COMPSCI 742 押题

Bo Pang/庞磅

Last Update: 2023 年 11 月 7 日

目录

第一编

网络测量

必读书目

0.1 Web workload characterization

0.2 Pinging in the rain

0.3 Modern website complexity

0.4 Comparative analysis of Web and P2P traffic

0.5 Interference effects in Wi-Fi networks

第 1 章　网络测量

1.1　网络测量中的概念

互联网测量可根据网络监控器的位置（边缘网络与核心网络）、使用的测量/分析工具（基于硬件与基于软件）、探测机制（被动与主动）以及视点数量（单视点与多视点）进行分类。

1.1.1　了解主动与被动等测量方法，以及边缘与核心等有利位置

被动网络测量是通过监听通过路由器或主机的所有流量来进行的。主动测量包括生成从一台主机到另一台主机的特殊探测流量。探测流量可能包含几乎没有有效载荷的小型 UDP 数据包。

注意，Google Analytics 被视为主动测量。

1.1.2　给定一个场景，确定测量方法和有利位置

在对 WWW2007 网站进行测量时，采用了被动测量和主动测量两种方法。

因为本研究收集并分析了服务器端数据和客户端数据。在服务器端，数据是从服务器日志中收集的，而客户端数据则是从 Google Analytics 服务中收集的。收集和分析服务器日志是观察服务器的一种方式，因此是被动的。

使用 Google Analytics 服务会在网页中注入 JS 部分，从客户端收集数据。它不需要用户额外参与或干预，但会主动向 Google Analytics 服务发送数据。因此，这是一种主动测量。

视点包括服务器端和客户端，它们都是边缘视点。因为服务器端和客户端都处于互联网网络的边缘。

这项工作考虑了多种观点。研究采用了服务器端和客户端测量技术来描述网站访问者的使用行为。在服务器端，研究分析了服务器的使用情况和流量。在客户端，还研究了多种用户行为，如提示、浏览网站的偏好和页面深度等。

本次测量研究进行了离线分析。研究分析了网络服务器上的文件，以研究服务器的性能，而 Google Analytics 则报告了客户端的行为。所有分析都是离线完成的。

测量中使用到的软件工具主要有：访问日志、文件列表、谷歌分析服务、Cookie 强化日志、服务器插件。

1.2　网站的复杂性

1.2.1　研究的意义

这项研究的意义在于揭示了网站复杂性对用户体验的影响，尤其是页面加载时间对用户满意度的直接影响。

1.2.2 影响网站性能的因素及其原因

内容层面

- **对象数量和大小**：研究发现，网站加载的对象数量是影响页面渲染和加载时间的最重要因素。在所有等级范围内，网页请求的对象数量中位数超过 40 个，20% 的网页请求的对象数量超过 100 个。新闻网站加载的对象数量明显多于其他网站。对象的大小也是一个重要的考量，但其影响相对较小（页面 8）。每个对象都需要单独的 HTTP 请求，因此对象数量的增加会导致更多的网络延迟和服务器处理时间，从而增加页面加载时间；而大对象（如高分辨率图片或大型 JavaScript 文件）会占用更多的带宽，导致加载时间增加。

- **内容类型**：网站加载的内容类型也影响性能。各种内容类型在不同等级范围内的贡献相似。图片在对象比例中占主导地位，但在字节比例中占较小比例。儿童和青少年网站的 Flash 内容比例明显高于其他网站。不同类型的内容（如 Flash 或 JavaScript）可能需要额外的客户端处理时间，这会影响页面的可用性和响应速度。

服务层面

- **服务器数量**：25-55% 的网站从至少 10 个服务器加载内容。新闻网站从明显多于其他网站的服务器获取内容。服务器数量增加可能会导致更多的网络延迟和更复杂的服务器处理，这可能会导致加载时间的不确定性和波动。此外，客户端可能需要开启多个 HTTP/TCP 连接到许多不同的服务器，这也会增加页面的加载时间。

- **非源内容（服务器、对象、时间）**：超过 60% 的网站从至少 5 个非源源加载内容。非源内容贡献了超过 35% 的下载字节。然而，非源内容对下载时间的影响相对较低，因为浏览器的优化减少了它们的影响（页面 1、2）。图片是由源码提供的主要对象类型，但 Javascript 占非源码对象的很大一部分。广告和分析服务是最常见的非源对象，而内容分发网络（CDNs）则贡献了大部分字节。这些第三方服务的集成对网站性能有显著影响（页面 4）。由于浏览器需要解析和执行来自不同源的内容，这可能会引入额外的延迟。尽管浏览器优化可以减少这种影响，但过多的非起源内容仍然可能导致性能问题。

