文章编号:1000-5471(2013)12-0112-05

基于兴趣值的适应度模型的研究①

杨俊涛

西南大学 计算机与信息科学学院,重庆 400715

摘要:提出了一种基于兴趣的网络模型,通过给每个网络中已存在的节点分配一个兴趣值,兴趣值与该节点的适应度和节点的度有关. 重新建立网络模型,实验结果表明,该网络具有幂律分布.

关 键 词:复杂网络;适应度网络;兴趣值

中图分类号:TP301 **文献标志码**:A

最近,随着科技的发展,对复杂网络的研究显得越来越重要. 特别是,国外两个开创性的研究掀起了复杂网络研究的浪潮. 一是 1998 年 Watts 和 Strogatz 在 Nature 杂志上发表文章[1],引入了小世界(Small-World)网络模型,它描述了一个完全规则的网络到完全随机网络的转变. 小世界网络既具有与规则网络类似的聚类特性,又具有与随机网络类似的较小的平均路径长度. 二是 1999 年 Barabási 和 Albert 在 Science 上发表的文章指出,许多实际的复杂网络的连接度分布具有幂律形式. 由于幂律分布没有明显的特征长度,这一类网络又被称为是无标度(Scale-Free)网络[2].

在 BA 无标度网络中越早加入到网络中的节点的度越高. 然而,在许多现实的网络关系结构当中,节点的度及其增长速度并非只与该节点加入到网络中的时间 t 有关. 比如社会网络当中的某些人具有比较强的沟通能力,他可以比较容易地把一些一次随机邂逅变成一个持续的社会连接;而在 WWW 上的某些网站通过好的内容和市场推广,可以在较短时间内获得大量的超文本连接甚至超越一些老的网站;一些高质量的科研论文在短时期也有可能获得大量的引用次数. 因此,在 BA 模型基础上人们做了各种各样的扩展,其中较为突出的就是适应度模型. 适应度模型通过给网络中的每一个节点设置一个适应度 η 与节点的度的乘积来建立网络模型[3],并得到结论,在有限节点支撑下,该模型与原始 BA 模型一样具有幂律分布特性,在无限支撑下就会表现出“赢者通吃”现象.

文献[4]的研究发现许多真实网络的度有如下特点,真实网络的增长中,新加入的节点并不总是有相同的边数连接到网络中已存在的节点上,而且绝大多数点的度小于 4,见表 1. 文献[5]通过总结分析文献[4]提出了加权适应度模型,通过改变新加入节点连接到网络中的边数来生成网络,且满足每个新节点所带的 n 条边(n 是一个正整数且 $n=\max\{0.26-P(n=1),0.38-P(n=2)\}$. 当 $m=2$ 时,$n=\max\{0.26-P(n=1),0.38-P(n=2),0.14-P(n=1)\}$;当 $m=3$ 时,$P(k)$是当前网络的度分布)连接到网络中已存在的节点上. 文献[6]在文献[5]的基础上改变了新加入节点的边数 n 的选取规则,改为新加入的节点的边数依据表 1 中度为 1,度为 2,度为 3 的节点所占的比重来确定边数 n 的值.

① 收稿日期:2013-04-18

作者简介:杨俊涛(1985-),男,河北唐山人,硕士研究生,主要从事复杂网络技术的研究工作.

表1 网络参数

模型		节点数 N	平均度 $\langle k\rangle$	度分布(%) $P(k=1)$	$P(k=2)$	$P(k=3)$
其他网络模型		11 122	5.4	26	38	14
适应度模型	$m=2$	11 122	3	0	33.9	21.7
	$m=3$	11 122	5	0	0	25.3

对于兴趣网络的研究，大多学者都在研究节点间的兴趣相似度问题，例如文献[7－8]认为网络中相似的节点互相连接，它具有社区结构，节点及其友邻节点就构成了兴趣网络. 并且以2个节点之间的相同文件数目作为相似度，相似度超过某一阈值就认为两个节点有共同兴趣. 文献[9]在此基础上对阈值的选取做了一些改进优化，过滤了尽可能多的伪相似情况.

就兴趣模型而言，目前的研究常采用聚类分析法和线性回归分析方法，进行用户兴趣挖掘[10]，并给出了兴趣公式

$$\alpha_i = f_{\text{class}}(C_i) / \sum_{j=1}^{s} f_{class}(C_j) \tag{1}$$

其中 α_i 为反馈几率，$f_{\text{class}}(C_i)$ 为用户对 C_i 的兴趣.

文献[11]进一步研究了群体兴趣图谱的结构与稳定性的机理，并揭示群体兴趣的群聚现象及其规律. 研究表明，群体中大部分人的兴趣是一致的，群体兴趣网络的入度分布具有幂律特征，群体兴趣图谱基本稳定.

