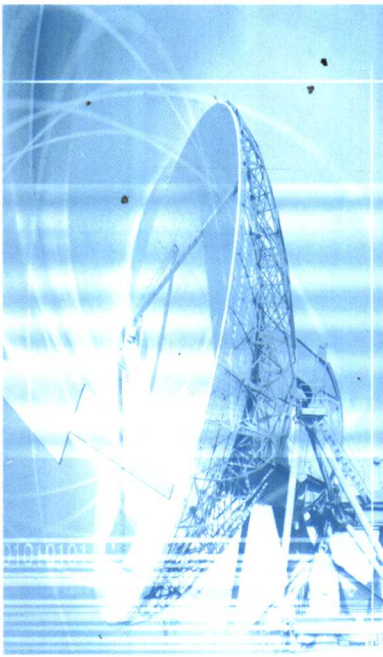
WANGLUO
XINXI JIANSUO
JISHU YU ANLI



网络信息检索技术与案例

JISHU YU ANLI

谢新洲 主编

北京图书馆出版社

网络信息检索技术与案例

谢新洲　主编

北京图书馆出版社

图书在版编目(CIP)数据

网络信息检索技术与案例/谢新洲主编. —北京:北京图书馆出版社,2005.5

ISBN 7-5013-2778-5

Ⅰ.网… Ⅱ.谢… Ⅲ.计算机网络—情报检索 Ⅳ.G252.7

中国版本图书馆 CIP 数据核字(2005)第 021673 号

书名 网络信息检索技术与案例

著者 谢新洲 主编

出版 北京图书馆出版社 (100034 北京西城区文津街 7 号)

发行 010-66139745 66175620 66125153

66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com

经销 新华书店

印刷 济宁市火炬书刊印务中心

开本 850×1168 毫米 1/32

印张 13.375

版次 2005 年 5 月第 1 版 2005 年 5 月第 1 次印刷

字数 250(千字)

书号 ISBN 7-5013-2778-5/G·618

定价 28.00 元

主　编：谢新洲

副主编：周　静　胡　英　徐守杰

撰稿人：谢新洲　周　静　胡　英　徐守杰
刘秋宏　王卫丹　龚文涛　熊松韫
田　原　王玉梅　马　明　季　琳
王艳芳　吴淑燕　黄俊平　虞惠达
肖　雯　郭吴昊　吴　健　李　毅
郑丽颖　兰　敏　闫　静　林海殷

目　录

1 网络信息检索的发展

网络信息资源是指利用互联网等方式,以文字、图像、声音、视频、多媒体等形式可以为用户共享的数字化信息资源。互联网从产生之初的军事领域的运用,发展为学术交流的平台,到现在已经演变成大众传播的舞台。我们正在见证一个互联网爆炸式增长的时代。著名互联网研究机构 Netcraft 在 2000 年 4 月的一个调查[①]显示,新的域名正在以每秒一个的速度被注册。互联网信息资源的爆炸式增长足以说明网络信息资源已经成了一种重要的信息资源,互联网已经成为人们主要的信息来源,网络信息检索因此也成为了人们日常生活、学习和工作中不可或缺的部分。

1.1 网络信息资源的类型

网络信息资源以其内容丰富、形式多样著称,以下按网络信息资源的主题对其进行划分,介绍人们日常生活、学习和工作中经常使用的网络信息资源。

新闻信息资源:互联网的出现改变了人们获取新闻信息的方式,互联网在同一时间向全世界范围内传播最新发生的新闻,人们可以不受地域限制获取世界上任何地区的新闻。世界各国主要的新闻网站是人们获取网络新闻信息的主要途径。

---

① Tony Gill. Metadata and the World Wide Web.
< http://www. getty. edu/research/conducting _ research/standards/intrometadata/2_articles/gill/index. html >

商业信息资源：商业信息资源是互联网上又一重要的资源。互联网上的企业名录、产品信息、商贸信息、金融信息和经济统计信息是一个企业获取商业信息的重要来源。

法律信息资源：互联网上具有大量免费的法律法规文献。人们可以通过互联网了解国家最新的立法，并可以通过互联网获取法律咨询服务。网络法律信息资源将是人们生活工作中的一种重要资源。

