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



## 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
- 6 -- 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 Developer Day2009

#### Google 垂直搜索使用Sitemap和Blog Ping



9

## 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 < 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?siteM
     ap=
  - Yahoo: http://search.yahooapis.com/SiteExplorerServi
     ce/V1/ping?sitemap=
  - Ask.com: http://submissions.ask.com/ping?sitemap=Google

Google Developer Day2009

## 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

							_		
🗋 Google 网站管理员工具 >	× (+)	and states in the second second	and the second second second	and the second	Sile of the second		_		
← → C ☆ http://0.v	webmaster-console.fe.wmx-prod.ag.borg.googl	e.com:4011/we	ebmasters/tools	s/sitem	aps?siteUr	🕨 🗗 🔑	-		
🗎 Imported From IE 🛛 🔌 Moma	🙋 新浪 🌜 多维 🍬 i18n 🥠 VideoMGQ 👖 Tec	hWeb <mark>評</mark> 中 🔗	🛛 Groups 🛛 📕 Mo	ndrian	🚞 house	» 🗀 其他书绘	EN I		
Google <sup>-</sup> 网站管理	员工具 ··· <b>y</b> ···· <del>·</del> ··············· C	•••ى " ،			<u>我的</u> /	<u>帐户 帮助 退出</u>	-		
We're changing! <u>Check out our new look!</u>									
<u>控制台</u> > Sitemaps									
HBT 2-4	Sitemans (?)								
城还	ovport i com-								
设直	export. com •								
故障诊断	提交 Sitemap 可以帮助 Google 了解在您的网站上涨	没有发现的网页。	<u>有关创建和提交</u>	Sitemap	的详情。				
统计信息。	Citoman 壮太								
	Sitemap 10.00								
链接	Sitemap 4.33 总网址数: 25629188 已编入索引的网址: 19996534								
链接 Sitemaps	总网址数: 25629188 已编入索引的网址: 19996534						I		
链接 Sitemaps <sup>要试</sup>	总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export.ucom	m/				提交 Sitemap	ш		
链接 Sitemaps 概述 工具	总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export.ucom 我的 Sitemap	m/				提 <u>交 Sitemap</u> 1 – 5 / 5	m		
链接 Sitemaps <sup>概述</sup> 工具	总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export	m/	<u>最新下载时间</u>	<u>状态</u> i	 	提 <u>交 Sitemap</u> 1 – 5 / 5 止	ш		
链接 Sitemaps 概述 工具	<ul> <li>         Sitemap (A)         <ul> <li></li></ul></li></ul>	m/ <u>格式</u> 视频	<u>最新下载时间</u> 8 分钟前	<u>状态</u> [ 确定 5		提 <u>交 Sitemap</u> 1 – 5 / 5 止 <u>详细信息</u>	II		
链接 Sitemaps 概述 工具	Sitemap 4.33 总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export.to	m/ <u>格式</u> 视频 Sitemap	<u>最新下载时间</u> 8 分钟前 2009-5-15	<u>状态</u> 确定 5 警告 1	<u>二提交的网</u> 切 5000 14188	提 <u>交 Sitemap</u> 1 – 5 / 5 止 <u>详细信息</u> <u>详细信息</u>	E		
链接 Sitemaps 概述 工具	Sitemap 4Xas 总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export	m/ 格式 视频 Sitemap Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前	<u>状态</u> j 确定 5 错误 5	<u>二提交的网</u> 址 5000 14188 9200000	提交 Sitemap 1-5/5 止 <u>详细信息</u> <u>详细信息</u> <u>详细信息</u>	B		
