DOI:10.13878/j.cnki.jnuist.2018.04.003

佘滢[1]　刘芳[1]　范正平[1]

互联网信息特征研究综述

摘要

随着互联网的高速发展,网络上信息呈爆炸性增长.理解互联网上信息的特征对信息的快速获取具有重要意义.本文首先介绍了互联网信息载体的特征,然后分别按照互联网信息在时间上的演化特征、信息的大小特征、信息的内容特征以及信息的类型特征从不同的角度对互联网信息的特征进行了分析、归纳和总结,最后,对互联网信息特征潜在的应用给出了一个可能的实际例子.

关键词

互联网;万维网;信息特征;幂律分布

中图分类号 TP316

文献标志码 A

收稿日期 2018-06-08

资助项目 国家自然科学基金(61473321)

作者简介

范正平(通信作者),男,博士,副教授,主要研究方向为互联网建模及应用.fanzhp@mail.sysu.edu.cn

1　中山大学　数据科学与计算机学院,广州,510006

0　引言

随着互联网走进千家万户成为人们生活中不可或缺的一部分,绝大多数人不再只是网络信息的被动接收者,越来越多的政府部门、商家企业和互联网普通用户成为信息发布者,这在一定程度上加速了网络信息的增长.政府部门通过互联网打破部门间信息壁垒,减少了人工办理事务的繁琐流程,网络的透明化也有利于人民群众行使监督权,从而提高了行政办事的效率.商家企业通过互联网进行线上贸易和广告宣传,打破了地域的局限性,降低了经营成本.普通网民可以通过互联网在社交网络上发表看法,分享自己所见所闻所感,互联网已日益成为人们的精神文化家园.

著名未来学家托夫勒在《权力的转移》一书中提到:“世界已经告别了依靠暴力和金钱控制的时代,未来世界政治的魔方将控制在‘信息强权’手里.”信息化发展水平已成为各国经济发展和社会进步的重要驱动力.因此,在互联网高速发展的今天,研究网络信息的特征,并基于此特征为网络信息管理、网络信息分析与信息增长预测提供服务,已成为越来越多学者致力研究的重点.

1　互联网信息载体的特征

万维网是互联网上最重要的信息载体,网络上的信息呈现及表达绝大多数都是基于万维网的.1999 年 Albert 等[1]构建了一个 Robot,将在一个文件中发现的所有 URLs 添加到它的数据库中并递归地跟踪这些 URL 以检索相关的文档和 URL.经统计发现,搜索引擎大约覆盖网络 38%内容,更重要的是,当分析网页的链入和链出概率时,结果表明无论是链入还是链出概率都服从幂律分布,即少数网页有较大的链入/链出数,但大多数网页都只有很少的链入/链出数,这与随机网络中的二项式分布有显著的不同.

Huberman 等[2]通过 Alexa 和 Inforseek 爬取网站的规模分布数据,发现网站的规模的分布和网站中网页的分布满足幂律特征.

Border 等[3]借助 Altavista 收集超过 2 亿个页面和 15 亿链接发现页面的链入分布和链出分布都服从幂律分布.

2　互联网信息自身的特征

2.1　互联网信息随时间的演化特征

互联网中的信息近年来一直呈现出增长的趋势.Egghe 等[4]基于

20 个在线数据库分析发现互联网上科技类信息满足幂律分布增长,而社会科学与人文类信息则符合 Gompertz 增长分布.Bar-Ilan[5] 基于收集的 100 d 与“疯牛病”有关的新闻项目的信息,定量地分析了新闻组中热点主题的增长情况.Seetharam 等[6] 研究了 1950—1990 年间全世界食品科学与技术文献的变化趋势,发现其增长趋势可以很好地用 Gompertz 函数描述.