第 2 章　齐普夫定律

在论文《1.3 Characterization of Content Delivery Applications》、《3.1* Workload Characterization of a Large Systems Conference Web Server》中，作者都提到了齐普夫分布。

论文《7.4 A Tale of the Tails: Power-laws in Internet Measurements》中，详细介绍了齐普夫定律。论文揭示了互联网测量中的幂律。幂律分布中的“重尾”和“长尾”现象称作“尾巴的故事”。在互联网测量中，这些分布常常表现出尾部的数据值比正态分布或指数分布中的要多，这意味着在分布的尾部有一些非常大的值出现的概率非常高。

- **重尾 (Heavy Tails)**：如果一个概率分布的尾部不是指数级地减少，那么这个分布就被称为重尾。重尾分布强调了大值的存在，这些较少出现的大值的变化对分布的影响比频繁出现的小值的变化更大。
- **长尾 (Long Tails)**：长尾现象是幂律关系的一个表现，它体现了低频事件在统计上比高斯分布要多得多。例如，搜索引擎中使用的关键词就表现出长尾特性，即存在一长串不常用的关键词，尽管每个关键词的使用频率不高，但它们加起来可能占搜索引擎看到的关键词搜索的很大一部分。

幂律分布

幂律特性通常出现在高方差分布中，在这种分布中，观测值的数量级跨度很大，尤其是在分布有明显偏斜的情况下。与广泛用于电信系统数学建模的指数分布相比，幂律分布的衰减速度更慢。幂律的存在表明，任意大数值可能以不可忽略的概率出现，因此，如果大型数据集中存在足够多的此类样本，与其将这些极端值视为” 异常值” 而忽略不计，不如研究其统计普遍性。

幂函数以 $f(x) = \alpha \times x^{-\eta}$ 的形式出现。其中，α 和 η 是正数常量，η 称为标度指数 (scaling exponent)。对幂函数两边取对数得出 $\log(f(x)) = -\eta \log(x) + \log(\alpha)$。这个表达式呈现线性关系，斜率为 $-\eta$，y 轴截距为 $\log(\alpha)$。在对数-对数标尺上绘制时，该函数显示为一条直线。这种现象通常被认为是幂律关系的显著特征。

幂律分布是在计算机科学文献中常用来描述某些数据集特性的一种分布。幂律分布的特点是，当你观察数据集中的大数值时，数据的分布遵循一个幂函数的形式。这里的幂函数表示为 $f(x) \sim x^{-\eta}$，其中 $\sim$ 符号表示随着 x 增加到很大的值时（趋向于无穷大），$f(x)$ 与 $x^{-\eta}$ 的比值趋近于某个正常数 c。

当 x 趋于无穷大时（即在数据集的 “尾部”），比例 $f(x)/x^{-\eta}$ 趋向于一个正常数 c。在实际应用中，这意味着数据的大值不会像正态分布那样迅速减少，而是以一种可预测的方式缓慢地减少。这是幂律分布的一个关键特征，通常用来描述社会、科技和自然现象中的许多类型的数据。在数据的尾部，即我们关注较大数值时，这个性质特别显著，因此这部分被称为“分布的尾部”。

幂律在许多自然发生的现象中被观察到（例如，地震、降水、地形），以及在许多人类行为中（例如，引文、城市人口、财富）。在信息系统的许多方面也观察到幂律，包括软件系统和计算机网络。早期例子包括虚拟内存系统中的内存引用行为、数据库查询以及文件系统中的文件使用模式。互联网和网络的几个特性也被声称表现出幂律特性，例如网站的访问者数量、网页的超链接数量、网页对象的大小、互联网上路由器的链接数量、在线社交网络上用户的朋友数量。

帕累托分布

帕累托分布（Pareto Distribution），这是一种在互联网流量测量中常见的幂律分布（power-law distribution）。帕累托分布可以用来描述那些大事件发生概率低但影响巨大的现象，常见于经济、社会科学和许多自然现象中。

