第21卷第11期
2003年11月

情 报 科 学

Vol. 21, No. 11
November, 2003

·专 论·

网络信息资源的分布规律*

马费成 裴 雷

(武汉大学信息资源研究中心，武汉 430072)

摘 要 网络信息资源是一切投入互联网络的电子信息资源的统称。本文首先讨论了网络信息资源的概念，然后分别研究了网络信息资源的集中分散规律、生产规律和时间分布规律等。最后，本文还提出了网络信息资源研究面临的一些问题。

关键词 网络信息资源 分布规律

中图分类号 G250.72 **文献标识码** A **文章编号** 1007-7634 (2003) 11-1121-04

The Analyses of Distribution Laws of Networks Information Resources

Mai Feicheng Pei Lei

(Information Resource Research Center of Wuhan University, Wuhan430072)

Abstract Networks information resource is all the electrical information resources in networks. This thesis discussed the define of networks information resources. Then, it described the distribution law of information resource, the produce law and the time－based development law. At last, the thesis brings to light the problems to the study of networks information resources.

Keywords Network information resource Distribution laws

1 网络信息资源的概念

网络信息资源是一切投入互联网络的电子信息资源的统称，本文将网络信息资源界定为狭义的网络信息，并认为网络信息资源是可在计算机技术、通信技术及多媒体技术相互融合而形成的网络上发布、查询与存取利用的信息资源的总和。

从内容上看，在网络信息资源中，搜索引擎信息与网络数据库的信息占有很大比重。搜索引擎不提供完整信息，只提供全文的前200字作为索引，而且信息的更新周期不确定，有许多信息原文在网络上已经不存在，但它的索引仍然存在于搜索引擎；同样，网络数据库也有许多著录条目并不提供全文信息，只提供文摘或题录。同时，在形式上，网络互联使网页之间相互链接，彼此之间相关信息的查找十分方便。这样，在统计网站的信息总量时，信息的边界就相对模糊。统计中，主要存在网络链接与网络载文、网络标题与网络信息正文以及网络信息资源的信息质量等问题。此外，网络信息资源的研究缺乏统一的标准体系和有效的统计方式，所以网络信息资源的研究一般停留在定性方面。

但元数据库的研究表明，网络信息资源的定量研究已经具备了一定的理论基础和技术条件。许多机构都在定期或不定期地发布与网络信息有关的各类数据，如中国互联网络信息中心，为网络信息资源的研究提供了丰富的数据。本文在研究网络信息资源分布时，利用中国互联网络信息中心提供的数据，半定量地分析了网络信息资源的分布特征和规律。

2 网络信息资源的集中与分散规律

集中与分散的规律是科学文献分布的最普遍的规律，揭示这一规律最富盛名的成果就是布拉德福定律。国外近几年对网络信息资源分布的研究成果表明，网络条件下信息资源分布仍然满足集中与分散的规律，但是网络条件下的信息发布环境与传统期刊出版条件相比有很大不同，这种集中与分布的程度是否会有所变化？本文将采用布拉德福的方法处理和研究网络信息的分布。

通过一些工具性网站5月1日～5月3日中

* 本文为国家自然科学基金项目《网络计算环境下信息资源共享及其效率研究》(编号：70173018) 成果之一。

收稿日期：2003－09－20

“网络经济”这个条目命中的网页数量的集中与分散分布发现，其分布的近似曲线与布拉德福分布曲线比较接近（如图1所示）。前半段数据较少，表明在门户网站搜索引擎的信息搜集中，信息资源集中的网站比较集中，而且信息共享形成一些信息集中的网站群。同时，由于信息检全率有限，对于非核心

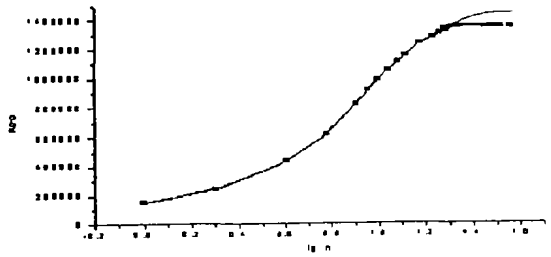

图1 以网络经济为主题的网页分布曲线

网站的信息统计数量明显不足，表现出的格鲁斯下垂格外明显。而在内容分析中，虽然网络经济的相关网页约150000篇，但根据搜狐的统计结果，相关网页超过10页的网络经济相关网站只有120个，超过20页的相关网站不过23个，最多的海脉咨询也不过2000页左右，大量相关网页分散在其它网站，网络信息分布更加分散。同时，对“远程教育”的研究也表明基本符合这一趋势。此外，网络条件下的信息资源分布不够稳定，各网站的信息丰裕程度和信息组织能力变化很快。比如新浪网的搜索引擎6月份升级以后，它的网络信息搜集能力大大超过同类网站，而且可能连续两天的搜索结果相差很大。但是笔者尝试在三个不同的时段对互联网进行统计，其分布图形基本一致。

