Google 抓取工具（用户代理）概览

查看 Google 抓取网页时使用哪些漫游器

“抓取工具”是一个统称，泛指通过跟踪从一个网页指向另一个网页的链接自动发现并扫描网站的任何程序（如漫游器或“蜘蛛”程序）。Google 的主要抓取工具叫作 Googlebot。下表列出了在引荐来源网址日志中经常会看到的 Google 抓取工具的相关信息，以及在 robots.txt、漫游器元标记和 X-Robots-Tag HTTP 指令中指定这些抓取工具的方式。

下表显示了 Google 的各种产品和服务所使用的抓取工具：

- 当您为网站编写抓取规则时，应在 robots.txt 文件中的 `User-agent:` 行内使用用户代理令牌，以便与某种类型的抓取工具匹配。如表中所示，某些抓取工具有多个令牌，但您只需使用与相应抓取工具匹配的 1 个令牌，即可应用规则。此列表并不完整，但涵盖了您在自己的网站上可能会看到的大多数抓取工具。
- 完整的用户代理字符串是对抓取工具的完整描述，会出现在请求和您的网站日志中。

> 这些值可能会被假冒。如果您需要验证访问者是否为 Googlebot，则应使用反向 DNS 查找。

抓取工具	用户代理令牌（产品令牌）	完整的用户代理字符串
APIs-Google	`APIs-Google`	`APIs-Google (+https://developers.google.com/webmasters/APIs-Google.html)`
AdSense	`Mediapartners-Google`	`Mediapartners-Google`
AdsBot Mobile Web Android （检查 Android 网页广告质量）	`AdsBot-Google-Mobile`	`Mozilla/5.0 (Linux; Android 5.0; SM-G920A) AppleWebK (KHTML, like Gecko) Chrome Mobile Safari (compatible AdsBot-Google-Mobile; +http://www.google.com/mobile/adsbot.html)`
AdsBot Mobile Web （检查 iPhone 网页广告质量）	`AdsBot-Google-Mobile`	`Mozilla/5.0 (iPhone; CPU iPhone OS 9_1 like Mac OS X AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13B143 Safari/601.1 (compatible; AdsBot-Googl Mobile; +http://www.google.com/mobile/adsbot.html)`
AdsBot （检查桌面设备网页广告质量）	`AdsBot-Google`	`AdsBot-Google (+http://www.google.com/adsbot.html)`
Googlebot	• `Googlebot-`	`Googlebot-Image/1.0`

Images	• Image • Googlebot	
Googlebot News	• Googlebot-News • Googlebot	Googlebot-News
Googlebot Video	• Googlebot-Video • Googlebot	Googlebot-Video/1.0
Googlebot（桌面版）	Googlebot	• Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html) • Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; Googlebot/2.1; +http://www.google.com/bot.html) Chrome/W.X.Y.Z‡ Safari/537.36 或（很少使用）： • Googlebot/2.1 (+http://www.google.com/bot.html)
Googlebot（智能手机版）	Googlebot	Mozilla/5.0 (Linux; Android 6.0.1; Nexus 5X Build/MMB29P) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/*W.X.Y.Z*‡ Mobile Safari/537.36 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
Mobile AdSense	Mediapartners-Google	（各类移动设备）(compatible; Mediapartners-Google/2.1; +http://www.google.com/bot.html)
Mobile Apps **Android** （检查 Android 应用页面广告质量。遵守 AdsBot-Google 漫游器规则。）	AdsBot-Google-Mobile-Apps	AdsBot-Google-Mobile-Apps
Feedfetcher	FeedFetcher-Google 不遵循 robots.txt	FeedFetcher-Google; (+http://www.google.com/feedfetcher.html)

	规则 - 查看原因	

Google Read Aloud	`Google-Read-Aloud`
	不遵循 robots.txt 规则 - 查看原因
		• 现用代理：
		`Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/41.0.2272.118 Safari/53 (compatible; Google-Read-Aloud; +https://support.google.com/webmasters/answer/1061`
		• 曾用代理（已弃用）：
		`google-speakr`

Duplex on the Web	`DuplexWeb-Google` 可能会忽略 * 用户代理通配符 - 查看原因	`Mozilla/5.0 (Linux; Android 8.0; Pixel 2 Build/OPD3.170816.012; DuplexWeb-Google/1.0) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/74.0.3729.131 Mobile Safari/537.36`
Google Favicon （检索各种服务的网站元素）	`Google Favicon` 对于用户发起的请求，会忽略 robots.txt 规则	`Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/49.0.2623.75 Safari/537. Google Favicon`

‡ 用户代理中的 **Chrome/W.X.Y.Z**

表中的用户代理字符串中有时候会出现 **Chrome/W.X.Y.Z** 字符串，W.X.Y.Z 实际上是代表该用户代理使用的 Chrome 浏览器版本的占位符：例如，41.0.2272.96。随着时间的推移，此版本号会增大，以便与 Googlebot 使用的最新 Chromium 发行版本相符。

如果您要搜索日志或过滤服务器以查找使用此格式的用户代理，您或许应该用通配符表示版本号，而不是指定确切的版本号。

robots.txt 中的用户代理

如果 Google 在 robots.txt 文件中识别出多个用户代理，它会跟踪最具体的用户代理。如果您希望 Google 的所有抓取工具都能够抓取您的网页，则您根本不需要 robots.txt 文件。如果您希望禁止或允许 Google 的所有抓取工具访问您的某些内容，您只需将 Googlebot 指定为用户代理即可。例如，如果您希望 Google 搜索范围包括您的所有网页，并且您的网页能够显示 AdSense 广告，您就不需要 robots.txt 文件。同样，如果您希望 Google 的所有抓取工具都不能访问您的某些网页，您可以禁止用户代理 Googlebot，这会一并禁止 Google 的所有其他用户代理。

不过，如果您希望更加精确地控制抓取范围，您可以将设置进一步具体化。例如，您可能希望 Google 搜索范围包括您的所有网页，但不希望 Google 抓取您个人目录中的图片。在这种情况下，您可以使用 robots.txt 禁止用户代理 Googlebot-image 抓取 /personal 目录中的文件（同时允许 Googlebot 抓取所有文件），具体如下：

```
User-agent: Googlebot
Disallow:

User-agent: Googlebot-Image
Disallow: /personal
```

再举个例子，假设您希望自己的所有网页上都显示广告，但不希望这些网页出现在 Google 搜索结果中。这时，您可以禁止 Googlebot，同时允许 Mediapartners-Google，具体如下：

```
User-agent: Googlebot
Disallow: /

User-agent: Mediapartners-Google
Disallow:
```

漫游器元标记中的用户代理

某些网页会使用多个漫游器 `meta` 标记为不同的抓取工具分别指定指令，如下所示：

```
<meta name="robots" content="nofollow"><meta name="googlebot" content="noindex">
```

在此示例中，Google 会使用所有否定指令，因此 Googlebot 将同时遵循 `noindex` 和 nofollow 指令。详细了解如何控制 Google 抓取您的网站并将其编入索引。