本文所做的主要工作是为网络中已存在的每个节点分配一个兴趣值，兴趣值的选取综合考虑了网络中已存在的节点 i 的当前的度 $k_i(t)$ 和不随之间变化的适应度 η_i，重新考虑网络的演化.

1 兴趣模型

BA无标度网络模型认为，如果节点的度大，那么该节点就应该对新加入网络中的节点有较强的吸引力，而适应度模型认为只要节点适应度大，在若干次加点后就应该有较高的度，但在现实生活中我们经常能看见一些公司的先进技术往往会被其他公司通过一些手段所淘汰. 如当年铱星公司的卫星电话业务，其设计理念庞大、先进，可以提供语音电话和数据业务，但因营销失误，在2000年宣布破产；才华横溢的年轻科学家、发明家特斯拉，他不同意爱因斯坦的理论，与新兴的量子物理学格格不入，而且，当时特斯拉的设想都太过超前，因此遭受排斥、诋毁. 最后，特斯拉死时身边只有成吨的资料. 这些例子都在提醒着我们，适应度高的点不一定就会有较大的度.

无论是BA无标度网络模型还是适应度模型，这些模型均都没有考虑新加入节点的兴趣问题. 以淘宝的买家卖家为例，某淘宝的卖家有一件商品，销量非常好，按BA无标度网络模型的理解就认为该商品对所有的买家都有同样的吸引力. 这显然是不对的，有些买家注重商品的销量，会选择买销量好的. 但是不排除有些买家特立独行，不喜欢销量好的商品，所以这一销量好的商品不会对这些买家有吸引力. 再以人际交往为例，在一个朋友圈中某人甲健谈、幽默、朋友多，按照适应度模型的理解，任何一个新加入到该朋友圈的人都会对甲感兴趣并且与他成为朋友的概率较大. 这显然是不对的，有些脾气秉性与甲完全不同的人乙可能就会觉得甲嘴贫、不着调，且不愿与甲交往，所以甲、乙二人成为朋友的概率就较低.

综上所述，在模型建立的时候，不能僵化地认为节点有较大的度和适应度就会对网络中所有的节点有较强的吸引力(连接概率大)，所以考虑节点的兴趣是很有必要的. 本文在综合考虑了节点的度、适应度提出了基于兴趣值的适应度模型.

1.1 兴趣值的选取

考虑到相关文献对群体兴趣的研究[11]，大部分人的兴趣是一致的，且服从幂律分布的特性，延伸到复

杂网络中可以理解为新加入的节点对网络中以存在的节点 i 的兴趣值与网络中节点 i 的适应度 η_i（η_i 不随时间变化）和节点度 k_i 成正相关，只有一小部分的节点不满足此特性. 联想到现实生活中，大部分人的兴趣爱好是相同的，只有少部分人比较特殊. 所以本文提出了如下的兴趣值的选取方式.

对于网络中已存在的任意点 i，新加入的节点对点 i 的兴趣值服从正态分布 $X \sim N(\mu, \sigma^2)$，并且假定若随机数为负，则兴趣值取为 0.

之所以选取的兴趣值要服从正态分布 $X \sim N(\mu, \sigma^2)$ 是根据了正态分布的特点. 即 X 的值落在期望 μ 附近的概率很大，而 X 的值离期望 μ 的值越远的概率越小. 这正符合了群体兴趣的分布特性，即大多数人的兴趣是一致的，只有少部分人比较特殊.

1.2　兴趣模型的建立

不妨把一个正在演化中的网络分为两个集合，集合 A 中的点是已经加入到网络中的点，集合 B 中的点为还没有加入到网络中的点. *BA* 无标度模型与适应度模型的共同问题在于对于集合 A 中的任意一个点 p，集合 B 中的任意一个点 q 与 p 的优先连接概率是一样的，这完全没有考虑集合 B 中点的差异性. 本文所引入的兴趣模型是在适应度模型的基础上，修正了 q 与 p 的优先连接概率都是一样的问题. 在网络演化的过程中每个节点对于任一新加入网络的节点的吸引程度（兴趣值）都是动态变化的，这符合网络的实际演化过程. 根据上述思想，基于兴趣值的适应度模型构造方法如下：

1）增长：从一个具有 m_0 个节点的网络开始，每次引入一个新的节点并且连接到 m 个已存在的节点上，这里 $m \leqslant m_0$. 每个节点的适应度 η_i 按概率分布 $\rho(\eta)$ 选取，η_i 不随时间变化，且由文献[3] 可知适应度依赖于一个动态指数. 随机生成一个符合正态分布 $X \sim N(\mu, \sigma^2)$ 的随机数作为新增节点对于网络中已经存在的老节点的兴趣值（因为符合正太分布的随机数与群体兴趣的大多数人的兴趣是一致的相似，X 的值在期望 μ 附近的概率较大，这样可以使生成的兴趣网络更加符合现实网络）.

