●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

摘 要 以PC显卡主题内容为研究对象,利用Google检索并统计分析网络主题内容数量分布和变化情况,可得出内容呈指数增长的基本判断。利用SPSS工具分别就数据进行拟合分析,可求得指数方程。统计分析中出现的误差,是有其原因的。图3。表4。参考文献3。

关键词 网络文献内容 增长规律 指数增长 网络文献计量学
分类号 G350

ABSTRACT With "PC display card" as the research object and using Google to search network subject contents and get statistical data, the authors get the result that network contents grow exponentially. Then, they derive an exponential equation by using the SPSS tool. 3 figs. 4 tabs. 3 refs.

KEY WORDS Network information contents. Growth law. Exponential growth. Network informetrics.
CLASS NUMBER G350

在互联网飞速发展的今天,如何研究网络文献内容主题的增长规律,并利用其规律为网络信息管理服务,为网络信息分析与预测服务,已成为网络计量学研究重点之一。在普赖斯的研究中,他通过对相关专著和论文的分别统计,然后从数据的增长、变化中发现规律,进而经过分析、预测、得出结论。本文首先考察网络内容主题相关网站(网页)数量按年代的增长规律,进而对相关内容主题经过粗略的细分,通过比较法考察细分的内容主题与上级内容主题的增长关系,来考察、验证网络内容主题的增长规律是否真实可靠,最后得出网络文献内容主题按指数规律增长的结论。

1 数据收集的工具、范围与方法

1.1 搜索引擎、方法的选择

虽然有些学者指出了使用商业搜索引擎进行链接分析的种种弊端,但在目前还没有其他可用的工具的情况下,我们不得不使用商业搜索引擎来进行搜索统计。经过对常用搜索引擎的研究,我们选取Google作为本例搜索统计数据来源。此外,考虑到互联网的发展状况以及计算机显卡技术发展的实际情况,本例在统计期选取时,1990年前以10年为一个单位直至1950年,1995年后是个人计算机及互联网飞速发展的时期,故逐年进行检索、统计分析。

1.2 关于显卡网络内容主题的统计

利用搜索引擎Google,对显卡有关内容进行分年检索,结果如表1。检索方法是利用Google的高级检索方法,分别输入网卡,1950年;网卡,1960年;……网卡,2003年。以1996年为例:进入http://www.google.com后选择高级检索,在"包含以下全部的字词"后填入"网卡",在"包含以下的完整字句"后填入"1996年",其他的限定条件用Google的默认值。

表1 对显卡有关内容的检索结果

年代分布	合计文献数量	环比指数	相对1995年的值	占总量(%)
1950	3080	100	15.7	0.38
1960	3050	99	15.6	0.38
1970	3620	119	18.5	0.45

* 本文为教育部人文社科重点基地重大项目(编号02JAZJD870003)"文献计量与内容分析的综合研究"的研究成果之一。

续表

年代分布	合计文献数量	环比指数	相对1995年的值	占总量(%)
1980	6700	185	34.2	0.82
1990	11400	170	58.2	1.4
1995	19600	172	100	2.4
1996	22300	114	113.8	2.7
1997	26000	117	132.7	3.2
1998	36400	140	185.7	4.5
1999	45600	125	232.7	5.6
2000	72200	158	368.4	8.9
2001	98200	136	501	12.1
2002	169000	172	862.2	20.8
2003	296000	151	1510.2	36.4
合计	813150			100

检索时间:2003年12月22日22:30—22:40

由表1可见,就增长速度而言,网卡相关的网络内容在20世纪80年代以前,都处于缓慢增长阶段,80年代初开始到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甚至没达到统计总量的1%,而1980年到2000年,单年的量依然没能达到总量的10%,2000年后,总量上有了质的变化,都达到和超过了12%。

1.3 关于显卡细分网络内容主题的统计

关于显卡这个内容主题,我们将其粗略的划分为6个方面即显卡技术、显卡新闻、显卡评测、显卡市场、显卡驱动与显卡历史。利用搜索引擎Google,就显卡有关内容进行分年检索,结果如表2。检索方法是利用Google的高级检索方法,分别输入网卡技术,1950年;网卡,1960年;……网卡,2003年。以1996年的显卡技术为例:进入http://www.google.com后选择高级检索,在"包含以下全部的字词"后填入"网卡技术",在"包含以下的完整字句"后填入"1996年",其他的限定条件用Google的默认值。

