复杂网络性质探讨及在垃圾邮件过滤中的运用

李 渊[1] 廖闻剑[1,2] 彭艳兵[1,2] 程 光[3]
(武汉邮电科学研究院 武汉 430074)[1] (烽火通信科技有限公司 南京 210019)[2]
(东南大学江苏省网络重点实验室 南京 210096)[3]

摘 要 基于描述社会网络中幂律分布和小世界效应的网络理论,社会计算能够定量分析社会行为的规律。首先通过幂律分布特征从统计意义上区分了网络中两类度数有差异的节点,这样的方法可以用于垃圾邮件过滤。考虑小世界效应后得到网络平均距离变化缓慢的动态性质,该性质指出了一种平均距离相对固定的网络模型构造思路。最后以邮件数据为实验对象,验证了节点分类的方法对垃圾邮件过滤的有效性。

关键词 社会网络,幂律分布,小世界效应,垃圾邮件过滤
中图法分类号 TP393.01 **文献标识码** A

Discussion of the Characters of Social Network and the Application of Spam Filtering

LI Yuan[1] LIAO Wen-jian[1,2] PENG Yan-bing[1,2] CHENG Guang[3]
(Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China)[1]
(Fiberhome Communication Technology Co. Ltd, Nanjing 210019, China)[2]
(Key Laboratory of Computer and Network Technology of Jiangsu Province, Southeast University, Nanjing 210096, China)[3]

Abstract Based on the network theory's capability to describe both the power-law distribution and the small-world effect in the social networks, the societal computing can measure the complex behavior in the society. Firstly the statistical method is proposed to distinguish two types of nodes with different active degrees through the character of the power-law distribution, which can be deployed as spam filter. Considering the small-world effect, the average distance between any two nodes changed lightly while the number of nodes varied significantly. The characters of different active nodes were validated in the spam filtering procedure from the original electrical mail records at last.

Keywords Social networks, Power-law distribution, Small-world effect, Spam filter

1 引言

作为社会计算的研究内容,社会现象需要量化分析,基于复杂理论建立的网络模型对此进行必要的抽象。例如社会关系网络中的大部分节点彼此并不邻接,但绝大部分节点经过少数几步即可到达,这种现象在生命科学、物理学、信息科学中均有报道,被称为"小世界效应"[1]。D. J. Watts, M. E. J. Newman, J. Kleinberg 等人研究发现了小世界效应两个互相独立的特性:高的聚集度与短的平均路径[2,3]。其中 D. J. Watts 和 S. H. Strogatz 在不改变节点数与边数的情况下[4],将规则网络的边按一定概率重新连接,在由规则向随机变化的过程中,得到了 W-S 网络模型,该模型能够表现小世界效应的两个特点。

值得注意的是,社会网络中的节点同时还服从幂律分布:度数为 k 的节点出现的概率正比于 k 的一个负数幂[5,6]。这种现象同样广泛存在,如互联网中超链接网络、飞机航班网络等。建立更真实的社会网络模型,需要同时考虑幂律分布与小世界效应。

另一方面,电子邮件作为现代技术衍生出的沟通媒介,具有社会属性。如今邮件数目激增,垃圾邮件与大量群发商业邮件的界限愈发模糊,寻找一个垃圾邮件快速过滤方法具有实际意义。分析邮件网络的固有特征,继而确定从邮件相对数量上区分两者的理论基础是本文的论证思路。

第 2 节描述了服从幂律分布网络具有的两个性质,并提出从统计意义上区分其中两类节点的方法;第 3 节随之阐明了当该网络引入小世界效应时表现出的一种动态特性,从而确定了一种平均距离相对固定的网络建模思路;最后通过程序验证邮件网络确实服从幂律分布,并且发送次数大于 44 的节点是垃圾邮件发送者的概率很高。

2 无尺度网络的两个性质

节点度数服从幂律分布的网络称为无尺度网络[7]。假设某社会网络对应的图中有 N 个节点,其中节点 i 的度数 d_i 等于与 i 连接的边数,d_{avg} 代表图中节点平均度数。各节点的度

