DOI:10.3969/j.issn.1673-4785.201304039
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130603.1601.008.html

互联网关注力的模型分析

仇建平
(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:随着基于互联网的社会媒体及其应用的迅速发展,互联网用户的“关注”对建立在其基础上的“虚拟经济”具有越来越重要的意义,通过把人看作是传播的内容,把信息资源看作是对象,互联网可以被看作是一个人类关注力在信息资源之间分配和流动的网络.利用搜集和分析互联网用户行为的数据,构建了基于互联网的关注力转移网络,给出了一个描述互联网用户关注力的动力学模型.实验结果表明,相比 Web 1.0 站点而言,Web 2.0 站点更受关注,网站的关注力增长小于流量增长,存在着“规模不经济”的现象,相比搜索引擎和广告联盟,广告网络和垂直网络更受关注.

关键词:互联网;虚拟经济;关注力模型;长尾分布
中图分类号:TP393.4 **文献标志码**:A **文章编号**:1673-4785(2013)04-0339-05

中文引用格式:仇建平.互联网关注力的模型分析[J].智能系统学报,2013, 8(4): 339-343.
英文引用格式:QIU Jianping. Analysis of attention model based on Internet[J]. CAAI Transactions on Intelligent Systems, 2013, 8(4): 339-343.

Analysis of attention model based on Internet

QIU Jianping
(Institute of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

Abstract: With rapid development of Internet-based social media and their applications, Internet users' attention has more and more important significance to a fictitious economy. In this paper, users are regarded as content conveyed in Internet while information resources are regarded as targets. Along this line, Internet is regarded as a net that distributes people's attention stream among information resources. By collecting and analyzing the behavior data of Internet users, this paper demonstrates a flow network of transporting attention and gives a dynamic model describing Internet users' attention. The experimental results show that, Web2.0 attracts more attention than the Web1.0 Website. The results of this paper also demonstrate that growth of attention to the website is slower than growth of traffic, and thereby creating a "diseconomies of scale" phenomenon. We were able to demonstrate that ad network and vertical niche attract more attention than search engine and affiliate network.

Keywords: Internet; fictitious economy; attention model; long-tailed distributions

互联网对人们生活和工作的影响越来越深入.随着信息传递速度和效率的大幅度提升,特别是社交媒体、电子商务和智慧终端的快速发展,人们越来越多地参与到了互联网上丰富的社会活动中,互联网用户的需求已经从单一、从众,向多样、个性、品位等方面进行转变[1].与此同时,遭遇到了 2 个突出的矛盾:1)可以获取的信息量爆炸性地增长和甄别与选择信息能力的局限性之间的矛盾;2)同时共现的信息量的极度丰富和关注力的局限性之间的矛盾.前者表现为人们难以从数万亿网页、数亿商品或数百万图书中高效率地找到自己喜欢或者需要的对像,后者表现为人们往往无法从以网页、视频、搜索引擎、邮箱和手机应用等为媒介或载体的广告中获得有价值的信息.

收稿日期:2013-04-16. **网络出版日期**:2013-06-03.
基金项目:山西省自然科学基金资助项目(2012011011-5);山西省重大科技专项基金资助项目(20121101001);山西省回国留学人员科研资助项目(2013-097);大学生创新创业专项基金资助项目(XJ2012007).
通信作者:仇建平. E-mail:choujianp@yahoo.com.cn.

在虚拟的互联网世界中,实体经济中的“消费者”转换为虚拟经济中的“用户”.“用户”是一个主动的概念[2],浏览网页、听音乐、看视频、下载软件等行为都是用户根据自己的兴趣主动做出的,是一种自由意志支配下的消费行为[3-4].随着这些行为数据的爆发性增长,新的模型不断涌现.一般的建模方法是从某些假设出发,引入变量,借助某些原理、定律,导出数学模型,再用数据来检验其正确性,有了数学模型,系统控制的问题就变成了变量和参数的控制.但要建立这样的数学模型有时非常困难,主要表现在:1)互联网结构复杂:节点数目巨大,网络结构呈现多种不同特征;2)互联网的进化:节点或链路会随着时间的变化而产生或消失;3)互联网链接的多样性:节点之间的链路权重存在差异,且有可能存在方向性;4)多重复杂性融合:即以上多重复杂性相互影响,导致更为难以预料的结果.实际的网络会受到多种因素的影响和作用,各种网络之间密切的联系也会使它们相互产生影响,从而加大对网络分析的难度.