2）优先连接：一个新节点与一个网络中已经存在的节点 i 相连接的概率 Π_i，满足如下关系

$$\Pi_i = \frac{X_i}{\sum_j X_j} \tag{2}$$

其中 X_i 为新加入节点对网络中已存在的节点 i 的兴趣值.

与适应度模型不同，在基于兴趣值的适应度模型中，适应度、节点度高的节点只会对网络中大多数节点有较高的连接概率，而对于个别节点的连接概率可能很小.

2　实验及数据分析

2.1　实验及数据

本次实验工具选取 Matlab 2010b，令 $X \sim N(\mu, \sigma^2)$ 中的 $\mu_i = \eta_i k_i$，原因是产生的随机数的值落在该正态分布的期望值附近的概率很大. 这满足兴趣值与适应度、节点度的正相关特性，同时也不排除个别节点对于适应度和节点度的非正相关特性. 方差决定了分布的幅度，如果方差太小，所产生的随机数就会与期望很接近；若是方差太大，则会使产生的随机数太过于分散. 通过反复实验，令方差 $\sigma^2 = 1$ 时产生的实验数据最为满意.

选取了本次实验的前 500 个节点中适应度最高的 10 个节点，并统计了这些点在网络发展到 1 000 个和 10 000 个节点时刻的度，见表 2，图 1. 因兴趣值是新加入的节点根据于网络中已有的节点的适应度和当前的度所产生的兴趣，所以随着网络的变化，兴趣值也是在不断变化的，故不再表中记录. 适应度是用 matlab 程序中的函数随机产生的一组服从指数分布的随机数[3] 并分配给网络中的各个节点. 根据公式（2）可知，节点的连接概率是与随机产生的服从正太分布的随机数有关的，所以不同时刻、同一个节点的连接概率也是不一样的，故无法在表中体现.

表2 基于兴趣值的网络规模

节点属性	适应度大到小									
	第一	第二	第三	第四	第五	第六	第七	第八	第九	第十
点序号	272	251	165	465	192	387	90	495	342	438
适应度	10.84	10.29	9.37	9.35	9.32	9.12	8.68	8.50	8.22	8.21
1 000点	35	46	130	21	70	27	119	26	25	12
10 000点	932	1 094	1 704	314	1 165	595	1 316	377	261	196

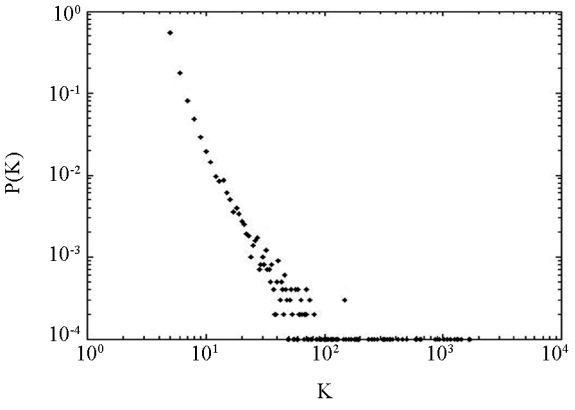

图1 节点度的概率分布图

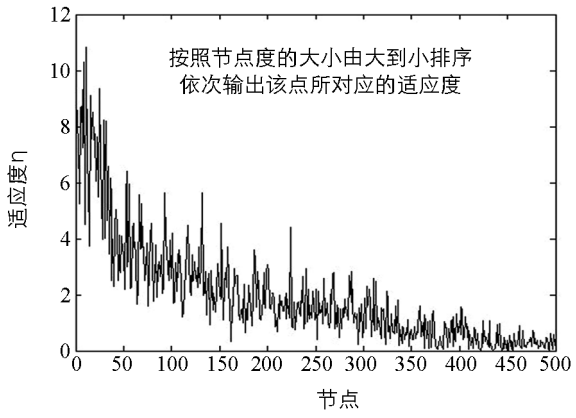

图2 节点适应度分布图

2.2 分 析

通过分析总结表2不难发现，节点度最大的节点并不是适应度最高的节点. 例如本次实验中度最大的节点序号为165号，它的度为1 704，而该节点的适应度值为9.37，且通过多次实验记录发现，适应度不高的点比适应度高的节点所拥有更大的度的现象并不罕见，这可以说明适应度高的节点不一定就有高的节点度.

