搜索引擎频繁抓取导致网站挂掉

从上周开始，我博客就经常出现了Bad Request (Invalid Hostname)
错误，询问网站服务器商只得知网站的并发过高，从而被服务器商限制网站访问。可是我天天都会去看网站的流量统计，没有一点异常，怎么可能会并发过高？后来我查看了一下网站的搜索引擎抓取网站的日志，发现每分钟都有大量的页面被搜索引擎抓取！难怪网站的并发过高！！

但是大家都知道搜索引擎收录网站对我们来说是件好事，我们不能禁止掉所有的搜索引擎抓取，所以可以设置一些爬取规则来限制。通过我流量来源分析，每天从百度，google来的流量很多，而其他的搜索引擎几乎没流量导入，像这些没有带来流量的搜索引擎我是可以屏蔽掉。我想到的第一种方法就是在网站根目录写robots.txt文件:

```
User-agent: Googlebot
Disallow: /wp-
Allow: /wp-content/uploads/
Disallow: /?
Disallow: /feed
Disallow: /*/*/feed
Disallow: /trackback
Disallow: /*/*/trackback
Disallow: /*.php$
Disallow: /*.css$

User-agent: Baiduspider
Disallow: /wp-
Allow: /wp-content/uploads/
Disallow: /?
Disallow: /feed
Disallow: /*/*/feed
Disallow: /trackback
Disallow: /*/*/trackback
Disallow: /*.php$
Disallow: /*.css$

User-agent: *
Disallow: /
```

正规的搜索引擎一般会遵循robots.txt文件规范的，上文只允许百度、google爬取博客。但是总有那么些搜索引擎不遵循robots.txt文件规范，也就是说这样设置是没用的。那些搜索引擎照常

爬取你网站！不遵循robots.txt协议的代表：iAskSpider SohuAgent wget、OutfoxBot。之前我认为微软的必应搜索引擎应该会遵循robots.txt协议，但是我设置了上述robots.txt文件规范，居然还发现日志里面有大量的bingbot！

2014-11-13 17:38:14 157.55.39.39 /archives/1112/comment-page-2
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:37:09 157.55.39.39 /archives/928/comment-page-10
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:34:53 157.55.39.60 /archives/896
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:30:09 157.55.39.60 /archives/268
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:27:59 157.55.39.40 /archives/857
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:27:46 207.46.13.99 /archives/740/comment-page-1
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)
2014-11-13 17:25:51 157.55.39.60 /archives/category/hadoop/page/2
Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)

那么第二招限制搜索引擎爬取就是可以在你网站根目录写一个.htaccess，来限制：

```
SetEnvIfNoCase User-Agent "^Yisou" bad_bot
SetEnvIfNoCase User-Agent "^Easou" bad_bot
SetEnvIfNoCase User-Agent "^Youdao" bad_bot
SetEnvIfNoCase User-Agent "^msn" bad_bot
SetEnvIfNoCase User-Agent "^bingbot" bad_bot
Deny from env=bad_bot
```

这样可以在底层就限制了搜索引擎的爬取。

第三招限制搜索引擎的方法：很多网站服务器应该支持屏蔽某个IP，这种方法从效果上来说应该是最好的，从底层就限制了，但是这个方法有个弊端，那就是你得知道你需要屏蔽的IP地址，目前我博客已经屏蔽掉必应的部分IP，希望这些方法能够给网站减负！