在计量分析的过程中，虽然数据有限，而且许多门户网站使用同样的搜索引擎，搜索结果的形式也不太一样，但网络信息资源分布与文献分布的差异比较明显。搜索引擎的共享使许多网站都具有相同的信息搜索能力，像百度搜索引擎被100多个地方网站和30多家专业网站采用，对网络信息资源的分布影响很大。其次，摆脱了版面与经费限制，也使网络信息离散程度加剧。再次，信息审查同科学期刊的差异也影响了网络信息资源的分布。科学期刊都有严格的审查制度，而且科技文献发表有一定的成本，其结果必然会有一个均衡。而信息一旦进入零成本，低限制的发布环境，无疑将造成信息激增，同时分布更加分散。另外，用户对网络信息资源的浏览也是造成网络信息资源集中分散的重要原因，我们可以用实际数据进行分析。根据中国互联网络信息中心2001年7月的统计数据，平均每个商业网站每天的浏览量为5542个页面，是企业网站的10倍左右，是政府机构网站的6倍左右。从企业网站的浏览量在各行业分布看，零售批发贸易业的浏览量占24.1%，其次为电脑/通信设备/网络设备/软件业，占18.8%，再次为机械及工业制品占7.4%。科学研究和综合技术服务业及公关、咨询、广告和市场研究等服务业等的浏览量也分别占5.8%与4.7%。从浏览范围上看，大多数网络信息用户的浏览范围局限于少数几个核心网站。这种状况主要受搜索成本的影响。用户获取信息资源必须付出一定的搜索成本，而一旦得到，今后就可以在不付出任何搜索成本的条件下使用该信息资源。所以，今后用户不愿进行新的信息搜索，就形成了较高的用户忠诚度。另一方面，不管是资金、技术、人力资源还是市场机会，都集中在少数优秀的网络信息资源，导致马太效应特别突出，更使得网络信息用户的浏览范围局限于少数几个核心网站。核心网站具有良好的市场前景，许多的网站都愿意与这些网站链接，提高网站的信息网罗程度，提高网络信息资源的信息质量，更加吸引网络信息用户。

3 网络信息资源的时间分布规律

网络信息资源时间分布的突出表现就是信息资源随时间增长的老化规律。在考查网络信息资源的增长情况时，发现根据中国互联网络信息中心的历年域名注册统计，截止到2002年12月31日，我国CN下注册的域名数为179544个，与半年前相比增加了53398个，增长率为42.3%，与去年同期相比增长了41%，同1997年10月第一次调查相比，域名总数已是当初4066个的44.2倍。而我国WWW站点数为371600个，半年内增加78387个，增长率为26.7%，和去年同期相比增长34.1%。下面通过历次CN下注册的域名数的调查数据如图2所示。

从图2可以看出，图像的前半段与科学文献的逻辑增长曲线比较相似，整体上呈指数增长趋势。$t \in (4, 5)$时，CN域名增长出现第一个拐点，这说明CN域名经过3～4年的发展，已经逐渐被人们所认可，域名增长出现相对稳定的时期。同传统文献增长一样，文献增长先进入一个稳定期，然后又进入一个急剧增长的时期，这样交错出现几个急剧增长和几个相对稳定的时期。$t \in (8, 9)$时，CN域

名增长出现第二个拐点，体现了政策对 CN 域名注册服务的巨大推动作用，同时广大网民对 CN 域名价值的认可促进 CN 域名注册进入新的急剧增长时期。而波动的周期仅 30 个月——这反映了网络信息资源的高速波动增长趋势。所以，CN 域名的增长基本符合逻辑增长，CN 域名先出现一段高速增长，然后逐步趋于缓和，甚至回落，2002 年下半年又开始大幅回升。对这一涨落趋势，可以作出如下分析：过去的几年中，我国网络信息资源发展很快，经历了较长时间的持续高速增长。但随着计算机与网络拥有率逐渐增加，网络市场逐渐趋于成熟，市场增长空间相对减小，网络信息资源的增长进入一个相对稳定的阶段；另一方面我国网络经济的发展经过一个不正常发展后趋于冷静，许多网络公司相继倒闭或被兼并，整体上减少了网络信息资源的数量。去年，通过政策的进一步推动，网络信息资源又进入了一个高速增长的时期。这体现在图形上，就是一段缓慢增长甚至停滞后又开始一段高速增长。

