Google
Developer
Day2009

使用Sitemap和Blog Ping让你的网站更容易被搜索

李钢江

2009年6月

Google Developer Day2009

内容简介

- Sitemap和Blog Ping简介
- 怎样生成Sitemap
- 怎样提交Sitemap
- 怎样使用Blog Ping
- 使用Google网站管理员工具查看网站情况

SITEMAP和BLOG PING简介

什么是Sitemap？

- Sitemap协议是一个公开的协议：
  - Google®, Microsoft®, Yahoo®在内的多个搜索引擎采用：
  - http://www.sitemaps.org

- 使用 Sitemap
  - 通知搜索引擎有哪些可供抓取的网页（但是搜索引擎并不保证索引所有的网页）
  - 提供网页的详细信息，例如更新频率，新闻日期，视频长度等

- 提交方式：robots.txt，Sitemap Ping，谷歌网站管理员工具等

Sitemap格式

- Sitepmap是一个 XML 文件，采用UTF-8编码:

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

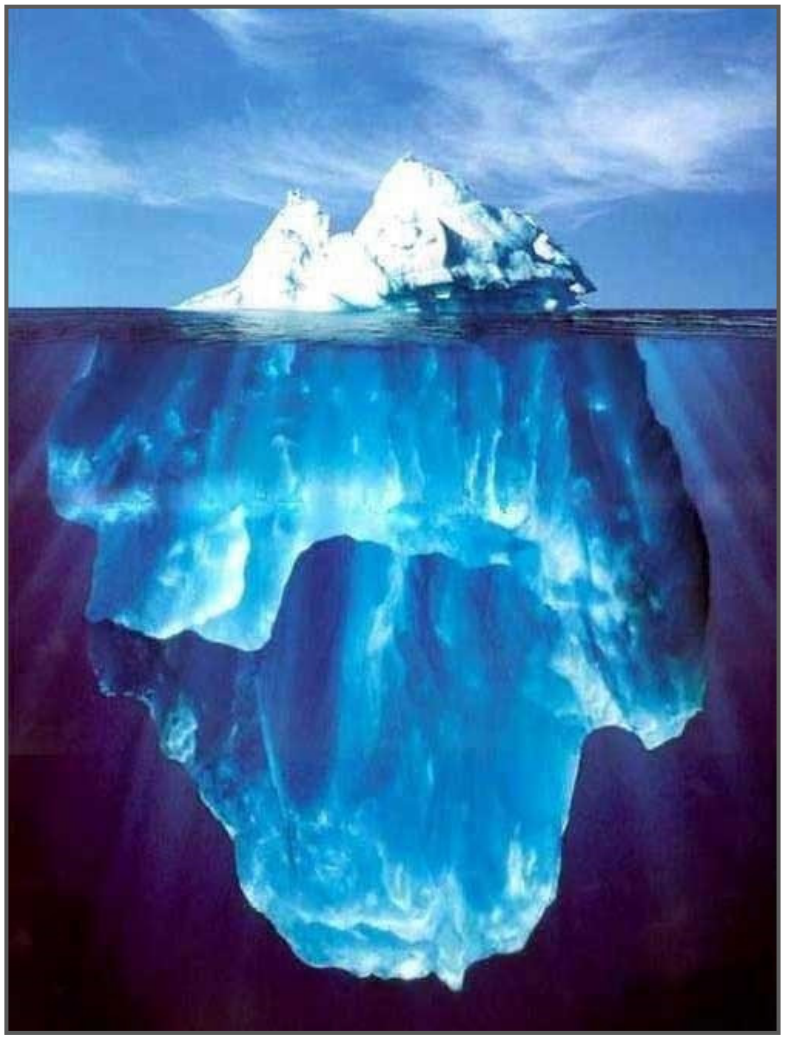
- Google对Sitemap的扩展：
  - -- 视频Sitemap
  - -- 移动Sitemap
  - -- 代码Sitemap
  - -- 新闻Sitemap
  - -- 地理Sitemap