链接 Sitemaps 概述 工具	Sitemap 4X8 总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export.com/sitemap 文件名 export/google/sitemap/page1.xml export/google/sitemap/sitemap_hd.xml export/google/sitemap/sitemap_index0.xml export/google/sitemap/sitemap_index1.xml	m/ 格式 视频 Sitemap Sitemap 索引 Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前 8 分钟前	<u>状态</u> 确定 5 错误 1 错误 1	<u>二提交的网</u> 址 5000 14188 9200000 10000000	提交 Sitemap 1 – 5 / 5 止 <u>详细信息</u> <u>详细信息</u> <u>详细信息</u> <u>详细信息</u>	E		
链接 Sitemaps <sup>概述</sup> 工具	Sitemap 4X8 总网址数: 25629188 已编入索引的网址: 19996534 我的 Sitemap (5) http://export.t	m/ 格式 视频 Sitemap Sitemap 索引 Sitemap 索引 Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前 8 分钟前 1 小时前	<u>状态</u> [ 确定 5 答告 1 错误 1 错误 6	<u>二提交的网</u> 址 5000 14188 9200000 10000000 5410000	提交 Sitemap 1-5/5 止 <u>详细信息</u> <u>详细信息</u> <u>详细信息</u> <u>详细信息</u> <u>详细信息</u>	H		
链接 Sitemaps <sup>概述</sup> 工具	Sitemap 4Xas         总网址数: 25629188         已编入索引的网址: 19996534         我的 Sitemap (5)         http://export.to	m/ 格式 视频 Sitemap Sitemap 索引 Sitemap 索引 Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前 8 分钟前 1 小时前	<u>状态</u> [ 确定 9 错误 1 错误 1 错误 6	<u>二提交的网</u> 地 5000 14188 9200000 10000000 5410000	提交 Sitemap 1-5/5 止 <u>详细信息</u> <u>详细信息</u> <u>详细信息</u> <u>详细信息</u>	H		
链接 Sitemaps 概述 工具	Sitemap vxs         总网址数: 25629188         已编入索引的网址: 19996534         我的 Sitemap (5)         http://export.t	m/ 格式 视频 Sitemap Sitemap 索引 Sitemap 索引 Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前 8 分钟前 1 小时前	<u>状态</u> 确定 5 错误 1 错误 1 错误 6	<u>二提交的网</u> 切 5000 14188 9200000 10000000 5410000	提交 Sitemap 1-5/5 止 详细信息 详细信息 详细信息 详细信息 详细信息	E		
链接 Sitemaps 概述 工具	Sitemap vxs         总网址数: 25629188         已编入索引的网址: 19996534         我的 Sitemap         文件名         export/google/sitemap/page1.xml         export/google/sitemap/sitemap_hd.xml         export/google/sitemap/sitemap_index0.xml         export/google/sitemap/sitemap_index1.xml         export/google/sitemap/sitemap_index1.xml         export/google/sitemap/sitemap_index2.xml         删除所选项目       重新提交选中项目         季 下载此表         季 下载所有网站数据	m/ 格式 视频 Sitemap Sitemap 索引 Sitemap 索引 Sitemap 索引	<u>最新下载时间</u> 8 分钟前 2009-5-15 11 分钟前 8 分钟前 1 小时前	<u>状态</u> [ 第二章 [ 第二章 [ 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	<ul> <li>         - 提交的网は        </li></ul>	提交 Sitemap 1-5/5 止 <u>详细信息</u> <u>详细信息</u> <u>详细信息</u> <u>详细信息</u>	m		

