**robots协议**

您检索的robots.txt，一般指robots协议，其他释义请[点击此处](http://baike.baidu.com/view/1280732.htm?force=1" \t "_blank)。

Robots协议（也称为爬虫协议、机器人协议等）的全称是“网络爬虫排除标准”（Robots Exclusion Protocol），网站通过Robots协议告诉搜索引擎哪些页面可以抓取，哪些页面不能抓取。

**目录**

1[简介](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa" \l "1)

2[原则](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#2)

3[功能](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3)

▪ [文件写法](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3_1)

▪ [文件用法](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3_2)

▪ [其它属性](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3_3)

▪ [标签](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3_4)

▪ [注意事项](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#3_5)

4[位置](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#4)

5[产生](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#5)

6[影响](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#6)

7[搜索引擎](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#7)

▪ [淘宝封杀](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#7_1)

▪ [京东封杀](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#7_2)

8[违反事例](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#8)

▪ [BE违规抓取eBay](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#8_1)

▪ [360搜索无视robots协议](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#8_2)

▪ [百度起诉360](http://baike.baidu.com/link?url=nUfkoOqCrEj6ujnRYTufJtnX9rjyEB3W4uMgkwXhxteFvusSqlYmdU-XBuiR4UTJBEJoeRL7YV78uNj4ehxjGKQvj0uaMCKMCv3JPCEiXL9OGvnkti40_Lf6jzC0ocxyvsKMK_iAuiX20PmJ5zVDaa#8_3)

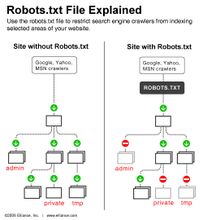
**1****简介**

robots.txt文件是一个[文本](http://baike.baidu.com/view/300107.htm" \t "_blank)文件，使用[任何](http://baike.baidu.com/view/1168473.htm)一个常见的文本编辑器，比如[Windows](http://baike.baidu.com/view/4821.htm)系统自带的Notepad，就可以创建和编辑它[1]。[robot](http://baike.baidu.com/view/248611.htm" \t "_blank)s.txt是一个协议，而不是一个[命令](http://baike.baidu.com/view/446604.htm" \t "_blank)。robots.txt是[搜索引擎](http://baike.baidu.com/view/1154.htm" \t "_blank)中访问网站的时候要查看的第一个文件。robots.txt文件告诉[蜘蛛程序](http://baike.baidu.com/view/995492.htm" \t "_blank)在服务器上什么文件是可以被查看的。

当一个搜索蜘蛛访问一个[站点](http://baike.baidu.com/view/391109.htm" \t "_blank)时，它会首先[检查](http://baike.baidu.com/view/386571.htm)该站点[根目录](http://baike.baidu.com/view/1061154.htm)下是否存在robots.txt，如果存在，搜索机器人就会按照该文件中的内容来确定访问的范围；如果该文件不存在，所有的搜索蜘蛛将能够访问网站上所有没有被口令保护的页面。百度官方建议，仅当您的网站包含不希望被[搜索引擎收录](http://baike.baidu.com/view/3192508.htm" \t "_blank)的内容时，才需要使用robots.txt文件。如果您希望搜索引擎收录网站上所有内容，请勿建立robots.txt文件。

如果将[网站](http://baike.baidu.com/view/4232.htm)视为酒店里的一个房间，robots.txt就是主人在房间门口悬挂的“[请勿打扰](http://baike.baidu.com/view/1884479.htm)”或“欢迎打扫”的提示牌。这个文件告诉来访的搜索引擎哪些房间可以进入和参观，哪些房间因为存放贵重物品，或可能涉及住户及访客的隐私而不对搜索引擎开放。但robots.txt不是[命令](http://baike.baidu.com/view/446604.htm" \t "_blank)，也不是[防火墙](http://baike.baidu.com/view/3067.htm)，如同守门人无法阻止窃贼等恶意闯入者。