邱均平等[7] 归纳统计了中国互联网络信息中心发布的 26 次中国互联网发展统计报告和 5 次中国互联网络信息数量调查报告中的相关数据,从域名数、网站数和网页数 3 个层面深入分析了互联网信息的增长变化和分布情况.结果表明,1997 年至 2009 年底,CN 域名数增长情况可分为 4 个阶段,如图 1 所示.第 1 阶段为起步阶段,增速较缓,对应于图中的[0,10)部分;第 2 阶段为初步发展阶段,增速较快,对应于图中的[10,15)部分;第 3 阶段为蓬勃发展阶段,呈快速增势,对应于图中的[15,21)部分;第 4 阶段为发展稳定阶段,增长减缓,对应于图中的[21,24)部分.类似地,可使用指数模型来描述网页数的增长情况,如图 2 所示.对于互联网上信息随时间的增长趋势,可用乘数扩张模型来描述[8].该方法类似于银行货币扩张率,并且基于“信息转发假设”和“信息创新假设”来刻画互联网中的信息增长趋势.乘数扩张模型包括:

1)网络信息总量的增长模型:$I_m = n[(m-1)n+1]A$;

2)网上真实信息量的增长模型:$I_r = mnA$;

3)网上泡沫信息量的增长模型:$I_f = I_m - I_r = n(n-1)(m-1)A$;

4)网络信息的乘数扩张率:

$$\frac{I_m}{I_r} = \frac{n[(m-1)n+1]A}{mnA} = \frac{(m-1)n+1}{m} \approx n \text{ 以及}$$

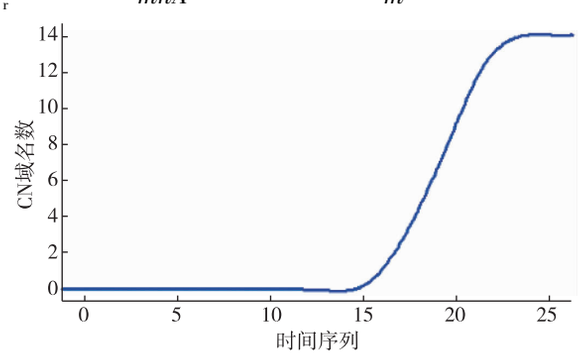

图 1 CN 域名数的逻辑曲线增长拟合[7]

Fig. 1 Logistic curve fitting for CN domain name growth[7]

$$\frac{I_f}{I_r} = \frac{n(n-1)(m-1)A}{mnA} = \frac{(m-1)(n-1)}{m} \approx n-1,$$

其中,I_m 表示网络信息总量,n 表示网络节点数,A 表示网络节点的平均信息创新能力,m 表示网络的信息转发次数.

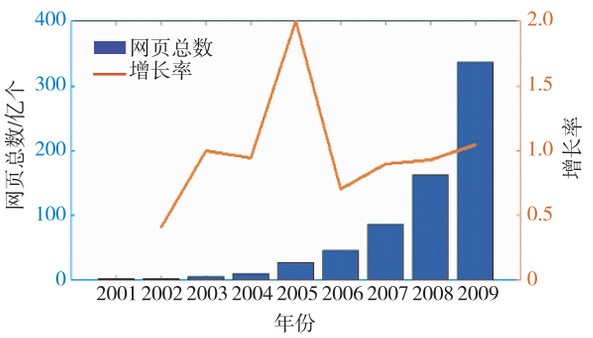

图 2 网页总数增长情况(2001—2009)[7]

Fig. 2 Growth of the total web pages during 2001-2009[7]

2.2 互联网信息在大小方面的特征

互联网中保存着海量的文件.2004 年,陈华等[9] 基于“天网”文件搜索引擎所搜集维护的 FTP 站点文件目录信息,考察了海量 FTP 文件和目录的分布特征.具体的研究方法是将“天网”收集的 1 000 多万个 FTP 文件和目录看作一个数据集合,因每个 FTP 文件具有文件数量、文件大小、文件类型、文件存放目录深度(即文件父目录的个数)和文件命名方式等属性,因此可用来统计分析这个数据集合各个属性所呈现的分布特征.对统计的 839 个服务器,共 14 587 171 个文件进行了分析,结果发现,FTP 服务器所含文件无论是大小还是数量,整体上都近似 Pareto(帕累托)分布,具有明显的重尾分布特征,如图 3、图 4 所示.