表2 显卡细分网络内容主题统计结果

年代	显卡技术	显卡新闻	显卡评测	显卡市场	显卡驱动	显卡历史	合计
1950	1640	1340	292	1250	261	2060	6843
1960	1530	1320	316	1210	237	1870	6483
1970	2050	1610	677	1660	347	2020	8364
1980	3660	3010	1050	3070	720	3750	15260
1990	6480	4990	1900	5630	1230	5460	25690
1995	11900	9640	3510	11300	2360	8530	47240
1996	13200	11000	3960	12800	2780	9340	53080
1997	15200	13100	4750	15200	3260	10200	61710
1998	23800	19100	10400	20400	4720	12300	90720
1999	28600	25000	14000	25500	7540	14300	114940
2000	49700	40100	28000	40900	12300	24200	195200

续表

年代	显卡技术	显卡新闻	显卡评测	显卡市场	显卡驱动	显卡历史	合计
2001	70800	56900	48000	62300	17800	28900	284700
2002	101000	115000	87200	95500	29400	25400	453500
2003	171000	182000	124000	199000	49000	28800	753800
合计	500560	380610	328055	495720	131955	177130	

检索时间：2003 年 12 月 22 日 22:45—23:30

从表 2 可以看出，和显卡内容主题相比，其 6 个分类主题的在相对应的统计年限中，存在着基本相同的增长规律，即基本都是 1960 年比 1950 年有个轻微的下降后都经历平稳、快速、飞速发展三个阶段，当然也存在 1960 年显卡评测的内容主题不降反升和 1970 年显卡历史的内容主题依然没超过 1950 年的两个特例。

2　数据处理与原因分析

2.1　合计数据处理及分析

本例在对数据进行分析时，是以统计期（统计单位）进行分析。即实际分析的是网络内容在统计期（统计单位）内的增长规律。

表 3　显卡内容主题数据

年代分布	合计文献量一	合计文献量二	年代分布	合计文献量一	合计文献量二
1950	3080	6843	1997	26000	61710
1960	3050	6483	1998	36400	90720
1970	3620	8364	1999	45600	114940
1980	6700	15260	2000	72200	195200
1990	11400	25690	2001	98200	284700
1995	19600	47240	2002	169000	453500
1996	22300	53080	2003	296000	753800

说明：合计文献量一为显卡内容主题的统计量，合计文献量二为细分的显卡内容主题统计合计量

仅仅从表 3 的数据变化我们可以看出，就总量而言，无论对于显卡内容主题还是显卡内容主题细分的统计和都存在基本一致的变化趋势。以年代为坐标横轴，利用表 3 的数据，绘制合计文献量一、合计文献量二数据绘制数据点折线图如图 1 所示。

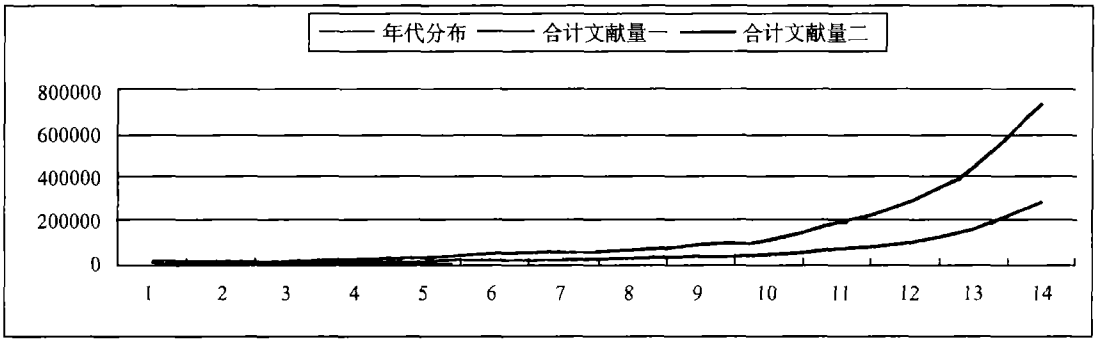

图 1　合计文献量一、合计文献量二数据点折线图

从数据折线图可以看出，两条折线从 1950 到 1980 数据实现翻番，但就整体数据量的变化而言，出于缓慢增长阶段。从 1981 到 1999 年，两条折线不再表现为前一阶段的重叠，而是逐渐分离。在此阶段，

