

基于语言分类的 www 网络中社会经济指标幂律分布

聂 锐¹, 黄传峰^{1,2}

(1.中国矿业大学管理学院, 江苏徐州 221008; 2.枣庄学院, 山东枣庄 277100)

〔摘要〕 从语言角度对 www 网络的社会经济指标以及它们之间的相互关系进行了分类研究。通过统计模型计算发现, 这些以网民/网页使用的语言为分类的指标, 在不同尺度上近似地遵守幂律分布, 反映了网络世界中社会、经济现象的分布与演化特征。指出语言已成为 www 网络中社会与经济活动的自然边界, 形成了 www 网络中的语言群落和语言市场, 它们的发展状况和规模受到所处的相应社会经济水平的较强约束, 因此, 在企业通过网络进行全球化经营中语言的重要性也越来越显著。同时, 通过初步估计, 给出了规模在前几位的网络语言市场, 并论证了它们均服从幂律分布。

〔关键词〕 www 网络, 幂律分布, 网络语言群落, 网络语言市场

〔中图分类号〕 TP393

〔文献标识码〕 A

〔文章编号〕 1000-7857(2005)07-0040-04

Power Law of Social and Economical Indexes Based on Languages on the Web

NIE Rui¹, HUANG Chuan-feng^{1,2}

(1.School of Management, China University of Mining and Technology, Xuzhou 221008, Jiangsu Province, China;

2.Zaozhuang University, Zaozhuang 277100, Shandong Province, China)

Abstract: From a new angle of language, we studied the distributions and correlativity of the social and economical indexes on the Web. Based on the computation to the model of statistics, we found out that these indexes obey the power law approximatively on different scales, they are the reflections of the distributions and evolving characters of the social and economical phenomena on the Web. We pointed out that languages have been the nature boundaries of social and economical on the Web, and so, web language communities and web language markets have been shaped. Their evolutive status and size are restricted strongly by the real language economical level correspond to languages. At the same time, we have estimated the upper web language markets' size at present, and discovered that these markets' size obeys the power law, too.

Key Words: world wide web, power law, web languages community, web language market

CLC Number: TP393

Document Code: A

Article ID: 1000-7857(2005)07-0040-04

收稿日期: 2005-01-24

作者简介: 聂锐, 男; 江苏省徐州市解放南路中国矿业大学管理学院院长, 教授, 研究方向为管理科学理论与方法、战略管理、产业组织理论; E-mail: adean@cumt.edu.cn

4 结论

1) 通过对驯化后好氧颗粒污泥的 PAM 降解研究发现, 微生物对 PAM 的降解是有限度的, 即只能部分降解 PAM, 而不能实现 PAM 的完全矿化。

2) 研究表明, 采用好氧颗粒污泥处理油田含聚废水是可行的。

3) 驯化前后的微生物研究表明, 对 PAM 起主导降解作用的优势微生物与接种颗粒污泥内的优势微生物是完全不同的两种微生物, 对模拟含聚废水中 PAM 起主要降解作用的微生物经鉴定为产碱假单胞菌。

参考文献(References)

- [1] 程林波, 张鸿涛. 废水中聚丙烯酰胺的生物降解试验研究初探[J]. 环境保护, 2004, (1): 20~23
- [2] Jeanine L, Kay-Shoemaker, Mary E, et al. Polyacrylamide as an

organic nitrogen source for soil microorganisms with potential effects on inorganic soil nitrogen in agricultural soil [J]. *Soil Biol. Biochem.* 1998, 30(8/9): 1 045~1 052