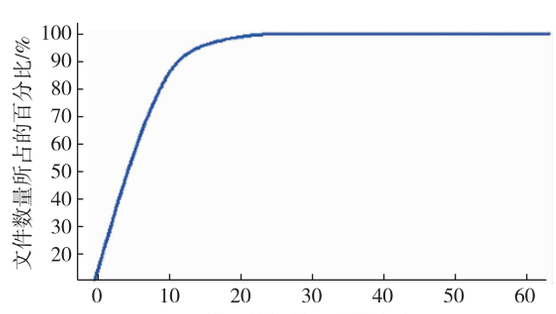

图 3 服务器与所含文件数量的关系[9]

Fig. 3 Relationship between servers and the included file number[9]

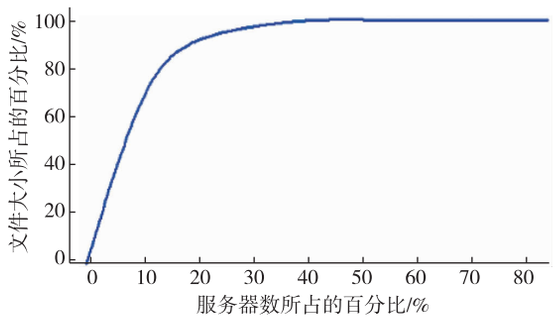

图 4 服务器与所含文件大小的关系[9]

Fig. 4 Relationship between servers and the included file size[9]

在研究文件大小的分布特征时,通过将一个文件的大小看作一个随机变量,并对小于 3 MB 的文件以 0.1 MB 为间隔单位进行频数统计,从而得到一个关于文件大小的累计分布函数,如图 5 所示.注意到:互联网上的 FTP 文件中的 90%以上的文件不到 1 MB,95%以上的文件不超过 2 MB.这启发我们在实际生活中,为了提高性能,减少大量的系统资源的耗费,可以对小文件进行一定的处理,例如把不能识别类型的且大小小于某个值的文件不建立索引.

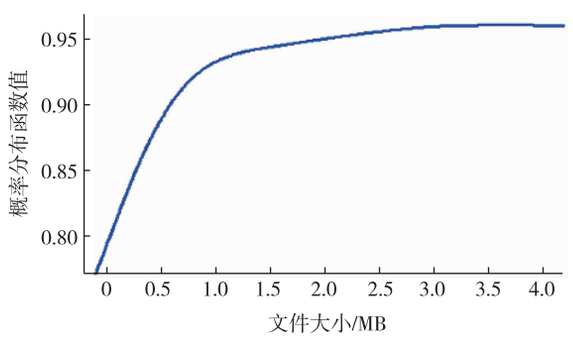

图 5 文件大小的分布函数[9]

Fig. 5 Distribution function of the file size[9]

如果定义文件目录的深度为文件父目录的个数,并定义根目录的层数为 0,以此类推.图 6 给出了不同目录层次下所含文件数量的概率分布.可以看出,文件搜索引擎可以只考虑搜索 15 层目录以下的文件,因为它们包括了 98%的文件.

2.3 互联网信息在内容方面的特征

互联网中的信息在维度上包括的信息十分广泛,如包含科技、娱乐、体育等各个方面的内容.

阎劲松[10]对仅存在正式交流过程的单一网站、对某一学科主题领域的网站和域层面的网站数量进行了研究.特别地,对“纳米科技”这一学科,分别以

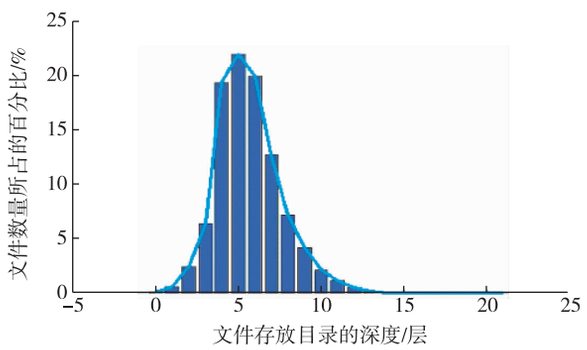

图 6 文件存放目录的深度[9]

Fig. 6 Depth of file storage directory[9]