**2****原则**

[](http://baike.baidu.com/picview/9274458/9307430/0/6c224f4a20a44623c2e0e0e29822720e0cf3d70c.html)

robots协议原理

Robots协议是国际互联网界通行的道德规范，基于以下原则建立：

1、[搜索](http://baike.baidu.com/view/8638.htm" \t "_blank)技术应服务于人类，同时尊重信息提供者的意愿，并维护其隐私权；

2、网站有义务保护其使用者的个人信息和[隐私](http://baike.baidu.com/view/350273.htm" \t "_blank)不被侵犯。

**3****功能**

[Robots协议](http://baike.baidu.com/view/9274458.htm)用来告知搜索引擎哪些页面能被抓取，哪些页面不能被抓取；可以屏蔽一些网站中比较大的文件，如：[图片](http://baike.baidu.com/view/19159.htm)，音乐，[视频](http://baike.baidu.com/view/16215.htm)等，节省[服务器](http://baike.baidu.com/view/899.htm)带宽；可以屏蔽站点的一些[死链接](http://baike.baidu.com/view/1880779.htm)。方便搜索引擎抓取网站内容；设置[网站地图](http://baike.baidu.com/view/20280.htm)连接，方便引导蜘蛛爬取页面。

**文件写法**

User-agent: \* 这里的\*代表的所有的搜索引擎种类，\*是一个[通配符](http://baike.baidu.com/view/451.htm" \t "_blank)

Disallow: /admin/ 这里定义是禁止爬寻[admin](http://baike.baidu.com/view/11220.htm" \t "_blank)目录下面的目录

Disallow: /require/ 这里定义是禁止爬寻require目录下面的[目录](http://baike.baidu.com/view/637.htm" \t "_blank)

Disallow: /ABC/ 这里定义是禁止爬寻ABC目录下面的目录

Disallow: /cgi-bin/\*.htm 禁止访问/cgi-bin/目录下的所有以".htm"为后缀的URL(包含[子目录](http://baike.baidu.com/view/1765050.htm" \t "_blank))。

Disallow: /\*?\* 禁止访问网站中所有包含问号 (?) 的网址

Disallow: /.jpg$ 禁止抓取网页所有的.jpg格式的[图片](http://baike.baidu.com/view/19159.htm" \t "_blank)

Disallow:/ab/adc.html 禁止爬取ab文件夹下面的adc.html文件。

Allow: /cgi-bin/　这里定义是允许爬寻cgi-bin目录下面的目录

Allow: /tmp 这里定义是允许爬寻tmp的整个目录

Allow: .htm$ 仅允许访问以".htm"为后缀的URL。

Allow: .gif$ 允许抓取网页和gif格式图片

Sitemap: 网站地图 告诉爬虫这个页面是网站地图

**文件用法**

例1. 禁止所有搜索引擎访问网站的任何部分

User-agent: \*

Disallow: /

实例分析：[淘宝网](http://baike.baidu.com/view/1590.htm" \t "_blank)的 Robots.txt文件

User-agent: Baiduspider

Disallow: /

User-agent: baiduspider

Disallow: /

很显然淘宝不允许百度的[机器人](http://baike.baidu.com/view/2788.htm" \t "_blank)访问其网站下其所有的目录。

例2. 允许所有的[robot](http://baike.baidu.com/view/248611.htm" \t "_blank)访问 (或者也可以建一个空文件 “/robots.txt” file)

User-agent: \*

Allow:　/

例3. 禁止某个搜索引擎的访问

User-agent: BadBot

Disallow: /

例4. 允许某个搜索引擎的访问

User-agent: Baiduspider

allow:/

例5.一个简单例子

在这个例子中，该网站有三个目录对搜索引擎的访问做了限制，即搜索引擎不会访问这三个目录。

需要注意的是对每一个目录必须分开声明，而不要写成 “Disallow: /cgi-bin/ /tmp/”。

User-agent:后的\*具有特殊的含义，代表“any robot”，所以在该文件中不能有“Disallow: /tmp/\*” or “Disallow:\*.gif”这样的记录出现。