学术信息资源：网络学术信息资源主要是收录高质量学术期刊的网络全文数据库、网上免费的电子期刊，以及用于网上学术交流的 Working Paper 和 E - Print。这类信息资源主要针对大学及研究机构。

娱乐信息资源：互联网上有许多休闲娱乐信息，包括电影、音乐、游戏、足球、购物信息和旅游信息等。这类信息已经成了人们日常生活中的一部分。

除此之外，还有许多重要的网络信息资源，比如政府信息资源、教育信息资源、就业信息资源、广告信息资源等等。互联网上的信息资源可以说是包罗万象，人们总能发现自己需要的信息资源。

1.2 网络信息资源的现状

网络信息资源是一种重要的国家战略资源，网络信息资源的建设受到了世界各国的重视。早在 1993 年美国就启动了国家信息基础设施（NII）计划，将使美国公民享用广泛的信息资源及信息服务作为信息资源建设的主要目标之一。而我国是从 20 世纪 90 年代中期开始进入互联网的迅猛发展、网上中文信息资源快速增长的阶段。以下从互联网上网页的数量、网络服务器的数量、网络主机的数量及搜索引擎标引网页的数量四个角度来对网络信息资

源的现状作一个统计性描述。

1.2.1 网页数量

网络信息资源主要以网页形式存在于互联网中，在20世纪90年代以后，互联网的迅速发展，使网页资源成为了网络信息资源的主流。

根据著名互联网市场研究公司GlobalReach在2004年9月公布的数据，全球网页的数量已经达到3130亿。GlobalReach的专家对网页内容按语言划分，并对排名前十种语言的网页数量进行了统计，其中英语网页占68.40%，日语网页占5.90%，德语网页占5.80%，中文网页占3.90%，法语网页占3.00%，西班牙语占2.40%，俄语占1.90%，意大利语占1.60%，葡萄牙语占1.40%，韩国语占1.30%，其他语种占4.60%（见图1.1）。

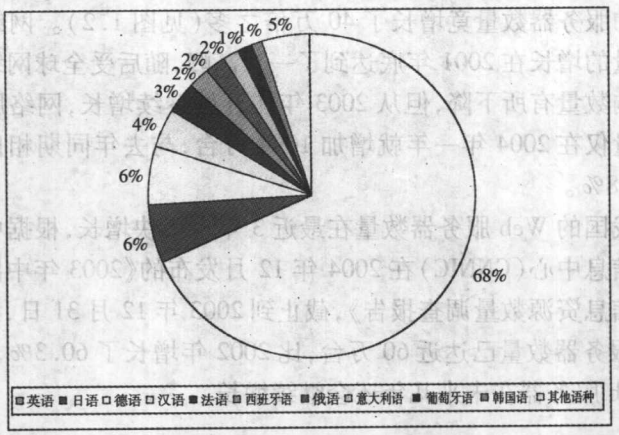


图1.1 按语种全球网页的分布

数据来源：http://global-reach.biz/globstats/refs.php3

在全球的网络信息资源中，英语网络资源超过一半之多，依然是网络信息资源的主流语种，但其他语种的网络信息资源已经取

得了一定发展，日语、德语、汉语和法语信息资源已经在全球网络信息资源中占了一席之地。随着网络信息资源重要性的凸现，世界其他各国都会加快本语种网络信息资源建设的步伐，以提高本国在网络时代的地位。

1.2.2 服务器数量

互联网上 Web 服务器数量的增长反映了网络信息资源的变化趋势，1993 年 6 月美国麻省理工大学的 Matthew Gray 教授对互联网进行的一次调查[①]中只发现了 130 个 Web 服务器。在同年 9 月，美国开始启动国家信息基础设施（NII）计划，互联网开始引起商业界和新闻媒体的注意，Web 服务器开始快速增长。Netcraft 在 2004 年 12 月进行的互联网 Web 服务器数量的调查中收到 HTTP 请求响应的服务器数量达 5700 万台，在短短 10 年多时间里互联网上的服务器数量竟增长了 40 万倍之多（见图 1.2）。网络服务器数量的增长在 2001 年底达到了一个高峰，随后受全球网络泡沫的影响数量有所下降，但从 2003 年初开始继续增长，网络服务器的数量仅在 2004 年一年就增加 1094 万台，与去年同期相比增长了 23.8%。