“年”和“月”为单位统计了 1980—2006 年的网页数量.结果表明,当以“年”为尺度时,纳米科技领域的网页累积数随着时间呈指数增长.但这一指数曲线回归方程比较粗糙,对纳米科技兴起初期的网络增长情况刻画得不够细致,没有反映出后期网页累积数增长速度减小的趋势,而且样本数目少,代表性不强.因此,应以“月”为尺度对网页累积数进行统计研究.将观测期划分为 3 个阶段(不同阶段的分割点大致在第 30 个月和第 132 个月).结果发现,在观测期第 1 期内,以月为尺度的纳米科技领域的网页累积数会随时间推移,按逻辑曲线增长,即增长速度先加快后减慢,直到接近一个固定值.在观测期第 2 期内,以月为尺度的纳米科技领域的网页累积数会随时间推移,按指数曲线增长.在观测期第 3 期内,以月为尺度的纳米科技领域的网页累积数会随时间推移,按二次曲线增长.通过对以“月”为尺度的数据进行分析发现,这 3 个阶段的变化过程与纳米科技发展的实际情况相符,分别经历兴起期、快速发展期和成熟期.该结果可推广到一般的学科领域,在诞生期,该学科的网页累积数的增长服从逻辑曲线增长规律;随后进入发展期,相应的网页累积数呈指数增长趋势快速增长;当步入成熟期后,网页累积数增长又开始趋缓.

段运[11]使用 Altavista 搜索引擎进行关键字检索,结合信息计量学理论,以网络累积网页数据为依据,同样对纳米技术网页的增长规律进行了时间序列的统计研究,结果发现,在 1991—2008 年间,纳米技术网络学术信息较好地按照指数模型增长,基本符合普赖斯科学文献指数增长定律.

在知识管理主题领域,苏金燕等[12]利用 Altavista 搜索引擎采集数据后首先对网络学术信息的主要构成成分进行分析,分别以“年”和“月”为尺

度进行增长曲线拟合.结果发现,按“年”为尺度时,知识管理主题领域网络学术信息的增长模型为指数增长模型;按“月”为尺度时,该领域网络学术信息以三次方模型或者指数增长模型快速增长.

类似地,在竞争情报领域,张晋朝等[13]以每月网络信息增加量为研究依据,对特定时间域内检索出来的信息分别在不进行去重处理和进行去重处理后进行曲线拟合,发现指数曲线拟合效果较好.

在IT领域,邱均平等[14]利用搜索引擎Google统计分析了“PC显卡”这一内容的数量分布和变化情况.首先,在Google搜索引擎下对显卡有关内容进行分年检索,分析检测到的数据,结果表明:就增长速度而言,显卡相关网络内容在20世纪80年代以前都处于缓慢增长阶段,80年代初到90年代末处于稳定增长期,进入21世纪后,相关内容进入了飞速发展时期.就总量而言,直到1980年,当年的相关内容总量没达到总量的1%,1980年到2000年,单年的量没达到总量的10%,2000年后,总量上有了质的变化,都至少达到12%.进一步地,将显卡这个内容粗略地划分为6个方面:显卡技术、显卡新闻、显卡评测、显卡市场、显卡驱动和显卡历史.结果表明,这6个细分后的内容存在着基本相同的增长规律.6条数据点折线存在着基本一致的增长幅度、趋势和规律,但“显卡历史”这一内容主题在2000年到2003年基本没有变化,如图7所示.

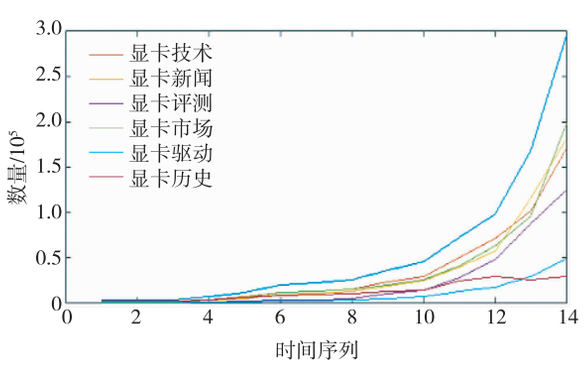

图7 显卡内容主题细分数据点折线[14]

Fig. 7 Literature quantity searched by Google on different subjects of graphics card[14]

总的说来,在6个细分类目中,“显卡技术”与“显卡市场”数量上相差不大,是属于数量最大的一类,“显卡新闻”与“显卡评测”次之,“显卡驱动”与“显卡历史”数量最少,这从一定程度上反映了网络用户的关注程度.