什么是Blog Ping

- Blog Ping可以通知Google博客搜索引擎某一博客内容发生更新
  - 这些更新将被发布出去，并与其他的搜索引擎共享
- 适用于
  - 希望自己的博客内容能被Google的博客搜索中搜到的博客
  - 希望向其博客作者们提供Blog Ping服务的博客服务提供商
- 提交方式
  - XML-RPC Call
  - REST

为什么要使用Sitemap和Blog Ping？

- 互联网上能被搜索引擎发现的内容就像是冰山上的一角，更多的是：
  - 使用Javascript的动态页面
  - 非HTML类型的链接
  - 网页表单后面的“隐藏”链接
  - 没有被链接的网页
  - 爬虫爬取时遇到暂时的链接错误
- 让搜索引擎早日发现变化的内容



Google 垂直搜索使用Sitemap和Blog Ping

新闻内容

图片内容

新闻Sitemap

图片Sitemap

Google

Blog Ping

视频Sitemap

博客内容

视频内容

怎样生成SITEMAP

生成Sitemap的方法

使用工具自动生成Sitemap

- 优点：
  - 上手快
  - 格式准确
- 缺点:
  - 灵活性差
  - 覆盖率低
  - 无法自动生成一些元数据

自己编程生成Sitemap

- 优点：
  - 灵活性高
  - 覆盖率高
  - 包含更多的元数据
- 缺点：
  - 需要编程经验
  - 容易犯格式错误

使用Google Sitemap Generator生成Sitemap

- 适用环境：Windows + IIS, Linux + Apache
- 下载地址：http://code.google.com/p/googlesitemapgenerator
- 主要功能：
  - 通过Web服务器插件，监控新的网页地址
  - 自动计算Sitemap元数据
  - 自动提交Sitemap
  - 支持Web Sitemap、移动搜索Sitemap、代码搜索Sitemap和Blog Ping
  - 功能强大而且友好的配置界面

使用Java生成Sitemap

- SitemapGen4j 是一个用于创建Sitemap的开源Java库：http://code.google.com/p/sitemapgen4j/

```
WebSitemapGenerator wsg = new WebSitemapGenerator(
  "http://www.example.com", myDir);
// Configure the URL with lastmod=now, priority=1.0,
// changefreq=hourly
WebSitemapUrl url = new WebSitemapUrl.Options(
  "http://www.example.com/index.html").lastMod(new Dat
  e()).priority(1.0).changeFreq(ChangeFreq.HOURLY).bui
  ld();
wsg.addUrl(url);
wsg.write();
```

使用Java生成Sitemap索引文件

- 一个Sitemap最多包含5万个URL 。当URL数目超过5万个时，需要采用Sitemap索引文件

```
WebSitemapGenerator wsg = new WebSitemapGenerator(
  "http://www.example.com", myDir);
for (int i = 0; i &lt; 60000; i++)
  wsg.addUrl("http://www.example.com/doc"+i+".html");
// generate sitemap1.xml and sitemap2.xml
wsg.write();
// generate the sitemap_index.xml
wsg.writeSitemapsWithIndex();
```

常见问题

- 编码错误
  - 非法UTF-8字符
  - 没有转义：&, <, >, “, ‘
- Sitemap太大
  - 超过50000个URL
  - 超过10M字节（压缩前）
  - 索引文件包含>1000个Sitemap文件
- 其它
  - 提交次数太多
  - 日期格式错误：采用W3C编码
  - …

怎样提交SITEMAP

使用Http请求提交Sitemap

- Java代码：

```
String urlStr = String.format("http://www.google.com/w
  ebmasters/tools/ping?sitemap=",  URLEncoder.encode("
  http://example.com/sitemap.xml.gz", "utf-8"))
InputStream inputStream = new InputStreamReader(
  new URL(urlStr).openStream())
BufferedReader in = new BufferedReader(inputStream);
System.out.println(in.readLine()); // Check the result
in.close();
```