1 背景及相关工作

为了解决这些问题,人们转换视角:把人看作是传播的内容,把信息资源看作是对象.同生态系统依赖能量流动,经济系统依赖货币流动一样,互联网依赖关注力流动.互联网上用户的点击行为实际上体现为关注力的流动,其中节点是信息资源,有权重的链路则指示着关注力的流动.这样,从信息在用户间的流动,转换为用户在信息间的流动,即把原来网络的节点变成了链路,原来的链路变成了节点.这一转换的好处在于:1)网络中的信息量是无限的,不易测量,而关注力相对于信息量来说却是有限的,易测量;2)由于信息资源可以被无限地复制,同一类信息资源不同内容与不同类信息资源,都是网络上不同的节点;3)用户的关注力是一个严格的守恒量.总的关注力是稀缺的,可变的就是其在信息资源上的分配和流动.在这种思路下,互联网可以被看作是一个人类集体关注力在信息资源之间分配和流动的网络,即关注力流网络.关注力流按生成内容可分为:1)UGC(user generated content)用户生成内容,生成网状关注力流;2)non-UGC 网站生成内容,生成树状关注力流.无论是树形结构还是网状结构,如果计算互联网用户关注力的分布,最后都会得到长尾分布[5-6].如图 1 所示,消费者的关注力在不同的区域相差很大,关注力集中的程度随着颜色的加深而增加.

图 1 关注力区域的分布

Fig.1 Distribution area of attention

1998 年 Watts[7] 和 Strogatz[8] 提出了小世界网络模型.他们分析了具有“小世界特征”的社会网络的演化过程,对小世界网络的聚类系数和最短路长度等进行了探讨,发现小世界网络模型的聚类系数比随机图模型要大得多,同时具有较小的平均最短路长度.2013 年 Barabasi[9] 研究表明,目前互联网大约有 1 万亿个文件,包括 140 亿个页面及其附带的图片、视频和其他文件,但其中绝大多数与其他页面或文件之间的联系并不紧密,而互联网中搜索引擎、门户网站等少数网页(主导节点)具有非常大的链接数.这些少数主导节点成为整个互联网相互联系的桥梁,从而使得用户最多只需 19 次点击即可到达任何一个网页.Barabasi 认为,互联网的这种“小世界”特性源于人性,即无论是在现实生活还是虚拟世界中,人类都喜欢群居.Barabasi 从多种水平对网络进行了分析后发现,无论规模有多大,“19 次点击”的规则仍然适用于互联网.

2 关注力模型

根据 2011 年底的 CNNIC 数据[10],我国互联网用户平均每周上网时间为 18 h,平均每天上网时间为 2.67 h,互联网用户在信息生产、交易和消费的所有环节都留下了数据记录,而且 90%的互联网用户仅仅访问网站,从不贡献内容,9%的互联网用户偶尔参与,只有 1%的互联网用户生产绝大多数内容,因此用户关注力相对于信息量来说是稀缺、可跟踪、可分析的.

本文中把互联网用户的关注力定义为关注某网站、同时忽略其他网站的选择性关注.一个互联网用户的关注力 X 取决于网站内容 n,并随着网站内容 n 线性变化,如式(1)所示[11].

$$X = anY. \tag{1}$$

式中:a 是正的常量系数,Y 是均值为 1 的噪声.

如果互联网用户的关注力超过了阈值 θ,那么

互联网用户将继续关注该网站,θ 为反映互联网用户个性化喜好的阈值,与网站内容给互联网用户带来的愉悦感、理想主义、归属感、增进自己的社会地位等因素相关.如果互联网用户的关注力小于 θ,则互联网用户的关注力转移到其他网站,其概率 P_n 如式(2)所示.

$$P(anY < \theta) = P, aY/\theta < 1/n. \tag{2}$$

式中:aY/θ 的累积分布函数为 F,则网站最终获得的关注力如式(3)所示:

$$F(1/n) = F(0) + F'(0)/n + O(1/n^2). \tag{3}$$

式中:$F'(0)$为一常数.