图2是在网络结束增长后，选取了网络中的前500个节点依据节点度的大小由大到小排序并输出了这些节点所对应的适应度的值，由图2中可以看出，随着节点度的逐渐减少，节点的适应度是呈现出剧烈波动下降的，甚至在图的最开始还呈现出了适应度上升的趋势，这说明兴趣模型虽然总体上服从节点度随着适应度的值减小而减小这一规律，但是在很多情况下还是表现出了与这一规律相背离的情况，这与真实网络是相符的. 如在市场竞争中，许多拥有先进技术的不知名小企业被其他公司打压排挤等.

而且还有一个现象，网络结束增长后，度较大的点都是在网络增长到1 000个节点时度较大的节点. 也就是说，当适应度高的节点有较高的度为依托时才会在网络成长中获得较多的节点. 例如在此网络中节点度最大的是165号节点，此节点在网络成长到1 000时的度为130，是表2中度数最大的点，且经过数次实验，均有此现象.

3 结束语与启示

在考虑兴趣的适应度模型里，出现了虽然适应度很高，但是节点的度和度的增长速度都不高的现象，虽然这种现象与适应度模型不一致，但是却是有现实意义的，正如在引言中所陈述的，适应度高的节点不一定就能获得较多的节点，适应度高的节点要想在网络发展中获得较多的节点必须要有一个较大的节点的度为依托.

本实验只考虑了新节点对网络中已存在的节点的兴趣因素，还没有加入行业间的排挤、打压等商业因素的影响，高适应度的节点的度没有低适应度的节点的度大的现象就已经显现了. 联想到现实，近年来我国虽然投入了大量的人力、物力、财力来搞科技研发创新，但是对科技创新发掘保护推广、对知识产权的

保护和营造良好的市场环境等方面做得远远不够. 这就如同在这个网络中，虽然有很多节点的适应度很高，但是由于没有较大的节点度为依托，所以往往在网络成长中不能获得较多的节点，正如文献[12]讲到，对于一个企业来说，新产品的推广和新产品的研发同样重要，缺一不可. 所以对科技创新的发掘、保护、推广和对知识产权保护的重要性丝毫不亚于研发与创新.

参考文献：

[1] WATTS D J. The"new"Science of Networks [J]. Annual Review of Sociology，2004(30)：243－270.

[2] ALBERT-LÁSZLÓ BARABÁSI，RÉKA ALBERT. Emergence of Scaling in Random Networks [J]. Science，1999，286(5439)：509－512.

[3] BIANCONI G，BARABÁSI A L. Competition and Multiscaling in Evolving Networks [J]. Europhys Lett，2001，54(4)：436－42.

[4] ZHOU S，MONDRAGÓN R J. Accurately Modeling the Internet Topology [J]. Phys Rev E，2004，70(6)：066106.

[5] CHEN X. Weighted Fitness Model in Complex Networks [C]//Engineering and Technology (S-CET). IEEE：2012 Spring Congress，2012：1－3.

[6] CHEN X. Priority Weighted Fitness Model in Complex Networks [C]//Computer Science & Service System (CSSS)，IEEE：2012 International Conference，2012：2197－2200.

[7] LE FESSANT F，HANDURUKANDE S，KERMARREC A M，et al. Clustering in Peer-to-Peer File Sharing Workloads [M] Peer-to-Peer Systems Ⅲ. Heidelberg：Springer Berlin Heidelberg，2005：217－226.

[8] ZHAO S，STUTZBACH D，REJAIE R. Characterizing Files in the Modern Gnutella Network：A Measurement Study [C]//Electronic Imaging 2006. International Society for Optics and Photonics，2006.

[9] 牛尔力，孙晓辉，陈　君，等. Gnutella 中基于兴趣的社区结构研究 [J]. 计算机工程，2009，35(5)：76－79.

[10] 林鸿飞，杨元生. 用户兴趣模型的表示和更新机制 [J]. 计算机研究与发展，2002，39(7)：843－847.

[11] 张　宁. 群体兴趣网的统计特性研究 [J]. 上海理工大学学报，2008，30(3)：243－248.

[12] 廖俊安. 基于市场化的新产品推广策略研究 [J]. 中国商贸，2010(9)：32－33.

On Fitness Model Based on Value of Interest

YANG Jun-tao

School of Computer and Information Science，Southwest University，Chongqing 400715，China

Abstract：A network model has been given based on interest，interest value assigned to each already existing node in the network，and the interest value of the node related to degree and fitness of the node studied. So we can re-establish the network model，and the experimental results show that the network has a power-law distribution.

Key words：complex networks；fitness model；value of interest

责任编辑　周仁惠