目录

[Robots.txt文件 2](#_Toc145321460)

[基本格式 2](#_Toc145321461)

[Sitemap指令 2](#_Toc145321462)

[Sitemap 4](#_Toc145321463)

[站点地图索引示例 4](#_Toc145321464)

[标记 5](#_Toc145321465)

# Robots.txt文件

Robots.txt 档案能够告诉搜寻引擎检索器，可存取网站上的哪些网址。这个档案主要用来避免网站因要求过多而超载，而不是让特定网页无法出现在 Google 搜寻结果。

简单来说，Robots.txt文件是用来告诉搜索引擎，网站上的哪些页面可以抓取，哪些页面不能抓取。

robots.txt 文件应位于网站的根目录下。因此，对于网站 www.example.com，robots.txt文件的路径应为www.example.com/robots.txt。【**domain.com/robots.txt】**

#### 创建流程

1. 创建一个名为robots.txt 的文件。

文件必须命名为 robots.txt，且网站只能有 1 个 robots.txt 文件。

robots.txt 文件必须是采用 UTF-8 编码（包括 ASCII）的文本文件。Google 可能会忽略不属于 UTF-8 范围的字符，从而可能会导致 robots.txt 规则无效。

1. 向robots.txt文件添加规则。
2. 将robots.txt文件上传到网站的根目录。

robots.txt文件必须位于其要应用到的网站主机的根目录下。例如，若要控制对 https://www.example.com/下所有网址的抓取，就必须将robots.txt文件放在 https://www.example.com/robots.txt下，一定不能将其放在子目录中（例如 https://example.com/pages/robots.txt 下）。

robots.txt 文件可以位于子网域（例如 https://website.example.com/robots.txt）。

robots.txt 文件仅适用于所在的协议、主机和端口内的路径。也就是说，https://example.com/robots.txt 中的规则仅适用于 https://example.com/ 中的文件，而不适用于子网域（如 https://m.example.com/）或备用协议（如 http://example.com/）。

1. 测试robots.txt文件。

### 基本格式

Sitemap: [URL location of sitemap

User-agent: [bot identifier]

[directive 1]

[directive 2]

[directive ...]

User-agent: [another bot identifier]

[directive 1]

[directive 2]

[directive ...]

简而言之，通过在文件中指定user-agents（用户代理）和directives（指令）来为搜索引擎蜘蛛分配抓取规则。

## Sitemap指令

使用此指令来标记网站地图所的位置。如果对网站地图不熟悉，它通常会包含需要被搜索引擎抓取&索引的所有页面链接。

下面就是一个使用Sitemap指令的robots.txt文件：

Sitemap: **https://www.domain.com/sitemap.xml**

User-agent: \*

Disallow: /blog/

Allow: /blog/post-title/

在这个例子中，搜索引擎可以访问： /blog/allowed-post

但是它不能访问：

/blog/another-post

/blog/yet-another-post

/blog/download-me.pdf

#### User-agents（用户代理）

Google: Googlebot

Google Images: Googlebot-Image

Bing: Bingbot

Yahoo: Slurp

Baidu: Baiduspider

DuckDuckGo: DuckDuckBot

#### Disallow指令

使用此指令来规定搜索引擎不要访问特定路径的文件和页面

例如，如果想阻止所有搜索引擎访问博客及其所有帖子：

User-agent: \*

Disallow: /blog

#### Allow指令

使用此指令来规定搜索引擎需要访问特定路径的文件和页面——即使在一个被disallow指令屏蔽的路径中也可以使用。如果屏蔽特定文章以外的所有文章页面：

User-agent: \*

Disallow: /blog

Allow: /blog/allowed-post

部分情况disallow指令会与allow指令互相冲突，对与谷歌和必应来说，它们会遵循指令字符较长的那一个，如：

Disallow: /blog/ (6 字符)

Allow: /blog (5 字符)

如果disallow指令和allow指令长度一样，那么限制范围较小的指令会胜出，这种情况就是allow指令。

/blog (没有斜杠为后缀) 也是可以被抓取的。

严格来说，这只适用于谷歌和必应搜索引擎。其它搜索引擎会遵循第一条指令，这种情况就是disallow指令。

# Sitemap

每个文本文件最多可包含50,000个 URL，且不得大于50MB（52,428,800 字节），必须使用UTF-8编码。

站点地图索引文件中引用的站点地图必须与站点地图索引文件位于同一目录中，或者位于网站层次结构中的更低一级目录中。

例如，如果站点地图索引文件位于https://example.com/public/sitemap\_index.xml，它只能包含位于相同或更深目录中的站点地图，如<https://example.com/public/shared/>...