通过简化和省略高次项后,式(3)可重写为

$$G(n) = \sum_{i=0}^{\infty} P_{n+i},$$

取极限则可得式(4):

$$-G'(n)/G(n) = F'(0)/n. \tag{4}$$

根据式(4)可得:

$$G(n) \sim n^{-F'(0)},$$

$$P_n = -G'(n) \sim n^{-F'(0)-1}.$$

式中:P_n满足长尾分布,可得:

$$P_n \sim \frac{1}{n} + \left(\frac{n}{n_{\max}}\right)^{-k+1}.$$

式中:k>1,生成内容越多,互联网用户的关注力转移到其他网站的概率越低.用户关注力转移网络如图 2 所示 .

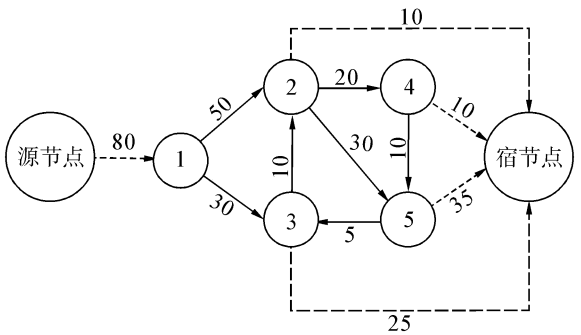

图 2　关注力转移网络

Fig.2　Network of transporting attention

图 2 中实线圆环代表网站 1 到网站 5,边代表用户关注力流,箭头指向代表关注力流动方向,边的权重(边上的数字)为从某网站转移到另一网站的人数,其转移矩阵为

$$\boldsymbol{A} = \begin{bmatrix} 0 & 50 & 30 & 0 & 0 \\ 0 & 0 & 0 & 20 & 30 \\ 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 \\ 0 & 0 & 5 & 0 & 0 \end{bmatrix}.$$

用户的关注力在网络上流动,由于关注力的守恒性,在模型中加入虚线圆环代表源节点和宿节点,使每个节点的关注力的进出相等,其用户转移矩阵扩充为

$$\boldsymbol{B} = \begin{bmatrix} 0 & 80 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 50 & 30 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 20 & 30 & 10 \\ 0 & 0 & 10 & 0 & 0 & 0 & 25 \\ 0 & 0 & 0 & 0 & 0 & 10 & 10 \\ 0 & 0 & 0 & 5 & 0 & 0 & 35 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

设 m_{ij}为用户关注力由 i 站点流向 j 站点的概率,

$$m_{ij} = X_{ij} / \sum_{k=1}^{n+1} X_{ik}, \forall i, j = 0, 1, \cdots, n.$$

式中:X_{ij}为由 i 站点转移到 j 站点的用户关注力,在本文中简化为转移人数,可得转移概率矩阵:

$$\boldsymbol{M} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{5}{8} & \frac{3}{8} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} \\ 0 & \frac{2}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{8} & 0 & 0 \end{bmatrix}.$$

网站 i 的流量 A_i 为

$$A_i = \sum_{k=1}^{n+1} X_{ik}, \forall i, j = 0, 1, \cdots, n.$$

关注力 $C_i = G_i \sum_{k=1}^{n} u_{ik}, \forall i, j = 0, 1, \cdots, n.$ $\boldsymbol{u}_{ij}$为式(5)的元素.

$$\boldsymbol{U} = \frac{1}{1 - \boldsymbol{M}} = 1 + \boldsymbol{M} + \cdots + \boldsymbol{M}^{\infty}. \tag{5}$$

图 2 对应的 $\boldsymbol{u}_{ij}$为

$$\boldsymbol{u}_{ij} = \begin{bmatrix} 1 & 1 & \frac{3}{4} & \frac{7}{16} & \frac{1}{4} & \frac{1}{2} \\ 0 & 1 & \frac{3}{4} & \frac{7}{16} & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & \frac{42}{41} & \frac{16}{7} & \frac{4}{14} & \frac{2}{28} \\ 0 & 0 & \frac{12}{41} & \frac{42}{41} & \frac{4}{41} & \frac{8}{41} \\ 0 & 0 & \frac{3}{164} & \frac{21}{329} & \frac{165}{164} & \frac{21}{41} \\ 0 & 0 & \frac{3}{82} & \frac{21}{164} & \frac{1}{82} & \frac{42}{41} \end{bmatrix}.$$

式中：$G_i=\dfrac{\sum_{j=1}^{n}X_{0j}u_{ij}}{u_{ii}}$，$X_{0j}$为从源节点流向 j 节点的关注流.