我国的 Web 服务器数量在最近 3 年也飞快增长，根据中国互联网信息中心（CNNIC）在 2004 年 12 月发布的《2003 年中国互联网络信息资源数量调查报告》，截止到 2003 年 12 月 31 日，我国的 Web 服务器数量已达近 60 万台，比 2002 年增长了 60.3%。我国的 Web 服务器在未来几年还会继续增长。

① Matthew Gray. Web Growth Summary. < http://www.mit.edu/people/mkgray/net/web-growth-summary.html >

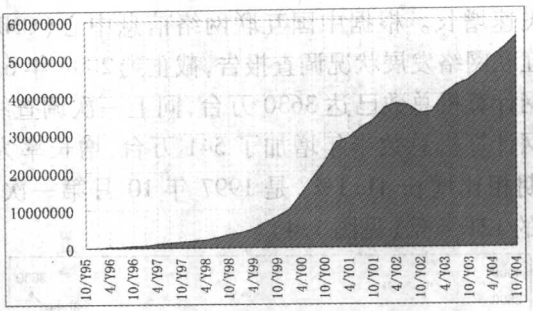


图1.2 1995年10月至2004年10月网站服务器数量的增长(单位:台)

数据来源:http://www.netcraft.com/survey/Reports/index.html

1.2.3 网络主机数量的飞速增长

网络上除了服务器数量的不断增长外,全球上网主机的数量也在不断增加,根据互联网系统联盟(ISC)公布的调查结果,截止到2004年7月,ISC通过DNS发现的网络主机数量达到了285 139 107台,与去年同期的调查结果相比,增加网络主机82 769 218台,增长率为40.9%,是1993年1月主机数量的217倍(见图1.3)。

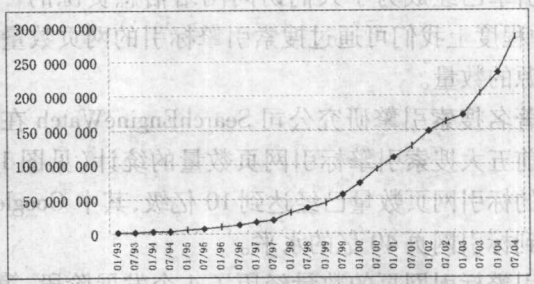


图1.3 1993年1月至2004年7月全球网络主机数量的增长(单位:台)

数据来源:http://www.isc.org/index.pl?/ops/ds/reports/2004-07/

同时,我国上网计算机的数量从1997年以来的8年时间中

也发生了飞速增长。根据中国互联网络信息中心(CNNIC)第十四次中国互联网络发展状况调查报告,截止到2004年6月30日,我国的上网计算机总数已达3630万台,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541万台,增长率为17.5%,和去年同期相比增长41.1%,是1997年10月第一次调查结果29.9万台的121.4倍(见图1.4)。

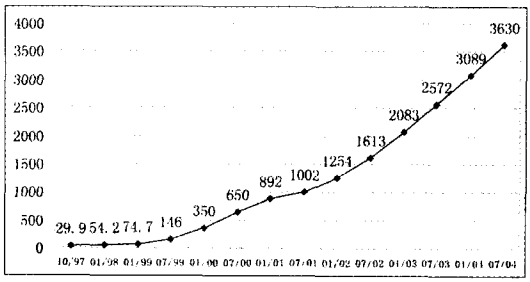


图1.4 我国历次调查上网计算机总数(单位:万台)