本文受国家 973 计划(2009CB320505),江苏省科技支撑计划(BE2011173)资助。
李 渊(1984—),男,硕士,主要研究方向为社会计算与网络理论;廖闻剑(1970—),男,博士,硕士生导师,主要研究方向为互联网海量信息分析与挖掘技术;彭艳兵(1975—),男,博士,主要研究方向为中文信息挖掘和网络行为分析;程 光(1973—),男,博士,教授,主要研究方向为网络行为学、网络测量、网络管理。

数分布服从幂律，即度数为 d 的节点存在的概率正比于 d 的一个负数幂：$P(k)=C*d^{-\gamma}$，其中 $C>0$，通过 $\lim_{t->\infty}(\sum_{d=0}^{t}C*d^{-\gamma})=1$ 确定 $C=(\sum_{d=1}^{\infty}d^{-\gamma})^{-1}$，$\gamma$ 为分布参数。社会网络中如果存在幂律分布，则有确定的 γ 值与之对应。

2.1 图的稀疏度与 γ 的关系

幂律分布参数 γ 能够描述图的边数。该特征是从度数上将节点分类的基础。由于边数和节点总数决定了图的稀疏度 β，下面讨论稀疏度 β 与幂律分布参数 γ 的关系。

性质 1 稀疏度 β 由节点总数 N 和幂律分布参数 γ 确定。

无尺度网络中，边数 E 与稀疏度 β 的表达式说明幂律分布的参数 γ 与图的稀疏度有直接关系。

$$E=\frac{N\sum_{k=1}^{\infty}d^{1-\gamma}}{2\sum_{k=1}^{\infty}d^{-\gamma}} \qquad \beta=\frac{\sum_{k=1}^{\infty}d^{1-\gamma}}{(N-1)\sum_{k=1}^{\infty}d^{-\gamma}}$$

图 1 反映了三者关系。N 越大，β 受 γ 的影响越明显。$N=100$，β 的极大值是极小值的 5.6 倍；$N=100000$，β 的极大值是极小值的 176.9 倍；N 和 γ 的增加都可以让图更稀疏。

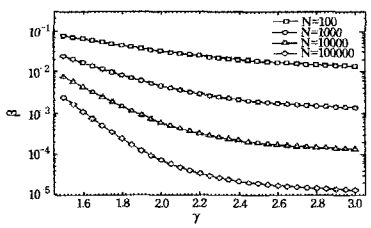

图 1 稀疏度 β 的变化

2.2 无尺度网络中的两类节点

无尺度网络中极小一部分节点由于其高度数支撑着整个网络，称这些节点为 I 类节点；剩下的称为 II 类节点。虽然理论上节点度数服从幂律分布的图中，不存在一个准确的阈值 T 来区分高度数节点与低度数节点，但从统计意义上来看，高度数节点如此之少，以至于任意抽取的节点属于 I 类节点的事件属于小概率事件，可以认为该事件发生的概率低于小概率水平 α[8]。

这样，I 类节点在所有节点中的比例为 α，II 类节点的比例为 $1-\alpha$。考虑到幂律分布的特征，取单边 p-值作为标准。小概率水平 α 属于后验概率，由于网络的类型不同，该值可能发生变化。在对待具体网络时，可以用实际数据来检验 α 值，然后调整 α。在下面的实验中，α 分别取不同的值 0.025，0.0025，0.00025。令分类阈值 T 等于 II 类节点度数的最大值。

性质 2 分类阈值 T 由小概率水平 α 和幂律分布参数 γ 确定，与节点总数 N 无关。

临界度数 T 由下列不等式组确定：

$$\frac{\sum_{d=1}^{T-1}d^{-\gamma}}{\sum_{d=1}^{\infty}d^{-\gamma}}<\alpha \qquad \frac{\sum_{d=1}^{T}d^{-\gamma}}{\sum_{d=1}^{\infty}d^{-\gamma}}\geqslant\alpha$$

