专题:IPv6 技术

基于搜索引擎的 IPv6 网络分析 *

刘 辉 叶绍志 黄 晖 李 星
(清华大学电子工程系 北京 100084)

摘 要 本文首先介绍了 IPv6 搜索引擎发展的国内外背景、网络指南针 IPv6 搜索引擎的主要技术特点和实现;然后基于网络指南针搜索引擎运行过程中得到的数据,从多个角度分析了全球 IPv6 网络的发展,包括站点的分布、规模、类型等;最后对 IPv6 的发展趋势进行了展望。

关键词 IPv6 搜索引擎 网络指南针

1 引言

1999 年, CERNET 在中国教育科研网范围内组建了 IPv6 试验床,在试验床的 pTLA(pseudo Top Level Aggregation) 地址范围内开始分配地址,同时开始进行有关 IPv6 各种特性的研究与开发。"网络指南针联盟"是 CERNET 网络中心自主开发的网络搜索引擎联盟,同时也是教育科研网"网络指南针"搜索引擎的第二版本,提供一系列不同资源的搜索服务,自 2000 年 7 月开始正式运行。2000 年诺基亚资助"网络指南针"进行 IPv6 搜索引擎的开发。网络指南针 IPv6 搜索引擎从 2001 年 5 月份开始提供正式稳定的 Web 服务。

分析一个网络发展的状况有很多方法,如分析路由表项、DNS 服务器、用户行为等。在半年的运行时间内,网络指南针积累了大量关于全球 IPv6 Web 站点的数据,从搜索引擎的角度分析 IPv6 在全球的发展状况,在国内外还比较少,并具有一定的实际意义。

2 网络指南针 IPv6 搜索引擎的实现

目前,网络指南针搜索引擎小组正在进一步地研究开发 IPv6 搜索引擎,传统的搜索引擎系统包括数据的采集、索引和查询三个部分。IPv6 搜索引擎和 IPv4 搜索引擎的最大的不同之处在于数据的采集。要访问基于 IPv6 协议栈的 Web 服务器,就必须要使用支持 IPv6 的网络蜘蛛 (搜索引擎中用来在网络上采集数据的一种软件),网络指南针 IPv6 搜索引擎所采用的主要采集程序是 IPv6 Wget(Wget 是一个 Linux 环境下用于从 World Wide Web 上提取文件的工具,这是一个 GPL 许可证下的自由软件,其作者为 Hrvoje Niksic),并考虑到现在 IPv6 Web 服务器数量少,文件总数相对不多,采集链路不稳定的情况,对采集程序做了一定的优化,主要有:超时设定、重试次数、时间标签和采集深度,以保证采集的快速有效。

IPv6 和 IPv4 搜索引擎的另一明显差别是 IPv6 Web 站点数量较少而且现在尚没有一个覆盖范围比较广的站点数据库,为解决这一问题,网络指南针小组从采集所得到的数据中分析出指向其他 IPv6 站点的超链接,从而获得更多的 IPv6 站点,将这些站点加入网络蜘蛛的采集目标中,就可以迅速地增加 IPv6 搜索引擎的数据量和覆盖范围。

3 网络指南针 IPv6 搜索引擎数据分析

3.1 IPv6 站点的国家分布情况

从 6bone 得到的数据显示,现在已经有 1000 多个 IPv6 站点在上面注册,搜索引擎采集到 100 多个提供 Web 服务的 IPv6 站点的数据,对这些站点的所在国家进行统计,得到 IPv6 站点的国家分布情况(如图 1 所示)。

从图 1 可以看出日本的迅猛发展,欧洲共同体的整体发展

* 网络指南针 IPv6 搜索引擎项目由 Nokia 资助开发完成

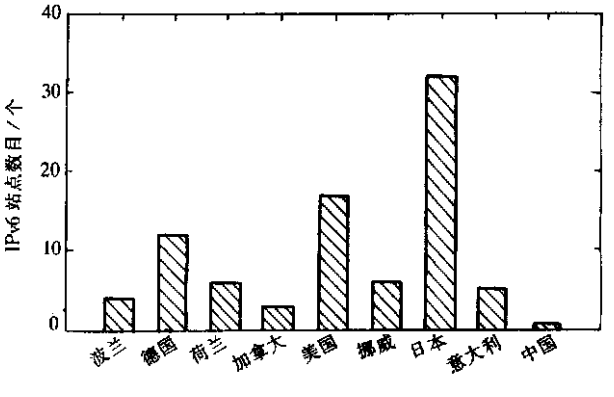

图 1　IPv6 站点的国家分布

优势以及美国这个网络大国对待 IPv6 不冷不热的态度。IPv6 对于亚洲地区，尤其对于中国来说具有特别的意义，中国所有的 IPv4 地址总和还比不上美国一个著名大学分配的 IPv4 地址数量，技术的差别造成了网络资源分配的巨大不平等。为了在下一代网络中占据主动权，亚洲地区在 IPv6 的发展上投入了大量的人力、物力，同时也取得了很大的发展。