合计文献量一从1980年的6700增长到1999年的45600,合计文献量二从1980年的15260增长到1999年的114940,前者增长为原来的6.8倍,后者增长为原来的7.5倍。从2000年到2003年,折线差距进一步拉大并在2003年出现不正常的差距。在此期间,合计文献量一从1999年的45600增长为2003年的296000,合计文献量二从1999年的114940增长为2003年的753800,两者基本都增长为原来的6.5倍。

从上面的数据分析以及我们对图形的基本判断中,可以初步认为,"显卡"网络内容呈指数增长规律。下面我们利用Spss11.5 for Windows英文版统计分析软件对两组数据对我们的判断进行拟合分析。选取Spss曲线回归中Exponential形式对两组数据分别进行拟合。得出合计文献量一的增长指数方程为:$Y = 1727.72^{*} e^{0.349t}$,其复相关系数为0.99155;合计文献量二的指数增长方程为:$Y = 3671.75^{*} e^{0.367t}$,其复相关系数为0.99297(Y代表对应的合计文献量,t代表年1950为1,1960为2,……2003为14)。两个复相关系数都达到了0.99的水平,说明指数增长方程对其实际增长的拟合效果非常好;其增长指数一个为0.349,一个为0.367,说明其增长趋势基本一致(增长率相差不到2%),且增长非常迅速,每个检索统计期的增长率都在35%左右。

2.2 细分数据处理及分析

依据表2的相关数据我们绘制图2(显卡内容主题细分数据点折线图)。从图2中我们可以看出,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,在数据折线图上不是表现为指数曲线而是条平稳的直线。此外,在6个细分类目中,"显卡技术"与"市场"数量上相差不大,是属于数量最大的一类,"显卡新闻"与"评测"次之,"显卡驱动"与"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。

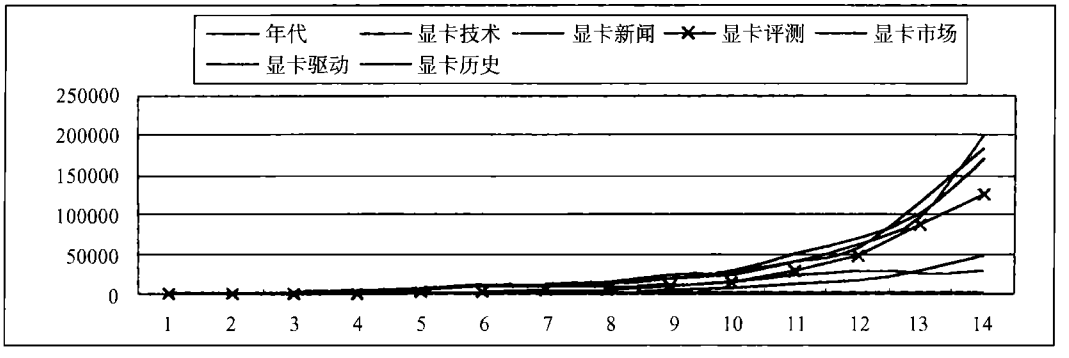

图2 显卡内容主题细分数据点折线图

同样,我们利用Spss for Windows英文版统计分析软件求各个细分内容主题的指数增长曲线方程与复相关系数,分别如表4。从表4中我们看出显卡相关的6个细分主题在各个检索统计期内的增长都在20%以上,除"显卡历史"这个细分主题其他5个细分内容主题的统计期增长率都在35%以上,"显卡评测"的检索统计期增长率甚至达到47.3%。复相关系数中除"显卡历史"复相关系数外都在0.99以上,说明这些方程对其增长状况拟合的非常好。

表4 显卡细分网络内容主题指数增长方程及其复相关系数

网络内容主题	增长方程	复相关系数
显卡技术	$Y = 908.42^{*} e^{0.365t}$	0.99304
显卡新闻	$Y = 690.15 * e^{0.381t}$	0.99224
显卡评测	$Y = 155.08 * e^{0.473t}$	0.99596
显卡市场	$Y = 710.36 * e^{0.383t}$	0.99022
显卡驱动	$Y = 135.53 * e^{0.413t}$	0.99393
显卡历史	$Y = 1494.83 * e^{0.233t}$	0.97595