从表达式可以看出 T 值与网络中的节点数目无关，这个性质在应用中很重要。图 2 给出了 T 值随 α 与 γ 变化的曲线，表明 T 随 γ 变大而减小。考虑最常见的单边小概率水平 $\alpha=0.0025$，γ 分别取 2.2、2.4、2.6 和 2.8 时，T 值分别为 91、44、27 和 18。现已发现具有幂律分布的社会网络中，γ 值大多集中在 2 到 3 之间，本文第 3 节的实验也表明，正常邮件网络的 γ 值在 2.4 左右，因此确定两类节点的区分阈值为 44。

用统计意义上的小概率事件的观点解决了无尺度网络中没有明显阈值区分两类节点的问题。如果网络行为中包含大量异常行为不断增加 I 类节点的数量，那么可以用一个该预设的阈值量化 I 类节点的异常情况。在电子邮件网络里面一类重要的网络异常行为是发送垃圾邮件，第 4 节将会分析垃圾邮件的发送行为。

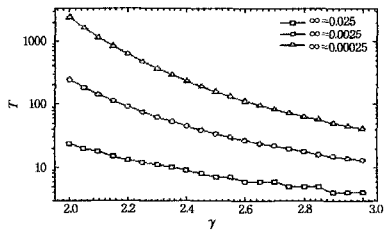

图 2 不同小概率水平下的阈值

3 无尺度网络的小世界效应

短的平均路径是小世界效应的另一个特点。图中任意两个节点间的距离 L 定义为节点间最短路径上边的数目，L_{avg} 代表平均距离。无尺度网络会产生小世界效应，R. Cohen 和 S. Havlin 的工作证明了在无尺度网络中的存在极短平均距离 $L\propto\log(\log N)$[10]。关于网络中节点数目变化的讨论具有研究意义[9]。

3.1 平均距离随节点数变化的性质

性质 3 如果无尺度网络中初始状态下节点数目为 N_1，随后节点数目变化为 N_2，那么平均距离的变化为 $\Delta L<\log(\log_{N_1}N_2)$。

设初始网络中平均距离为 L_1；变化以后平均距离为 L_2；为保证一般性，令 $N_2=k*N_1^q$，$(1<k<N_1, q>0)$，由 $L_1=\log_a(\log_a N_1)$，$L_2=\log_a(\log_a N_2)$，得 $\frac{L_2}{L_1}=\log_{\log_a N_1}\log_a N_2$，计算得 $\Delta L<L_1*\log_{\log_a N_1}^{(q+1)}$，化简得 $\Delta L<\log(\log_{N_1}N_2)$。

该性质保证随着无尺度网络中节点数目发生变化，平均距离的改变很缓慢。比如节点数由 1000000 增加到 6000000000 时，无论原始平均距离为多少，其变化不应该超过 0.5，往往只有 0.2 到 0.3。

以往关于网络模型的研究并没考虑以平均距离相对稳定作为初始条件，而性质 3 意味着可能存在一种网络，其平均距离随节点数目的变化很小。树状模型证实了这种可能性，这也提供了一种新的建模思路。

3.2 树状模型

讨论一棵层数为 t 的 z 叉树，总节点数为 N。根据小世界特性，平均距离与网络中的节点总数 N 有如下关系：

$$L_{avg}\propto\log N$$

同时考虑度数与节点总数的关系得到表 1 和表 2，从 d_{avg} 和 L_{avg} 的具体数值可看出两者变化的规律。分析树状模型下 L_{avg} 的表达式可知，一旦 t 值确定，那么 N 或者 d_{avg} 的改变对 L_{avg} 影响不大，即使向模型中增加节点，也只是增加节点平均度数。

表1 d 对 N 的变化敏感

d_{avg}	t=5	t=6	t=7
N=10000	6.087	4.449	3.556
N=1000000	15.64	9.823	7.041
N=6000000000	90.09	42.46	24.79

表2 L 对 t 变化敏感，对 N 变化不敏感

L_{avg}	t=5	t=6	t=7
N=10000	9.2153	10.8425	12.438
N=1000000	9.7268	11.5467	13.3379
N=6000000000	9.9551	11.9035	13.8319

