

【哥飞SEO教程】多语言网站 robots.txt 设置指南：如何正确阻止不希望被抓取的页面

原创 我是哥飞 哥飞 2024-08-10 12:18:20 广东

大家好，我是哥飞。
今天哥飞在例行查看 Google Search Console （下文简称 GSC）数据时，发现最近哥飞上线的一个新网站，未编入索引的页面数量有点儿多。



再看具体原因，发现被 `noindex` 标记的页面有一百多个。

<div><div>i</div><div>网页未被编入索引的原因</div><div>未编入索引的网页无法显示在 Google 中</div></div>				
原因	来源 ?	验证 ↓	趋势	网页
被“noindex”标记排除了	网站	! 未启动		105

这就有点奇怪了，于是继续点进去看，到底是哪些网址出现问题了。

受影响的网页数

105



示例 ?



网址

上次抓取日期

https://domain.ai/ja/people/c2f4f6

2024年8月4日

https://domain.ai/tw/people/fcaa45

2024年8月4日

https://domain.ai/ja/people/fcaa45

2024年8月3日

https://domain.ai/hu/people/7e009b

2024年8月3日

https://domain.ai/ko/people/41eedc

2024年8月3日

看到这些网址列表，哥飞终于知道原因了。

之前哥飞要求小伙伴们把 /people/ 页面都暂时禁止抓取，在 robots.txt 设置了禁止抓取，也在页面里 meta 信息中返回了 noindex 标记。

所以默认语言下的 people 目录的确没抓取，但是多语言下的被抓取了。

而 robots.txt 是这么写的：

← → ↻  https://domain.ai/robots.txt

```
User-Agent: *  
Allow: *  
Disallow: /people/
```

看出问题来了吗？

上面的这种写法，只会禁止默认语言下的 `/people/` 目录下的页面。

但这个网站是加了多语言支持的，并且是用子目录形式放多语言的，上面的规则无法禁止像日语 `/ja/people/`、韩语 `/ko/people/` 等语言下的 `people` 目录。

那么，为了达到禁止多语言下的 `people` 目录，你的 `robots.txt` 需要改成下面这样才行：

- `User-Agent: *`
- `Allow: /`
- `Disallow: /people/`
- `Disallow: /ja/people/`
- `Disallow: /fr/people/`
- `Disallow: /ko/people/`
- `Disallow: /zh/people/`

有多少种语言就写多少行禁止规则。

另外注意，别偷懒，如果你写成下面这样，很有可能会出现不可预知的问题：

- `User-Agent: *`
- `Allow: /`
- `Disallow: /*/people/`

举个例子，你有一个页面是 `/abc/def/people/` 也会被这个规则覆盖到，而这个页面其实你是想被抓取的。

所以，最好的方式是手动列出每一种语言，并且注意定期更新，如果你的网站增加了别的语言的支持，这里也需要增加一行。

注意，这里的 `people` 只是举例，你需要自己判断你的网站哪些页面不给抓取。一般来说，所有不是拿来获取流量的页面，都不应该被抓取。

好友@Aladdin 补充说，如果你用的是 Next.js 框架，在 `app` 目录下面可以用 `robots.ts/js` 来动态生成，里面就可以写逻辑了。

好友@涛涛 补充了参考文档：

<https://nextjs.org/docs/app/api-reference/file-conventions/metadata/robots#generate-a-robots-file>