以图 2 网络为例，运算结果如图 3 所示，图中黑点为节点 2 的流量 $A_2=60$，关注力 $G_2=125$，通过最小二乘法获得 $\gamma=1.45>1$.

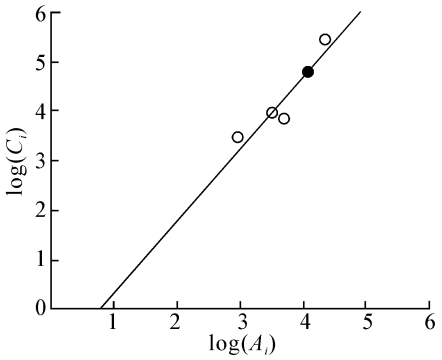

图 3 运算结果

Fig.3 Operation results

3 实验结果及分析

实验用到的数据是通过使用一个 Python 语言编写的网络爬虫程序，从谷歌广告计划（Google ad planner）[12]上获得世界排名前 1 000 的网站名单，使用 Alexa 分析这些网站间的关注流并构建网络，调用 AlchemyAPI 侦测网站类型.通过该方法获得的数据集比通过其他方法获得的数据集更稠密，如图 4 所示[12].

Upstream Sites
Which sites died users visit immediately preding youtube.com?

% of Unique Visits	Upstream Site
17.57%	google.com
13.84%	face book.com
2.13%	yahoo.com
1.72%	google.com.tr
1.48%	google.co.in
1.14%	wikipedia.org
1.08%	google.co.uk
0.96%	twitter.com
0.85%	google.de
0.85%	t.co

图 4 相关数据

Fig.4 Related data

图 5 展示了世界流量排名前 1 000 的网站构成的“关注力流”网络，其中圆形点代表网站，点的大小反映了取对数值后网站的日流量，点到中心的距离代表关注力的大小，即越靠近中心，则该点代表的网站受到的关注力越大.灰色圆形点表示 Web 2.0 站点，黑色圆形点表示 Web 1.0 站点.箭头表示用户的“关注力”在网站间的流动方向，一共12 888条，从图 5 可知，Web 2.0 站点更受关注.

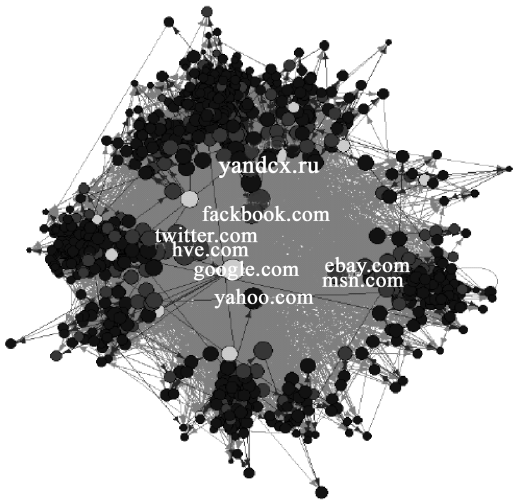

图 5 世界流量排名前 1 000 的网站构成的“关注力流”网络

Fig.5 Attention network of the 1 000 most-visited sites on the web

如图 6 所示，横轴是取对数值后的网站流量 A_i，纵轴为取对数值后的关注力 G_i，$\gamma=0.92$，γ 小于 1 表明用户对网站的关注力增长小于流量增长，存在着“规模不经济”的现象.