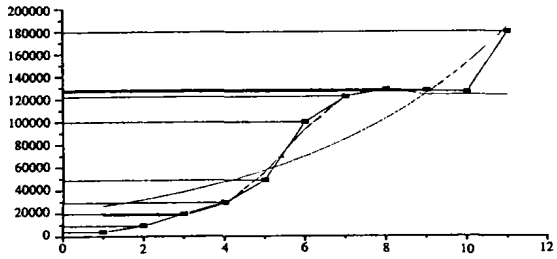

图 2　网络信息资源的增长图像

传统信息资源的老化规律一般用半衰期和普赖斯指数来界定，但对网络信息资源的考察，其时间参数肯定要有很大变化，至少普赖斯指数必须重新界定。在对网络信息资源的尝试性研究中发现，百度网络搜索引擎所包含的“网络经济”的条目从 5 月 3 日的 145000 条到 8 月 26 日的 291000 条新增一倍的时间只有 114 天。如果信息用户对网络信息资源的消费能力一定，那么必定由相当数量的网络信息资源得不到或很少被利用，也就是所谓的网络信息资源的老化。但由于网络信息资源的老化随网络信息资源的不同有不同的内涵，所以一般而言，网上新闻和公司及机构主页的老化可以从它的内容更新情况界定。科技文献等研究性信息资源的老化可以从它的浏览和链接情况界定，而网络数据库既要通过内容更新考察资源优化情况，又要通过被利用的情况考察内容的老化情况。

网页或数据库记录更新速度越快说明网络信息资源的老化速度越快，网络信息资源的质量越好。根据中国互联网络信息中心 2001 年 7 月的统计数据，网页更新时间在一个月以上的高达 88.10%，而更新时间在一周以内的仅占 6.89%，更新时间在一周到一个月之间的占 5.01%。从数据库更新情况看，虽然有超过 35%的网站每周更新记录超过 10%，但也有 35%的网站每周只更新 1～5%的记录，还有超过 20%的网站每周对记录不作更新或更新比例在 1%以下。用户最常使用的产品数据库，每周更新记录在 1～5%之间的占 41.1%，10～20%之间占 22.1%，每周更新记录超过 10%的网站还不足 35%。以上数据说明我国网络信息资源普遍存在质量较低的特点。而网络信息资源链接或引用少说明网络信息资源已经老化，反之，说明网络信息资源正被广泛利用。网络信息资源的链接依据网上网页之间的相互链接或再链接情况，引用则通过传统科技文献的引文中网络信息资源的引用情况。根据中国互联网络信息中心 2001 年 7 月的统计数据，没有进行过任何网站链接的网站占 47.8%，在进行链接的网站中，链接 2～5 家的比例最大，为 22.7%，其次为链接 6～10 家，占 11%，另有极少部分网站链接超过了 20 家。虽然近年传统文献对网络信息资源的引用情况呈上升势头，但总体上说，网络信息资源的被引率不高。

综合上面的分析，网络信息资源的时间分布的基本规律一是高速增长，并呈逻辑增长态势；二是老化速度很快，但仍旧具有价值。

4　网络信息资源生产者的分布规律

这里的网络信息生产者主要指在网上生产和发布信息的个人，包括一般信息生产者和机器作者。迄今研究信息生产分布规律最有代表性的就是科学文献生产的普赖斯定律和洛特卡定律。普赖斯定律认为，全体科学工作者人数的平方根撰写了所有科学论文的一半；洛特卡定律被称为科学生产率的平方反比分布，即撰写 n 篇论文的作者数是写 1 篇论文作者数的 n^2 分之一。网络信息生产完全不同于科学生产，大量的非科研人员参与信息的生产，成分复杂、目的多样，科学生产率已经丧失了它原来的评价意义。即便如此，是否可以采用普赖斯和洛特卡的方法来研究网络信息生产者的分布动态，是否也存在科学文献生产中所谓的“核心生产者”呢？

由于网络信息生产的复杂性，加之信息组织方面的滞后，很难通过整个网络来考察网络信息的生产者规律，相对而言校园BBS提供了一些可供研究的数据。由于网络信息尤其是BBS类信息更新频繁，我们只能考察某一短暂时段中的信息生产状态。下面通过珞珈山水BBS中5月5日的发文篇数进行统计。5月5日发文总量3307篇，作者1171人，而其发文情况如下：