J. Kleinberg 在讨论社会联系的强弱性的时候提出，在强联系图(社会网络)里，节点度数一般在 10 与 20 之间。如果一个人有 500 个朋友，常发生联系的人数期望会小于 50[11]。因此在表 1 的几组解中考察与目前全球人口数相当的 N(=6000000000)值，确定 t=6,7,8 符合条件。即使在最宽松的条件下，d_{avg} 也不应该超过 42。这样的 d_{avg} 不仅表现了社会网络中个体交互范围的正常值，还与前文确定的分类阈值 T=44 吻合。当然，该数值在社会网络中是否具有普适性还需要进一步论证。

考虑到小概率水平的变化，分类阈值也会发生改变。实验结果表明 T=44 时效果最好，下面就此展开讨论。

4 验证实验及应用

目前，群发商业邮件和垃圾邮件界限模糊。从数量上看，商业邮件群发范围过广，也是一种骚扰，应划归为垃圾邮件的范畴。本文为从数量上区分两者提供了一种统计学方法。下面分析某服务提供商 5 天(包括 3 个工作日和 2 个休息日)的原始邮件交互记录，验证其幂律分布性质以及从数量上过滤垃圾邮件的有效性。该服务运营商每天的邮件数量在亿数量级；进行对比验证的垃圾邮件过滤引擎为自行开发产品，过滤效果较好。实验中剔除了个人相关的任何数据，以保护隐私。表 3 是这次实验中，正常邮件与垃圾邮件的具体数据，可以看到垃圾邮件占有比率在 85%到 90%，与现有报道相符[12]。

表3 实验样本的统计信息

	工作日			休息日	
	I	II	III	IV	V
正常邮件	19059375	18943994	18958028	13600900	10863900
spam	87.29%	86.16%	86.65%	89.69%	90%
分布参数 γ	2.487	2.452	2.328	2.427	2.309

图 3 为工作日 I 的数据中正常邮件的节点度数分布情况，横纵坐标均为对数坐标，可以看到常规邮件数据存在明显的幂律分布。图 4 为工作日 I 包含垃圾邮件的节点度数分布情况，仍可观察到原有轮廓。垃圾邮件发送节点的目的是最大程度地广播消息，这种异常行为大量存在干扰幂律分布，使度数分布变得杂乱。

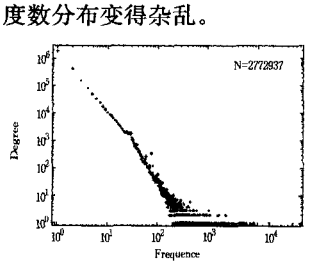

图 3 正常邮件数据的幂律分布

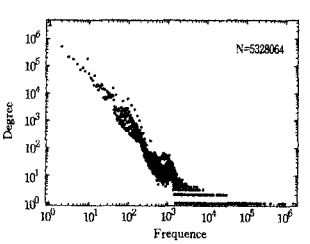

图 4 整体数据集存在大量异常行为

在正常邮件网络中，I 类节点是负责维护正常邮件通信的节点，如通知新用户、退信等工作，这样的节点允许存在。但如果群发数量超过了分类阈值，则应认定为异常行为。在 α 分别取 0.025，0.0025，0.00025 时，对应的分类阈值分别为 9，44，223。如果降低 α 的小概率水平，会降低误中率，但也会减小垃圾邮件的绝对检出数量。

将每天的正常数据作为抽样估计邮件数据整体的度数分布规律，并采取加权滑动平均的方式来确定 γ，这将避免随机误差及抽样误差，其中权重为每天邮件记录的数目在整体中所占的比例。设第 i 天邮件总数为 S_i，幂律参数为 γ_i，则 n 天内总体参数计算公式为：$\overline{\gamma_n}=\sum_{i=1}^{n}\gamma_i \cdot S_i / \sum_{i=1}^{n} S_i$，第 $n+1$ 天的参数计算公式为：$\overline{\gamma_{n+1}}=(\gamma_{n+1}+\overline{\gamma_n} \cdot \sum_{i=1}^{n} S_i)/(S_{n+1}+\sum_{i=1}^{n} S_i)$。因此在求出第 $n+1$ 天的参数 γ_{n+1} 后，需要记录的历史数据只有 3 个：$\overline{\gamma_n}$，$\sum_{i=1}^{n} S_i$ 和 S_{i+1}。表 4 给出了 n=10 时 $\overline{\gamma_i}$ 的取值情况。