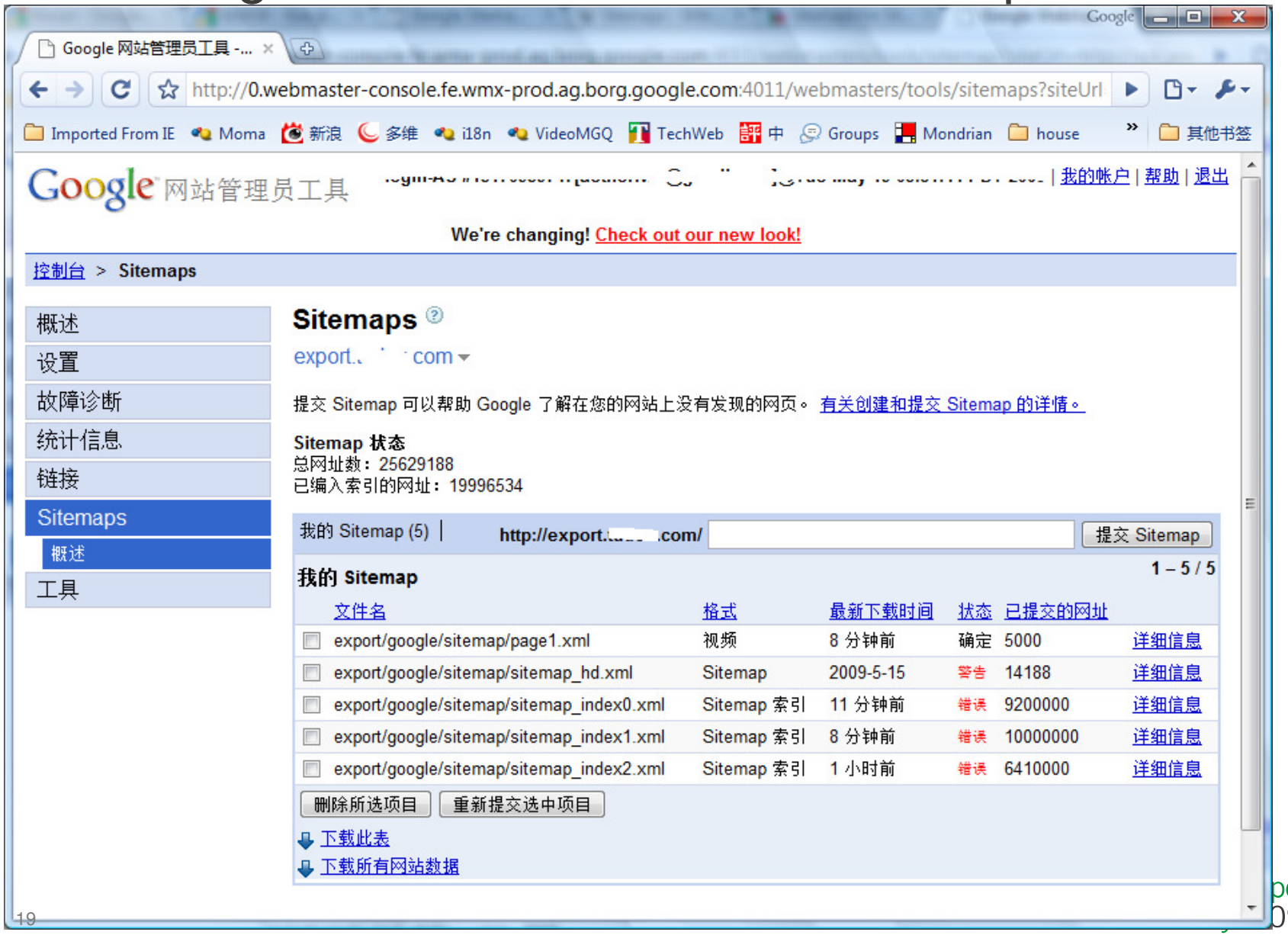
- Http服务器地址
  - Google: http://www.google.com/webmasters/tools/ping?sitemap=
  - Microsoft: http://webmaster.live.com/ping.aspx?siteMap=
  - Yahoo: http://search.yahooapis.com/SiteExplorerService/V1/ping?sitemap=
  - Ask.com: http://submissions.ask.com/ping?sitemap=