表1 5月5日珞珈山水BBS部分发文情况

发文篇数	作者数	累积篇数	累积作者数	%	数据点斜率
77	1	77	1	0.085	—
64	1	141	2	0.171	−3.848
58	1	199	3	0.256	−4.093
50	1	249	4	0.342	−1.953
41	3	372	7	0.598	−2.816
38	1	410	8	0.683	−1.748
37	1	447	9	0.767	−4.344
31	1	478	10	0.854	−6.081
28	1	506	11	0.939	−0.932
26	2	558	13	1.11	−2.545
25	1	583	14	1.196	−1.895
23	1	606	15	1.281	−0.823
21	3	669	18	1.537	−2.005
20	2	709	20	1.708	−2.160
19	2	747	22	1.879	−1.865
18	2	783	24	2.05	−1.613
17	4	851	28	2.391	−2.683
16	3	899	31	2.647	−1.681
15	3	944	34	2.904	−1.675
14	4	1000	38	3.245	−1.607
12	5	1060	43	3.672	−0.803
11	8	1148	51	4.355	−1.960
10	7	1218	58	4.953	−1.326
9	12	1326	70	5.978	−1.784
8	8	1390	78	6.661	−0.918
7	16	1502	94	8.027	−1，398
6	6+	1538+	100+	8.54+	—

注：6+表示由于不完全统计，实际数量将大于显示数据。

从上表可以看出，按普赖斯定律全部生产人员的开根号所得人数生产了全部信息的一半。5月5日的发文人数为1171，开根号得34.2，而实际上前34人发文只有944篇，不足全部发文（3307篇）的1/3。所以，BBS条件下的信息生产是不符合普赖斯定律的，也就是说，BBS中信息生产者的分布不如科学文献的生产集中。

同样，按照洛特卡分布曲线绘制方法将原来的数据取对数坐标转化为分布曲线，发现它也近似一条直线，而且它的斜率约为−2.18。通过各数据点的斜率分析也可以看出，斜率与科学文献生产分布的−2相比，变动性更大，但基本维持在−2上下浮动。

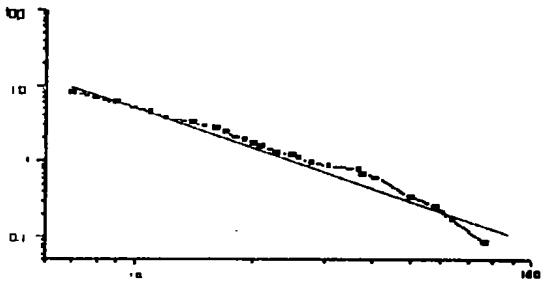

图3 网络信息生产者分布曲线

当然，上面分析的BBS的情况可能并不能代表网络信息资源生产者的频率分布，而且所选数据也不够完整。但是，一个新的结论不可忽视：著者的集中程度比传统文献低，著者分布更加广泛而且相对均衡。这一点与蔡明月教授的网络条件下网页分布的结论恰恰相反。蔡明月教授在分析我国台湾地区“数字图书馆”的网页内容分布时，发现核心区（总数开根号所得的数目）的网页制作单位生产的网页数量远远大于网页数量的一半，而我们从BBS的分析中发现核心区的信息生产者所生产的信息远远不足网络信息总量的一半。可见网络条件下的信息生产表现出不同的规律。

5 结 语

与其他信息资源相比，网络信息资源内容丰富，在传统信息资源的基础上实现了多媒体信息资源与文献信息资源的融合；网络信息资源具有开放性和互动性，信息来源广泛，信息传递及时，为集体异地研究提供了有效的信息平台；网络信息传播速度快，基本克服了传统信息资源环境下的传输时滞问题；网络信息资源可以使用自然语言检索，便于信息资源的获取和信息利用的普及；网络信息内容具有很强的关联性，便于系统查找相关全文资料；利用网络信息资源易于形成专家库资源，便于与同行有效沟通，集聚社会智力成果。但是，网络信息资源的信息组织方式与分布特征也给网络信息资源研究带来许多不便。网络信息资源数量巨大，难以保证网络信息资源研究中所搜集的信息资源的完整性；网络异构性与无序组织，难以形成统一的网络信息分类标准；信息非线性组织，难以统计分析网络信息资源的分布状况；信息分布式存储，难以控制管理新增与老化的网络信息资（下转第1169页）