### 站点地图索引示例

以下示例是两个XML格式的站点地图索引：

<?xml version="1.0" encoding="UTF-8"?>

<sitemapindex xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

<sitemap>

<loc>https://www.example.com/sitemap1.xml.gz</loc>

</sitemap>

<sitemap>

<loc>https://www.example.com/sitemap2.xml.gz</loc>

</sitemap>

</sitemapindex>

<?xml version="1.0" encoding="UTF-8"?>

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">

<url>

<loc>https://www.example.com/</loc>

<lastmod>2017-06-14T19:55:25+02:00</lastmod>

<changefreq>daily</changefreq>

<priority>0.80</priority>

</url>

<url>

<loc>https://www.example.com /blog/</loc>

<changefreq>weekly</changefreq>

<priority>0.60</priority>

<image:image>

<image:loc>https://cdn.example.com / files/001.jpg</image:loc>

<image:title>Example </image:title>

<image:caption> Example image </image:caption>

</image:image>

</url>

</urlset>

## 标记

#### 必需的标记

sitemapindex XML 树的父标记。它包含所有其他标记。

sitemap 文件中列出的每个站点地图的父标记。它是sitemapindex标记的第一个也是唯一一个直接子标记。

urlset 封装文件并引用当前的协议标准。

url 页面的URL。如果Web 服务器需要，此URL必须以协议（例如 http）开头并以斜杠结尾。该值必须小于 2,048 个字符。每个 URL 条目的父标记。其余标签是该标签的子标签。

loc 站点地图的位置（网址）。它是sitemap标记的第一个也是唯一一个子级。一个站点地图索引文件最多可以包含50,000个loc标记。

#### 可选标记

lastmod 页面的最后修改日期。有助于 Google 安排何时抓取您的站点地图。应采用 W3C 日期时间格式，允许省略时间部分，并使用 YYYY-MM-DD。

changefreq 页面可能更改的频率。此标签被视为提示而不是命令，可能与它们抓取页面的频率不完全相关。always，hourly，daily，weekly，monthly，yearly，never。

priority 该URL 相对于您网站上其他 URL 的优先级。有效值范围为 0.0 到 1.0。该值不会影响页面与其他网站上的页面的比较方式，只会让爬虫知道哪些网页相对更被重视。默认优先级是0.5。

**示例：**

<?xml version="1.0" encoding="UTF-8"?>

<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"

<url>

<loc>https://www.mdpi.com/</loc>

<lastmod>2023-09-08T09:58:09+00:00</lastmod>

<priority>1.00</priority>

</url>

<url>

<loc>https://www.mdpi.com/about/journals</loc>

<lastmod>2023-09-08T09:58:09+00:00</lastmod>

<priority>0.80</priority>

</url>

#### 在GSC提交Sitemap的原因

1. 超过500个页面的中大型网站。网站越大，Googlebot 错过某些页面的可能性就越大，尤其是对于电子商务网站或市场等大型网站。
2. 站点有大量内容页面、独立页面或未链接在一起的页面的存档。
3. 新网站，没有很多/任何反向链接。由于Googlebot依赖链接来发现页面，因此它通常还使用来自其他域的链接来访问网站，如果其他网站尚未开始链接到新网站，则需要更长的时间才能建立索引。
4. 网站有大量多媒体内容，例如视频、图像，或者在 Google 新闻中显示。在某些情况下，Google 可以使用站点地图来获取额外的上下文。

每个文本文件最多可包含 50,000 个 URL，且不得大于 50MB（52,428,800 字节）。