- [3] Rachid El-Mamouni, Jean-Claude Frigon, Jalal Hawari, et al. Combining photolysis and bioprocesses for mineralization of high molecular weight polyacrylamides[J]. *Biodegradation*, 2002, 13(4): 221~227
- [4] 廖广志, 石梅, 李蔚, 等. 以部分水解聚丙烯酰胺和原油为营养源的微生物筛选及性能评价[J]. 石油学报, 2003, 24(6): 59~63
- [5] 雷乐成, 陈琳, 张瑞成. 油田含聚废水的 COD 构成分析及生物可降解性研究[J]. 给水排水, 2002, 28(6): 44~47
- [6] 杨云霞, 张晓健. 我国油田采出水处理回注的现状和技术发展[J]. 给水排水, 2000, 26(7): 32~35
- [7] 田利, 邹明珠, 许宏鼎. 采油污水中部分水解聚丙烯酰胺浓度的测定[J]. 吉林大学学报(理学版), 2003, 41(2): 224~227

(责任编辑 李慧政)

1 引言

信息技术的迅速普及和 www 网络的指数增长导致并形成了人类生存的新的活动空间——虚拟空间(Cyberspace),这一巨大空间有着与现实地理世界不同的结构和分布形态。对于 www 网络的结构、规模、连通性的测量方法及其增长和演化机理正在被人们从诸多方面进行着广泛的探索。

一般来说,当 Web 站点已经明显地成为社会和经济发展的的重要组成部分时,它们的规模、结构和内容可能反映出现实世界的人口、城市、网站等地理上的分布状况和社会经济活动的状态。Faloutsos 等人(1999)证明,尽管网站的出现是随机的,但站点的规模和它们之间的超链接以及 Web 的增长都服从幂律分布^[1];也有人从多个角度证明,互联网络和 www 网络都是无标度(scale-free)的复杂网络(Albert et al. 1999^[2], Huberman & Adamic 1999^[3], Scharnhorst 2003^[4])。Adamic & Huberman (2000) 通过对 www 网络中网站的分布、访问者数量的统计特征,得出网络交易市场是一个“赢者通吃”的市场的结论^[5];Shiode & Batty(2000) 则从国别的角度讨论了 Web 站点在现实世界中的地理分布,他们认为不同国家(域)的站点分布服从幂律分布,并为站点的分布与地理经济的关系提供了一个解释^[6]。

为了进一步研究 www 网络中经济活动的分布状况和特征,本文探讨以语言分类下 www 网络中的各指标的统计分布,并进一步分析了语言对 www 网络边界的影响和网络语言交易市场的形成问题。

2 数据来源与说明

本文所谓语言是指被应用于 Web 页面的现实生活中人们使用的自然语言。使用具有相同语言的 Web 页面的网民构成了一个虚拟空间中相互联系较为密切的社会群体。本文共选择了 35 种在 www 网络中已经被网民使用的语言,并相应统计了 www 网络上使用这些语言的网民人数、总人口、实际总 GDP、Web 域名数量和超链接数量等指标值(附录 A 略)。其中,网民人数、总人口源于 Global Reach (2004.9);GDP 数据源于 Nationmaster (2002) 和 CIA World Factbook (2002);Web 域名数量和超链接数量源于搜索引擎 AltaVista (1999)。

关于所使用的数据说明:① 每一语言的网民数量包括双语人群,也就是能同时以两种或两种以上语言上网的人,因此按语言累计的网民人数略大于实际网民人数;② 使用每一语言的总人口不是以这种语言为官方语言的国家的人口之和,而是根据不同国家内不同民族使用的第一语言分别计算;③ 由于难以将 GDP 在一国内按语言划分,因此,对于每一语言的 GDP 总值取近似,即等于以此语言为官方语言的国家的 GDP 之和;④ 域名数量是指由以该语言页面构成的 Web 站点数量,总的链接数量是指该语言网页中超级链接数量(包括入边链接和出边链接)。这两者的计算是借助于不同国家域名的不同,通过搜索引擎 AltaVista 对不同域名的 Web 站点数量、同一域内和域间超级链接数量进行搜索,然后再依照国家官方语言进行分类语言统计。