使用robots.txt提交Sitemap

- 在robots.txt文件中添加以下行

  Sitemap: http://example.com/Sitemap.xml
- 此指令不受 User-agent 行的影响
- 如果使用Sitemap 索引文件，则无需列出每个单独的 Sitemap

使用Google网站管理员工具提交Sitemap



怎样使用BLOG PING

通过XML-RPC请求发送Blog Ping

```
POST /RPC2 HTTP/1.0
User-Agent: request
Host: blogsearch.google.com
Content-Type: text/xml
Content-length: 447

<?xml version="1.0"?>
<methodCall>
 <methodName>weblogUpdates.
  extendedPing</methodName>
 <params>
  <param>
   <value>Official Google B
  log</value>
  </param>
  <param>
   <value>http://googleblog.
  blogspot.com/</value>
  </param>
  <param>
   <value>http://googleblog.
  blogspot.com/atom.xml</va
  lue>
  </param>
 </params>
</methodCall>
```

从Blog Ping服务器的XML-RPC应答

```
HTTP/1.1 200 OK
Connection: close
Content-Length: 451
Content-Type: text/xml
Date: Sun, 30 Sep 2001 20:0
  2:30 GMT
Server: Apache

<?xml version="1.0"?>
<methodResponse>
 <params>
  <param>
   <value>
    <struct>
     <member>
      <name>flerror</name>
      <value>
       <boolean>0</boolean>
      </value>
     </member>
     <member>
      <name>message</name>
     <value>Thanks for the
  ping.</value>
     </member>
    </struct>
   </value>
  </param>
 </params>
</methodResponse>
```

使用Apache XML-RPC开源Java库发送Blog Ping

- Java库地址：http://ws.apache.org/xmlrpc/index.html

```
// Build RPC Client
XmlRpcClientConfigImpl config = new XmlRpcClientConfigImpl();
config.setServerURL(new URL("http://blogsearch.google.com/ping/
  RPC2"));
XmlRpcClient client = new XmlRpcClient();
client.setConfig(config);
// Build request param list
String siteName = "Official Google Blog";
String siteUrl = "http://googleblog.blogspot.com/";
String pageUrl = "http://googleblog.blogspot.com/";
String rssUrl = "http://googleblog.blogspot.com/atom.xml"
String tags = "IT|News"; // Optional, '|' seperated tags.
Object[] params = new Object[]{siteName, siteUrl, pageUrl, rssU
  rl, tags};  // Note order!
// Send request and get result
Map result = (Map)client.execute("weblogUpdates.extendedPing",
  params);
boolean sucess = !result.get("flerror");
String message = result.get("message");
```

使用REST发送Blog Ping

- 地址：http://blogsearch.google.com/ping
- 参数:
  - name:博客名称
  - url：博客地址
  - changesURL：博客RSS, RDF, ATOM Feed的地址
- 返回值：
  - 成功：Thanks for the ping.
  - 失败：错误消息
- 示例：

  http://blogsearch.google.com/ping?name=Official+Google+Blog&url=http%3A%2F%2Fgoogleblog.blogspot.com%2F&changesURL=http%3A%2F%2Fgoogleblog.blogspot.com%2Fatom.xml

使用标准Java库发送REST Blog Ping

```
// Construct the url and encode it properly.
String urlStr = String.format("http://blogsearch.google.com/pin
  g?name=%s&url=%s&changesURL=%s",
  URLEncoder.encode("Official Google Blog", "utf-8"),
  URLEncoder.encode("http://googleblog.blogspot.com", "utf-8"),
  URLEncoder.encode("http://googleblog.blogspot.com/atom.xml",
  "utf-8"))

// Send request and get result
URL url = new URL(urlStr);
BufferedReader in = new BufferedReader(new InputStreamReader(ur
  l.openStream()));
String message = in.readLine();
in.close();

// Check the result message.
final String successMsg = "Thanks for the ping."
if (successMsg.equals(message)) {
  // success!
} else {
  System.err.println(message);
}
```