表4 滑动平均值 γ 逐步稳定

天数	1	2	3	4	5
γ	2.427	2.373691	2.417195	2.426225	2.405391
天数	6	7	8	9	10
γ	2.409698	2.393716	2.401793	2.409593	2.412488

在滑动平均计算公式里，每日 γ_i 近似以权重为系数对总体参数 γ 产生影响。权重即为每天邮件数量与总体邮件数量的比值。随着时间累积，权重减小很快，因此总体参数 γ 会逐步稳定。实验样本中，γ=2.412。表 5 给出了过滤结果。

表5 不同小概率水平下的检测结果

	α=0.025,T=9	α=0.0025,T=44	α=0.00025,T=223
正常邮件节点	101026	8837	685
异常邮件节点	714952	209971	38254
误中率	12.38%	4.40%	1.76%

T=44 时误中率和垃圾邮件的绝对检出数量都相对平衡。过滤结果意味着当一个邮件发送节点每天发送的邮件数目大于 44 时，认为其不是正常节点的概率为0.9975。所以，正常节点发送邮件数目应该小于该值，只有一些邮件运营商负责处理邮件信息的节点，或者一些会员邮件发送节点时，少量商业广告服务节点可以大于这个值。误中邮件几乎全部是合法商业广告邮件和会员邮件的事实可以证明这一点。

与采用机器学习算法基于垃圾邮件内容进行过滤的传统方法相比[13,14]，基于统计的方法不需要完整的邮件信息，占用空间小；对 1 天的邮件记录数据集只需要 10～15 分钟处理并生成任务记录，速度快；不依赖机器消除词语多义性。据报道，目前规模最大、最有效的过滤系统来自美国。Yao Zhao 等人运用 240 台计算机进行 1.5 小时的运算得到的数据约有 0.44%的误中率[15]。尽管与最精确的方法相比误中率较大，但是如此小的开销完全能胜任服务器前端过滤的要求，对分析过后的数据进一步筛选，而不是一开始就对海量数据进行复杂运算，可以节约大量计算能力。

现在群发商业邮件与垃圾邮件的界限模糊，大量商业邮件在网络中没有限制地扩散，其实已经与垃圾邮件无异。根据邮件网络的固有特性得到的 T 值表明，正常行为不应该超

过该值，合法的商业邮件群发数量也应该参考该值。

结束语 本文分析了服从幂律分布的网络的3个特性，提出了一种节点分类的统计学方法。社会网络中存在幂律分布，其中的节点可以被阈值 T 分成两类。T 值的确定与参数 γ 以及观察者对小概率事件概率 α 的取值相关，而与网络中节点规模 N 无关。网络中大量的异常行为会影响应有的分布规律，导致超过 T 值的节点数远远高于正常值。通过程序分析邮件服务商提供的样本数据验证了该方法的有效性。

在如何有效地建立网络模型方面仍有改进的空间。另外，僵尸网络与垃圾邮件之间有很密切的联系，利用超图理论通过检出的垃圾邮件发送节点寻找僵尸网络也将是一个可行的研究方向。

参 考 文 献

[1] Kleinberg J. The small-world phenomenon: An algorithmic perspective[C]// ACM Symposium on Theory of Computing. 2000, 32

[2] Newman M E J. Models of the small world[J]. Journal of Statistical Physics, 2000, 101: 819-841

[3] Watts D J. The "New" Science of Networks[J]. Annual Review of sociology, 2004, 30: 243-270

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

[5] Newman M E J, Strogatz S H, Watts D J. Random graphs with arbitrary degree distributions and their applications[J]. Physical Review E, 2001, 64

[6] Clauset A, Shalizi C R, Newman M E J. Power-law distributions in empirical data[J]. SIAM Review, 2009, 51: 661-703