User-agent: \*

Disallow: /cgi-bin/

Disallow: /tmp/

Disallow: /~joe/

Robot特殊参数：

允许 Googlebot：

如果您要拦截除[Googlebot](http://baike.baidu.com/view/1848615.htm)以外的所有漫游器不能访问您的网页，可以使用下列语法：

User-agent:

Disallow: /

User-agent: Googlebot

Disallow:

Googlebot 跟随指向它自己的行，而不是指向所有漫游器的行。

“Allow”扩展名：

Googlebot 可识别称为“Allow”的 robots.txt 标准扩展名。其他搜索引擎的漫游器可能无法识别此扩展名，因此请使用您感兴趣的其他搜索引擎进行查找。“Allow”行的作用原理完全与“Disallow”行一样。只需列出您要允许的目录或页面即可。

您也可以同时使用“Disallow”和“Allow”。例如，要拦截[子目录](http://baike.baidu.com/view/1765050.htm" \t "_blank)中某个页面之外的其他所有页面，可以使用下列条目：

User-agent: Googlebot

Allow: /folder1/myfile.html

Disallow: /folder1/

这些条目将拦截 folder1 目录内除 myfile.html 之外的所有页面。

如果您要拦截 Googlebot 并允许 Google 的另一个漫游器（如 Googlebot-Mobile），可使用”Allow”规则允许该漫游器的访问。例如：

User-agent: Googlebot

Disallow: /

User-agent: Googlebot-Mobile

Allow:

使用 \* 号匹配字符序列：

您可使用星号 (\*) 来匹配字符序列。例如，要拦截对所有以 private 开头的子目录的访问，可使用下列条目：　User-Agent: Googlebot

Disallow: /private\*/

要拦截对所有包含问号 (?) 的网址的访问，可使用下列条目：

User-agent: \*

Disallow: /\*?\*

使用 $ 匹配网址的结束字符

您可使用 $[字符](http://baike.baidu.com/view/263416.htm)指定与网址的结束字符进行匹配。例如，要拦截以 .asp 结尾的网址，可使用下列条目：　User-agent: Googlebot

Disallow: /\*.asp$

您可将此模式匹配与 Allow 指令配合使用。例如，如果 ? 表示一个会话 ID，您可排除所有包含该 ID 的网址，确保 Googlebot 不会抓取重复的网页。但是，以 ? 结尾的网址可能是您要包含的网页版本。在此情况下，可对 robots.txt 文件进行如下设置：

User-agent: \*

Allow: /\*?$

Disallow: /\*?

Disallow: / \*?

一行将拦截包含 ? 的网址（具体而言，它将拦截所有以您的域名开头、后接任意字符串，然后是问号 (?)，而后又是任意字符串的网址）。

Allow: /\*?$ 一行将允许包含任何以 ? 结尾的网址（具体而言，它将允许包含所有以您的域名开头、后接任意字符串，然后是问号 (?)，问号之后没有任何字符的网址）。

尽管robots.txt已经存在很多年了，但是各大搜索引擎对它的解读都有细微差别。Google与百度都分别在自己的站长工具中提供了robots工具。如果您编写了robots.txt文件，建议您在这两个工具中都进行测试，因为这两者的解析实现确实有细微差别[1]。

**其它属性**

1. Robot-version: 用来指定robot协议的版本号

例子: Robot-version: Version 2.0

2.Crawl-delay：雅虎YST一个特定的扩展名，可以通过它对我们的抓取程序设定一个较低的抓取请求频率。您可以加入Crawl-delay:xx指示，其中，“XX”是指在crawler程序两次进入站点时，以秒为单位的最低延时。