在描述帕累托分布时，通常会用到互补累积分布函数（CCDF），它用来表示一个事件发生的概率大于某个值 x。数学上，CCDF 用 $P[X > x]$ 表示，并且满足与 $-\kappa$ 成反比，这里的 κ 是形状参数（shape parameter）。这个性质表明，随着 x 的增加，发生大于 x 的事件的概率以 x 的 $-\kappa$ 次幂衰减。

在双对数坐标轴（doubly logarithmic scales）上绘制 CCDF 时，如果数据遵循幂律分布，那么会在图上显示为一条直线。确定一个数据集是否遵循对数正态、幂律或其他相关分布并不容易，识别的金标准是在双对数坐标上看到一条直线。

此外，幂函数的积分本身也是一个幂函数，这意味着如果你对帕累托分布的密度函数进行积分，你得到的累积分布函数（CDF）仍然是幂律形式的。因此，可以很容易地展示帕累托分布和幂律分布之间的关系，通过 $\kappa = \eta - 1$ 这一关系式，其中 η 是之前讨论的幂律分布的指数。

指数分布（exponential distribution）和对数正态分布（lognormal distribution）的尾部行为（即大值事件的概率衰减）和帕累托分布是不同的。帕累托分布的尾部比指数分布更厚，也就是说，在大事件发生的概率方面下降得更慢。

2.1 什么是齐普夫定律

GPT 的解释

齐普夫定律（Zipf's Law）是由美国语言学家乔治 · 金斯利 · 齐普夫（George Kingsley Zipf）提出的一种经验法则。它描述了一种在自然语言文本中出现的词频分布现象，即在给定的语料库中，任何词的频率与它在频率表中的排名成反比。

具体来说，齐普夫定律指出，语料库中最常见的词出现的次数大约是第二常见的词的两倍，是第三常见的词的三倍，以此类推。换句话说，第 n 个最常见的词在语料库中出现的频率大致成 1/n 的规模。这个定律在自然语言的各种现象中都有体现，如单词的使用频率、城市人口排名、网站访问量等。

数学上，齐普夫定律可以用以下的公式表示：

$$f(n) = \frac{C}{n^a}$$

这里的 $f(n)$ 表示排名为 n 的项的频率，C 是一个常数，a 通常接近于 1。

齐普夫定律不仅仅适用于语言学，它在许多自然和社会科学领域都有广泛的应用，显示出一种普遍的幂律分布特征。例如，它也被用来描述城市规模、公司规模、收入分布等的统计特性。这个定律的一个有趣的特点是，尽管它非常简单，但它却能非常准确地描述现实世界中的各种现象。

论文 7.4 的解释

幂律的另一个经典例子是齐普夫分布，它最早用于模拟书面文本中的词频，后来又被用于模拟图书馆图书、电影租赁和网络对象的偏斜流行度。齐普夫分布是一种离散分布，在等级-频率域中由齐普夫定律定义，该定律指出，当项目按受欢迎程度降序排列（R）时，项目的频率（F）与项目的等级成反比。

2.2 齐普夫定律的数学表示

齐普夫分布是一种离散分布，在等级-频率域中由齐普夫定律定义：如果我们将一些项目按照它们的流行度排名，排在第 R 位的项目的频率 F 和它的排名 R 的关系可以用下面的公式表示：

$$F \propto R^{-\theta}$$

2.3 齐普夫定律的参数

在一个对数-对数排名-频率图上，齐普夫分布呈现为一条直线，这条直线的斜率是 $-\theta$。在这种图上，两个坐标轴（排名和频率）都是对数刻度，因此原本呈幂律分布的关系在这样的图上显示为直线关系。θ 通常接近 1，但也可以有不同的值。意味着排名每增加 10 倍（一个对数单位），频率会减少到原来的大约 1/10。当 $\theta = 1$ 时，我们得到一个纯粹的齐普夫分布。

在实际的互联网度量中，可能会观察到齐普夫分布的“退化形式（Degenerate forms）”，这种情况下分布的行为是分段线性的，或者说只有图中的一部分是线性的。这意味着实际观察到的数据可能在某些排名区间内遵循齐普夫分布，在其他区间则不遵循。

例如，人们发现在点对点文件分享系统中，文件的流行度呈现出齐普夫分布的特征，但最受欢迎的文件与预期的直线有所偏离。这种偏离可能是因为用户通常按照“最多获取一次”的方式分享文件，也就是说，用户一旦下载了某个文件，就不太可能再去下载同一个文件，这导致了最流行的文件的实际频率低于齐普夫分布所预期的频率。