3 语言指标的幂律分布及解释

许多自然或者人文现象,包括城市规模、收入高低、词频、地震级别、站点规模、超级链接数量等,都遵守幂律分布

(Power-law) (Barabási 2001^[7], Huberman & Adamic 2002^[8])。所谓幂律分布是指小事件极其普遍,而大事件相当稀少;在不同场合下,它又被称为 Zipf 定律或是 Pareto 分布,是同一个事件在不同观察下的不同表现形式(Adamic 2000^[9])。其一般形式是: $P(r)=Kr^{-q}$, 这里, $P(r)$ 是事件 r 的规模 ($P(r)>P(r+1)$), q 是这个分布的参数, K 是比例常数,它在双对数坐标中表现为一直线。如果事件 r 依 λ 增长,则此事件的规模尺度为: $p(\lambda r)=K(\lambda r)^{-q}=\lambda^{-q}Kr^{-q}=\lambda^{-q}p(r)$ 。因此,事件分布形态与事件规模无关,其中,对于较成熟的系统,参数 q 一般取值范围在 1~3 之间,但对于较不成熟正在快速演化的系统, q 会较大。这个关系对任意 r 可用 $P(r)r^q=K$ 表示,它表示了系统处于稳定状态时规模与名次的关系式,该式隐含的一些稳定形态与增长过程是相容的。

在 www 网络中,基于语言的指标虽然数据总量较少(只有 35 种语言),但初步看来也是小规模事件较多,而大规模事件较少,且指标规模在较广泛的尺度内存在(见表 1)。因此,我们猜想它们与幂律分布相近。基于此,对这些数据(附录 A 略)做了幂律分布的回归分析与拟合。首先以降序对语言分类指标 Web 域名数量(Domain)、网民人口(Access)、实际人口(Population)、实际 GDP 和总的超级链接数量(Total Link)等数据进行了排序。然后,分别对指标进行 $\log[P(r)]$ 对 $\log[r]$ 的线性回归,这里, r 是某一指标的降序名次, $p(r)$ 为指标规模大小,参数 $\log K$ 和 q 是截距和斜率,分别得到拟合曲线: $\log P(r) = \log K - q \log r$, 相关的参数见表 2。最后,给出了这些数据的完全图形分析,在双对数坐标中画出这些分布(见图 1)。图中直线为幂律分布的拟合线,它们表现了数据间的联系,刻画了语言分类下各指标的理论和实际的排序与大小的关系。

表 1 Domains 指标的累积分布

Scale	频率	累积(%)
0	2	5.71
10 000	2	11.43
50 000	2	17.14
100 000	4	28.57
500 000	9	54.29
1 000 000	2	60.00
5 000 000	12	94.29
10 000 000	1	97.14
50 000 000	0	97.14
100 000 000	1	100.00

表 2 语言指标的幂律分布拟合的回归参数

分布	截距 $\log K$	斜率 $-q$	相关系数 r^2	置信水平 (95%)
域名数量	8.03	2.21	0.77	1.86E-11
超级链接数量	8.04	1.82	0.87	2.69E-15
网民人口	2.86	1.79	0.90	6.39E-18
GDP	4.58	1.85	0.89	6.01E-17
实际人口	3.55	1.79	0.82	1.03E-13

从上述图表可以看出以下 5 点。

1) 数据表明,在 www 网络中,根据网民/页面使用的语言不同,网民或网页已经被分割为分属于不同语言的集合,而且,这些语言集合数量很少(本文统计共有 35 个),我们称之为语言集群。因此,对于 www 网络中的经济活动而言,它们也是基于这些语言集群,故基于 www 网络的全球网络交易市场被语言分割成为数不多的、相应的交易市场,我们

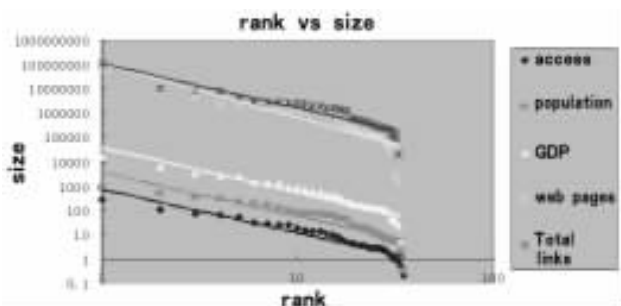


图 1 双对数坐标中语言指标的规模与排名关系

称之为语言市场，它们对应着现实世界中基于地理分割的交易市场(地理市场)。

2) 语言分布的各指标，尤其是前 30 项都非常好地符合了幂律分布。这表明在 www 网络中，几个应用最广泛的语言构成了网络中语言使用的主体，它们包含了 www 网络中绝大多数网民、网页、链接和经济活动(见表 3)。因此，企业通过 www 网络进行全球化经营时，需要将产品策略和营销策略的重点放在开拓前几位语言市场上。