对显卡细分后的内容与其上一级内容的增长,也可通过统计在同一年代二者合计发表的论文数量来表示,如图8所示.可以看出,无论是对于显卡内容(合计文献量一)还是显卡内容的细分主题(合计文献量二),都存在基本一致的变化趋势,验证了“显卡”网络内容呈指数增长的规律.

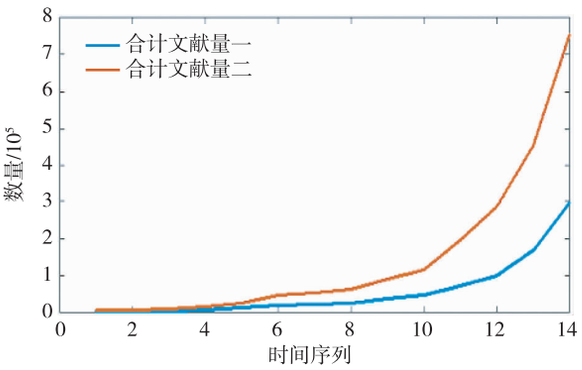

图8 合计文献量一、合计文献量二数据点折线[14]

Fig. 8 Literature quantity searched by Google on subject of graphics card[14]

2.4 互联网信息在不同文件类型方面的特征

互联网中的信息有不同的类型,如有的是文本文件,有的是视频文件等.因此可对互联网中的信息按照不同的文件类型来进行研究,我们课题组在这方面也做了一些工作.特别地,因互联网中视频文件占据着网络信息流的很大一部分,对其进行分析,有助于互联网服务提供商更好地配置文件缓存,从而减少网络的拥塞状况,提高用户体验.因此,这里只对视频文件的特征进行了研究.Youtube是世界上最大的UGC网站,提供1亿多个不同的视频且每天有65 000个上传量.Daum UCC是在韩国最流行的UGC服务,其最高流速为800 kb/s,每周访问量达1 500万次.通过抓取YouTube和Daum网站上的视频信息,统计了不同视频的时间长度分布.结果表明,视频文件时长分布类似帕累托分布,其中大多数视频文件时长在5 h以内,有超过1/3的视频文件时长不足1 h,如图9所示.

3 总结

互联网上的信息一直以来都呈现出爆炸性的增长,同时,网络上的内容多式多样,既包括科学技术类,也包括文学艺术类;网络中的信息文件无论是大小还是文件类型都存在非常大的差异.本文从不同的角度对互联网上信息的特征进行了分析和总结,

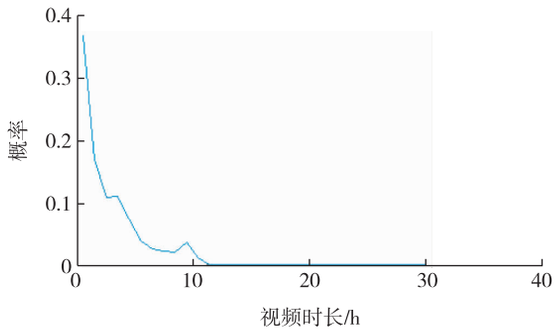

图 9 视频文件时长的概率分布特征

Fig. 9 Probability distribution characteristics of video file duration

结果表明,网络中的信息在不同的阶段有不同的增长模式,信息文件的大小基本上满足幂律分布特征.互联网上信息的这种特征在实际中有着重要的应用价值.例如近年来,内容分发网络受到了广泛的关注[15-17].在内容分发网络的缓存部署问题中,文件被访问的时间演化特征、文件的大小分布特征等对内容分发网络协议的性能有着重要的影响.利用信息的这些特征,可极大地提高内容分发网络协议的性能.目前对这方面的研究已有一部分工作,但只是刚刚起步,对互联网信息的特征是如何影响内容分发网络协议性能的机理目前还尚未清楚,需要继续进行深入的研究.