2.3 增长原因分析

显卡内容主题的飞速增长首先与显卡技术的发展分不开的。考虑到本例统计的是网络文献内容主题，那么我国互联网的飞速发展，上网计算机的猛增，也是其飞速增长的原因之一。

(1)内容主题的发展。MDA，CGA - > MGA - > EGA - > MCGA - VGA - > 8514/A - > 3D 卡，在显卡的整个发展过程内，经历了上面几种卡的转变，而当时的转变却大都是围绕在 IBM PC，IBM 老大哥的主导而定。(就像目前的电脑硬件的走向是由 INTEL 呼风唤雨一样)在 1981 年，IBM 推出了 PC 时，IBM 提供了两种显示卡，一种是"单色显示卡(简称 MDA)，一种是"彩色绘图卡"(简称 CGA)[1]。1996 年 Voodoo 诞生了。Voodoo 并不是历史上第一款 3D 加速卡，但它却是第一个有应用价值的 3D 显卡产品。自此一系列与其相关的文献开始剧增[2]。

同样，显卡的发展和个人计算机的发展与应用是紧密相连的，从 1946 年世界上第一台个人计算机的诞生到 1981 年 IBM 公司个人电脑的推出，30 多年间计算机相关技术研究集中在集成电路技术、芯片技术的发展之上，显示技术的发展只是到了 80 年代后，随着个人计算机技术的飞速增长发展，经历了从单色到彩色，从 2D 到 3D 的发展。当然，相关的内容主题出现的次数也随之增长。

(2)与我国互联网发展状况的发展基本一致。本例选取的检索词为"显卡""1996 年"等，包含的基本上是中文信息，考虑到现阶段中文信息量主要由我国提供，所以我们可以基本认为我们统计的结果基本代表了我国的显卡内容主题的发展状况。

截止到 2003 年 6 月 30 日，我国的上网计算机总数已达 2572 万台，同上一次调查结果相比，我国的上网计算机总数半年增加了 489 万台，增长率为 23.5%，和去年同期相比增长 59.5%，是 1997 年 10 月第一次调查结果 29.9 万台的 86 倍(如图 3 所示)[3]。同样，显卡内容主题从 2002 年的 169000 增长到 2003 年的 296000，增长率达到 51%，可以看出两者基本持平。同样我们可以看出，我国的上网计算机数从 1997 年 7 月的 29.9 万台，增长到 2003 年 7 月的 2572 万台，同期关于显卡的内容主题从 26000 剧增到了 296000。即网卡相关内容主题的增长与我国上网计算机数的增长存在基本一致的趋势。

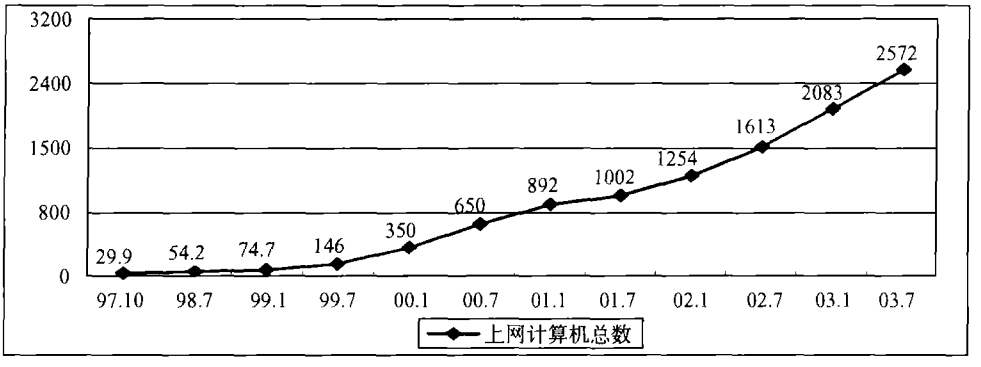

图 3　1997 年 10 月至 2003 年 7 月我国上网计算机台数变化曲线

2.4 误差分析

在实证分析的各个环节，如检索工具选择，搜索引擎选择，数据处理工具、方法等各个方面都无法完全避免出现误差。下面仅就数据检索方法、数据检索收集单位的设置、内容主题的细分 3 个方面进行具体分析。