表 3 各指标前 5 位语言的所占比重

网民人口	GDP	实际人口	域名数量	超级链接数量
73.2%	75.28%	63.45%	80.36%	81.5%

3) 由于数据集中许多指标(网民数量、GDP、现实总人口)都包括着能同时使用两种以上语言的网民，因此，语言市场的边界不是绝对的，而是模糊的、可跨越的。存在着许多可以同时出现在不同语言市场中的网民和企业，这是虚拟空间的交易主体的形态和交易过程的特征所决定的。

4) 由图 1 可以看出，这些分布中没有一个是完全的线性排序形式，因为它们都稍微凹向原点，每个指标首部(upper)和尾部(lower)却都在拟合线下方。事实上，当分别对首部和尾部区域的数据(前后各 5 项)作幂律分布拟合，也发现首部的幂律分布指数略微小于总体分布幂律指数，且置信度很高；相反，尾部的幂律指数却大大高于总体分布指数，并且可信度太低，甚至有些指标通不过检验；但同时对于尾部区域做 rank-size 线性回归分析的结果却是高度相关的，这表明尾部呈线性关系。这一现象可以用 Simon 模型^[10]来解释。① 新的事件或目标以一个规则但随机速率生成，并有最小规模；② 这些已经存在事件的增长率是完全随机的；③ 这个增长率独立于目标集的大小，但是，平均活动增长与大小成比例。随着事件数量增长，它们的分布收敛于一个稳定状态 $P(r) = Kr - q, q = 1/(1-s)$ ；这里 $s < 1$ ，是事件的增长率，在这个稳定状态，它趋于 0。这是一个非常有益的解：当增长率接近于 1 的时候，它意味高度线性相关，表示事件正处于早期不成熟的快速发展阶段。因此，对于语言分布模型而言，网络中前若干位的语言系统经过近 10 年的发展已经进入较稳定状态(q 趋向于 1)，增长率已经大大降低(接近于 0)，而后几位的语言则刚刚在网络中出现，增长率较高(接近于 1)。为了验证这一点，我们计算了不同语言的网民数量的年度增长率(1996-2004)，结果也证明了这一结论(附录 B 略)。

5) 对于处于稳定状态的系统来说，其幂律分布形式可以等价地写成： $p(r)r^{-q}=K$ 。因此，可以在 rank-size 图中看到 www 网络中不同语言离稳定状态有多远(即与理论值的偏

离大小)。基于以上分析，可以得出结论：网络中的各语言系统仍然处于增长阶段，但位于前部的大部分语言系统已经超越了指数增长阶段，正在趋于稳定状态；而后几位规模较小的语言系统却依然处于快速发展阶段。这一结论也可以被各语言的网民数量年度增长率部分得到验证。当然，这个稳定状态的临近也标志着 www 网络中以语言为边界的、由使用同一语言的网民和相关的各种社会经济机构构成新社会群落(Social Community)的状态趋于稳定，将之称为网络语言群落(Web Social Community)；而由基于同一网络语言群落的各种交易活动和交易主体构成的交易体系和环境，称之为网络语言市场(Web Language Market)。网络语言群落的稳定也预示着网络语言市场的逐步形成和完善，并将有一个较大发展。事实上，基于电子商务的全球网络交易的繁荣和蓬勃发展也证实了这一点。相信随着人类社会活动的逐步网络化，网络语言市场也将在更大范围内和更大规模上替代传统地理市场的主导地位。