3. Visit-time:只有在visit-time指定的时间段里,robot才可以访问指定的URL,否则不可访问.

例子: Visit-time: 0100-1300 #允许在凌晨1:00到13:00访问

4. Request-rate: 用来限制URL的读取频率

例子: Request-rate: 40/1m 0100 - 0759 在1:00到07:59之间,以每分钟40次的频率进行访问

Request-rate: 12/1m 0800 - 1300 在8:00到13:00之间,以每分钟12次的频率进行访问

**标签**

Robots.txt文件主要是限制整个站点或者目录的搜索引擎访问情况，而Robots Meta标签则主要是针对一个个具体的页面。和其他的[META标签](http://baike.baidu.com/view/740572.htm" \t "_blank)（如使用的语言、页面的描述、[关键词](http://baike.baidu.com/view/10083.htm)等）一样，Robots Meta标签也是放在页面中，专门用来告诉搜索引擎ROBOTS如何抓取该页的内容。

Robots Meta标签中没有大小写之分，name=”Robots”表示所有的搜索引擎，可以针对某个具体搜索引擎写为name=”BaiduSpider”。content部分有四个指令选项：index、noindex、follow、nofollow，指令间以“,”分隔。

index指令告诉搜索机器人抓取该页面；

follow指令表示搜索机器人可以沿着该页面上的链接继续抓取下去；

Robots Meta标签的缺省值是index和follow，只有inktomi除外，对于它，缺省值是index、[nofollow](http://baike.baidu.com/view/1584081.htm" \t "_blank)。

**注意事项**

上述的robots.txt和Robots Meta标签限制[搜索引擎机器人](http://baike.baidu.com/view/2884260.htm" \t "_blank)（ROBOTS）抓取站点内容的办法只是一种规则，需要搜索引擎机器人的配合才行，并不是每个ROBOTS都遵守的。目前看来，绝大多数的搜索引擎机器人都遵守robots.txt的规则，而对于Robots　META标签，支持的并不多，但是正在逐渐增加，如著名搜索引擎GOOGLE就完全支持，而且GOOGLE还增加了一个指令“archive”，可以限制GOOGLE是否保留[网页快照](http://baike.baidu.com/view/522994.htm" \t "_blank)。

**4****位置**

robots.txt文件应该放置在网站根目录下。举例来说，当spider访问一个网站时，首先会检查该网站中是否存在robots.txt这个文件，如果 Spider找到这个文件，它就会根据这个文件的内容，来确定它访问权限的范围。

**wordpress的robots位置**

没有在[wordpress](http://baike.baidu.com/view/23618.htm)网站根节目上传过robots.txt,当搜寻引擎和用户拜访某个文件时，wordpress程序会主动生成一个robots.txt给搜寻引擎和用户;若是我们上传编写的robots.txt到网站根节目，用户和搜寻引擎蛛蛛拜访的就是我们上传的文件，wordpress就不会再产生那个文件了。只有服务器找不到robots的时候wordpress才会生成这个文件。

**5****产生**

robots.txt并不是某一个公司制定的，而是早在20世纪93、94年就早已出现，当时还没有Google。真实Robots协议的起源，是在互联网从业人员的公开邮件组里面讨论并且诞生的。即便是今天，互联网领域的相关问题也仍然是在一些专门的邮件组中讨论，并产生（主要是在美国）。

1994年6月30日，在经过搜索引擎人员以及被搜索引擎抓取的网站站长共同讨论后，正式发布了一份行业规范，即robots.txt协议。在此之前，相关人员一直在起草这份文档，并在世界互联网技术邮件组发布后，这一协议被几乎所有的搜索引擎采用，包括最早的[altavista](http://baike.baidu.com/view/93295.htm" \t "_blank)，infoseek，后来的[google](http://baike.baidu.com/view/105.htm" \t "_blank)，[bing](http://baike.baidu.com/view/2491177.htm" \t "_blank)，以及中国的百度，[搜搜](http://baike.baidu.com/view/101297.htm" \t "_blank)，[搜狗](http://baike.baidu.com/view/24982.htm)等公司也相继采用并严格遵循。

Robot，又称Spider，是搜索引擎自动获取网页信息的电脑程序的通称。Robots协议的核心思想就是要求Robot程序不要去检索那些站长们不希望被直接搜索到的内容。将约束Robot程序的具体方法规范成格式代码，就成了Robots协议。一般来说，网站是通过Robots.txt文件来实现Robots协议。