检查Blog Ping发布结果

- 访问 http://blogsearch.google.com/changes.xml 查看 Blog Ping 结果
  - 包含最近几分钟Google收集的Blog Ping，可能需要几个小时的下载时间

```
<weblogUpdates version="2" updated="Wed, 30 May 2006 14:10:00
  GMT" count="1384779">
  <weblog name="Some Blog"
          url="http://googleblog.blogspot.com"
          rssUrl="http://googleblog.blogspot.com/atom.xml"
          when="1"/>
  ...
</weblogUpdates>
```

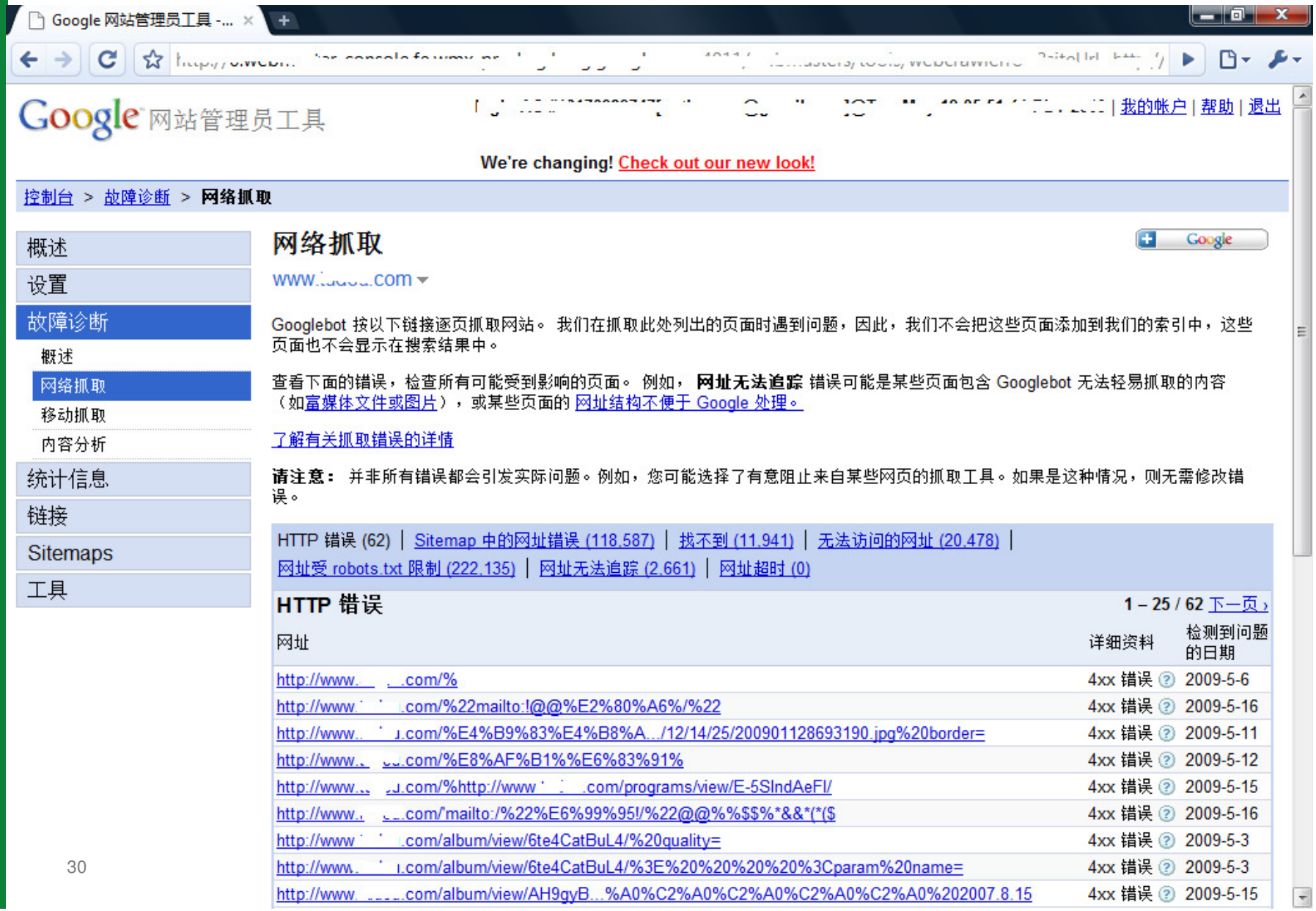
使用GOOGLE网站管理员工具查看网站情况

Google网站管理员工具简介

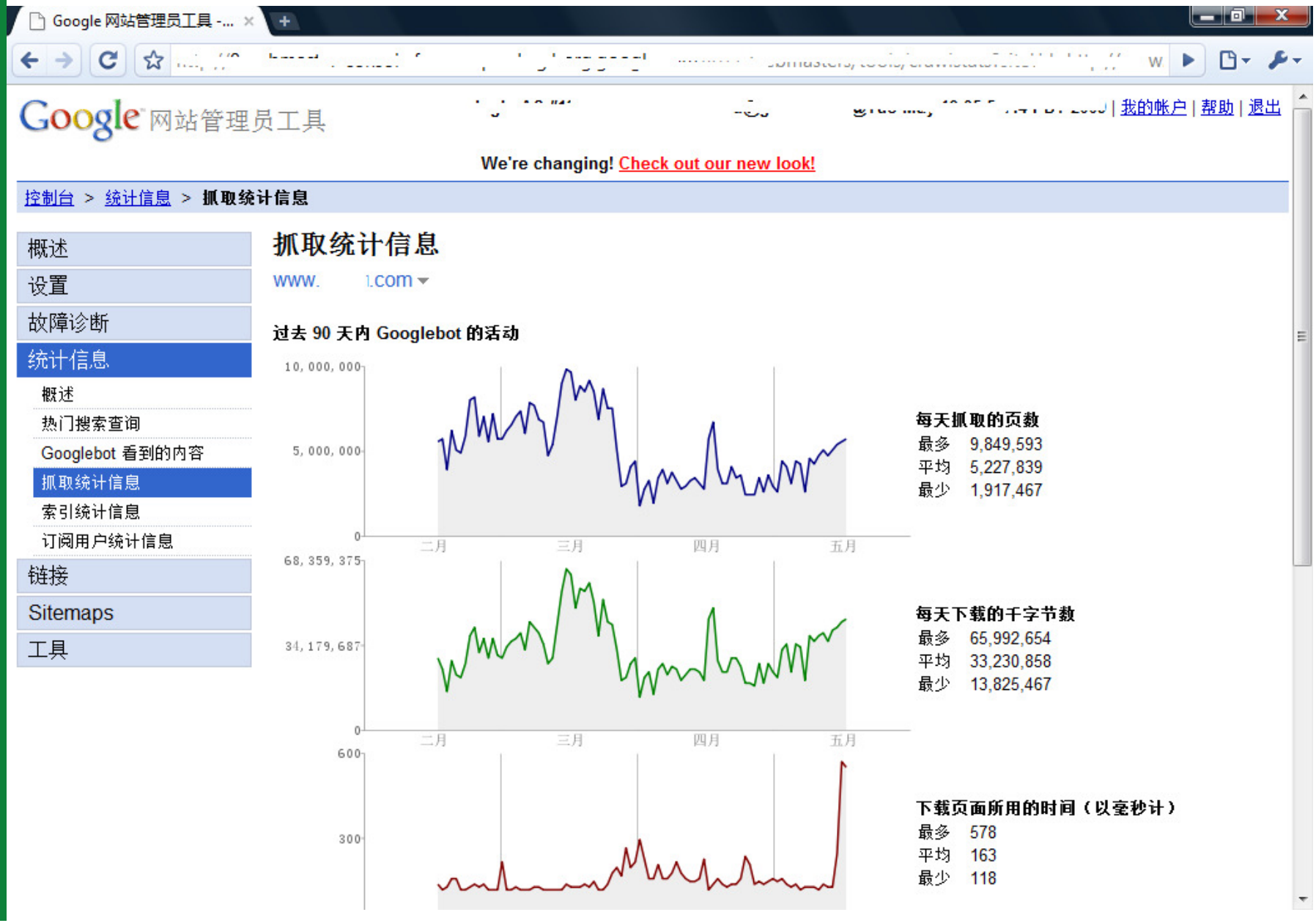
怎样验证网站

- 通过添加新文件
  - 404页面需要返回一个错误响应
- 通过在Homepage添加元数据

查看网络抓取情况


Google 网站管理员工具 -...
Google 网站管理员工具
我的帐户 | 帮助 | 退出
We're changing! Check out our new look!
控制台 > 故障诊断 > 网络抓取
概述
设置
故障诊断
概述
网络抓取
移动抓取
内容分析
统计信息
链接
Sitemaps
工具
网络抓取
Google