4 网络语言市场的规模估计与分布

由于网络基础设施、教育及社会发展水平的差异，网络语言市场的发展也很不平衡。同时，全球电子商务虽然在近几年有迅猛发展，但由于各国和地区经济发展水平的限制，绝大多数国家尚没有形成电子商务发展状况的统计报告体系。全球性电子商务水平的数据也是极其缺乏的，现有的也大都粗略估计值，而且数值相差较大。因此，目前尚无法准确统计出各网络语言市场的总体规模和发展状况。但是，根据 Forrester Research(2001)的调查报告，经过处理，给出了前几位语言市场规模(电子商务交易总额)的近似估计值(见表 4)。然而，我们发现它们也是遵从了幂律分布，其中幂指数 $q=1.5$ ，相关系数 $R^2 = 0.9727$ (见图 2)。因此，它也从另一个角度证明了可以用幂律分布来解释 www 网络中语言群落和网络语言市场的规模、结构及其发展的许多现象。

表 4 前 9 位语言市场的交易规模

语言	排名	电子商务规模(亿美元)
English	1	3 845.7
Japanese	2	880.3
Spanish	3	488.8
German	4	386.5
Chinese	5	356.2
French	6	206.4
Korean	7	205.7
Italian	8	142.4
Dutch	9	98.3

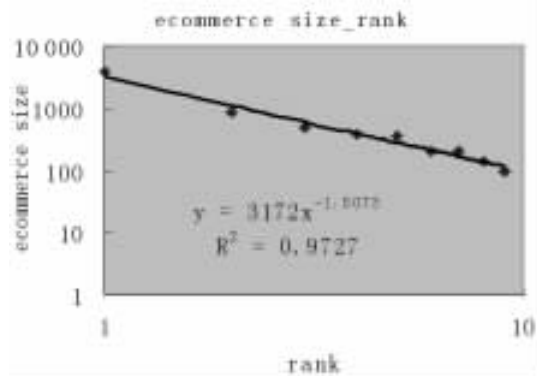


图 2 双对数坐标中的语言市场

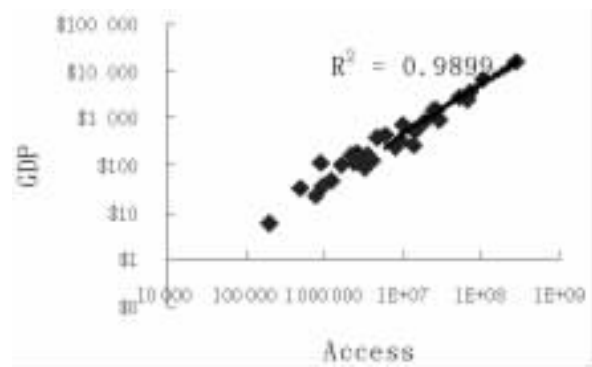


图3 GDP与Access的相关性分析

5 语言指标间的相关性

为了研究语言分布模型各指标之间的关系，以及对网络语言市场的影响，我们进一步考察了语言分布各指标间的相关性问题。在表5中给出了语言分布模型各指标间的相关系数值，图3表明了指标GDP与指标Access间极强的相关性。

表5 语言指标间的相关系数表

	网民人口	实际人口	GDP	域名数量	超级链接数量
网民人口	1				
实际人口	0.688	1			
GDP	0.995	0.709	1		
域名数量	0.927	0.427	0.922	1	
超级链接数量	0.930	0.441	0.922	0.996	1

由相关系数表可以看出：一个语言的网路发展水平与其现实的经济发展水平高度相关；网民数量与GDP水平和网页数量、总链接数量高度相关，与现实总人口数量显著相关；网页数量与GDP、网民数量和总链接高度相关；而现实总人口与现实GDP及网民数量显著相关，但与网页数量及总链接数量相关性并不显著；总链接数量与GDP、网页数量和网民数量高度相关。这说明网络语言群落和网络语言市场的发展总体水平是受到经济发展水平的强约束的，而与语言人口数量分布的关系相对较小，这也与www网络的发展现状相吻合。因此，一个企业要想通过www网络进入某一网络语言市场进行全球化经营，首先要考虑这一网络语言市场的经济的成熟度，而不能只看它的人口规模。