自有搜索引擎之日起，Robots协议已是一种目前为止最有效的方式，用自律维持着网站与搜索引擎之间的平衡，让两者之间的利益不致过度倾斜。它就像一个钟摆，让互联网上的搜索与被搜索和谐相处。

[](http://baike.baidu.com/picview/9274458/9307430/0/55e736d12f2eb9389ec93176d5628535e4dd6f89.html)

维基百科上描述的robots协议的最原始起源

**6****影响**

Robots协议是网站出于安全和隐私考虑，防止搜索引擎抓取敏感信息而设置的。搜索引擎的原理是通过一种爬虫spider程序，自动搜集互联网上的网页并获取相关信息。而鉴于[网络安全](http://baike.baidu.com/view/17495.htm" \t "_blank)与隐私的考虑，每个网站都会设置自己的Robots协议，来明示搜索引擎，哪些内容是愿意和允许被搜索引擎收录的，哪些则不允许。搜索引擎则会按照Robots协议给予的权限进行抓取。

Robots协议代表了一种契约精神，互联网企业只有遵守这一规则，才能保证网站及用户的隐私数据不被侵犯。违背Robots协议将带来巨大安全隐忧——此前，曾经发生过这样一个真实的案例：国内某公司员工郭某给别人发了封求职的电子邮件，该Email存储在某邮件服务公司的服务器上。因为该网站没有设置robots协议，导致该Email被搜索引擎抓取并被网民搜索到，为郭某的工作生活带来极大困扰。

如今，在中国国内互联网行业，正规的大型企业也都将Robots协议当做一项行业标准，国内使用Robots协议最典型的案例，就是淘宝网拒绝百度搜索、京东拒绝一淘搜索。不过，绝大多数中小网站都需要依靠搜索引擎来增加流量，因此通常并不排斥搜索引擎，也很少使用Robots协议。

北京市汉卓律师事务所首席律师赵虎表示，Robots协议是维护互联网世界隐私安全的重要规则。如果这种规则被破坏，对整个行业就是灭顶之灾。

**7****搜索引擎**

百度对robots.txt是有反应的，但比较慢，在减少禁止目录抓取的同时也减少了正常目录的抓取。

原因应该是入口减少了，正常目录收录需要后面再慢慢增加。

Google对robots.txt反应很到位，禁止目录马上消失了，部分正常目录收录马上上升了。/comment/目录收录也下降了，还是受到了一些老目标减少的影响。

搜狗抓取呈现普遍增加的均势，部分禁止目录收录下降了。

总结一下：Google似乎最懂站长的意思，百度等其它搜索引擎只是被动的受入口数量影响了。

**淘宝封杀**

2008年9月8日，淘宝网宣布封杀百度爬虫，百度忍痛遵守爬虫协议。因为一旦破坏协议，用户的隐私和利益就无法得到保障，搜索网站就谈不到人性关怀。

**京东封杀**

2011年10月25日，京东商城正式将一淘网的搜索爬虫屏蔽，以防止一淘网对其的内容抓取。

**8****违反事例**

**BE违规抓取eBay**

在12年前，美国加州北部的联邦地方法院就在著名的eBayVS. Bidder's Edge案中（NO.C-99-21200RMW,2000 U.S Dist. LEXI 7282），引用robots协议对案件进行裁定。 Bidder's Edge（简称BE）成立于1997年，是专门提供拍卖信息的聚合网站。12年前，BE利用“蜘蛛”抓取来自eBay等各个大型拍卖网站的商品信息，放在自己的网站上供用户浏览，并获得可观的网站流量。

对于eBay来说，来自BE蜘蛛每天超过十万次的访问，给自己的服务器带来了巨大的压力。而虽然eBay早已设置了robots协议禁止BE蜘蛛的抓取，但BE却无视这个要求——原因很简单，据估算，BE网站69%的拍卖信息都来自eBay, 如果停止抓取eBay内容，这意味着BE将损失至少三分之一的用户。