数据来源:http://www.cnnic.net.cn/index/0E/00/11/index.htm

1.2.4 搜索引擎标引网页数量的急剧攀升

搜索引擎已经成为了人们访问网络信息资源的一个主要途径,在某种程度上我们可通过搜索引擎标引的网页数量"一窥"网络信息资源的数量。

根据著名搜索引擎研究公司SearchEngineWatch在2003年9月对全球前五大搜索引擎标引网页数量的统计(见图1.5),主流搜索引擎的标引网页数量已经达到10亿级,其中Google在今年2月已经达到标引网页40亿的水平。

搜索引擎标引网页的数量经历了4个发展阶段:第一阶段是1997年12月至1999年6月,当时主要搜索引擎是AltaVista、Inktomi和Northern Light,最高标引网页数量达1.5亿;第二阶段是1999年9月到2000年6月,AllTheWeb和Google加入竞争,到

2000 年 6 月，Google 标引的网页数量已经达到 5 亿；第三阶段是 2002 年 6 月至 2002 年 12 月，在这一阶段，各大主要搜索引擎标引数量都有了飞速增长，其中，AllTheWeb 达到了标引 20 亿的水平，而 Google 和 Inktomi 都称已经达到了标引 30 亿个页面的水平；第四阶段是 2003 年 8 月至今，Google 的标引数量已经达到了 40 亿。①

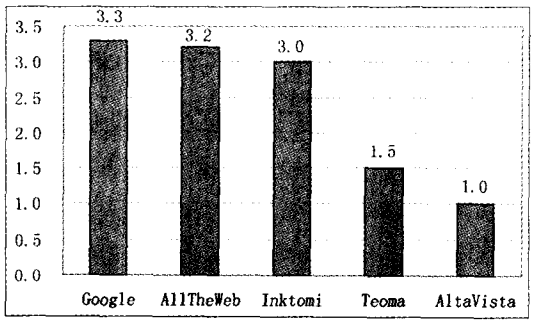


图 1.5 全球前五大搜索引擎标引网页数量(单位：十亿)
数据来源：http://searchenginewatch.com/reports/article.php/2156481

1.3 网络信息资源的用户

网络信息资源的爆炸式增长是网络用户对信息资源需求不断增长的必然表现，或者说网络用户的需求推进了网络信息资源的不断发展。无论是全球网页数量和网络服务器数量的增长，还是搜索引擎标引网页数量的增长，都是在某种程度上为了更好地满足网络用户的信息需求。本节从全球网络用户的数量、网络用户信息检索需求和网络用户查找网页途径 3 个角度来介绍网络信息

① Danny Sullivan. Search Engine Sizes. <http://searchenginewatch.com/reports/article.php/2156481>

资源用户的现状。

1.3.1 网络用户数量的飞速增长

近年来,全球网络用户的数量不断攀升,根据 GlobalReach 最新的统计数据,截止到2004 年9 月30 日,全球网络用户的数量已经达到8.5 亿人,与2003 年同期相比,增加了1.21 亿个网络用户,增长率为16.6%,是1996 年网络用户数量的17 倍。GlobalReach 的专家预测,到2005 年底全球网络用户将达到11 亿,到时将占全球总人口的19%(见图1.6)。

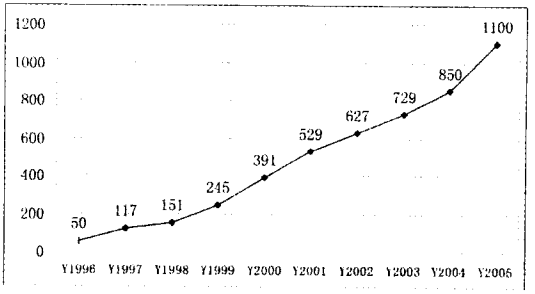


图1.6 1996 年至2005 年全球网络用户增长趋势(单位:百万人)
数据来源:http://global - reach. biz/globstats/evol. html

同时,我国上网用户人数也在不断增长,根据中国互联网络信息中心(CNNIC)最新调查数据,截止到2004 年6 月30 日,我国的上网用户总人数为8700 万人,同上一次调查相比,我国上网用户总人数半年增加了750 万人,增长率为9.4%,和去年同期相比增长27.9%,同1997 年10 月第一次调查结果62 万上网用户人数相比,现在的上网用户人数已是当初的140.3 倍(见图1.7)。我国网络用户的数量已经处于世界第二位,仅次于美国。据全球顶级咨询机构摩根斯坦利的预测,我国将在未来5 年内网络用户数量超过美国,成为世界上网络用户第一大国。