等方式保持着与客户的沟通状态。此外，产业分析栏目中的企业排名机制，通过某一行业企业的排名及相关信息分析，便于发展新客户。

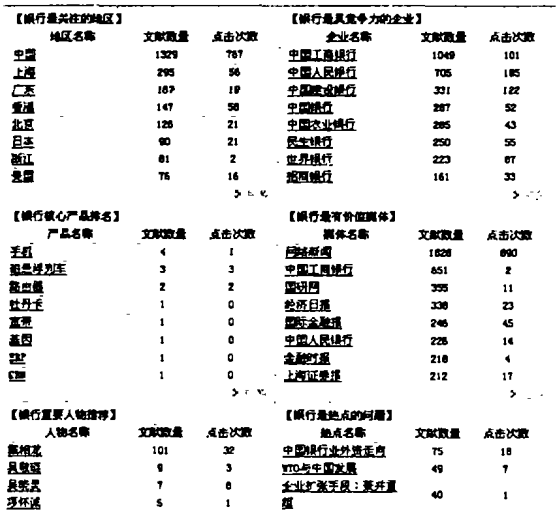

图2　对手分析

3.5　发掘新的投资领域、衍生金融产品

新政策的出台或行业结构的变化往往会带来新的投资方向和对新金融产品的需求。银行企业要把握商机就必须加强对政策信息的分析和对其他行业信息的研究。比如，康凯始终把握信息的行业和地区属性，通过对相关数据的研究，可以发掘最有价值的产业和投资最热门的地区，从而为银行企业开辟新的投资领域指引方向。另外，通过观测行业动向（包括金融业和其他相关行业），可为银行企业开发金融产品提供技术信息和策略支持。

4　结　　语

WTO对中国银行业的冲击是毋庸置疑的，但我们不能以“悲观”二字一概论之。毕竟在新的知识经济环境下，无论是国资银行，还是在华外资银行，乃至外国银行，大家都处于企业知识开发利用的同一起跑线上，面对激烈的行业竞争，成败的关键在于能否把握知识经济的脉搏，用竞争情报这一利器寻找契机，开辟一片新的天地。

参考文献

1　包昌火，谢新洲著. 竞争情报与企业竞争力. 北京：华夏出版社，2001. 4
2　Michael E. Porter 著，陈小悦译. 竞争战略. 北京：华夏出版社，2002. 4
3　邱均平，马海群. 论知识管理与信息管理. 中国图书馆学报，1999 (6)
4　任海平，王延飞，等. 知识经济与情报研究. 情报学报，2001，20 (5)：592
5　中国竞争情报网. http：//www. chinaci. com
6　北京康凯信息咨询公司. http：//211. 99. 143. 151
7　FULD&COMPANY，INC，http：//www. fuld. com/

（责任编辑：徐　波）

（上接第1124页）

源；信息动态发布，难以对每个网络信息资源的基本单元——网站实时监控；文本链接复杂，难以理清网络信息资源之间错综复杂的关系；网络安全差，信息污染严重，对网络信息资源分布研究的结果影响很大。此外，还存在网络信息资源内容的差异性与内容的交叉引用导致研究的困难。网络条件下，信息资源的知识产权保护突破了原有的知识产权法律体系，在很多领域存在侵权或高频率转载事件。同时，对著者或信息生产者没有专稿专投的限制，而且技术方面的网络备份与网络镜像技术加剧了这种现象，出现了同一文章在不同的网络地址下，甚至不同时间的网站记录中重复命中。网络信息资源的交叉分布更加大了网络信息资源分布的研究难度。

参考文献

1　马费成，李　纲，查先进. 信息资源管理. 武汉：武汉大学出版社，2000
2　焦玉英，符绍宏，何绍华. 信息检索. 武汉：武汉大学出版社，2000
3　邱均平. 文献计量学. 北京：科学技术文献出版社，1998
4　王翠萍. 我国网络信息资源分布. 情报科学，2002 (7)
5　王曰芬，甘利人，李　媚. 网上信息资源分布及查询案例分析. 情报科学，1999 (1)
6　蔡明月. 资讯计量与网路计量. 新世纪图书馆，2003 (2)
7　安新颖. 网络信息组织研究. 现代情报，2003 (2)
8　王　林. 网络条件下的信息资源与信息资源组织建设. 现代情报，2003 (2)
9　徐仁杏. 浅谈网络信息资源与DC元数据. 现代情报，2003，(4)
10　http：//www. cnnic. net. cn/develst
11　http：//www. cnnic. net. cn/+j/rep2001. shtml
12　http：//www. cnnic. net. cn/tj/repll. shtml
13　http：//www. cnnic. net. cn/mapinfo. rep2001－11/

（责任编辑：赵立军）