数次沟通交涉未果后，2000年2月，忍无可忍的eBay终于一纸诉状，将BE告上联邦法庭，要求禁止BE的违规抓取行为。3个月后，受理此案的[美国](http://baike.baidu.com/view/2398.htm" \t "_blank)联邦法官Ronald M. Whyte在经过多方调查取证后做出裁定，认定BE侵权成立，禁止了BE在未经eBay允许的情况下，通过任何自动查询程序、网络蜘蛛等设置抓取eBay的拍卖内容。

在当时的庭审中，双方争议的焦点主要集中在“网站是否有权设置robots协议屏蔽其他网站蜘蛛的抓取”。被告BE认为，eBay的网站内容属于网民自创，因此是公共资源，eBay无权设立robots协议进行限制。然而，法院对这一说辞却并不认同。在法官看来：“eBay 的网站内容属于私有财产，eBay有权通过robots协议对其进行限制。”违规抓取的行为无异于“对于动产的非法侵入”。

也正是出于这一判断，即使当年BE还只是搜索了eBay计算机系统里的一小部分数据，其违反robots协议的抓取行为，仍然被判为侵犯了eBay将别人排除在其计算机系统以外的基本财产权。

作为美国历史上第一个保护互联网信息内容的法律裁定，eBay与BE的这起纠纷，成为网络侵权案的标志性案例，并在当时引发了美国互联网产业乃至社会的广泛讨论。SearchEngine Watch的知名专栏作家DannySullivan 曾专门在文章中指出，robots协议是规范搜索引擎爬虫行为的极少数约定之一，理应遵守，它不仅仅让整个互联网的开放性变成可能，最终也让整个互联网用户受益。

**360搜索无视robots协议**

2012年8月，360综合搜索被指违反robots协议：

[](http://baike.baidu.com/picview/9274458/9307430/0/c2fdfc039245d688ae55d325a4c27d1ed31b24df.html)

360甘冒行业道德风险 撕毁Robots自律协议

2012年8月29日，上线不到十日即轻松登上国内第二大搜索引擎宝座的[奇虎360](http://baike.baidu.com/view/1458023.htm" \t "_blank)搜索遭遇滑铁卢，在[百度](http://baike.baidu.com/view/262.htm" \t "_blank)一项打击违规抓取和匿名访问的措施上线后，[360搜索](http://baike.baidu.com/view/9199096.htm)通过山寨加绑架用户的模式被彻底揭穿，这个故事也正是对所谓“搜索引擎技术无用论”的当头棒喝。

“360综合搜索已超出Robots协议的‘底线’，不仅未经授权大量抓取百度、[google](http://baike.baidu.com/view/105.htm" \t "_blank)内容，还记录国内知名网游的后台订单、优惠码等，甚至一些用户的电子邮箱、帐号、密码也被360通过浏览器悄然记录在案”。2012年8月28日业内一位有10年搜索工程师经验的专业人士表示。[2]

奇虎360搜索8月16 日上线，已连续遭到业界“违规”的警告。不仅UI设计，搜索结果直接剽窃、篡改百度、google等搜索引擎，还不顾国际通行的Robots协议，抓取百度、google等搜索引擎的内容，很多商业网站和个人信息的私密内容也被抓取泄密。据悉，这种匿名访问和违规抓取其他网站内容的行为，不仅严重违反了行业底线，也伤害到了用户的体验。在国外，这种行为甚至会受到法律方面的制裁。“这已经严重触碰了Robots协议的底线!这些让如百度这样踏踏实实做技术的公司进行反击行动，是必然的!”