(1)数据检索方法原因。回首我们分析数据统计的方法时我们能发现，在进行数据收集时，为了统计的方便，我们只是限定了如："显卡""1996 年"，对于相应的内容主题出现的位置、出现的作用、意义都没有进行规定。我们可以这样认为，像"'显卡''1950 年'"这样的内容主题，都是 1980 年 PC 显卡开始发展后，现在相关的文章中出现的对历史的回顾。同样，回顾互联网的历史我们也可以发现，在 50 年代的互联网雏形上是没有显卡内容主题文章(网站)存在的。

同样，因为内容主题划分简单，加上简单的数据统计方法。可能出现统计的数据的大量重叠。例如，一篇文章既讨论了显卡新闻又讨论了"显卡历史"还讨论了"显卡技术"，那么它将被在 3 个细分主

题中统计出来。同样,如果一个内容主题讨论了2002年的"显卡技术",又讨论了1999年的"显卡技术",那么它将在两年的"显卡技术"统计数据中得到体现。

(2)数据收集单位的设置。本例在选择数据的收集单位时,1995年前以10年为一个统计单位,1995年后是逐年统计,这样本来是基于计算机显卡技术、互联网的发展状况来考虑的,但是这样的单位选择对1995年的发展状况表现得不够充分。

(3)内容主题的细分方面。本例设计"两步走"的方法,目的是想通过比较,以验证"第一步"检索统计结果的准确性。但是设计"第二步"的时候,也仅仅依据了网络上内容受关注程度进行的划分,并没有科学的分类依据,可能会在一定程度上加剧"第二步"检索结果的重复率,并且在其合计中得到放大。

本例仅仅是试图从最简单的方法,最简单的计算和最基本的分析中去把握总体的网络内容主题的增长与发展趋势。针对本方法的缺陷与不足,进一步的研究可以从如下几个方面进行改进:

(1)选取具有统计功能的搜索引擎。现在具有统计功能的所有引擎不多,我们只发现了alltheweb具有统计功能,但考虑到它对中文信息统计的天然不足,固没有选取。

(2)扩大统计分析的语种。进而可以比较不同语种的增长差异,在研究内容主题增长规律的基础上,从中发现各国在特定内容主题的研究进展差异、优劣、突破口等。

(3)内容主题进一步细分。本例在就显卡内容主题进行细分时,仅仅是根据经验进行了粗略的划分,后来的研究可以依据《中国图书分类法》或者《杜威十进分类法》进行更加完善的细分,以进行检索、统计比较分析。

参考文献

1 显卡的发展史 .http://www.kwanwa.com.cn/jszcw7.htm
2 图形加速卡的过去、现在和未来.http://wwwb.pconline.com.cn/pchardware/diyheaven/question/0616gpu.htm
3 中国互联网络信息中心.2003.12.2日中国互联网络宏观状况. http://www.cnnic.net.cn/html/Dir/2003/12/02/1627.htm

邱均平 武汉大学中国科学评价研究中心主任,教授、博士生导师。通信地址:武汉。邮编430072。

殷之明 武汉大学中国科学评价研究中心硕士生。通信地址同上。

(来稿时间:2004-04-15)

(上接第9页)

参考文献

1 Neil Jarvis. Enterprise Lnformation Managemenf in the Digital Economy. http://WWW.eds.com
2 Robert Hawley, Nigel Horne. A practical Agenda for the Information Age. http://WWW.unisys.com/execmag/1997-08
3 关家麟,刘绿茵.建设国家科技文献信息资源共建共享体系的若干思考.见:信息化与信息资源管理学术研讨会议文集.北京:科学技术文献出版社,2003
4 程鹏.社会信息化与可持续发展关系分析.情报学报,1997(6)
5 何丽珈.可持续发展战略对21世纪图书馆发展的启示.图书馆论坛,2002(2)
6 胡昌平,谷斌.网络信息资源的社会化组织与开发构想.中国图书馆学报,2002(4)
7 王勇.网络信息资源开发中的多网合作模式.中国图书馆学报,2002(5)
8 黄如花.国内外信息组织研究述评.中国图书馆学报,2003(4)
9 贾君枝.市场环境中网络信息资源配置影响因素.中国图书馆学报,2003(2)
10 王翠萍.我国网络信息资源分布.情报科学,2004(2)
11 胡昌平,杨曼.国家网络信息资源组织的系统化实施.情报杂志,2003(1)

胡昌平 武汉大学信息资源研究中心教授,博士生导师。通信地址:武汉。邮编430072。

(来稿时间:2004-08-30)