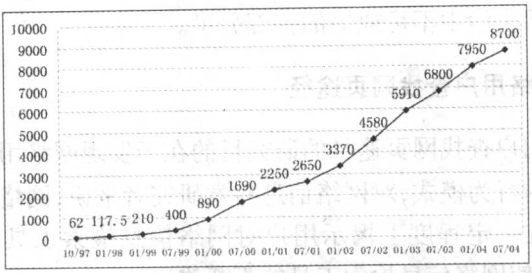


图 1.7　我国历次调查上网用户总数(单位:万人)

数据来源:http://www.cnnic.net.cn/index/0E/00/11/index.htm

1.3.2　网络用户信息检索需求不断提高

今天,我们已经越来越依赖网络来获取各种各样的信息,按照传统的信息需求理论,用户的信息需求是无限的,而网络用户对网络信息的需求同样也是无限的。根据 SearchEngineWatch 在 2003 年 2 月做的关于网络用户每天在全球主要搜索引擎检索次数的统计,全球每天有 6.25 亿人次的网络用户检索需求,其中每天有 2.5 亿人次的网络用户使用 Google 来检索网络信息,全球网络用户每天花在网络信息检索上的时间超过 1 亿分钟(见图 1.8)。可见,网络用户的信息检索需求在不断提高,网络信息检索已经成为

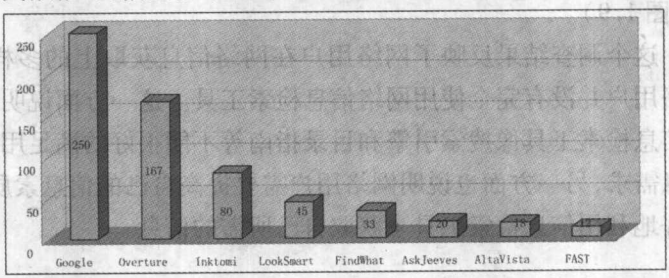


图 1.8　全球主要搜索引擎每日提交用户检索请求次数(单位:百万人次)

数据来源:http://searchenginewatch.com/reports/article.php/2156461

了人们日常生活中不可或缺的一部分了。

1.3.3 网络用户查找网页途径

网络用户查找网页途径的研究目的在于发现网络用户查找网络信息时的行为模式,对网络信息检索研究者来说,网络用户检索途径可以在一定程度上揭示用户对网络信息检索工具的满意程度,从而促使网络信息检索工具的不断改进。

美国乔治亚技术研究所 GVU 中心从 1994 年到 1998 年对全球网络用户查找网页途径进行了十次调查,GVU 总结了十种途径,第一种是通过网页之间的链接,第二种是通过搜索引擎,第三种是通过目录指南,第四种是通过新闻组(Usenet),第五种是通过 email 后的签名,第六种是通过图书,第七种是通过杂志和报纸,第八种是通过电视广告,第九种是通过朋友,最后一种是其他途径。[①] 在 1998 年的第十次调查中,共有 3291 份有效问卷,调查结果显示,网络用户最常用的查找网页的途径是从一个网页链接跳到另一个网页链接,占调查对象的 88.3%,其次是选择搜索引擎来查找的用户占了 84.8%,而另一种查找网络信息的主要工具目录指南占了 58%,分别排在朋友推荐及杂志和报纸之后(见图 1.9)。

这个调查结果反映了网络用户在网络信息获取上的多样性,网络用户并没有完全使用网络信息检索工具。这一方面说明了网络信息检索工具像搜索引擎和目录指南等不能很好的满足用户的信息需求,另一方面也说明网络用户需要提高自己的信息素质,以更好地利用信息检索工具来获取自己所需的信息。

---

① GVU's 10th WWW User Survey. (http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/use/q52.htm)