这位搜索工程师告诉记者，所谓Robots协议，通俗的讲就是网站会通过robots.txt协议来自主控制是否愿意被搜索引擎收录，或者指定搜索引擎只收录指定的内容。而搜索引擎会按照每个网站主给与自己的权限来进行抓取。这就好比，一个正常的人去到别人家里，需要先敲门，得到许可以后才能进入客厅。除非有主人的进一步许可和邀请，否则你不能擅自进入内室，或者在别人家里四处溜达。当然，强盗或者小偷例外。[3]

对于奇虎360综合搜索被爆出无视国际通行的robots协议，抓取百度、Google等搜索引擎内容，导致众多网站出于安全和隐私考虑不允许搜索引擎抓取的内网信息被泄露一事，资深互联网观察家洪波指出，做搜索就要遵守搜索行业公认的游戏规则，无视规则，肆意违反规则才是真正的不正当竞争，这种行为不能从法律和政府监管上及时制止，将引发行业大乱。

百度并没有禁止所有爬虫抓取问答、知道和贴吧的内容，百度只是禁止了不规矩的、存在潜在安全风险的爬虫，这恰是保护市场秩序，保护用户隐私的合理举措。2008年淘宝也曾禁止百度爬虫，而百度则严格遵守robots协议，停止抓取淘宝网内容，并没有以淘宝不正当竞争为借口违反robots协议。

中国工程院院士高文：Robots协议是搜索引擎在抓取网站信息时要遵守的一个规则，是国际互联网界通行的道德规范。网站主人通过Robots协议明示自动抓取程序，服务器上什么内容可以被抓取，什么内容不可以被抓取。这是保证互联网行业健康发展的自律准则，尽管不是强制命令，但业界都应该遵守。中国[互联网企业](http://baike.baidu.com/view/2336558.htm)要想在国际舞台受到更多关注、得到更大的发展，没有理由不遵守这个游戏规则。如果不遵守，不仅损害了违规企业自身的形象，也损害了中国互联网行业的整体声誉，得不偿失。

浙江大学计算机学院院长庄越挺：互联网网站页面，如同广阔农村中的一个菜园，各有其主。一般而言，访客进去逛逛无可厚非，但是如果主人在边界立下界碑：未经允许不得入内，这就意味着主人的意愿成为外界是否获准入园参观的标准。Robots协议就是这样一块界碑，它虽然不具法律效应，但是人们都普遍遵循。未经允许入园就参观不仅违反了游戏规则，也有违道德标准。同样的道理，违反Robots协议，等同于违背了搜索引擎的行业规范，以这种方式获取资源是一种不道德的竞争。

[哈尔滨工业大学](http://baike.baidu.com/view/13897.htm)教授刘挺：Robots协议是互联网中网站为搜索引擎所制定的内容抓取规则，体现了业界同行之间的相互信任。如果不遵循Robots协议，将会破其业界信任，使得网站的内容不愿意被抓取的内容被搜索引擎索引，最终伤害用户的利益。而清华大学计算机系教授马少平则指出，如果不遵守Robots协议，肆意抓取网站的内容，网站的信息被任意泄漏，将对互联网的良性发展产生巨大的破坏作用。

**百度起诉360**

百度诉奇虎360违反“Robots协议”抓取、复制其网站内容侵权一案，2013年10月16日上午在北京市第一中级人民法院开庭审理。百度方面认为，360搜索在未获得百度公司允许的情况下，违反业内公认的Robots协议，抓取百度旗下百度知道、百度百科、百度贴吧等网站的内容，已经构成了不正当竞争，并向奇虎索赔1亿元。

百度公关部郭彪向媒体表示，Robots协议是网站信息和网民隐私保护的国际通行规范之一，理应得到全球互联网公司的共同遵守。不尊重Robots协议将可能导致网民隐私大规模泄露。

郭彪表示，更严重的是，奇虎360还利用360浏览器等客户端，强行抓取网民的浏览数据和信息到搜索服务器，完全无视Robots协议。这一做法目前已经导致大量企业内网信息被泄露。2012年年底，百度工程师通过一个名为“鬼节捉鬼”的测试，证明了360浏览器存在私自上传“孤岛页面”等隐私内容到360搜索的行为。

360方面则认为，360搜索索引这些内容页面并不涉嫌侵犯百度的权益，实际上还为百度带来了大量的用户和流量，百度应该感谢360。