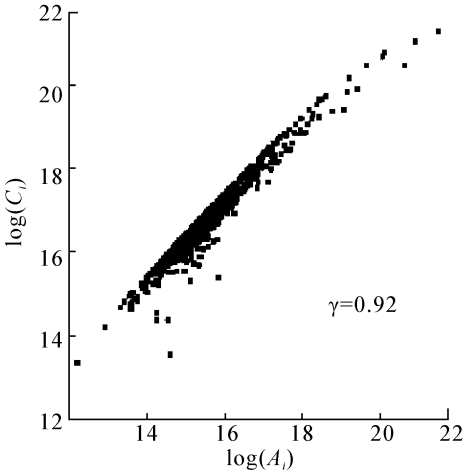

图 6 世界流量排名前 1 000 的网站流量分布

Fig.6 Attention stream of the 1 000 most-visited sites on the web

从流量来源来看，如图 7 所示，纵轴表示来源网站的类型，分别是广告网络（Ad network）、垂直网络（vertical niche）、门户网站（portal）、搜索引擎（search engine）、广告联盟（affiliate network）；横轴表示某个流量来源关注力的变动比例.黑色条块越往右表示某种流量来源关注力的提升，反之则表示关注力越低，黑色条块中的白线表示中位数.从图 7 中可以看出，搜索引擎和广告联盟所受的关注力较低，而广告网络和垂直网络则较高.

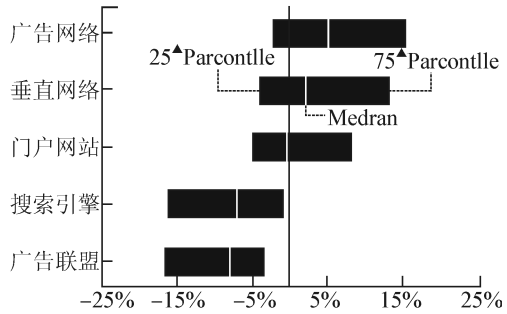

图 7　流量来源分析

Fig.7　Source analysis of stream

4　结束语

社会网络的快速发展带来了理论研究和实际应用上的巨大挑战,数据产生、组织和流通方式产生了革命性的变化,这些数据背后潜藏着巨大的商业机会.本文针对互联网的新发展,通过搜集互联网用户行为数据,推导了基于互联网的关注力动力模型,并通过实验进行了验证分析.本文作为一个探索性工作,初步勾勒了全球互联网用户“关注力流”的概貌,为更深入地探讨“虚拟经济”奠定了基础.

参考文献:

[1]苏萌,柏林森,周涛. 个性化:商业的未来[M]. 北京:机械工业出版社, 2012: 1-20.

[2]PAN W, AHARONYM N, PENTLAND A S. Composite social network for predicting mobile apps installation[C]// Proceedings of the 25th AAAI Conference on Artificial Intelligence. Cambridge, USA, 2011: 821-827.

[3]ZHANG C J, ZENG A. Behavior patterns of online users and the effect on information filtering[J]. Physica A, 2012, 391: 1822-1830.

[4]GUO S, WANG M, LESKOVEC J. The role of social networks in online shopping: information passing, price of trust, and consumer choice[C]//Proceedings of the 12th ACM Conference on Electronic Commerce. New York, USA, 2011: 157-166.

[5]HUBERMAN A, PIROLLI P L, PITKOW J E, et al. Strong regularities in world wide web surfing[J]. Science, 1998, 280(5360): 95-96.

[6]DENNIS M.WILKINSON. Strong regularities in online peer production[C]//Proceedings of the 9th ACM Conference on Electronic Commerce. Chicago, USA, 2008: 302-309.

[7]WATTS D. Network, dynamics, and the small-world phenomenon[J]. Sociol, 1999, 105: 2063-2064.

[8]STROGATZ S. The emerging science of spontaneous order [M]. New York, USA: Hyperion press, 2003: 312-319.

[9]BARABASI A L. Network science[J]. Philosophical Transactions of the Royal Society A, 2013, 371: 1471-2962.

[10]孟凡新. 互联网时代的眼球经济:中国网民注意力聚焦何处?[EB/OL]. [2012-10-25]. http://www.cnnic.cn/research/fxszl/fxswz/201207/t20120719_32346.html.

[11]ROBERTS J, HANN I H, SLAUGHTER S. Understanding the motivations, participation and performance of open source software developers: a longitudinal study of the apache projects[J]. Management Science, 2006, 52(7): 984-999.

[12]Google. The 1000 most-visited sites on the web[EB/OL]. [2012-10-25]. http://www.google.com/adplanner/static/top1000.

作者简介:

仇建平,男,1973 年生,讲师,主要研究方向为复杂网络、人工智能,承担山西省自然科学基金、山西省重大科技专项、山西省回国留学人员科研资助项目等项目多项,发表学术论文 8 篇,其中被 EI 检索 5 篇.