3.2　IPv6 站点类型情况

除了常见的几种站点，如 com、net、org 等，很多由国家实验室或商业企业联合实验室建立的 IPv6 站点用来进行 IPv6 技术的研究开发，这类站点被归入 IPv6 实验站点类别。不同国家的后缀名有不同规定，如日本的商业站点后缀为 co，学术站点后缀为 ad，按照站点的类型分别归入 com、net、org、edu 四大类型。

通过分析 IPv6 站点类型的分布(如图 2 所示)，可以从一个侧面了解不同领域、行业的人们对于 IPv6 发展的不同态度和举措，有助于判断当前 IPv6 技术所处的阶段和形势，以及进一步发展面临的机会与挑战。

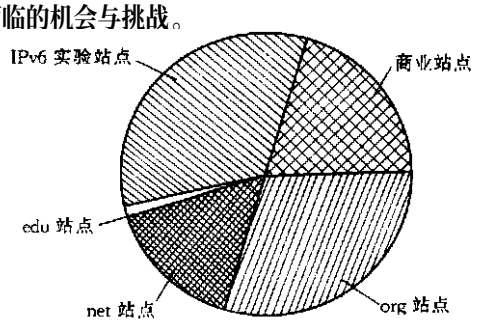

图 2　IPv6 站点的类型分布

从图 2 可以看到，IPv6 站点中学术和非盈利机构站点的数目所占比例较大，说明目前 IPv6 站点的建立和发展还主要是处于科研和测试阶段；商业站点已经占有相当的份额，相当一部分是商业机构投资于 IPv6 技术的研发，其中已经有一些采用 IPv6 技术的产品投入市场，如 Nokia 与 CERNET 在 IPv6 方面的合作。

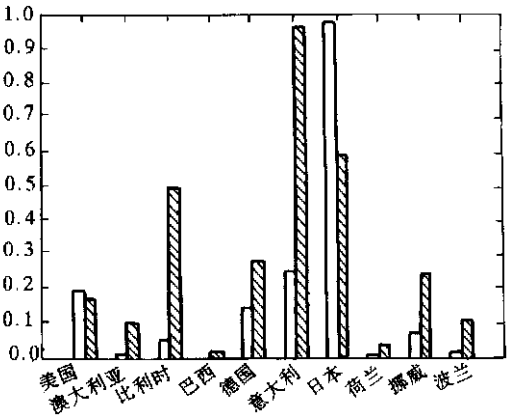

图 3　IPv6 站点规模的国家分布

3.3　IPv6 站点规模情况

关于站点规模，有很多度量方法，这里根据搜索引擎的数据特点，采用了两种度量方式，一个国家拥有的所有 IPv6 站点的页面数总和以及平均页面数(页面数总和除以站点数)。

图 3 中每个国家的两个立柱，左边代表总页面数，右边代表平均页面数。日本的页面总数居于首位，这与日本的发展速度之快十分吻合。其它国家站点规模的分布情况与 IPv6 站点的分布趋势大致相似，在全球范围内仍然呈现分布不均的趋势，三个地域(亚洲、欧洲、北美)的 IPv6 分布集中，欧洲一些国家的 IPv6 站点数目虽然并不是很多，而且规模比较小，但地区分布比较广泛，IPv6 技术的发展比较普及，其它地区很少甚至没有。有一个特例，意大利的总页面数不是很突出，但平均页面数遥遥领先，原因是意大利有几个十分庞大的 IPv6 站点。这个特例其实反映了 IPv6 网络的规模仍然很小，分配到不同的国家就更加有限，使得个别站点对总体数据的影响较大，随着 IPv6 网络像 IPv4 一样在全球范围内普及开来，图 3 所示的反常现象就不会再出现。

图 4 的数据分别来自 2001 年的 10 月和 11 月(左边代表 10

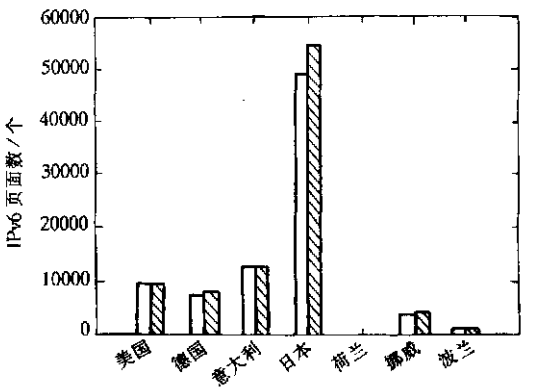

图 4　IPv6 站点规模的增长

月 27 日,右边代表 11 月 11 日),反映了在短短的一个月内 IPv6 动态增长的情况。可见日本在 IPv6 领域具有良好的开端,而且始终保持全速前进的势头,增长速度远远超过其他地区。

