生成词同现网络的语料和本文采用的微博文本不同，更加规范并且经过人工分词。对微博词同现网络和这类规范文本词同现网络的参数做一下对比，见表 4。

表 4 词同现网络的基本数据

网络	E	N	<k>	C	C_{random}	L	L_{random}
CW1	48401	16492	4.912	$9.79*10^{-3}$	$4.2*10^{-4}$	3.78	5.69
CW2	168129	37438	6.621	$9.6*10^{-3}$	$2*10^{-4}$	3.54	5.04
CW3	294709	64569	9.128	$1.195*10^{-2}$	$1.283*10^{-4}$	3.34	5.26
CPD12	$0.10*10^{7}$	$0.71*10^{5}$	28.44	0.535	$3.99*10^{-4}$	2.75	3.34

其中 CPD12 是《人民日报（1988 年上半年）》第 1~2 月分词语料库的词同现网络数据，引自[11]。比较表 4 中 CW1，CW2 和 CW3 的参数可以发现，随着网络规模的上升，网络的平均度<k>和聚类系数 C 会随着增大，新的词会被加入到原有的语言当中，原来很少使用的词越来越被人们熟知并使用。相反地，发现网络的平均最短路径 L 随着网络的增大有减小的趋势，这说明网络中词与词之间的跳转更加的容易了，越来越多的词被人们拿来一起使用。把本文构建的微博词同现网络 CW3 和《人民日报》规范文本且经过人工分词处理的语料库构建的词同现网络 CPD12 对比可以发现，网络的平均度<k>和聚类系数 C 要小的多，这也是符合预期的，在微博语言网络的度分布情况就可以看出，一半以上的节点的度都是 1 或者 2，在微博中有更多的新词或者语言的不规范使用情况。不仅如此，CW3 的平均最短路径 L 也比 CPD12 的要大，这说明微博中任意两个词之间有联系的距离要比规范文本要远，这与微博语言使用的不规范也是有关的。

5 结束语

本文基于新浪微博的大规模语料库，构建了 3 个不同规模的词同现网络，并通过实验揭示了微博词同现网络上的小世界效应和无标度特性。虽然微博文本存在着开放性和随意性的特征，但在词同现网络上表现出了类似的复杂网络特性。还对其复杂网络参数做了相关性分析，验证了节点的度是决定一个词网络参数的主要因素。最后与由规范文本构建的词同现网络做了对比，发现其复杂网络参数跟其词汇的使用情况是相关的。本文从定量分析的角度验证了微博的语体特点，验证了复杂网络作为语言研究的手段在微博这一新兴语言载体形式上是有效的。但是，作为一种新的微博研究方法，本文也存在不足之处，一方面语料来源问题，本文构建的微博语言网络，其语料来自爬虫抓取的一段时间内的微博，在内容方面涉及到的范围太广，进一步工作可以尝试抓取某一话题的相关微博或是某条热门微博的评论等；另一方面考虑到微博依存句法分析的难度，本文构建的是词同现网络，难免忽视了语言本身的词语之间的依赖关系，微博依存句法网络的构建与分析是今后研究的重要课题。

参考文献

[1] Akshay Java, Xiaodan Song, Tim Finin et al. Why we twitter: understanding microblogging usage and communities[C]//Proceedings of the Joint 9th WebKDD and 1st SNA-KDD Workshop 2007,2007:56-65.

[2] Haewoon Kwak, Changhyun Lee, Hosung Park et al. What is Twitter, a social network or a news media[C]//Proceedings of the 19th international conference on World wide web,2010:591–600.

[3] 邹艳菁. 基于语料库的中文微博话语特征研究初探[J], 中国报业, 2012, 18:101-103.

[4] 邬智慧. 中文微博的语体特征研究[D], 华中师范大学, 2012.

[5] Watts Duncan J, Strogatz Steven H. Collective dynamics of “small-world” networks[J]. nature, 1998, 393(6684):440-442.

[6] Barabási Albert László,Albert, Réka.Emergence of scaling in random networks[J]. science, 1999, 286(5439):509-512.

[7] Ramon Ferrer i Cancho,Richard V. Solé. The small world of human language[C]//Proceedings of the Royal Society of London. Series B: Biological Sciences, 2001, 268(1482):2261-2265.

[8] Adilson E. Motter, Alessandro P.S. de Moura, Ying-Cheng Lai et al. Topology of the conceptual network of language[J]. Physical Review E, 2002, 65(6):065102.

[9] Mariano Sigman,Guillermo A. Cecchi. Global organization of the Wordnet lexicon[C]//Proc. of the National Academy of Sciences,2002, 99(3):1742-1747.

[10] 韦洛霞，李勇，康世勇等. 汉语词组网的组织结构与无标度特性[J]. 科学通报，2005, 50(15):1575-1579.

[11] 刘知远，孙茂松. 汉语词同现网络的小世界效应和无标度特性[J]. 中文信息学报, 2007,21(6): 52-58.

[12] Haitao Liu. The complexity of Chinese syntactic dependency networks[J]. Physica A: Statistical Mechanics and its Applications, 2008, 387 (12):3048-3058.

[13] 刘海涛, 冯志伟. 自然语言处理的概率配价模式理论[J]. 语言科学, 2007, 6(3):32-41.

[14] 陈芯莹，刘海涛. 汉语句法网络的中心节点研究[J]. 科学通报, 2011, 56(10):735-740.

[15] Rada Mihalcea,Paul Tarau. TextRank:Bringing order into texts[C]//Proceedings of EMNLP 2004,2004:404-411.

作者联系方式：姓名 地址 邮编 电话（最好手机） 电子邮箱

马宏炜 杭州电子科技大学计算机学院 310018 18868879331 mhwqq1988@foxmail.com