Googlebot 按以下链接逐页抓取网站。 我们在抓取此处列出的页面时遇到问题，因此，我们不会把这些页面添加到我们的索引中，这些页面也不会显示在搜索结果中。
查看下面的错误，检查所有可能受到影响的页面。 例如， 网址无法追踪 错误可能是某些页面包含 Googlebot 无法轻易抓取的内容（如富媒体文件或图片），或某些页面的 网址结构不便于 Google 处理。
了解有关抓取错误的详情
请注意： 并非所有错误都会引发实际问题。例如，您可能选择了有意阻止来自某些网页的抓取工具。如果是这种情况，则无需修改错误。
HTTP 错误 (62) | Sitemap 中的网址错误 (118,587) | 找不到 (11,941) | 无法访问的网址 (20,478) | 网址受 robots.txt 限制 (222,135) | 网址无法追踪 (2,661) | 网址超时 (0)
HTTP 错误
1 – 25 / 62 下一页 ›
网址
详细资料
检测到问题的日期
http://www. .com/% 4xx 错误 2009-5-6
http://www. .com/%22mailto:!@@%E2%80%A6%/%22 4xx 错误 2009-5-16
http://www. .com/%E4%B9%83%E4%B8%A.../12/14/25/200901128693190.jpg%20border= 4xx 错误 2009-5-11
http://www. .com/%E8%AF%B1%%E6%83%91% 4xx 错误 2009-5-12
http://www. .com/%http://www. .com/programs/view/E-5SlndAeFl/ 4xx 错误 2009-5-15
http://www. .com/"mailto:/%22%E6%99%95!/%22@@%%$$%*&&*(*($ 4xx 错误 2009-5-16