6 结论

1) 通过考察www网络中语言的各种指标可知：www网络中的网民和网页已经形成了以语言为界限的新群落结构，它们近似地服从幂律形式的分布。较大规模的网络语言群落的发展水平较为成熟，而较小规模的语言群落的发展则正处于快速增长状态。进一步比较语言指标之间的相互

关系则可以发现，不同网络语言群落以及网络语言市场的规模和发展水平与现实世界中的经济发展水平高度相关，而与现实中语言人口分布的相关程度并不显著。

2) 建立在网络语言群落结构之上的网络交易市场也被语言分割为不同的网络语言市场，它们之间有着不同的市场结构、规模、交易密度、发展状态、文化和消费习惯。同时，它们还有着许多不同于地理市场的特征：拥有相同或相近的文化背景的消费群体；市场主体的虚拟性，能够在多个语言市场同时存在；地理上的全球性、产品的多样性和较强的竞争性；营销层次较少；较高的信息透明度等等。因此，对于参与此类网络语言市场经营的企业来说，语言已经越来越成为全球化经营能否成功的重要因素，而不仅仅是在企业网站上增加几个多语言平台的问题了。

3) 由于语言及其指标的多样性和www网络发展与经济应用进程的短暂，导致了全球性数据统计的匮乏和获取的困难，因此本文的结论或许不能全面、准确地说明www网络中的语言问题，还需要对以下方面进行进一步的研究：数据统计方法的改进；时间序列数据的检验；语言群落的聚集系数等的结构刻画；网络语言市场的内部结构、特征以及语言在企业全球化经营中的影响与作用。

参考文献(References)

[1] Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the Internet topology [A]. In: ACM SIGCOMM'99 [C]. Cambridge, 1999, 251~262

[2] Albert R, Jeong H, Barabasi A L. Diameter of the World-Wide Web [J]. *Nature*, 1999, 401:130

[3] Huberman B A, Adamic L A. Growth dynamics of the World-Wide Web[J]. *Nature*, 1999, 40 : 450~457

[4] Scharnhorst A. Complex Networks and the Web: Insights From Nonlinear Physics [J]. *JCMC* 2003, 8 (4): <http://www.ascusc.org/jcmc/vol8/issue4/scharnhorst.html>

[5] Adamic L A, Huberman B A. The nature of markets in the World Wide Web [J]. *Quarterly Journal of Electronic Commerce*, 2000, 1(1): 5~12

[6] Shiode N, Batty M. Power Law Distributions in Real and Virtual Worlds [A] In: presented at INET 2000 [C]. Yokohama, Japan: 2000, http://www.isoc.org/inet2000/cdproceedings/2a/2a_2.htm

[7] Barabási A-L. The physics of the Web [J]. *Physics World*, 2001, 14(7): 33~38

[8] Huberman B A, Adamic L A. Zipf's law and the Internet [J]. *Glottometrics*, 2002, 3:143~150

[9] Lada A, Adamic. Zipf. Power-laws, and Pareto - a ranking tutorial [J]. Tech. Rep, Xerox Palo Alto Research Center, 2000. <http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html>

[10] Simon H A. Models of Man[M]. New York: John Wiley & Sons, 1957

(责任编辑 胡春华)

·读者·作者·编者·

本刊自本期起正文页码增加16页

为了适应我刊投稿数量日益增加的形势，进一步扩大我刊刊载容量，满足广大读者的阅读需求，我社自2005年第7期起，将《科技导报》正文由原来的4个印张扩充至5个印张，即增加16个页码。扩版后，每期《科技导报》论文的刊载容量将由过去的16万字增加至20万字，但刊物的定

价仍保持每册6元不变。按我社发展规划，自2006年第1期起，《科技导报》将进一步扩大刊载容量，在本期正文增加至80个页码的基础上，再增加1个印张，每期页码扩充至96页，为《科技导报》下一步改为半月刊创造条件。

《科技导报》编辑部