齐普夫分布与帕累托分布的关系

齐普夫分布可视为帕累托分布的离散解释。它可以通过变换帕累托分布的坐标轴来表示。因此，齐普夫分布可以写成：$R \propto F^{\frac{-1}{\theta}}$。也就是说，齐普夫分布中的 θ、帕累托分布中的 κ、幂律分布中的 η 的关系为：

$$\kappa = \eta - 1 = \frac{1}{\theta}$$

这个关系表明，这些不同的分布实际上是通过参数相互转换的不同表达方式。

这表明在齐普夫分布中，少数非常流行的项目会得到大量的引用，体现了一个强烈的倾斜或偏态分布。在许多现实世界的数据集中，通常是小部分的元素占据了大部分的活动或资源。这是齐普夫分布的一个典型特征，也是帕累托法则的表现。

形状参数 θ 决定了分布的倾斜程度。不同的 θ 值将导致不同程度的倾斜。在许多实证研究中观察到了类似的倾斜现象，例如文中未给出的图 2，可能展示了某些主机对 Web 服务器的请求情况，其中少数主机发起了大多数请求。这种现象在文献中通常被称为帕累托法则、帕累托原理或者 80/20 规则，即 20% 的原因往往会产生 80% 的结果。

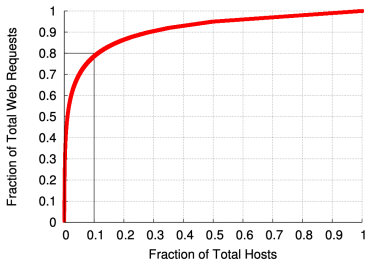

图 2.1: Fig. 2. 帕累托原则：图中显示的是 WWW 2007 会议网络服务器在一年时间内的请求分布情况。我们观察到，排名前 10% 的主机占了网络请求总量的 80%。这体现了帕累托原则，即大部分网络请求是由少数主机发出的。

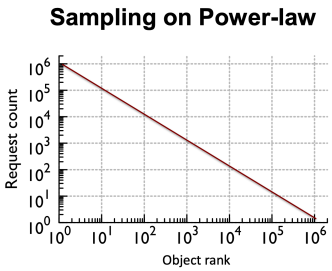

图 2.2: 从图表中识别齐普夫定律并计算参数

2.4 从图表中识别齐普夫定律并计算其参数

在双对数坐标纸上绘制数据，即对数-对数图表。横轴为元素的排名（R），纵轴为相应的频率（F）。对数变换是为了线性化幂律关系，齐普夫定律预测这样的图应该呈现出一条直线。

对数据点进行线性拟合，以获得最佳拟合直线。这条线的斜率（在对数-对数图上）将与齐普夫分布的参数 θ 相关。

斜率 m 通常会是负数，因为随着排名的增加，频率通常会下降。在齐普夫定律中，这个斜率 m 与 θ 相关，通常有 $m = -1/\theta$。因此，通过测量这个斜率，你可以计算出 $\theta = -1/m$。

2.5 齐普夫定律的含义

在互联网测量中，齐普夫定律的应用包括但不限于 Web 缓存的有效性，它依赖于 Web 对象及其大小的非均匀流行度分布。Web 访问已被证明遵循齐普夫定律，这在 Web 缓存架构的设计中非常重要，因为它允许设计者计算出近似的缓存大小以实现期望的命中率。适当的缓存大小和适当的替换策略可以实现高缓存命中率。齐普夫定律对于预测对象被访问的概率也很有用。

2.5.1 Facebook Haystack 系统的分布式缓存案例研究

2.5.2 不同缓存级别的内容受欢迎程度如何变化?

第 3 章　WiFi 干扰效应

3.1　Physical-layer characteristics interferers 物理层特征干扰源

3.1.1　Spectrogram 频谱图

3.1.2　Dutycycle 占空比

3.2　频谱图和占空比如何影响 WiFi 流量

3.3　基于场景的实验装置分析干扰对数据、视频和语音等各类流量的影响

第 4 章　交通流量

4.1　了解并计算用于研究互联网流量的主机级和流量级的各种指标

4.1.1　Flow size

4.1.2　Flow duration

4.1.3　Flow rate

4.1.4　Transfer volume

4.1.5　Transfer rate

4.1.6　Heavy hitters

第二编

通信原理

第 5 章　无线通信

第 6 章　蜂窝网络

第 7 章　卫星网络

第 8 章　星链

第 9 章　海底光缆