## 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

<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>
<value>

<boolean>0</boolean> </value> </member> <member> <name>message</name> <value>Thanks for the ping.</value> </member> </struct> </value> </param> </params> </methodResponse>

> Google Developer Day2009

## 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&changesUR L=http%3A%2F%2Fgoogleblog.blogspot.com%2Fatom.xml

> Google Developer Day2009

## 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);
                                                               Google
                                                               Developer
}
                                                               Dav2009
25
```

## 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>

. . .

## 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

	Through I for strapstray, then it	CONTRACTOR INCOME.	eren Maren - Marcard Rouse Press	Google 🗖 🗖 📈					
🗋 Google 网站管理员工具 🗙	¢	Annual Contract							
← → C ☆ http://C		argana ogicio	·····	,**** 🕨 🗗 🖋					
Google 网站管理员工具 "」 **									
We're changing! <u>Check out our new look!</u>									
<u>控制台</u> > 概述									
概述	概述								
设置	www.ia Ja.com 🔻								
故障诊断	编制索引   热门搜索查询 »								
统计信息	索引状态:	✓ 您网站的网页包含在	Google 索引中。 请参阅 <u>索引统计</u>						
链接		您尚未提交任何 Sitemap。 提交 Sitemap 可帮助 Google 找到抓取工具可能							
Sitemaps		A 我不到的两页。 创建 关于 Google 如何将ì	并提交列有您网站网班的 Sitemap 这些网页编入索引的相关数据。 <u>详</u>						
工具									
	网络抓取错误		上》	R更新日期 2009-5-18					
	HTTP 错误	<u>▲</u> 62	<u>详情 »</u>						
	Sitemap 中的网址错误	118,587	<u>详情 »</u>						
	<u>找不到</u>	11,941	<u>详情 »</u>						
	无法访问的网址	20,478	<u>详情 »</u>						
	<u> 网址受 robots.txt 限制</u>	A 222,135	<u>详情 »</u>						
	网址无法追踪	<b>1</b> 2,661	<u>详情 »</u>						
	<u>网址超时</u>	<b>√</b> 0							
	合计:	375,864							

#### 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

C Google 网站管理员工具 ×	+						
← → C ☆ hupg/our		11el 1447					
Google <sup>-</sup> 网站管理员	员工具 (Ji Li Li Contra Contra Li Contra Li Contra Co						
We're changing! Check out our new look!							
<u>控制台</u> > <u>故障诊断</u> > <b>网络抓取</b>							
概述	网络抓取	E Google					
设置	WWW.lacos.com 🔻						
故障诊断 概述	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 页面也不会显示在搜索结果中。	加到我们的索引中,这些					
网络抓取 移动抓取	查看下面的错误,检查所有可能受到影响的页面。 例如, <b>阿址无法追踪</b> 错误可能是某些页面包含 Googlebot (如 <u>富媒体文件或图片</u> ),或某些页面的 <u>网址结构不便于 Google 处理。</u>	无法轻易抓取的内容					
内容分析	了解有关抓取错误的详情						
统计信息	<b>请注意:</b> 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 误。	<b>《种情况,则无需修改错</b>					
链接							
Sitemaps	HTTP 错误 (62)   <u>Sitemap 中的网址错误 (118,587)</u>   <u>技不到 (11,941)</u>   <u>无法访问的网址 (20,478)</u>   网址受 robots txt 限制 (222,135)   网址无法追踪 (2,661)   网址超时 (0)						
工具		1_25/62下一页。					
	网址	详细资料 检测到问题 详细资料 的日期					
	http://wwwcom/%	4xx 错误 ② 2009-5-6					
	http://www.icom/%22mailto:!@@%E2%80%A6%/%22	4xx 错误 ② 2009-5-16					
	http://wwwJ.com/%E4%B9%83%E4%B8%A/12/14/25/200901128693190.jpg%20border=	4xx 错误 ② 2009-5-11					
	http://wwwuu.com/%E8%AF%B1%%E6%83%91%	4xx 错误 ② 2009-5-12					
	http://wwwJ.com/%http://www.'com/programs/view/E-5SIndAeFI/	4xx 错误 ② 2009-5-15					
	http://www.rcom/mailto:/%22%E6%99%95!/%22@@%%\$\$%*&&*(*(\$	4xx 错误 ② 2009-5-16					
	http://www i com/album/view/6te4CatBuL4/%20quality=	4xx 错误 ② 2009-5-3					
30	http://www.i.com/album/view/6te4CatBuL4/%3E%20%20%20%20%3Cparam%20name=	4xx 错误 ② 2009-5-3					
	http://wwwcom/album/view/AH9gyB%A0%C2%A0%C2%A0%C2%A0%C2%A0%202007.8.15	4xx 错误 ② 2009-5-15					

#### 查看网络抓取统计



# 问答

## 更多信息,访问code.google.com



Google Developer Day2009