3.4 IPv6 站点的引用率

本文中对一个站点引用率的判断依据是其它站点对该站点的引用次数,表 1 中仅列举了根据搜索引擎采集的页面进行分析得到的排名前十名的站点。

表 1 前十名站点的引用率

排 名	站 点	链接数
1	www.freebsd.org	999514
2	www.tac.eu.org	4930
3	carmen.cselt.it	3397
4	bd.wide.ad.jp	2364
5	whois.iprg.nokia.com	1110
6	www.gnome.org	982
7	ee-staff.ethz.ch	597
8	www.ietf.org	446
9	publish.aps.org	432
10	hp.vector.co.jp	350

从表 1 中可以看出:引用率最高的站点主要还是来自日本、美国等 IPv6 技术发展先进的国家;引用率最大的站点的规模并不是非常大,说明站点的重要性主要还是取决于站点的内容;有一个有趣的事实,www.freebsd.org 的引用率比其它站点高出两个数量级,造成这个巨大差距的原因是 FreeBSD 是一种广泛流行的 UNIX 操作系统,也是 IPv6 实现的最主要平台。

4 结论

通过对网络指南针 IPv6 搜索引擎采集到的全球 IPv6 站点网的数据进行分析,可以得到最全面的关于 IPv6 站点的发展状况的静态统计和分析。另外,由于搜索引擎的动态更新,可以对 IPv6 站点的动态变化进行即时的跟踪和分析,观察 IPv6 站点发展的趋势。因此搜索引擎作为信息检索理论在网络上的应用,对于分析网络增长行为具有特殊的意义。本文通过对“网络指南针联盟”采集到的 IPv6 站点数据进行分析,得出以下结论:

(1) 作为下一代网络的 IP 协议,IPv6 的发展与地区 IPv4 网络的技术基础和发展程度有着密切联系,但更为重要的因素是该国家(地区)的发展战略和策略选择。亚洲地区的地址耗尽促使日本、中国等地区成为 IPv6 技术的积极实施者。欧洲地区雄厚的资金和技术实力,以及它们希望在下一代网络中占据主动的目标,使得欧洲在 IPv6 的发展方面具有整体的优势。美国的 IPv4 地址还可以维持一段时间,因此并不需要竭尽全力来发展新的协议,但其强大的技术基础不可小视。

(2) 目前 IPv6 网络还远远不能够和 IPv4 抗衡,主要的 IPv6 站点目的还是为了研究和测试。从 IPv4 到 IPv6 的过渡,不仅是技术上的革新,更意味着巨大的商业战略和商业投资,尽管新兴的网络社会充满了商机和希望,但风险更大,因此 IPv6 的商业化程度还远远不够,大多数的企业积极研发,静观发展,当 IPv6 成为 Internet 的必需品时,他们就会群起争夺 IPv6 产品的市场,这就是 IPv6 实验站点比例较高的原因。

(3) 在搜索引擎采集数据的过程中,发现相对于 IPv4 的网页来说,IPv6 的网络具有明显的不稳定性,一个原因是 IPv6 网络处于成长期,增长速度较快,截至 2002 年 1 月 20 日网络指南针已经采集到 103 个站点的 15 万网页;另一个原因是实验站点比重较大,许多站点尚未提供成熟稳定的网络服务。

搜索引擎采集到的是一部分已经提供 WWW 服务的 IPv6 站点数据,从一定程度上可以反映整个 IPv6 网络的发展和增长情况。但是由于实验周期比较短,采集到的站点数量有限,链路状况不能保持稳定,本文的结论也存在局限性。但从 WWW 服务的角度来分析 IPv6 站点对于分析 IPv6 网络的增长来说是一种独特的方法,将对网络发展的研究产生积极的作用。

Research and Data Analysis of IPv6 Search Engine

Liu Hui, Ye Shaozhi, Huang Hui, Li Xing
(Department of Electronic Engineering, Tsinghua University, Beijing 100084)

Abstract Firstly this paper provided background of IPv6 search engine (SE) development, with the introduction of main technology character and developing process of compass IPv6 SE. Secondly using data from compass SE, it analyzes varies aspect of global IPv6 network increasing behavior, including IPv6 sites distribution, magnitude and dynamic change. At last it presents some conclusion from data analysis, and concludes by anticipation of IPv6 development.

Key words IPv6, search engine, network compass

(收稿日期:2002-01-26)