[7] Newman M E J. Power laws, Pareto distributions and Zipf's law [J]. Contemporary Physics, 2005, 46: 323-351

[8] Iversen G R, Gergen M. 统计学[M]. 吴喜之，等译. 北京：高等教育出版社，2002: 235-237

[9] Arbesman S, Kleinberg J, Strogatz S. Superlinear Scaling for Innovation in Cities[J]. Physical Review E, 2009, 79

[10] Cohen R, Havlin S. Scale-Free Networks Are Ultrasmall[J]. Physical Review Letters, 2009, 90

[11] Easley D, Kleinberg J. Networks, Crowds, and Markets: Reasoning About a Highly Connected World[M]. Cambridge University Press, 2010: 63

[12] Symantec Corp. Symantec Announces August 2011 Symantec Intelligence Report[EB/OL]. http://www.symantec.com/about/news/release/article.jsp?prid=20110823_01, 2011-08-23

[13] 张铭峰，李云春，李巍. 垃圾邮件过滤的贝叶斯方法综述[J]. 计算机应用研究，2005(8): 14-19

[14] 王斌，潘文锋. 基于内容的垃圾邮件过滤技术综述[J]. 中文信息学报，2005(8): 1-10

[15] Zhao Y, et al. BotGraph: Large Scale Spamming Botnet Detection[C]// Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation (USENIX, Berkeley, CA). 2009: 321-334

(上接第 114 页)

的动态交互、数据的高效管理和组织、上下文随机存储数据一致性等优点，充分利用磁盘并发所带来的性能优势已经在语义网络、数据库等 I/O 密集型的应用程序中得到了广泛的应用。本文通过语义信息对存储能效进行了研究，实验表明基于语义信息的存储优化了系统性能(读写速度和响应时间)、降低了存储能耗，当然论文仅降低了单个磁盘的能耗，提高了单个磁盘的读性能，对于系统各项 I/O 性能的提高有待进一步研究，有关理论和计算结果还需进行实验验证。

参 考 文 献

[1] Huang Hai, Huang Wan-da, Shin G K. FS2: Dynamic Data Replication in Free Disk Space for Improving Disk Performance and Energy Consumption[C]// Proceedings of the 20th ACM Symposium on Operating Systems Principles (SO-SP). NewYork: ACM, 2005: 263-276

[2] 仇德成. 网络存储 cache 替换与磁盘调度算法研究[D]. 兰州：兰州大学，2007

[3] Gurumurthi S, Sivasubramaniam A, Kandemir M, et al. DRPM: Dynamic Speed Control for Power Management in Server Class-Disks[C] // Proceedings of the International Symposium on Computer Architecture(ISCA). New York: ACM, 2003: 169-181

[4] 王娟. 对象存储系统中元数据管理研究[D]. 武汉：华中科技大学，2010

[5] 夏鹏. 文件系统语义分析技术研究[D]. 武汉：华中科技大学，2011

[6] 肖亮. 基于服务质量的对象存储优化研究[D]. 武汉：华中科技大学，2009

[7] 吴晨涛. 对象存储系统中热点数据的研究[D]. 武汉：华中科技大学，2010

[8] Papathanasiou E A, Scott L M. Energy Efficient Prefetching and Caching[C]// Proceedings of the USENIX 2004 Annual Technical Conference(USENIX). Berkeley, CA, USA: USENIX, 2004: 255-268

[9] Zhu Qing-bo, Chen Zhi-feng, Tan Lin, et al. Hibernator: Helping Disk Arrays Sleep through the Winter[C]// Proceedings of the 20th ACM Symposium on Operating Systems Principles (SOSP). NewYork: ACM, 2005: 177-190

[10] Sivathanu M, Prabhakaran V, Popovici F I. Semantically-Smart Disk Systems[J]. Proceedings of the Second USENIX Conference on File and Storage Technologies (FAST '03), 2003, 33 (4): 73-78

[11] Sivathanu M. Semantically-Smart Disk Systems[D]. NewYork: University of Wisconsin-Madison, 2001