参考文献

References

[1] Albert R, Jeong H, Barabási A L. Internet: diameter of the World-Wide Web[J]. Nature, 1999, 401(6): 130-131

[2] Huberman B A, Adamic L A. Internet: growth dynamics of the World-Wide Web[J]. Nature, 1999, 401(16): 23-25

[3] Broder A, Kumar R, Maghoul F, et al. Graph structure in the Web[J]. Computer Networks, 2000, 33(1): 309-320

[4] Egghe L, Rao I K R. Classification of growth models based on growth rates and its applications[J]. Scientometrics, 1992, 25(1): 5-46

[5] Bar-Ilan J. The 'mad cow diseas', USENET newsgroups and bibliometric laws[J]. Scientometrics, 1997, 39(1): 29-55

[6] Seetharam G, Rao I K R. Growth of food science and technology literature: a comparison of CFTRI, India and the world[J]. Scientometrics, 1999, 44(1): 59-79

[7] 邱均平,马凤.我国网络信息的增长变化和分布状况研究[J].情报杂志,2011,30(3):1-7
QIU Junping, MA Feng. Study on growth and distribution of network information in China[J]. Journal of Intelligence, 2011, 30(3): 1-7

[8] 侯经川,赵蓉英.网络信息的增长机制研究[J].情报学报,2003,22(3):267-272
HOU Jingchuan, ZHAO Rongying. On the increasing mechanism of web information[J]. Journal of the China Society for Scientific and Technical Information, 2003, 22(3): 267-272

[9] 陈华,王继民,韩近强,等.互联网上 FTP 文件的分布特征及启示[J].计算机工程与应用,2004,40(1):129-133
CHEN Hua, WANG Jiming, HAN Jinqiang, et al. FTP files' distribution characteristics and their implications[J]. Computer Engineering and Applications, 2004, 40(1): 129-133

[10] 阎劲松.网络信息的增长规律研究[D].兰州:兰州大学管理学院,2007
YAN Jingsong. Study on the growth law of web information[D]. Lanzhou: School of Management, Lanzhou University, 2007

[11] 段运.纳米技术主题领域网络信息增长模型研究[J].中华医学图书情报杂志,2010,19(6):62-64
DUAN Yun. Growth model of network information on nanotechnology[J]. Chinese Journal of Medical Library and Information Science, 2010, 19(6): 62-64

[12] 苏金燕,周春雷,罗力.网络学术信息增长模型分析:以知识管理主题领域为例[J].情报杂志,2009,28(5):103-106
SU Jinyan, ZHOU Chunlei, LUO Li. Study on the growth law of web scholarly information in knowledge management field[J]. Journal of Intelligence, 2009, 28(5): 103-106

[13] 张晋朝,李改霞.网络知识增长的对数透视研究:以竞争情报为例[J].图书情报工作,2011,55(2):78-82
ZHANG Jinchao, LI Gaixia. Research on the logarithmic perspective of network knowledge growth: a case of competitive intelligence[J]. Library and Information Service, 2011, 55(2): 78-82

[14] 邱均平,殷之明.网络文献内容增长规律的实证研究:以 PC 显卡相关内容主题的增长为例[J].中国图书馆学报,2005,31(1):15-20
QIU Junping, YIN Zhiming. A study of the increase of network information contents[J]. Journal of Library Science in China, 2005, 31(1): 15-20

[15] Liu D Z, Huang K B. Mitigating interference in content delivery networks by spatial signal alignment: the approach of shot-noise ratio[J]. IEEE Transactions on Wireless Communications, 2018, 17(4): 2305-2318

[16] Haghighi A A, Heydari S S, Shahbazpanahi S. Dynamic QoS-aware resource assignment in cloud-based content-delivery networks[J]. IEEE Access, 2018, 6(99): 2298-2309

[17] Bottger T, Cuadrado F, Tyson G, et al. Open connect everywhere: a glimpse at the Internet ecosystem through the lens of the Netflix CDN[J]. ACM SIGCOMM Computer Communication Review, 2018, 48(1): 28-34

A survey of developments on the characteristics of the Internet information

SHE Ying[1] LIU Fang[1] FAN Zhengping[1]
1 School of Data and Computer Science,Sun Yat-Sen University,Guangzhou 510006

Abstract With the rapid development of the Internet,the information keeps explosive growth.It is clear that understanding the characteristics of the Internet information is important to,for example,fast acquisition of the information from the Internet.A survey on the development of the characteristics of the Internet information is presented in this paper.Firstly,the characteristic of the carrier of the Internet information is discussed.Secondly,the characteristics of the Internet information in terms of time evolution,information scale,information content,and types of information files,are analyzed and summarized.Finally,the possible application of the characteristics of the Internet information is discussed.

Key words Internet;World Wide Web;information characteristics;power-law distribution