http://www. .com/album/view/6te4CatBuL4/%20quality= 4xx 错误 2009-5-3
http://www. .com/album/view/6te4CatBuL4/%3E%20%20%20%20%3Cparam%20name= 4xx 错误 2009-5-3
http://www. .com/album/view/AH9gyB...%A0%C2%A0%C2%A0%C2%A0%C2%A0%202007.8.15 4xx 错误 2009-5-15

查看网络抓取统计


Google 网站管理员工具 -...
Google 网站管理员工具
我的帐户 | 帮助 | 退出
We're changing! Check out our new look!
控制台 > 统计信息 > 抓取统计信息
概述
设置
故障诊断
统计信息
概述
热门搜索查询
Googlebot 看到的内容
抓取统计信息
索引统计信息
订阅用户统计信息
链接
Sitemaps
工具
抓取统计信息
www. .com
过去 90 天内 Googlebot 的活动
10,000,000
5,000,000
0
二月
三月
四月
五月
每天抓取的页数
最多 9,849,593
平均 5,227,839
最少 1,917,467
68,359,375
34,179,687
0
每天下载的千字节数
最多 65,992,654
平均 33,230,858
最少 13,825,467
600
300
下载页面所用的时间（以毫秒计）
最多 578
平均 163
最少 118

问答

更多信息，访问**code.google.com**

Google
Developer
Day2009